

The Jackson Laboratory

## The Mouseion at the JAXlibrary

---

Faculty Research 2020

Faculty Research

---

6-22-2020

### **Approaches for integrating heterogeneous RNA-seq data reveal cross-talk between microbes and genes in asthmatic patients.**

Daniel Spakowicz

Shaoke Lou

Brian Barron

Jose L Gomez

Tianxiao Li

*See next page for additional authors*

Follow this and additional works at: <https://mouseion.jax.org/stfb2020>

---

---

**Authors**

Daniel Spakowicz, Shaoke Lou, Brian Barron, Jose L Gomez, Tianxiao Li, Qing Liu, Nicole Grant, Xiting Yan, Rebecca Hoyd, George M. Weinstock, Geoffrey L Chupp, and Mark Gerstein

---

METHOD

Open Access



# Approaches for integrating heterogeneous RNA-seq data reveal cross-talk between microbes and genes in asthmatic patients

Daniel Spakowicz<sup>1,2,3,4†</sup>, Shaoke Lou<sup>1†</sup>, Brian Barron<sup>1</sup>, Jose L. Gomez<sup>5</sup>, Tianxiao Li<sup>1</sup>, Qing Liu<sup>5</sup>, Nicole Grant<sup>5</sup>, Xiting Yan<sup>5</sup>, Rebecca Hoyd<sup>3</sup>, George Weinstock<sup>2</sup>, Geoffrey L. Chupp<sup>5</sup> and Mark Gerstein<sup>1,6,7,8\*</sup> 

\* Correspondence: [mark@gersteinlab.org](mailto:mark@gersteinlab.org)

<sup>†</sup>Daniel Spakowicz and Shaoke Lou contributed equally to this work.

<sup>1</sup>Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA

<sup>6</sup>Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA

Full list of author information is available at the end of the article

## Abstract

Sputum induction is a non-invasive method to evaluate the airway environment, particularly for asthma. RNA sequencing (RNA-seq) of sputum samples can be challenging to interpret due to the complex and heterogeneous mixtures of human cells and exogenous (microbial) material. In this study, we develop a pipeline that integrates dimensionality reduction and statistical modeling to grapple with the heterogeneity. LDA(Latent Dirichlet allocation)-link connects microbes to genes using reduced-dimensionality LDA topics. We validate our method with single-cell RNA-seq and microscopy and then apply it to the sputum of asthmatic patients to find known and novel relationships between microbes and genes.

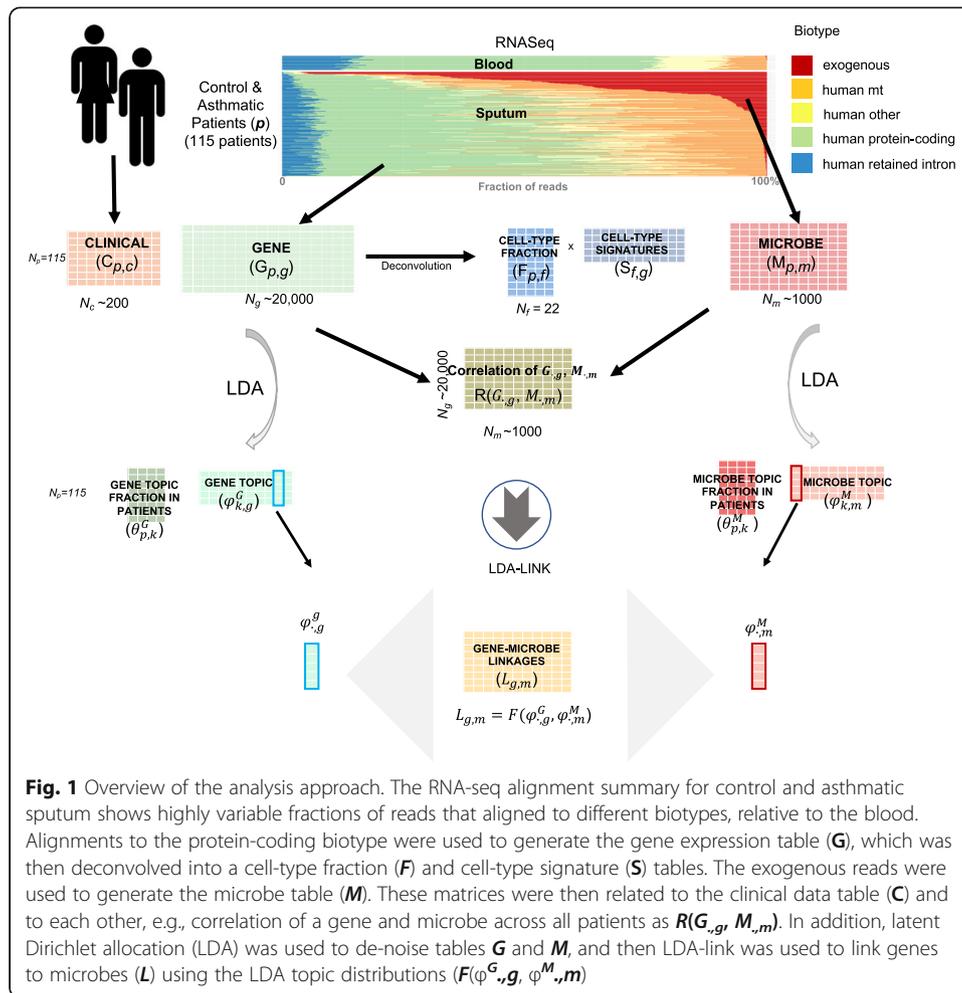
## Introduction

### Linking high-dimensional, heterogeneous datasets

RNA sequencing (RNA-seq) has become a standard method of analyzing complex communities. Depending on the sample type, these data can be very heterogeneous. A key problem tackled in this paper is dealing with the heterogeneity and noise in RNA-seq data in samples such as sputum. This can be appreciated by comparing sputum RNA-seq to a more traditional experiment, e.g., blood RNA-seq, where the sample can be collected consistently and that contains relatively well-defined cell types (Fig. 1). In the blood, the vast majority of RNA-seq reads align to the human genome, and the goal is often to relate the expression of the genes to a phenotype. By contrast, sputum may be less consistently collected, its cell types are less well-defined, and it may contain RNA from microbes and other organisms that act as cryptic indicators of the environment. But within this complexity is an opportunity—the interactions between immune cells and microbes may provide clinically meaningful information from a non-invasive sample. Moreover, generating RNA-seq data from the sputum can be technically easier than other methods to study microbe and immune cell interactions, such as by culturing sputum microbes or by flow sorting sputum immune cells. Here, we



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.



**Fig. 1** Overview of the analysis approach. The RNA-seq alignment summary for control and asthmatic sputum shows highly variable fractions of reads that aligned to different biotypes, relative to the blood. Alignments to the protein-coding biotype were used to generate the gene expression table ( $G$ ), which was then deconvolved into a cell-type fraction ( $F$ ) and cell-type signature ( $S$ ) tables. The exogenous reads were used to generate the microbe table ( $M$ ). These matrices were then related to the clinical data table ( $C$ ) and to each other, e.g., correlation of a gene and microbe across all patients as  $R(G_{p,g}, M_{p,m})$ . In addition, latent Dirichlet allocation (LDA) was used to de-noise tables  $G$  and  $M$ , and then LDA-link was used to link genes to microbes ( $L$ ) using the LDA topic distributions ( $F(\varphi_{p,g}^G, \varphi_{p,m}^M)$ )

present a strategy for dealing with the complexity that uses a number of supervised and unsupervised techniques such as cell-type signatures and latent Dirichlet allocation (LDA). These techniques can produce a low-dimensional representation of common groups of genes, microbes, or other features that tend to increase or decrease in abundance together. Our approach is useful when the heterogeneity comes from the sample type (e.g., sputum) and especially when the samples derive from a heterogeneous population of individuals, such as patients with asthma.

### Interactions between the host and microbes in the lung

Asthma is a disease of the airway that can present with diverse clinical phenotypes. Much work has focused on identifying subgroups of the disease and how each subgroup responds to treatment. For example, Yan et al. introduced transcriptional endotypes of asthma, and the Severe Asthma Research Program defined five subtypes of asthma [1, 2]. Some of these subgroups respond differently to environmental and microbial triggers, such as fungal spores. Some fungi have well-defined effects on asthma, but the role of many microbes remains uncertain. A simplified model assigns microbes

to one of three categories: pathogenic organisms that cause inflammation, beneficial organisms that reduce inflammation, and commensal organisms that have no effect on inflammation. The majority of the organisms in the lungs are expected to have no effect, and severe asthmatics are expected to have more pathogenic and fewer beneficial microbes. However, the reality is more complex; organisms can interact with the lung in myriad ways, such as through innate responses to pathogens, small-molecule interactions, and others [3, 4].

#### **Inferring immune cell fractions from RNA-seq data**

The pathology of microbes is often inferred by the number and type of immune cells observed in samples, such as sputum total leukocyte counts [5, 6]. A standard method for counting immune cells in sputum samples uses microscopy, but the resolution is limited to a few cell types [7]. Other cell-counting methods such as flow sorting can be challenging because of the viscosity and highly variable cell numbers in sputum. An alternative strategy uses cell type-specific expression patterns to deconvolve RNA-seq reads from mixtures of cells into fractions of different immune cells [8]. This deconvolution also effectively de-noises heterogeneous datasets by greatly reducing the number of dimensions. Importantly, the RNA needed for this analysis can be purified without poly-A enrichment—here, we use human ribosomal RNA depletion—which allows for the simultaneous analysis of microbial and human transcripts.

#### **Supervised deconvolution and the microbiome**

While deconvolution to cell fractions effectively de-noises human RNA-seq data, an equivalent supervised dimensionality reduction method does not exist for microbes. One can reduce the dimensions by collapsing microbial strains to different taxonomic ranks (e.g., genus or family); however, taxonomy is notoriously imprecise at defining behavior. For example, many bacteria in the genus *Escherichia* are human commensals, whereas *Escherichia coli* OH157:H7 causes hemorrhagic colitis. Alternatively, one can group sequences by the metabolic pathways observed, although this requires high-depth sequencing. Here, we propose a method to reduce the dimensionality of microbes by first linking the microbes to human genes and then applying supervised gene dimensionality-reduction methods (e.g., deconvolution to cell types).

