A novel network approach reveals tissue-specific cellular modulators of the immune-fibrotic axis in systemic sclerosis

Jaclyn N. Taroni[1], J. Matthew Mahoney[2,§], Casey S. Greene[3], Viktor Martyanov[1], Tammara A. Wood[1], Romy B. Christmann[4], Harrison W. Farber[5], Robert A. Lafyatis[6], Christopher P. Denton[7], Monique E. Hinchcliff[8], Patricia A. Pioli[9], Michael L. Whitfield[1, §]

Affiliations:
[1] Department of Genetics, Geisel School of Medicine at Dartmouth, Hanover, NH 03755 USA
[2] Department of Neurological Sciences, University of Vermont, Burlington, VT 05405 USA
[3] Department of Systems Pharmacology & Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104 USA
[4] Division of Rheumatology [5] Pulmonary Center, Department of Medicine, Boston University School of Medicine, Boston, MA, 02118, USA
[6] Division of Rheumatology and Clinical Immunology, Department of Medicine, University of Pittsburgh Medical Center, Pittsburgh. PA, 15261 USA
[7]Division of Medicine, University College London, London, UK
[8] Division of Rheumatology, Department of Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL, 60611, USA
[9] Department of Obstetrics and Gynecology, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756 USA

Running Title: Multi-tissue functional genomic networks of SSc

[§]To whom correspondence should be addressed:

Michael L. Whitfield, Ph.D.
Department of Genetics
Geisel School of Medicine at Dartmouth
7400 Remsen
Hanover, NH 03755
michael.whitfield@dartmouth.edu

J. Matthew Mahoney, Ph.D.
Department of Neurological Sciences
University of Vermont College of Medicine
HSRF 426
149 Beaumont Avenue
Burlington, VT 05405
John.M.Mahoney@uvm.edu

Abstract

Systems biology uses genome-scale data to determine the molecular mechanisms behind normal and disease states. We developed a novel computational method, which allowed for the first time, identification of a disease-associated signature across multiple tissues and organs in systemic sclerosis (SSc), a rare and sometimes fatal autoimmune disease. We find a common immune-fibrotic axis associated with the most severe disease phenotypes, including pulmonary fibrosis and pulmonary arterial hypertension. We evaluated disease-associated gene-gene interactions in the context of tissue-specific functional genomic networks and utilized differential network analysis to gain important insights into the role of pro-fibrotic macrophages in SSc. Using a novel cell type-aware multi-network approach, we find that genes modulated by immunosuppressive treatment occupy privileged positions in the skin-specific network. In total, this study not only presents a set of putative therapeutic targets for SSc, but a framework for multi-tissue functional genomic studies of complex human disease.

60 **Introduction**

61 Recent advances in integrative genomics have yielded powerful tissue- and cell type- specific
62 functional networks which model the interaction of genes in these specialized contexts or
63 'microenvironments' (1). These tools hold promise for understanding how genes may contribute to
64 human diseases (2) that arise, in part, out of an aberrant interplay of cell types and tissues. Network
65 biology has played a crucial role in our understanding of complex human diseases such as cancer
66 (3,4), and more recently, in disorders where the interactions between multiple tissues are dysregulated
67 (e.g., alopecia areata) (5). Many network biology approaches rely on assumptions about the relative
68 homogeneity of samples under study or the analysis of non-specific network models, which are not
69 necessarily appropriate for the study of autoimmune diseases. Here, we investigate a debilitating and
70 deadly *systemic* autoimmune disease by analyzing multiple tissue-specific networks (1) in parallel and
71 applying a novel cell type-aware multi-network measure that has implications for immunosuppressive
72 treatment.

73 Analytical approaches that leverage biological 'big data' can be especially fruitful in rare and
74 heterogeneous diseases (6) in which the risk of mortality is significant and no approved treatments
75 exist. We performed an integrative, multi-tissue analysis for systemic sclerosis (SSc; scleroderma), a
76 disease for which all of these tenants are true. SSc is characterized by abnormal vasculature, adaptive
77 immune dysfunction (autoantibody production), and extracellular matrix (ECM) deposition in skin and
78 internal organs. The etiology of SSc is unknown, but it is of complex genetic risk (7) and may arise from
79 immune activation following a wide array of insults such as cancer (8), infection (9), or dysbiosis (10).
80 SSc is clinically heterogeneous with some patients experiencing rapidly progressive skin and internal
81 organ disease, while others have stable disease that is largely limited to skin. Furthermore, 'traditional'
82 biomarkers, such as autoantibodies, are not always indicative of clinical phenotype, disease
83 progression, or outcome. Thus, there is much to be gained from applying a data-driven analytical
84 approach to genome-scale SSc gene expression data and from doing so in a manner that
85 encompasses severe pathophenotypes that are not well understood at the molecular level such as
86 pulmonary fibrosis (PF) and pulmonary arterial hypertension (PAH).

87 Gene expression has been studied in multiple affected tissues from patients with SSc, including

88 skin (11-13), blood (14,15), esophagus (16), and lung (17,18). We identified a pathogenic signature—a
89 common 'immune-fibrotic process' (IFP)—that is present in all tissues analyzed and is correlated with
90 all severe pathophenotypes (i.e., PF, PAH). Coexpression module identification was used as a 'rough'
91 estimate of the biological signal within a tissue, which is measured with considerable noise due to small
92 sample sizes and in some cases lack of control samples. We then refined our estimate of that signal
93 through 'consensus clustering' and arrived at a 'disease consensus' gene set that comprised the IFP.
94 We then analyzed the SSc disease consensus genes in functional genomic networks (1) that encode
95 critical features of tissue microenvironments. We validated the networks by dissecting their features as
96 they relate to genetic risk and response to therapy.

97 The IFP implicates activated macrophages (MØs) as central to immune-related processes in all
98 solid organs studied. The plasticity of this cell type cannot be understated and is an active research
99 area of great interest (19-21). Using differential network analysis, we were able to distinguish skin-
100 specific and lung-specific gene-gene interactions highly relevant to the plasticity of MØs and to SSc
101 pathobiology. Because the lung data sets sampled distinct PF disease stages at steady state, we were
102 able to develop a model of PF progression that emphasizes the role of lipid-laden macrophages, or
103 foam cells, in early fibrotic disease.

104 Although prior work has examined which cell types may contribute to an inflammatory signature in
105 SSc skin ((12), most recently in (22)), to our knowledge, no other study has considered the interaction
106 of cell types in this tissue systematically. Cell types act in concert to form the structure and function of
107 their resident tissue. As a result, a tissue-specific network is an estimate of the gene-gene interactions
108 within the constituent cell types and of the interactions between cell types in the tissue. We sought to
109 characterize the cell type 'contributions' to the tissue-level network structure in skin and developed a
110 statistic we termed 'cell type centrality' to characterize biologically meaningful genes in a cell type-
111 aware manner. We validated this measure by demonstrating that genes that decrease during response
112 to mycophenolate mofetil (MMF) have high cell type centrality and, in particular, act as 'bridges'
113 between the T cell and MØ networks. Thus, MMF response genes have a 'network signature' that is
114 subtly related to the tissue context. This implicates cell type-specific hubs as a rich set of putative
115 targets for drug discovery, particularly in light of recent efforts to profile drug effects on immune cell

116 subsets (23).

117 In sum, we applied a data-driven approach to and developed novel computational methods for a
118 multi-tissue functional genomic study of SSc. We employed an integrative genomics approach to
119 genome-scale expression data from multiple tissues and we applied differential network analysis of
120 tissue-specific networks and a novel multi-network measure to reveal tissue-specific cellular modulators
121 of the immune-fibrotic axis in severe manifestations of this disease. We demonstrate the benefit of
122 utilizing functional genomic networks, built using data from thousands of experiments, to analyze whole
123 tissue biopsies from disease states that are challenging to study experimentally. Importantly, our
124 network results have experimental support from independent clinical trials, animal models, and *ex vivo*
125 studies. Thus, this study has implications not only for precision medicine in SSc, but can be seen as a
126 roadmap for the use of tissue- and cell type-specific networks to evaluate complex multi-tissue human
127 disease.

128

129 **Results**

130 We performed an integrative analysis of a compendium of SSc gene expression data containing
131 10 independent datasets from 4 tissues—skin, lung, esophagus and peripheral blood mononuclear
132 cells (PBMCs) (Table I). A total of 573 samples from 321 subjects were analyzed. These datasets
133 included lung biopsies from early SSc-associated pulmonary fibrosis (SSc-PF) (18) and late or end-
134 stage PF (17), PBMC (12,15) and lung samples from patients with SSc-associated and idiopathic PAH
135 (SSc-PAH and IPAH, respectively) (17). Skin biopsies from patients with limited (LSSc, UCL) and
136 diffuse cutaneous (11-13) SSc (lcSSc and dcSSc, respectively), which are two distinct clinical forms of
137 the disease, were also included (see Table SI for summary of patient cohorts). Our goal was to
138 understand the fundamental processes that occur in end-target and peripheral tissues of patients with
139 SSc and to identify the presence or absence of common expression patterns that underlie the
140 molecular subsets in esophagus and skin and other severe SSc-associated pathophenotypes. Analysis
141 of multiple tissue biopsies from patients with the same pathophenotype allowed us to explore whether
142 these tissues were perturbed in a similar manner on a 'genomic' scale.

143 We recently developed a methodology called Mutual Information Consensus Clustering (MICC) to
144 identify conserved coexpression across SSc skin datasets (24). Here, we have adapted MICC to
145 examine coexpression patterns across datasets from different tissues in a multi-tissue MICC analysis.
146 Briefly, datasets were partitioned into sets of highly correlated genes, or coexpression modules, using
147 weighted gene correlation network analysis (WGCNA) (Fig 1). These modules can be tested for
148 differential expression in pathophenotypes of interest, yielding gene sets associated with each
149 pathophenotype. However, because WGCNA is a clustering procedure, gene sets derived from
150 different data sets necessarily differ from each other. (This is expected due to both biological
151 phenomena, such as tissue type, and technical variation.) The MICC procedure allowed us to detect
152 imperfectly conserved coexpression between datasets in a formal manner. In brief, MICC determines
153 similarity between modules (from different data sets) using an information-theoretic criterion and
154 aggregates these similarities into an 'information graph' that encodes all module-module similarities
155 across all data sets studied. Specifically, two modules will share an edge in the information graph if
156 there is significant overlap between them (see Methods). Then, community detection is performed on

157 the information graph to identify groups of modules that are densely interconnected; these groups
158 correspond to 'consensus clusters' (see (24) for a detailed mathematical account of MICC; a glossary
159 of terms used in this paper is provided as Text S1). A consensus cluster represents a collection of
160 modules that share a significant number of coexpressed genes. These shared coexpression patterns
161 suggest that conserved and correlated biological processes are occurring in these tissues at the time of
162 biopsy. Consensus clusters in the information graph can be examined for an enrichment of modules
163 differentially expressed in SSc-associated pathophenotypes or 'pathophenotype-associated modules'
164 (Fig 1).

165 The addition of pathophenotype information and the identification of pathophenotype-associated
166 modules allow us to determine: 1) if a core set of processes is perturbed in SSc tissues and 2) if
167 distinct gene expression patterns underlie specific clinical phenotypes. Prior studies of gene expression
168 from the same disease suggest that molecular signatures of disease are more prominent than tissue-
169 specific signals (e.g., Huntington's disease expression signatures in blood and brain are more
170 concordant than muscle expression from two distinct diseases) (25). In SSc, we find similar patterns of
171 gene expression between clinically-involved forearm and clinically-uninvolved back samples (26), which
172 allowed for the identification of gene expression-based 'intrinsic subsets' (11-13). Molecular
173 heterogeneity is a hallmark of disease as intrinsic subsets have been identified in two different end-
174 target organs (16). This suggests that these conserved patterns are reproducible features of SSc
175 expression data. Multi-tissue MICC allows us to leverage our knowledge of these SSc-associated
176 pathophenotypes to inform us about less well-studied tissues and to identify the presence or absence
177 of common SSc-associated perturbations. Using basic set operations (see Methods), consensus
178 clusters in the information graph correspond to disease-associated gene sets that can be used to query
179 tissue-specific functional genomic networks (1). By submitting the same query to multiple tissue- and
180 cell type-specific networks, we are able to compare gene-gene interactions and to investigate
181 divergence of pathophenotypes in the different tissues sampled (Fig 1).

