Deconvolution of sputum RNA-seq from asthma patients reveals microbe and immune cell interactions

Daniel J Spakowicz*
Shaoke Lou*
...
Geoff Chupp
Mark Gerstein

Abstract

Asthma is a highly heterogeneous disease and many of its clinical manifestations are resistant to treatment. Induced sputum from asthmatic patients is a non-invasive sample that has been shown to containing features that correlate with clinical parameters, however it poses experimental challenges because of its heterogeneity. We analyze sputum-derived RNA-seq data from 115 asthmatic and control patients and deconvolve it to its component microbes and human cells using a variety of methods including single-cell sequencing. We demonstrate that the deconvolved data gives stronger correlations with clincal features than in aggregate. Microbes and immune cell fractions are correlated with each other and with clinical features. Finally, we build a model to link microbes and human cell expression pathways to clinical features using machine learning on latent dirichlet allocation topics and show a subset of patients with severe asthma associated with microbes XXX and YYY and host expression of ZZZ. This demonstrates the utility of evaluating sputum as a mixture of many interacting cell as a means of understanding asthma heterogeneity.

Introduction

Asthma afflicts over 300 million people worldwide and approximately 30 million in the United States. For reasons that are largely unknown, the prevalence of asthma has risen to epidemic proportions over the past five decades [@masoli_global_2004], resulting in roughly 15 billion dollars in health expenditures in the US each year [@lugogo_epidemiology_2006]. While the understanding of disease pathogenesis has increased in recent years, the morbidity related to asthma remains high, accounting for 10 million school absences each year and limitations to

physical activity reported by approximately half of asthma patients [@bousquet_public_2005]. Efforts to develop better therapeutics are hampered by the heterogeneity of the disease, the source of which remains poorly understood. Recently suggested as a potential source of this heterogeneity is the airway microbiome [@huang_microbiome_2015].

Sputum has long been studied for its ability to report on asthma phenotypes, typically by determining by counting the relative amounts of different cell types present (e.g. [PMC462047]). In some cases, mRNA levels of single genes have been used to infer the presennce of difficult to quantify cell types, such as Th17 cells [doi.org/10.1186/1465-9921-7-135]. More recently, studies have demonstrated a link with another cell type present in the sputum: microbes. Sputum RNAseq data can therefore be considered a complex mixture of expression by different cell types, each with a different profile, and disease phenotype may be a direct consequence of the relative abundances of one cell type or another.

Here we deconvolve RNA-seq of the sputum of asthmatic patients to identify and quantify human cell types as well as the non-human community of organisms, and show that the relative amounts of these cells are correlated with clinical features. This work speaks to the phenotypic heterogeneity of asthmatic patients and may provide insight into the biological mechanisms that drive those differences.

Methods

Sample collection and sequencing

Sputum induction was performed with hypertonic saline, the mucus plugs dissected away from saliva, the cellular fraction separated and the RNA purified as described previously. Briefly, RNA was purified using the All-in-One purification kit (Norgen Biotek) and its integrity assayed by Agilent bioanalyzer (Agilent Technologies, Santa Clara, CA). Ribosomal depletion was performed with the RiboGone-Mammalian kit (Clontech Cat. Nos. 634846 & 634847). Samples were fragmented to an insert size of 150-200 bp cand the SMARTer Stranded RNA-Seq Kit (Cat. Nos. 634836) used to generate the cDNA library. The cDNA library is then amplified and indexed adapters are added. Libraries that meet appropriate cut-offs for both are quantified by qRT-PCR using a commercially available kit (KAPA Biosystems) and insert size distribution determined with the Perkin Elmer LabChip GX or Agilent Bioanalyzer. Samples with a yield of ≥0.5 ng/ul are used for sequencing.

Flow Cell Preparation and Sequencing: Sample concentrations are normalized to 10 nM and loaded onto Illumina High-output flow cell at a concentration that yields 200 million passing filter clusters per lane. Samples are sequenced using 75bp paired-end sequencing on an Illumina HiSeq 2500 according to Illumina protocols. The 6bp index is read during an additional sequencing read that automatically follows the completion of read 1. Data generated during

sequencing runs are simultaneously transferred to the YCGA high-performance computing cluster. A positive control (prepared bacteriophage Phi X library) provided by Illumina is spiked into every lane at a concentration of 0.3% to monitor sequencing quality in real time.

Ten ng RNA was amplified using random primers and the WT-Ovation Pico RNA amplification System (NuGen, San Carlos, CA). Samples were sequenced using an Illumina HiSeq 4000 with 2x125 bp reads, with an average of 40 million reads per sample.

Table 1. Patient Characteristics

Group	Control		Asthma	
Asthma Severity		MILD	MODERATE	SEVERE
Num Indiv	16	21	41	44
Mean Age (St. Dev)	44.6 (15.2)	44.3 (17.2)	50.1 (16.6)	46 (12.8)
Perc Female	62.5	76.2	80.5	70.5
BMI	26.7	24.9	28.9	33.2
BDR	3.5	5.0	8.8	12.8
ACT Score	19.4	20.0	17.5	12.7
Median Percent Eosinophils	1.7	5.1	7.5	9.5
Perc White	88	76	83	45
Perc Black	6	14	10	32
Percent Other	0	10	7	23

Table 1: Patient characteristics

RNAseq processing by exceRpt

An adapted version of the software package exceRpt [@_excerpt_????], was used to process and conservatively search for exogenous sequences within RNA-seq data. Briefly, RNA-seq reads are subjected to quality-assessment using the FastQC software v.0.10.1 [@_babraham_????] both prior to and following 3' adapter clipping. Adapters were removed using FastX v.0.0.13 [@_fastx-toolkit_????]. Identical reads were counted and collapsed to a single entry and reads containing N's are removed. Clipped, collapsed reads are mapped directly to the human reference genome and pre-miRNA sequences using STAR [@dobin_star:_2012]. Reads that did not align are mapped against a ribosomal reference library of bacteria, fungi and archaea, compiled by Ribosome Database Project [@cole_ribosomal_2014]. Remaining reads are aligned to genomes of other organisms including bacteria, fungi, plants and viruses, retreived from GenBank [@benson_genbank_2013].

Microbial abundance counting and normalization

Reads mapping to taxa were normalized to the total sequencing output for each sample and presented as the number reads mapping to each taxon per million sequencer reads. Reads that mapped to multiple reference genomes were assigned to the phylogenetic tree node shared by all genomes to which the read mapped. For example, if a read mapped equally-well to two species in the genus Bacterioides, the read would be assigned to the genus node.

Single-cell RNAseq

##SKL feature selection to identify Esophil percentage and severity associated Microbial
There are evidences show the exogenous microbes will affect the development of Asthma in different, even adverse way: some bad and some are good (need further elucidate here). To further investigate the severity associated exogenous microbes, we evaluate the importance of microbes on the performance of classification of different severities.

Firstly, we filtered the microbes that has been detected in less than 5(10) samples, and had 148(112) microbes left for further analysis. Considering the definition of severity is according the xxx medicine dosage, it will be affected by the patient's person preferences and doctors advices, which leads the intermediate group less accurate. Hence, we took only two extreme severity classes: mild and severity, and try to use microbes to predict the severity.

Due to the limit number of samples and a large number of features, the 'overfitting' will easily affect the results. To reduce the risk of 'overfitting' and identify the most associated microbes, we combined different feature selection frameworks to discriminate the most importances microbes.

Pathogen to host linkage identification (or maybe some other header)

[[SKL: some data briefs, the first draft will put the important process step, then will refine before submit to Mark. mixtured with some results now, will clean (todo)
Dataset I used to do this analysis is: gencode-vstnorm-combat_clinical.Rdata, there are some values (RPM) are negative, we only remove negative expression values, which means the low expressed gene are removed]]

LDA analysis of microbes and gene expression

The gene expression values for the bulk RNA-Seq and exogenous RNA are scaled down to to reduce computation intensity when do samping. Simply, the RPM expression value are convert to integer and then divided by 10, and max value was set to 1000 if it is greater than that. Then LDA model with 10 topics are optimized using Gibbs sampling.

The direct pearson correlation of bulk RNA-Seq and exogenous RNA are calculated, if the correlation is greater than 0.4, we use this as the positive links between gene and microbes, because the hypothesis is that the strong linkage can be detected by linearly correlation though we don't know the mechanism of this interaction. Then we define the negative dataset if the absolute correlation is less than 0.05.

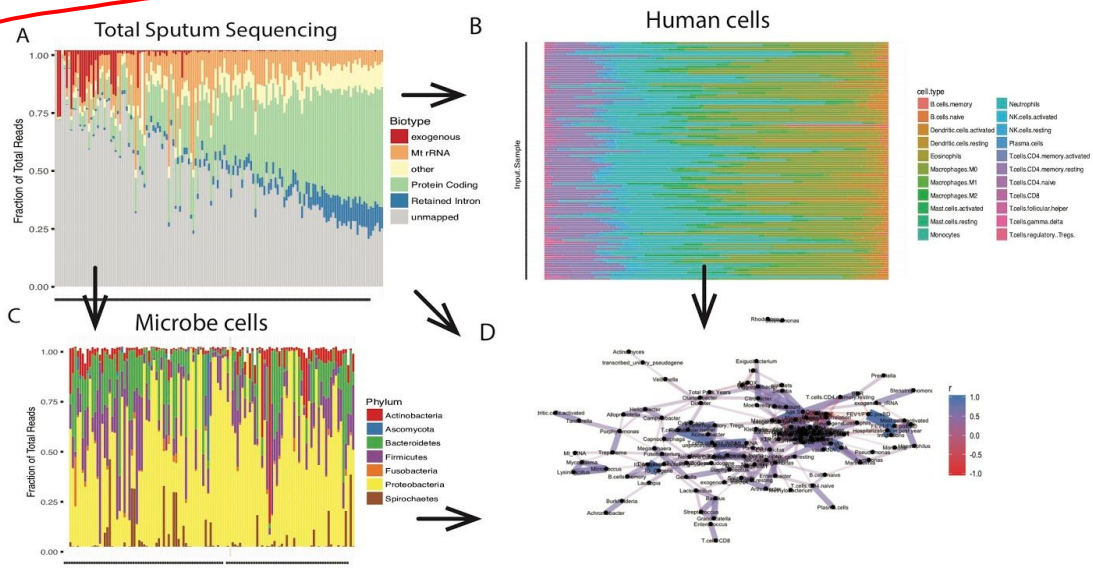
This results in 302 interactions, and 650398 negative controls in the dataset. Since the dataset is extremely unbalanced, and also we need to use non-linear linkage detection algorithm, randomForest becomes the best choice, also because it can handle unbalanced dataset in some degrees. In addition, we tested downsampling and upsampling techniques in parallel. Compared with directly using unbalanced dataset, downsampling seems removed some biases of unbalancement, but it cannot fully take advantage of all the negative dataset. In the final model, we adopt the upscaling technique and tested using the cross-validations. The positive dataset is upscaled to a very high levels, we choose records with half of unique positive data, and testing with the remaining records with another half unique positive data. This can potentially avoid information leaking from the training model. The AUC and AUPR for this 10-times 2-fold cross-validation, is 0.9943860 and 0.9961792 in average respectively.

