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

Milica Vukmirovic*[1,11], Xiting Yan*[1,2], Kevin F. Gibson[3], Mridu Gulati[1], Jonas C. Schupp[1], Giuseppe DeIuliis[1], Taylor S. Adams[1], Buqu Hu[1], Antun Mihaljinec[1], Tony N. Woolard[1], Heather Lynn[1,8], Nkiruka Emeagwali[1], Erica L. Herzog[1], Edward S. Chen[4], Alison Morris[3], Joseph K. Leader[14], Yingze Zhang[3], Joe G. N. Garcia[8], Lisa A. Maier[5], Ronald G. Collman[9], Wonder P. Drake[6], Michael J. Becich[13], Harry Hochheiser[13], Steven R. Wisniewski[3], Panayiotis V. Benos[15], David R. Moller[4], Antje Prasse[10,12], Laura L. Koth[7], Naftali Kaminski**[1]

[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] Department of Medicine, 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] Department of Medicine, Division of Respirology, McMaster University, Hamilton, ON Canada

[12] Fraunhofer ITEM, Hannover, Germany

[13] Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, PA/US

[14] Department of Radiology, University of Pittsburgh School of Medicine, Pittsburgh, PA/US

[15] Department of Computational and Systems Biology and Department of Computer Science, University of Pittsburgh, Pittsburgh, PA, US

On behalf of the GRADS Investigators
* Equally contributing authors
** Corresponding author
Address correspondence: 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, R01 HL127349, U01 HL137159, CTSI U54 grant 9 UL1 TR000005, CDC NMVB 5U24 OH009077, CTSA UL1 TR002535, R21 LM012884, German Network for Lung Research (DZL/ BREATH) and Fraunhofer CIMD Forschungscluster.

Table of Contents

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.

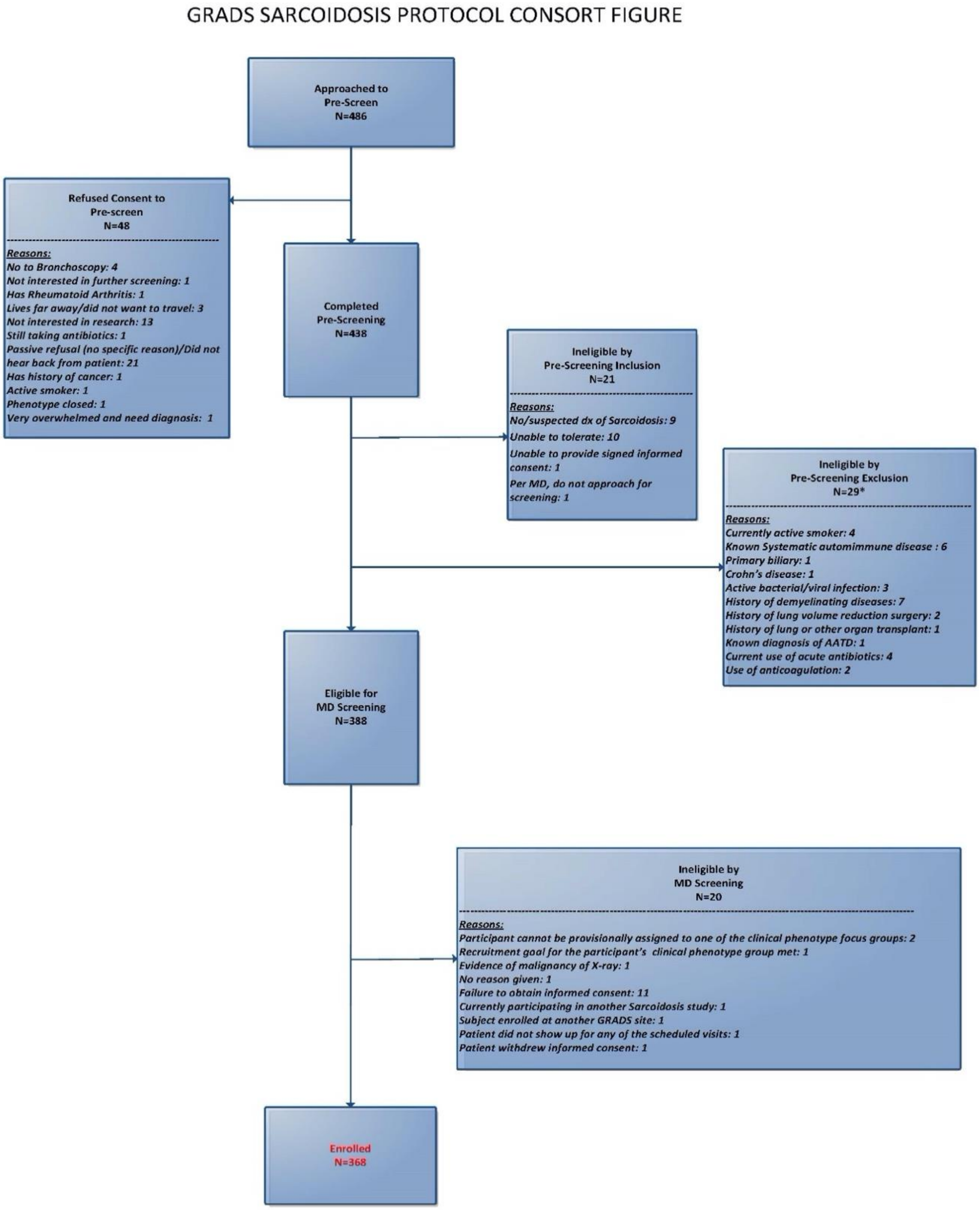
Sample preparation and RNA sequencing

All patients involved in this study have signed a consent to participate in this study in accordance with institutional IRB protocols (**Figure S1**). 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 S1** and Supplemental 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.


