**Transcriptomics of Bronchoalveolar Lavage Cells Identifies New Molecular Endotypes of Sarcoidosis - Online Supplementary Materials**

Milica Vukmirovic\*1,11, Xiting Yan\*1,2, Kevin F. Gibson3, Mridu Gulati1, Jonas C. Schupp1, Giuseppe DeIuliis1, Taylor S. Adams1, Buqu Hu1, Antun Mihaljinec1, Tony Woolard1, Heather Lynn1,8, Nkiruka Emeagwali1, Scott M. O'Neal3, Erica L. Herzog1, Edward S. Chen4, Alison Morris3, Joseph K. Leader3, Yingze Zhang3, Joe G. N. Garcia8, Lisa A. Maier5, Ron Colman9, Wonder P. Drake6, Michael Becich3, Harrison Hochheiser3, Steven R. Wisniewski3, Panayiotis V. Benos3, David R. Moller4, Antje Prasse10, Laura L. Koth7, Naftali Kaminski1

1 Section of Pulmonary, Critical Care and Sleep Medicine, Department of Internal Medicine, Yale University School of Medicine, New Haven, CT/US

2 Department of Biostatistics, Yale School of Public Health, New Haven, CT/US

3 University of Pittsburgh, School of Medicine, Pittsburgh, PA/US

4 Johns Hopkins University, Baltimore, MD/US

5 National Jewish Health - Denver, CO/US,

6 Vanderbilt University, Nashville, TN/US

7 University of California San Francisco, San Francisco, CA/US

8 University of Arizona Health Sciences, Tucson, AZ/US

9 University of Pennsylvania School of Medicine, PA/US

10 Hannover Medical School, Hannover (MHH), Germany

11 McMaster University, Hamilton, Canada

On behalf of the GRADS Investigators

\* Equally contributing authors

Address correspondence: [naftali.kaminski@yale.edu](mailto:naftali.kaminski@yale.edu)

300 Cedar Street, PO Box 208057, New Haven, CT 06520-8057

Phone (203) 737-4612 / Fax (203) 785-6094

**Sources of support**

This work is supported by NIH grants: U01 HL112707, U01 HL112694, U01 HL112695, U01 HL112696, U01 HL112702, U01 HL112708, U01 HL112711, U01 HL112712, UL1 RR029882, UL1 RR025780, R01 HL110883, R01 HL114587, CTSI U54 grant 9 UL1 TR000005, CDC NMVB 5U24 OH009077, CTSA UL1 TR002535, R21 LM012884

Table of Contents

[Supplementary Website 2](#_Toc43137684)

[Sample preparation and RNA sequencing 2](#_Toc43137685)

[Sequencing Data Preprocessing 4](#_Toc43137686)

[***Data Quality Assessment*** 4](#_Toc43137687)

[***Mapping and FPKM Calculation*** 4](#_Toc43137688)

[***Data Cleaning and Batch Effect Assessment*** 4](#_Toc43137689)

[Supervised Analysis 5](#_Toc43137690)

[Unsupervised Analysis 6](#_Toc43137691)

[***Cluster Analysis using Chosen Modules*** 9](#_Toc43137692)

[Validation Analysis 10](#_Toc43137693)

[***Freiburg Cohort*** 10](#_Toc43137694)

[***Validating the WGCNA results*** 10](#_Toc43137695)

[***Validating the supervised analysis results for PFTs*** 10](#_Toc43137696)

# **Supplementary Website**

We have generated a supplementary website (<https://yale-p2med.github.io/SARC_BAL/>) for this article from which data, analytical codes, paper supplement, results of supervised analysis, results of unsupervised analysis can be downloaded.