The model using all the upscaled dataset are trained and predict all the combination of gene to microbes. The linkages discovered by our model has 1399 genes and 45 microbes involved by using prob=0.9 as the cut off. (3219 genes and 61 microbes for 0.5 as the cutoff, no need mention in the main text, just for reference)

Results

The RNA isolated from sputum samples from 113 patients were sequenced with a median of 47.5 million reads per sample. 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 1A, green bars), which is consistent with other RNA sequencing efforts on samples of this type [/ref]. A median of 0.7% of the input reads aligned to exogenous sources, with some samples containing as much as 28.1% exogenous reads. A large portion of the reads remained unmapped to any references (median = 27.8%, min = 5.6, max = 90)(Figure 1A, grey bars).

Figure 1. RNAseq alignment summary for control and asthmatic sputum, showing A) fractions of reads that aligned to different biotypes as well as unmapped reads. The protein-coding biotype was deconvolved to fractions of human cell types (B), and the exogenous reads into fractions of different microbes (C).



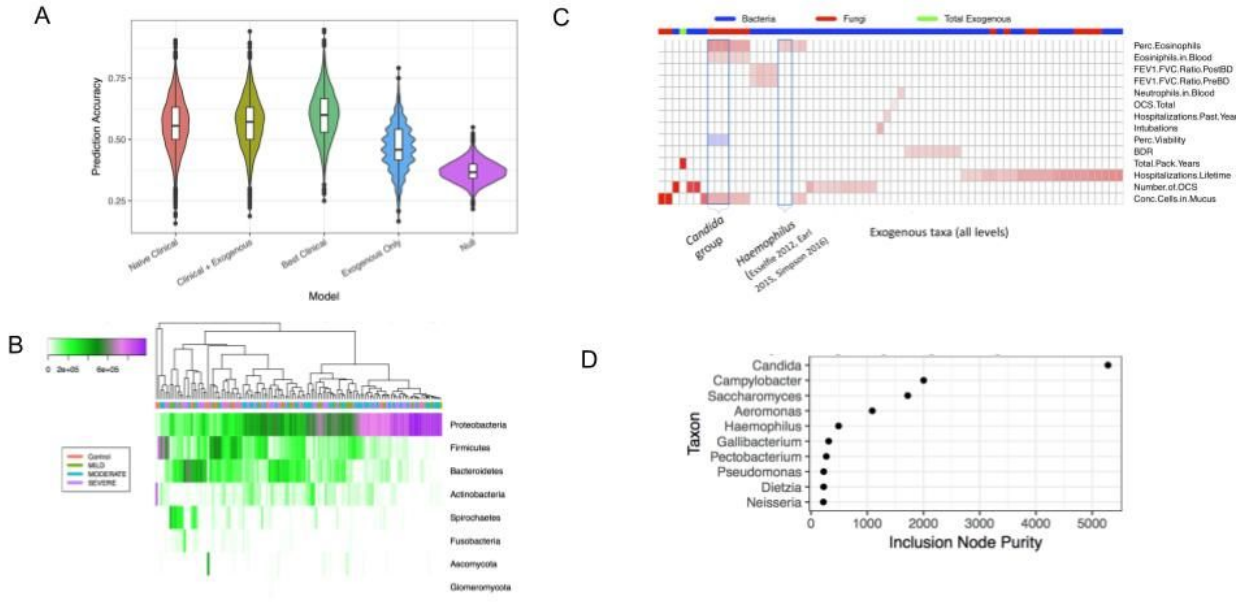
The reads aligning to protein-coding regions of the genome were deconvolved to component cells (Fig1B). The deconvolved cell fractions included an average of XX cell types per sample, with the majority of the cells (YY%) being neutrophils (Z%), macrophages () and eosinophils, which significantly correlates with microscopy-based measurements collected on a subset of the samples (Figure S1). Reads that did not align to the human reference genome were aligned to exogenous rRNA databases and reference genomes [/cite{exceRpt}] (Figure 1C). Observed were bacteria, archaea and fungi, with bacteria being the most abundant. The fractions of human cell-types and microbe cell types were then correlated with clinical varaiables (Figure 1D), described in more detail below.

Human-aligned reads

Exogenous reads

The alpha diversity of these samples was not signficantly different between the different asthma severity groups (Figure 2), as defined by the amount of fluticasone or equivalent per day to

control symptoms (mild = >200 ug, moderate = 200-800 ug, severe = >800 ug). The Fisher's alpha, Shannon and Simpson diversity metrics showed slightly higher diversity in mild asthmatics relative to other groups, though with a wide distrubution of values for both asthmatics and controls. Other studies have observed lower alpha diversity in asthmatics (n = 6) relative to controls (n = 8) using transcriptomics of nasal cavity swabs from children and adolescents [@castro-nallar_integrating_2015]. It is possible that the differences observed here are due to the wider distribution of patient ages, sampling of the sputum rather than the nasal cavity, or larger number of samples. Notably, more fungi were observed in our study than by Castro-Nallar et al.



The dominant phyla observed in the samples was Proteobacteria, followed by Firmicutes and Bacteroidetes (Figure 3). The abundance of Proteobacteria is in contrast to observations from the gut where Bacterioides predominate [@turnbaugh_human_2007]. Also notable was the presence of two phyla of fungi among the eight most abundant overall, though in lower abundance than many of the bacterial phyla. Though the asthma severity categories were not significantly different in their alpha diversities, significant category enrichment when clustering by the beta-deversity metric Bray-Curtis distance was observed (dengrogram cut height 0.7, fisher's exact test p-value = `r round(fishtmp, 2)`)(Figure 7). In particular, one of the three major clusters observed was significantly depleted in control samples (permutation test p-value = 0.0085) and significantly enriched in moderate asthmatics (permutation test p-value = 0.0058). This group has moderate levels of both Proteobacteria and Firmucutes, but the highest Bacteroidetes, Spirochaetes and Fusobacteria levels in the cohort. However, this cluster was not signficantly different from the other clusters in any continuous clincal variables after

accounting for multiple hypothesis testing. We therefore sought to identify if specific taxa were significantly correlated with the clinical parameters by regression approaches.

Microbial ribosomal RNA abundances at all taxonomic levels were correlated with continuous clinical variables. After controlling for the effects of age, body mass index and gender `r nrow(betas.red)` (`r round(nrow(betas.red)/ length(continuous), 2) * 100`% of total) clinical variables were significantly associated with one or more of `r ncol(betas.red)` `r round(ncol(betas.red) / (ncol(texo)-1), 2) *100`% of total) exogenous taxa (FDR < 0.05)(Figure 4). This included the total signal for exogenous sequences, which was strongly positively associated with the total pack years of smoking for the patients. Interestingly, none of the individual taxa were associated with the total pack years of smoking, perhaps suggesting an overall effect of smoking as increasing the total microbial load without selecting for a subset of the organisms. Alternative explanations include that the chronic inflammation in asthmatic lungs has phenotypic overlap with the inflammation caused by smoking, leading the environment to be similar between smoking and non-smoking asthmatics [@larsen_chronic_2015].

Microbial taxa were associated with healthier metrics as well as with potentially pathogenic roles. In the case of spirometry lung function metrics, the ratio of Forced Expiratory Volume in one second (FEV1) to the Forced Vital Capacity (FVC) was positively associated with members of the genus _Pseudomonas._ This suggests that _Pseudomonas_ is associated with a reduction in obstructive defects to the airway, which was true both before and after treatment with bronchodilators. However, _Pseudomonas_ was not associated with the response to bronchodilators (BDR); rather three bacterial groups were: the family Ruminococcaceae, the genus Fusobacteria and its family and order, as well as the genus _Prevotella._ This result agrees with previously reported observation that _Prevotella_ does not promote Toll-like receptor 2-independent lung inflammation, whereas members of the phylum Proteobacteria did, including _Haeomophilus_ and _Moraxella_ [@larsen_chronic_2015].

There were far greater numbers of taxa correlated with negative health effects. Interestingly, roughly one third of the significant correlations were with fungi, highlighting the importance of analyzing more than the 16S of bacteria. In particular, the number of hospitalizations that the patient has experienced correlated with both fungi and bacteria in roughly equal proportions. Proteobacteria taxa such as _Escherichia coli_ were observed, as well as the fungal orders of Glomeraleas and Pleosporales. Glomerales is an order of arbuscular mycorrhizal fungi not known to be associated with humans. The order Pleosporales contains a known human pathogen but has not been associated with the lungs or asthma.

Fungal and bacterial taxa were also correlated with the concentration of cells in the mucus as well as the percent of eosinophils in both the sputum and the blood. _Haemophilus_, which has been reported to increase inflammation, was positively correlated with the percent of eosinophils in the sputum but not in the blood, nor the overall concentration of cells in the mucus. However, the fungal genus _Candida_ was associated with all three. Pulmonary candidiasis has long been associated with allergic bronchial asthma and inflammation [@masur_pulmonary_1977].