GRADS SARCOIDOSIS PROTOCOL CONSORT FIGURE
Approached to Pre-Screen N=486
Refused Consent to Pre-screen N=48
Reasons:
No to Bronchoscopy: 4
Not interested in further screening: 1
Has Rheumatoid Arthritis: 1
Lives far away/did not want to travel: 3
Not interested in research: 13
Still taking antibiotics: 1
Passive refusal (no specific reason)/Did not hear back from patient: 21
Has history of cancer: 1
Active smoker: 1
Phenotype closed: 1
Very overwhelmed and need diagnosis: 1
Completed Pre-Screening N=438
Ineligible by Pre-Screening Inclusion N=21
Reasons:
No/suspected dx of Sarcoidosis: 9
Unable to tolerate: 10
Unable to provide signed informed consent: 1
Per MD, do not approach for screening: 1
Ineligible by Pre-Screening Exclusion N=29*
Reasons:
Currently active smoker: 4
Known Systematic automimmune disease : 6
Primary biliary: 1
Crohn's disease: 1
Active bacterial/viral infection: 3
History of demyelinating diseases: 7
History of lung volume reduction surgery: 2
History of lung or other organ transplant: 1
Known diagnosis of AATD: 1
Current use of acute antibiotics: 4
Use of anticoagulation: 2
Eligible for MD Screening N=388
Ineligible by MD Screening N=20
Reasons:
Participant cannot be provisionally assigned to one of the clinical phenotype focus groups: 2
Recruitment goal for the participant's clinical phenotype group met: 1
Evidence of malignancy of X-ray: 1
No reason given: 1
Failure to obtain informed consent: 11
Currently participating in another Sarcoidosis study: 1
Subject enrolled at another GRADS site: 1
Patient did not show up for any of the scheduled visits: 1
Patient withdrew informed consent: 1
Enrolled N=368

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 S2a**). 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 S2b 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.

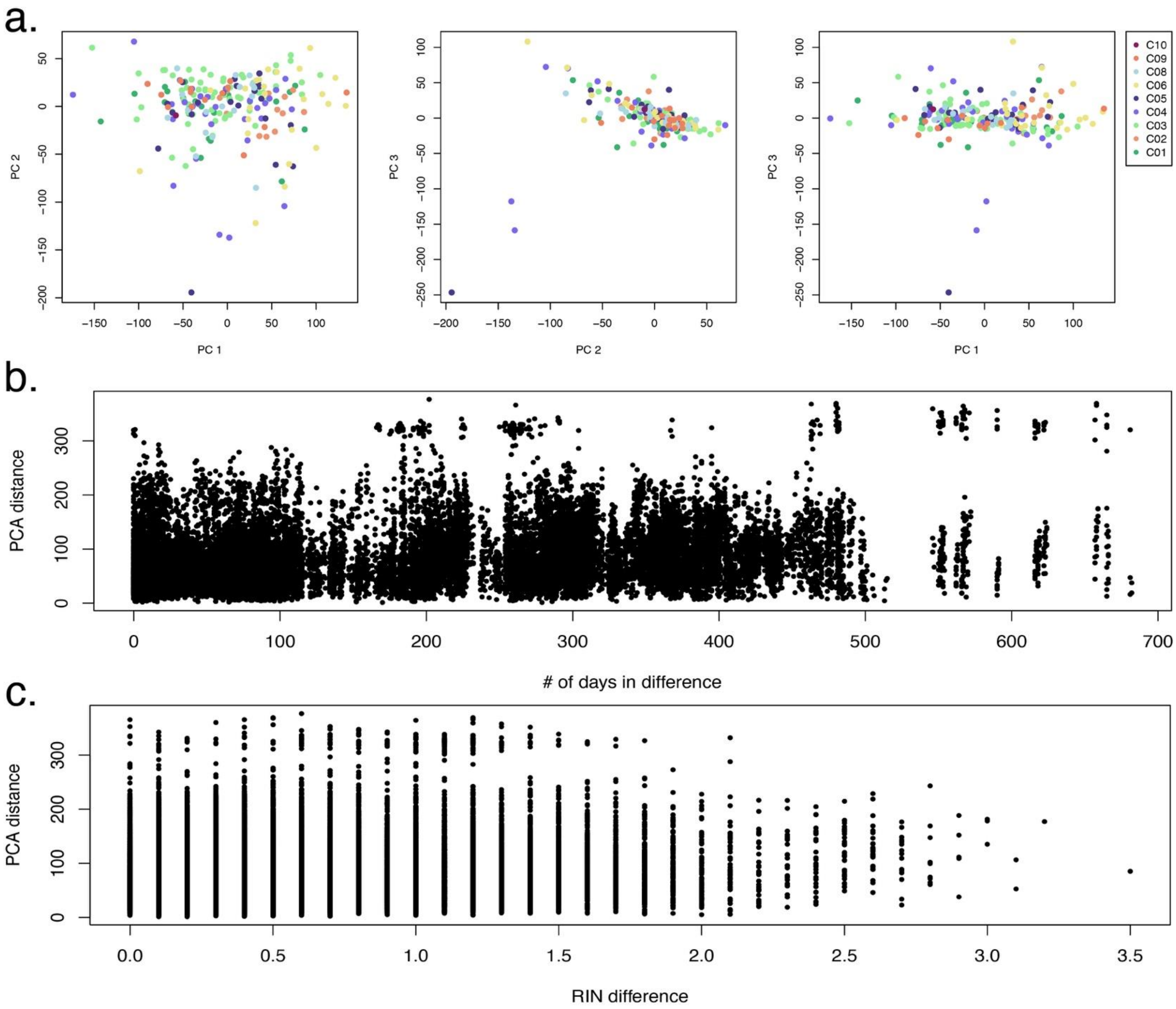


Figure S2: Technical effect examination of a). sample collection center, b). sequencing date and c). RNA integrity number.

Supervised Analysis

The supervised analysis identified gene signatures associated with 24 clinical traits (age, gender, race, FVC, FVC% predicted, FEV1, FEV1% predicted, DLCO, DLCO% predicted, FEV1/FVC ratio, bronchial wall thickening, bronchiectasis severity, ground glass, honeycombing, reticular abnormality, traction bronchiectasis, mediastinal lymphadenopathy, hilar lymphadenopathy, Scadding, total BAL cell count, macrophage %, eosinophil %, lymphocyte % and neutrophil %) 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 Supplemental 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 (Supplemental 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 MetaCore^{TM} 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 S3**). To decide exactly which samples to exclude, we applied WGCNA with trimming of 0, 1, 2, 3,…, 6 branches from the top of the

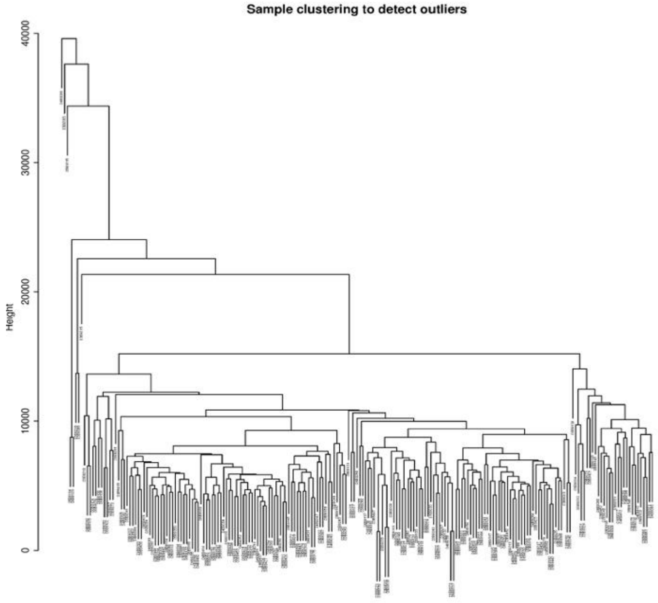


Figure S3: Hierarchical clustering tree of all 209 samples to identify potential outliers for the WGCNA analysis.

clustering tree and compared their clustering results (**Figure S4**). 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.