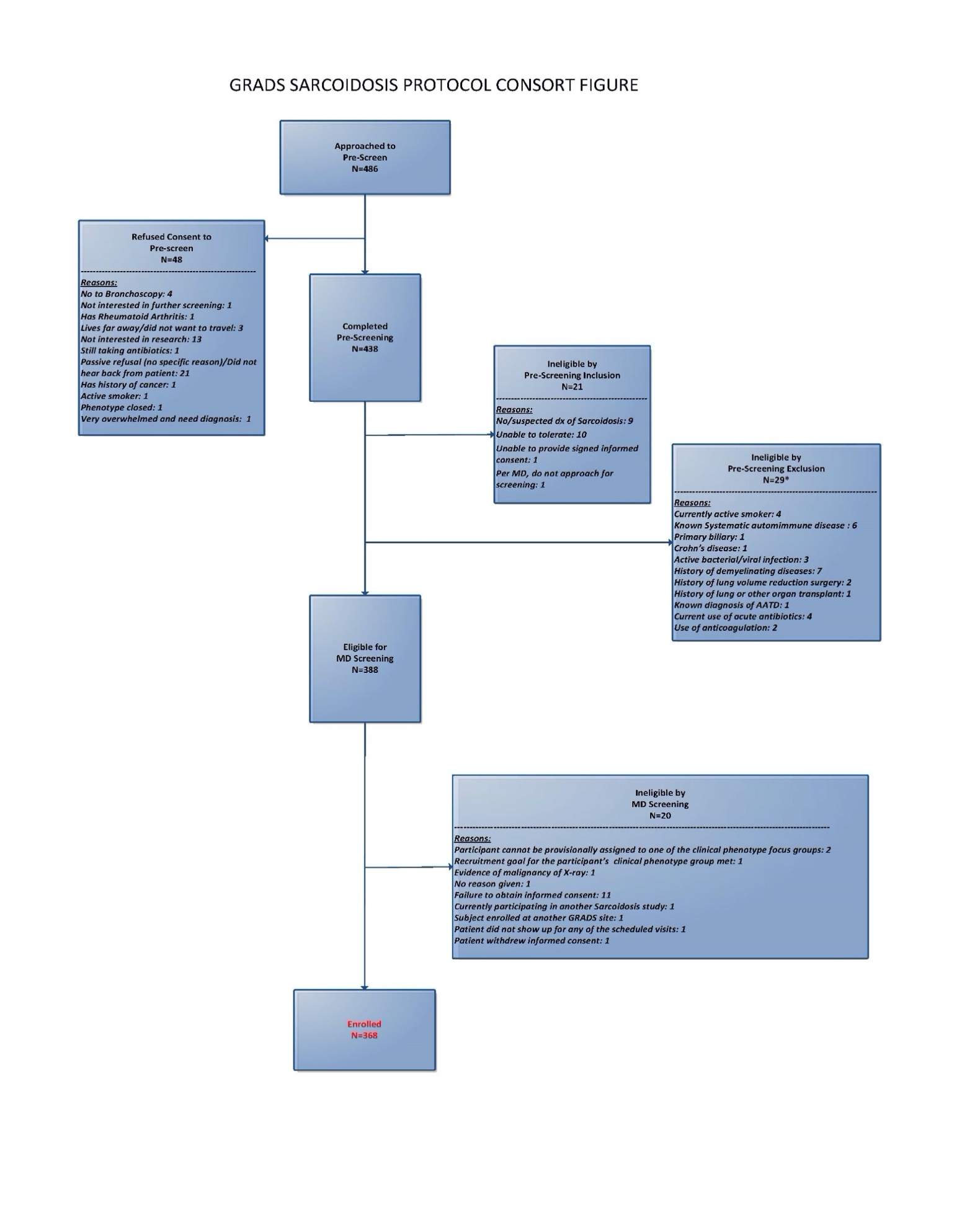
# **Sample preparation and RNA sequencing**

All patients involved in this study have signed a consent to participate in this study in accordance with institutional IRB protocols (**Figure S*1***). Total RNA was extracted from BAL samples using Qiazol following Qiagen’s miRNeasy protocol (Qiagen 217004) and using QiaCube. RNA quantity and quality was assessed using NanoDrop (Thermo Scientific) and TapeStation 2200 (Agilent). RNA Integrity Number (RIN) over 6.5 and yield over 1ug of total RNA were criteria for acceptable quality to be submitted for sequencing (**Table S*1*** and Supplement Table E1). cDNA libraries were made from 1ug of total RNA upon Poly-A selection using Dynabeads® mRNA DIRECT™ Micro Purification Kit (Ambion 61021) and fragmentation using the AB Library Builder™ System (Life Technologies 4463592) with the Ion Total RNA-Seq Kit for AB Library Builder™ System (Life Technologies 4482416). The cDNA was amplified and barcoded using the Ion Xpress™ RNA-Seq Barcode 1-16 Kit (Life Technologies 4475485). cDNA was loaded onto Ion PI™ Chip Kit v2 BC (Life Technologies 4484270) using the Ion Chef™ System (Life Technologies 4484177) with the Ion PI™ IC 200 Kit (Life Technologies 4488377). Sequencing was performed using Ion Proton™ System for Next-Generation Sequencing (Life Technologies 4476610) using the Ion PI™ IC 200 Kit (Life Technologies 4488377) to obtain RNA-Seq depth of ~ 30 million single-end reads/sample with an average read length of 150bps. Successfully sequenced samples were samples whose cDNA libraries passed quality control and had depth of sequencing of ~ 30 million single-end reads/sample.