Model for the percent eosinophils

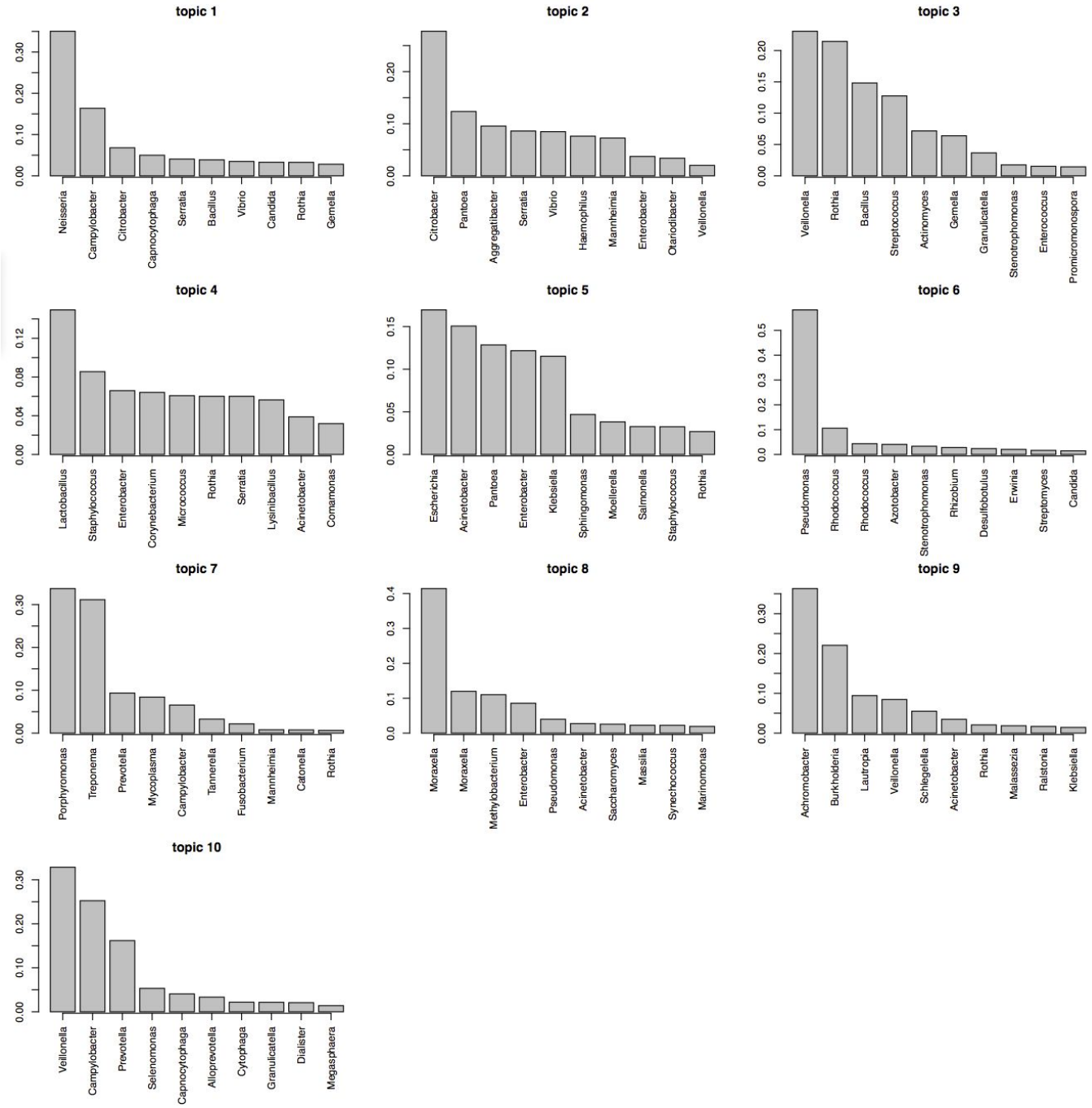
To further explore the association of exogenous microbes with the percent of eosinophils found in the sputum we used a machine learning approach. A random forest model was applied to the 150 genera with the most variance in the dataset. In the context of this large number of genera, _Candida_ is shown to have the greatest influence in the model (Figure 5). The next most influential genera were _Campylobacter_ and another yeast genus, _Saccharomyces_. _Campylobacter_ has been associated with chronic diseases of various types including asthma [@doorduyn_novel_2008], while _Saccharomyces_ has been shown to be protective against the development of asthma-like symptoms in mice [@fonseca_oral_2017]. Moreover, Fonseca _et al._ observed the protective effect of _Saccharomyces_ to be mediated in part through decreased airway eosinophils. In the present study, the contrary is observed, in that each partial dependence plots for each of those taxa has an overall positive slope (Figure 8). However, the particular strain used in the mouse model study, _Saccharomyces cerevisiae_ UFMG A-905, could not be unambiguously identified in this study, in that reads aligned to several _S. cerevisiae_ genomes equally well.

One of the benefits of analyzing the bulk sputum by RNAseq, in addition to being able to survey both the bacteria and fungi, is the ability to simultaneously view the human transcriptome signal. Future work will analyze the human reads to determine if particular pathways are associated with the microbial taxa observed. For example, are the same patterns relating clinical and exogenous sequences observable in the human transcriptome signal, such as in inflammation response pathways? This has the potential to speak directly to the mechanisms by which the microbial taxa are having an effect, and perhaps shed light on the mechanisms and role of microbes in asthma heterogeneity.

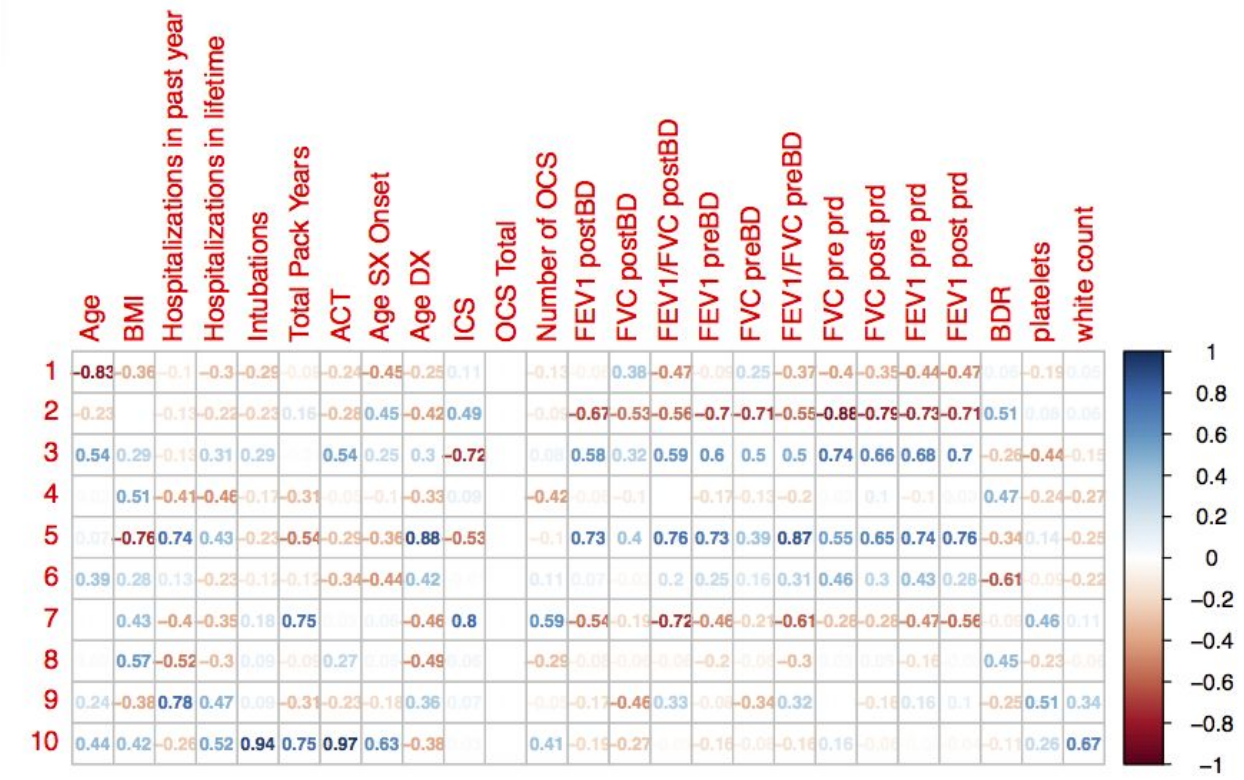
Exogenous/Bulk topics correlation with clinical information

[[SKL2DS: could you help something on the explanation of clinical ?]]

Exo microb dist for each topic:

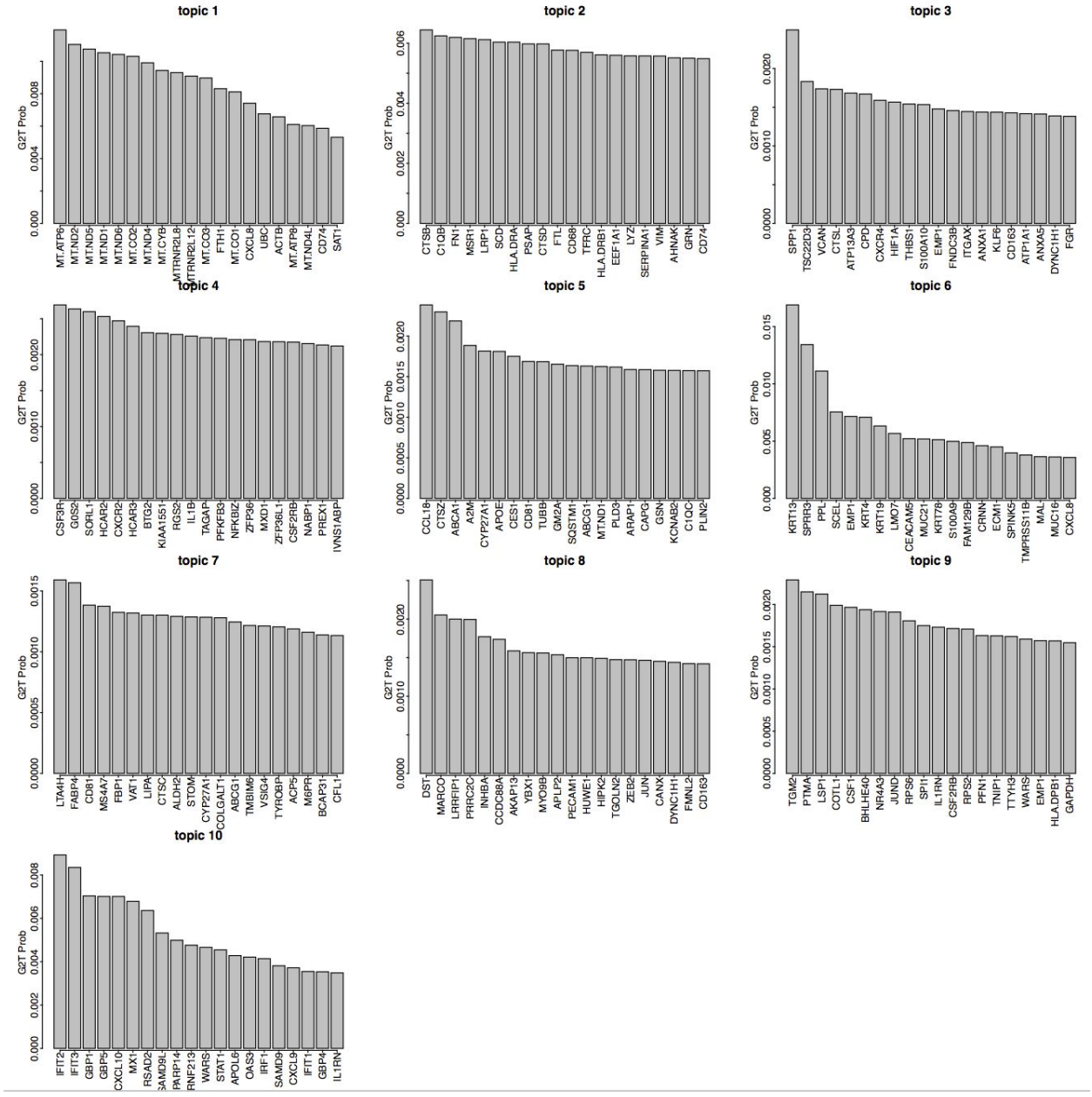


exo_topic2clinical:

