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

Daniel Spakowicz*[1,2,3,4], Shaoke Lou*[1], Brian Barron[1], Tianxiao Li[1], Jose L Gomez[5], Qing Liu[5], Nicole Grant[5], Xiting Yan[5], George Weinstock[2], Geoffrey L Chupp[5], Mark Gerstein[1,6,7,8]

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

ABSTRACT (337 words)

Sputum induction is a non-invasive method to evaluate the airway environment, particularly for asthma. RNA sequencing (RNAseq) can be used on sputum, but it can be challenging to interpret because sputum contains a complex and heterogeneous mixture of human cells and exogenous (microbial) material. In this study, we developed a methodology that integrates dimensionality reduction and statistical modeling to grapple with the heterogeneity. We use this to relate bulk RNAseq data from 115 asthmatic patients with clinical information, microscope images, and single-cell profiles. First, we mapped sputum RNAseq to human and exogenous sources. Next, we decomposed the human reads into cell-expression signatures and fractions of these in each sample; we validated the decomposition using targeted single-cell RNAseq and microscopy. We observed enrichment of immune-system cells (neutrophils, eosinophils, and mast cells) in severe asthmatics. Second, we inferred microbial abundances from the exogenous reads and then associated these with clinical variables --- e.g., *Haemophilus* was associated with increased white blood cell count and *Candida*a, with worse lung function. Third, we applied a generative model, Latent Dirichlet allocation (LDA), to identify patterns of gene expression and microbial abundances and relate them to clinical data. Based on this, we developed a method called LDA-link that connects microbes to genes using reduced-dimensionality LDA topics. We found a number of known connections, e.g. between *Haemophilus* and the gene IL1B, which is highly expressed by mast cells. In addition, we identified novel connections, including *Candida* and the calcium-signaling gene CACNA1E, which is highly expressed by eosinophils. These results speak to the mechanism by which gene-microbe interactions contribute to asthma and define a strategy for making inferences in heterogeneous and noisy RNAseq datasets.

INTRODUCTION

Linking high-dimensional, heterogeneous datasets

RNA sequencing (RNAseq) 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 RNAseq data in complex samples such as sputum. This can be appreciated by comparing sputum RNAseq to a more traditional experiment, e.g. blood RNAseq, where the sample can be collected consistently and that contains relatively well-defined cell types (Figure 1). In blood, the vast majority of RNAseq 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 defined, and it may contain RNA from microbes and other organisms that act as cryptic indicators of the environment. This combination of variables and dimensions often requires researchers to collapse the dimensions to appropriately de-noise the analysis. Here, we present such a strategy that uses a number of supervised and unsupervised techniques such as single-cell 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 many 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 Respiratory Phenotype consortium defined five subtypes of asthma [1]. Some of these subgroups respond differently to environmental and microbial triggers, such as fungal spores. Some fungi have well-defined effects in asthma, but the role of many microbes remains contentious. A simplified model assigns microbes to one of three categories: pathogenic organisms that cause inflammation, beneficial organisms that reduce inflammation, and those 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.

Inferring immune cell fractions from RNAseq 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 [2, 3]. A standard method for counting immune cells in sputum samples uses microscopy, but the resolution is limited to a few cell types [4]. 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 RNAseq reads from mixtures of cells into fractions of different immune cells [5]. 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 knockdown – 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 RNAseq data, an equivalent method does not exist for microbes. Although we can map microbe reads onto their genomes, this approach is imperfect because the genome databases are incomplete and assigning a read to a single genome can be complicated if it matches more than one equally well. 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 the relatively well-defined gene dimensionality-reduction methods (e.g., deconvolution to cell types).

In this paper, we use RNAseq of sputum samples from asthmatic patients to demonstrate dimensionality-reduction strategies and identify microbe-host relationships. We map RNAseq 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 sputum. Finally, we relate the human genes and microbes using a method we call LDA-link, which identifies relationships between genes, microbes, and cell types. These methods represent a general strategy for dealing with heterogeneous RNAseq 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 sample using RNAseq. The median read depth per sample was 47.5 million, which meets depth recommendations for analyses of this type [6]. We processed these reads through the exceRpt pipeline [7], 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 [2], 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[3]) 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 (Figure 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 (Figure 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 RNAseq reads and relate them to clinical phenotypes. We conceptualized the clinical information and RNAseq alignments as a series of tables (**Figure 1**). The clinical table includes patient data collected at the clinic, ***C,*** including age, weight, lung function tests, etc, with rows indexed by patient (*p*) 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 1,000 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**($\boldsymbol{X}_{\cdot,\boldsymbol{i}}$, $\boldsymbol{Y}_{\cdot,\boldsymbol{j}}$), where $\boldsymbol{X}_{\cdot,\boldsymbol{i}}$ is the *i*th column of table ***X*** and $\boldsymbol{X}_{\cdot,\boldsymbol{j}}$ 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 **R**($\boldsymbol{C}_{\cdot,\boldsymbol{age}}$, $\boldsymbol{M}_{\cdot,\boldsymbol{m}}$) across all patients. Similarly, we correlate the expression of a gene (e.g., *TLR4)* with microbe *Candida* **R**($\boldsymbol{G}_{\cdot,\boldsymbol{TLR4}}$, $\boldsymbol{M}_{\cdot,\boldsymbol{Candida}}$).

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 ***G*** ($\boldsymbol{N_p} \times \boldsymbol{N_g}$) into a cell-type fraction table, ***F*** ($\boldsymbol{N_p} \times \boldsymbol{N_f}$) , and a cell-type signatures table, **S** ($\boldsymbol{N_f} \times \boldsymbol{N_g}$) . 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 a topic distributions in patients ($\boldsymbol{\theta}^G$, $\boldsymbol{N_p} \times \boldsymbol{N_k}$) across a smaller number ($\boldsymbol{N_k}$=10) of topics and gene topic (and $\boldsymbol{\varphi}^G$, $\boldsymbol{N_k} \times \boldsymbol{N_g}$). This can also be done to the microbe table M and get $\boldsymbol{\theta}^M$ and $\boldsymbol{\varphi}^M$, and the gene and microbe topic can be correlated (e.g. **R(**$\boldsymbol{\theta^G_{:,g}}$, $\boldsymbol{\theta^M_{:,m}}$ **)** 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 [8, 9], using the de-noised inputs for each gene and microbe ($\boldsymbol{\varphi}^G$ and $\boldsymbol{\varphi}^M$). We call this method 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 regards 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 RNAseq reads as a mixture of immune cell types, each with a distinct expression profile, we deconvolved the Gene table (***G***) into a table of fractions of component cells type (***F***) and cognate cell-type signatures (***S***) by solving the formula ***G*** ~ ***F*** * ***S***. This method relies on knowing the signature gene-set in each cell type, which derived from the blood immune cell high quality profiles. To validate that we could apply these cell expression profiles to sputum, we generated several additional datasets including single-cell RNAseq (scRNAseq), microscopy, and unsupervised decomposition, and then compared the results to the deconvolution table ***F***. (Figure 2A, schema).

Evaluation of deconvolution results by scRNAseq

First, we performed scRNAseq on a cohort of similar sputum samples (five control and five asthmatic patients). The single-cell sequences clustered into four groups (Figure 2B, first and second panels). To determine whether the reference profiles that we used to deconvolve the bulk RNAseq recapitulate those found in the single-cell clusters, we co-clustered the reference profiles with the scRNAseq data (Figure 2B, third panel). The reference profiles split into the groups by lineage; for example, those in the lymphoid progenitor line co-clustered with cluster 2, and the myeloblast progenitor line co-clustered with cluster 4. This result suggests that the reference profiles accurately represent the cell types in sputum. The myeloid lineage cluster showed a significant difference in the number of cells between asthmatics and controls (Figure 2C). From this analysis, we concluded that (1) the blood-derived cell profiles appropriately fit the sputum cell types and (2) no additional cell types are needed to deconvolve the sputum bulk RNAseq data.

Evaluation of deconvolution results by microscopy

Second, we evaluated a subset 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. 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 $\boldsymbol{F}_f$ correlated well with the microscopy counts (Figure 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 ($\boldsymbol{R}(\boldsymbol{F}_{\cdot,\boldsymbol{f}}, \boldsymbol{C}_{\cdot,\boldsymbol{c}})$ for all patients). We found that the changes in fractions of several cell types were highly correlated with clinical features (Figure 2E). For example, the fraction of T-regulatory cells negatively correlated with the number of hospitalizations per year, suggesting a beneficial role of these cells in the management of asthma.

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-patients table, $\boldsymbol{\theta}^G$ ($\boldsymbol{N_p} \times \boldsymbol{N_k}$) with N_k=10 topics, as well as corresponding gene-topic table, $\boldsymbol{\varphi}^G(\boldsymbol{N_k} \times \boldsymbol{N_g})$, that are analogous to the supervised deconvolution tables ***F*** and ***S***. We correlated the cell-type fractions table with the gene topics fraction table ($\mathbf{R}(\boldsymbol{F}_{\cdot,\boldsymbol{f}}, \boldsymbol{\theta}_{\cdot,\boldsymbol{k}})$ for all patients, and found agreement between LDA and the cell-signature-based deconvolution for only the most prominent cell type, neutrophils (Figure 2D, topic 4). The top genes associated with topic 4 were enriched in the neutrophil chemotaxis pathway (Figure S8 B).