#### **Unsupervised decomposition including LDA**

A number of algorithms have been developed to infer higher-order structures within high-dimensional datasets. Latent Dirichlet allocation (LDA) was first used in text mining applications to learn the set of topics in documents and has since been used in a broad range of applications including marketing, genomics, drug-pathway relationships, and others [9–12]. Unlike some other unsupervised decomposition methods (e.g., principal component analysis (PCA), non-negative matrix factorization (NMF)), the result is not a linear decomposition. The topic distributions from LDA are governed by a series of hyper-parameters that define distributions from which the “words” (in this case, genes or microbes) are drawn. This non-linear character has advantages in complex, sparse, or noisy datasets and RNA-seq such as sputum RNA-seq or single-cell RNA-seq.

In this paper, we use RNA-seq of sputum samples from asthmatic patients to demonstrate dimensionality-reduction strategies and identify microbe-host relationships. We map RNA-seq reads onto human or microbial genomes and relate the resulting abundance matrices to each other and to clinical data. Further, we deconvolve the human reads into fractions of the various cell types that make up the sputum. Finally, we relate the human genes and microbes using a method we call the LDA-link, which identifies relationships between genes, microbes, and cell types. These methods represent a general strategy for dealing with heterogeneous RNA-seq data that is applicable to other sample types beyond sputum.

## Results

### Sequencing and processing with the extracellular RNA processing toolkit (exceRpt) pipeline

We collected induced sputum samples from 115 patients with heterogeneous asthma phenotypes and sequenced these samples using RNA-seq. The median read depth per sample was 47.5 million, which meets the depth recommendations for analyses of this type [13]. We processed these reads through the exceRpt pipeline [14], which conservatively matches reads to genomes in a sequential order designed to reduce experimental artifacts. In brief, we first aligned the quality-filtered reads to the UniVec database of common laboratory contaminants [3] and then aligned the remaining reads to human ribosomal sequences before aligning them to the human genome. We excluded samples with a low ratio of transcript alignments to intergenic sequence alignments and then aligned the remaining reads to the comparably large sequence space of non-human genomes. We first aligned reads to the relatively well-curated ribosomal databases of bacteria, fungi, and archaea (e.g., Ribosomal Database Project [4]) and then to curated genomes of bacteria, fungi, viruses, plants, and animals. The percent of reads mapping to different biotypes was highly heterogeneous; a median of 60% of the reads aligned to the human reference genome and 50% to annotated transcripts (Fig. 1, green bars). A median of 0.7% of the input reads aligned to exogenous sources, with some samples containing as much as 28.1% exogenous reads. As a control, we applied the same protocol to blood samples, which demonstrated more homogeneity than sputum (Fig. 1, top, “blood”).

### Overview of the analysis approach

The goal of the analysis was to infer meaningful relationships between the numbers and origins of the RNA-seq reads and relate them to clinical phenotypes. We conceptualized the clinical information and RNA-seq alignments as a series of tables (Fig. 1). The clinical table includes patient data collected at the clinic,  $C$ , including age, weight, and lung function tests, with rows indexed by patient ( $N_p = 115$ ) and roughly 200 clinical variables ( $N_c$ ). Alignments to human protein-coding regions created the gene table,  $G$ , with  $N_p$  rows, as above, and roughly 20,000 genes ( $N_g$ ). Alignments to exogenous genomes created the microbe table ( $M$ ) with  $N_p$  rows and roughly 1000 microbes ( $N_m$ ). Given these three tables ( $C$ ,  $G$ , and  $M$ ), the basic analysis framework is to correlate columns or rows within or between tables. We represent this by a matrix of correlations,  $R(X_{,i}, Y_{,j})$ , where  $X_{,i}$  is the  $i$ th column of table  $X$  and  $Y_{,j}$  is the  $j$ th column of table  $Y$ .

This correlation is summed over the other index, usually  $p$ . For example, we test the relationship between age and the abundance of each microbe  $\mathbf{R}(C_{\text{age}}, M_{\text{m}})$  across all patients. Similarly, we correlate the expression of a gene (e.g., *TLR4*) with microbe (e.g., *Candida*)  $\mathbf{R}(G_{\text{TLR4}}, M_{\text{Candida}})$ .

Individual correlations can be difficult to interpret, particularly in heterogeneous, sparse, or noisy datasets. Organizing the genes into relevant pathways or cell types can reduce the dimensionality and de-noise the analysis. To this end, we deconvolved the gene table ( $\mathbf{G}$ ,  $N_p$  by  $N_g$ ) into a cell-type fraction table ( $\mathbf{F}$ ,  $N_p$  by  $N_f$ ) and a cell-type signature table ( $\mathbf{S}$ ,  $N_f$  by  $N_g$ ). However, an analogous supervised method does not exist for the microbes. Therefore, we applied an unsupervised dimensionality-reduction approach, latent Dirichlet allocation (LDA), which provides topic distributions in patients ( $\theta^G$ ,  $N_p$  by  $N_k$ ) across a smaller number ( $k = 10$ ) of gene topic ( $\phi^G$ ,  $N_k$  by  $N_g$ ). This can also be done to the microbe table  $\mathbf{M}$  and get  $\theta^M$  and  $\phi^M$ , and the gene and microbe topic can be correlated (e.g.,  $\mathbf{R}(\theta_{\text{g}}^G, \theta_{\text{m}}^M)$  over all patients).

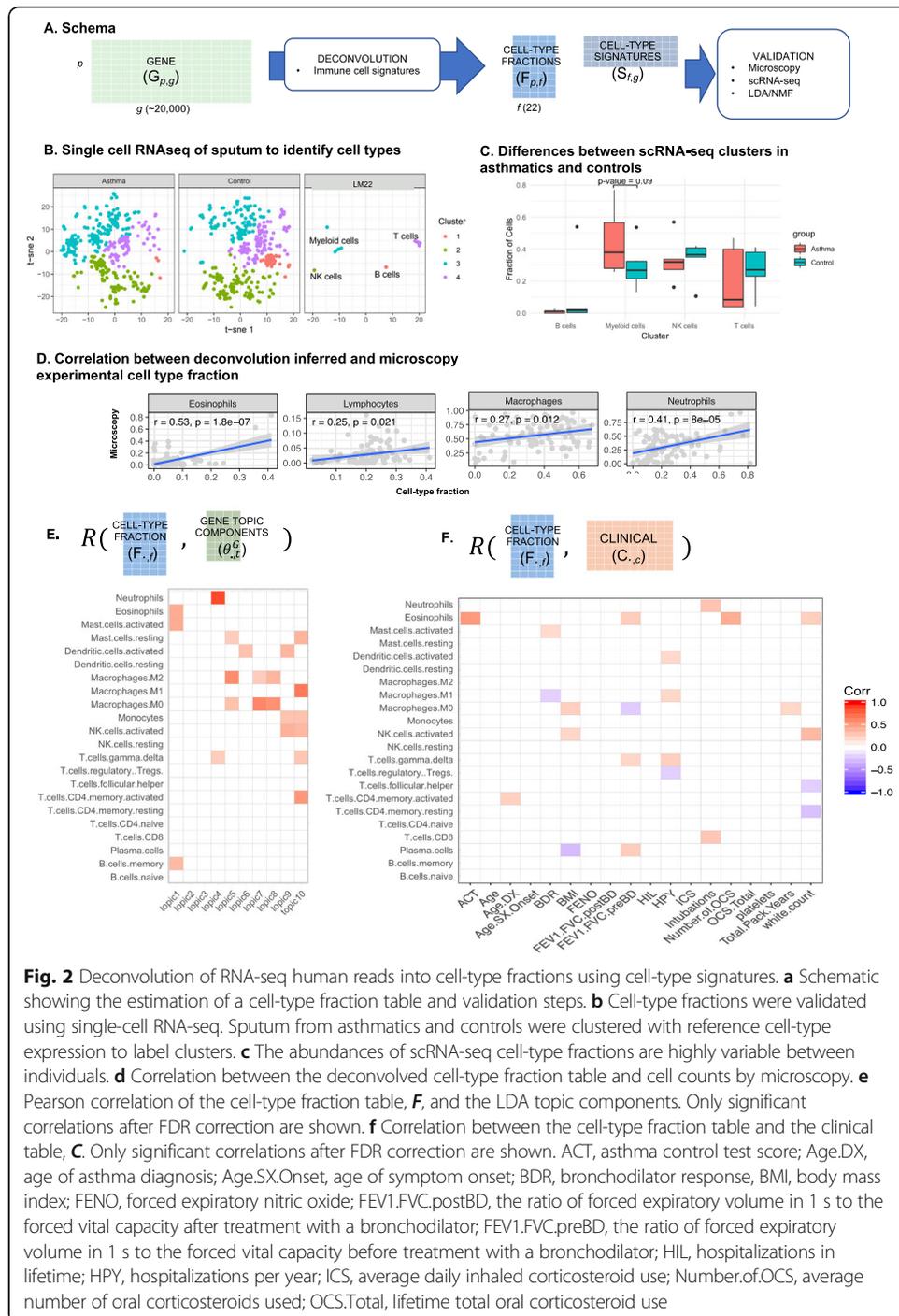
The framework described above is useful for identifying linear relationships, but non-linear relationships are also possible. For example, a microbe sensed by a human immune cell could lead to the activation of a transcription factor and the expression of several genes, each of which would have a non-linear relationship to microbe abundance. To identify such relationships, we applied a non-linear ensemble learning algorithm [15, 16], using the de-noised inputs for each gene and microbe ( $\phi^G$  and  $\phi^M$ ). We call this method the LDA-link. Further, we relate the gene and microbe links identified to cell fractions and thereby relate how the host is responding to microbes with regard to immune cell type response with a particular gene.