**Table S1**: Sample filtering using RIN and RNA quality metrics.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STAGE (PHENOTYPE) | # SUBJECTS | # BAL SAMPLES with RNAs | # BAL SAMPLES PASSING QC | # SUCCESSFULLY SEQUENCED |
| TOTAL, n | 318 | 261 | 219 | 215 |
| Non-acute, Stage I, untreated | 36 | 34 | 26 | 26 |
| Acute Sarcoidosis, untreated | 16 | 16 | 15 | 14 |
| Remitting, untreated | 54 | 48 | 44 | 42 |
| Stage II-III, untreated | 50 | 48 | 40 | 42 |
| Stage II-III, treated | 49 | 45 | 36 | 36 |
| Stage IV, untreated | 32 | 18 | 13 | 13 |
| Stage IV, treated | 46 | 23 | 19 | 19 |
| Multi-organ | 35 | 29 | 26 | 24 |

**Figure S1**: Consort figure.



# **Sequencing Data Preprocessing**

## ***Data Quality Assessment***

The Torrent Suite™ Software (V5.0.5) was used to generate the raw sequencing bam file without alignment. These bam files were further converted into fastq files using the bam2fastx component from tophat2 (V2.0.12). The pre-alignment metrics provided in the Torrent Suite™ Software run reports, including bead loading, Ion Sphere™ Particle (ISP) density, total number of reads, filtering numbers, and mean read length. We used these quality thresholds provided by the company to filter out low quality sequencing runs. The samples in these low-quality sequencing runs were sequenced again.

The raw fastq files were assessed for sequencing reads quality using FastQC(1) to identify possible sequencing adapter or polymer contamination. The distribution of the base quality score along the read positions was also considered to control the quality. For this data set, all samples that pass the sequencing run filtering based on the sequencing run report passed the FastQC quality control.