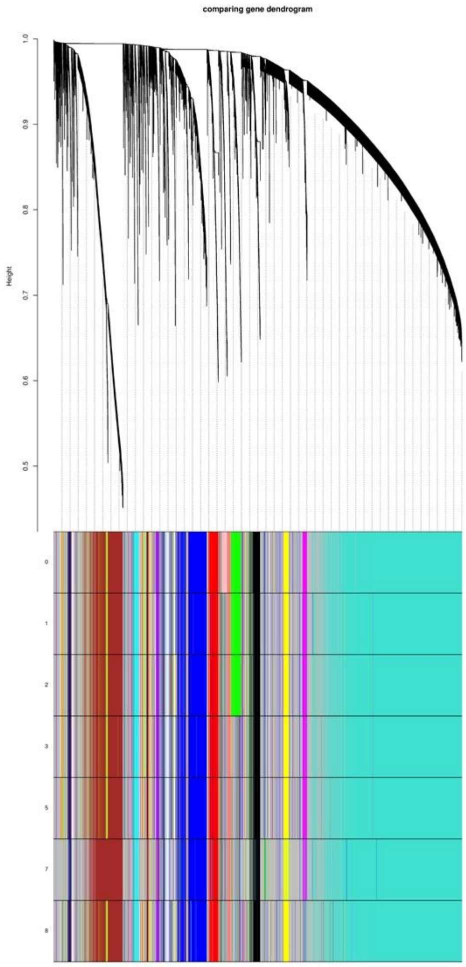


Figure S4: Comparison of the clustering results by trimming 0, 1, 2, …, 6 branches from the top of the clustering tree in Figure S3. The dendrogram on top shows the hierarchical clustering tree of all genes by WGCNA analysis. The 7 color bars on bottom show the clustering results of all these genes after 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.

Correlating gene modules with clinical traits

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**

S5. The Modules 1, 4, 18, 33 and 47 were chosen for further clustering analysis due to their significant correlation (p value<0.05) with the highest number of clinical traits or unique combination of the clinical traits. The priority was given to the modules with a largest number of genes in the module. This analysis was not adjusted for demographics, smoking, cell differentials or specific treatment because none of these had strong association with our chosen gene modules. The distribution of treatment type and the time of last treatment can be found in Table S2.

Table S2. Distribution of the treatment type and the time of last treatment in GRADS cohort.

Systematic corticorsteroids

Treatment	Currently taking	Within last 90 days but not currently	Past but not within last 90 days	Never
Prednisone	32 (15.5%)	21 (10.1%)	94 (45.4%)	60 (29.0%)
Medrol	0 (0%)	1 (0.5%)	9 (4.3%)	197 (95.2%)
Dexamethasone	0 (0%)	0 (0%)	8 (3.9%)	199 (96.1%)

Immune Suppressive Agents

Treatment	Currently taking	Within last 90 days but not currently	Past but not within last 90 days	Never
Adalimumab (Humira)	5 (2.4%)	1 (0.5%)	4 (1.9%)	197 (95.2%)
Azathioprine (Imuran)	5 (2.4%)	1 (0.5%)	6 (2.9%)	195 (94.2%)
Chlorambucil (Leukeran)	0 (0%)	0 (0%)	1 (0.5%)	206 (99.5%)
Colchicine	1 (0.5%)	3 (1.4%)	0 (0%)	203 (98.1%)
Cyclophosphamide (Cytoxan)	0 (0%)	0 (0%)	2 (1.0%)	205 (99.0%)
Cyclosporine (Gengraf,Neoral,Sandimmune)	0 (0%)	0 (0%)	3 (1.4%)	204 (98.6%)
Etanercept (Enbrel)	0 (0%)	0 (0%)	3 (1.4%)	204 (98.6%)
Hydroxychloroquine (Plaquinil)	13 (6.3%)	3 (1.4%)	17 (8.2%)	174 (84.1%)
Infliximab (Remicade)	2 (0.9%)	13 (6.3%)	0 (0%)	192 (92.8%)
IVIG	0 (0%)	0 (0%)	1 (0.5%)	206 (99.5%)
Leflunamide (Arava)	2 (1.0%)	1 (0.5%)	4 (1.9%)	200 (96.6%)
Methotrexate (Rheumatrex)	28 (13.5%)	4 (1.9%)	28 (13.5%)	147 (71.1%)
Mycophenolate mofitil (CellCept)	7 (3.4%)	10 (4.8%)	0 (0%)	190 (91.8%)
Pentoxyfiline (Trental)	3 (1.4%)	204 (98.6%)	0 (0%)	0 (0%)

Antibiotics

Treatment	Currently taking	Within last 90 days but not currently	Past but not within last 90 days	Never
Augmentin	0 (0%)	154 (74.4%)	47 (22.7%)	6 (2.9%)
Avelox	0 (0%)	0 (0%)	16 (7.7%)	191 (92.3%)

Azithromycin	0 (0%)	15 (7.2%)	62 (30.0%)	130 (62.8%)
Bactrim DS	1 (0.5%)	4 (1.9%)	45 (21.8%)	157 (75.8%)
Ciprofloxacin	0 (0%)	3 (1.4%)	49 (23.7%)	155 (74.9%)
Clarithromycin	0 (0%)	1 (0.5%)	16 (7.7%)	190 (91.8%)
Clindamycin	0 (0%)	0 (0%)	16 (7.7%)	191 (92.3%)
Doxycycline	1 (0.5%)	4 (1.9%)	44 (21.3%)	158 (76.3%)
INH	0 (0%)	0 (0%)	4 (1.9%)	203 (98.1%)
Levaquin	0 (0%)	3 (1.5%)	29 (14.0%)	175 (84.5%)
Minocycline	0 (0%)	0 (0%)	9 (4.3%)	198 (95.7%)
Pyrazinamide	0 (0%)	0 (0%)	2 (1.0%)	205 (99.0%)
Rifampin	0 (0%)	0 (0%)	2 (1.0%)	205 (99.0%)

