A Genomic Meta-Analysis of Clinical Variables and Their Association with Intrinsic Molecular Subsets in Systemic Sclerosis

Jennifer M. Franks[1,2&], Diana M. Toledo[2], Viktor Martyanov[1,2#], Yue Wang[1,2], Suiyan Huang[4], Tammara A. Wood[1,2], Cathie Spino[4], Christopher P. Denton[7], Emma Derret-Smith[7], Jessica K. Gordon, Robert Spiera, Robyn Domsic[5], Monique Hinchcliff[6], Dinesh Khanna[3,4], Michael L. Whitfield[1,2]

[1] Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756, USA
[2] Department of Molecular and Systems Biology, Geisel School of Medicine at Dartmouth, Lebanon NH 03756, USA
[3] Division of Rheumatology, Department of Medicine, University of Michigan, Ann Arbor, MI, USA.
[4] University of Michigan Scleroderma Program, Ann Arbor MI 48105
[5] University of Pittsburgh, Pittsburgh PA
[6] Yale University, New Haven CT
[7]Division of Medicine, University College London, NW3 2PF, UK.
[&]Current Affiliation: Department of Genome Sciences, University of Washington, Seattle, WA
[#]Current Affiliation: Celdara Medical LLC, Lebanon, NH 03766

Keywords: Systemic sclerosis, Machine Learning, Gene Expression, meta-analysis, clinical associations

Financial Support: NIH (National Institute of Allergy and Infectious Diseases Clinical and Autoimmunity Center of Excellence grant 5-UM1-AI-110557 to the University of Michigan for ASSET data and National Institute of Arthritis and Musculoskeletal and Skin Diseases grants K24-AR-063120 to Dr. Khanna.

Corresponding authors:
Michael L. Whitfield, Ph.D.
HB 7261, 1 Medical Center Drive
Lebanon, NH 03756
Phone number: (603)-650-1835
Fax number: (603)-650-1188
Email: michael.whitfield@dartmouth.edu

Dinesh Khanna, MD, MS
Division of Rheumatology, Department of Internal Medicine
University of Michigan
Ann Arbor, Michigan
Tel.: 734-764-7606
Fax: 734-763-4151
E-mail: khannad@med.umich.edu

Abstract (< 250 words)

Introduction. Four intrinsic molecular subsets (Inflammatory, Fibroproliferative, Limited, Normal-like) have previously been identified in systemic sclerosis (SSc) and are characterized by unique gene expression signatures and pathways. The intrinsic subsets have been linked to improvement with specific therapies. Here, we investigated associations between baseline demographics and intrinsic molecular subsets in a meta-analysis of published datasets.

Methods. Publicly available gene expression data from skin biopsies of 311 SSc patients measured by DNA microarray were classified into the intrinsic molecular subsets. RNA-sequencing data from 84 participants from the ASSET trial were used as a validation cohort. Baseline clinical demographics and intrinsic molecular subsets were tested for statistically significant associations.

Results. Males were more likely to be classified in the fibroproliferative subset. SSc patients who identified as African-American/Black were 2.5x more likely to be classified as fibroproliferative compared to White/Caucasian patients. Patients sera positive for anti-RNA pol I and RNA pol III autoantibodies were enriched in the inflammatory subset, while Scl-70 was enriched in the fibroproliferative subset. Average Modified Rodnan skin score (mRSS) was statistically higher in the inflammatory and fibroproliferative subsets compared to normal-like. The average disease duration for inflammatory subset was less than fibroproliferative and normal-like intrinsic subsets.

Conclusions. We identified multiple statistically significant differences in baseline demographics between the intrinsic subsets which may represent underlying features of disease pathogenesis (e.g. chronological stages of fibrosis) and have implications for treatments that are more likely to work in certain SSc populations.

Introduction

Systemic sclerosis (SSc) is a deadly autoimmune disease of unknown etiology and complex clinical phenotype. It is characterized by skin fibrosis, internal organ dysfunction, vascular damage, and immunologic abnormalities. SSc clinical subtypes are defined according to the extent of skin involvement[1]. For patients with limited cutaneous SSc (lcSSc), skin fibrosis is restricted to the arms, legs, and face. In diffuse cutaneous SSc (dcSSc), skin fibrosis extends to include the torso and typically coincides with increased disease severity[2]. Previously, four intrinsic molecular subsets (inflammatory, fibroproliferative, limited, normal-like) have been defined in SSc, characterized by unique biological processes and gene expression signatures[3-5]. The molecular subtypes have been demonstrated across multiple tissues[6,7] and validated in multiple studies [3,4,8-10] demonstrating the systemic nature of the disease.

Intrinsic subset is consistent across different skin biopsy sites within a single patient, regardless of clinically affected or unaffected status [9]. The inflammatory subset is defined by up-regulation of immune system processes including inflammatory, stress, and defense responses[4]. The fibroproliferative subset is characterized by increased expression of proliferative processes including cell cycle and mitosis. The normal-like subset is composed of samples from SSc patients, whose gene expression most closely resembles that of healthy controls, notably missing inflammatory and proliferative signatures[9,11]. The limited subset consists exclusively of patients with lcSSc and is the least molecularly characterized. Importantly, patients with lcSSc can also be assigned to the inflammatory and normal-like subsets. The intrinsic subsets are clinically meaningful and have been linked to improvement and long-term outcomes with different treatments[5,8,12,13].