However, the remaining topics were comprised of multiple cell types. This suggests that LDA can identify distinct but partially overlapping features in ***G***. According to the clustering of $\boldsymbol{\theta}^G$, a subgroup of severely asthmatic patients was highly correlated with topic four (**Figure S8A**). The top-weighted genes in topic 4 were enriched for the pathways “neutrophil chemotaxis” and “asthma-related genes” (**Figure S8B**). These pathways were not enriched in the analogous cell-type-signatures 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 components 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 [10]; however, we did not observe this relationship for any cells in the cell-type-fractions table (***F***). Another method using a very different algorithm than LDA, non-negative matrix factorization (NMF), showed strong agreement with LDA (Figure S2, Nmf.1). This supports the use of supervised deconvolution methods as picking out interpretable signals that are different than those identified by unsupervised methods. Unsupervised decomposition should be considered a set of features distinct 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 (**Supplemental Table X**). 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 Bacterioidetes predominate [11]. 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.

Microbes correlations with clinical information and cell fractions

We correlated the microbe abundances to clinical information ($\boldsymbol{R}(\boldsymbol{M}_{\cdot,m}, \boldsymbol{C}_{\cdot,c})$ for all patients) (**Figure 3A**). *Haemophilus* was associated with increased total white blood cell numbers, as has been described previously [12]. *Candida* was associated with worse lung function test results (e.g., forced expiratory volume and forced vital capacity), which supports the association with a severe form of asthma characterized by eosinophilia [13].

We next correlated microbe abundances to human immune cell fractions ($\mathbf{R}(\boldsymbol{M}_{\cdot,m}, \boldsymbol{F}_{\cdot,f})$ for all patients) (**Figure 3B**). Several correlations demonstrated results with strong literature precedence. For example, studies have previously shown that *Haemophilus* associates with eosinophilia [14], and we observed a significant correlation between *Haemophilus* and the fraction of eosinophils. We also observed a significant correlation between *Haemophilus* and activated mast cells, suggesting an alternative route to *Haemophilus*-induced inflammation [15]. Moreover, the fungal genus *Candida* was also significantly correlated with eosinophils, even more strongly than *Haemophilus*. Pulmonary candidiasis has long been associated with allergic bronchial asthma and inflammation [16], however few lung microbiome studies have examined both bacterial and fungal signals. This highlights the need for a more comprehensive search of the lung microbiome and demonstrates the power of an RNAseq-based method that can report on all kingdoms with the same sample preparation.

Dimensionality reduction for microbes: clustering and networks

We attempted to de-noise the microbe table ($\boldsymbol{M}^{phylum}$) with a variety of dimensionality-reduction techniques. First, we collapsed the microbes by taxonomy, grouping them to the rank of phylum, and then hierarchically cluster the patients based on their phylum abundance (Figure 3C **HierClust**($\boldsymbol{M}^{phylum}$)). The hierarchical clustering showed that the phylum distributions formed three clusters of patients. We related these clusters to the clinical variable “asthma severity” and observed that cluster 2 was enriched for patients identified as having moderate or severe asthma. This cluster was characterized by the highest relative abundance of the phylum Proteobacteria (Figure 3C). Notably, the genus Haemophilus belongs to this phylum, consistent with the correlations observed at the genus rank (Figures 3A, 3B).

Similarly, we could de-noise the microbe table using a co-abundance network, by correlating the genus-level abundances ($\mathbf{R}(\boldsymbol{M}_{\cdot,m}, \boldsymbol{M}_{\cdot,m})$ and identifying significant modules (**Supplemental Figure Z**). 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 ($\mathbf{R}(\boldsymbol{\varphi}^{M}_{\cdot,m}, \boldsymbol{\varphi}^{M}_{\cdot,m})$ for all the topics) (**Figure 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 microbes and human cells in the airway remains contentious [17]. We feel this is partly due to the heterogeneous and noisy data from airway samples, where it is often difficult to find strong correlations using standard algorithms. We therefore sought to link genes to microbes via a new method called LDA-link.

LDA-link connects genes to microbes using a combination of linear correlation, unsupervised decomposition and an ensemble learning classifier. We hypothesized that the only strongest gene and microbe correlations would be observable through the noise in the RNAseq data. Therefore, we used these strong links as a training set to find other links, after taking steps to reduce the noise in the data. We reduced the noise using LDA and then identified links using a random forest classifier, described in more detail below and in the methods section.

To define the training, set we first related columns between the gene and microbe tables (**R**($G_{\cdot,g}$, $M_{\cdot,m}$)), yielding many low-scoring correlations. However, a relatively small number were strong (**R** > 0.4) and highly significant (p < 1E-5 after FDR correction) (**Figure 4A**). We selected the very strong correlations as true-positive links between genes and microbes in the training set, and non-correlated pairs (-0.05 < **R** < 0.05) as true-negative links. The genes involved in these strong correlations were enriched for pathways related to microbial interactions in the airway, including “Asthma & Bronchial Hypersensitivity” and “Respiratory Syncytial Virus Bronchiolitis” (**Figure 4B**), suggesting that the small set of strong linear correlations were relevant to asthma.

Next, we trained a random forest classifier on the linear correlations described above. To reduce the noise in the data, the features used as inputs to the classifier were the LDA topics for each gene and microbe ($\varphi^{G}_{\cdot,g}$, $\varphi^{M}_{\cdot,m}$). That is, for each gene-microbe pair, we concatenated the gene and microbe topics into a single vector (length 20). The Gini index showed the most important features in defining links between genes and microbes were gene topics #7 and #8, and microbe topic #1 (**Figure 4 C-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. It is tempting to speculate that these genes are strong predictors of a link between genes and microbes because they indicate when the presence of a microbe has triggered an inflammatory response.

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 summarizes a subset of the connections, showing in most cases several genes linked to each microbe (**Figure 5A**, for a complete list see **Supplemental Table X**). Notably, both fungi and bacteria showed these links, further highlighting the need to evaluate more than bacteria when performing microbiome experiments in the airway. The gene lactotransferrin was linked to *Aeromonas,* which has been associated with gastroenteritis and skin infections and has been previously reported to bind lactoferrin [18]. *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 [19]. *Haemophilus* was observed to be linked to NFKB Inhibitor Zeta, which is induced by the bacterial cell wall component lipopolysaccharide [20]. 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 expression of IL1B [21]. In addition to single gene-microbe pairs, we layered on pathway and cell deconvolution data to identify larger-scale effects of microbes.

Microbes were linked to genes that are enriched in pathways relating to auto-immunity and inflammation as well as cytokine receptors and their interactions (**Figure 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 signatures table ($\boldsymbol{S}_{f,g}$) to identify the immune cell types that are related to each microbe (**Figure 5C**). We observed the *Haemophilus*-IL1B linkage in monocytes and mast cells. Samples containing *Haemophilus* triggered more activated mast cells according to its cell fraction (**Figure 5C inset**) [22-25]. 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.

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 RNAseq 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 combined unsupervised and supervised approaches to de-noise the data while retaining interpretability.

The field is increasingly appreciating the role of the airway microbiome in the development of disease. Commensal microbiota have been shown in other contexts to be strong regulators of host immune system development and homeostasis [26]. Disturbances in the composition of commensal bacteria can result in imbalanced immune responses and 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 [27]). Investigating the microbiota in the lower respiratory tract is a relatively new field in comparison to 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 [11]. A limited number of reports have investigated the changes in the lung microbiota between healthy, non-smoking and smoking individuals as well as in patients suffering from cystic fibrosis, chronic obstructive pulmonary disease, or asthma [2, 28-30]. Despite emerging data on the airway microbiota, little is known about the role of the lung microbiome in modulating pulmonary mucosal immune responses. LDA-link can find relationships between microbes and genes and link them to immune cells and their responses.

The linkages identified here suggest major 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 linkages is needed, these results represent observations that would be missed by analyses that do not deconvolve RNAseq data into cell fractions, or that analyze only human RNAseq reads. We expect LDA-link to be broadly useful in relating heterogeneous or noisy RNAseq data.

METHODS

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 [1]. 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 & 634847) and cDNA was created with the SMARTer Stranded RNAseq Kit (Cat. Nos. 634836). Samples were sequenced using an Illumina HiSeq 4000 with 2x125 bp reads, with an average of 47.5 million reads per sample.

RNAseq processing by exceRpt

An adapted version of the software package exceRpt [7] was used to process the sputum RNAseq data. Briefly, RNAseq reads were subjected to quality assessment using 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/). Identical reads were counted and collapsed to a single entry and reads containing N's were removed. Clipped, collapsed reads were mapped directly to the human reference genome (hg19) and pre-miRNA sequences using STAR [31]. Reads that did not align were mapped against a ribosomal reference library of bacteria, fungi, and archaea, compiled by the Ribosome Database Project [32], and then to genomes of bacteria, fungi, plants, and viruses, retreived from GenBank [32]. In cases where RNAseq 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 [7, 33].