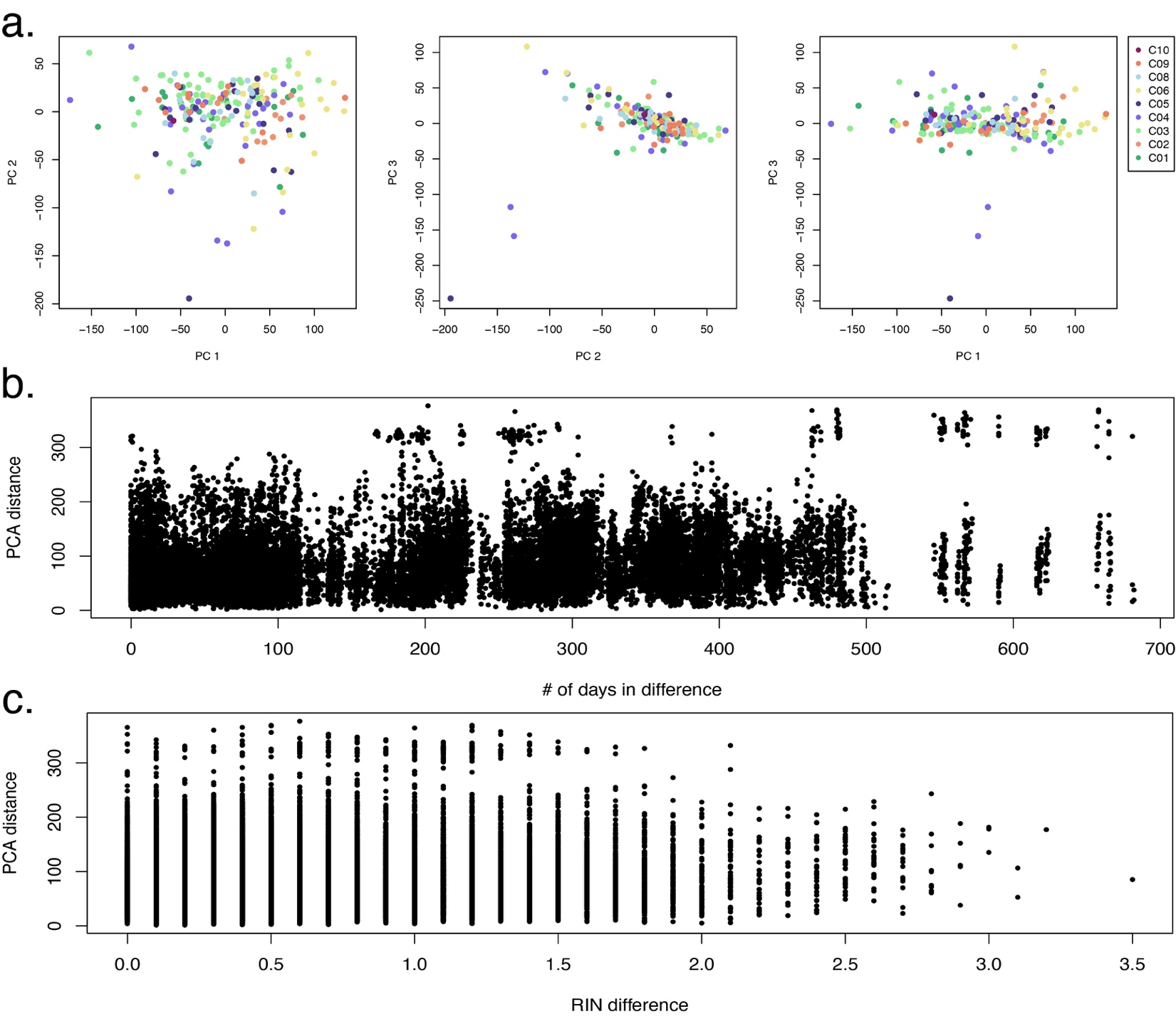
## ***Mapping and FPKM Calculation***

The sequencing reads in the fastq files were mapped onto the human genome (UCSC hg38) using a two-stage mapping strategy suggested by the manufacturer. In the first stage, all raw reads were mapped to hg38 using STAR (2) with gene annotation and the --b2-very-sensitive option. The unmapped reads from the first stage were further mapped to hg38 using bowtie2 (3) with local alignment and the –very-sensitive-local option. Cufflinks (4) was used to calculate the Fragments Per Kilobase of transcript per Million mapped reads (FPKMs) as the estimated gene expression levels.

## ***Data Cleaning and Batch Effect Assessment***

The principal component analysis (PCA) was applied to identify potential outlying sequencing reactions. There were 240 sequencing reactions for 215 samples, among which 15 reactions were identified as outliers by PCA and thus removed from further analysis. Among these 15 reactions, 11 of them also had low mapping rate, low numbers of expressed genes, and low numbers of mapped reads. For the other 4 reactions, 3 of them were shown to cluster with the PBMC samples instead of the other BAL samples, indicating that they were actually PBMC samples mislabeled as BAL samples. After the outlier removal, we had 225 high quality sequencing reactions in total, which included 31 repeated reactions from 15 BAL samples. Among the repeated reactions, we kept the one with comparable number of mapped reads to the other samples (~30 million single-end reads/sample). If multiple replicates qualified, we kept the reaction with the highest mapping rate. After this cleaning, we kept 209 sequencing reactions for 209 unique BAL samples.

In addition to data cleaning, PCA was also used to examine the data for possible batch effect due to multiple technical factors including sequencing date, sample collection centers, and the RNA integrity number (RIN). The batch effect assessment was done mainly using two ways: data visualization and the sample-sample PCA distance. To examine the effect of the sample collection center, we visualized the data using the PCA projection plot and labeled the samples based on their sample collection centers (**Figure S*2*a**). For the sequencing date and the RNA integrity number (RIN), we calculate the Euclidean distance between any two samples using the top 3 PCs and plotted this distance against the number of days in differences in their sequencing date and the difference in their RINs, respectively (**Figure S*2*b** **c**). None of these visualizations showed significant effect of these three technical factors so we proceeded to downstream analysis without any data adjustment for these technical factors.



**Figure S2**: Technical effect examination of a). sample collection center, b). sequencing date and c). RNA integrity number.

# **Supervised Analysis**

The supervised analysis identified gene signatures associated with clinical traits using non-parametric test. The Wilcoxon Rank Sum test and Kruskal-Wallis test were used for categorical clinical traits with two categories and more than two categories, respectively. The Spearman’s Rho test was used for continuous clinical traits. The false discovery rate (FDR) was calculated to control for multiple testing error. Genes with an FDR<0.05 were defined to be the significant associated genes. When no genes achieve this global significance, genes with a fold change (FC) >2 and a p value<0.05 were considered as significant.