182

183 **The multi-tissue information graph reveals that SSc pathophenotypes share immune and**
184 **fibrotic coexpression patterns**

185 All datasets were partitioned into coexpression modules using WGCNA. The resulting 549
186 modules were input into MICC and the 10-partite information graph was constructed (Fig 2, Table II).
187 Community detection was performed to identify groups of modules that are significantly conserved
188 across data sets. The output of this procedure is groups of modules from different datasets that
189 significantly mutually overlap each other. Each such group of modules indicates a shared or
190 'consensus' molecular process between datasets and is termed 'consensus cluster.' As in Mahoney et
191 al., multi-tissue MICC identifies clusters that are dysregulated in SSc as well housekeeping processes
192 such as DNA repair and metabolism of proteins (1A and 3A respectively) (24) that are common to all
193 samples (Table III). Additional discussion of pathophenotype-enriched consensus clusters can be found
194 in Text S2.

195 Clusters 4A and 4B in the information graph are densely interconnected and contain modules with
196 increased expression in *all SSc-associated pathophenotypes of interest*: the inflammatory and
197 proliferative subsets, PAH, and PF (Fig 2B). Edges in the information graph represent overlap between
198 coexpression modules in different data sets, so we identified the intersection of genes between
199 adjacent modules. We then asked if the edge gene sets were similar to known biological processes by
200 computing the Jaccard similarity between edges and canonical pathways from MSigDB (27)(see
201 Methods). Edges in 4A are similar to immune processes such as antigen processing and presentation
202 and cytotoxic T cell and helper T cell pathways (Table III). This cluster also contains modules from all
203 tissues including PBMCs (Fig 2B). Altered immunophenotypes have been reported in SSc-PAH and
204 SSc-PF (14,15,17,18). Here, we find that the upregulated immune processes in these severe
205 phenotypes have substantial overlap with each other and the inflammatory subsets in esophagus and
206 skin (Fig 2B and S1). Notably, 4A is composed of modules with increased expression in PAH in PBMCs
207 and in lung, and a module upregulated in end-stage PF (Fig S1). *This demonstrates a commonality*
208 *between the inflammatory component of SSc and the most severe end-organ complications at the*
209 *expression level.* Edges in 4B overlap with processes that are typically thought of as pro-fibrotic,
210 including ECM receptor interaction, collagen formation, and TGF-β signaling (Table III). Cluster 4B
211 consists of skin inflammatory- and fibroproliferative-associated modules as well as lung PAH-, late PF-
212 and early PF-associated modules (Fig 2B and S1). These results validate and expand what we have

213 found in our prior meta-analysis of skin data (24): the immune-fibrotic processes (IFP) observed in SSc
214 intrinsic subsets are *connected* to and, furthermore, *are found in all* other tissues and SSc-associated
215 pathophenotypes.

216 To understand how the IFP in these phenotypes are functionally related, we identified the
217 consensus genes in the combined 4A and 4B clusters and determined that consensus genes from 4A
218 and B on average are much more correlated to their respective module eigengene than non-consensus
219 genes (Fig S2). Thus, consensus genes are highly central within their respective dataset gene-gene
220 correlation networks and our procedure identifies sets of genes that capture disease-specific variation.
221 Importantly, the union of these tissue consensus gene sets is enriched for genes upregulated in PF in
222 both lung data sets ($p < 2.2e-16$, as reported by R), in PAH lung samples ($p = 7.88x10^{-5}$), and in PAH in
223 both PBMC data sets ($p = 3.20x10^{-15}$) (genes identified with SAM (28), FDR <5%, Fisher's exact test,
224 two-sided p-values). Therefore, the tissue consensus strategy finds differentially expressed genes
225 highly relevant to SSc-associated complications. This positive control validates our approach, as the
226 tissue consensus genes (defined by our procedure) are enriched for genes differentially expressed
227 within each pathophenotype and data set, as defined by an auxiliary, rigorous statistical method (SAM
228 (28))that is standard in genome-scale gene expression analysis. Thus, the consensus gene sets allow
229 us to rigorously extrapolate from specific pathophenotypes and tissues to all tissues studied. We
230 analyzed the union of the tissue consensus gene sets, or 'SSc disease consensus genes,' by
231 submitting them as queries to functional genomic networks to investigate their functional significance.
232

233 **The lung microenvironment provides a distinct milieu for pro-fibrotic processes**

234 The GIANT functional networks infer functional relationships between genes by integrating
235 publicly available data including genome-wide human expression experiments, physical and genetic
236 interaction data, and phenotype and disease data (1). In these networks, genes are nodes and edges
237 are weighted by the estimated probability of a tissue-specific relationship between them. The edge
238 weight corresponds to the probability of interaction. We filtered these networks for high confidence
239 edges between gene pairs (probability of interaction > 0.5). We queried multiple tissue-specific and cell
240 type-specific networks to interrogate consensus gene functions across tissues included in this study.

241 We queried the GIANT lung network to investigate the function of the SSc disease consensus
242 genes in pulmonary manifestations of SSc, which are the leading cause of mortality ('consensus lung
243 network'; Fig. 3 and S3). We find interconnected functional modules related to both immune
244 (interferon/antigen presentation and innate immune/NF-κB/apoptotic processes) and fibrotic (response
245 to TGF-β and ECM disassembly/wound healing) processes (Fig. 3A). Thus, there is an abundance of
246 biological evidence for the coupling of the IFP in pulmonary manifestations of SSc.

247 Our analysis includes two lung datasets derived from both 'early' disease (open lung biopsies
248 from SSc-PF patients obtained for diagnostic purposes (18)) and end-stage or 'late' disease (SSc-PF
249 patients that underwent lung transplantation (17)). Comparison at different disease stages allows us to
250 interrogate how genes that have increased expression (SAM, PF vs. Normal comparison, FDR < 5%) at
251 each of these stages are distributed throughout the consensus lung network and to infer the order of
252 molecular events in SSc-PF progression.

253 Genes overexpressed in SSc-PF are distributed throughout the consensus lung network and
254 therefore are predicted to participate in all of the molecular processes identified in the network. Genes
255 that contain putative SSc risk polymorphisms are found in two functional modules (Fig 3A, gene
256 symbols in bold): *TNFAIP3* and *PLAUR* are in the innate immune response/NF-κB/apoptotic processes
257 module and *CAV1* is in the response to TGF-β module. Interestingly, *TNFAIP3* and *PLAUR* are also
258 upregulated in late SSc-PF. Quantification of the distribution of SSc-PF differentially expressed genes
259 throughout the consensus lung network (Fig 3B) demonstrates that molecular processes can be
260 associated with disease stage or transitions. For instance, the cell cycle module contains only early
261 SSc-PF genes, the innate immune response/NF-κB/apoptotic processes module contains more late
262 SSc-PF genes, and the response to TGF-β module contains genes from both disease stages.

263 Differences between the lung and skin consensus networks can offer additional insight into SSc-
264 PF pathogenesis. For high probability edges (> 0.5) that are present in both lung and skin, there are
265 gene pairs that are more likely to interact in one tissue than the other (Fig 4A). To identify *highly lung-*
266 *specific interactions*, we performed differential network analysis (see methods). We display a binary
267 representation of the tissue-specific interactions (Fig 4B), where a cell is red if the probability of the
268 interaction in lung is higher than the probability of interaction in all tissues and skin (lung-specific) or

269 blue if it is skin-specific (see Fig S4 for a fully labeled heatmap). The number of tissue-specific edges in
270 each functional module is quantified in Fig. 4C. The matrix and barplot illustrate that most functional
271 modules are less likely to interact in lung than in skin, with the exception of the cell cycle module. Of
272 particular interest is the relationship between the phagolysosome/ECM disassembly genes and
273 response to TGF-β genes, as differential connectivity can be observed in the phagolysosome/ECM
274 disassembly module (Fig 4B and 4C). Thus, even though ECM disassembly and TGF-β module genes
275 are differentially expressed in both lung and skin in SSc, they are differentially connected to each other
276 suggesting that the microenvironment strongly determines the functional consequences of upregulating
277 profibrotic genes.

278 We summarized the lung-specific biological processes in the differential lung network (Fig S5,
279 derived only from the interactions in red in Fig 4B). Many functional modules were identified; one of the
280 largest (Fig 4D and S5C) includes *NPC2*, *S100A4*, and *CTSB*, which encode protein products that are
281 highly expressed in normal lung-resident macrophages (LR-MØs) (29,30). Thus, a major difference
282 between the consensus lung and skin networks can be attributed, in part, to the presence of LR-MØs in
283 lung.

284

285 **Hub and bridge genes are highly relevant to the pathogenesis of pulmonary fibrosis**

286 Some genes occupy privileged positions within molecular networks and these genes often have
287 crucial biological function (31). Module hub genes are connected to a significant fraction of genes within
288 a molecular module, whereas bridge genes are genes that connect to multiple molecular modules and
289 thus 'bridge' them. We examined the hub and bridge genes within both the consensus and differential
290 lung networks for their possible roles in PF pathogenesis. We analyzed genes that were important in
291 both the consensus lung and differential lung networks in order to identify groups of genes of interest to
292 SSc-PF; we have highlighted the hubs and bridges of the full consensus lung network (Fig. 3C-E and
293 3F, respectively). The hubs of several of the functional modules in the consensus lung network show
294 increased expression at different disease stages (Fig 3C-E). Information regarding selected genes from
295 the consensus lung network, including functional module, network position, and differential expression,

296 is presented in Table IV. Salient features of the consensus and differential lung networks are described
297 below.
298
299 The gene Niemann-Pick disease, type C2 (*NPC2*) is upregulated in early disease and is connected to
300 cathepsins L and B (*CTSL, CTSB*) and *GLB1* in the consensus lung network (Fig 3C). This signal is
301 retained in the differential lung network (Fig 4D and S5C) and, as described above, we infer that this is
302 an LR-MØ signature. Interestingly, it has been shown that CTSL and CTSB regulate NPC2 secretion,
303 TNF-α production, and cholesterol trafficking genes in an animal model of obesity (32). This suggests
304 that lipid and cholesterol trafficking in LR-MØs may be upregulated in early SSc-PF and therefore may
305 play a role in inflammation.
306 *CD44* and *PLAUR* (also known as uPAR), which have been implicated in idiopathic PF, are
307 connected to one another and bridge multiple functional modules (Fig 3F). CD44 is a hyaluronic acid
308 receptor that can interact with many other ligands found in the ECM. Notably, anti-CD44 treatment of
309 primary idiopathic PF fibroblasts attenuates invasion of these cells and reduces bleomycin-induced lung
310 fibrosis in a mouse model (33). Moreover, Grove *et al*. found that pulmonary fibroblasts from patients
311 with idiopathic PF overexpress uPAR and that uPAR ligation results in a hypermotile phenotype (34).
312 The innate immune response/NF-κB signaling/apoptotic process module contains mostly genes
313 up in late SSc-PF including *CYR61* and *TM4SF1* (Fig 3A-B and S3). *TNFAIP3* (A20), a negative
314 regulator of NF-κB signaling and inhibitor of TNF-mediated apoptosis, is a hub of this module and is
315 increased in late SSc-PF (Fig 3E). The innate immune response/NF-κB signaling/apoptotic process and
316 interferon/antigen presentation modules are bridged by *TNFSF10,* which is also known as TRAIL (TNF-
317 related apoptosis inducing ligand, Fig 3F). These results suggest that the balance of apoptosis is
318 altered in late SSc-PF. The upregulation of genes with anti-apoptotic function was not reported in the
319 original study (17), which demonstrates the strength of both the MICC method and the study of gene-
320 gene interactions.
321
322

323 **Multi-tissue MICC recovers SSc risk genes that occupy privileged positions in the consensus**
324 **skin network**

325 Skin fibrosis is present in nearly every patient with SSc, though the severity is variable. To
326 identify potential mechanisms contributing to skin fibrosis, we queried the GIANT skin network with the
327 disease consensus genes and analyzed the resulting network ('consensus skin network'), first
328 identifying functional modules. As in Mahoney *et al*. (24), we observed connectivity between immune-
329 related and fibrotic functional modules, now using a skin-specific functional genomic network and the
330 SSc disease consensus genes (Fig 5A and S6). We found key differences between the consensus lung
331 and skin networks. Unlike the lung network, the skin network implicates the involvement of the adaptive
332 immune system, specifically T lymphocytes, in inflammation (Fig 3 and 5A). In addition, the ECM
333 disassembly genes and response to TGF-β genes are more likely to interact in skin than in lung (Fig 4).
334 Consistent with these results, one module indicative of fibrosis—the extracellular matrix/TGF-β
335 response module—was identified in skin (Fig 5A, violet module). These findings suggest that TGF-β
336 signaling and ECM remodeling are more tightly coupled in skin, whereas in lung they are functionally
337 distinct.