Data tables notation

We use the following notation to define matrices associated with *p* patients (115) (Figure 1):

C: Clinical table ($N_p \times N_c$), c is the clinical index

G: Gene table ($N_p \times N_g$), bulk-RNA seq table before deconvolution,

M: Microbe abundance table ($N_p \times N_c$)

F: Cell fractions table ($N_p \times N_f$), resulting from the deconvolution of $\boldsymbol{G}_{p,g}$

S: Cell signatures table ($N_f \times N_g$), resulting from the deconvolution of $\boldsymbol{G}_{p,g}$

$\boldsymbol{\theta}^G$: Patient topic table ($N_p \times N_k$) after LDA inference based on gene table $\boldsymbol{G}_{p,g}$

$\boldsymbol{\varphi}^G$: Gene topic table ($N_k \times N_g$) after LDA inference based on gene table $\boldsymbol{G}_{p,g}$

$\boldsymbol{\theta}^M$: Patient topic table ($N_p \times N_k$) after LDA inference based on microbe table $\boldsymbol{M}_{p,m}$

$\boldsymbol{\varphi}^M$: Microbe topic table ($N_k \times N_m$) after LDA inference based on table $\boldsymbol{M}_{p,m}$

L: gene microbe linkage table ($N_g \times N_m$) 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 (lm22) using the CIBERSORT tool [5]. Briefly, a pre-defined set of characteristic gene expression patterns for each cell type was used to identify the fraction of each cell type given a mixture of expression by solving for the equation:

$$\mathbf{G} = \mathbf{F} * \mathbf{E}$$

Where **G** is the Gene table of human protein-coding gene expression from the exceRpt pipeline, **F** is the Cell Fraction table, and **E** is the characteristic gene expression calculated within CIBERSORT. 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 **E** and **F** to **G** demonstrated that all samples were significant at α = 0.05.

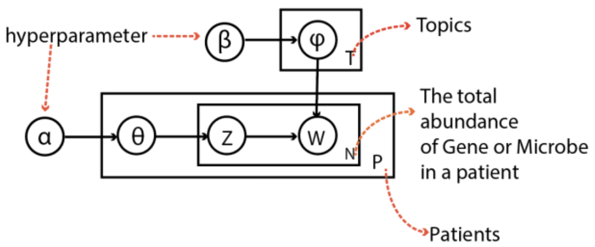
Following the solution of **F**, a Cell Signature table **S** was calculated to estimate the expression of g genes, as opposed to the reduced set appropriate for the characteristic expression evaluation, by solving the equation:

$$\mathbf{S} = \mathbf{G} * \mathbf{F}$$

Decomposition though a generative model

The Gene table **G** was decomposed using LDA and Non-negative Matrix Factorization (NMF).
For LDA, the abundance values for bulk RNAseq and exogenous RNA were scaled down to reduce computation intensity during sampling. More simply, the RPM expression values were converted to integers, and then divided by 10. The max value was set to 1,000.

Given each patient (p), all of the genes and microbes were treated like corpus of words in the traditional LDA application. The word (w) was gene or microbe, and the word count was gene expression or microbe abundances. We built LDA models for genes and microbes, respectively.



$$\varphi_k \sim Dirichlet\ (\beta)$$
$$\theta_p \sim Dirichlet\ (a)$$
$$Z_{p,n} \sim Multinomial\ (\theta_p)$$
$$w_{p,n} \sim Multinomial\ (\varphi_{z_{p,n}})$$

Given $p, w, k, v, N_p, \mathrm{N}_w, N_k,\ N_v,\ \alpha, \beta, \mathrm{Z}, \theta, \varphi, \mathrm{W},$ where p, w, k, v denote a patient, a word in a document, a topic and a word in the corpus respectively; N_p is the number of documents(patients), N_w is the number of words (gene or microbe) in a document, N_k is the number of topics (set as 10), N_v is the corpus for all the documents; $\alpha\ \ (N_k \text{ dimensional vector})\text{and } \beta$ (N_v-dimensional vector) are the hyper parameters for $\theta\ \big(N_p\ \times$ $N_k, \text{the distributoin of topics in documents}\big)$ and $\varphi\ (N_k \times N_v, \text{the distribution of word for topics})$ W is an N_w-dimensional vector that denotes the word (gene or microbe expression) in a document (patients). Z is the N_w-dimensional vector of integers between 1 and N_k for the topic of word in a document.

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

$$\mathcal{P}(Z, W;\ \alpha, \beta) = \int_{\varphi} \prod_{i=1}^{N_k} \mathcal{P}(\varphi_i; \beta) \prod_{j=1}^{N_p} \prod_{t=1}^{N_w} \mathcal{P}\left(W_{j,t} \middle| \varphi_{Z_{j,t}}\right) d\varphi \int_{\theta} \prod_{i=1}^{N_p} \mathcal{P}(\theta_i; \alpha) \prod_{j=1}^{N_w} \mathcal{P}\left(Z_{i,j} \middle| \theta_i\right) d\theta$$

$$= \prod_{k=1}^{K} \frac{\Delta(n_{\cdot,k} + \beta)}{\Delta(\beta)} \prod_{s=1}^{S} \frac{\Delta(n_{s,\cdot} + \alpha)}{\Delta(\alpha)}$$

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

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 nth word token in the mth document, and also assume that its word symbol is the vth word in the vocabulary, the conditional probability can be inferred as follows:

$$\mathcal{P}\left(Z_{m,n} = k | Z_{\neg(m,n)}, W;\ \alpha, \beta\right) = \frac{\mathcal{P}(Z,W;\alpha,\beta)}{\mathcal{P}\left(Z_{\neg(m,n)},W;\alpha,\beta\right)} = \frac{\mathcal{P}(w,z)}{\mathcal{P}\left(w_{m,n}, w_{\neg(m,n)},\ z_{\neg(m,n)}\right)} = \frac{\mathcal{P}(w,z)}{\mathcal{P}\left(w_{\neg(m,n)},\ z_{\neg(m,n)}\right)} \cdot \frac{1}{\mathcal{P}\left(w_{m,n}=t\right)}$$

$$\propto \frac{\mathcal{P}(w,z)}{\mathcal{P}\left(w_{\neg(m,n)},\ z_{\neg(m,n)}\right)}$$

$$= \frac{\prod_{k=1}^{K} \frac{\Delta\left(n_{\cdot,k} + \beta\right)}{\Delta(\beta)} \prod_{p=1}^{P} \frac{\Delta\left(n_{p,\cdot} + \alpha\right)}{\Delta(\alpha)}}{\prod_{k=1}^{K} \frac{\Delta\left(n_{\neg(m,n),k} + \beta\right)}{\Delta(\beta)} \prod_{p=1}^{P} \frac{\Delta\left(n_{p}, \neg(m,n) + \alpha\right)}{\Delta(\alpha)}}$$

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

After sampling, the expectation of the θ (doc → topic) and φ(topic → word) matrix can be inferred as follows given the symmetric hyper-parameters α and β were used:

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

$$\varphi_{k,v} = \frac{n_{k,v} + \beta}{\sum_{i=1}^{V} n_{k,i} + N_v \beta}$$

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

Single-cell RNAseq

Sputum cells were separated on a Fluidigm C1 medium-sized channel. The mRNA was purified from approximately 500pg-1ng 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 200-500bp DNA fragments by sonication (Covaris, Massachusetts, USA), and then indexed and size validated by LabChip GX. Two nM libraries were loaded onto Illumina version 3 flow cells and sequenced using 75bp single-end sequencing on an Illumina HiSeq 2000 according to Illumina protocols. Data were cleaned, processed, aligned, and quantified following the SINCERA pipeline [34].

Pathogen-to-host linkage 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:

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

Gene-microbe correlations with p-values less than 1e-5 (absolute correlation greater than 0.4) were chosen as the positive links in a training set. Negative links in the training set were defined as an absolute correlation of less than 0.05. This approach resulted in 302 positive and 650,398 negative links. A random forest algorithm was trained on this set, which can accommodate the highly unbalanced dataset as well as potentially identify non-linear links between genes and microbes. Down-sampling and up-sampling techniques were tested but did not significantly improve the model. In the final model, we adopted the upscaling technique and tested it using cross-validation. The positive dataset was upscaled to very high levels. We use 2-fold cross validation to validate the performance. Simply, we randomly select half training data to train the model, and use the remaining records to test the performance and repeat this for ten times. The AUC and AUPR were 0.994 and 0.996 on average, respectively.

Microbe co-abundance network

The raw abundance M and LDA microbe topic matrices φ^M , which represent the microbe's weight to each topic, were generated.

The correlation network between different microbes was calculated using Pearson correlation. The cutoff to define a co-abundance edge was 0.8 for $\boldsymbol{R}(\varphi^M_{\cdot,m}, \varphi^M_{\cdot,m})$ and 0.3 for $\boldsymbol{R}(, \varphi^M_{\cdot,m}$. The microbe network modules, which were densely connected themselves but sparsely connected to other modules, were clustered based on between-ness [35] and the other algorithms; we also tested label propagation and fast greedy algorithms [36].