Studies that first assigned intrinsic subsets in SSc used unsupervised, agglomerative methods to determine the number of intrinsic subsets and each sample's membership in a subset [3,4,8,14,15]. To classify patients in clinical trials or for diagnostic purposes, we previously developed a supervised machine learning classifier to assign individual samples to intrinsic molecular subsets. Our method uses a multinomial elastic net classifier for classification using objective molecular genomic data. Here, we extend the use of this method to classify all publicly available gene expression data from SSc skin samples.

Although overall survival and treatment strategies for SSc are improving, there are only two therapies currently approved for SSc treatment, nintedanib and tocilizumab, which are approved for treatment of SSc-associated interstitial lung disease. Although overall survival and treatment strategies for SSc are improving, SSc remains a challenge to treat and patient stratification in SSc could increase the possibility of success[16]. Often, the statistical power in these clinical trials is compromised by the extreme clinical and molecular heterogeneity, which we address in this study. The use of genomic data and intrinsic subsets may help improve patient outcomes by identifying therapies with higher potential for success in each individual patient. For example, the inflammatory intrinsic subset has been associated with response to immune-modulating therapies[8,12,17]. Identifying clinical variables that are associated with intrinsic subsets may allow clinical trials to refine inclusion criteria to decrease genetic heterogeneity in study cohorts. Ultimately, we hope this will increase power for clinical trials and lead to the identification of treatments that are effective in SSc. Additionally, longitudinal tracking of intrinsic subset assignment may provide insight into SSc pathogenesis and overall disease trajectory.

Most published genomic studies are limited in sample size and therefore underpowered to detect associations, though some clinical associations with SSc intrinsic subsets have been reported. To directly address this issue, we undertook a genomic meta-analysis of intrinsic subsets in SSc. By aggregating multiple genomic studies, we greatly increase the statistical power to detect novel associations. The overall goal of the study is to identify clinical covariates associated with SSc intrinsic subsets which may provide important insight into disease treatment or pathogenesis.

Methods

DNA microarray data preprocessing.

Raw gene expression data for each study (Table 1) were downloaded from NCBI GEO and processed using the following pipeline. Each dataset was processed independently. GenePattern[18] was used to impute missing values using k-nearest neighbors imputation with default settings. Probes were collapsed to genes by the maximum expression value using the appropriate annotation file for each dataset and platform. Genes were median centered across arrays within the dataset. Samples were classified using GLMnet as previously described[19]. Due to substantial differences in data distributions between Affymetrix data and the training data for GLMnet, feature specific quantile normalization (FSQN) was performed prior to classification as previously described[19-21].

RNA-seq data preprocessing.

RNA-sequencing was performed on skin biopsies from 84 participants in the ASSET (Abatacept Systemic SclErosis Trial) trial. Normalized RPKM values were classified into the intrinsic gene expression subsets using FSQN and a support vector machine[22]. Normalized RPKM values were classified into intrinsic gene expression subsets. Gene expression from forearm biopsies at baseline was used for classification, with the exception of one patient whose baseline gene expression sample failed quality control metrics and the three-month forearm sample was used instead.

Clinical data processing.

Age was coded in years. Disease duration was coded in months. Sex was classified as Male, Female or Unknown/Not Reported. Race was coded as follows: White (identifying as Caucasian or White), Black (identifying as Black, African-American, or African), Asian (identifying as Asian or Southeast Asian), Other (American Indian, Alaska Native or other), or Unknown/Not Reported. Ethnicity was coded independently from race as Hispanic (identifying as Hispanic or Latinx), Non-Hispanic, or Unknown/Not Reported. Pulmonary function tests were used in the meta-analysis only if they were reported as % FVC and % DLCO. Autoantibodies were coded as individual tests positive/negative/missing for Scl-70, anti-RNA polymerase III, or CENPB in the meta-analysis. Additional information for anti-RNA polymerase I was also available for patients in the ASSET cohort.

Statistical analyses.

We did not impute any missing values in clinical data. Associations between baseline clinical demographics and intrinsic molecular subsets were tested pairwise using Fisher's Exact Test or Chi-squared tests for categorical variables and Wilcoxon Rank Sum test was used for continuous variables. For comparing intrinsic subsets, ANOVA tests with Tukey's correction for multiple hypotheses were used in pairwise comparisons of continuous variables. P-values less that 0.05 were considered significant.

Results

We first identified all publicly available DNA microarray gene expression datasets generated from SSc skin to form a discovery cohort for clinical features associated with intrinsic subset assignment. Genomic studies were excluded if (1) there was no published individual-level patient clinical information, (2) if we were unable to obtain any clinical information about the study participants from the investigators, or (3) there were fewer than five individuals in the study. Following these criteria, we were able to include 13 genomic datasets in our investigation of clinical demographics associated with intrinsic subset (Table S1, S2). We then restricted our analyses to only include individuals with a diagnosis of SSc and classified as either lcSSc or dcSSc. SSc patients with morphea, sine scleroderma, and polymyositis overlap were excluded. For each SSc patient, only the baseline forearm sample (pre-treatment) was retained for further analysis. A baseline back/flank sample was used if there was no forearm sample. These criteria resulted in a study population of 311 SSc patients for our meta-analysis (Table 1). The majority of the SSc patients included in this analysis were white (44.37%), female (72.67%), and classified as dcSSc (74.60%). The average age of subjects in this study was 50.33 years with an average disease duration of 3.2 years (38.77 months).