### Analysis of human-aligned reads

Working toward the hypothesis that we can conceptualize human-aligned sputum RNA-seq reads as a mixture of immune cell types, each with a distinct expression profile, we deconvolved the gene table ( $\mathbf{G}$ ) into a table of fractions of component cells type ( $\mathbf{F}$ ) and cognate cell-type signatures ( $\mathbf{S}$ ) by solving the formula  $\mathbf{G} = \mathbf{FS}$ . This method relies on knowing the signature gene set in each cell type, for which we used high-quality reference sets from experimentally isolated and sequenced circulating immune cells [8]. To validate that these circulating immune cells behaved similarly to those in sputum, we generated additional datasets including single-cell RNA-seq (scRNA-seq) and microscopy, and then compared the results to the deconvolution table  $\mathbf{F}$  and to unsupervised decomposition (Fig. 2a, schema).

### Evaluation of deconvolution results by scRNA-seq

First, we performed scRNA-seq on the sputum of a cohort of similar patients (five control and five asthmatics). The single-cell sequences clustered into four groups (Fig. 2b, first and second panels). In order to identify the clusters, we co-clustered the reference cell types with the scRNA-seq data (Fig. 2b, third panel); the reference cells grouped by lineage each with a different scRNA-cluster. For example, the reference cells of lymphoid origin (e.g., T cells, NK cells) clustered together with one of the scRNA-seq clusters, and the reference cells of myeloblast origin (e.g., neutrophils, eosinophils, macrophages) clustered together and with another scRNA-seq cluster. In



aggregate, the asthmatic group had fewer B cells and more myeloid cells (Fisher’s exact test,  $p$  values  $2e-8$  and  $7e-7$  for B cells and myeloid cells, respectively). However, the number of cells in each cluster was highly variable between individuals; for example, nearly all of the B cell cluster was found in one individual (Fig. 2c, control B cell outlier). For these five patients, the number of cells in the myeloid lineage cluster showed the most distinct separation between asthmatics and controls but was not significant (Mann-Whitney-Wilcoxon signed-rank test,  $p$  value = 0.09) (Fig. 2c). From this analysis,

we concluded that (1) the blood-derived cell profiles appropriately fit the sputum cell types, (2) no additional cell types were needed to deconvolve the sputum bulk RNA-seq data, and (3) the high variability would require larger sample sizes to detect differences between asthmatics and controls.

#### ***Evaluation of deconvolution results by microscopy***

Second, we evaluated a random subset (~50%) of the samples by microscopy and manually counted the number of neutrophils, eosinophils, lymphocytes, and macrophages. We found good agreement with  $F$ , when cell counts could be directly compared, i.e., neutrophils and eosinophils were both present in  $F$  and counted by microscopy ( $F$  test  $p$  values =  $8e-5$  and  $1.8e-7$  for neutrophils and eosinophils, respectively). In cases where the deconvolution method gave higher resolution (e.g., M0, M1, and M2 macrophages versus one type of macrophage by microscopy), the aggregation of the relevant columns in  $F$  correlated well with the microscopy counts ( $F$  test  $p$  values = 0.012 and 0.021 for macrophages and lymphocytes, respectively) (Fig. 2d).

#### ***Association of cell fractions with clinical features***

Having validated the deconvolution of sputum samples (table  $F$ ), we then correlated the cell fractions with clinical features ( $\mathbf{R}(F_{,f}, C_{,c})$  for all patients). We found that the changes in fractions of several cell types were highly correlated with clinical features (Fig. 2e). For example, the fraction of T-regulatory cells negatively correlated with the number of hospitalizations per year ( $p$  value = 0.002, all correlations false discovery rate corrected, see the “Methods” section).

#### ***Evaluation of deconvolution results by unsupervised decomposition***

We compared the signal captured by cell-type deconvolution to an unsupervised decomposition method: LDA. Using LDA, we factored the gene expression table into ten topics that conceptually represent gene expression programs. This resulted in a gene-topic-fraction-in-patient table ( $\theta^G$ ,  $N_p$  by  $N_k$ ) with  $N_k = 10$  topics, as well as the corresponding gene-topic table ( $\phi^G$  [ $N_k$  by  $N_g$ ]), that is analogous to the supervised deconvolution tables  $F$  and  $S$ . We correlated the cell-type fraction table with the gene topics fraction table ( $\mathbf{R}(F_{,f}, \theta, k)$  for all patients and found agreement between LDA and the cell signature-based deconvolution for the most prominent cell type, neutrophils ( $p$  value =  $1.7e-34$ ) (Fig. 2d, topic 4). Gene set enrichment analysis of the genes in this topic showed enrichment in the neutrophil chemotaxis pathway (Additional file 1: Fig. S8 B). The top genes in each topic can be found in Additional file 1: Fig. S7 and the complete membership in Additional file 2: Table S1.

However, the remaining topics comprised multiple cell types. This suggests that LDA can identify distinct but partially overlapping features in  $G$ . According to the clustering of  $\theta^G$ , a subgroup of severely asthmatic patients associated with topic 4 (Additional file 1: Fig. S8A). The top-weighted genes in topic 4 were enriched for the pathways “neutrophil chemotaxis” and “asthma-related genes” (Additional file 1: Fig. S8B). These pathways were not enriched in the analogous cell-type signature table  $S$ , suggesting that LDA topics are distinct from the cell-type signatures, but are also clinically relevant. Moreover, the top-weighted genes in topic 1 of the gene topic component table were

mitochondrial genes, and topic 1 was strongly correlated with age. This link shows strong support in the literature, as reactive oxygen species produced by the mitochondria reduce their function over time [17]; however, we did not observe this relationship for any cells in the cell-type fraction table ( $F$ ). Another method using a very different algorithm than LDA, non-negative matrix factorization (NMF), showed strong agreement with LDA ( $p$  value  $< 2.2e-16$ ) (Additional file 1: Fig. S2, Nmf.1). This supports the use of supervised deconvolution methods for identifying signals that are different than those from unsupervised methods. Unsupervised decomposition should be considered a distinct set of features from those found through deconvolution.

### Analysis of exogenous reads

After filtering out contaminants and human reads, we assembled the set of reads that aligned to exogenous genomes into a microbe table ( $M$ ). The exogenous sequences aligned to mostly bacteria and fungi, although we also observed a few arthropod and helminth reads. The dominant phyla observed were from the bacterial kingdom: Proteobacteria, Firmicutes, and then Bacteroidetes. The abundance of Proteobacteria is in contrast to observations from the gut where Bacteroidetes predominate [18]. Also notable was the presence of two phyla of fungi among the eight most abundant overall, although this was in lower abundance than many of the bacterial phyla.

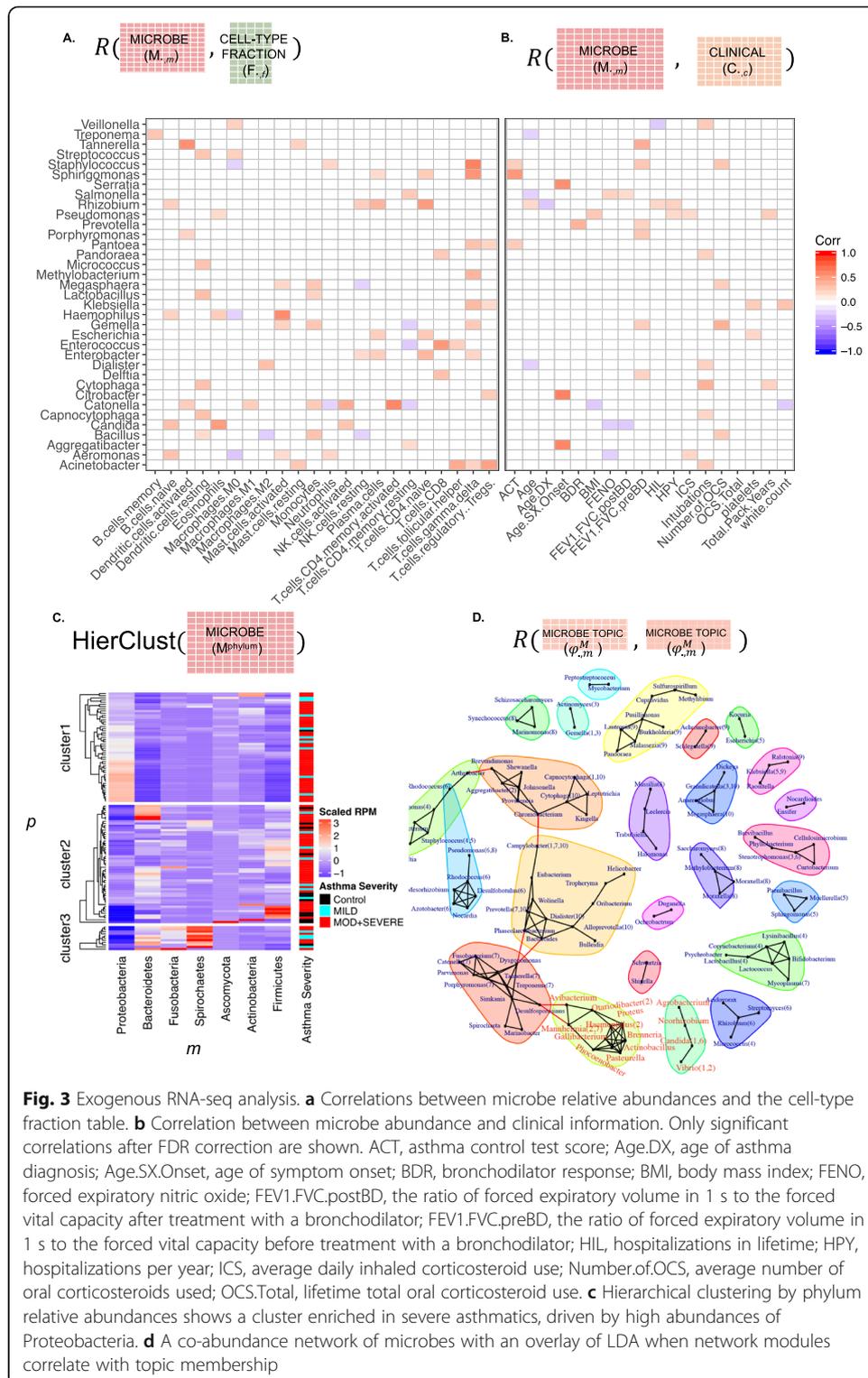
### Microbe correlations with clinical information and cell fractions

We correlated the microbe abundances to clinical information ( $R(M, m, C, c)$  for all patients) (Fig. 3a). *Klebsiella* was associated with increased total white blood cell counts ( $p$  value = 0.001), as has been described previously [19]. *Candida* was associated with worse lung function test results (e.g., forced expiratory volume and forced vital capacity ( $p$  value = 0.02)).