Reflux

Treatment	Currently taking	Within last 90 days but not currently	Past but not within last 90 days	Never
Aciphex (rabeprazole)	1 (0.5%)	1 (0.5%)	6 (2.9%)	199 (96.1%)
Nexium (esomeprazole)	5 (2.4%)	1 (0.5%)	38 (18.4%)	163 (78.7%)
Prevacid (lansoprazole)	4 (1.9%)	2 (1.0%)	31 (15.0%)	170 (82.1%)
Prilosec (omeprazole)	37 (17.9%)	10 (4.8%)	40 (19.3%)	120 (58.0%)
Protonix (pantoprazole)	7 (3.4%)	11 (5.3%)	0 (0%)	189 (91.3%)

H2 Blockers

Treatment	Currently taking	Within last 90 days but not currently	Past but not within last 90 days	Never
Axid (nizantidine)	0 (0%)	0 (0%)	1 (0.5%)	206 (99.5%)
Pepcid (famotidine)	2 (1.0%)	4 (1.9%)	39 (18.8%)	162 (78.3%)
Zantac (ranitidine)	10 (4.8%)	6 (2.9%)	47 (22.7%)	144 (69.6%)

Inhale Steroids

Treatment	Currently taking	Within last 90 days but not currently	Past but not within last 90 days	Never
Aerobid	0 (0%)	0 (0%)	2 (1.0%)	205 (99.0%)
Advair	17 (8.2%)	8 (3.9%)	53 (25.6%)	129 (62.3%)
Azmacort	0 (0%)	0 (0%)	10 (4.8%)	197 (95.2%)
Flovent	5 (2.4%)	4 (1.9%)	42 (20.3%)	156 (75.4%)
Pulmicort	4 (1.9%)	20 (9.7%)	0 (0%)	183 (88.4%)
Serevent	1 (0.5%)	8 (3.8%)	0 (0%)	198 (95.7%)
QVAR	4 (1.9%)	12 (5.8%)	0 (0%)	191 (92.3%)

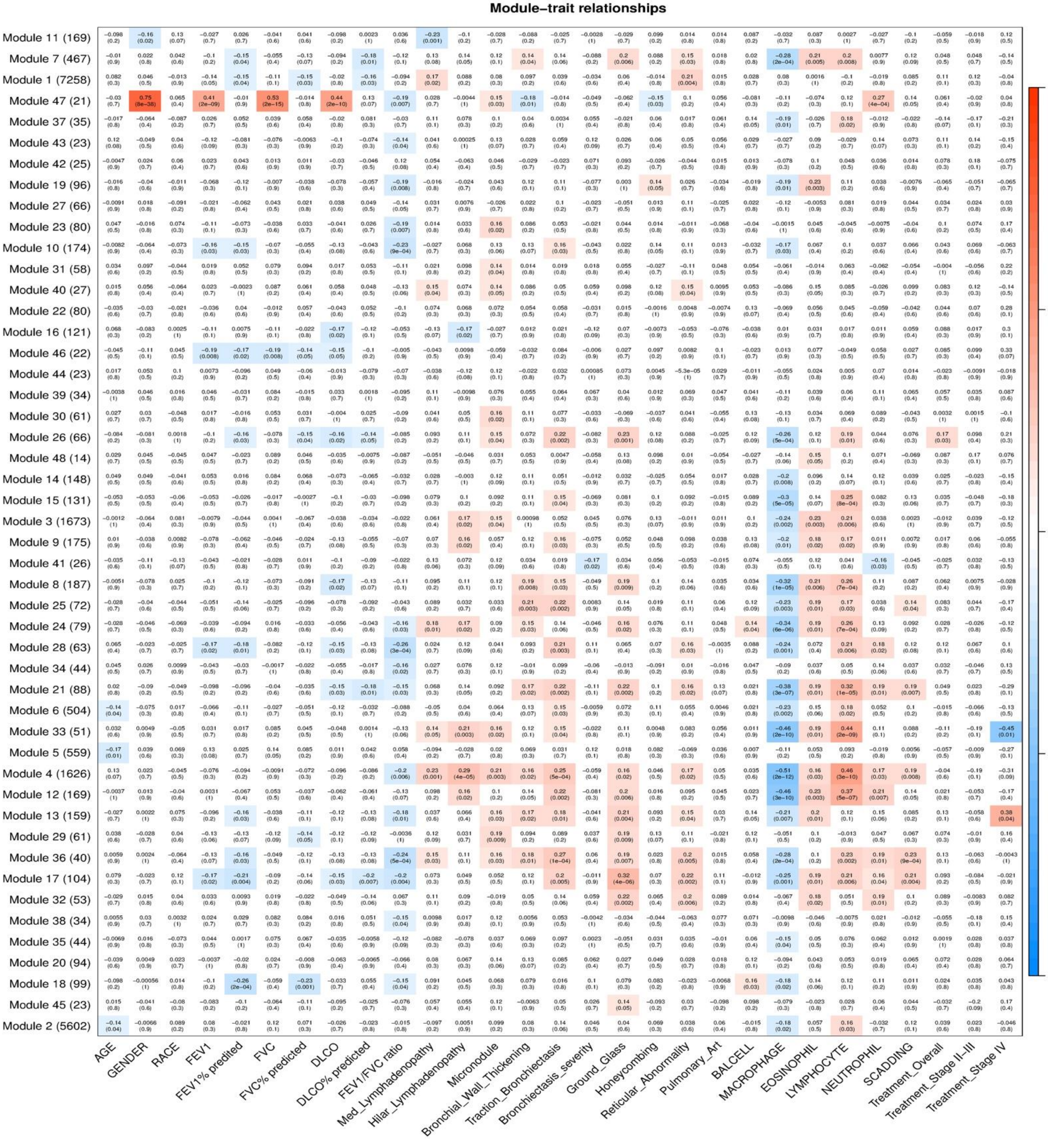


Figure S5: Heatmap showing the correlation of the 48 identified gene modules and the given clinical traits including the demographics, PFTs, CT scan features, cell differentials and the SCADDING staging.

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 2,289 clinical traits collected under the GRADS protocol, including 204 environmental factors. 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 Supplemental 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 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. The PFTs% predicted values in GRADS cohort were calculated using the Hankinson’s race specific reference equations (6). In Freiburg cohort, these values were determined using the GLI reference equations (7). Our validation analysis is impacted by this difference because the PFTs% predicted values were not directly compared for validation. Instead, we calculated the correlation of genes and gene modules with FVC% predicted and FEV1% predicted in each cohort separately and compared these correlations between the two cohorts. In addition, we assessed the correlations and associations using non-parametric approaches including Spearman correlation and Wilcoxon Rank Sun test, which are robust to such difference.

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 and visualized in Figure 4) 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. To assess the significance of validation, we conducted hypergeometric test on the overlap of the results between the two cohorts in Table 2. The p values can be round in the column title of Table 2. Due to the small number of overlapping features (12) which corresponds to a small sample size for the hypergeometric test, we defined endotypes with a less stringent threshold ($p<0.1$) as significantly validated.