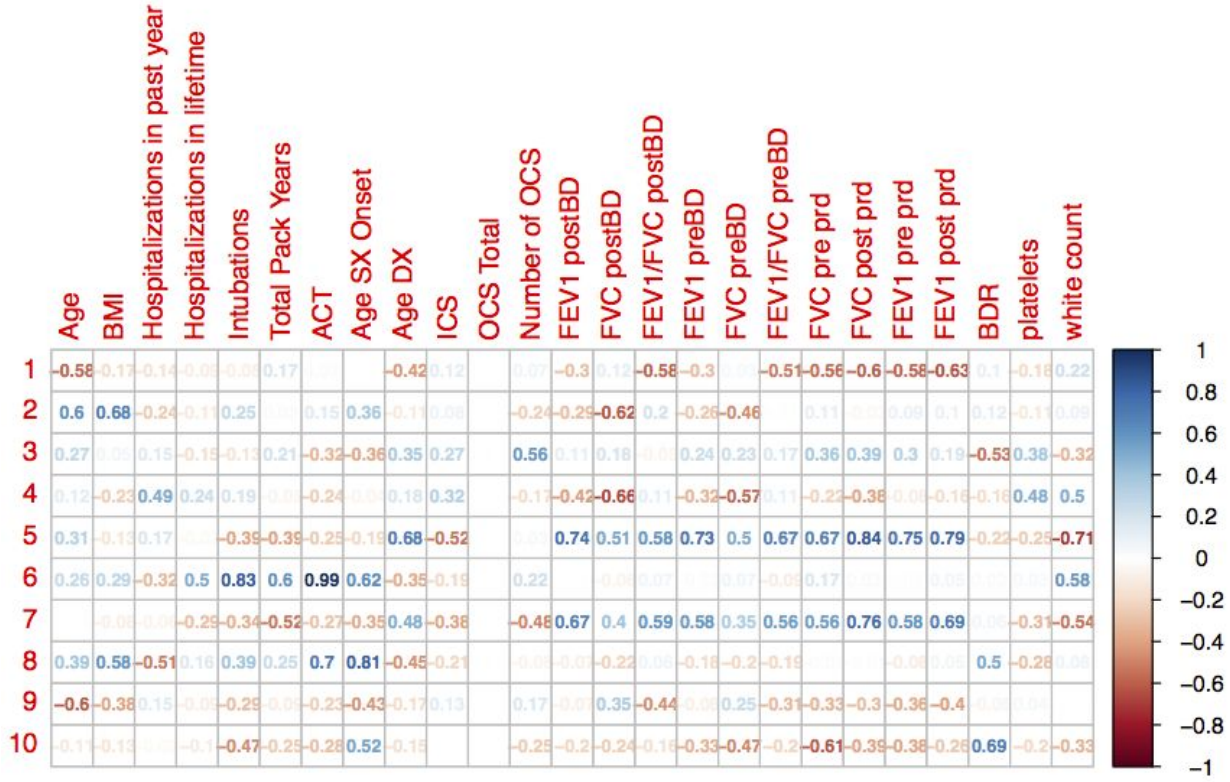


We take 122 samples with clinical information, and correlated the sample to topic distribution with 25 continuous clinical informations, only complete obs pair are considered.

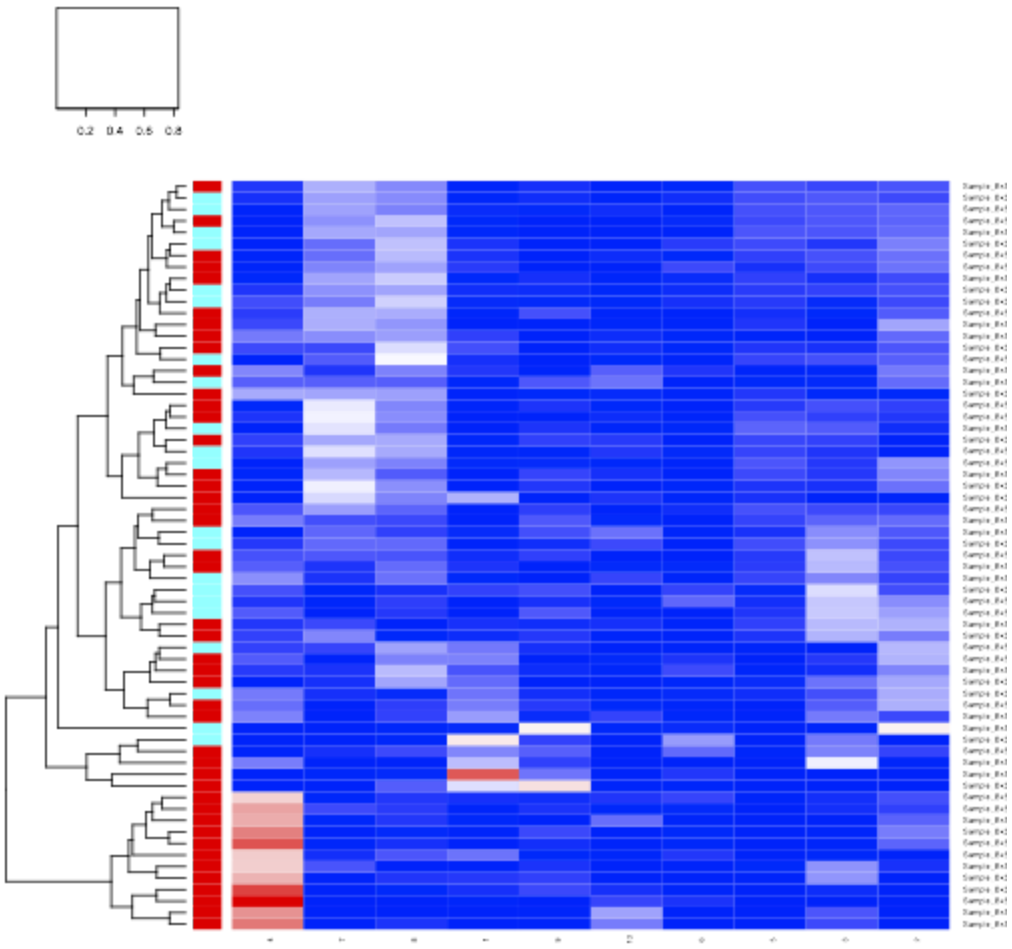
Bulk gene 2 topic distribution:



bulk_topic2clinnical

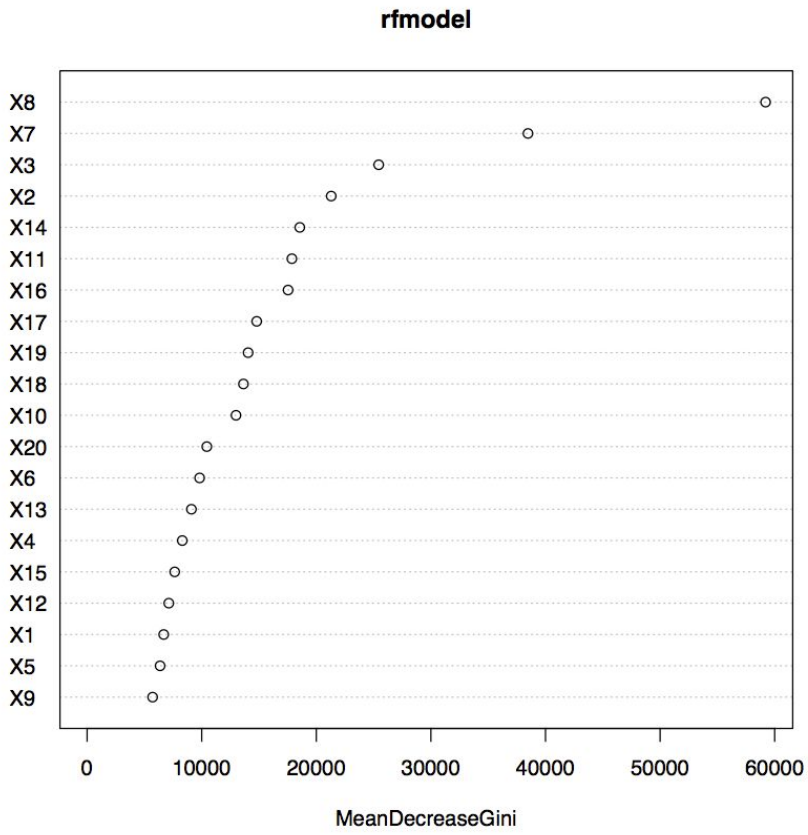


Bulk topics clustering for mild and severe samples



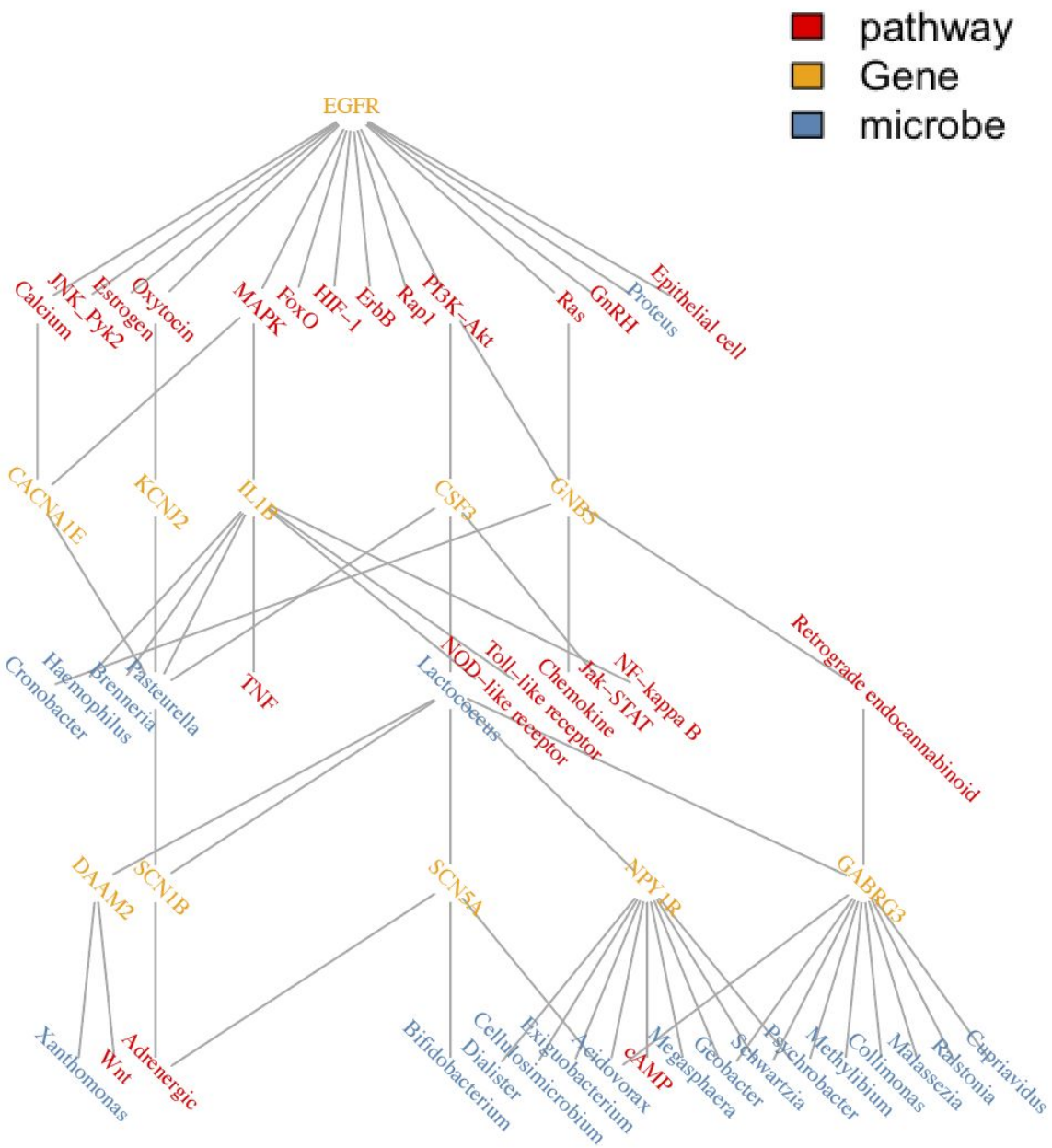
Todo: top genes in topic 4, redraw the heatmap feature