We next correlated microbe abundances to human immune cell fractions ( $R(M, m, F, p)$  for all patients) (Fig. 3b). Several correlations demonstrated results with strong literature precedence. For example, studies have previously shown that *Haemophilus* associates with eosinophilia [20], and we observed a significant correlation between *Haemophilus* and the fraction of eosinophils ( $p$  value = 0.005). We also observed a significant correlation between *Haemophilus* and activated mast cells ( $p$  value =  $9e-12$ ), suggesting an alternative route to *Haemophilus*-induced inflammation [21]. Moreover, the fungal genus *Candida* was also significantly correlated with eosinophils ( $p$  value =  $1e-8$ ), even more strongly than *Haemophilus*.

### Dimensionality reduction for microbes: clustering and networks

We attempted to de-noise the microbe table ( $M$ ) with a variety of dimensionality-reduction techniques. First, we collapsed the microbes by taxonomy, grouping them to the rank of phylum ( $M^{\text{phylum}}$ ), and then hierarchically cluster the patients based on their phylum abundance (Fig. 3c,  $\text{HierClust}(M^{\text{phylum}})$ ). The hierarchical clustering showed that the phylum distributions formed three clusters of patients. We related these clusters to the clinical variable “asthma severity,” defined by the amount of fluticasone or equivalent per day used to control symptoms (mild,  $< 200 \mu\text{g}$ ; moderate,  $200-800 \mu\text{g}$ ; severe,  $> 800 \mu\text{g}$ ). We observed that one of the clusters was enriched for



patients identified as having moderate or severe asthma (Fisher’s exact test,  $p$  value = 0.005). This cluster was characterized by the highest relative abundance of the phylum Proteobacteria (Fig. 3c). Notably, the genus *Haemophilus* belongs to this phylum, consistent with the correlations observed at the genus rank (Fig. 3a, b).

Similarly, we could de-noise the microbe table using a co-abundance network, by correlating the genus-level abundances ( $\mathbf{R}(M, m, M, m)$ ) and identifying significant modules (Additional file 1: Fig. S5). An interpretation of these modules is that they define metabolic niches, where microbes either directly compete for metabolites or there is interdependency in metabolite production. Such networks could be created from other tables, such as the topic distribution of microbes ( $\mathbf{R}\phi_{\cdot,m}^M, \phi_{\cdot,m}^M$ ) for all the topics (Fig. 3d). These modules represent another unit that could be related to the clinical information ( $C$ ) and the cell-type fractions ( $F$ ).

#### LDA-link for the identification of links between genes and microbes

How much cross-talk exists between the airway microbes and human cells remains contentious [22]. We feel this is partly due to the heterogeneous and noisy data from airway samples, where it is often challenging to find strong correlations using standard algorithms. We therefore sought to link genes to microbes via a method we dubbed the LDA-link.

As is often the case, these RNA-seq data lack a gold standard set of labels for the links between genes and microbes that can train a machine learning algorithm. Therefore, the LDA-link uses a strategy similar to self-supervised methods, which create a pseudo-gold standard, train a supervised model using the pseudo-gold standard, and then apply the model to the remainder of the data [23–30]. In this case, the training set was defined by a small number of very strong linear correlations and an equal number of very low correlations.

Specifically, we first related columns between the gene and microbe tables ( $\mathbf{R}(G, g, M, m)$ ), yielding many low-scoring correlations. A relatively small number of the correlations were strong ( $\mathbf{R} > 0.4$ ) and highly significant ( $p < 1e-5$  and  $FDR < 0.016$ ) (Fig. 4a). We selected these strong correlations as true-positive links between genes and microbes in the self-supervised training set and non-correlated pairs ( $-0.05 < \mathbf{R} < 0.05$ ) as the true-negative links. We tested the validity of this true-positive label by gene set enrichment analysis. The genes involved in these strong correlations were enriched for pathways related to microbial interactions in the airway, including “asthma and bronchial hypersensitivity” and “respiratory syncytial virus bronchiolitis” (Fig. 4b), suggesting that the small set of strong linear correlations was relevant to asthma. These gene-microbe links were then used to train a random forest classifier.

An important idea for the method is that the input features for the random forest classifier were the unsupervised LDA topic distributions for each gene and microbe ( $\phi_{\cdot,g}^G, \phi_{\cdot,m}^M$ ), rather than the “direct” gene counts and microbe abundances that were used to identify the linear correlations. That is, for each gene-microbe pair, we concatenated the gene and microbe topics into a single vector (length 20) that were the features used by the random forest classifier. The most important features in defining links between genes and microbes in the training set were gene topics 7 and 8 and microbe topic 1 (Fig. 4c–f). The genes that comprise the most influential gene topic 8 are enriched for the pathway “inflammatory response,” and specifically the cytokines IL2 and IL6.

We applied the trained model to all gene-microbe pairs (excluding the training set), using the LDA topic distributions as inputs. The confidence scores of the machine learning approach were strongly bimodal. Whereas the linear correlation showed many

weak to mid-strength and a few strong interactions, the LDA-link resulted in many low-confidence interactions, few mid-strength, and a larger number of high-confidence interactions. Altogether, we identified 1883 high-confidence (probability > 0.95) gene-microbe interaction pairs, as opposed to the linear correlations, which yielded 862 ( $p$  value <  $1e-5$ ).

### Cross-talk between genes and microbes defined by LDA-link

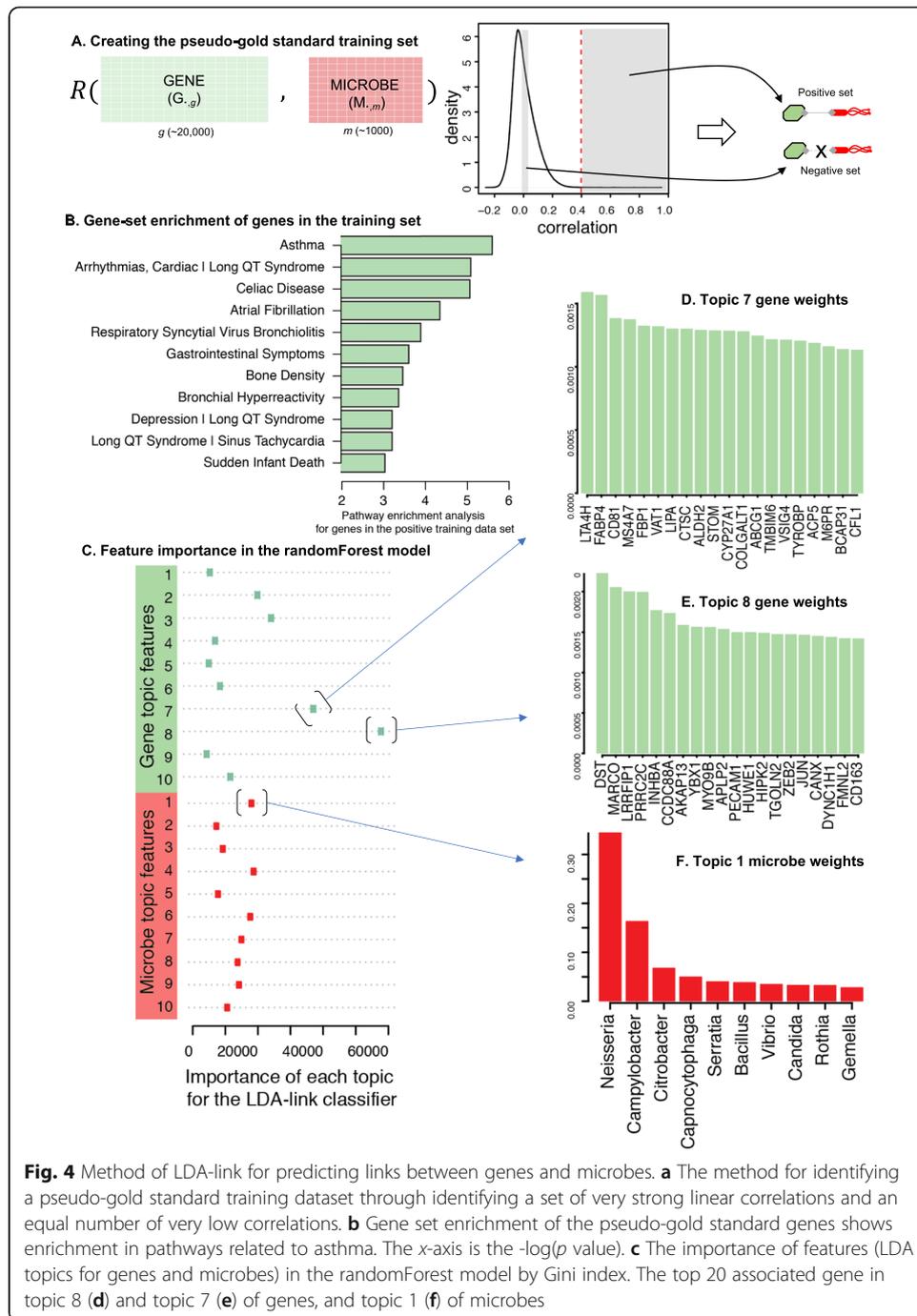
LDA-link identified connections between genes and microbes reported elsewhere in the literature as well as novel observations. A bipartite graph summarizing a subset of the connections between genes and microbes shows that in most cases, several genes are linked to each microbe (Fig. 5a, for a complete list, see Additional file 3: Table S2). Several of the observed links showed strong literature precedence. For example, the gene lactotransferrin was linked to *Aeromonas* [31], *Burkholderia* was linked to gene MUC6 [32], *Haemophilus* was linked to NFKB Inhibitor Zeta [33], and *Pasteurella* was linked to IL1B [34]. To the best of our knowledge, several novel links were also found, such as between *Haemophilus* and IL1B and between *Candida* and GCSAML. In addition to gene-microbe pairs, we layered on the pathway and cell deconvolution data to identify larger-scale effects of microbes.