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.

DATA ACCESS

Sputum bulk-cell RNAseq data can be found under the bioproject SRRXXXXX and sputum single-cell RNAseq data at SRRYYYYYY.

ACKNOWLEDGMENTS

This work was supported by a National Library of Medicine fellowship to DS (5T15LM007056-28) and an NHLBI grant to GC (1R01HL118346-01). The authors would like to thank the support of the Yale High Performance Computing services (Grace, Ruddle, Farnam) and Yale Center for Genome Analysis.

DISCLOSURE DECLARATION

The authors declare no conflicts of interest. [[[check with GC]]]

ABBREVIATIONS

LDA: Latent Dirichlet Allocation; PCA: Principle Component Analysis;

FIGURES:

Figure 1. RNAseq alignment summary for control and asthmatic sputum, showing fractions of reads that aligned to different biotypes. Alignments to the protein-coding biotype were used to generate the gene expression matrix (G), which was then deconvolved into a cell fraction matrix (F) and cell expression (E). The exogenous reads were used to generate a microbial profile matrix (M). These matrices were then related to the clinical phenotype matrix (P) for biological insight.

Figure 2. Deconvolution of RNAseq human reads into cell fractions using cell signature deconvolution. A) Schematic showing the imputation of a cell fraction matrix and cell-specific expression matrix. B) Imputed cell fractions were validated using microscopy; Cell fractions were then correlated with SARP cluster for two major cell type: (C) Machrophases.M0 and (D) Mast cell activiated. E) the cell fraction of LM22 gene signature deconvolution are correlated with the topic distribution of samples from LDA analysis. F) G) tSNE analysis and clustering using single cell RNAseq from Asthmatic patient and control. H) The fraction of single cells for different cell types clusters between Asthmatic patient and Controls.

Figure 3. Exogenous RNAseq analysis. (A). The correlations between microbes abundance and cell fraction based on LM22 signature (B) The correlation between microbes abundance and clinical information. (C) The microbes abundance shows clear patterns that associated with Asthmatic severity. (D) The co-abundance network and overlay with the associated topics of microbes.

Figure 4. Prediction of cross-talk between microbe and gene. (A) The diagram to combine linear and LDA-based non-linear algorithms to identify gene microbe linkages. (B) simple correlation to identify strong linkages between microbes and genes. Gene set over represent analysis for genes. X-axis is the -log(p-value) . (C) the importance of features (LDA topics for gene and microbes) in the RandomForest model by Gini index. The top 20 associated gene in topics 8 (D) and topic 7 (E) of genes, and topic 1 (F) of microbes.

Figure 5. The linkage between microbes and genes reflects the heterogeneity of different cell types. (A). Linkages between microbes and genes. (B) The linkages indicated by the cell proportion of certain types.

SUPPLEMENTAL INFORMATION

Figure S1:

The distribution of cell fraction of sample in different group.

Figure S2

The heatmap between NMF component and cell fraction from LM22.

Figure S3

Overall view of correlation of all the extracellular organism with clinical features.

Figure S4

Heatmap of the correlation of topics (gene and microbe) with clinical information.

Figure S5

Co-abundance network based on correlation of abundance.\

Figure S6

Top associated microbes in microbe topics

Figure S7

Top associated genes in gene topics

Figure S8

(A) The topic distribution of patient. (B) The gene enrichment analysis of top genes in topic 4.

Figure S9

Main pathways get involved by microbe linked genes.

REFERENCES

1. Yan X, Chu JH, Gomez J, Koenigs M, Holm C, He X, et al. Noninvasive analysis of the sputum transcriptome discriminates clinical phenotypes of asthma. Am J Respir Crit Care Med. 2015;191(10):1116-25. Epub 2015/03/13. doi: 10.1164/rccm.201408-1440OC. PubMed PMID: 25763605; PubMed Central PMCID: PMCPMC4451618.
2. Huang YJ, Nariya S, Harris JM, Lynch SV, Choy DF, Arron JR, et al. The airway microbiome in patients with severe asthma: Associations with disease features and severity. J Allergy Clin Immunol. 2015;136(4):874-84. Epub 2015/07/30. doi: 10.1016/j.jaci.2015.05.044. PubMed PMID: 26220531; PubMed Central PMCID: PMCPMC4600429.
3. Gibson PG, Girgis-Gabardo A, Morris MM, Mattoli S, Kay JM, Dolovich J, et al. Cellular characteristics of sputum from patients with asthma and chronic bronchitis. Thorax. 1989;44(9):693-9. Epub 1989/09/01. doi: 10.1136/thx.44.9.693. PubMed PMID: 2588203; PubMed Central PMCID: PMCPMC462047.
4. 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(2 Pt 1):475-8. Epub 2000/02/15. doi: 10.1164/ajrccm.161.2.9903097. PubMed PMID: 10673188.
5. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12(5):453-7. Epub 2015/03/31. doi: 10.1038/nmeth.3337. PubMed PMID: 25822800; PubMed Central PMCID: PMCPMC4739640.
6. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74. Epub 2012/09/08. doi: 10.1038/nature11247. PubMed PMID: 22955616; PubMed Central PMCID: PMCPMC3439153.

7. Rozowsky J, Kitchen RR, Park JJ, Galeev TR, Diao J, Warrell J, et al. exceRpt: A Comprehensive Analytic Platform for Extracellular RNA Profiling. Cell Syst. 2019;8(4):352-7 e3. Epub 2019/04/09. doi: 10.1016/j.cels.2019.03.004. PubMed PMID: 30956140.

8. 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. Epub 2015/05/26. doi: 10.1016/j.ejpb.2015.05.012. PubMed PMID: 26004819.

9. Welling SH, Refsgaard HHF, Brockhoff PB, Clemmensen LH. Forest Floor Visualizations of Random Forests. arXiv e-prints [Internet]. 2016 May 01, 2016. Available from: https://ui.adsabs.harvard.edu/abs/2016arXiv160509196W.

10. Payne BA, Chinnery PF. Mitochondrial dysfunction in aging: Much progress but many unresolved questions. Biochim Biophys Acta. 2015;1847(11):1347-53. Epub 2015/06/09. doi: 10.1016/j.bbabio.2015.05.022. PubMed PMID: 26050973; PubMed Central PMCID: PMCPMC4580208.

11. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. Nature. 2007;449(7164):804-10. Epub 2007/10/19. doi: 10.1038/nature06244. PubMed PMID: 17943116; PubMed Central PMCID: PMCPMC3709439.

12. Essilfie AT, Simpson JL, Horvat JC, Preston JA, Dunkley ML, Foster PS, et al. Haemophilus influenzae infection drives IL-17-mediated neutrophilic allergic airways disease. PLoS Pathog. 2011;7(10):e1002244. Epub 2011/10/15. doi: 10.1371/journal.ppat.1002244. PubMed PMID: 21998577; PubMed Central PMCID: PMCPMC3188527.

13. Bousquet J, Chanez P, Lacoste JY, Barneon G, Ghavanian N, Enander I, et al. Eosinophilic inflammation in asthma. N Engl J Med. 1990;323(15):1033-9. Epub 1990/10/11. doi: 10.1056/NEJM199010113231505. PubMed PMID: 2215562.

14. 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(5):598-605. Epub 2003/04/12. doi: 10.1165/rcmb.2002-0138OC. PubMed PMID: 12689921.

15. 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(7):1843-51. Epub 2010/06/29. doi: 10.1002/eji.201040559. PubMed PMID: 20583030; PubMed Central PMCID: PMCPMC3581154.

16. Masur H, Rosen PP, Armstrong D. Pulmonary disease caused by Candida species. Am J Med. 1977;63(6):914-25. Epub 1977/12/01. PubMed PMID: 343588.

17. Mathieu E, Escribano-Vazquez U, Descamps D, Cherbuy C, Langella P, Riffault S, et al. Paradigms of Lung Microbiota Functions in Health and Disease, Particularly, in Asthma. Front Physiol. 2018;9:1168. Epub 2018/09/25. doi: 10.3389/fphys.2018.01168. PubMed PMID: 30246806; PubMed Central PMCID: PMCPMC6110890.

18. Ascencio F, Ljungh A, Wadstrom T. Characterization of lactoferrin binding by Aeromonas hydrophila. Appl Environ Microbiol. 1992;58(1):42-7. Epub 1992/01/01. PubMed PMID: 1311545; PubMed Central PMCID: PMCPMC195170.

19. 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(7):4188-99. Epub 2004/06/24. doi: 10.1128/IAI.72.7.4188-4199.2004. PubMed PMID: 15213163; PubMed Central PMCID: PMCPMC427436.

20. Park CY, Heo JN, Suk K, Lee WH. Sodium azide suppresses LPS-induced expression MCP-1 through regulating IkappaBzeta and STAT1 activities in macrophages. Cell Immunol. 2017;315:64-70. Epub 2017/04/11. doi: 10.1016/j.cellimm.2017.02.007. PubMed PMID: 28391993.