Scadding staging, PFTs% predicted, age, CT scan features with severity measurement and BAL cell differentials were considered as continuous. Race, gender, sex and CT scan features without severity measurement were considered as categorical. For disease severity, Scadding stage II, III and IV were compared to Scadding stage I separately. Similarly, all the 8 clinically defined phenotype groups were also compared to the non-acute stage I group separately. For PFTs% predicted (FEV1% predicted, DLCO% predicted, FVC% predicted), samples with PFT% predicted higher than 80% were compared to those from 50% to 80%. In addition, patients with obstructive lung disease (FEV1/FVC ratio < 70%) were compared to those with restrictive lung disease (FEV1/FVC ratio >70% and FVC% predicted<80%). These separate comparisons were conducted using Wilcoxon Rank Sum test. The detailed results and the actual gene lists can be found on our supplementary website (<https://yale-p2med.github.io/SARC_BAL>). The summary of the globally significant genes is presented in the Supplementary Table E2. This analysis is not adjusted for cell differentials.

The total number of significant genes associated with each clinical trait as well as the overlap between each two clinical traits is shown in Figure 2a. For the clinical traits included in the analysis, the percentage of missing values was very low (<3%). Entries on diagonal show the total number of genes significantly associated with each clinical trait and the numbers of positively (followed by +) and negative (followed by -) correlated genes for the same clinical trait. Off the diagonal, each entry describes the total number of genes significantly associated with both given clinical traits and the number of genes with the described correlation directions for trait in the row and column in the parentheses, respectively. The GeneGo Metacore (Thomson Reuters) was applied to the lists of significant genes to identify significant (FDR<0.05) enriched pathways (Figure 2b). The detailed results and the actual gene lists can be found on our supplementary website (<https://yale-p2med.github.io/SARC_BAL>). In Figure 2b, genes significantly associated with each clinical trait are represented by bars on the left with the length of each bar proportionate to the number of genes. These genes were further divided into positively and negatively correlated genes represented by bars in the middle with purple bars for negative correlation and yellow bars for positive correlation. The lengths of these bars are also proportionate to the number of corresponding genes. Each set of negatively or positively correlated genes was further connected to pathways (represented by bars on the right) that were significantly (FDR<0.05) enriched for genes in the given set. Only the top 5 significant (FDR<0.05) pathways with at least 3 overlapping genes are shown.

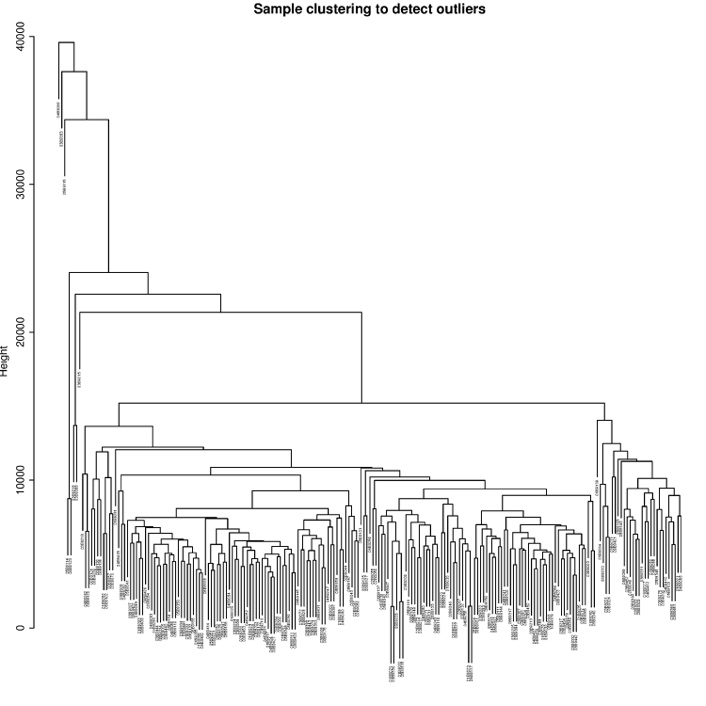
# **Unsupervised Analysis**

The unsupervised analysis of the data consists of two parts. In the first part, we applied the WGCNA(5) to identify gene modules and assess their correlation with the following clinical traits: demographics, PFTs, CT scan variables, phenotypes, treatment and BAL cell differentials. In the second part, we chose 5 gene modules that had significant correlation (p value<0.05) with highest number of clinical traits. Genes from each module were used to cluster the patients into subgroups using K-means clustering (Figure 4). Among the identified clusters, the two extreme clusters with the largest differences in their gene expression profiles shown in Figure 4 were compared for all patient characteristics collected under GRADS study protocol for a better understanding of the clinical relevance for these modules (Supplement Table E3). Chi-square test and Wilcoxon rank sum test were used to assess the significance for categorical and continuous patient characteristics, respectively, amongst the clusters for chosen gene modules. The MetaCoreTM of GeneGO, Inc. was applied to identify significant enriched pathways for each gene module identified by the unsupervised analysis.