Feature Importance:

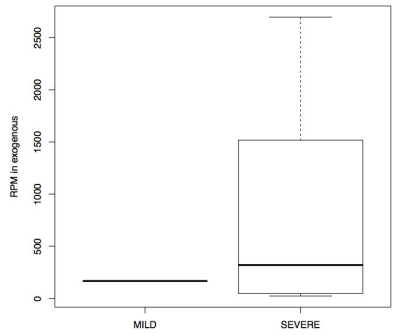


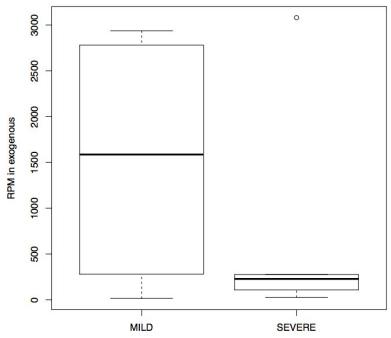
The feature importance in our final pathogen-to-host model, X1-x10 is 10 features from gene, x11-x20 is 10 features from exo topic probability.

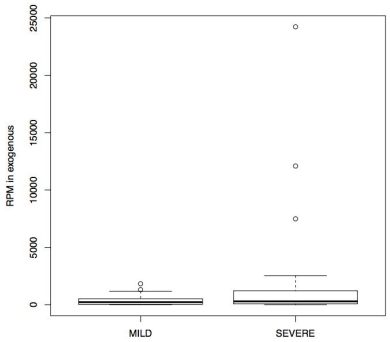
Gene linkage (very interesting, need more biology insight, put more figures)


pathway
Gene
microbe
EGFR
Calcium
JNK_Pyk2
Estrogen
Oxytocin
MAPK
FoxO
HIF−1
ErbB
Rap1
PI3K−Akt
Ras
GnRH
Proteus
Epithelial cell
CACNA1E
KCNJ2
IL1B
CSF3
GNB5
Haemophilus
Cronobacter
Brenneria
Pasteurella
TNF
Lactococcus
Toll−like receptor
NOD−like receeptor
Chemokine
Jak−STAT
NF−kappa B
Retrograde endocannabinoid
DAAM2
SCN1B
SCN5A
NPY1R
GABRG3
Xanthomonas
Wnt
Adrenergic
Bifidobacterium
Dialister
Cellulosimicrobium
Exiguobacterium
Acidovorax
cAMP
Megasphaera
Geobacter
Schwartzia
Psychrobacter
Methylibium
Collimonas
Malassezia
Ralstonia
Cupriavidus

IL1B related microbe rpm in mild and SEVER patients (0 removed)

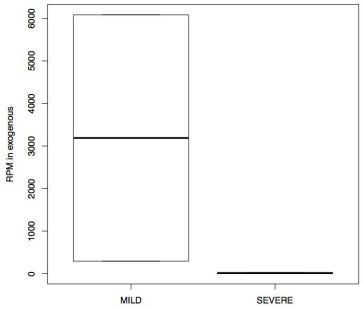


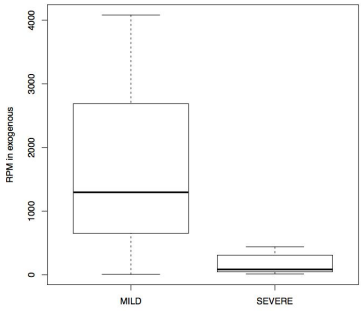


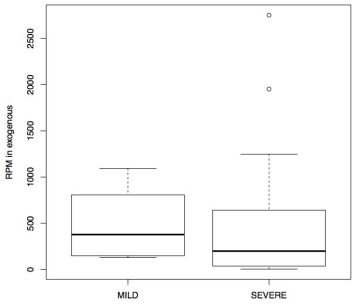


(Pasteurella) (Brenneria), (Haempophilus)

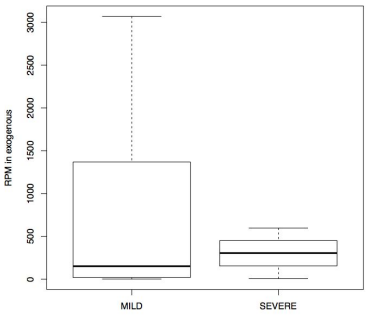
NPY1R (Psychrobacter, Geobacter,Megasphaera:)

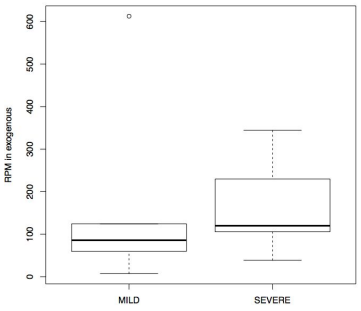


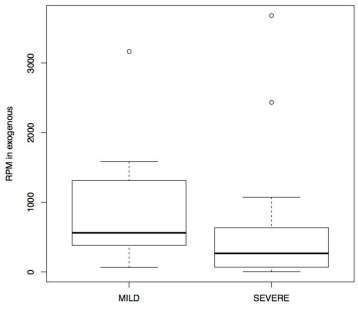




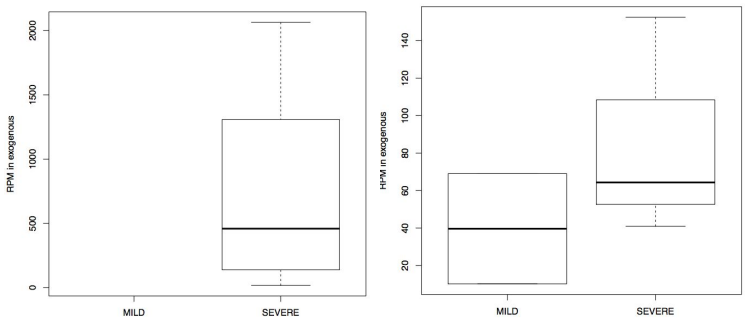
(Acidovorax,Exiguobacterium,Dialister:







(Bifidobacterium,Lactococcus)



The high confident predictions for gene and microbe interactions are defined using 0.9 as cutoff. The signaling pathway related linkage are extracted and then combined with microbe informations, as shown in Fig xxx. IL1 is an important gene that related to Asthma, which is linked to Pasteurella (severe > mild), Brenneria(mild > severe), Haempophilus (sever> mild).

NPY1R \cite{22705097}

DAAM2 \cite{22424883}

WHICH?

Discussion

The role of the airway microbiome in the development of disease is being increasingly appreciated. Commensal microbiota have been shown in other contexts to be strong regulators of host immune system development and homeostasis [@round_gut_2009]. Disturbances in the composition of commensal bacteria can result in imbalanced immune responses and affect an individual's susceptibility to various diseases, including inflammatory (IBD and colon cancer), autoimmune (e.g., celiac disease, arthritis), allergic (e.g., asthma and atopy) and metabolic (e.g., diabetes, obesity, metabolic syndrome) (reviewed in [@shreiner_gut_2015]). Investigation of 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 [@turnbaugh_human_2007]. 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 (CF), Chronic Obstructive Pulmonary Disease (COPD) or Asthma [@erb-downward_analysis_2011; @hilty_disordered_2010; @huang_airway_2015; @morris_comparison_2013]. Despite emerging data on airway microbiota, little is known about the role of the lung microbiome in modulating pulmonary mucosal immune responses. The lung

microbiota in humans has been observed to include on the order of hundreds of bacterial species per person and exhibits exceptional inter-individual diversity that relate to the clincal heterogenity of asthma [@zemanick_airway_2011].

Traditional methods for the analysis of airway microbiota involve the amplification of ribosomal RNA (rRNA) gene fragments and then sequencing the mixture of amplicons, however, recently studies have shown that this signal is confounded by environmental DNA. For example, swabbing ATM buttons in different neighborhoods in New York City demonstrated the ability to distinguish neighborhoods by food preferences, such as chicken and fish [@bik_microbial_2016]. In addition, the primers used to amplify the rRNA fragments has been shown to bias the results, most strongly in that a single kingdom (typically bacteria) is sampled in each experiment. In contrast, RNA is more environmentally labile and therefore more likely to be observed only if isolated from intact, metabolically active cells. In addition, deep sequencing of the total RNA present in a sample, so-called meta-transcriptomics, avoids biases introduced by specific primer amplification and enables discovery of organisms from multiple kingdoms.