21. Hildebrand D, Bode KA, Riess D, Cerny D, Waldhuber A, Rommler F, et al. Granzyme A produces bioactive IL-1beta through a nonapoptotic inflammasome-independent pathway. Cell Rep. 2014;9(3):910-7. Epub 2014/12/02. doi: 10.1016/j.celrep.2014.10.003. PubMed PMID: 25437548.

22. Chapman SJ, Khor CC, Vannberg FO, Rautanen A, Segal S, Moore CE, et al. NFKBIZ polymorphisms and susceptibility to pneumococcal disease in European and African populations. Genes Immun. 2010;11(4):319-25. Epub 2009/10/03. doi: 10.1038/gene.2009.76. PubMed PMID: 19798075; PubMed Central PMCID: PMCPMC3051152.

23. Baldwin AS, Jr. The NF-kappa B and I kappa B proteins: new discoveries and insights. Annu Rev Immunol. 1996;14:649-83. Epub 1996/01/01. doi: 10.1146/annurev.immunol.14.1.649. PubMed PMID: 8717528.

24. Motoyama M, Yamazaki S, Eto-Kimura A, Takeshige K, Muta T. Positive and negative regulation of nuclear factor-kappaB-mediated transcription by IkappaB-zeta, an inducible nuclear protein. J Biol Chem. 2005;280(9):7444-51. Epub 2004/12/25. doi: 10.1074/jbc.M412738200. PubMed PMID: 15618216.

25. Yamazaki S, Muta T, Matsuo S, Takeshige K. Stimulus-specific induction of a novel nuclear factor-kappaB regulator, IkappaB-zeta, via Toll/Interleukin-1 receptor is mediated by mRNA stabilization. J Biol Chem. 2005;280(2):1678-87. Epub 2004/11/04. doi: 10.1074/jbc.M409983200. PubMed PMID: 15522867.

26. Round JL, Mazmanian SK. The gut microbiota shapes intestinal immune responses during health and disease. Nat Rev Immunol. 2009;9(5):313-23. Epub 2009/04/04. doi: 10.1038/nri2515. PubMed PMID: 19343057; PubMed Central PMCID: PMCPMC4095778.

27. Shreiner AB, Kao JY, Young VB. The gut microbiome in health and in disease. Curr Opin Gastroenterol. 2015;31(1):69-75. Epub 2014/11/14. doi: 10.1097/MOG.0000000000000139. PubMed PMID: 25394236; PubMed Central PMCID: PMCPMC4290017.

28. Erb-Downward JR, Thompson DL, Han MK, Freeman CM, McCloskey L, Schmidt LA, et al. Analysis of the lung microbiome in the "healthy" smoker and in COPD. PLoS One. 2011;6(2):e16384. Epub 2011/03/03. doi: 10.1371/journal.pone.0016384. PubMed PMID: 21364979; PubMed Central PMCID: PMCPMC3043049.

29. Morris A, Beck JM, Schloss PD, Campbell TB, Crothers K, Curtis JL, et al. Comparison of the respiratory microbiome in healthy nonsmokers and smokers. Am J Respir Crit Care Med. 2013;187(10):1067-75. Epub 2013/03/16. doi: 10.1164/rccm.201210-1913OC. PubMed PMID: 23491408; PubMed Central PMCID: PMCPMC3734620.

30. Hilty M, Burke C, Pedro H, Cardenas P, Bush A, Bossley C, et al. Disordered microbial communities in asthmatic airways. PLoS One. 2010;5(1):e8578. Epub 2010/01/07. doi: 10.1371/journal.pone.0008578. PubMed PMID: 20052417; PubMed Central PMCID: PMCPMC2798952.

31. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21. Epub 2012/10/30. doi: 10.1093/bioinformatics/bts635. PubMed PMID: 23104886; PubMed Central PMCID: PMCPMC3530905.

32. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 2014;42(Database issue):D633-42. Epub 2013/11/30. doi: 10.1093/nar/gkt1244. PubMed PMID: 24288368; PubMed Central PMCID: PMCPMC3965039.

33. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3):R46. Epub 2014/03/04. doi: 10.1186/gb-2014-15-3-r46. PubMed PMID: 24580807; PubMed Central PMCID: PMCPMC4053813.

34. 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(11):e1004575. Epub 2015/11/26. doi: 10.1371/journal.pcbi.1004575. PubMed PMID: 26600239; PubMed Central PMCID: PMCPMC4658017.

35. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys Rev E. 2004;69(2). PubMed PMID: WOS:000220255500019.

36. Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Phys Rev E. 2004;70(6). PubMed PMID: WOS:000226299200018.

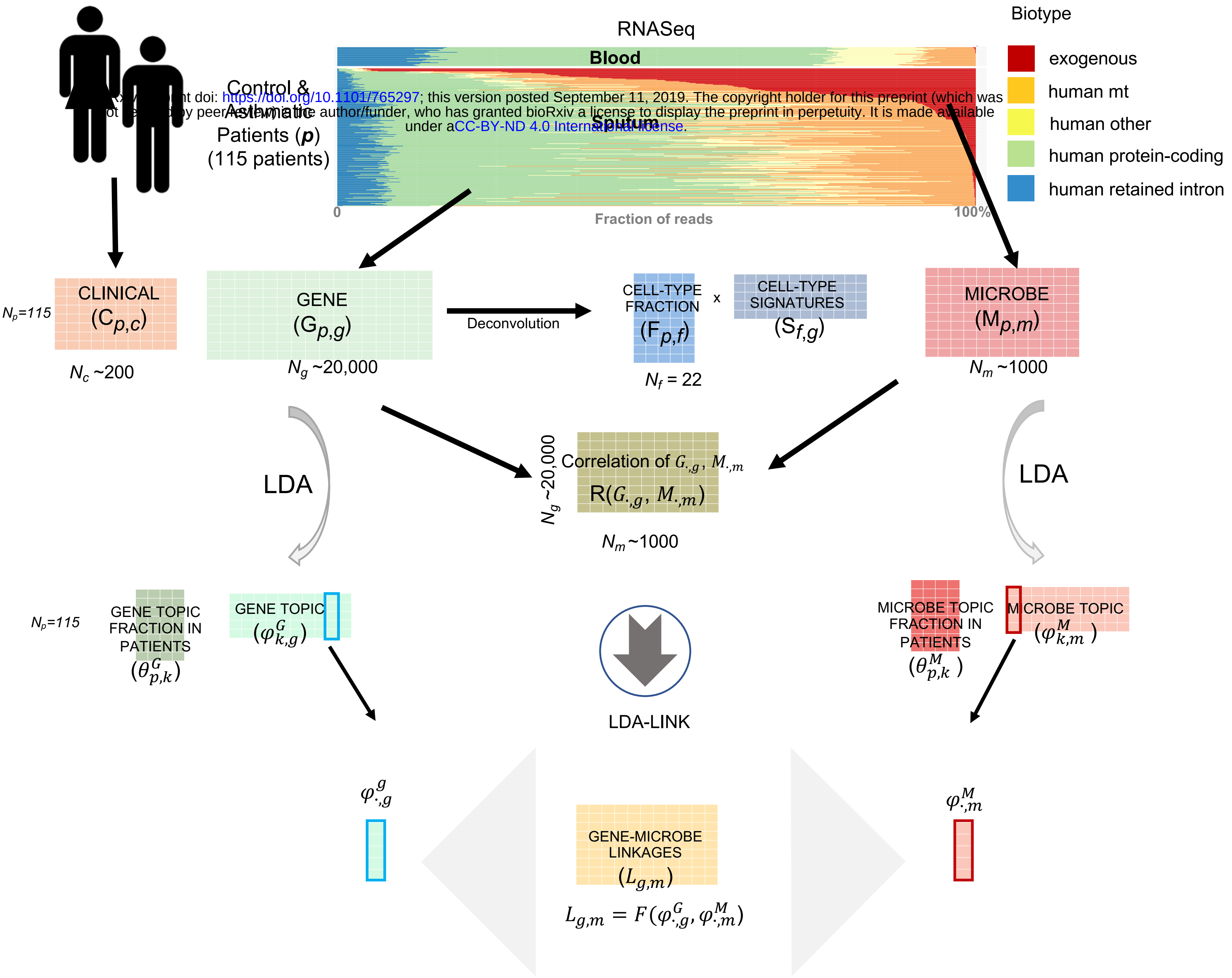

Control &
Asthmatic
Patients (p)
(115 patients)
RNASeq
Blood
Sputum
0
Fraction of reads
100%
Biotype
🟥 exogenous
🟧 human mt
🟨 human other
🟩 human protein-coding
🟦 human retained intron
N_p=115
CLINICAL
$(C_{p,c})$
N_c ~200
GENE
$(G_{p,g})$
N_g ~20,000
Deconvolution
CELL-TYPE FRACTION
$(F_{p,f})$
x
CELL-TYPE SIGNATURES
$(S_{f,g})$
N_f = 22
MICROBE
$(M_{p,m})$
N_m ~1000
LDA
Correlation of $G_{\cdot,g}$, $M_{\cdot,m}$
$R(G_{\cdot,g}, M_{\cdot,m})$
N_g ~20,000
N_m~1000
LDA
N_p=115
GENE TOPIC FRACTION IN PATIENTS
$(\theta^G_{p,k})$
GENE TOPIC
$(\varphi^G_{k,g})$
LDA-LINK
MICROBE TOPIC FRACTION IN PATIENTS
$(\theta^M_{p,k})$
MICROBE TOPIC
$(\varphi^M_{k,m})$
$\varphi^g_{\cdot,g}$
GENE-MICROBE LINKAGES
$(L_{g,m})$
$L_{g,m} = F(\varphi^G_{\cdot,g}, \varphi^M_{\cdot,m})$
$\varphi^M_{\cdot,m}$