***WGCNA Analysis***

Identifying outliers

We applied the weighted gene co-expression network analysis (WGCNA) to the 209 BAL samples using the WGCNA R package (5). Genes expressed (FPKM>0.01) in less than 10% of the 209 samples were also removed before the WGCNA analysis, which kept 22,307 genes for WGCNA analysis. The threshold chosen for the gene expression represents the minimum FPKM level that could be robustly detected by the sequencing protocol in this dataset. Since WGCNA results can be sensitive to outliers, we did hierarchical clustering of all the 209 samples using the 22,307 genes to identify possible outliers for the WGCNA analysis, which showed that there are potentially 6 branches in the tree, including 8 samples, that could have big impact on the WGCNA results (**Figure S*3***). To decide exactly which samples to exclude, we applied WGCNA with trimming of 0, 1, 2, 3,…, 6 branches from the top of the clustering tree and compared their clustering results (**Figure S*4***). The comparison showed that all 6 branches have heavy impact on the WGCNA results and thus we excluded all 8 samples from further WGCNA analysis.



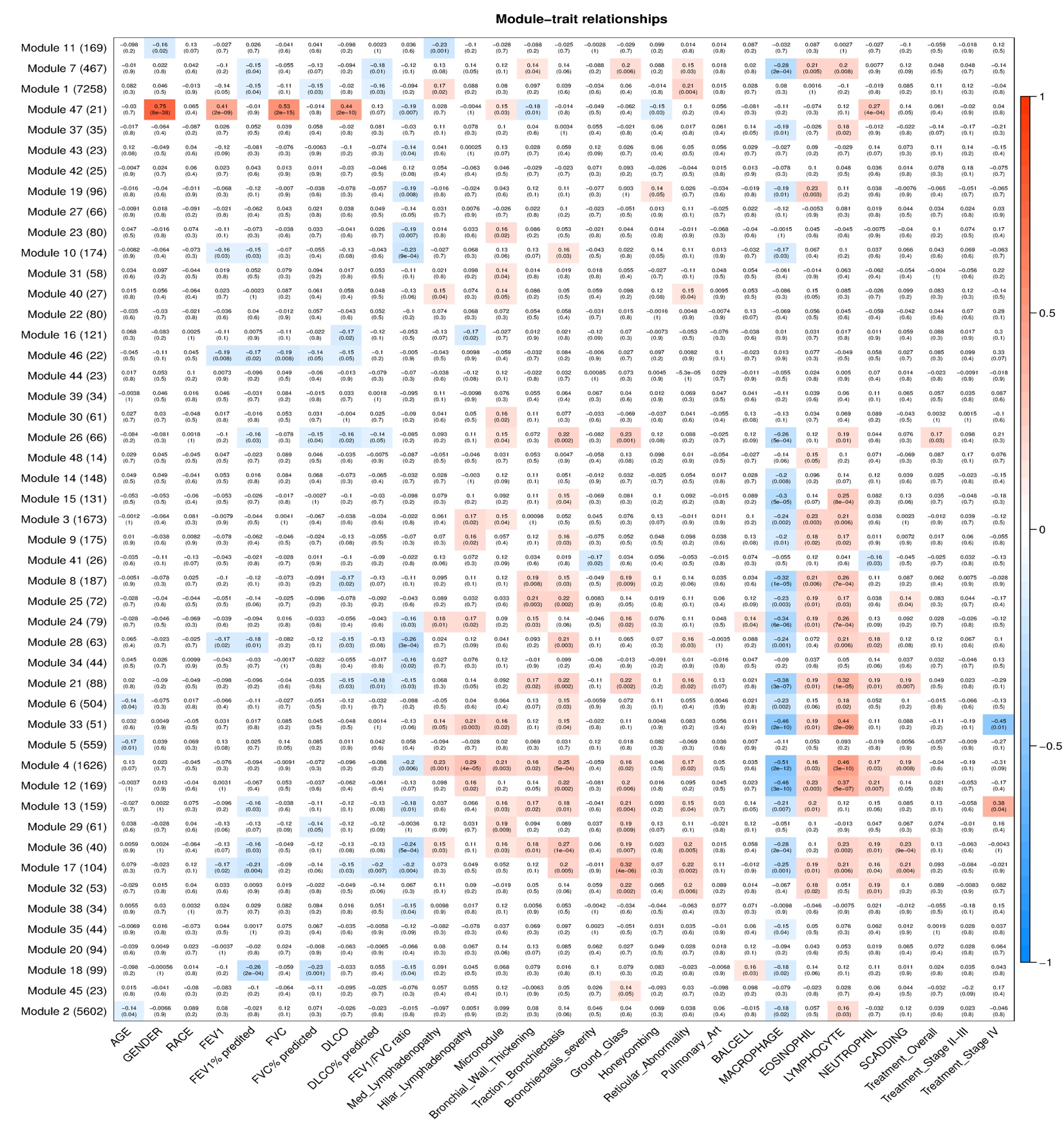
**Figure S3**: Hierarchical clustering tree of all 209 samples to identify potential outliers for the WGCNA analysis.