338 In the consensus lung network, only 3 genes with putative SSc-risk polymorphisms interacted
339 with the disease consensus genes (Fig 3); however, in skin, more risk genes are associated with the
340 disease consensus genes (Fig 5A, gene symbols in bold). Importantly, we recovered 11 out of 41
341 previously identified risk genes (24) (27%) and they were distributed throughout the functional modules
342 (Fig 5A, gene symbols in bold). This finding reflects results obtained in our prior analysis of skin alone
343 (24). The significant distinction from Mahoney, *et al*. is that we did not specifically query the network
344 with risk genes. Rather, these genes are either tissue consensus genes (*GRB10*, *PTPN22*, *TLR2*,
345 *SCHIP1*, *HLA-DPA1*, *HLA-DPB1*, *HLA-DQB1*, *CAV1*, and *CD247*) or are highly likely to interact with
346 tissue consensus genes and are pulled in by the functional network step in our pipeline (*PLAUR* and
347 *TNFAIP3*; see Methods). The disease consensus genes are the *major disease signal* in SSc
348 expression data and thus genes with polymorphisms contribute to this signal.

349 The functional significance of risk genes in SSc is largely unknown. In the skin consensus
350 network, genes with putative SSc-risk polymorphisms were more likely to be bridges—genes that

351 interact with genes in multiple functional modules (see Methods). Indeed, 3 out of 9 genes that contain
352 SSc-risk polymorphisms (degree > 1) are bridges (Fisher's exact test, p = 0.05). As shown in Fig 5B,
353 *TNFAIP3* and *HLA-DQB1* have common edges with genes in the interferon, T cell proliferation, and
354 cytokine modules. *TNFAIP3* and *PLAUR* have common edges with the phagocytosis and ECM
355 modules (Fig 5B). This suggests that SSc risk genes are more likely to bridge multiple functional
356 modules and likely interact, in skin, with many genes contributing to differential expression in SSc.
357 Furthermore, the disproportionate representation of putative risk genes in bridge positions suggests
358 that bridge nodes may be crucial to *skin disease modulation*.
359

360 **Interactions among skin fibroblasts, MØs, and T lymphocytes give rise to SSc skin disease**
361 **processes and are evident at the functional network level**

362 Although the structure in the consensus skin network is representative of SSc disease processes
363 (Fig 5), SSc skin biopsies are nevertheless composed of multiple cell types. With GIANT cell type-
364 specific networks, we can determine if a functional module that we have detected in the skin network
365 persists in a cell type network of interest and then attribute the signal in whole tissue to this particular
366 cell type. We chose to analyze keratinocyte, skin fibroblast, MØ, and T lymphocyte networks because
367 we expect that these are representative of the complex mixture of cell types found in an SSc skin
368 biopsy.

369 Cell types within a tissue are not independent; rather, they interact to produce the structure and
370 function of the tissue. As a result, the GIANT skin network is not simply a 'sum' of cell type networks,
371 but instead is simultaneously an estimate of the gene-gene interactions within the constituent cell types
372 and of the interactions between cell types in the tissue. To dissociate these within/between cell type
373 interactions and to identify which cell types may contribute to various components of the skin network
374 structure, we calculated whether functional modules detected in skin had significantly more weight
375 within or between them in cell type-specific networks than at random (Fig. 6; see Methods). The T cell
376 module detected in skin is only significant in the T lymphocyte common network (Fig 6B). This suggests
377 that we are capturing the presence of these cells in the skin network and that our test for module
378 preservation is able to detect *relevant features* in common cell type-specific networks. Another

379 important control is the lack of a cell type-specific cell cycle signal (Fig 6B); we would expect all cell
380 types to have a cell cycle program that is not significantly different than all other tissues and, therefore,
381 this module would be removed in the global average subtraction step (see Methods). With the
382 exception of the cell cycle module, each of the skin modules is significantly represented in at least one
383 cell type (Fig. 6B). Of particular interest is the MØ ECM module, as we believe this cell type may be
384 responsible for the secretion of TGF-β in lung and we have previously inferred an activated MØ
385 functional module in our meta-analysis of SSc skin data (24).

386 There is significant weight off diagonal in several of the common cell type-specific networks,
387 which indicates that genes in these functional modules are interconnected. For instance, keratinocytes,
388 skin fibroblasts and MØs all have densely interconnected interferon/antigen presentation and
389 cytokine/NF-κB signaling modules (Fig 6B). This makes biological sense, as we expect that genes that
390 participate in these responses would functionally interact. Keratinocytes and skin fibroblasts have
391 densely interconnected cytokine/NF-κB signaling and ECM/growth factor signaling modules (Fig 6B).
392 This suggests that these skin-resident cell types are capable of coupling the inflammatory and fibrotic
393 responses; i.e., we infer that an inflammatory stimulus (e.g., interferon) could lead to a fibrotic response
394 in these cell types. Many blocks of the compressed matrix have significant weight in multiple cell types,
395 which suggests that *functional modules in skin arise from interactions among cell types*.

396

397 **A cell type-aware network measure prioritizes multi-functional genes with experimental support**
398 **in model systems**

399 Because the interaction of cell types is implicated in SSc disease signatures, we sought to
400 understand the relative *importance* of genes from the skin network in the cell type-specific networks.
401 We developed a measure—termed 'cell type centrality (CTC)—to identify genes that bridge multiple cell
402 types (see Methods). Briefly, we computed the maximum degree (max degree) of each node across all
403 cell types and we defined the CTC to be the average of the max degree of a node's neighbors. Thus, a
404 node has high CTC if many of its neighbors have high degree (i.e. are important) in a cell-type network.
405 Because the degree structure of the cell-type networks varies (e.g. the MØ network has more edges
406 than any other network), we quantile-normalized the degree in each network prior to finding the

407 maximum degree. The significance of a node's CTC was calculated by permutation testing (see
408 Methods).

409 CTC is biologically motivated; we hypothesize that genes that facilitate functional interactions
410 between cell types are likely to be important to SSc skin biology. CTC is a network-based measure of
411 functional interactions across cell types. Genes with significant CTC ('significant genes'; 146 in total)
412 were found in all of the functional modules depicted in Fig. 7A (blue and green nodes, see File S4 'cell
413 type pruned skin nodes'), with the exception of the cell cycle module (fully labeled network in Fig S7).
414 We next characterized the local network properties of the significant genes. The degree distribution of
415 these genes is higher than that of the network in general (Fig S9). However, this plot revealed that CTC
416 captured and allowed for the prioritization of biologically interesting genes that have a relatively low
417 degree in the skin network. For instance, the significant gene *CCL2* is known to be important in the
418 sclerodermatous graft-vs-host disease (sclGVHD) inflammatory mouse model of skin disease as
419 Greenblatt *et al*. showed that antibody blockade of CCL2 prior to graft ameliorated skin disease (35). In
420 addition, genes that have been shown to play a role in systemic autoimmunity and fibrosis in animal
421 models such as *TRIM21 (36,37)* and *SGK1 (38)* were also identified as significant despite having low
422 degree in the skin network. Furthermore, significant genes had higher bridgeness scores than other
423 genes in the network (Fig S8; Mann Whitney U, $p < 2.2x10^{-16}$ reported by R), strongly suggesting their
424 multi-functionality and importance for skin disease modulation.

425 The CTC statistic identifies genes that potentially bridge multiple cell type networks, but it does
426 not quantify which cell types are bridged by those genes. Each gene is labeled by the cell type network
427 in which it achieves maximum degree (cell type label). For example, genes that are most highly
428 connected in the fibroblast network can be labeled 'fibroblast genes', as we assume that these genes
429 have the most influence within fibroblasts. To understand the cell types bridged by genes with
430 significant CTC, we characterized genes in the skin network in terms of their 'neighborhood
431 composition'—i.e., what proportion of a genes neighbors are most highly connected in each of the four
432 cell type-specific networks. For example, a gene that bridges MØs and fibroblasts has many neighbors
433 that have max degree in the MØ and fibroblast networks. Consequently, most of its neighbors will be
434 labeled as MØ or fibroblast genes.

435 Clustering of the proportion matrix (shown in Fig. 7B) revealed sets of genes with distinct
436 neighborhood composition that we then examined for the over- or under-representation of significant
437 genes. Clusters that lacked significant genes contained nodes that were disproportionately highly
438 connected to keratinocyte and skin fibroblast genes (Fig. 7C, clusters 3 and 7). Conversely, clusters
439 that were enriched for significant genes were indicative of neighborhoods dominated by one (MØ,
440 cluster 11) or both immune cells (MØ and T lymphocyte, cluster 15) emphasizing the importance of
441 these cell types in the SSc skin disease signal. Moreover, these clusters contained highly multi-
442 functional receptors (*CD44*, *ICAM1* in cluster 11), and interleukin cell surface receptors and soluble
443 ligands (*IL15*, *IL7R* in cluster 15).
444
445 **Immunosuppressive therapy disrupts MØ – T cell interaction genes in the network**
446 We next looked at gene expression data from a small clinical study of the immunosuppressive
447 agent mycophenolate mofetil (MMF; Hinchcliff *et al.*) to determine if genes with significant CTC were
448 modulated during treatment. We hypothesized that genes that are downregulated during treatment in
449 patients with a clinically significant reduction of modified Rodnan skin score (mRSS—a measure of skin
450 disease severity) (39) would be genes that are more likely to modulate functions in multiple cell types
451 present in skin, i.e., genes with significant CTC. Clinical improvers were defined as patients that had a
452 decrease in MRSS of at least 5 points between base and 12 month time points while on MMF (39).
453 Baseline and post-treatment biopsies from responders (n=6; flank and arm biopsies) were compared to
454 identify genes modulated during response to treatment (base vs. 12 months, paired t-test, FDR < 0.05).
455 As genes with increased expression in SSc manifestations tend to be conserved compared with those
456 that have decreased expression (see section on the information graph and Mahoney *et al*. (24)), we
457 focused on genes that were downregulated during response to MMF (up in baseline biopsies of
458 improvers). See Table SII for detailed clinical information.
459 Of the 578 genes in the cell type-pruned skin network (Fig. 7), 48 were downregulated during
460 response to MMF (Fig. 7, green and yellow nodes; hypergeometric $p = 2.82 \times 10^{-10}$). Twenty-three of
461 these genes had significant cell type centrality including *STAT1*, *IFI16*, *TNFSF10*, *THBS1*, and *PLAU*.
462 (Fig. 7, green nodes); therefore, genes with significant CTC are more likely to be altered during

463 response to immunosuppressive therapy than genes in the network in general (Fisher's exact test, p =
464 0.0004, OR 3.03). An enrichment of non-significant MMF response genes was identified in the T
465 lymphocyte cluster 17 of the proportion matrix, which contained genes highly connected to T
466 lymphocyte genes (Fig. 7C). Genes that fit both criteria were over-represented in the MØ – T interaction
467 cluster 16 (Fig 7C, Fisher's exact test, p = 0.00056), which was indicative of genes that bridge the MØ
468 and T cell networks. Additionally, the significant MMF response genes themselves are highly connected
469 in the MØ network beyond what would be expected by chance (Fisher's exact test, p = 0.0012) and are
470 disproportionately connected to the ECM/TGF-β response module (Fig. S10, Mann Whitney U, p =
471 2.68×10^{-5}). MMF is a general immunosuppressive agent that is thought to inhibit lymphocyte
472 proliferation (40). We assert that the disruption of the genes that capture the interaction between these
473 two immune cell types is not only informative about the action of MMF in SSc skin, but it also places
474 MØs at the center of skin disease pathogenesis. Furthermore, significant CTC genes represent a rich
475 set of putative targets for clinical intervention, perhaps with a more specific agent.
476

477 **Discussion**

478 We performed the first-of-its-kind multiple tissue, multiple clinical manifestation study of SSc gene
479 expression data. In this rare disease characterized by clinical and molecular heterogeneity, functional
480 genomic analyses are a powerful tool to generate hypotheses. These gene expression data capture a
481 steady-state picture of molecular processes occurring in tissues at the time of biopsy. Our pipeline
482 allowed us to identify imperfectly conserved coexpression across datasets from different clinical
483 centers, tissues, and platforms. Fundamentally, we found that the gene expression patterns underlying
484 SSc-associated complications and the inflammatory intrinsic subsets are similar and are the result of
485 immunological abnormalities. In particular, coexpression modules from PBMC datasets, a mixture of
486 innate and adaptive immune cells, had significant overlap with modules associated with all severe
487 pathophenotypes included in this study, whereas fibrotic processes were largely associated with solid
488 tissue modules. As with our prior studies, we did not find a strong association with autoantibody
489 subtype and the co-expression modules identified here.