Figure 1

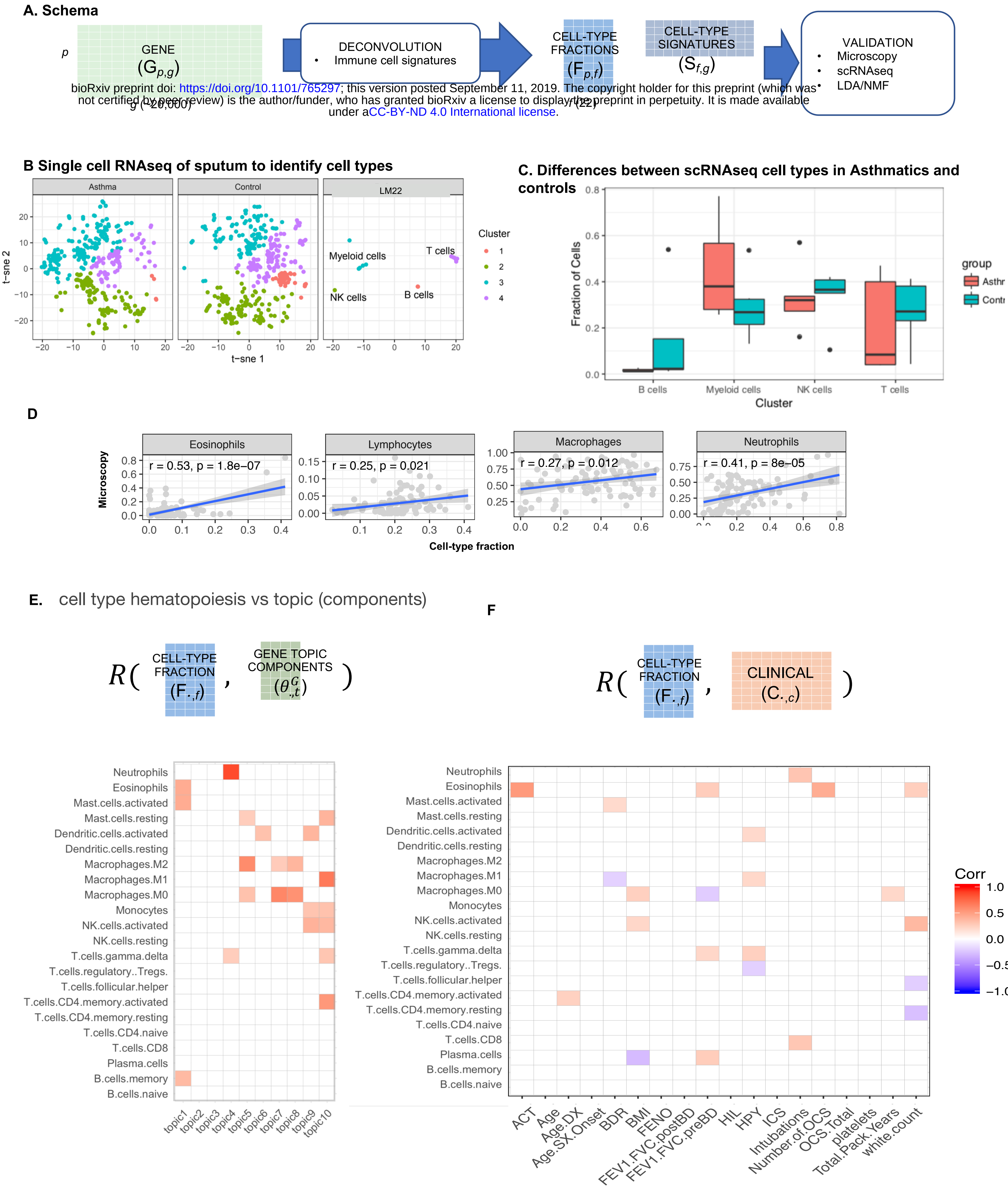

A. Schema
p
GENE
$(G_{p,g})$
g (~20,000)
DECONVOLUTION
• Immune cell signatures
CELL-TYPE FRACTIONS
$(F_{p,f})$
f (22)
CELL-TYPE SIGNATURES
$(S_{f,g})$
VALIDATION
• Microscopy
• scRNAseq
• LDA/NMF
B Single cell RNAseq of sputum to identify cell types
Asthma
Control
LM22
t−sne 2
t−sne 1
20
10
0
−10
−20
Myeloid cells
T cells
NK cells
B cells
Cluster
1
2
3
4
C. Differences between scRNAseq cell types in Asthmatics and controls
Fraction of Cells
0.8
0.6
0.4
0.2
0.0
B cells
Myeloid cells
NK cells
T cells
Cluster
group
Asthr
Contr
D
Microscopy
Eosinophils
r = 0.53, p = 1.8e−07
Lymphocytes
r = 0.25, p = 0.021
Macrophages
r = 0.27, p = 0.012
Neutrophils
r = 0.41, p = 8e−05
Cell-type fraction
E. cell type hematopoiesis vs topic (components)
$R($ CELL-TYPE FRACTION $(F_{.,f})$, GENE TOPIC COMPONENTS $(\theta^{G}_{.,t})$ $)$
F
$R($ CELL-TYPE FRACTION $(F_{.,f})$, CLINICAL $(C_{.,c})$ $)$
Neutrophils
Eosinophils
Mast.cells.activated
Mast.cells.resting
Dendritic.cells.activated
Dendritic.cells.resting
Macrophages.M2
Macrophages.M1
Macrophages.M0
Monocytes
NK.cells.activated
NK.cells.resting
T.cells.gamma.delta
T.cells.regulatory..Tregs.
T.cells.follicular.helper
T.cells.CD4.memory.activated
T.cells.CD4.memory.resting
T.cells.CD4.naive
T.cells.CD8
Plasma.cells
B.cells.memory
B.cells.naive
topic1
topic2
topic3
topic4
topic5
topic6
topic7
topic8
topic9
topic10
ACT
Age
Age.DX
Age.SX.Onset
BDR
BMI
FENO
FEV1.FVC.postBD
FEV1.FVC.preBD
HIL
HPY
ICS
Intubations
Number.of.OCS
OCS.Total
platelets
Total.Pack.Years
white.count
Corr
1.0
0.5
0.0
−0.5
−1.0