To remove and examine the effect of cell differentials on our validation results, we adjusted the gene expression data in GRADS and Freiburg cohorts using a linear regression model to remove the BAL cell differential effect. The adjusted gene expression was used to cluster patients in both cohorts in the same way as the unadjusted gene expression data. Clinical traits significantly associated with the identified patient clusters were also identified in both cohorts and compared for validation again in the same way as the unadjusted gene expression data. The validation results using adjusted data

(Table S3) showed that modules 47, 4, 18, and 1 were validated (hypergeometric test p<0.05), indicating the robustness of our validation to BAL cell differentials of these modules.

In addition, the significant associated clinical traits in the unadjusted data in Table 2 disappeared after the data adjustment for most endotypes except for the endotype of gender and PFT (basal). This suggests that there is a correlation between important clinical traits of Sarcoidosis and BAL cell differentials, which is consistent with the fact that BAL cell differentials are also indicative of disease severity. Therefore, by removing the cell differential effect on gene expression, we also removed the effect of clinical traits important to Sarcoidosis in the expression data. The BAL differentials should be considered as disease relevant effects instead of technical effects to avoid removing important disease effect.

Validating the supervised analysis results

We also applied the same supervised analysis to the 12 overlapping clinical traits that are available in both GRADS and Freiburg cohorts. For each trait, the two sets of identified associated genes were compared between the two cohorts and the significance of overlap was assessed using chi-square test. In this analysis, due to the small sample size of Freiburg cohort, we consider genes with a nominal p value<0.05 as significant genes in each cohort and only genes identified in both cohorts with the same association direction (both negatively or positively correlated) in the two cohorts were considered as overlapping genes in this analysis. We found that genes associated with Scadding, Neutrophil %, Lymphocypte %, FVC, FVC% predicted, FEV1% predicted and FEV1/FVC ratio from the two cohorts significantly overlapped (Table S4).

Table S3. Comparison of each endotype's association with the 12 overlapping clinical traits in GRADS and Freiburg cohorts. The p values in the column titles assess the significance of the validation based on the hypergeometric test.

	Module 47 Gender module (p<0.01)		Module 4 Hilar Lymphadenopathy and Acute Lymphocytic Inflammation (p<0.01)		Module 33 Multiorgan involvement with increased immune response (p=1)		Module 18 Chronic sarcoidosis (p<0.01)		Module 1 Extraocular organ involvement and PI3K activation (p<0.01)	
	GRADS (A vs B) P value	**Freiburg (A vs B) P value**	**GRADS (B vs C) P value**	**Freiburg (B vs C) P value**	**GRADS (C vs D) P value**	**Freiburg (B vs C) P value**	**GRADS (C vs D) P value**	**Freiburg (A vs C) P value**	**GRADS (A vs B) P value**	**Freiburg (B vs C) P value**
SCADDING	0.05	0.11	0.07	0.14	0.63	0.18	0.08	0.22	0.32	0.51
AGE	0.98	0.59	0.38	0.56	0.79	0.29	0.05	0.31	0.06	0.60
GENDER	**5.0x10^{-4}**	**5.0x10^{-4}**	0.68	1.00	1.00	0.68	0.51	0.42	0.36	0.77
MACROPHAGE	0.70	0.05	0.40	0.14	**4.6x10^{-4}**	0.83	0.68	0.72	0.40	0.55
LYMPHOCYTES	0.71	0.05	0.33	0.16	**2.4x10^{-4}**	0.81	0.75	0.91	0.30	0.55
NEUTROPHILS	0.24	0.47	0.71	0.79	0.12	0.71	0.60	0.78	0.84	0.11
EOSINOPHILS	0.73	0.84	0.14	0.64	0.96	0.93	0.69	0.25	0.42	0.35
FVC	**2.4x10^{-13}**	**3.0x10^{-3}**	0.24	0.97	0.26	1.00	0.12	0.82	0.70	0.66
FEV1	**2.0x10^{-8}**	**8.0x10^{-3}**	0.68	0.92	0.65	0.65	0.21	0.85	0.41	0.90
FVC% predicted	0.92	0.14	0.70	0.54	0.32	0.15	**0.04**	0.98	0.26	0.60
FEV1% predicted	0.92	0.75	0.58	0.58	0.41	0.18	0.21	0.60	0.11	0.86
FEV1/FVC ratio	0.69	**0.03**	0.15	0.46	**8.6x10^{-3}**	0.51	0.45	0.29	0.36	0.32

Table S4. Validation of supervised analysis for the 10 overlapping clinical traits between GRADS and Freiburg cohorts.

Traits	# of associated genes (GRADS)	# of associated genes (Freiburg)	# of overlapping genes	Chi-square p value
SCADDING	2,394	1,682	431	**1.2x10^{-73}**
AGE	1,373	556	26	0.05
GENDER	1,075	409	14	0.10
Macrophage %	3,487	445	89	0.15
Eosinophil %	2,674	492	67	0.87
Neutrophil %	2,420	755	72	**0.04**
Lymphocyte %	5,310	300	31	**3.4x10^{-10}**
FVC% predicted	2,612	622	16	**7.3x10^{-15}**
FEV1% predicted	3,243	348	39	**1.5x10^{-2}**
FEV1/FVC ratio	1,928	375	20	**6.4x10^{-3}**

Genes associated with BAL macrophage and eosinophil differentials

Genes increasing with increased macrophage fraction in BAL

Among genes most associated (Spearman's rho > 0.2, FDR < 0.05) with increased macrophage fraction were the known alveolar macrophage markers 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 fraction, potentially reflecting monocyte differentiation into macrophages and modulation of macrophage function (8). Similar to genes associated with lymphocyte differential, genes associated with decreased macrophage fraction 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 (9, 10). Some of the functional associations are most informative when looked at together (Figure 2b). The increase in BAL macrophages fraction was associated with an increase in development and fibrosis related pathways such as PI3K/AKT, MAPK, BMP7 and K-RAS signaling (Figure 2b, supplementary website).

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 (11). 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 (12). Decrease in IRS2 led to pulmonary inflammation and accumulation of eosinophils in allergic lung inflammation and remodeling (13).

Table S5. Breakdown of patients based on PFTs% predicted severity.