490 Our strategy addresses both common disease perturbations and tissue-specific differences and
491 links abnormalities at the molecular (mRNA) level to abnormalities in cellular function and SSc-
492 associated pathophenotypes. We identified an SSc disease consensus gene set that was enriched for
493 genes that have increased expression in pulmonary disease, and studied the functional interaction of
494 the consensus genes in gene-gene networks that are tissue- and cell type-specific. As such, we are
495 able to identify genes that bridge molecular processes in both skin and lung and to build models of
496 tissue-specific disease progression that are difficult to capture experimentally. In addition, we show that
497 these networks capture important SSc pathobiology and, as a proof of principle, demonstrate that
498 genes that have important network features are more likely to be altered during therapeutically-
499 modulated improvement in skin disease. Although it is clear from this work that the coupling of immune
500 and fibrotic processes is a hallmark of SSc, we found subtle organ-specific functional differences that
501 we attribute, in part, to the plasticity of the myeloid cell lineage.

502

503 **Networks capture a fundamental disease signature that is altered during response to treatment**

504 The gene-gene networks presented above capture fundamental disease processes in SSc
505 tissues; this is illustrated by the distribution of genes with increased expression in SSc-PF throughout
506 the consensus lung network (Fig. 3). We applied network analysis techniques to characterize functional
507 modules in skin and lung and to identify important genes in these tissues. For skin, we were able to add
508 cell type-specific information to our analysis due to the existence of cell type-specific networks. We
509 identified genes with salient network features in both skin and lung.

510 For example, *CD44* is a highly connected bridge node in both tissues and has significant CTC in
511 skin. This pleiotropic receptor has been found to play a role in type 2 diabetes (T2D) using an
512 expression-based association study (41), in addition to its role modulating fibroblast migration (33).
513 Furthermore, blocking CD44 ameliorates disease in an animal model of T2D in part through reducing
514 MØ infiltration of adipose tissue (41). Shedding of cell surface CD44 (through antibody-blocking) also
515 reduces arthritis severity in a murine model (42,43). Intriguingly, SSc patients with higher soluble CD44
516 in their serum have less severe disease (44). *PLAUR* was also identified as an important bridge in both
517 tissues and contains a putative SSc-risk polymorphism. Significantly, this gene is predicted to interact
518 with genes that participate in inflammatory and fibrotic functional modules. We postulate that these
519 genes may be drivers of disease in dermal and pulmonary fibrosis and thus, warrant further
520 mechanistic study in SSc and may prove to be rational targets for clinical intervention in multiple
521 manifestations of SSc.

522 To date, the majority of work on SSc expression data has not systematically considered the
523 interaction of cell types captured in the biopsies included in the study. Here, we queried specific
524 functional networks of relevant skin cell types—keratinocyte, skin fibroblast, MØ and T lymphocyte—to
525 address this issue. We found that genes that composed the functional modules we detected in skin had
526 non-random interactions between them in cell type-specific networks and that we could attribute these
527 functional modules to the interaction of cell types. On a more granular level, we used the cell type-
528 specific network features of the genes found in skin to guide our analysis by utilizing a novel metric
529 termed cell type centrality (CTC).

530 We demonstrated that CTC captures information about the ability of a gene to modulate function
531 in multiple cell types and to participate in multiple molecular processes. We found that genes in the cell

532 type-pruned consensus skin network were found in the MMF response signature and that genes with
533 significant CTC were more likely to be downregulated during clinical improvement than other genes in
534 the network. In particular, significant MMF response genes tended to bridge MØ and T lymphocyte
535 genes. An independent clinical trial of the immunosuppressant abatacept (Orencia™) in SSc
536 demonstrated that some patients show clinical improvement when the interaction between T
537 lymphocytes and antigen-presenting cells is specifically disrupted (45). This demonstrates the strength
538 of our pipeline: we captured relevant disease processes that are modulated during treatment, speak to
539 the pathobiology of the disease, and are therefore of translational value.
540

541 **A model of SSc-PF pathogenesis**

542 In the present work, we have identified disease-relevant gene expression patterns in a data-
543 driven manner and investigated interactions between disease genes that are learned from a
544 compendium of publicly available human expression data. In addition, we leveraged the difference in
545 disease stage sampled in the two lung datasets incorporated into our study. These two datasets
546 contained patients with different types of interstitial pneumonia (see Methods), which may limit
547 interpretation of these results, as it is not known if these different patterns have similar underlying
548 molecular alterations. We can draw non-intuitive conclusions through the combination of this data-
549 driven analytical approach and mechanistic insight supported by seemingly disparate literature. We
550 present a model of SSc-PF pathogenesis in Fig. 8.

551 ***Early SSc-PF*** Christmann and coworkers identified an upregulation of interferon- and TGF-β-regulated
552 genes in biopsies from early SSc-PF (18). It was also noted that there was more CCL18 at the protein-
553 level and a higher level of CD163 transcript in SSc-ILD lungs, suggestive of the presence of
554 alternatively activated MØs (18). However, it is unclear which cell types are responsible for the
555 interferon signature or if there is evidence of two distinct populations of MØs. We found that gene sets
556 upregulated in alternatively activated human MØs and MØs treated with free fatty acids are enriched in
557 early SSc-PF patients and that there is no evidence of enrichment of pro-inflammatory, interferon-
558 stimulated MØ signatures (Fig S11) (19). In addition, we identified an LR-MØ signature in our

559 differential network analysis that consisted of genes upregulated in early SSc-PF and that participate in
560 lipid and cholesterol trafficking (Fig. 4D S6, differential lung network).

561 MØ-derived foam cells (lipid-laden MØs) have been noted in histological specimens of chronic
562 fibrotic lung disease and animal models (46). Combined with the results of our network analyses, this
563 observation suggests that the lung network indicates the presence of foam cells and alternatively
564 activated MØs in the biopsies from early SSc-PF patients. The LR-MØ genes are located in the ECM
565 disassembly/wound healing functional module of the full consensus lung network, which is highly
566 interconnected with the interferon/antigen presentation module suggesting a mechanistic connection.
567 For instance, IFN-α priming has been shown to increase transcript levels of lipid receptors in MØs and
568 to increase foam cell formation (47). In the bleomycin injury mouse model of pulmonary fibrosis, foam
569 cells have been observed to upregulate markers associated with alternative MØ activation and to
570 secrete TGF-β (46). Oxidized phospholipid treatment also caused alternative activation and TGF-β
571 secretion in human MØs (46). Therefore, we hypothesize that an initial insult that provokes a type I
572 interferon response in the lung may contribute to lipid uptake by the MØs of SSc-PF patients and
573 ultimately alternative activation and TGF-β secretion (Fig. 8).

574 ***Mid to Late SSc-PF*** TGF-β signaling is a hallmark of profibrotic disease, and was noted in the initial
575 analysis of both lung data sets included in this work (17,18). Similarly, we find genes from both stages
576 of SSc-PF in the response to TGF-β module of the lung network, but also that the type I interferon
577 signature may be sustained into late disease (Fig. 4, S4 bar plot lung modules). The functional module
578 most associated with late disease is the innate immune, NF-κB, and apoptotic processes module (Fig.
579 4, S4 bar plot lung modules). This module is connected to the response to TGF-β module in part
580 through components of the fibrinolysis pathway such as PAI-1 (*SERPINE1*), an inhibitor of fibrinolysis
581 (Fig. 3 and S8 bridges). PAI-1 is upregulated in late SSc-PF and is known to be important in pulmonary
582 fibrosis as mice overexpressing this gene deposit more collagen after inflammatory lung injury with
583 bleomycin (48). The role of fibrinolysis in ameliorating disease was also shown in a rabbit model of
584 bleomycin-induced lung fibrosis, where aerosolized administration of uPA reduced fibrosis (49,50). One
585 mechanism by which fibrinolysis may contribute to the resolution of fibrosis is through the induction of

586 fibroblast apoptosis (51). Notably, both TGF-β1 and PAI-1 have been shown to inhibit lung fibroblast
587 apoptosis (51).

588 We found evidence for a shift in the balance of apoptosis, perhaps in myofibroblasts (52), in our
589 network analyses (Fig. 8). Long-lived myofibroblasts are thought to continually deposit collagen and
590 contribute to persistent fibrosis (53). Moreover, impaired phagocytosis of apoptotic cells, or
591 efferocytosis, has been observed in the alveolar MØs of idiopathic PF patients (54). We note genes
592 involved in efferocytosis, specifically in the receptors (*CD44*) and endocytic machinery associated with
593 this process, in the consensus lung network (Fig. 3 and 8) (55). If the shift in apoptosis and
594 efferocytosis occurs, we speculate that the fibrotic and inflammatory processes in our network will also
595 be altered. Efferocytosis by alveolar MØs plays a key role in the resolution of inflammation in the lung
596 through the subsequent release of TGF-β (56). We hypothesize that following initial injury, TGF-β
597 signaling, antifibrinolytic factors, and the disruption of apoptosis and efferocytosis contribute to
598 progressive fibrosis in SSc-PF and, ultimately, end-stage disease (Fig. 8).

599

600 **The plasticity of the myeloid lineage may drive tissue-specific SSc disease processes**

601 By performing a meta-analysis of SSc tissue gene expression data, we are able to observe and
602 infer, in a genome-wide manner, commonalities between a patient's cytokine milieu and the complex
603 mixture of cell types in a tissue at the time of biopsy. Overwhelmingly, we detected a MØ signature
604 associated with severe disease. We found evidence for the contribution of LR-MØs to SSc-PF
605 progression, consistent with the alternative activation of MØs and TGF-β production. In our prior
606 analysis of a single tissue, we inferred activated MØs as modulators of the inflammatory subset in skin
607 (24). Our current study further supports this hypothesis, as genes in the ECM and TGF-β response
608 functional module in skin likely function together in a coherent manner in the MØ-specific network (Fig.
609 8). Simultaneously, we observed a coupling of TGF-β signaling and ECM disassembly for which there
610 is less evidence in lung (Fig. 4), where an LR-MØ signature distinguishes the functional relationships of
611 SSc disease consensus genes (Fig. 5). We posit that the differences in fibrotic responses in the two
612 tissues (networks) are due, in large part, to innate differences in tissue-resident MØs that have been
613 observed (57,58), as well as the interactions between infiltrating monocytes and tissue-resident cell

614 types (e.g., alveolar epithelial cells vs. keratinocytes). Because MØ phenotype and function are plastic
615 and readily modulated by the local tissue microenvironment, it is likely that differential activation of MØs
616 in these tissues is the result of exposure to distinct cytokine milieus. Alternatively, we cannot rule out
617 that the MØ changes we observe are a secondary response to SSc pathology. Regardless, therapies
618 that target MØ effectors such as IL6R have shown promise in clinical trials (59) suggesting they play a
619 central role in SSc pathogenesis.

620 The importance of the innate immune system in fibrosis has been demonstrated in Rag2$^{-/-}$ mice
621 (which lack mature lymphocytes) exposed to bleomycin-induced dermal and pulmonary injury (60,61).
622 Here, in a complementary study and in a systematic manner, we have now analyzed data from SSc
623 patients with diverse pathophenotypes. In the information graph, we observe that PAH-associated
624 modules from PBMCs (14,15) have significant overlap with inflammatory subset-associated modules
625 from skin and esophagus (Fig. 2). Indeed, in Pendergrass *et al. (14)*, our group observed that PBMCs
626 from lcSSc patients have significant enrichment in myeloid- and MØ-related gene sets as compared to
627 healthy controls and Christmann *et al*. (62) expanded upon this, showing that transcripts highly
628 expressed in lcSSc-PAH CD14^{+} monocytes were induced in IL-13-stimulated cells (alternatively
629 activated). We assert that this MØ polarization is a significant part of the immune-fibrotic signal we find
630 in these data and, therefore, is likely *a common driver* of the pathobiology of SSc complications studied
631 here. An independent study also identified MØs and dendritic cells (DCs) as possible sources of an
632 'inflammatory' signature in lesional SSc skin (22). We cannot rule out that DCs contribute to our results,
633 as they are observed to be important in the Stiff Skin Syndrome mouse model (63), although some
634 skin-resident DCs have been shown to be transcriptionally similar to peripheral blood monocytes in
635 humans (64). We speculate that the circulation of peripheral myeloid cells contributes to the multi-organ
636 nature of SSc. Future studies may use *in silico* and cell sorting techniques to deconvolute SSc
637 expression data to identify changes in cell proportion and transcriptome throughout disease course and
638 to finely phenotype myeloid cells from SSc patient tissue samples.