Acknowledgements

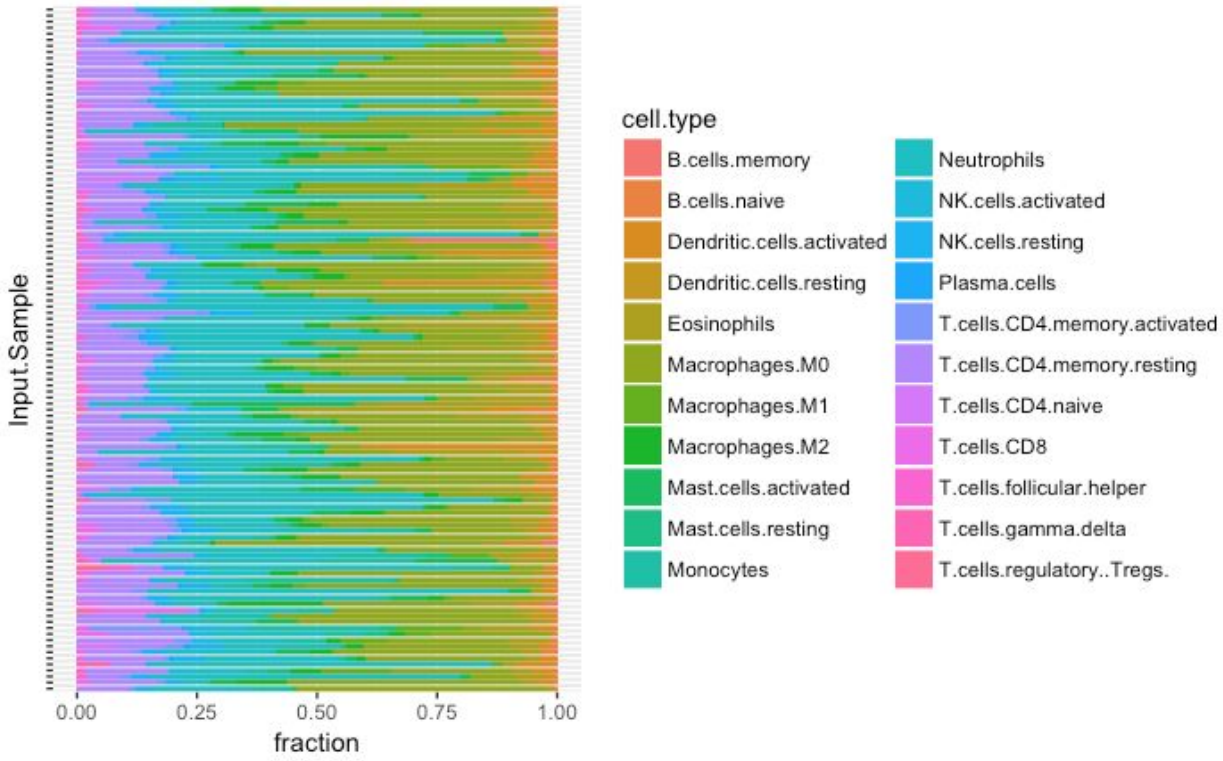
This work was supported by a National Library of Medicine fellowship to DS (5T15LM007056-28) and an NHLBI grant to GC (1R01HL118346-01). The authors would like to thank the support of the Yale High Performance Computing services (Grace, Ruddle, Farnam) and Yale Center for Genome Analysis.

References

Supplemental Information

Figure S1 bulkseq analysis: focus on deconvolution based on cytospin;
Note: first version bulkreq, combat-data based on RSEM dataset; second version firstly vst normalized.
- GSEA
- deconvolution (cytospin)

- deconvolution (scRNAseq)
- deconvolution (cibersort)



- agreement with NMF signatures

Fig1a: summaries of data

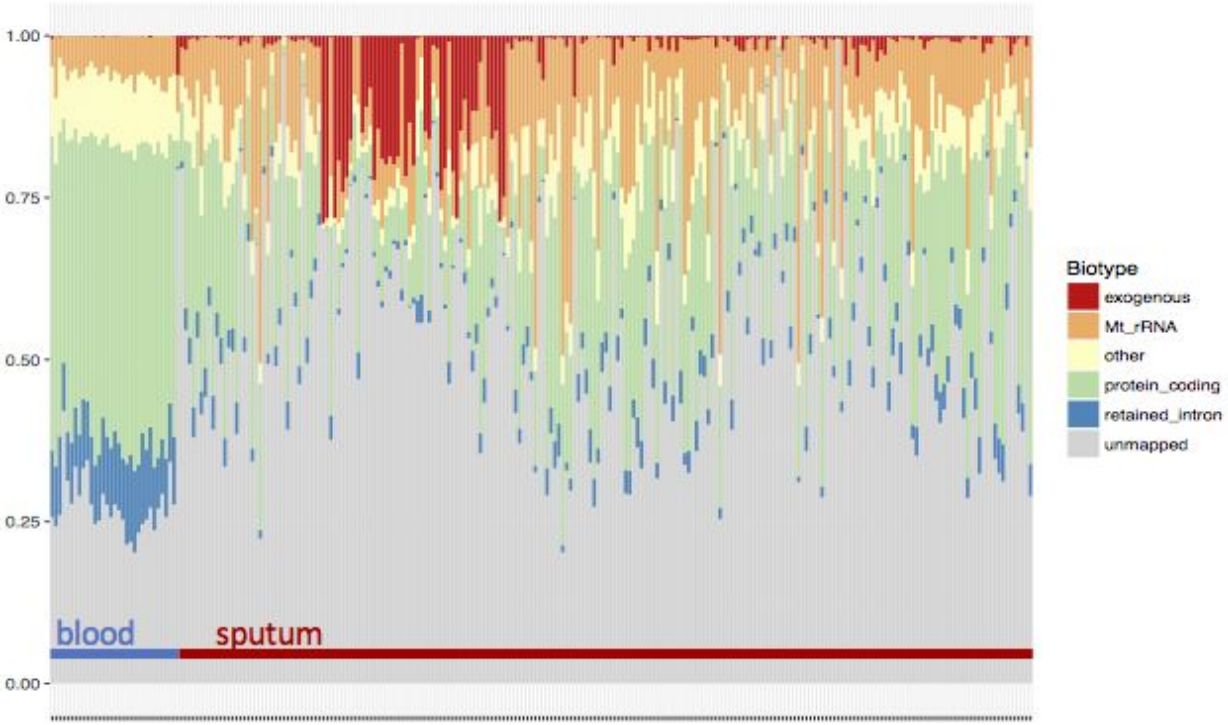


Fig 1b: cytospin data comparison; boxplot for different samples by severity or TEA clusters (supplementary)
Try all 4, maybe mild versus severe, others left in supp

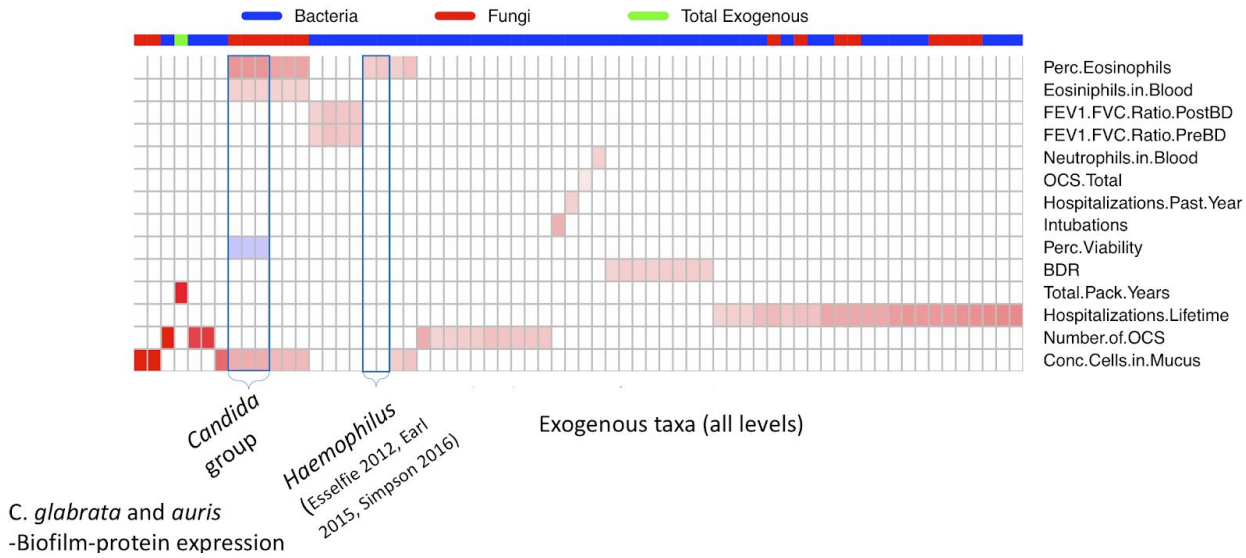
Fig 1c: deconvolution and differential expression Gene with/without deconv
Fig 1d: GSEA analysis using one of the important cell lines
Fig 1e: DEG genes network analysis from different cell lines
(extra: whether we need to associate with other clinical information FFV etc ?)
Question to ask: 1. How cytospin shows different cell contents in sputum? Why deconvolution are usefull?Whether all the cell type are associated with asthma? How to infer the gene set and DEGs genes? How these genes associated from different cell-lines

Figure 2 single cell analysis:

- A summary of single cell data
- B ImmGen distances analysis
- C - difference between # of T-gamma delta cells in asthmatics v controls
- D - cell type signatures

*** Figure 2 exogenous seq summary?

Figure 3 exogenous seq



Question: common reads are put as the parents taxonomy value; unbalanced number of microbial from the same genus; medicines; growth curve for baterial with a break point

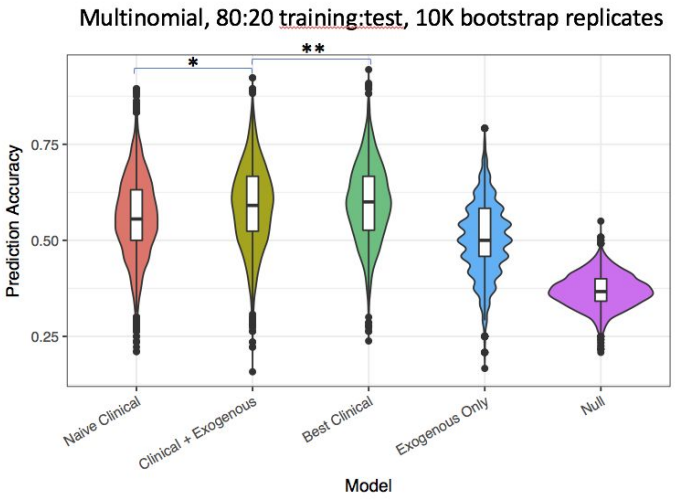
Predicting asthma severity with exogenous sequences

- Naïve clinical model :

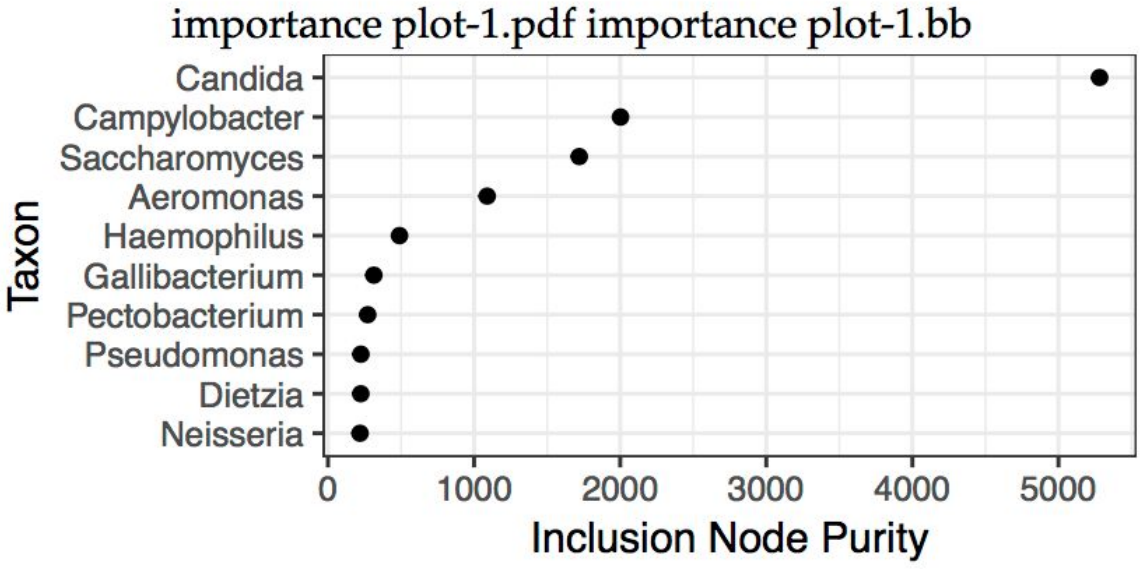
Asthma Severity ~ Age + BMI + Gender +
FeNO+ % Pred FEV1 Post BD

- Best clinical model (step-wise AIC):

Asthma Severity ~ Gender +
% Pred FEV1 Post BD +
Hospitalizations in Lifetime + ACT Score



Comments: use AUC; try not do bootstrap



Comments: forward search from a null dataset to do regression;

Fig 3 : correlations -- cells with clinical, microbes with clinical, cells with microbes??