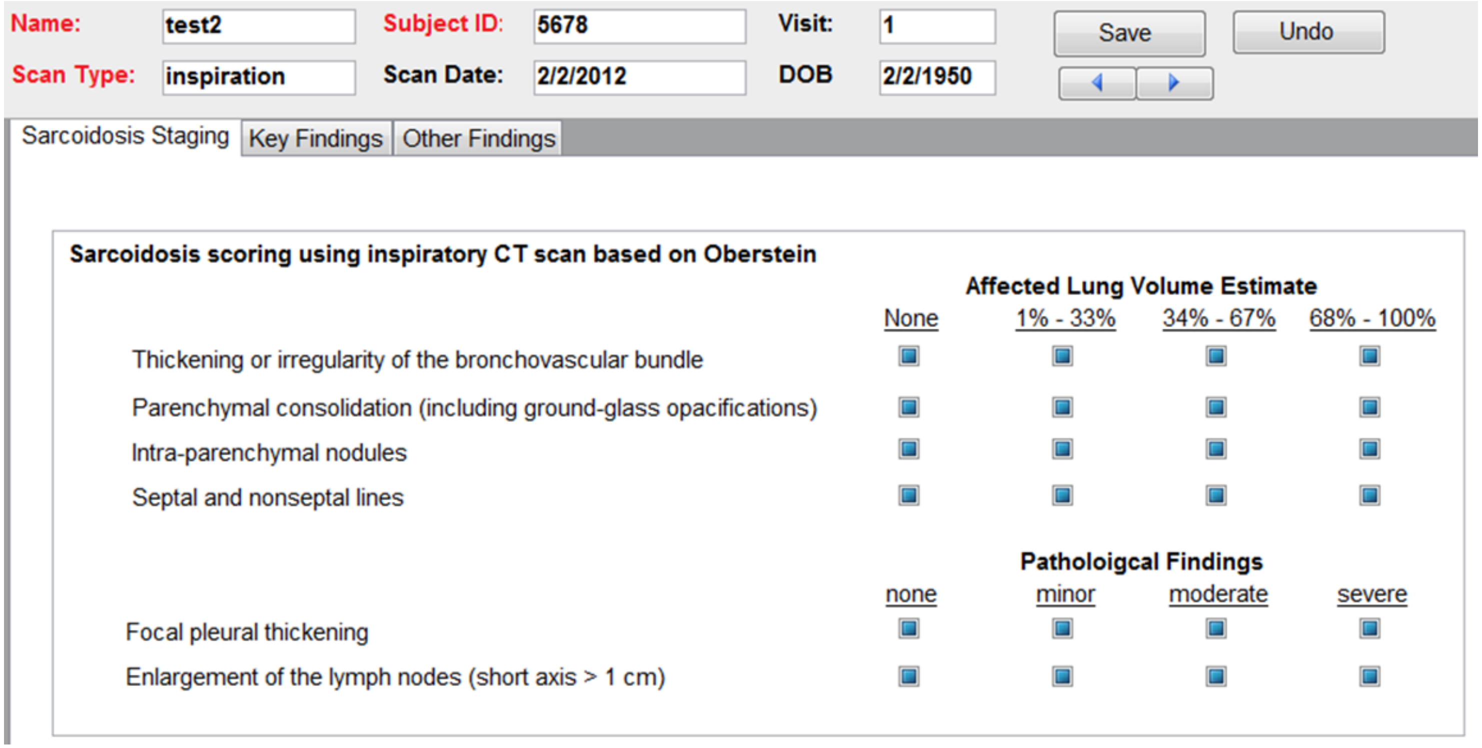
PHENOTYPE GROUPS	1. MULTIORGAN	2. NON ACUTE STAGE I UNTREATED	3. STAGE II-III TREATED	4. STAGE II-III UNTREATED	5. STAGE IV TREATED	6. STAGE IV UNTREATED	7. ACUTE UNTREATED	8. REMITTING UNTREATED
TOTAL, n	23	25	34	42	19	12	14	40
FEV1% PRED severity								
Mild, n (%)	21 (91.3%)	19 (76.0%)	22 (64.7%)	33 (78.6%)	7 (36.8%)	9 (75.0%)	12 (85.7%)	38 (95.0%)
Moderate, n (%)	2 (8.7%)	3 (12.0%)	5 (14.7%)	5 (11.9%)	7 (36.8%)	0 (0.0%)	0 (0.0%)	1 (2.5%)
Moderately severe, n (%)	0 (0.0%)	0 (0.0%)	5 (14.7%)	1 (2.4%)	2 (10.5%)	1 (8.3%)	0 (0.0%)	0 (0.0%)
Severe, n (%)	0 (0.0%)	0 (0.0%)	2 (5.9%)	0 (0.0%)	3 (15.8%)	2 (16.7%)	0 (0.0%)	0 (0.0%)
Very severe, n (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
NA, n (%)	0 (0.0%)	3 (12.0%)	0 (0.0%)	3 (7.1%)	0 (0.0%)	0 (0.0%)	2 (14.3%)	1 (2.5%)
FVC% PRED severity								
Mild, n (%)	23 (100%)	20 (80.0%)	26 (76.5%)	37 (88.1%)	11 (57.9%)	9 (75.1%)	12 (85.7%)	39 (97.5%)
Moderate, n (%)	0 (0.0%)	2 (8.0%)	3 (8.8%)	1 (2.4%)	7 (36.8%)	1 (8.3%)	0 (0.0%)	0 (0.0%)
Moderately severe, n (%)	0 (0.0%)	0 (0.0%)	4 (11.8%)	1 (2.4%)	1 (5.26%)	1 (8.3%)	0 (0.0%)	0 (0.0%)
Severe, n (%)	0 (0.0%)	0 (0.0%)	1 (2.9%)	0 (0.0%)	0 (0.0%)	1 (8.3%)	0 (0.0%)	0 (0.0%)
Very severe, n (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
NA, n (%)	0 (0.0%)	3 (12.0%)	0 (0.0%)	3 (7.1%)	0 (0.0%)	0 (0.0%)	2 (14.3%)	1 (2.5%)
DLCO% PRED severity								
Mild, n (%)	21 (91.3%)	21 (84.0%)	21 (61.8%)	37 (88.1%)	14 (73.7%)	9 (75.1%)	11 (78.6%)	36 (90.0%)
Moderate, n (%)	2 (8.7%)	2 (8.0%)	8 (23.5%)	2 (4.8%)	5 (26.3%)	1 (8.3%)	1 (7.1%)	3 (7.5%)
Severe, n (%)	0 (0.0%)	0 (0.0%)	4 (11.8%)	0 (0.0%)	0 (0.0%)	1 (8.3%)	0 (0.0%)	0 (0.0%)
NA, n (%)	0 (0.0%)	2 (8.0%)	1 (2.9%)	3 (7.1%)	0 (0.0%)	1 (8.3%)	2 (14.3%)	1 (2.5%)

Definition of PFT severity:

FEV1% PRED and FVC% PRED: Mild (>70%), Moderate (60-69%), Moderately severe (50-59%), Severe (35-49%) and Very severe (<35%);

DLCO% PRED: Mild (>60%), Moderate (40-60%), and Severe (<40%).