639

640

641 **Conclusions**

642 The clinical heterogeneity and rarity of SSc makes it challenging to study, as all severe disease
643 manifestations typically are not present in an individual patient at a given time. In addition, there are
644 concerns about the ethics and feasibility of acquiring biopsies from multiple tissues from patients,
645 particularly in the case of the lung. In this study, we have utilized previously acquired data from multiple
646 tissues to examine the systemic nature of SSc. Our integrative analysis allowed us to leverage well-
647 studied tissues to inform us about SSc manifestations that are under-studied molecularly. This study
648 rigorously tests the observation that patients with severe disease have immunological and fibrotic
649 alterations at the gene expression level. Indeed, we were able to identify a common immune-fibrotic
650 axis that is modulated by the skin and lung microenvironments as captured by the tissue-specific
651 functional genomic networks. In particular, the tissue-specific context of pro-fibrotic MØs may greatly
652 contribute to the outcome in end-target tissues in SSc. Our network findings have experimental support
653 in animal models and *ex vivo* studies. Furthermore, we show that response genes to MMF treatment
654 have a network signature, which validates our cell type-aware network measure and indicates that
655 network approaches are a rich source of putative targets for therapeutic intervention. In sum, this work
656 provides a conceptual framework for multi-tissue functional genomic studies as well as a novel multi-
657 tissue overview of SSc pathogenesis.

658 **Methods**

659 **Patients and datasets**

660 Eight out of 10 datasets included in this study were previously published (see Table I) and
661 descriptions of the patient populations and criteria for inclusion can be found in those publications. We
662 used the patient disease label (e.g., PAH) as published in the original work for all of these sets. In Table
663 SI, we summarize the patient information we had access to on a per array basis as that is what is
664 required for comparison to the expression data. Below, we note some important characteristics (for the
665 purposes of this work) of the patient populations included. In the Bostwick lung dataset (17), all patients
666 with SSc-PF and idiopathic PF had usual interstitial pneumonia (UIP). This study used lung tissues
667 from patients who underwent lung transplantation (late disease). The Christmann lung dataset (18)
668 contains 5 patients with non-specific interstitial pneumonia (NSIP) and 2 patients with centrilobular
669 fibrosis (CLF). This study looked at early SSc-PF patients, used open lung biopsies, and specifically
670 avoided honeycombing areas. Some patients included in the PBMC dataset, including those with PAH,
671 also had interstitial lung disease, though exclusion of these patients does not significantly change the
672 interpretation as put forth in (14). As illustrated in Table SI, two datasets (ESO, LSSc) did not contain
673 healthy control samples and three datasets (UCL, LSSc, and PBMC) were comprised entirely of lcSSc
674 patients.

675

676 **Ethics statement on previously unpublished datasets**

677 The LSSc and UCL studies are previously unpublished. The samples from the LSSc dataset were
678 obtained at Boston University Medical Center (BUMC)/Boston Medical Center (BMC); the BUMC/BMC
679 Institution Review Board approved this study. The samples from the UCL dataset were obtained at
680 University College of London; the London-Hampstead NRES Committee approved this study. The
681 Dartmouth College CPHS approved this work. All subjects gave informed consent.

682

683 **Microarray dataset processing**

684 This work contains 10 datasets on multiple microarray platforms. Agilent datasets (Pendergrass,
685 PBMC, Milano, Hinchcliff, ESO, UCL, LSSc) used either Agilent Whole Human Genome (4x44K)

686 Microarrays (G4112F)(Pendergrass, PBMC, Milano, Hinchcliff, ESO, UCL) or 8x60K (LSSc). Data were
687 Log_2-transformed and lowess normalized and filtered for probes with intensity 2-fold over local
688 background in Cy3 or Cy5 channels. Data were multiplied by -1 to convert to Log_2(Cy3/Cy5) ratios.
689 Probes with >20% missing data were excluded. The Illumina dataset (Bostwick, HumanRef-8 v3.0
690 BeadChips) was processed using variance-stabilizing transformation and robust spline normalization
691 using the lumi R package. Dr. Christmann provided the raw data in the form of .CEL files. Dr. Feghali-
692 Bostwick provided Illumina BeadSummary files. Affymetrix datasets (Risbano, HGU133plus2 ;
693 Christmann, HGU133A_2) were processed using the RMA method as implemented in the affy R
694 package. Batch bias was detected in the ESO dataset. In order to adjust these data, missing values
695 were imputed via *k*-nearest neighbor algorithm using a GenePattern (65) module with default
696 parameters and the data were adjusted using ComBat (66) run as a GenePattern module to eliminate
697 the batch effect.

698 In order to compare datasets in our downstream analysis, duplicate genes must not be present in
699 the dataset and must be summarized in some way. First, we annotated each probe with its Entrez gene
700 ID. Agilent 4x44K arrays were annotated using the hgug4112a.db Bioconductor package. LSSc was
701 annotated using UNC Microarray Database with annotations from the manufacturer. Probes annotated
702 to lincRNAs (A19) were removed from the analysis. The Illumina dataset was annotated by converting
703 the gene symbols (provided as part of the BeadSummary file) to Entrez IDs using the org.Hs.eg.db
704 package. The Risbano PBMC dataset was annotated using the hgu133plus2.db package. The
705 Christmann dataset was annotated using an annotation file from the manufacturer. NAs and probes that
706 mapped to multiple Entrez IDs were removed in all cases. Probes that mapped to the same Entrez ID
707 were collapsed to the gene mean using the aggregate function in R, followed by gene median
708 centering.

709

710 **Clustering of microarray data and statistical tests for phenotype association**

711 The collapsed datasets were used to find coherent coexpression modules. We used Weighted
712 Gene Co-expression Network Analysis (WGCNA), a strong clustering method, which allows us to
713 automatically detect the number of coexpression modules and remove outliers. Each dataset was

714 clustered using the blockwiseModules function in WGCNA R package using the signed network option
715 and power = 12; all other parameters were set to default. The number of arrays and resulting co-
716 expression modules are summarized in Table II. Using the WGCNA coexpression modules also
717 reduces the dimensionality of the dataset, as it allows us to test for genes' association with, or
718 differential expression in, a particular pathophenotype of interest on the order of tens, rather than
719 thousands using the module eigengene. The module eigengene is the first principal component, and
720 represents the expression of all genes in a module and an idealized hub of the coexpression module.
721 We used the moduleEigengenes function in the WGCNA R package to extract the eigengenes. A
722 module was considered to be pathophenotype-associated if the module eigengene was significantly
723 differentially expressed in or significantly correlated with a pathophenotype of interest. Only 2-class
724 categorical variables were considered using a Mann-Whitney U test (i.e., all pulmonary fibrosis and
725 pulmonary arterial hypertension patients were grouped together regardless of underlying etiology). We
726 used Spearman correlation for continuous values. P-values were Bonferroni-corrected on a per
727 phenotype basis. See File S1 for complete output. In the main text, we discuss categorical
728 pathophenotypes as these are what were enriched at the consensus cluster level. We do find instances
729 coexpression modules that are associated with continuous pathophenotypes, such as pulmonary
730 function test measurements, but these were not apparent at the consensus cluster level of abstraction.
731

732 **Information graph construction and consensus cluster detection**

733 The 10-partite information graph was constructed as in Mahoney et al (24). In order to detect
734 clusters in the information graph, two methods of community detection were used. First, fast-greedy
735 modularity optimization (67) was used, which yielded large consensus clusters. The gene sets derived
736 from the clusters are dependent on the number of nodes (coexpression modules) within that cluster (as
737 detailed below). Therefore, smaller groups of nodes or clusters will result in more conservative gene
738 sets. In order to find sub-consensus clusters (i.e., smaller more densely connected sub-communities),
739 spin-glass community detection (igraph R package implementation, gamma.minus = 0.125, all other
740 parameters were set to default) was performed (68,69). Consensus and sub-consensus clusters as
741 displayed in Fig. 2 (top and right annotation tracks, respectively) are the result of consensus ordering—

742 first, ordering by fast-greedy community label and then, within each of those communities, ordering by
743 spin-glass community label—of the adjacency matrix followed by hierarchical clustering within the
744 smallest structure. These two community detection methods are based on modularity. By using both,
745 we refine the community structure to get smaller, more densely connected sub-consensus clusters. The
746 information graph adjacency matrix and node attributes are supplied in File S2.

747 Because an edge is only shared by coexpression modules from different datasets that share a
748 significant overlap in coexpressed genes, we can think of an edge 'encoding' that overlap as a gene
749 set. An edge gene set E_{ij} is defined as the overlap between the coexpression modules C_i and C_j in two
750 datasets

751 $$E_{ij} = C_i \bigcap C_j$$

752 The similarity of an edge E and a pathway P is calculated using Jaccard similarity

753 $$J(E,P) = \frac{\left|E \bigcap P\right|}{\left|E \bigcup P\right|}$$

754 The Jaccard similarity between the edge and pathway will be equal to 1 if the overlapping genes shared
755 between two modules are exactly the same set of genes annotated to the pathway, or zero if no genes
756 are shared between the two sets. As the Jaccard similarity is a statistic that ranges from 0 to 1, we then
757 identified pathways with disproportionally high similarity to edges within a cluster as compared to edges
758 outside of that cluster using a Mann-Whitney U test (with Bonferroni adjustment). We refer to pathways
759 that have significantly higher Jaccard similarity to the edges of a cluster as being 'encoded by' that
760 cluster. In addition, we tested for enrichment of phenotype-associated modules for each phenotype of
761 interest and each consensus cluster (fast-greedy labels) separately using a Fisher's exact test followed
762 by Bonferroni correction. The full results of this analysis are included as File S3.

763

764 **Tissue consensus gene sets**

765 In order to understand how the immune and fibrotic responses in these phenotypes are
766 functionally related, we found the consensus genes in the combined 4A and 4B clusters. Tissue
767 consensus gene sets were derived by considering all modules within 4A and 4B, finding their unions

768 within their dataset, and then computing their intersection across datasets from the same tissue of
769 origin. For example, the lung consensus gene set was derived by computing the union of the
770 Christmann and Bostwick modules separately, and then computing the intersection across these two
771 datasets.

772 As each tissue was considered separately (limited skin and diffuse skin were considered two
773 different tissues), 5 tissue consensus gene sets were generated; the union of these tissue consensus
774 datasets was used to query the functional genomic networks and is referred to as the 'SSc disease
775 consensus' gene set or genes throughout the text. For all genes in modules in clusters 4A and 4B, we
776 calculated the Pearson correlation to their respective module eigengene (kME). We compared the kME
777 of consensus genes to that of non-consensus genes using a Mann-Whitney U test. Table S2 contains
778 the tissue consensus genes from 4AB or the 'SSc disease consensus genes.'

779

780 **Querying GIANT functional networks, single tissue network analysis, and network visualization**

781 The GIANT functional genomic networks were obtained as binary (.dab) files from CSG. The
782 Dat2Graph function in the Sleipnir library for computational functional genomics (70) was used to query
783 the networks using the SSc disease consensus gene sets (as Entrez IDs). Edge weights were rescaled
784 to match the web prior (0.1) and low probability (< 0.5) edges were pruned. Networks are now available
785 for download from the GIANT webserver (giant.princeton.edu) (1). For each single tissue analysis
786 (consensus lung and consensus skin networks), we considered only the largest connected component
787 and then performed spin-glass community detection as implemented in the igraph R package (69) to
788 obtain the functional modules. We considered previously identified SSc-associated SNPs curated from
789 the literature as reported in Mahoney, *et al.* (24). All networks in this work were visualized using Gephi
790 (71). The network layout is determined by community membership, the strength of connections
791 between communities, and finally the interactions between individual genes. Because of the density of
792 the skin networks, we utilized locally adaptive network scaling as implemented in GRETNA (obtained
793 from github December 2014; https://github.com/sandywang/GRETNA) to prune edges for visualization
794 purposes ($\alpha = 0.05$) (72). Node attribute files and edge lists for all functional genomic networks in this
795 study may be found in File S4.