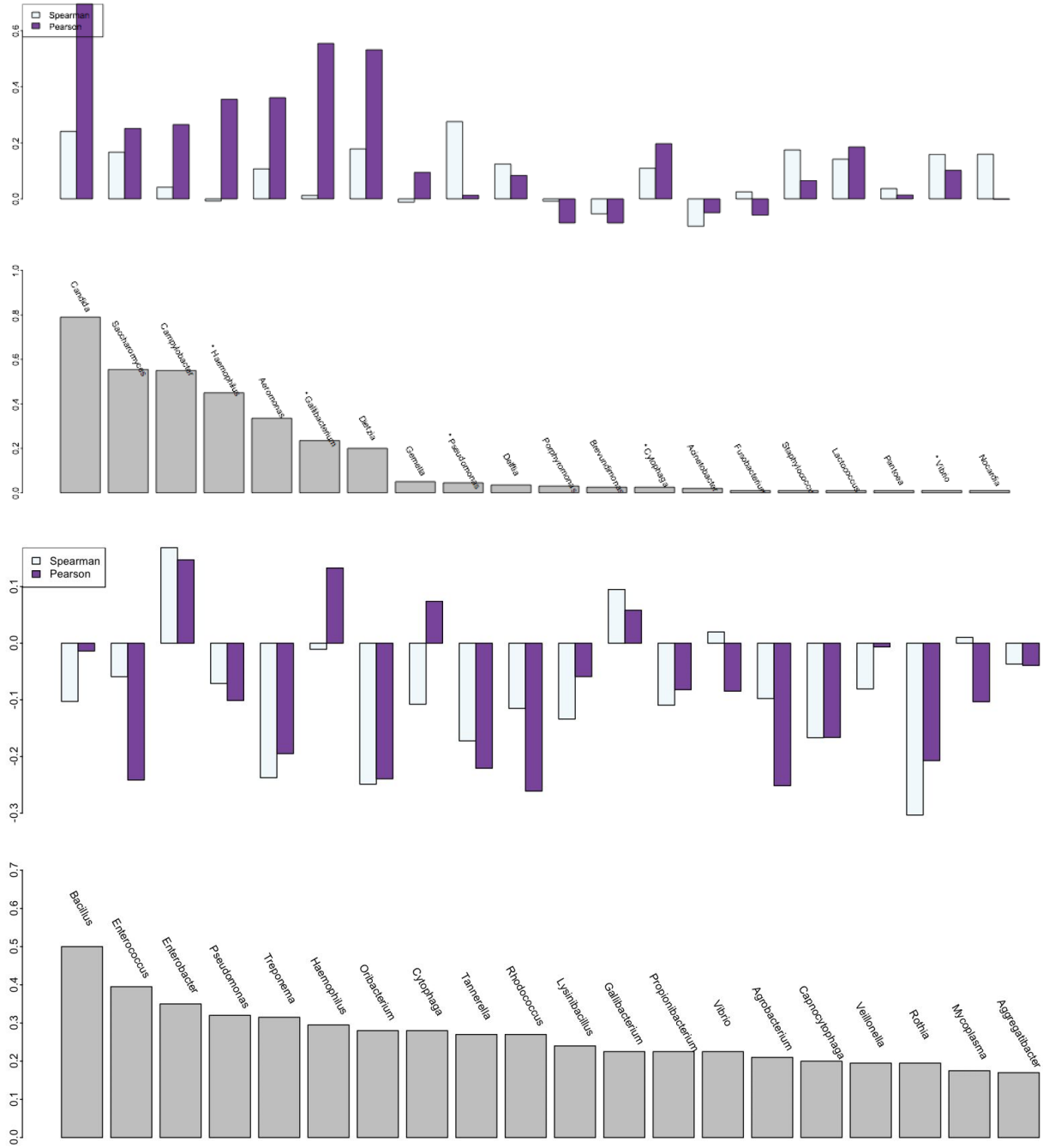


Figure 4 link exogenous to bulkseq

[[SKL: updated the model and trying three different way to train the model: down sampling, unbalanced data, and upscale sampling]]

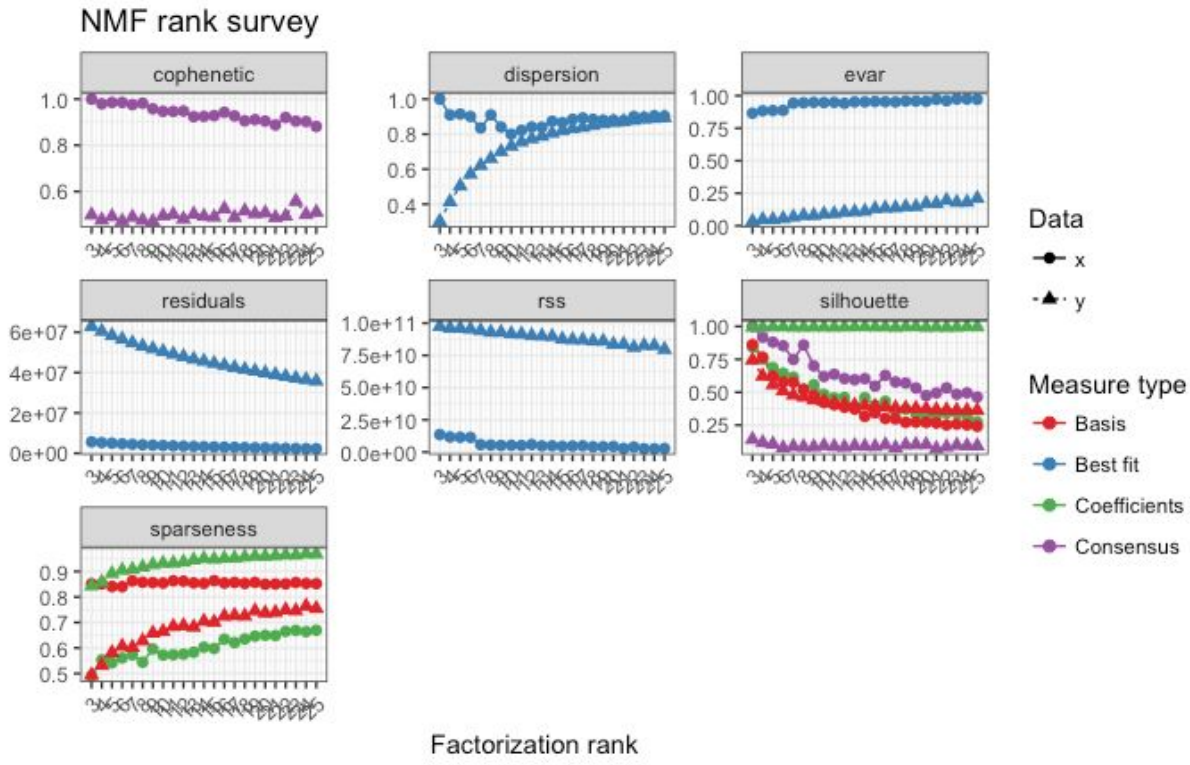
4a: schema or flowchart
4b: correlation with gsea
4c:

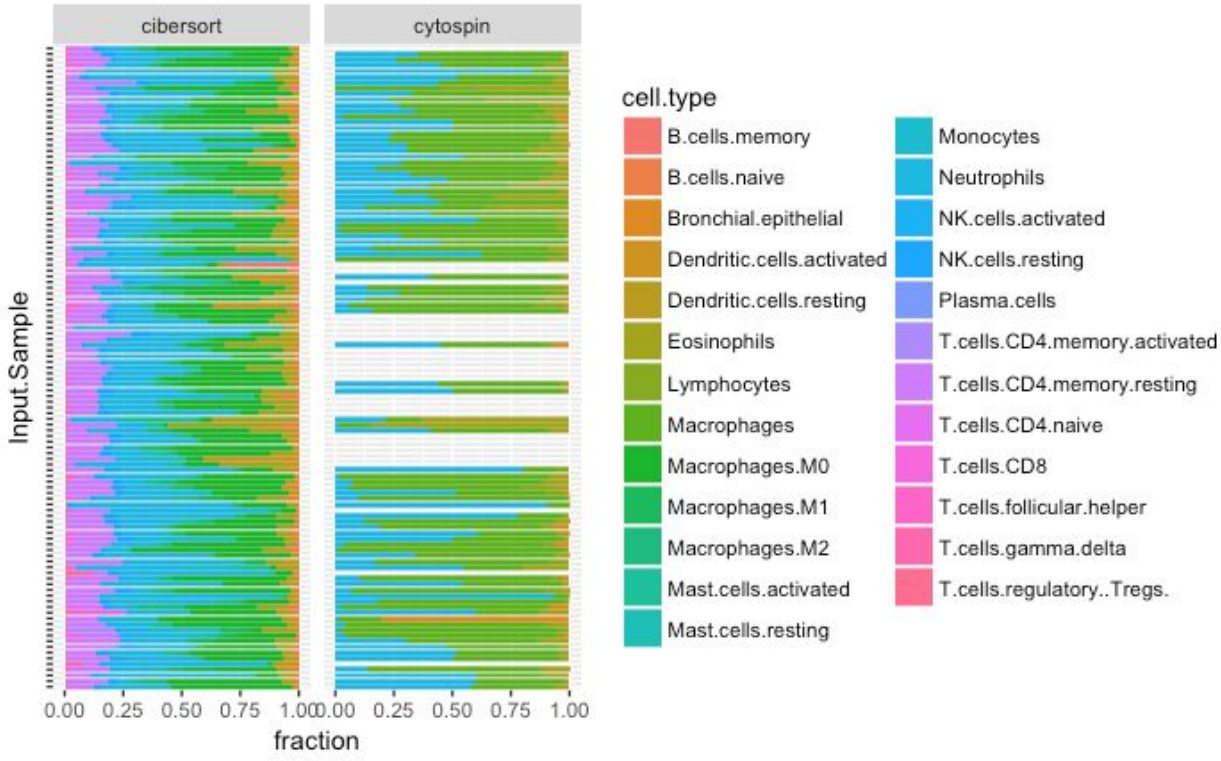
1. Pairwise correlation
2. Supervised learning
3. LDA and linking

Figure 5 downstream (TBD)

Supplemental Figures

Figure SX. Evaluating the optimal nmf rank compared to a randomized set (y).