Clinical demographics are associated with intrinsic subsets

In our study population, 311 patients with SSc were individually assigned to an intrinsic subset based on gene expression using a pre-trained classifier[19]. We tested corresponding baseline demographic data to identify clinical associations with the intrinsic subsets in a meta-analysis (Table 2, Figure 1) and validation cohort (Table S3). Of the 311 patients with SSc, 117 (37.6%) were assigned to the inflammatory subset, 105 (33.7%) were assigned to the fibroproliferative subset, 84 (27%) were assigned to the normal-like subset, and 5 (1.6%) were assigned to the limited subset. Of the 84 participants from the ASSET study used as validation, 33 (39%) were assigned to the inflammatory subset, 18 (21.4%) were assigned to fibroproliferative subset and 33 (39%) were assigned to the normal-like subset. No limited patients were included in the ASSET study.

As noted above, there were many more females than males in our study population (Figure 1A) and there were significant differences in the distribution of intrinsic subsets between the sexes (p=0.030, Fisher's Exact Test). Males were 2.41 times more likely to be fibroproliferative than females (p=0.0046, Fisher's Exact Test). Females and males were equally likely to be classified as inflammatory, normal-like, or limited. There was also a significant association with gender in the ASSET cohort (p=0.015, Fisher's Exact). Analysis of this trend showed that males were 3.99 times more likely to be classified as fibroproliferative (p=0.015, Fisher's Exact Test). Like the discovery cohort, males and females were equally likely to be assigned to the inflammatory and normal-like subsets (Figure 1C).

We identified a statistically significant difference in the average ages between intrinsic subsets (p=0.0041, ANOVA). The average age of patients was 52.92 years in the inflammatory subset, 47.41 years in the fibroproliferative subset, 49.50 years in the normal-like subset, and 60.6 years in the limited subset (Fig. 1B). Overall, fibroproliferative patients were significantly younger than inflammatory patients (p=0.011, Tukey's HSD, n=259), but no other pairwise comparisons of age were statistically significant. In the ASSET cohort, the average age of inflammatory patients (53.21 years) compared to the average age of fibroproliferative patients (46.56 years, (p=0.203, ANOVA) (Figure 1D)) demonstrates a larger absolute difference of means that in the discovery cohort, suggesting that smaller sample size in the ASSET cohort is likely responsible for the lack of statistical significance.

We investigated the distribution of intrinsic subsets within and between self-reported races (n=168). Although there were many more patients who identified as White/Caucasian in our study (n=138), there was also a sizeable number of patients who identified as African-American/Black (n=25). Five patients identified as Asian, and one patient identified as White and Asian. 17 patients identified as Hispanic or Latinx, and 64 patients identified as Non-Hispanic or Non-Latinx. In some studies, race and ethnicity were coded together and kept separate in other studies. We did not infer race for patients who identified as Hispanic or Latino, nor did we infer ethnicity for patients who did not explicitly report it. Thus, information on race was missing for 144 patients and information on ethnicity was missing for 230 patients. There was no significant relationship between ethnicity and intrinsic subset assignment for either the meta-analysis or the validation cohort (p=0.38, p=0.91 respectively, Fisher's Exact Test).

From a global analysis, there was no significant association when considering all races and all intrinsic subsets in the meta-analysis (p=0.11) although we did find a significant association with race in the ASSET cohort (p=0.0033, Fisher's Exact). The lack of a significant difference in the meta-analysis could be due to vastly different sample sizes between races, because notable trends were conserved in both study populations. There was an evident relationship between African-American/Black patients and the fibroproliferative subset. Forty-one out of 138 (29.71%) patients who identified as White/Caucasian were classified as fibroproliferative, whereas 13 out of 25 (52%) of patients identifying as African-American/Black were classified as fibroproliferative. Compared to White/Caucasian SSc patients, African-American/Black patients were 2.5 times more likely to be classified as fibroproliferative (p=0.0378, Fisher's Exact Test).

In the ASSET cohort, patients with SSc who identified as African-American or Black were also much more likely to be classified as fibroproliferative compared to other SSc patients (p=0.0062, Fisher's Exact). Patients with SSc who identified as White or Caucasian are more likely to be classified as inflammatory or normal-like (p=0.0037, Fisher's Exact Test).

Association with Autoantibodies

There was a substantial amount of missing clinical information regarding autoantibodies for patients with lcSSc in our cohort, and it has previously been reported that there is an association of autoantibodies with clinical subtype, so we restricted autoantibody analyses only to patients with dcSSc. A substantial number of inflammatory patients (29 out of 61) tested positive for anti-RNA polymerase III autoantibodies. Similarly, a large number of normal-like patients (9 out of 22) tested positive for anti-RNA polymerase III. Most fibroproliferative patients (18 out of 37) tested positive for Scl-70 autoantibodies. Although these results are interesting, the autoantibody analyses did not reach statistical significance in the validation cohort (p=0.24, Fisher's Exact).