796 We used exponentiated entropy to measure the 'bridgeness' of a gene (73). Exponentiated
797 entropy is a community-aware measure that allows us to identify genes connected to multiple
798 communities or functional modules. For instance, if a gene has a third of its edges each connecting to 3
799 communities, it will have an exponentiated entropy of 3, or 1 if it is completely connected to its own
800 community. The bridgeness of a vertex *v* is calculated

801 $$bridgeness_v = e^{-\sum_{i=1}^{n} x_i \log(x_i)}$$

802 where x_i represents the fraction of all edges connected to *v* that are connected to nodes in the
803 community *i*. We consider all genes with a bridgeness > 1.5 to be bridges. For our analysis of risk
804 genes in the skin network, we asked if genes with putative SSc-risk polymorphisms were more likely to
805 be in the 90th percentile (bridgeness > 3.05) than genes in the network that are not known to contain
806 risk polymorphisms (only considering genes with degree > 1) using a Fisher's exact test.
807

808 **Differential network analysis**

809 In order to perform differential network analysis (Fig. 4), we retained only nodes common to
810 consensus skin network and consensus lung largest connected component (see above). We then
811 subtracted global edge weight from each network and set all negative edges to zero, which resulted in
812 highly tissue-specific networks. To generate a differential network, we subtracted skin adjacency matrix
813 from the lung adjacency matrix. The analysis of the lung differential network was performed by
814 considering only the largest connected component of the network comprised solely of positive edges
815 (highly lung-specific) followed by spin-glass community detection (igraph).
816

817 **Cell type-specific network analysis**

818 GIANT cell type networks, like tissue networks, were queried with the SSc disease consensus
819 genes and rescaled to a common prior (0.1). We pruned low probability edges (< 0.5), subtracted
820 global (tissue naive) edge weights, and set negative edges to zero. We refer to these as 'cell type-
821 specific networks' as they are highly cell type-specific after global edges have been subtracted. For the
822 cell type network comparison to skin (Fig. 6), we retained only nodes common to all networks (global,

823 skin, keratinocyte, macrophage, and skin fibroblast) after pruning of low probability edge weights and
824 global subtraction in the case of the cell type networks. Then, we performed spin-glass community
825 detection in igraph on the skin network (no global subtraction) and used functional module labels (in
826 skin) to order the networks. We summed the weight within the on-diagonal and off-diagonal 'blocks'
827 (within and between functional modules/communities, respectively) in each cell type-specific network
828 and divided each by the sum of the edge weights in respective network (to obtain the proportion of
829 weight). We performed 10,000 randomizations of the adjacency matrix and computed the block
830 proportions to test the significance of the true block proportions; p-values below 0.001 were considered
831 significant. The cell-type pruned skin network (Fig. 7) was obtained by first identifying the largest
832 connected component in consensus skin network and removing all nodes that were not present in at
833 least one cell type-specific network after pruning low probability edges and global subtraction. Cell type
834 centrality was calculated by computing average quantile-normalized degree of a node's neighbors'
835 maximum quantile-normalized degree in a cell type-specific network and testing significance through
836 permutation testing; we performed 5000 permutations, randomly selecting *n* nodes where *n* is equal to
837 the node's degree in the cell type-pruned skin network. The cell type proportion matrix was clustered by
838 using 1-Earth Mover's Distance between genes as the adjacency matrix and performing fast-greedy
839 community detection iteratively in igraph.
840
841 **Differential expression and single sample gene set enrichment analysis**
842 In order to identify genes that were differentially expressed in SSc-PF, SSc-PF samples were
843 compared to normal controls in both datasets using SAM (1000 permutations, as implemented in samr).
844 Genes with an FDR < 5% were considered further. For the MMF response genes in the Hinchcliff
845 dataset, base and 12 month biopsies of patients that had a decrease in MRSS of at least 5 at 12
846 months were compared (see (74)) (39). We considered all genes in the collapsed dataset and
847 performed paired t-tests as implemented in Comparative Marker Selection in GenePattern. Genes with
848 an FDR < 5% were extracted. For the cell type enrichment analysis in early SSc-PF ('Christmann'
849 dataset), single sample gene set enrichment analysis (ssGSEA) was implemented in R, using the
850 weighting scheme as described in Bolen *et al*. (75). The enrichment scores that are the output of this

851 procedure are correlated with Counter count measures of blood cells (75). The gene sets used as input
852 to ssGSEA were WGCNA modules taken from a study of human macrophage transcriptomes (19).

Acknowledgements

Funding:

This work has been supported by grants from: the Scleroderma Research Foundation (SRF www.srfcure.org) to MLW and MH; grants from the Scleroderma Foundation to PAP and MLW; the Dr. Ralph and Marian Falk Medical Research Trust Catalyst Award to MLW; Dartmouth SYNERGY award to MLW; the Gordon and Betty Moore Foundation (GBMF4552) to CSG; National Institutes of Health (NIH; www.nih.gov) grants 2R01AR051089 to RL, P50 AR060780 and P30AR061271 to RL and MLW, K23 AR059763 to MH. JNT received support from NIH-NIGMS T32GM008704 and the John H. Copenhaver, Jr. and William H. Thomas, MD 1952 Junior Fellowship from Dartmouth Graduate Studies. CPD received support from the EULAR ODP programme. Funding agencies have had no role in the interpretation of the data or in writing the manuscript.

Author contribution

JNT, JMM, and MLW conceived of the study. JNT, CSG, VM, JMM, and MLW designed data analyses, performed analyses, and interpreted the results. TAW performed the microarray experiments. RBC, HWF, RAL, and CPD designed study cohorts included in this work and contributed samples and/or data. MEH provided clinical expertise and interpreted the results. PAP provided macrophage biology expertise and interpreted the results. JNT, PAP, JMM, and MLW wrote the paper. All authors read, revised, and approved the manuscript.

Conflicts of interest

CPD has been a consultant to Roche, GlaxoSmithKline, Actelion, Inventiva, CSL Behring, Takeda, Merck-Serono, MedImmune and Biogen. MLW and MH have filed patents for gene expression biomarkers in systemic sclerosis. MLW is a scientific founder of Celdara Medical LLC. MLW has served as consultant to GlaxoSmithKline, Bristol Myers Squib, EMD Serono, Biogen and Quintiles.

880 **References**

881 1. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, et al. Understanding
882 multicellular function and disease with human tissue-specific networks. Nat Genet 2015, Apr 27.2.
883 Gross AM, Ideker T. Molecular networks in context. Nature Biotechnology 2015;33(7):720-1.
884 3. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations.
885 Nat Methods 2013, Nov;10(11):1108-15.
886 4. Chen S, Wang Q, Wu Z, Li Y, Li P, Sun F, et al. Genetic association study of TNFAIP3, IFIH1, IRF5
887 polymorphisms with polymyositis/dermatomyositis in chinese han population. PLoS One
888 2014;9(10):e110044.
889 5. Chen JC, Cerise JE, Jabbari A, Clynes R, Christiano AM. Master regulators of infiltrate recruitment in
890 autoimmune disease identified through network-based molecular deconvolution. Cell Syst 2015, Nov
891 25;1(5):326-37.
892 6. Sharma A, Menche J, Huang C, Ort T, Zhou X, Kitsak M, et al. A disease module in the interactome
893 explains disease heterogeneity, drug response and captures novel pathways and genes. Human
894 Molecular Genetics 2015:ddv001.
895 7. Assassi S, Wu M, Tan FK, Chang J, Graham TA, Furst DE, et al. Skin gene expression correlates of
896 severity of interstitial lung disease in systemic sclerosis. Arthritis Rheum 2013, Nov;65(11):2917-27.
897 8. Joseph CG, Darrah E, Shah AA, Skora AD, Casciola-Rosen LA, Wigley FM, et al. Association of the
898 autoimmune disease scleroderma with an immunologic response to cancer. Science 2014, Jan
899 10;343(6167):152-7.
900 9. Farina A, Cirone M, York M, Lenna S, Padilla C, McLaughlin S, et al. Epstein-Barr virus infection
901 induces aberrant TLR activation pathway and fibroblast-myofibroblast conversion in scleroderma. J
902 Invest Dermatol 2014, Apr;134(4):954-64.
903 10. Arron ST, Dimon MT, Li Z, Johnson ME, A Wood T, Feeney L, et al. High rhodotorula sequences in
904 skin transcriptome of patients with diffuse systemic sclerosis. J Invest Dermatol 2014, Aug;134(8):2138-
905 45.
906 11. Milano A, Pendergrass SA, Sargent JL, George LK, McCalmont TH, Connolly MK, Whitfield ML.
907 Molecular subsets in the gene expression signatures of scleroderma skin. PLoS One 2008;3(7):e2696.
908 12. Pendergrass SA, Lemaire R, Francis IP, Mahoney JM, Lafyatis R, Whitfield ML. Intrinsic gene
909 expression subsets of diffuse cutaneous systemic sclerosis are stable in serial skin biopsies. J Invest
910 Dermatol 2012, May;132(5):1363-73.
911 13. Hinchcliff M, Huang CC, Wood TA, Matthew Mahoney J, Martyanov V, Bhattacharyya S, et al.
912 Molecular signatures in skin associated with clinical improvement during mycophenolate treatment in
913 systemic sclerosis. J Invest Dermatol 2013, Aug;133(8):1979-89.
914 14. Pendergrass SA, Hayes E, Farina G, Lemaire R, Farber HW, Whitfield ML, Lafyatis R. Limited
915 systemic sclerosis patients with pulmonary arterial hypertension show biomarkers of inflammation and
916 vascular injury. PLoS One 2010;5(8):e12106.
917 15. Risbano MG, Meadows CA, Coldren CD, Jenkins TJ, Edwards MG, Collier D, et al. Altered immune
918 phenotype in peripheral blood cells of patients with scleroderma-associated pulmonary hypertension.
919 Clin Transl Sci 2010, Oct;3(5):210-8.
920 16. Taroni JN, Martyanov V, Huang C, Mahoney JM, Hirano I, Shetuni B, et al. Molecular
921 characterization of systemic sclerosis esophageal pathology identifies inflammatory and proliferative
922 signatures. Arthritis Res Ther 2015, Jul 29;17(1).
923 17. Hsu E, Shi H, Jordan RM, Lyons-Weiler J, Pilewski JM, Feghali-Bostwick CA. Lung tissues in
924 patients with systemic sclerosis have gene expression patterns unique to pulmonary fibrosis and
925 pulmonary hypertension. Arthritis Rheum 2011, Mar;63(3):783-94.
926 18. Christmann RB, Sampaio-Barros P, Stifano G, Borges CL, de Carvalho CR, Kairalla R, et al.
927 Association of interferon- and transforming growth factor β-regulated genes and macrophage activation
928 with systemic sclerosis-related progressive lung fibrosis. Arthritis Rheumatol 2014, Mar;66(3):714-25.
929 19. Xue J, Schmidt SV, Sander J, Draffehn A, Krebs W, Quester I, et al. Transcriptome-based network
930 analysis reveals a spectrum model of human macrophage activation. Immunity 2014, Feb 20;40(2):274-
931 88.