Figure 2

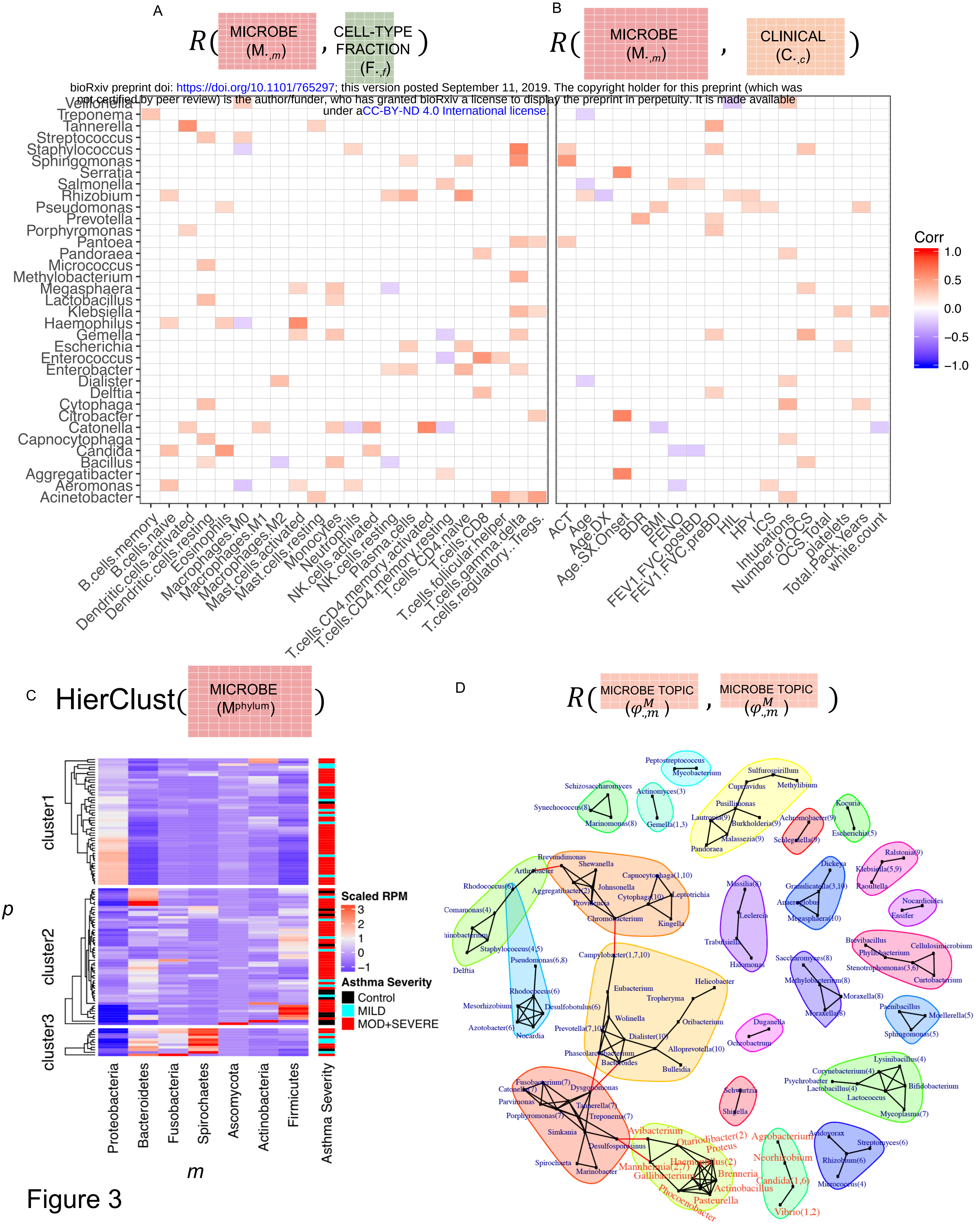

A
R(MICROBE (M.,m) , CELL-TYPE FRACTION (F.,f))
B
R(MICROBE (M.,m) , CLINICAL (C.,c))
Veillonella
Treponema
Tannerella
Streptococcus
Staphylococcus
Sphingomonas
Serratia
Salmonella
Rhizobium
Pseudomonas
Prevotella
Porphyromonas
Pantoea
Pandoraea
Micrococcus
Methylobacterium
Megasphaera
Lactobacillus
Klebsiella
Haemophilus
Gemella
Escherichia
Enterococcus
Enterobacter
Dialister
Delftia
Cytophaga
Citrobacter
Catonella
Capnocytophaga
Candida
Bacillus
Aggregatibacter
Aeromonas
Acinetobacter
B.cells.memory
B.cells.naive
Dendritic.cells.activated
Dendritic.cells.resting
Eosinophils
Macrophages.M0
Macrophages.M1
Macrophages.M2
Mast.cells.activated
Mast.cells.resting
Monocytes
Neutrophils
NK.cells.activated
NK.cells.resting
Plasma.cells
T.cells.CD4.memory.activated
T.cells.CD4.memory.resting
T.cells.CD4.naive
T.cells.CD8
T.cells.follicular.helper
T.cells.gamma.delta
T.cells.regulatory..Tregs.
ACT
Age
Age.DX
Age.SX.Onset
BDR
BMI
FENO
FEV1.FVC.postBD
FEV1.FVC.preBD
HIL
HPY
ICS
Intubations
Number.of.OCS
OCS.Total
platelets
Total.Pack.Years
white.count
Corr
1.0
0.5
0.0
−0.5
−1.0
C
HierClust(MICROBE (Mphylum))
p
cluster1
cluster2
cluster3
Proteobacteria
Bacteroidetes
Fusobacteria
Spirochaetes
Ascomycota
Actinobacteria
Firmicutes
Asthma Severity
m
Scaled RPM
3
2
1
0
−1
Asthma Severity
■ Control
■ MILD
■ MOD+SEVERE
D
R(MICROBE TOPIC (φ^M_.,m) , MICROBE TOPIC (φ^M_.,m))
Peptostreptococcus
Mycobacterium
Sulfurospirillum
Cupriavidus
Methylibium
Schizosaccharomyces
Actinomyces(3)
Synechococcus(8)
Marinomonas(8)
Gemella(1,3)
Pusillimonas
Lautropia(9)
Burkholderia(9)
Malassezia(9)
Pandoraea
Achromobacter(9)
Schlegelella(9)
Kocuria
Escherichia(5)
Brevundimonas
Arthrobacter
Shewanella
Capnocytophaga(1,10)
Johnsonella
Cytophaga(10)
Leptotrichia
Aggregatibacter(2)
Providencia
Chromobacterium
Kingella
Rhodococcus(6)
Comamonas(4)
hinobacterium
Staphylococcus(4,5)
Pseudomonas(6,8)
Delftia
Massilia(8)
Leclercia
Trabulsiella
Halomonas
Dickeya
Granulicatella(3,10)
Anaeroglobus
Megasphaera(10)
Ralstonia(9)
Klebsiella(5,9)
Raoultella
Nocardioides
Ensifer
Campylobacter(1,7,10)
Eubacterium
Tropheryma
Helicobacter
Wolinella
Oribacterium
Dialister(10)
Alloprevotella(10)
Phascolarctobacterium
Bacteroides
Bulleidia
Rhodococcus(6)
Mesorhizobium
Desulfobotulus(6)
Azotobacter(6)
Nocardia
Prevotella(7,10)
Saccharomyces(8)
Methylobacterium(8)
Moraxella(8)
Moraxella(8)
Brevibacillus
Phyllobacterium
Cellulosimicrobium
Stenotrophomonas(3,6)
Curtobacterium
Paenibacillus
Moellerella(5)
Sphingomonas(5)
Duganella
Ochrobactrum
Fusobacterium(7)
Catonella(7)
Dysgonomonas
Parvimonas
Tannerella(7)
Porphyromonas(7)
Treponema(7)
Simkania
Desulfosporosinus
Spirochaeta
Marinobacter
Schwartzia
Shigella
Corynebacterium(4)
Lysinibacillus(4)
Psychrobacter
Lactobacillus(4)
Bifidobacterium
Lactococcus
Mycoplasma(7)
Avibacterium
Otariodibacter(2)
Proteus
Agrobacterium
Haemophilus(2)
Mannheimia(2,7)
Gallibacterium
Brenneria
Actinobacillus
Phocoenobacter
Pasteurella
Neorhizobium
Candida(1,6)
Vibrio(1,2)
Acidovorax
Streptomyces(6)
Rhizobium(6)
Micrococcus(4)