NMF rank survey
cophenetic
dispersion
evar
residuals
rss
silhouette
sparseness
Factorization rank
Data
x
y
Measure type
Basis
Best fit
Coefficients
Consensus


cibersort
cytospin
Input.Sample
0.00
0.25
0.50
0.75
1.00
fraction
cell.type
B.cells.memory
B.cells.naive
Bronchial.epithelial
Dendritic.cells.activated
Dendritic.cells.resting
Eosinophils
Lymphocytes
Macrophages
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.

Todo:
~~1. DS to check exceRpt output -- is the Count data normalized? Also combat~~
~~Use combat test to check whether need batch effect. --~~
~~2. SKL batch effect check (paper)~~
~~3. Meeting at Friday 915am~~

1. Check rsem*.txt for single cell is rpkm or rpm (FPKM)
2. Week of after jan 1 for next meet Jan5 2pm chat

1a) DS try normalizing the data summary figure to remove unmapped
- change blood to red, sputum to blue
- split human gencode mapped reads into cell fractions

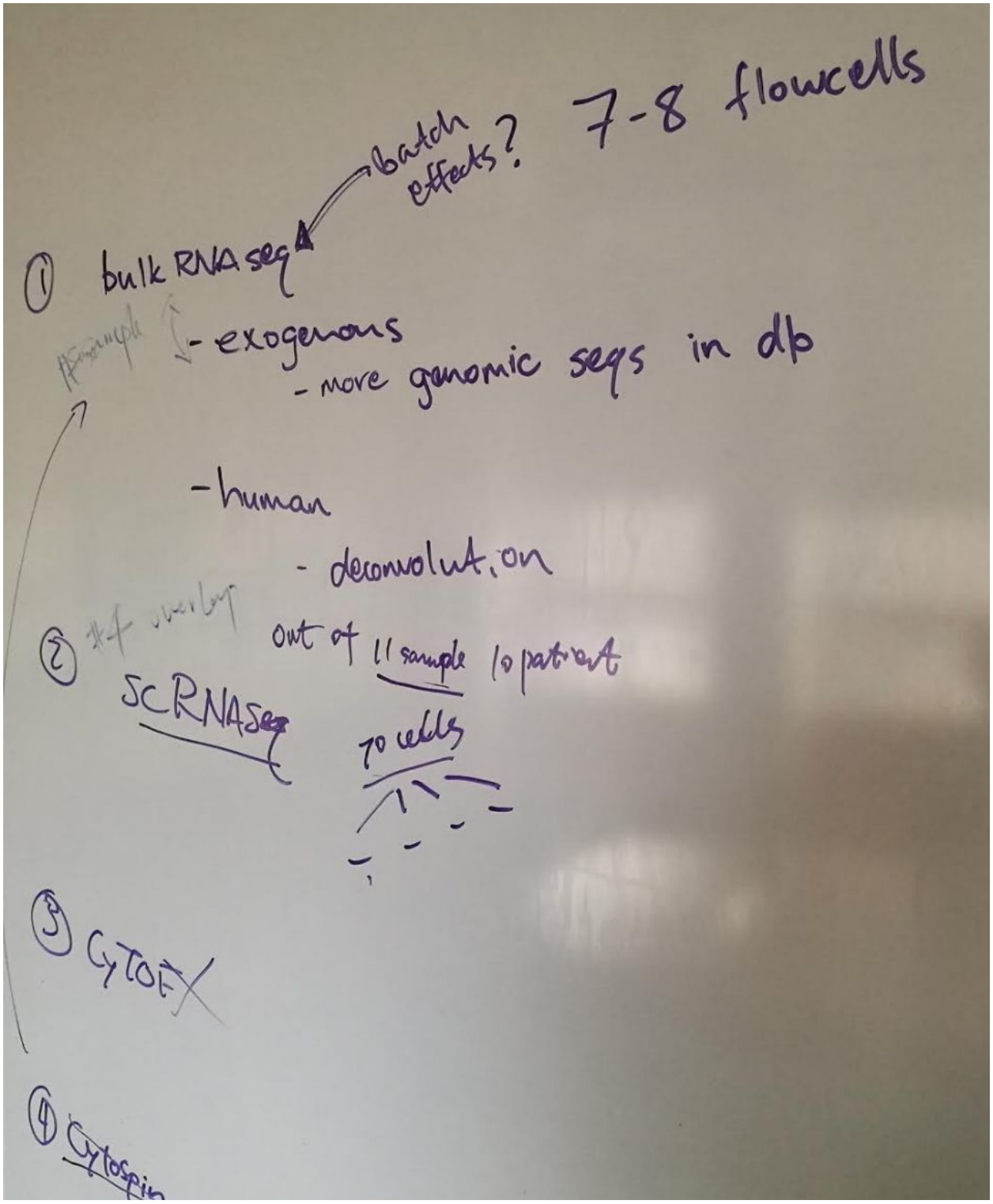
1b) DS - NEW How variable are the cell percentages for the human reads
All severe vs control+MILD deseq after accounting for different cell populations
- Try all splits (mild vs severe, fev1 split, etc)

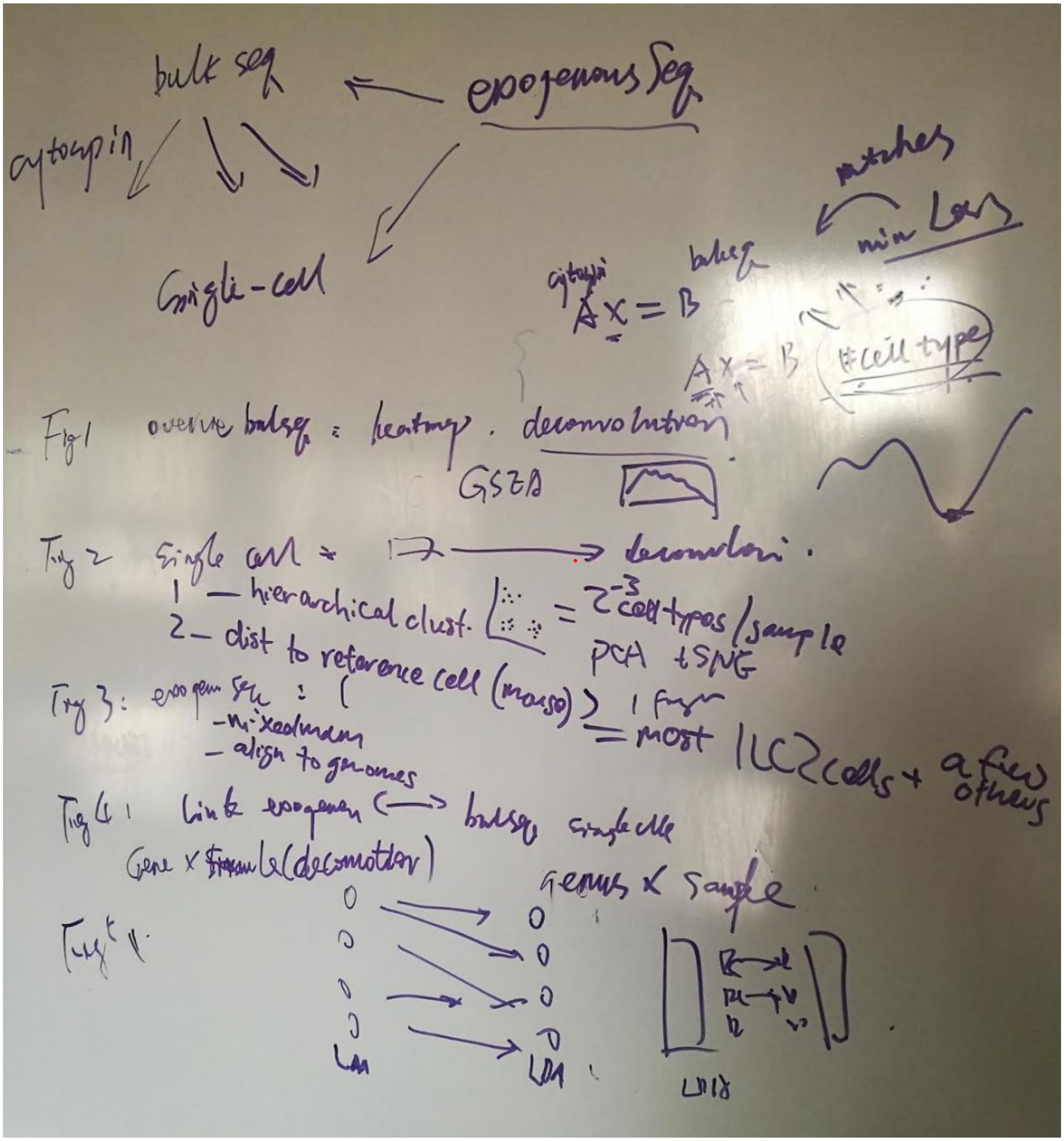
1c) DS - DESeq w+w/o controlling for cell populations (plotMA) -e
1d) SKL - LDA on individual cell types, bulk and plot (f-?)

2a) single cell t-sne
DS - Seurat
SKL: single to deconvoluted and single cell count, rpkm fpkm tsne

3)

4)


batch effects?
7-8 flowcells
① bulk RNA seq
- exogenous
- more genomic seqs in db
- human
- deconvolution
② 1k/sample
out of 11 sample 10 patient
SCRNASeq
10 cells
③ CyTOF
④ Cytospin


bulk seq
exogenous seq
cytospin
single-cell
matches
min Loss
AX = B
# cell type
Fig 1 overview bulkseq : heatmap . deconvolution
GSEA
Fig 2 single cell
1 — hierarchical clust.
2 — dist to reference cell (mouse)
2-3 cell types / sample
PCA tSNE
Fig 3: exogenous seq
- mixed model
- align to genomes
= most ILC2 cells + a few others
Fig 4 link exogenous <—> bulkseq single cell
genes x sample
LDA
LDA