Figure S6. An overview of GRADS CT scoring forms

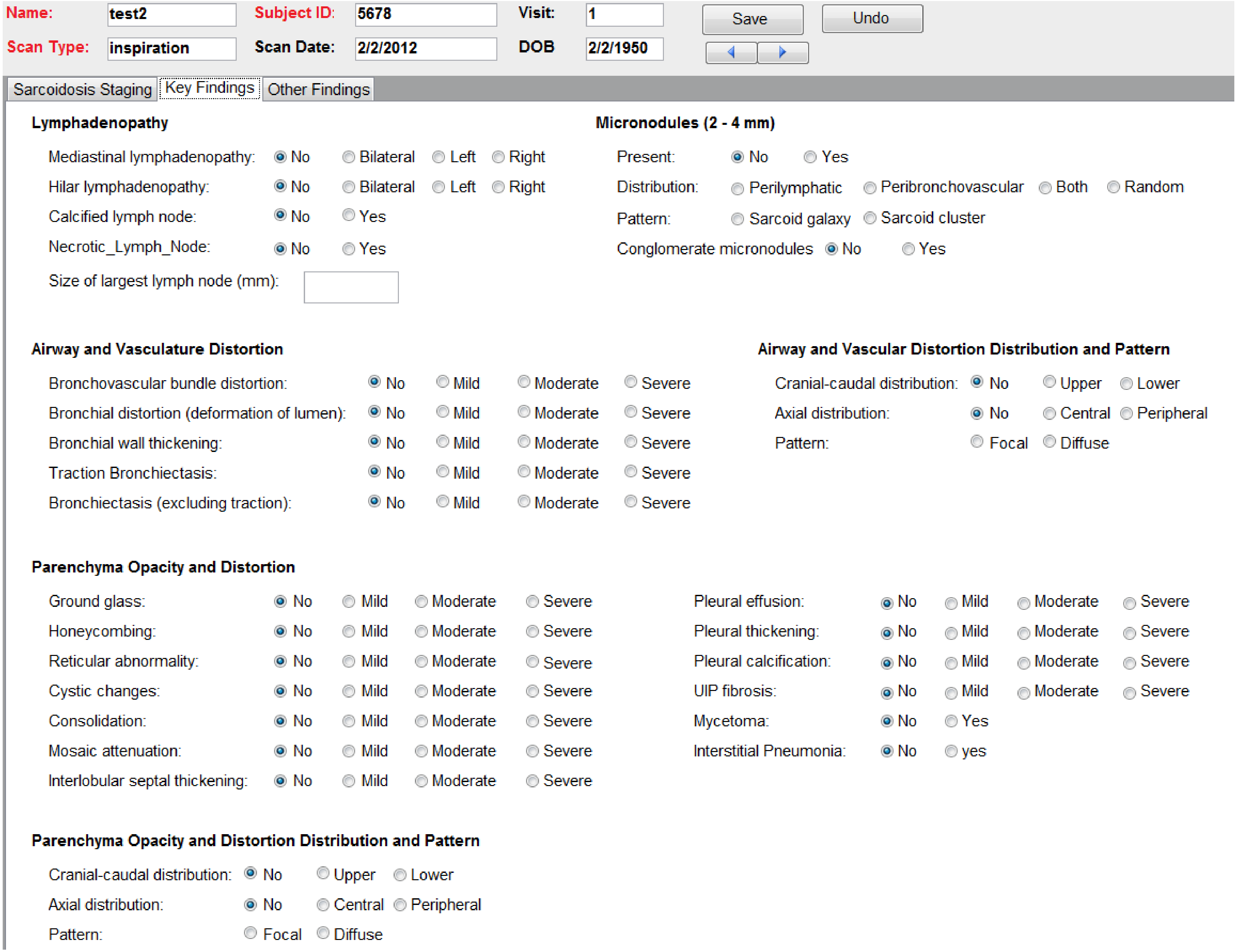


Name:	test2	Subject ID:	5678	Visit:	1
Scan Type:	inspiration	Scan Date:	2/2/2012	DOB	2/2/1950

Sarcoidosis scoring using inspiratory CT scan based on Oberstein

	Affected Lung Volume Estimate			
	None	1% - 33%	34% - 67%	68% - 100%
Thickening or irregularity of the bronchovascular bundle	□	□	□	□
Parenchymal consolidation (including ground-glass opacifications)	□	□	□	□
Intra-parenchymal nodules	□	□	□	□
Septal and nonseptal lines	□	□	□	□

	Patholoigcal Findings			
	none	minor	moderate	severe
Focal pleural thickening	□	□	□	□
Enlargement of the lymph nodes (short axis > 1 cm)	□	□	□	□



Name:	test2	**Subject ID:**	5678	**Visit:**	1
Scan Type:	inspiration	**Scan Date:**	2/2/2012	**DOB**	2/2/1950

Lymphadenopathy

Mediastinal lymphadenopathy:	● No	○ Bilateral	○ Left	○ Right
Hilar lymphadenopathy:	● No	○ Bilateral	○ Left	○ Right
Calcified lymph node:	● No	○ Yes		
Necrotic_Lymph_Node:	● No	○ Yes		
Size of largest lymph node (mm):				

Micronodules (2 - 4 mm)

Present:	● No	○ Yes		
Distribution:	○ Perilymphatic	○ Peribronchovascular	○ Both	○ Random
Pattern:	○ Sarcoid galaxy	○ Sarcoid cluster		
Conglomerate micronodules	● No	○ Yes		

Airway and Vasculature Distortion

Bronchovascular bundle distortion:	● No	○ Mild	○ Moderate	○ Severe
Bronchial distortion (deformation of lumen):	● No	○ Mild	○ Moderate	○ Severe
Bronchial wall thickening:	● No	○ Mild	○ Moderate	○ Severe
Traction Bronchiectasis:	● No	○ Mild	○ Moderate	○ Severe
Bronchiectasis (excluding traction):	● No	○ Mild	○ Moderate	○ Severe

Airway and Vascular Distortion Distribution and Pattern

Cranial-caudal distribution:	● No	○ Upper	○ Lower
Axial distribution:	● No	○ Central	○ Peripheral
Pattern:	○ Focal	○ Diffuse	