Figure 3

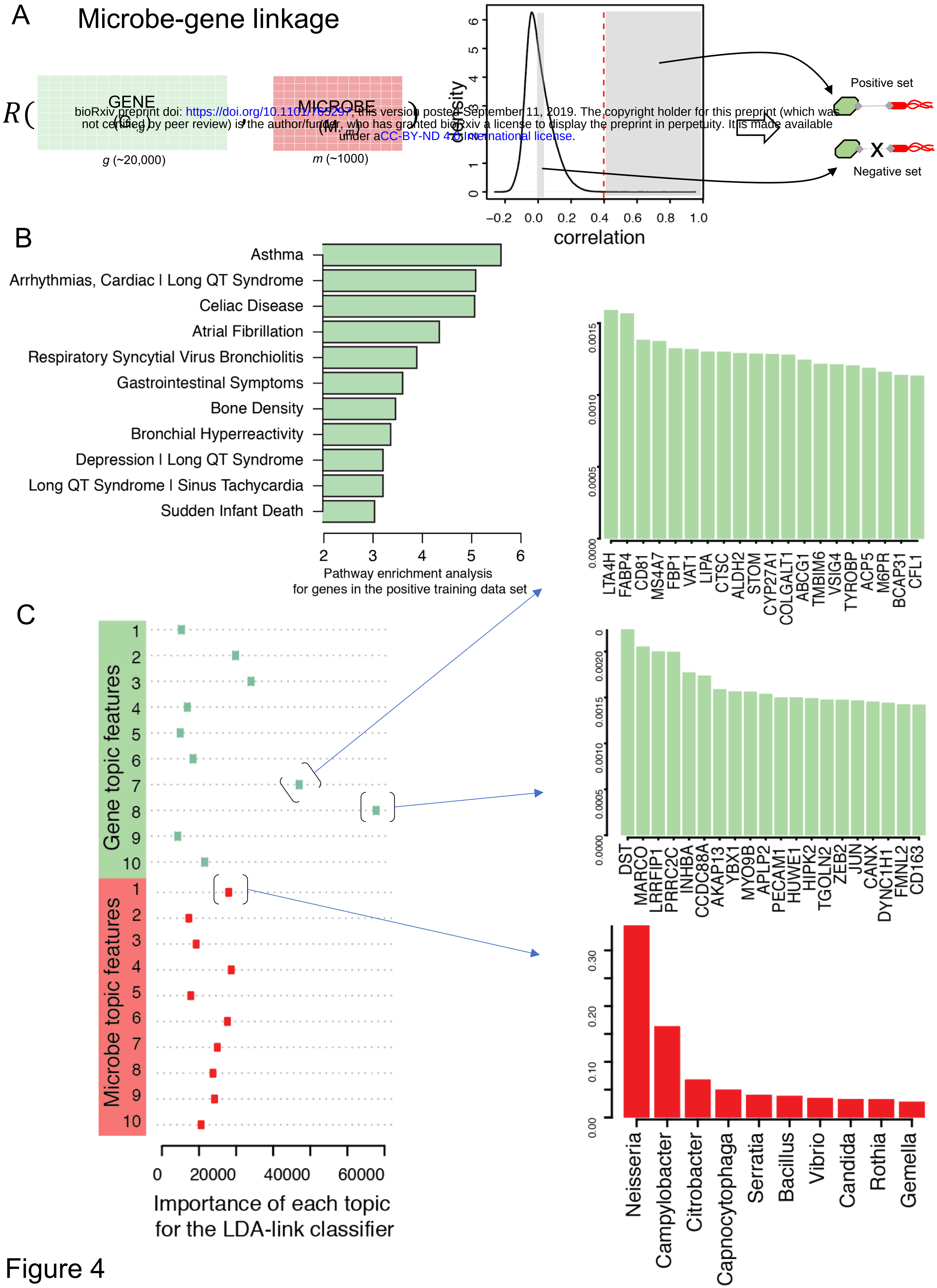

A
Microbe-gene linkage
R(
GENE
(G, g)
,
MICROBE
(M, m)
)
g (~20,000)
m (~1000)
density
0
1
2
3
4
5
6
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
correlation
Positive set
Negative set
B
Asthma
Arrhythmias, Cardiac | Long QT Syndrome
Celiac Disease
Atrial Fibrillation
Respiratory Syncytial Virus Bronchiolitis
Gastrointestinal Symptoms
Bone Density
Bronchial Hyperreactivity
Depression | Long QT Syndrome
Long QT Syndrome | Sinus Tachycardia
Sudden Infant Death
2
3
4
5
6
Pathway enrichment analysis
for genes in the positive training data set
C
Gene topic features
1
2
3
4
5
6
7
8
9
10
Microbe topic features
1
2
3
4
5
6
7
8
9
10
0
20000
40000
60000
Importance of each topic
for the LDA-link classifier
D
0.0000
0.0005
0.0010
0.0015
LTA4H
FABP4
CD81
MS4A7
FBP1
VAT1
LIPA
CTSC
ALDH2
STOM
CYP27A1
COLGALT1
ABCG1
TMBIM6
VSIG4
TYROBP
ACP5
M6PR
BCAP31
CFL1
E
0.0000
0.0005
0.0010
0.0015
0.0020
0
DST
MARCO
LRRFIP1
PRRC2C
INHBA
CCDC88A
AKAP13
YBX1
MYO9B
APLP2
PECAM1
HUWE1
HIPK2
TGOLN2
ZEB2
JUN
CANX
DYNC1H1
FMNL2
CD163
F
0.00
0.10
0.20
0.30
Neisseria
Campylobacter
Citrobacter
Capnocytophaga
Serratia
Bacillus
Vibrio
Candida
Rothia
Gemella

Figure 4

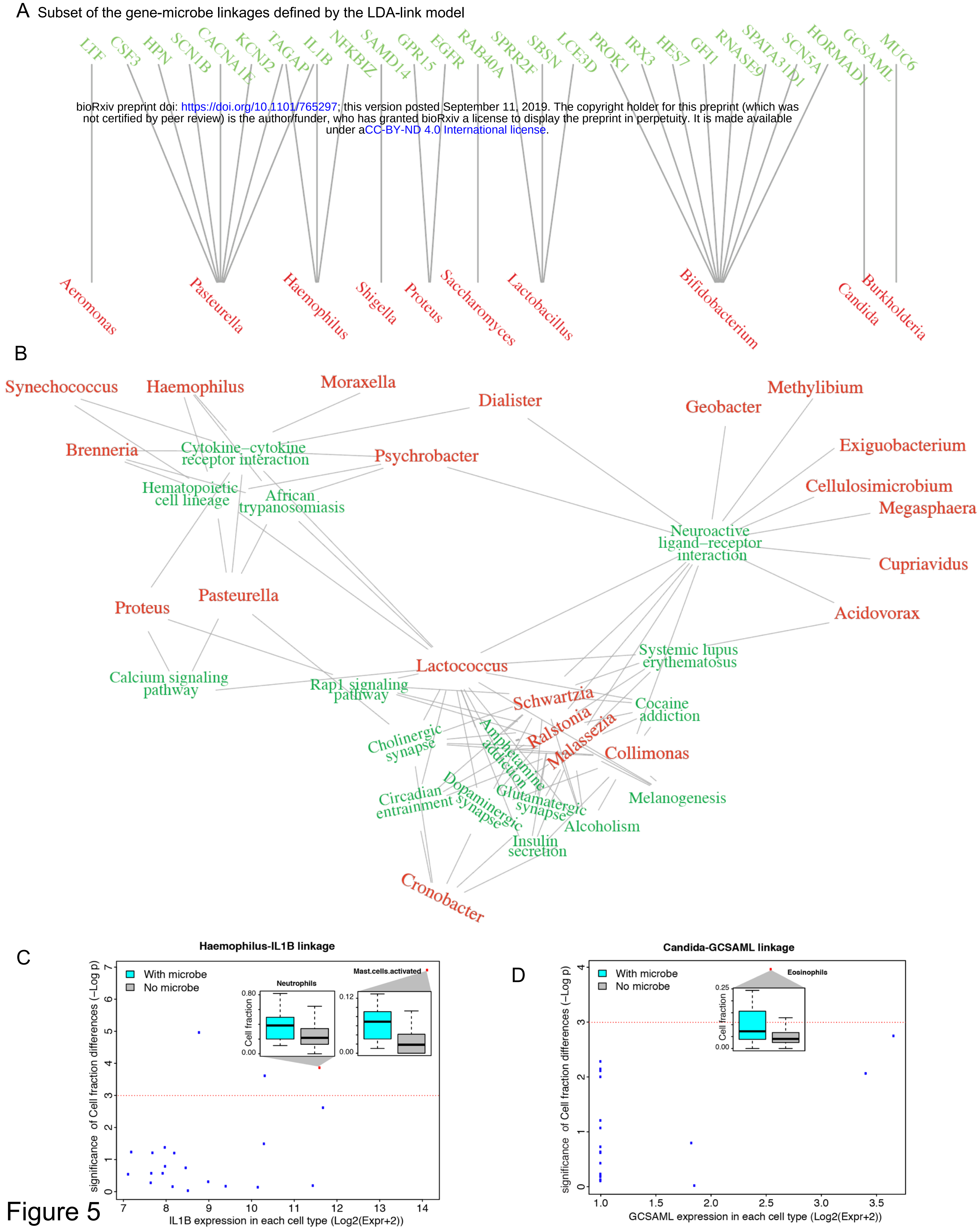

A Subset of the gene-microbe linkages defined by the LDA-link model
LTF
CSF3
HPN
SCN1B
CACNA1E
KCNJ2
TAGAP
IL1B
NFKBIZ
SAMD14
GPR15
EGFR
RAB40A
SPRR2F
SBSN
LCE3D
PROK1
IRX3
HES7
GFI1
RNASE9
SPATA31D1
SCN5A
HORMAD1
GCSAML
MUC6
Aeromonas
Pasteurella
Haemophilus
Shigella
Proteus
Saccharomyces
Lactobacillus
Bifidobacterium
Candida
Burkholderia
B
Synechococcus
Haemophilus
Moraxella
Dialister
Geobacter
Methylibium
Brenneria
Cytokine−cytokine receptor interaction
Psychrobacter
Exiguobacterium
Cellulosimicrobium
Megasphaera
Hematopoietic cell lineage
African trypanosomiasis
Neuroactive ligand−receptor interaction
Cupriavidus
Proteus
Pasteurella
Acidovorax
Systemic lupus erythematosus
Calcium signaling pathway
Lactococcus
Rap1 signaling pathway
Schwartzia
Cocaine addiction
Cholinergic synapse
Amphetamine addiction
Ralstonia
Malassezia
Collimonas
Circadian entrainment
Dopaminergic synapse
Glutamatergic synapse
Melanogenesis
Alcoholism
Insulin secretion
Cronobacter
C
Haemophilus-IL1B linkage
With microbe
No microbe
Neutrophils
Mast.cells.activated
Cell fraction
0.80
0.00
0.12
0.00
significance of Cell fraction differences (−Log p)
0
1
2
3
4
5
6
7
7
8
9
10
11
12
13
14
IL1B expression in each cell type (Log2(Expr+2))
D
Candida-GCSAML linkage
With microbe
No microbe
Eosinophils
Cell fraction
0.25
0.00
significance of Cell fraction differences (−Log p)
0
1
2
3
4
1.0
1.5
2.0
2.5
3.0
3.5
GCSAML expression in each cell type (Log2(Expr+2))

Figure 5