932 20. Amit I, Winter DR, Jung S. The role of the local environment and epigenetics in shaping
933 macrophage identity and their effect on tissue homeostasis. Nat Immunol 2015, Dec 17;17(1):18-25.
934 21. Okabe Y, Medzhitov R. Tissue biology perspective on macrophages. Nat Immunol 2015, Dec
935 17;17(1):9-17.
936 22. Assassi S, Swindell WR, Wu M, Tan FD, Khanna D, Furst DE, et al. Dissecting the heterogeneity of
937 skin gene expression patterns in systemic sclerosis. Arthritis Rheumatol 2015, Aug 3.
938 23. Kidd BA, Wroblewska A, Boland MR, Agudo J, Merad M, Tatonetti NP, et al. Mapping the effects of
939 drugs on the immune system. Nat Biotechnol 2016, Jan;34(1):47-54.
940 24. Mahoney JM, Taroni J, Martyanov V, Wood TA, Greene CS, Pioli PA, et al. Systems level analysis
941 of systemic sclerosis shows a network of immune and profibrotic pathways connected with genetic
942 polymorphisms. PLoS Comput Biol 2015;11(1):e1004005.
943 25. Dudley JT, Tibshirani R, Deshpande T, Butte AJ. Disease signatures are robust across tissues and
944 experiments. Mol Syst Biol 2009;5:307.
945 26. Whitfield ML, Finlay DR, Murray JI, Troyanskaya OG, Chi JT, Pergamenschikov A, et al. Systemic
946 and cell type-specific gene expression patterns in scleroderma skin. Proc Natl Acad Sci U S A 2003,
947 Oct 14;100(21):12319-24.
948 27. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set
949 enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.
950 Proc Natl Acad Sci U S A 2005, Oct 25;102(43):15545-50.
951 28. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing
952 radiation response. Proc Natl Acad Sci U S A 2001, Apr 24;98(9):5116-21.
953 29. Martinez FO, Helming L, Milde R, Varin A, Melgert BN, Draijer C, et al. Genetic programs
954 expressed in resting and IL-4 alternatively activated mouse and human macrophages: Similarities and
955 differences. Blood 2013, Feb 28;121(9):e57-69.
956 30. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, et al. Towards a
957 knowledge-based human protein atlas. Nature Biotechnology 2010;28(12):1248-50.
958 31. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: A network-based approach to human
959 disease. Nature Reviews Genetics 2011;12(1):56-68.
960 32. Hannaford J, Guo H, Chen X. Involvement of cathepsins B and L in inflammation and cholesterol
961 trafficking protein NPC2 secretion in macrophages. Obesity (Silver Spring) 2013, Aug;21(8):1586-95.
962 33. Li Y, Jiang D, Liang J, Meltzer EB, Gray A, Miura R, et al. Severe lung fibrosis requires an invasive
963 fibroblast phenotype regulated by hyaluronan and CD44. J Exp Med 2011, Jul 4;208(7):1459-71.
964 34. Grove LM, Southern BD, Jin TH, White KE, Paruchuri S, Harel E, et al. Urokinase-type
965 plasminogen activator receptor (upar) ligation induces a raft-localized integrin signaling switch that
966 mediates the hypermotile phenotype of fibrotic fibroblasts. Journal of Biological Chemistry 2014,
967 May;289(18):12791-804.
968 35. Greenblatt MB, Sargent JL, Farina G, Tsang K, Lafyatis R, Glimcher LH, et al. Interspecies
969 comparison of human and murine scleroderma reveals IL-13 and CCL2 as disease subset-specific
970 targets. Am J Pathol 2012, Mar;180(3):1080-94.
971 36. Espinosa A, Dardalhon V, Brauner S, Ambrosi A, Higgs R, Quintana FJ, et al. Loss of the lupus
972 autoantigen ro52/trim21 induces tissue inflammation and systemic autoimmunity by disregulating the il-
973 23-th17 pathway. J Exp Med 2009, Aug 3;206(8):1661-71.
974 37. Yoshimi R, Chang TH, Wang H, Atsumi T, Morse HC, Ozato K. Gene disruption study reveals a
975 nonredundant role for TRIM21/ro52 in nf-kappab-dependent cytokine expression in fibroblasts. J
976 Immunol 2009, Jun 15;182(12):7527-38.
977 38. Yang IV, Luna LG, Cotter J, Talbert J, Leach SM, Kidd R, et al. The peripheral blood transcriptome
978 identifies the presence and extent of disease in idiopathic pulmonary fibrosis. PLoS One
979 2012;7(6):e37708.
980 39. Khanna D, Furst DE, Hays RD, Park GS, Wong WK, Seibold JR, et al. Minimally important
981 difference in diffuse systemic sclerosis: Results from the d-penicillamine study. Ann Rheum Dis 2006,
982 Oct;65(10):1325-9.
983 40. Allison AC, Eugui EM. Mycophenolate mofetil and its mechanisms of action. Immunopharmacology
984 2000, May;47(2-3):85-118.

985 41. Kodama K, Horikoshi M, Toda K, Yamada S, Hara K, Irie J, et al. Expression-based genome-wide
986 association study links the receptor CD44 in adipose tissue with type 2 diabetes. Proc Natl Acad Sci U
987 S A 2012, May 1;109(18):7049-54.
988 42. Mikecz K, Brennan FR, Kim JH, Glant TT. Anti-CD44 treatment abrogates tissue oedema and
989 leukocyte infiltration in murine arthritis. Nat Med 1995, Jun;1(6):558-63.
990 43. Mikecz K, Dennis K, Shi M, Kim JH. Modulation of hyaluronan receptor (CD44) function in vivo in a
991 murine model of rheumatoid arthritis. Arthritis & Rheumatism 1999;42(4):659-68.
992 44. Komura K, Sato S, Fujimoto M, Hasegawa M, Takehara K. Elevated levels of circulating CD44 in
993 patients with systemic sclerosis: Association with a milder subset. Rheumatology (Oxford) 2002,
994 Oct;41(10):1149-54.
995 45. Chakravarty EF, Martyanov V, Fiorentino D, Wood TA, Haddon DJ, Jarrell JA, et al. Gene
996 expression changes reflect clinical response in a placebo-controlled randomized trial of abatacept in
997 patients with diffuse cutaneous systemic sclerosis. Arthritis Res Ther 2015;17:159.
998 46. Romero F, Shah D, Duong M, Penn RB, Fessler MB, Madenspacher J, et al. A pneumocyte-
999 macrophage paracrine lipid axis drives the lung toward fibrosis. Am J Respir Cell Mol Biol 2014, Nov
1000 19.
1001 47. Li J, Fu Q, Cui H, Qu B, Pan W, Shen N, Bao C. Interferon-α priming promotes lipid uptake and
1002 macrophage-derived foam cell formation: A novel link between interferon-α and atherosclerosis in
1003 lupus. Arthritis Rheum 2011, Feb;63(2):492-502.
1004 48. Eitzman DT, McCoy RD, Zheng X, Fay WP, Shen T, Ginsburg D, Simon RH. Bleomycin-induced
1005 pulmonary fibrosis in transgenic mice that either lack or overexpress the murine plasminogen activator
1006 inhibitor-1 gene. Journal of Clinical Investigation 1996;97(1):232.
1007 49. Gu nther A, Lu bke N, Ermert M, Schermuly RT, Weissmann N, Breithecker A, et al. Prevention of
1008 bleomycin-induced lung fibrosis by aerosolization of heparin or urokinase in rabbits. Am J Respir Crit
1009 Care Med 2003;168(11):1358-65.
1010 50. Chambers RC. Abnormal wound healing responses in pulmonary fibrosis: Focus on coagulation
1011 signalling. European Respiratory Review 2008, Dec;17(109):130-7.
1012 51. Horowitz JC, Rogers DS, Simon RH, Sisson TH, Thannickal VJ. Plasminogen activation induced
1013 pericellular fibronectin proteolysis promotes fibroblast apoptosis. Am J Respir Cell Mol Biol 2008,
1014 Jan;38(1):78-87.
1015 52. Moodley YP, Caterina P, Scaffidi AK, Misso NL, Papadimitriou JM, McAnulty RJ, et al. Comparison
1016 of the morphological and biochemical changes in normal human lung fibroblasts and fibroblasts derived
1017 from lungs of patients with idiopathic pulmonary fibrosis during fasl-induced apoptosis. J Pathol 2004,
1018 Apr;202(4):486-95.
1019 53. Uhal BD. The role of apoptosis in pulmonary fibrosis. European Respiratory Review 2008, Dec
1020 1;17(109):138-44.
1021 54. Morimoto K, Janssen WJ, Terada M. Defective efferocytosis by alveolar macrophages in IPF
1022 patients. Respir Med 2012, Dec;106(12):1800-3.
1023 55. Vachon E, Martin R, Plumb J, Kwok V, Vandivier RW, Glogauer M, et al. CD44 is a phagocytic
1024 receptor. Blood 2006, May;107(10):4149-58.
1025 56. Noguera A, Gomez C, Faner R, Cosio B, González-Périz A, Clària J, et al. An investigation of the
1026 resolution of inflammation (catabasis) in COPD. Respir Res 2012;13:101.
1027 57. Okabe Y, Medzhitov R. Tissue-specific signals control reversible program of localization and
1028 functional polarization of macrophages. Cell 2014, May 8;157(4):832-44.
1029 58. Gosselin D, Link VM, Romanoski CE, Fonseca GJ, Eichenfield DZ, Spann NJ, et al. Environment
1030 drives selection and function of enhancers controlling tissue-specific macrophage identities. Cell 2014,
1031 Dec 4;159(6):1327-40.
1032 59. Khanna D, Denton CP, Jahreis A, van Laar JM, Cheng S, Spotswood H, et al. OP0054 safety and
1033 efficacy of subcutaneous tocilizumab in adults with systemic sclerosis: Week 48 data from the
1034 fasscinate trial. Ann Rheum Dis 2015;74(Suppl 2):87-8.
1035 60. Eferl R, Hasselblatt P, Rath M, Popper H, Zenz R, Komnenovic V, et al. Development of pulmonary
1036 fibrosis through a pathway involving the transcription factor fra-2/AP-1. Proc Natl Acad Sci U S A 2008,
1037 Jul 29;105(30):10525-30.

1038 61. Aliprantis AO, Wang J, Fathman JW, Lemaire R, Dorfman DM, Lafyatis R, Glimcher LH.
1039 Transcription factor t-bet regulates skin sclerosis through its function in innate immunity and via IL-13.
1040 Proc Natl Acad Sci U S A 2007, Feb 20;104(8):2827-30.
1041 62. Christmann RB, Hayes E, Pendergrass S, Padilla C, Farina G, Affandi AJ, et al. Interferon and
1042 alternative activation of monocyte/macrophages in systemic sclerosis-associated pulmonary arterial
1043 hypertension. Arthritis Rheum 2011, Jun;63(6):1718-28.
1044 63. Gerber EE, Gallo EM, Fontana SC, Davis EC, Wigley FM, Huso DL, Dietz HC. Integrin-modulating
1045 therapy prevents fibrosis and autoimmunity in mouse models of scleroderma. Nature 2013, Nov
1046 7;503(7474):126-30.
1047 64. McGovern N, Schlitzer A, Gunawan M, Jardine L, Shin A, Poyner E, et al. Human dermal CD14
1048 cells are a transient population of monocyte-derived macrophages. Immunity 2014, Sep 18;41(3):465-
1049 77.
1050 65. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nat Genet
1051 2006;38(5):500-1.
1052 66. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using
1053 empirical bayes methods. Biostatistics 2007, Jan;8(1):118-27.
1054 67. Newman ME. Modularity and community structure in networks. Proceedings of the National
1055 Academy of Sciences 2006;103(23):8577-82.
1056 68. Reichardt J, Bornholdt S. Statistical mechanics of community detection. Physical Review E
1057 2006;74(1):016110.
1058 69. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal,
1059 Complex Systems 2006;1695(5):1-9.
1060 70. Huttenhower C, Schroeder M, Chikina MD, Troyanskaya OG. The sleipnir library for computational
1061 functional genomics. Bioinformatics 2008, Jul 1;24(13):1559-61.
1062 71. Bastian M, Heymann S, Jacomy M. Gephi: An open source software for exploring and manipulating
1063 networks. ICWSM 2009;8:361-2.
1064 72. Foti NJ, Hughes JM, Rockmore DN. Nonparametric sparsification of complex multiscale networks.
1065 PLoS One 2011;6(2):e16431.
1066 73. Nepusz T. [igraph] bridges between clusters; Available from:
1067 https://lists.nongnu.org/archive/html/igraph-help/2014-06/msg00065.html. Accessed 18 October 2014.
1068 74. Johnson ME, Mahoney JM, Taroni J, Sargent JL, Marmarelis E, Wu MR, et al. Experimentally-
1069 derived fibroblast gene signatures identify molecular pathways associated with distinct subsets of
1070 systemic sclerosis patients in three independent cohorts. PLoS One 2015;10(1):e0114017.
1071 75. Bolen CR, Uduman M, Kleinstein SH. Cell subset prediction for blood genomic studies. BMC
1072 Bioinformatics 2011, Jun;12(1):258.
1073 76. Jung KK, Liu XW, Chirco R, Fridman R, Kim HR. Identification of CD63 as a tissue inhibitor of
1074 metalloproteinase-1 interacting cell surface protein. EMBO J 2006, Sep 6;25(17):3934-42.
1075 77. Matsumoto T, Matsumori H, Taki T, Takagi T, Fukuda Y. Infantile gm1-gangliosidosis with marked
1076 manifestation of lungs. Acta Pathol Jpn 1979, Mar;29(2):269-76.
1077 78. Nacu N, Luzina IG, Highsmith K, Lockatell V, Pochetuhen K, Cooper ZA, et al. Macrophages
1078 produce tgf-beta-induced (beta-ig-h3) following ingestion of apoptotic cells and regulate MMP14 levels
1079 and collagen turnover in fibroblasts. J Immunol 2008, Apr 1;180(7):5036-44.
1080 79. Roy N, Deveraux QL, Takahashi R, Salvesen GS, Reed JC. The c-iap-1 and c-iap-2 proteins are
1081 direct inhibitors of specific caspases. EMBO J 1997, Dec 1;16(23):6914-25.
1082 80. Deveraux QL, Roy N, Stennicke HR, Van Arsdale T, Zhou Q, Srinivasula SM, et al. IAPs block
1083 apoptotic events induced by caspase-8 and cytochrome c by direct inhibition of distinct caspases.
1084 EMBO J 1998, Apr 15;17(8):2215-23.
1085 81. Todorovic V, Chen CC, Hay N, Lau LF. The matrix protein CCN1 (CYR61) induces apoptosis in
1086 fibroblasts. J Cell Biol 2005, Nov 7;171(3):559-68.
1087 82. Juric V, Chen C-C, Lau LF. Fas-mediated apoptosis is regulated by the extracellular matrix protein
1088 CCN1 (CYR61) in vitro and in vivo. Molecular and Cellular Biology 2009;29(12):3266-79.
1089 83. Franzen CA, Chen CC, Todorović V, Juric V, Monzon RI, Lau LF. Matrix protein CCN1 is critical for
1090 prostate carcinoma cell proliferation and trail-induced apoptosis. Mol Cancer Res 2009, Jul;7(7):1045-
1091 55.