Microbes were linked to genes that are enriched in pathways relating to autoimmunity and inflammation as well as cytokine receptors and their interactions (Fig. 5b). The microbes associated with cytokine pathways included *Synechococcus*, *Lactococcus*, *Dialister*, *Psychrobacter*, *Moraxella*, *Brenneria*, *Proteus*, *Haemophilus*, and *Pasteurella*. In addition, we related the cell-type signature table ( $S_{f,g}$ ) to identify the immune cell types that are related to each microbe (Fig. 5c). We observed the *Haemophilus*-IL1B link in monocytes and mast cells. Samples containing *Haemophilus* correlated with more activated mast cells (Spearman correlation,  $p$  value = 9–12) (Fig. 5c, inset) [35–38]. Similarly, the fungal genus *Candida* was linked to the gene GCSAML, which was highly expressed by eosinophils. The presence of *Candida* was associated with increased numbers of eosinophils in the airway (Spearman correlation,  $p$  value =  $1e-8$ ).

### Discussion

Heterogeneity and noise are common problems in biological datasets. Heterogeneity can derive from mixtures of different cell types, such as in sputum, or from sparsity, such as in microbiome or single-cell RNA-seq data. Unsupervised methods of dimensionality reduction can effectively eliminate these issues, but suffer from decreased interpretability. That is, variables are collapsed together for reasons that are often opaque. Supervised dimensionality reduction maintains interpretability because variables are collapsed using prior knowledge, such as the genes in a pathway or the expression patterns of a cell type. Here, we describe a method we dubbed LDA-link that combines supervised and unsupervised methods to de-noise the data while retaining interpretability.

The field of asthma is increasingly appreciating the role of the airway microbiome in disease. Commensal microbiota have been shown in other contexts to be strong regulators of host immune system development and homeostasis [39]. Disturbances in the composition of commensal bacteria can result in imbalanced immune responses and



**Fig. 4** Method of LDA-link for predicting links between genes and microbes. **a** The method for identifying a pseudo-gold standard training dataset through identifying a set of very strong linear correlations and an equal number of very low correlations. **b** Gene set enrichment of the pseudo-gold standard genes shows enrichment in pathways related to asthma. The x-axis is the  $-\log(p)$  value. **c** The importance of features (LDA topics for genes and microbes) in the randomForest model by Gini index. The top 20 associated gene in topic 8 (**d**) and topic 7 (**e**) of genes, and topic 1 (**f**) of microbes

affect an individual’s susceptibility to various diseases, including those that are inflammatory (e.g., inflammatory bowel disease and colon cancer), autoimmune (e.g., celiac disease and arthritis), allergic (e.g., asthma and atopy), and metabolic (e.g., diabetes, obesity, and metabolic syndrome) (reviewed in [40]). Investigating the microbiota in the lower respiratory tract is a relatively new field in comparison with the extensive work on the intestinal tract. In fact, the lung was excluded from the original Human Microbiome Project because it was not thought to have a stable resident microbiome [18]. Since this initial lag, reports have defined commensal microbes within healthy

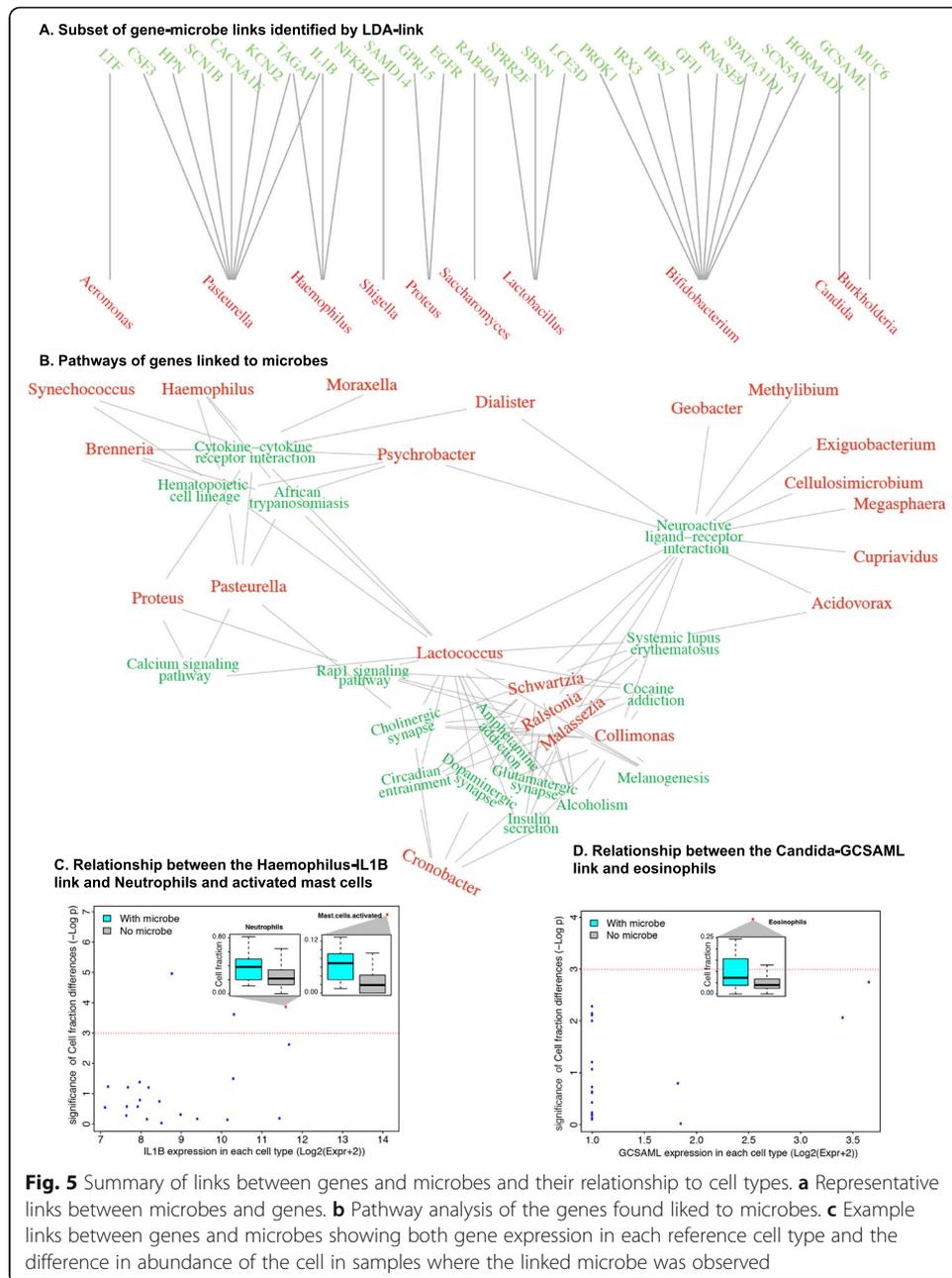
individuals and identified consistently enriched or depleted taxa in such conditions as asthma, chronic obstructive pulmonary disease, and cystic fibrosis, to name a few [5, 41–44]. LDA-link facilitates the observation of relationships between microbes and genes and links them to immune cells and their responses.

While several links imply that the presence of a microbe has triggered an inflammatory response, other interactions are also suggested. The gene lactotransferrin was linked to *Aeromonas*, which has been associated with gastroenteritis and skin infections and has been previously reported to bind lactoferrin [31]. *Burkholderia*, a gram-negative bacterial genus, is recognized as an important pathogen in the mucus-filled lungs of patients with cystic fibrosis; it was linked to gene MUC6, which encodes a secreted protein responsible for the production of mucin [32]. Classical pathogenic-type interactions included *Haemophilus* linked to NFKB inhibitor zeta, which is induced by the bacterial cell wall component lipopolysaccharide [33]. In addition, *Haemophilus* was linked to the cytokine interleukin 1 beta (IL1B), an important mediator of the inflammatory response. IL1B hypersensitivity is a hallmark of the asthma phenotype. *Pasteurella* was also linked to IL1B, and its toxin has been shown to induce the expression of IL1B [34].

A number of the links identified here have been described previously through different experimental methods. For example, a severe form of asthma has been characterized by eosinophilia [45], and pulmonary candidiasis has long been associated with allergic bronchial asthma and inflammation [46]. However, the methods used to uncover these relationships we argue are more challenging than total RNA-seq of sputum or provide less information. For example, asthmatic eosinophilia was identified by microscopic counting of cells [45], as we used to validate deconvolved cell counts (Fig. 2d). In the case of candidiasis, the observation was made by culturing [46]. By contrast from these RNA-seq data, significantly more information can be obtained, such as other cell types and other microbes. In addition, these links were not found via standard correlations. By using the LDA-link, we identified associations that could not be extracted from this heterogeneous dataset. Notably, both fungi and bacteria showed these links, further highlighting the need to evaluate more than bacteria when performing microbiome experiments in the airway.

Having gained confidence in the method by finding previously reported associations, we then turn to the novel findings. The links identified here suggest processes by which lung immune cells respond to microbes. We found that mast cells respond to *Haemophilus* and *Pasteurella* via IL1B and that eosinophils respond to *Candida* via GCSAML. While experimental validation of these links is needed, these results represent observations that would be missed by analyses that do not effectively de-noise RNA-seq data in a way that maintains the interpretability of cell types, or that analyze only human RNA-seq reads. Additional follow-up is needed to establish a causal connection between these correlations.

RNA-seq continues to be a common dataset collected in biological sciences. Increasingly, its applications include the identification of distinct entities within the data, such as microbes and immune cells. For example, the roles of both microbes and immune cells are being inferred using RNA-seq from tumor biopsies and other tissues [47, 48] where the interpretation can be challenging because of the sample (sparse, low-abundance microbes) and its preservation (e.g., RNA degradation in formalin-fixed paraffin-embedded tissues). We expect the LDA-link to be useful in this and many other such contexts.



## Methods

### Patient cohort

The patient population in this study received treatment at the Yale Center for Asthma and Airway Diseases before 2016. Included were patients classified as having mild, moderate, or severe asthma, as defined by the amount of fluticasone or equivalent per day to control symptoms (mild, > 200 µg; moderate, 200–800 µg; severe, > 800 µg). Most patients were young (median age 46), female (~ 75%), and White (60%). A single sample was used from each patient, not within 2 weeks of an exacerbation event. Control samples from 13 healthy individuals were included (Additional file 4: Table S3).