Autoantibodies in the ASSET clinical trial were measured at a single center in a consistent manner providing a complete, well-measured dataset. We tested for and found significant associations of autoantibodies with the intrinsic subsets (p=0.0465, Fishers Exact, n=81). In the validation cohort, there was a statistically significant difference in anti-RNA Polymerase III between the intrinsic subsets (p=9.25E-5, Fisher's Exact). 24 out of 33 (72.7%) inflammatory patients tested positive for anti-RNA polymerase III in comparison to only 5 out of 31 (16.2%) of normal-like patients, and 4 out of 17 (25.5%) of fibroproliferative patients. The ASSET clinical trial also tested for anti-RNA polymerase I autoantibodies and there was a statistically significant difference between the intrinsic subsets. (p=5.81E-5, Fisher's Exact). Like anti-RNA polymerase III, the inflammatory subset was much more likely to be positive for RNA polymerase I (22 out of 33) compared to the normal-like (5 out of 31) and fibroproliferative (4 out of 17) subsets. There were no statistically significant differences in anti-Scl-70 (p=0.111, Fisher's Exact) or anti-centromere (p=0.6043, Fisher's Exact) between the intrinsic subsets. Of the patients who were anti-Scl-70 positive in the discovery cohort, we found that 16 were inflammatory, 18 were fibroproliferative, and 6 were normal-like (Table 3). In the ASSET validation set, 3 were inflammatory, 5 were fibroproliferative, and 8 were normal-like.

Measures of SSc severity and correlation to intrinsic subsets

We tested measures of SSc severity between intrinsic subsets in patients with dcSSc in the meta-analysis (Table 3, Figure 2) and the validation cohort (Table S4). Modified Rodnan Skin Score (mRSS) is a standard outcome measure for skin involvement, calculated by assessing skin thickness (scored 0-3) across 17 body sites. There was a statistically significant difference in average mRSS between intrinsic subsets (Fig. 2A) (p=0.0027, ANOVA). Normal-like patients exhibited the lowest average mRSS (19.29), and this was significantly lower than the average for inflammatory patients (p=0.0017, Tukey's HSD) and fibroproliferative patients (p=0.047, Tukey's HSD). There was no difference between the average mRSS for inflammatory (average=24.67) and fibroproliferative (average=23.13) patients (p=.48, Tukey's HSD).

As seen in the meta-analysis, the inflammatory and fibroproliferative subsets in the ASSET cohort also showed significantly higher average mRSS than the normal-like subset (respectively: p=1.53E-5, p=0.0060, ANOVA Tukey's HSD) confirming these results (Figure 3A).

Forced vital capacity (FVC) and diffusing capacity of the lungs for carbon monoxide (DLCO) are two standard measures of lung involvement in SSc (Fig. 2B-C). Lower values indicate more severe disease activity. There were no statistically significant differences between intrinsic subsets for DLCO % Predicted (p=0.49, ANOVA) or FVC % Predicted (p=0.067, ANOVA) in the meta-analysis. Despite not reaching statistical significance, we observed a consistent trend of increased lung function (average DLCO and FVC, respectively) in inflammatory patients (66.09, 81.04) and slightly reduced lung function in normal-like (59.58, 71.9) and fibroproliferative patients (63.18, 73.0). A similar trend was observed for FVC in the ASSET cohort but also did not reach statistical significance (p=0.229, ANOVA) between the intrinsic subsets (Figure 3B). There were no statistically significant differences for DLCO (% corrected) in the ASSET cohort (p=0.135, ANOVA) (Figure 3C).

Evidence of a temporal relationship between intrinsic subsets

Next, we investigated the temporal spacing between the intrinsic subsets by quantifying disease duration in months from first non-Raynaud's symptom. Patients with lcSSc exhibited longer average disease duration than patients with dcSSc (p=1.55E-4, Wilcoxon Rank Sum). In order to reduce confounding by clinical subtype, we restricted the analysis of disease duration to only patients with dcSSc. We identified a statistically significant difference in average disease duration between the inflammatory, fibroproliferative, and normal-like intrinsic subsets (p=8.8E-4, ANOVA) (Figure 2D). Patients in the inflammatory subset had average disease duration of 14.85 months. This was statistically lower than both the fibroproliferative subset (p=0.0073, Tukey's HSD) and the normal-like subset (p=0.0042, Tukey's HSD). The average disease duration for the fibroproliferative subset was 35.59 months which was shorter than 41.39 months for the normal-like subset, but this did not reach statistical significance (p=0.78, Tukey's HSD). This difference in temporal distributions may reflect chronological stages of fibrosis.

There was not a significant difference in disease duration (p=0.416, ANOVA) in the ASSET cohort, as expected due to the recruitment criteria (disease duration no more than 36 months) of the clinical trial (Figure 3D).

Discussion

In this study, we performed the first large-scale genomic meta-analysis aimed at elucidating the clinical covariates associated with intrinsic subsets defined by gene expression in SSc. We utilized a powerful machine learning classification system in order to aggregate clinical data and summarize genomic information from multiple studies that were performed over time, on several different platforms, and in multiple independent laboratories. We analyzed clinical data from 311 individuals with SSc and found significant and important associations with intrinsic subset assignments. Combined with the 84 individuals in the validation cohort, we analyzed data for a total of 395 patients with SSc.