***Correlating gene modules with clinical traits***



**Figure S4**: Comparison of the clustering results by trimming 0, 1, 2, …, 6 branches from the top of the clustering tree in Figure S3. The dendrogram on top shows the hierarchical clustering tree of all genes by WGCNA analysis. The 7 color bars on bottom show the clustering results of all these genes after trimming off a given number of outliers (left of the color bars) in the hierarchical clustering tree of samples in Figure S3. Within each color bar, each color represents one identified gene module. The comparison between different color bars showed that when we trimmed >=8 outliers (the leftmost 8 samples in the hierarchical clustering tree in Figure S3), the clustering results became stable, justifying the need to remove 8 outliers which is consistent with observations from Figure S3.

In total, the WGCNA analysis identified 48 gene modules. The correlation between the eigen gene of these modules and part of the clinical traits collected under the GRADS protocol is shown in **Figure S*5***. The Modules 1, 4, 18, 33 and 47 were chosen for further clustering analysis due to their significant correlation (p value<0.05) with extensive clinical traits. This analysis is not adjusted for cell differentials.



**Figure S5**: Heatmap showing the correlation of the 48 identified gene modules and the given clinical traits including the demographics, PFTs, CT scan features, cell differentials and the SCADDING staging.

## ***Cluster Analysis using Chosen Modules***

For each of the 5 chosen modules, we applied K-means to cluster patients using its member genes only. The optimal number of clusters were chosen based on data visualization using heatmaps and multiple internal clustering criteria calculated by the Nbclust R package including the Dunn Index, the Silhouette Index, the Calinski and Harabasz Index and the Connectivity. The clinical relevance of these identified clusters or molecular endotypes was further evaluated by correlating the clustering results to all other clinical traits collected under the GRADS protocol. The Chi-square test and the Kruskal-Wallis test were used when correlating the clustering results to categorical and continuous patient characteristics, respectively. The summary of this analysis and results for each identified cluster is presented in the Supplementary Table E3.

# **Validation Analysis**

## ***Freiburg Cohort***

We validated our findings using a microarray expression dataset from an independent cohort of Sarcoidosis patients from Freiburg, Germany. The consents were collected following institutional IRB protocols. Bronchoscopy with bronchoalveolar lavage was performed in these patients to obtain the BAL cells. The gene expression profile of these BAL cells was quantified using the Affymetrix Human Gene 1.0 ST Arrays. The raw data was quantile normalized using the affy R package. Principal component analysis was conducted which found no outlier samples. All the 50 samples were processed in the same batch on the same day so there is no batch effect. In total, this dataset recruited 50 sarcoidosis patients. There were 12 clinical traits that were recorded in both Freiburg and GRADS cohorts including scadding staging, age, gender, PFTs, PFTs% predicted and BAL cell differentials. We were unable to obtain any CT imaging features in the Freiburg cohort and therefore these features were not validated using GRADS cohort.

## ***Validating the WGCNA results***

To validate the novel molecular endotypes of sarcoidosis defined in the GRADS cohort, we cluster the patients in the Freiburg cohort using genes from each of the 5 chosen gene modules individually. The two extreme clusters (indicated in column names of Table 2) were compared for each of the 12 overlapping clinical traits in Freiburg cohort, which can be considered as one type of association between the clustering results or molecular endotypes with the clinical traits. This correlation was compared between the GRADS and the Freiburg cohorts for validation in Table 2.