### Sample collection and sequencing

Sputum induction was performed with hypertonic saline, the mucus plugs were dissected away from the saliva, the cellular fraction was separated, and the RNA was purified as described previously [49]. Briefly, RNA was purified using the All-in-One purification kit (Norgen Biotek), and its integrity was assayed by an Agilent bioanalyzer (Agilent Technologies, Santa Clara, CA). Ribosomal depletion was performed with the RiboGone-Mammalian kit (Clontech Cat. Nos. 634846 and 634847), and cDNA was created with the SMARTer Stranded RNA-seq Kit (Cat. No. 634836). Samples were sequenced using an Illumina HiSeq 4000 with  $2 \times 125$  bp reads, with an average of 47.5 million reads per sample.

### RNA-seq processing by exceRpt

An adapted version of the software package exceRpt [14] was used to process the sputum RNA-seq data. Briefly, RNA-seq reads were subjected to quality assessment using the FastQC software v.0.10.1 (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) both prior to and following 3' adapter clipping. Adapters were removed using FastX v.0.0.13 ([http://hannonlab.cshl.edu/fastx\\_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/)). Identical reads were counted and collapsed to a single entry, and reads containing N's were removed. Clipped, collapsed reads were filtered through the Univec database of common laboratory contaminants and a human ribosomal database before being mapped to the human reference genome (hg19) and pre-miRNA sequences using STAR [50]. Reads that did not align were mapped against a ribosomal reference library of bacteria, fungi, and Archaea, compiled by the Ribosome Database Project [51], and then to genomes of bacteria, fungi, plants, and viruses, retrieved from GenBank [51]. In cases where RNA-seq reads aligned equally well to more than one microbe, a "last common ancestor" approach was used, and the read was assigned to the next node up the phylogenetic tree, as performed by similar algorithms [14, 52].

### Data tables notation

We use the following notation to define matrices associated with patients ( $N_p = 115$ ), genes ( $N_g = \sim 20,000$ ), microbes ( $N_m = \sim 1000$ ), cell fractions ( $N_f = 22$ ), reference cell-type signature genes ( $N_s = \sim 600$ ), and topics ( $N_k = 10$ ) (Fig. 1):

**C**: clinical table [ $N_p$  by  $N_c$ ] where  $c$  is the clinical index

**G**: gene table [ $N_p$  by  $N_g$ ] bulk RNA-seq table, FPKM expression of each gene for each patient

**M**: microbe abundance table [ $N_p$  by  $N_m$ ], relative abundances of each microbe for each patient

**F**: cell fraction table [ $N_p$  by  $N_f$ ] resulting from the deconvolution of  $G_{p,g}$

**S'**: reference cell-type expression table [ $N_f$  by  $N_s$ ] from the RNA-seq of experimentally isolated immune cells where  $s$  is the index of genes in the signature set

**S**: expanded all signature table [ $N_f$  by  $N_g$ ] resulting from solving  $S = F^{-1}G$

$\theta^G$ : patient topic table [ $N_p$  by  $N_k$ ] after LDA inference based on gene table  $G_{p,g}$

$\phi^G$ : gene topic table [ $N_k$  by  $N_g$ ] after LDA inference based on gene table  $G_{p,g}$

$\theta^M$ : patient topic table [ $N_p$  by  $N_k$ ] after LDA inference based on microbe table  $M_{p,m}$

$\phi^M$ : microbe topic table [ $N_k$  by  $N_m$ ] after LDA inference based on table  $M_{p,m}$

$L$ : gene-microbe link table [ $N_g$  by  $N_m$ ] predicted by LDA-link

### Dimensionality reduction

#### *Supervised deconvolution*

The gene table ( $G$ ) was deconvolved using the transcriptomes from 22 flow cytometry-sorted and sequenced immune cell types using the CIBERSORT tool [8]. Briefly, characteristic gene expression patterns for each cell type were used to estimate the fraction of each cell type in the sample by solving for the equation:

$$G = FS'$$

where  $G$  is the gene table of human protein-coding gene expression from the exceRpt pipeline,  $F$  is the cell fraction table, and  $S'$  is the reference cell-type gene expression signature from experimentally isolated and sequenced cell types. Support vector regression was used to perform variable selection, reducing the number of characteristic genes used to distinguish cell types and thereby reducing overfitting. The above equation was then solved to provide an estimate of  $F$ .  $P$  values for the fit of  $F$  to  $G$  demonstrated that all samples were significant at  $\alpha = 0.05$ . Following the solution of  $F$ , the cell type expression,  $S$ , expanded to all the other, non-signature genes, were inferred using the same equation.

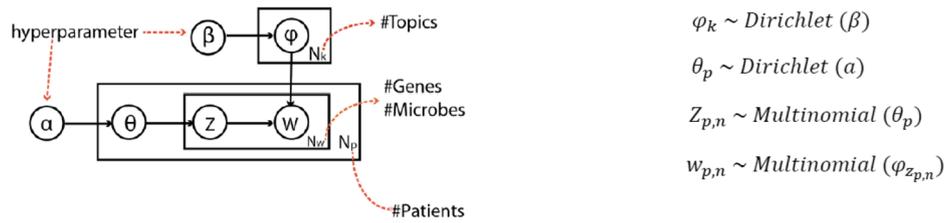
#### *Justification for unsupervised decomposition using LDA*

We compared LDA to several commonly used dimensionality reduction techniques. As a canonical method, principal component analysis (PCA) uses an orthogonal transformation to convert a data matrix into a set of linearly uncorrelated principal components, but it may give negative numbers. Similar to PCA, non-negative matrix factorization (NMF) can be represented by a linear combination of all the variables, but it requires all the matrices to have only non-negative elements.

LDA has two advantages over these methods that we demonstrate using a classic machine learning dataset. First, LDA provides a sparser decomposition that yields more robust and interpretable representations. By controlling the hyper-parameter  $\alpha$ , LDA achieves better sparseness of low-dimensional patterns (Additional file 1: Fig. S10A). Second, another important aspect of the decomposition is that it should preserve the distances (correlations) between two variables in the original data [53]. The correlations from LDA document-to-topic distributions more closely resemble the linear correlations than the feature matrix ( $W$ ) from NMF (Additional file 1: Figure S10B-C).

#### *Decomposition through a generative model*

The gene table,  $G$ , was decomposed using LDA and NMF. For LDA, the abundance values for bulk RNA-seq and exogenous RNA were scaled down to reduce computation intensity during the Gibbs sampling. More simply, the RPM expression values were converted to integers and then divided by 10. The max value was set to 1000. Each patient represents a document, and the genes and microbes were treated like a dictionary of words in the traditional LDA application. The word was gene or microbe, and the word count represents gene expression or microbe abundances. The dictionary (or corpus) includes  $V$  unique words (genes or microbes), and each word has an index ( $v$ ) in the dictionary; we built LDA models for genes and microbes, respectively.



In this model,  $N_p$  is the number of documents,  $N_w$  is the number of words in a document,  $N_k$  is the number of topics (set as 10), and  $N_v$  is the total number of words in the dictionary. The variables  $p$ ,  $w$ ,  $k$ , and  $v$  denote indices of documents, words in a document, topics, and words in a dictionary, respectively. The vector  $\alpha$ , of length  $N_k$ , and  $\beta$ , of length  $N_v$ , are the hyper-parameters;  $\theta$  [ $N_p$  by  $N_k$ ] is the topic distribution of documents; and  $\phi$  [ $N_k$  by  $N_v$ ] is the distribution of words in topics.  $z$  is a categorical integer between 1 and  $N_k$ , denoting the topic of a word in a document.

The joint distribution of the LDA model is  $\mathcal{P}(Z, W; \alpha, \beta)$ , and  $\phi$  and  $\theta$  are integrated out as:

$$\begin{aligned} \mathcal{P}(Z, W; \alpha, \beta) &= \int_{\phi} \overbrace{\prod_{i=1}^{N_k} \mathcal{P}(\phi_i; \beta) \prod_{j=1}^{N_p} \prod_{t=1}^{N_w} \mathcal{P}(W_{j,t} | \phi_{z_{j,t}})}^{\mathcal{P}(W|Z, \beta)} d\phi \int_{\theta} \overbrace{\prod_{i=1}^{N_p} \mathcal{P}(\theta_i; \alpha) \prod_{j=1}^{N_w} \mathcal{P}(Z_{i,j} | \theta_i)}^{\mathcal{P}(Z | \alpha)} d\theta \\ &= \prod_{k=1}^{N_k} \frac{\Delta(n_{\cdot,k} + \beta)}{\Delta(\beta)} \prod_{s=1}^{N_p} \frac{\Delta(n_{s,\cdot} + \alpha)}{\Delta(\alpha)} \end{aligned}$$

where

$$\Delta(\alpha) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}$$

The Gibbs sampling equation can be derived from  $\mathcal{P}(Z, W; \alpha, \beta)$  to approximate the distribution of  $\mathcal{P}(Z|W; \alpha, \beta)$  because  $\mathcal{P}(W; \alpha, \beta)$  is invariant to  $Z$ . Given  $Z_{m,n}$  denotes the topic of the  $n$ th word in the  $m$ th document and also assume that its word symbol is the  $v$ th word in the dictionary;  $\neg m, n$  denotes all the word except the  $n$ th one in the  $m$ th document. The conditional probability can be inferred as follows:

$$\begin{aligned} \mathcal{P}(Z_{m,n} = k | Z_{\neg(m,n)}, W; \alpha, \beta) &= \frac{\mathcal{P}(Z, W; \alpha, \beta)}{\mathcal{P}(Z_{\neg(m,n)}, W; \alpha, \beta)} = \frac{\mathcal{P}(w, z)}{\mathcal{P}(w_{m,n}, w_{\neg(m,n)}, z_{\neg(m,n)})} \\ &= \frac{\mathcal{P}(w, z)}{\mathcal{P}(w_{\neg(m,n)}, z_{\neg(m,n)})} \cdot \frac{1}{\mathcal{P}(w_{m,n} = t)} \\ &\propto \frac{\mathcal{P}(w, z)}{\mathcal{P}(w_{\neg(m,n)}, z_{\neg(m,n)})} \\ &= \frac{\prod_{k=1}^{N_k} \frac{\Delta(n_{\cdot,k} + \beta)}{\Delta(\beta)} \prod_{p=1}^{N_p} \frac{\Delta(n_{p,\cdot} + \alpha)}{\Delta(\alpha)}}{\prod_{k=1}^{N_k} \frac{\Delta(n_{\neg(m,n),k} + \beta)}{\Delta(\beta)} \prod_{p=1}^{N_p} \frac{\Delta(n_{p,\neg(m,n)} + \alpha)}{\Delta(\alpha)}} \end{aligned}$$

$$= \frac{\Delta(n_{\cdot, k} + \beta)}{\Delta(n_{\neg(m,n), k} + \beta)} \cdot \frac{\Delta(n_{p, \cdot} + \alpha)}{\Delta(n_{p, \neg(m,n)} + \alpha)}$$