Our primary results indicate that fibroproliferative patients may be younger and more likely to be male. Individuals of Black and African American ancestry were more likely to fall into this group and may have a slight increase in anti-Scl-70 antibodies. Patients in the inflammatory subset were more likely to be older, female, Caucasian, and sera-positive for anti-RNA pol I and III positive. These results show an enrichment for certain demographics within the intrinsic subsets but that stratification by autoantibodies, gender, or race alone would not be sufficient to predict an individual's molecular subset since they are distributed across groups. Some findings were not significant in both cohorts likely because of the difference in distribution of clinical subtypes and sample size between the discovery and validation populations.

Notably, we identified differences in disease severity between the intrinsic subsets. Inflammatory and fibroproliferative patients are more likely to have higher skin score, measured by mRSS, compared to normal-like patients. Patients in the fibroproliferative subset may have decreased lung function, a phenotype that has previously been noted[23]. This finding is particularly important given increased prevalence of African-American/Black patients with SSc belonging to the fibroproliferative subset. This is the first study to find a significant association of race and intrinsic subset. It has previously been noted that African-American/Black SSc patients have been linked to decreased lung function[24,25] and increased TGFβ gene expression signatures[23]. This result further establishes a plausible link between genomic signatures and phenotypic outcomes in a particular population. These findings may have clinical implications for identifying treatments more likely to work in this population, such as stem cell transplantation[26].

It has previously been suggested that normal-like patients may represent later stage disease[27], and our study supports a temporal relationship between the intrinsic subsets. These data support the inflammatory subset as earlier disease, and fibroproliferative as having an intermediate disease duration. Based on available longitudinal data, we believe the inflammatory and fibroproliferative subsets do not readily interconvert. The normal-like subset may represent a later disease stage in which the early inflammatory and fibroproliferative stages have previously burned out. This model makes biological sense because previous studies have been unsuccessful in capturing patients' changing subset over time[4], except in the context of treatment[12], and then typically only toward normal-like. Notably, in studies such as ASSET which only enrolled early dcSSc patients, all of the subsets are represented within the baseline biopsies despite the temporal relationship we can also find here.

We did not control for prior treatment in this study and that is a significant limitation to this analysis, however in the ASSET trial, all patients were not on background immunomodulatory therapy at the baseline visit. However, a major strength of this study is that only baseline samples are considered, and many of the samples were from “pre-treatment” individuals in clinical trials who experienced wash-out time prior to sample collection. Thus, we believe the results of this analysis may be indicative of natural disease history and supportive of an immune-fibrotic axis in SSc[7,9].

In conclusion, by leveraging data from multiple studies, we increased statistical power and identified multiple novel associations between clinical variables and intrinsic subsets in SSc. These associations may explain aspects of SSc pathogenesis and probe interesting biological questions, such as how fibroproliferative process impact lung function and how this manifests in certain populations. Finally, the results from this study provide additional clinical context for the intrinsic subsets. This may help future clinical trials refine inclusion criteria to reduce molecular heterogeneity and facilitate the identification of effective treatments for subgroups of patients with SSc.

Table 1: Clinical demographics of the overall discovery and validation study populations. *One patient in the validation population identified as both White and Black and one patient identified as both White and Asian.

	SSc Patients (n=311)	ASSET cohort (n=84)
Age (years) – Mean (SD)	50.33 (12.30)	50.92 (12.70)
Sex		
Female – no. (%)	226 (72.67)	62 (73.81)
Male – no. (%)	58 (18.65)	22 (26.19)
Race		
White – no. (%)	138 (44.37)	71 (84.52*)
Black – no. (%)	25 (8.04)	7 (8.33*)
Asian – no. (%)	5 (1.61)	6 (7.14*)
Other/Unknown/Not reported – no. (%)	143 (45.98)	2 (2.38)
Ethnicity		
Hispanic – no. (%)	17 (5.45)	10 (11.90)
Non-Hispanic – no. (%)	64 (20.58)	73 (86.90)
Clinical Subtype		
Limited – no. (%)	79 (25.40)	0 (0.0)
Diffuse – no. (%)	232 (74.60)	84 (100.0)
Disease Duration (months) – Mean (SD)	38.77 (63.45)	18.46 (10.39)
mRSS – Mean (SD)	20.02 (10.59)	22.18 (7.50)
FVC % Predicted – Mean (SD)	79.82 (19.44)	84.89 (14.93)
DLCO % Predicted – Mean (SD)	64.24 (20.11)	77.60 (18.42)
Autoantibodies		
Scl70 – no. (%)	55 (17.68)	16 (19.05)
Anti-RNA polymerase III – no. (%)	53 (17.04)	36 (42.86)
Anti-centromere – no. (%)	17 (5.45)	3 (3.57)