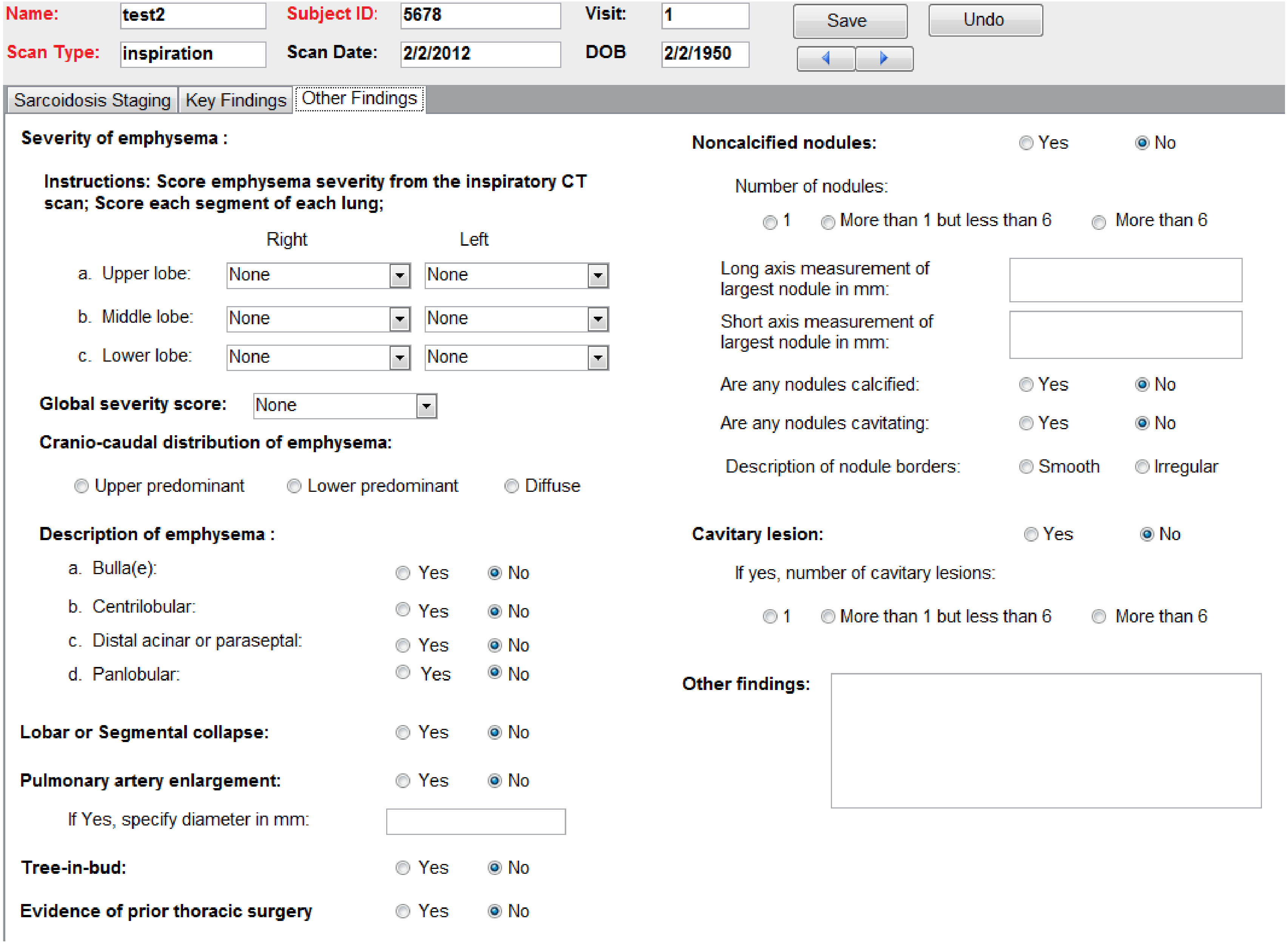
Parenchyma Opacity and Distortion

Ground glass:	● No	○ Mild	○ Moderate	○ Severe
Honeycombing:	● No	○ Mild	○ Moderate	○ Severe
Reticular abnormality:	● No	○ Mild	○ Moderate	○ Severe
Cystic changes:	● No	○ Mild	○ Moderate	○ Severe
Consolidation:	● No	○ Mild	○ Moderate	○ Severe
Mosaic attenuation:	● No	○ Mild	○ Moderate	○ Severe
Interlobular septal thickening:	● No	○ Mild	○ Moderate	○ Severe

Pleural effusion:	● No	○ Mild	○ Moderate	○ Severe
Pleural thickening:	● No	○ Mild	○ Moderate	○ Severe
Pleural calcification:	● No	○ Mild	○ Moderate	○ Severe
UIP fibrosis:	● No	○ Mild	○ Moderate	○ Severe
Mycetoma:	● No	○ Yes		
Interstitial Pneumonia:	● No	○ yes		

Parenchyma Opacity and Distortion Distribution and Pattern

Cranial-caudal distribution:	● No	○ Upper	○ Lower
Axial distribution:	● No	○ Central	○ Peripheral
Pattern:	○ Focal	○ Diffuse	



Name: test2 | Subject ID: 5678 | Visit: 1

Scan Type: inspiration | Scan Date: 2/2/2012 | DOB 2/2/1950

Severity of emphysema :

Instructions: Score emphysema severity from the inspiratory CT scan; Score each segment of each lung;

	Right	Left
a. Upper lobe:	None	None
b. Middle lobe:	None	None
c. Lower lobe:	None	None

Global severity score: None

Cranio-caudal distribution of emphysema:

○ Upper predominant ○ Lower predominant ○ Diffuse

Description of emphysema :

a. Bulla(e): ○ Yes ● No

b. Centrilobular: ○ Yes ● No

c. Distal acinar or paraseptal: ○ Yes ● No

d. Panlobular: ○ Yes ● No

Lobar or Segmental collapse: ○ Yes ● No

Pulmonary artery enlargement: ○ Yes ● No

If Yes, specify diameter in mm:

Tree-in-bud: ○ Yes ● No

Evidence of prior thoracic surgery ○ Yes ● No

Noncalcified nodules: ○ Yes ● No

Number of nodules:

○ 1 ○ More than 1 but less than 6 ○ More than 6

Long axis measurement of largest nodule in mm:

Short axis measurement of largest nodule in mm:

Are any nodules calcified: ○ Yes ● No

Are any nodules cavitating: ○ Yes ● No

Description of nodule borders: ○ Smooth ○ Irregular

Cavitary lesion: ○ Yes ● No

If yes, number of cavitary lesions:

○ 1 ○ More than 1 but less than 6 ○ More than 6

Other findings:

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. Hankinson JL, Odencrantz JR, Fedan KB. Spirometric reference values from a sample of the general U.S. population. Am J Respir Crit Care Med. 1999;159(1):179-87.

7. Zaiβ AW, Matthys H. A multiuser system for whole body plethysmographic measurements and interpretation. Lung. 1990;168(1):1185-92.

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

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

10. 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.

11. 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.

12. 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.

13. 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.