After Gibbs sampling, the expectation of the  $\theta$  (doc  $\rightarrow$  topic) and  $\phi$  (topic  $\rightarrow$  word) matrix can be inferred as follows given the symmetric hyper-parameters  $\alpha$  and  $\beta$  were used:

$$\theta_{p,k} = \frac{n_{p,\cdot}^k + \alpha}{\sum_{k'=1}^{N_k} n_{p,\cdot}^{k'} + N_k \alpha}$$

$$\phi_{k,v} = \frac{n_{\cdot,v}^k + \beta}{\sum_{v'=1}^{N_v} n_{\cdot,v'}^k + N_v \beta}$$

We instantiated the variables  $\theta$  and  $\phi$  to  $\theta_{p,t}^G, \theta_{p,t}^M$  and  $\phi_{k,g}^G, \phi_{k,m}^M$ , respectively, where  $\theta_{p,t}^G, \theta_{p,t}^M$  denotes the gene and microbe topic fraction in patient;  $\phi_{k,g}^G, \phi_{k,m}^M$  denotes the gene and microbe topic.

### Single-cell RNA-seq

Sputum cells were separated on a Fluidigm C1 medium-sized channel. The mRNA was purified from approximately 500 pg to 1 ng of total RNA using the Clontech SMARTer Ultra Low RNA Kit and poly-dA-selected using SPRI beads and dT primers. Full-length cDNA was sheared into 200–500 bp DNA fragments by sonication (Covaris, MA, USA) and then indexed and size validated by LabChip GX. Two-nanomolar libraries were loaded onto Illumina version 3 flow cells and sequenced using 75-bp single-end sequencing on an Illumina HiSeq 2000 according to Illumina protocols. Data were cleaned, processed, aligned, and quantified following the SINCERA pipeline [54]. Briefly, the data are processed to filter genes with poor expression across all cells and cells with poor expression across most genes ( $\geq 5$  FPKM in 2 cells of the sample,  $\geq 1000$  genes in a cell) and then transformed to z-scores. Clusters were created via principal component analysis and then t-SNE of the principal component loadings. The optimal number of clusters was identified by calculating the gap statistic (method = “Tibs2001SEmax”) using the cluster package in R.

### Pathogen-to-host correlation and link identification

Microbe relative abundances and gene TPM values were correlated as follows, with  $G_{\cdot, i}$  for  $i$  gene and  $M_{\cdot, j}$  for  $j$  microbe:

$$R(i, j) = \frac{\sum_{p=1}^{N_p} (G_{p,i} - \overline{G_{\cdot,i}}) (M_{p,j} - \overline{M_{\cdot,j}})}{\sqrt{\sum_{p=1}^{N_p} (G_{p,i} - \overline{G_{\cdot,i}})^2} \sqrt{\sum_{p=1}^{N_p} (M_{p,j} - \overline{M_{\cdot,j}})^2}}$$

Correlation  $p$  values were corrected by the method of Benjamini and Hochberg [55].

We employed a self-supervised learning framework to learn labels from the data. The links between gene and microbes that were very tightly correlated, ergo they were in one cluster, were labeled as the positive links after being confirmed by functional enrichment (gene-microbe correlations with  $p$  values less than  $1e-5$  (FDR = 0.016 and absolute correlation greater than 0.4)). Negative links in the training set were defined as

an absolute correlation of less than  $R(i,j) = 0.05$ . Each gene and microbe could be linked to many microbes or genes, respectively.

Cross-validation was performed using a strict left-out process, as shown in Additional file 1: Figure S11. Ten percent of the training data (true-positive and negative labels) were separated as the test set, and any microbes and genes in this test set were removed from the remaining training data, leaving slightly less than 90% of the pseudo-gold standard for training. A randomForest model was trained on the remaining labels and then tested on the left-out 10%. We repeated these procedures 10 times to estimate the receiver operator characteristics of the model; the area under the curve was approximately 0.7.

### Microbe co-abundance network

The raw abundance  $M$  and LDA microbe topic matrices  $\phi^M$ , which represent the microbe's weight to each topic, were generated. The correlation network between different microbes was calculated using the Pearson correlation. The cutoff to define a co-abundance edge was 0.8 for  $R(\phi_{\cdot,m}^M, \phi_{\cdot,m}^M)$  and 0.3 for  $R(M, m, M, m)$ . The microbe network modules, which were densely connected to themselves but sparsely connected to other modules, were clustered based on betweenness [56] and the other algorithms; we also tested label propagation and fast greedy algorithms [57]. We also compared the LDA topics with the microbes in the same clusters. If a microbe was the top 10 most highly contributed for a topic, then we labeled the topic number in the bracket. Some microbes may have multiple topic labels because they highly contribute to multiple topics.

### Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s13059-020-02033-z>.

**Additional file 1: Figures S1-S11.** Supplemental Figures.

**Additional file 2: Table S1.** Gene-topic membership. A tab-delimited text file showing the complete topic distributions for all genes.

**Additional file 3: Table S2.** Gene-microbe links. A tab-delimited text file showing all genes-microbe pairs and the confidence of their link.

**Additional file 4: Table S3.** Patient characteristics. A Microsoft Excel spreadsheet providing summary statistics for the patients included in the study.

**Additional file 5:** Review history.

### Acknowledgements

The authors would like to thank the support of the Yale High Performance Computing services (Grace, Ruddle, Farnam) and Yale Center for Genome Analysis.

### Peer review information

Yixin Yao was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

### Review history

The review history is available as Additional file 5.

### Authors' contributions

GC and MG conceived the study. GC, JLG, and NG collected the data. QL processed the samples. DS, SL, BB, TL, YX, and RH performed the analyses. GC, GW, and MG oversaw the analyses. DS, SKL, TL, MG, and GC wrote the manuscript. All authors reviewed the manuscript. The authors read and approved the final manuscript.

### Authors' information

Twitter handles: @dspakowicz (Daniel Spakowicz); @MarkGerstein (Mark Gerstein).

**Funding**

This work was supported by a National Library of Medicine fellowship to DS (5T15LM007056-28) and an NHLBI grant to GC (1R01HL118346-01).

**Availability of data and materials**

Sputum bulk and single-cell RNA-seq data are available through dbGaP under the NCBI BioProject PRJNA611097 [58]. Code for LDA-link and to regenerate all figures is available at <https://github.com/gersteinlab/decoasthma> [59]. The code is open-source under the MIT License in compliance with the open-source initiative.

**Ethics approval and consent to participate**

This study was approved by the Yale Institutional Review Board Human Investigation Committee (0102012268). All study participants gave written informed consent. All methods comply with the Helsinki Declaration.

**Competing interests**

The authors declare no conflicts of interest.

**Author details**

<sup>1</sup>Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA. <sup>2</sup>The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA. <sup>3</sup>Division of Medical Oncology, Ohio State University College of Medicine, Columbus, OH, USA. <sup>4</sup>Department of Biomedical Informatics, Ohio State University College of Medicine, Columbus, OH, USA. <sup>5</sup>Section of Pulmonary, Critical Care, and Sleep Medicine, Department of Internal Medicine, Yale University School of Medicine, New Haven, CT, USA. <sup>6</sup>Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA. <sup>7</sup>Department of Computer Science, Yale University, New Haven, CT, USA. <sup>8</sup>Department of Statistics and Data Science, Yale University, New Haven, CT, USA.