Table 2. Demographics across the discovery cohort SSc intrinsic subsets

SSc patients (n=311)	**Inflammatory** (n=117)	**Fibroproliferative** (n=105)	**Normal-like** (n=85)	**Limited** (n=5)
Age – Mean (SD)	52.92 (11.22)	47.41 (11.06)	49.50 (12.46)	60.6 (4.62)
Sex – no.				
Female	88	66	67	5
Male	16	29	13	0
Race – no.				
White	58	41	35	4
Black	7	13	4	1
Asian	3	2	0	0
Other/Unknown	49	49	46	0
Ethnicity – no.				
Hispanic or Latino	4	6	7	0
Non-Hispanic or Non-Latino	27	18	19	0
Unknown	86	81	58	5
Clinical Subtype – no.				
Diffuse	103	79	50	0
Limited	14	26	34	5

Table 3. Phenotypic Measures across the discovery cohort of dcSSc intrinsic subsets

dcSSc patients (n=232)	**Inflammatory** (n=103)	**Fibroproliferative** (n=79)	**Normal-like** (n=50)
Disease Duration (months) – Mean (SD)	14.85 (15.29)	35.59 (51.16)	41.393 (45.50)
mRSS – Mean (SD)	24.67 (9.47)	23.13 (8.12)	19.29 (8.78)
FVC % Predicted – Mean (SD)	81.04 (17.89)	73.00 (20.50)	71.79 (14.15)
DLCO % Predicted – Mean (SD)	66.09 (19.29)	59.58 (20.03)	63.18 (22.15)
Autoantibodies – no.			
anti Scl-70	16	18	6
Anti-RNA polymerase III	29	12	9
Anti-centromere	4	1	1

Figure 1: Clinical demographics of gender and age distribution stratified by intrinsic subsets in the SSc meta-analysis (A, B) and validation cohorts (C, D), respectively.

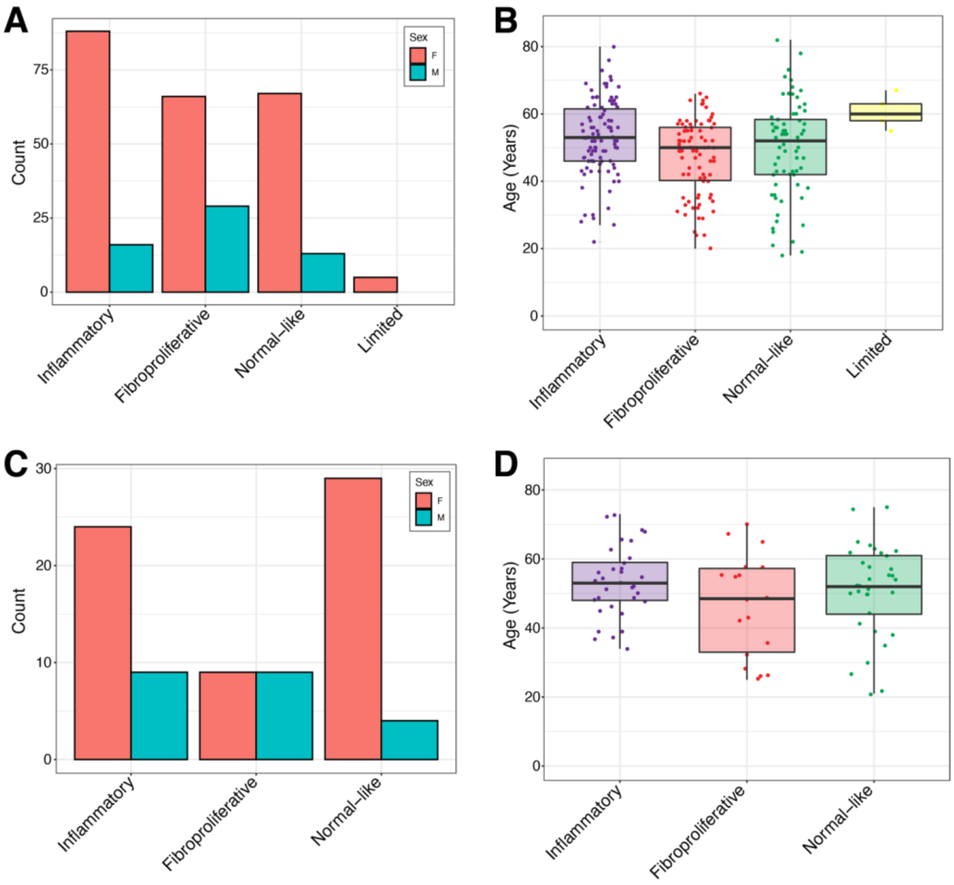


Figure 2: Measures of phenotypic severity in dcSSc patients of the discovery cohort stratified by intrinsic subset (A) mRSS, (B) FVC, (C) DLCO, (D) Disease duration.