## ***Validating the supervised analysis results for PFTs***

Since the novel molecular endotypes identified by WGCNA were strongly associated and confounded by the cell differentials, we applied the same supervised analysis to the PFTs that are available in both GRADS and Freiburg cohorts. For each PFT, the two sets of identified associated genes were compared between the two cohorts and the significance of the overlap was assessed using chi-square test.

We found that the significant genes associated with the FVC% predicted and the FEV1% predicted from the two cohorts significantly overlapped (p value=3.22 x 10-7 and 0.02, respectively).

**RESULTS**

***Genes increasing with increased macrophage fraction in BAL***

Among the genes most associated (Spearman’s rho > 0.2, FDR < 0.05) with increased macrophage fractions counts were the known alveolar macrophage marker SIGLEC11, ANXA1, ALOX5, CXCL5, ITGA5, LRP1, TREM, IRS2. Interestingly PECAM1, a known macrophage marker was among the most associated genes (Spearman’s rho 0.36, FDR <0.05) with increased macrophage counts, potentially reflecting monocyte differentiation into macrophages and modulation of macrophage function (6). Similar to genes associated with lymphocyte counts, genes associated with decreased macrophage counts overlapped with genes associated with increased Scadding stage (71), hilar lymphadenopathy (213), and bronchial wall thickening (254), but they also overlapped with genes associated with increased traction bronchiectasis (28) and reticular abnormalities (71) (Figure 2A) potentially reflecting unique transcriptional programs in macrophages in lung fibrosis. Among the overlapping decreased genes were SLC40A, PLXNC1, and CMKLR1 known to be involved in initiation and resolution of inflammation (7, 8). Some of the functional associations are most informative when looked at together (Figure 2b). The increase in alveolar macrophages ~~wa~~s associated with an increase in development and fibrosis related pathways such as PI3K/AKT, MAPK, BMP7 and K-RAS signaling (Figure 2b, Supplement Tables).

***Genes decreasing with increase in eosinophil fraction in BAL are associated with increase in airway thickness***

Mild increases in BAL eosinophil counts have been reported in progressive sarcoidosis (9). In our cohort, the BAL eosinophil fraction has an average of 0.24% and range from 0% to 5.5%. Out of 115 genes associated with BAL eosinophil fraction, 27 genes (9 positively and 18 negatively) were correlated with bronchial thickening and 10 were negatively correlated with reticular abnormality. CAMP, IRS2, ST3GAL2, SPIRE2, and FHL1 were negatively correlated with eosinophil counts, bronchial wall thickening and reticular abnormality. Although correlation between bronchial wall thickening and the eosinophil counts had marginal significance (p value= 0.056 and Spearman rho=0.135) in our data, common negatively correlated genes such as CAMP and IRS2 were identified. CAMP was previously shown to be decreased in severe sarcoidosis (10). Decrease in IRS2 led to pulmonary inflammation and accumulation of eosinophils in allergic lung inflammation and remodeling (11).

**REFERENCES**

1. S. A. FastQC: a quality control tool for high throughput sequence data. Available online at: <http://wwwbioinformaticsbabrahamacuk/projects/fastqc>. 2010.

2. 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.

3. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-9.

4. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28(5):511-5.

5. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.

6. Kral JB, Schrottmaier WC, Salzmann M, Assinger A. Platelet Interaction with Innate Immune Cells. Transfus Med Hemother. 2016;43(2):78-88.

7. Yoshimura T, Oppenheim JJ. Chemerin reveals its chimeric nature. J Exp Med. 2008;205(10):2187-90.

8. Konig K, Marth L, Roissant J, Granja T, Jennewein C, Devanathan V, et al. The plexin C1 receptor promotes acute inflammation. Eur J Immunol. 2014;44(9):2648-58.

9. Ziegenhagen MW, Rothe ME, Schlaak M, Muller-Quernheim J. Bronchoalveolar and serological parameters reflecting the severity of sarcoidosis. Eur Respir J. 2003;21(3):407-13.

10. Barna BP, Culver DA, Kanchwala A, Singh RJ, Huizar I, Abraham S, et al. Alveolar macrophage cathelicidin deficiency in severe sarcoidosis. J Innate Immun. 2012;4(5-6):569-78.

11. Dasgupta P, Dorsey NJ, Li J, Qi X, Smith EP, Yamaji-Kegan K, et al. The adaptor protein insulin receptor substrate 2 inhibits alternative macrophage activation and allergic lung inflammation. Sci Signal. 2016;9(433):ra63-ra.