Received: 10 October 2019 Accepted: 30 April 2020

Published online: 22 June 2020

**References**

1. Yan X, Chu J-H, Gomez J, Koenigs M, Holm C, He X, Perez MF, Zhao H, Mane S, Martinez FD, Ober C, Nicolae DL, Barnes KC, London SJ, Gilliland F, Weiss ST, Raby BA, Cohn L, Chupp GL. Non-invasive Analysis of the Sputum Transcriptome Discriminates Clinical Phenotypes of Asthma. *Am J Respir Crit Care Med* [Internet]. 2015 Mar 12 [cited 2015 May 11]; Available from: <http://www.atsjournals.org/doi/abs/10.1164/rccm.201408-1440OC>.
2. Moore WC, Meyers DA, Wenzel SE, Teague WG, Li H, Li X, D'Agostino R Jr, Castro M, Curran-Everett D, Fitzpatrick AM, et al. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. *Am J Respir Crit Care Med*. 2010;181:315–23.
3. Holgate ST. Innate and adaptive immune responses in asthma. *Nat Med*. 2012;18:673–83.
4. Shukla SD, Budden KF, Neal R, Hansbro PM. Microbiome effects on immunity, health and disease in the lung. *Clin Transl Immunol*. 2017;6:e133.
5. Huang YJ, Nariya S, Harris JM, Lynch SV, Choy DF, Arron JR, Boushey H. The airway microbiome in patients with severe asthma: associations with disease features and severity. *J Allergy Clin Immunol*. 2015;136:874–84.
6. Gibson PG, Girgis-Gabardo A, Morris MM, Mattoli S, Kay JM, Dolovich J, Denburg J, Hargreave FE. Cellular characteristics of sputum from patients with asthma and chronic bronchitis. *Thorax*. 1989;44:693–9.
7. Belda J, Leigh R, Parameswaran K, O'Byrne PM, Sears MR, Hargreave FE. Induced sputum cell counts in healthy adults. *Am J Respir Crit Care Med*. 2000;161:475–8.
8. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*. 2015;12:453–7.
9. Duitama J, Srivastava PK, Mandoiu II. Towards accurate detection and genotyping of expressed variants from whole transcriptome sequencing data. *BMC Genomics*. 2012;13(Suppl 2):S6.
10. Kumar L, Greiner R. Gene expression based survival prediction for cancer patients—a topic modeling approach. *PLoS One*. 2019;14:e0224446.
11. Liu B, Liu L, Tsykin A, Goodall GJ, Green JE, Zhu M, Kim CH, Li J. Identifying functional miRNA-mRNA regulatory modules with correspondence latent Dirichlet allocation. *Bioinformatics*. 2010;26:3105–11.
12. Pratanwanich N, Lio P. Exploring the complexity of pathway-drug relationships using latent Dirichlet allocation. *Comput Biol Chem*. 2014;53(Pt A):144–52.
13. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489:57–74.
14. Rozowsky J, Kitchen RR, Park JJ, Galeev TR, Diao J, Warrell J, Thistlethwaite W, Subramanian SL, Milosavljevic A, Gerstein M. *excite*: a comprehensive analytic platform for extracellular RNA profiling. *Cell Syst*. 2019;8:352–7 e353.
15. Welling SH, Clemmensen LK, Buckley ST, Hovgaard L, Brockhoff PB, Refsgaard HH. In silico modelling of permeation enhancement potency in Caco-2 monolayers based on molecular descriptors and random forest. *Eur J Pharm Biopharm*. 2015;94:152–9.
16. Welling SH, Refsgaard HHH, Brockhoff PB, Clemmensen LH. Forest floor visualizations of random forests. In: *arXiv e-prints*; 2016.
17. Payne BA, Chinnery PF. Mitochondrial dysfunction in aging: much progress but many unresolved questions. *Biochim Biophys Acta*. 2015;1847:1347–53.
18. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. *Nature*. 2007;449:804–10.
19. Essilfie AT, Simpson JL, Horvat JC, Preston JA, Dunkley ML, Foster PS, Gibson PG, Hansbro PM. Haemophilus influenzae infection drives IL-17-mediated neutrophilic allergic airways disease. *PLoS Pathog*. 2011;7:e1002244.
20. Ahren IL, Eriksson E, Egesten A, Riesbeck K. Nontypeable Haemophilus influenzae activates human eosinophils through beta-glucan receptors. *Am J Respir Cell Mol Biol*. 2003;29:598–605.

21. Galli SJ, Tsai M. Mast cells in allergy and infection: versatile effector and regulatory cells in innate and adaptive immunity. *Eur J Immunol.* 2010;40:1843–51.
22. Mathieu E, Escribano-Vazquez U, Descamps D, Cherbuy C, Langella P, Riffault S, Remot A, Thomas M. Paradigms of lung microbiota functions in health and disease, particularly, in asthma. *Front Physiol.* 2018;9:1168.
23. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8. 2013. pp. 1798–28. <https://doi.org/10.1109/TPAMI.2013.50>.
24. Jayaraman D, Grauman K. Learning image representations tied to egomotion from unlabeled video. *Int J Comput Vis.* 2017;125:136–61.
25. Kong X, Zhang J, Yu PS. Inferring anchor links across multiple heterogeneous social networks. In: *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*. San Francisco: Association for Computing Machinery; 2013. p. 179–88.
26. Lynch C, Khansari M, Xiao T, Kumar V, Tompson J, Levine S, Sermanet P: Learning latent plans from play. *arXiv e-prints* 2019:arXiv:1903.01973.
27. Pathak D, Agrawal P, Efros AA, Darrell T. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML'17)*. JMLR.org. 2017:2778–87.
28. Pathak D, Krähenbühl P, Donahue J, Darrell T, Efros AA. Context Encoders: Feature Learning by Inpainting," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas; 2016. p. 2536–44. <https://doi.org/10.1109/CVPR.2016.278>.
29. Raina R, Battle A, Lee H, Packer B, Ng AY. Self-taught learning: transfer learning from unlabeled data. In: *Proceedings of the 24th International Conference on Machine Learning*. Corvallis: Association for Computing Machinery; 2007. p. 759–66.
30. Wu J, Wang X, Wang WY. Self-supervised dialogue learning. *ACL.* 2019;(1):3857–67.
31. Ascencio F, Ljungh A, Wadstrom T. Characterization of lactoferrin binding by *Aeromonas hydrophila*. *Appl Environ Microbiol.* 1992;58:42–7.
32. Sajjan U, Keshavjee S, Forstner J. Responses of well-differentiated airway epithelial cell cultures from healthy donors and patients with cystic fibrosis to *Burkholderia cenocepacia* infection. *Infect Immun.* 2004;72:4188–99.
33. Park CY, Heo JN, Suk K, Lee WH. Sodium azide suppresses LPS-induced expression MCP-1 through regulating I $\kappa$ B $\alpha$  and STAT1 activities in macrophages. *Cell Immunol.* 2017;315:64–70.
34. Hildebrand D, Bode KA, Riess D, Cerny D, Waldhuber A, Rommler F, Strack J, Korten S, Orth JH, Miethke T, et al. Granzyme A produces bioactive IL-1 $\beta$  through a nonapoptotic inflammasome-independent pathway. *Cell Rep.* 2014;9:910–7.
35. Chapman SJ, Khor CC, Vannberg FO, Rautanen A, Segal S, Moore CE, Davies RJ, Day NP, Peshu N, Crook DW, et al. NFKB1 polymorphisms and susceptibility to pneumococcal disease in European and African populations. *Genes Immun.* 2010;11:319–25.
36. Baldwin AS Jr. The NF- $\kappa$ B and I $\kappa$ B proteins: new discoveries and insights. *Annu Rev Immunol.* 1996;14:649–83.
37. Motoyama M, Yamazaki S, Eto-Kimura A, Takeshige K, Muta T. Positive and negative regulation of nuclear factor- $\kappa$ B-mediated transcription by I $\kappa$ B $\alpha$ , an inducible nuclear protein. *J Biol Chem.* 2005;280:7444–51.
38. Yamazaki S, Muta T, Matsuo S, Takeshige K. Stimulus-specific induction of a novel nuclear factor- $\kappa$ B regulator, I $\kappa$ B $\alpha$ , via Toll/interleukin-1 receptor is mediated by mRNA stabilization. *J Biol Chem.* 2005;280:1678–87.
39. Round JL, Mazmanian SK. The gut microbiota shapes intestinal immune responses during health and disease. *Nat Rev Immunol.* 2009;9:313–23.
40. Shreiner AB, Kao JY, Young VB. The gut microbiome in health and in disease. *Curr Opin Gastroenterol.* 2015;31:69–75.
41. Erb-Downward JR, Thompson DL, Han MK, Freeman CM, McCloskey L, Schmidt LA, Young VB, Toews GB, Curtis JL, Sundaram B, et al. Analysis of the lung microbiome in the “healthy” smoker and in COPD. *PLoS One.* 2011;6:e16384.
42. Morris A, Beck JM, Schloss PD, Campbell TB, Crothers K, Curtis JL, Flores SC, Fontenot AP, Ghedin E, Huang L, et al. Comparison of the respiratory microbiome in healthy nonsmokers and smokers. *Am J Respir Crit Care Med.* 2013;187:1067–75.
43. Hilty M, Burke C, Pedro H, Cardenas P, Bush A, Bossley C, Davies J, Ervine A, Poulter L, Pachter L, et al. Disordered microbial communities in asthmatic airways. *PLoS One.* 2010;5:e8578.
44. Pulvirenti G, Parisi GF, Giallongo A, Papale M, Manti S, Savasta S, Licari A, Marseglia GL, Leonardi S. Lower airway microbiota. *Front Pediatr.* 2019;7:393.
45. Bousquet J, Chanez P, Lacoste JY, Barneon G, Ghavanian N, Enander I, Venge P, Ahlstedt S, Simony-Lafontaine J, Godard P, et al. Eosinophilic inflammation in asthma. *N Engl J Med.* 1990;323:1033–9.
46. Masur H, Rosen PP, Armstrong D. Pulmonary disease caused by *Candida* species. *Am J Med.* 1977;63:914–25.
47. Bhatt AP, Redinbo MR, Bultman SJ. The role of the microbiome in cancer development and therapy. *CA Cancer J Clin.* 2017;67:326–44.
48. Andrews MC, Reuben A, Gopalakrishnan V, Wargo JA. Concepts collide: genomic, immune, and microbial influences on the tumor microenvironment and response to cancer therapy. *Front Immunol.* 2018;9:946.
49. Yan X, Chu JH, Gomez J, Koenigs M, Holm C, He X, Perez MF, Zhao H, Mane S, Martinez FD, et al. Noninvasive analysis of the sputum transcriptome discriminates clinical phenotypes of asthma. *Am J Respir Crit Care Med.* 2015;191:1116–25.
50. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29:15–21.
51. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 2014;42:D633–42.
52. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 2014;15:R46.
53. Guo Y, Gifford DK. Modular combinatorial binding among human trans-acting factors reveals direct and indirect factor binding. *BMC Genomics.* 2017;18:45.
54. Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. SINCERA: a pipeline for single-cell RNA-seq profiling analysis. *PLoS Comput Biol.* 2015;11:e1004575.
55. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol.* 1995;57:289–300.
56. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. *Phys Rev E.* 2004;69:026113.
57. Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. *Phys Rev E.* 2004;70:066111.
58. Spakowicz D, Lou S, Barron B, Li T, Gomez JL, Liu Q, Grant N, Yan X, Hoyd R, Weinstock G, Chupp GL, Gerstein M. PRJNA611097: Cross-talk between microbes and genes inferred by RNA-seq of sputum from asthmatic patients (TaxID:

9606). RNA-seq of 115 sputum samples, single-cell RNA-seq of 11 sputum samples. National Center Biotechnol Information; (2020) ID: 611097.

59. Spakowicz D, Lou S, Barron B, Li T, Gomez JL, Liu Q, Grant N, Yan X, Hoyd R, Weinstock G, Chupp GL, Gerstein M. gersteinlab/decoasthma: decoasthma.v1.0 Zenodo (2020). <https://doi.org/10.5281/zenodo.3749840>.

### **Publisher's Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