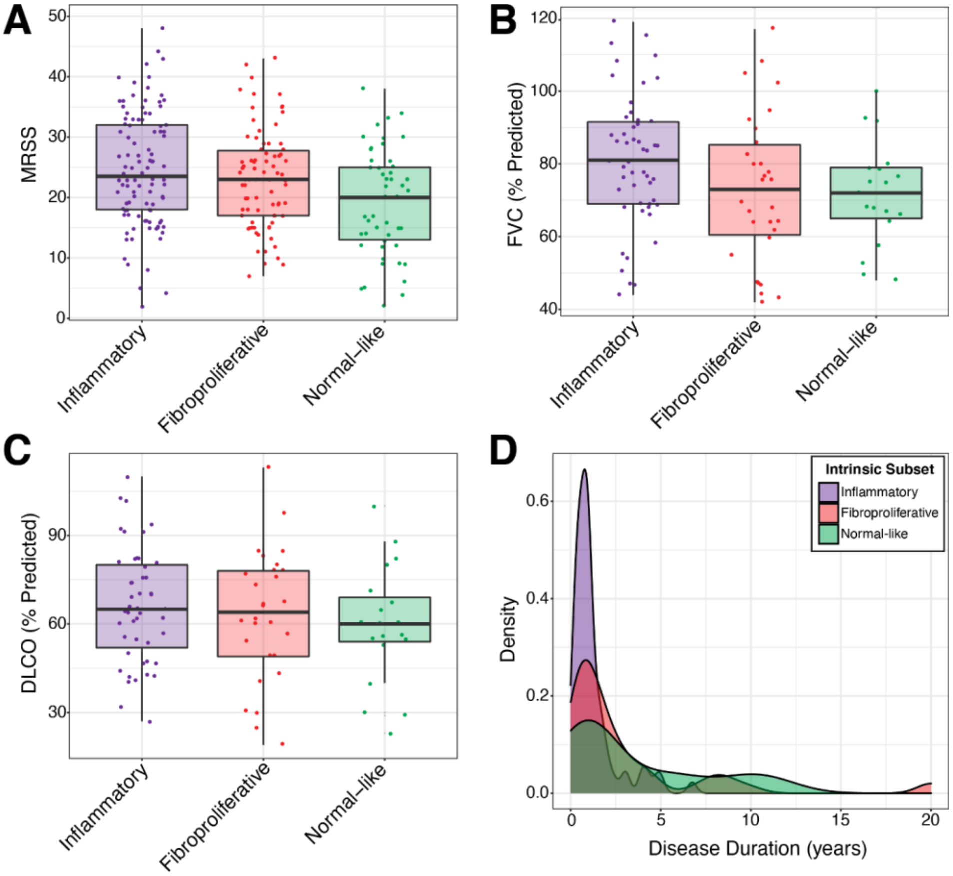
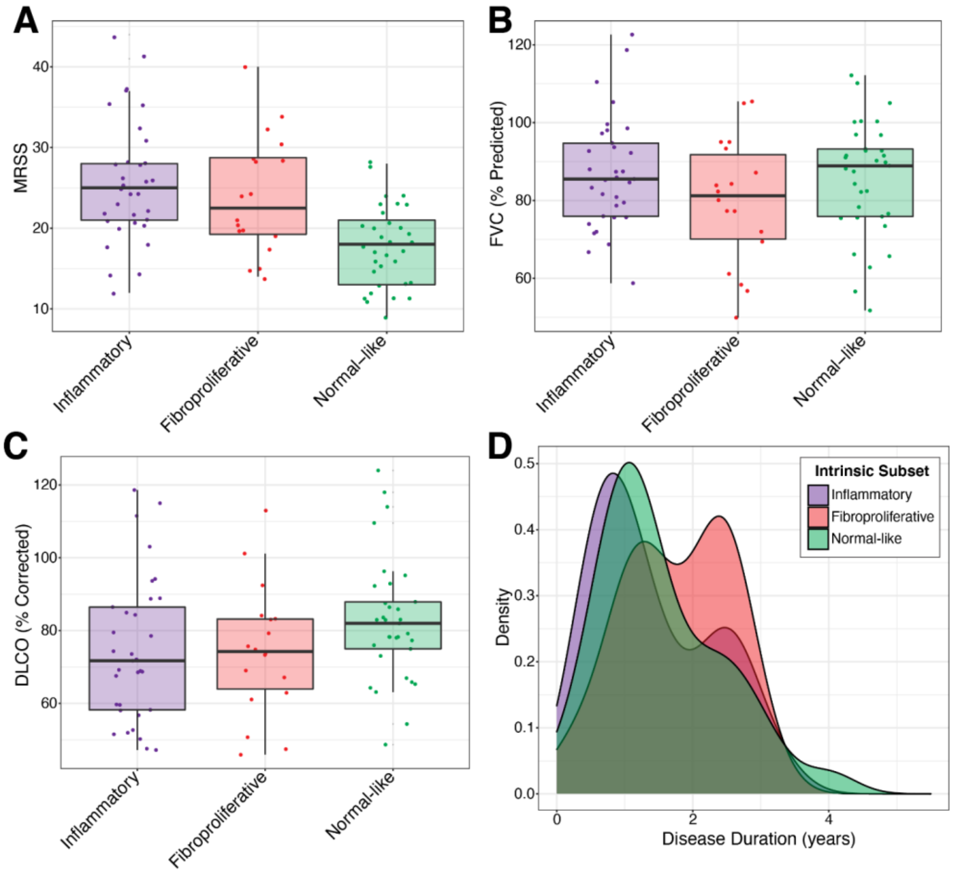


Figure 3: Measures of phenotypic severity in dcSSc patients of the validation cohort (ASSET) stratified by intrinsic subset (A) mRSS, (B) FVC, (C) Disease duration.



References

1. LeRoy EC, Black C, Fleischmajer R, et al. Scleroderma (systemic sclerosis): classification, subsets and pathogenesis. J Rheumatol 1988;15:202-5.
2. Varga J, Denton CP, Wigley FM, Alanore Y, Kuwana M. Scleroderma : from pathogenesis to comprehensive management. Second edition. ed. Cham: Springer; 2017.
3. Milano A, Pendergrass SA, Sargent JL, et al. Molecular subsets in the gene expression signatures of scleroderma skin. PLoS ONE 2008;3:e2696.
4. Pendergrass SA, Lemaire R, Francis IP, Mahoney JM, Lafyatis R, Whitfield ML. Intrinsic gene expression subsets of diffuse cutaneous systemic sclerosis are stable in serial skin biopsies. J Invest Dermatol 2012;132:1363-73.
5. Hinchcliff M, Toledo DM, Taroni JN, et al. Mycophenolate Mofetil Treatment of Systemic Sclerosis Reduces Myeloid Cell Numbers and Attenuates the Inflammatory Gene Signature in Skin. J Invest Dermatol 2018;138:1301-10.
6. Taroni JN, Martyanov V, Huang CC, et al. Molecular characterization of systemic sclerosis esophageal pathology identifies inflammatory and proliferative signatures. Arthritis Res Ther 2015;17:194.
7. Taroni JN, Greene CS, Martyanov V, et al. A novel multi-network approach reveals tissue-specific cellular modulators of fibrosis in systemic sclerosis. Genome Med 2017;9:27.
8. Hinchcliff M, Huang CC, Wood TA, et al. Molecular signatures in skin associated with clinical improvement during mycophenolate treatment in systemic sclerosis. J Invest Dermatol 2013;133:1979-89.
9. Mahoney JM, Taroni J, Martyanov V, et al. Systems level analysis of systemic sclerosis shows a network of immune and profibrotic pathways connected with genetic polymorphisms. PLoS computational biology 2015;11:e1004005.
10. Whitfield ML, Finlay DR, Murray JI, et al. Systemic and cell type-specific gene expression patterns in scleroderma skin. Proc Natl Acad Sci U S A 2003;100:12319-24.
11. Johnson ME, Mahoney JM, Taroni J, et al. Experimentally-derived fibroblast gene signatures identify molecular pathways associated with distinct subsets of systemic sclerosis patients in three independent cohorts. PLoS One 2015;10:e0114017.
12. Gordon JK, Martyanov V, Franks JM, et al. Belimumab for the Treatment of Early Diffuse Systemic Sclerosis: Results of a Randomized, Double-Blind, Placebo-Controlled, Pilot Trial. Arthritis & rheumatology 2018;70:308-16.
13. Gordon JK, Martyanov V, Magro C, et al. Nilotinib (Tasigna) in the treatment of early diffuse systemic sclerosis: an open-label, pilot clinical trial. Arthritis Res Ther 2015;17:213.
14. Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 2001;98:10869-74.
15. Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. Nature 2000;406:747-52.
16. Denton CP, Khanna D. Systemic sclerosis. Lancet 2017;390:1685-99.
17. Chakravarty EF, Martyanov V, Fiorentino D, et al. Gene expression changes reflect clinical response in a placebo-controlled randomized trial of abatacept in patients with diffuse cutaneous systemic sclerosis. Arthritis Res Ther 2015;17:159.
18. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nat Genet 2006;38:500-1.

19. Franks JM, Martyanov V, Cai G, et al. A Machine Learning Classifier for Assigning Individual Patients with Systemic Sclerosis to Intrinsic Molecular Subsets. Arthritis & rheumatology 2019.

20. Franks JM, Cai G, Whitfield ML. Feature Specific Quantile Normalization Enables Cross-Platform Classification of Molecular Subtypes using Gene Expression Data. Bioinformatics 2018.

21. Khanna D, Spino C, Johnson S, et al. Abatacept in Early Diffuse Cutaneous Systemic Sclerosis: Results of a Phase II Investigator-Initiated, Multicenter, Double-Blind, Randomized, Placebo-Controlled Trial. Arthritis & rheumatology 2020;72:125-36.

22. Franks JM, V.; Cai, G.; Wang, Y.; Wood T.A.; Whitfield, M.L. A Machine Learning Classifier for Assigning Patients with Systemic Sclerosis to Intrinsic Molecular Subsets. Submitted.

23. Sargent JL, Milano A, Bhattacharyya S, et al. A TGFbeta-responsive gene signature is associated with a subset of diffuse scleroderma with increased disease severity. J Invest Dermatol 2010;130:694-705.

24. Blanco I, Mathai S, Shafiq M, et al. Severity of systemic sclerosis-associated pulmonary arterial hypertension in African Americans. Medicine (Baltimore) 2014;93:177-85.

25. Morgan ND, Shah AA, Mayes MD, et al. Clinical and serological features of systemic sclerosis in a multicenter African American cohort: Analysis of the genome research in African American scleroderma patients clinical database. Medicine (Baltimore) 2017;96:e8980.

26. Franks JM, Martyanov V, Wang Y, et al. Machine learning predicts stem cell transplant response in severe scleroderma. Ann Rheum Dis 2020;79:1608-15.

27. Assassi S, Swindell WR, Wu M, et al. Dissecting the heterogeneity of skin gene expression patterns in systemic sclerosis. Arthritis & rheumatology 2015;67:3016-26.