1092 84. Rössig L, Haendeler J, Hermann C, Malchow P, Urbich C, Zeiher AM, Dimmeler S. Nitric oxide
1093 down-regulates MKP-3 mrna levels: Involvement in endothelial cell protection from apoptosis. J Biol
1094 Chem 2000, Aug 18;275(33):25502-7.
1095 85. Azab NA, Rady HM, Marzouk SA. Elevated serum TRAIL levels in scleroderma patients and its
1096 possible association with pulmonary involvement. Clin Rheumatol 2012, Sep;31(9):1359-64.
1097 86. Prunier C, Howe PH. Disabled-2 (dab2) is required for transforming growth factor beta-induced
1098 epithelial to mesenchymal transition (EMT). J Biol Chem 2005, Apr 29;280(17):17540-8.
1099 87. Penheiter SG, Singh RD, Repellin CE, Wilkes MC, Edens M, Howe PH, et al. Type II transforming
1100 growth factor-beta receptor recycling is dependent upon the clathrin adaptor protein dab2. Mol Biol Cell
1101 2010, Nov 15;21(22):4009-19.
1102 88. Smyth N, Vatansever HS, Murray P, Meyer M, Frie C, Paulsson M, Edgar D. Absence of basement
1103 membranes after targeting the LAMC1 gene results in embryonic lethality due to failure of endoderm
1104 differentiation. J Cell Biol 1999, Jan 11;144(1):151-60.
1105 89. Rice LM, Ziemek J, Stratton EA, McLaughlin SR, Padilla CM, Mathes AL, et al. A longitudinal
1106 biomarker for the extent of skin disease in patients with diffuse cutaneous systemic sclerosis. Arthritis
1107 Rheumatol 2015, Nov;67(11):3004-1

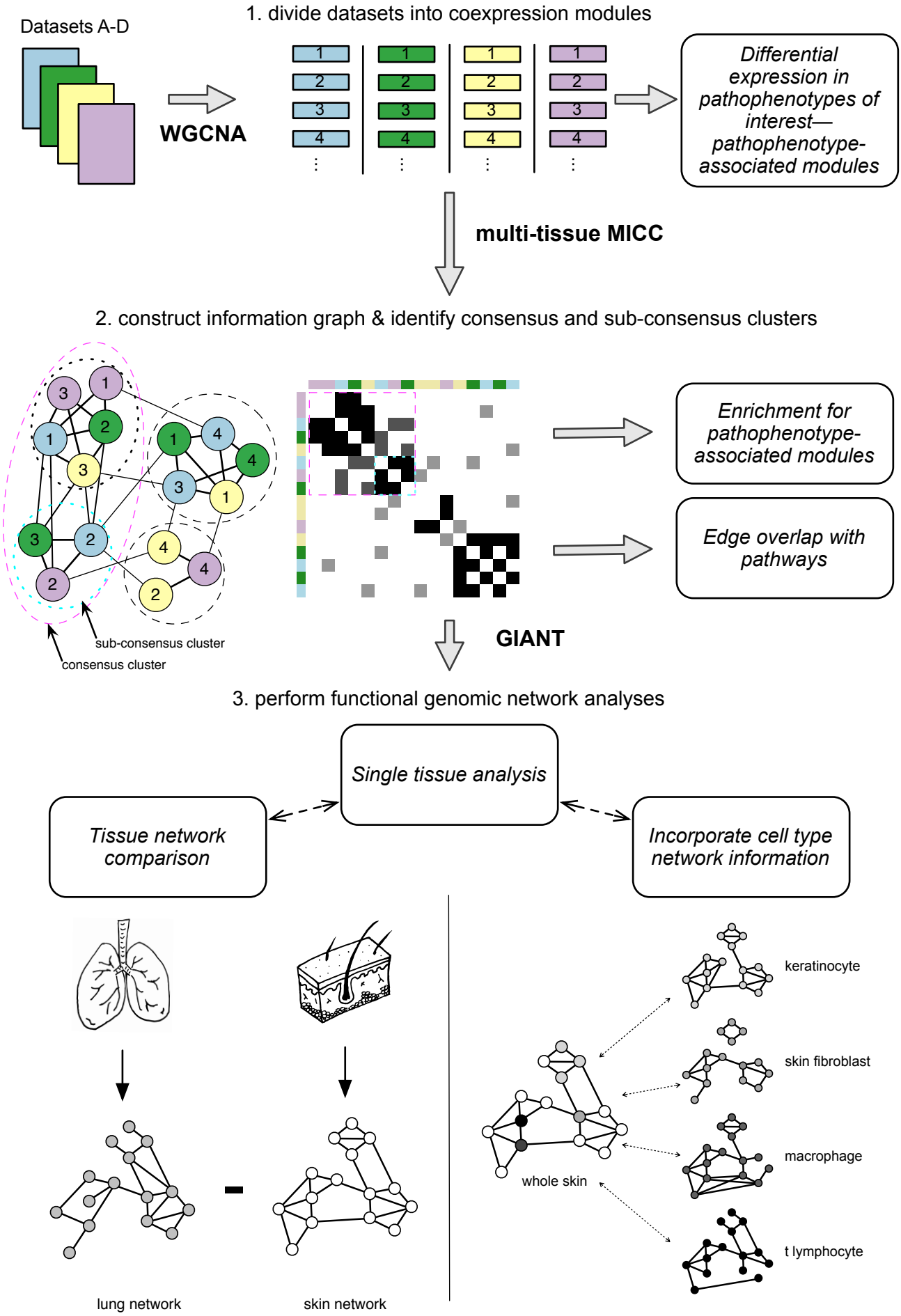
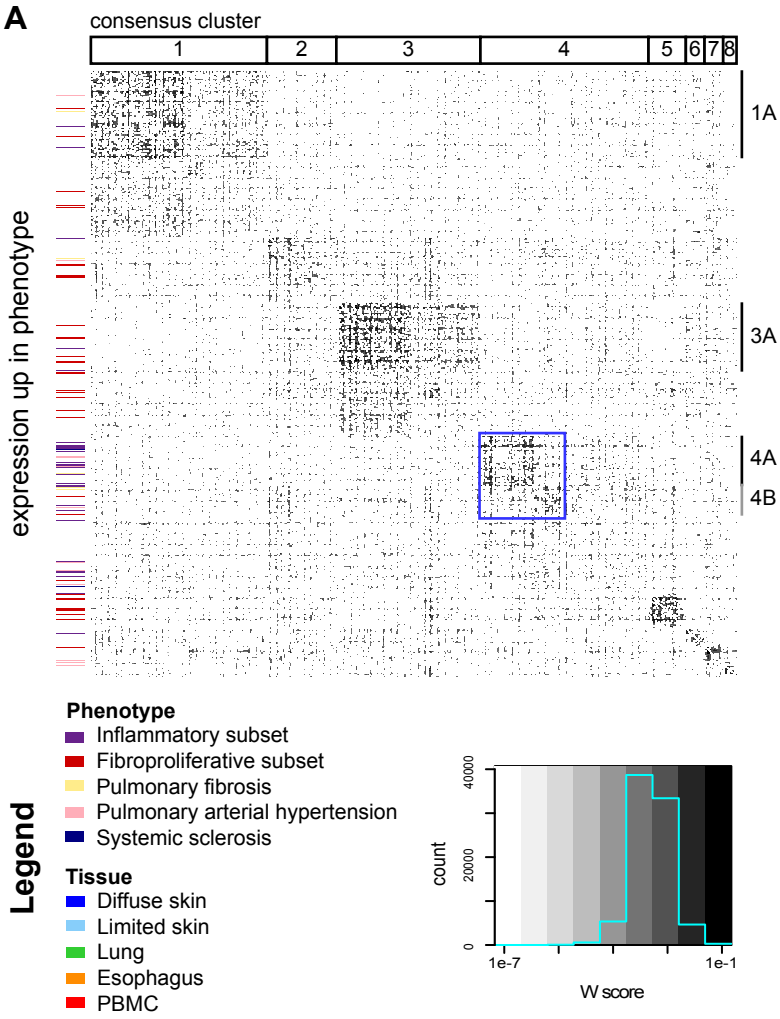


Fig 1. Schematic overview of analysis pipeline. Four datasets are shown for simplicity. Each microarray dataset was partitioned using WGCNA independently to obtain coexpression modules. Module eigengenes were tested for their differential expression in pathophenotypes of interest. Modules were compared across datasets using MICC to form the information graph and community detection algorithms were used to identify consensus or sub-consensus clusters in the information graph. Each cluster was examined for enrichment of pathophenotype-associated modules and edge overlap with canonical biological pathways. Gene sets derived from each cluster were used to query GIANT functional genomic networks. The resulting graphs allow for tissue- and cell type-specific interrogations of the gene sets. Differential network analysis was performed to compare the lung and skin networks. Cell type-specific networks and a novel multi-network approach were used to interpret the skin network results.



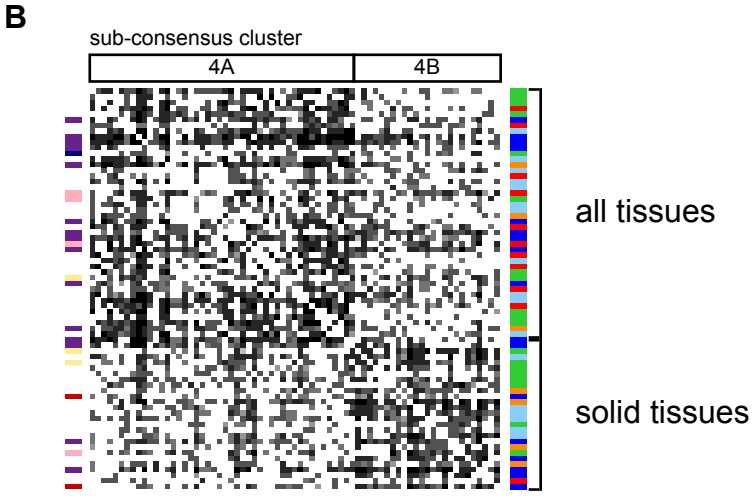


Fig 2. The multi-tissue information graph consensus clusters show that severe SSc-associated pathophenotypes have similar underlying expression patterns.

A. The full adjacency matrix of the information graph sorted to reveal consensus clusters. A darker cell color is indicative of a higher W score or larger edge weight. Consensus clusters and sub-consensus clusters are indicated by the annotation tracks above and on the right side of the matrix, respectively. Coexpression modules with expression that is up in a phenotype of interest are marked by the annotation bar on the left side of the matrix. If a module was up in SSc as well as another pathophenotype of interest, the other pathophenotype is shown.

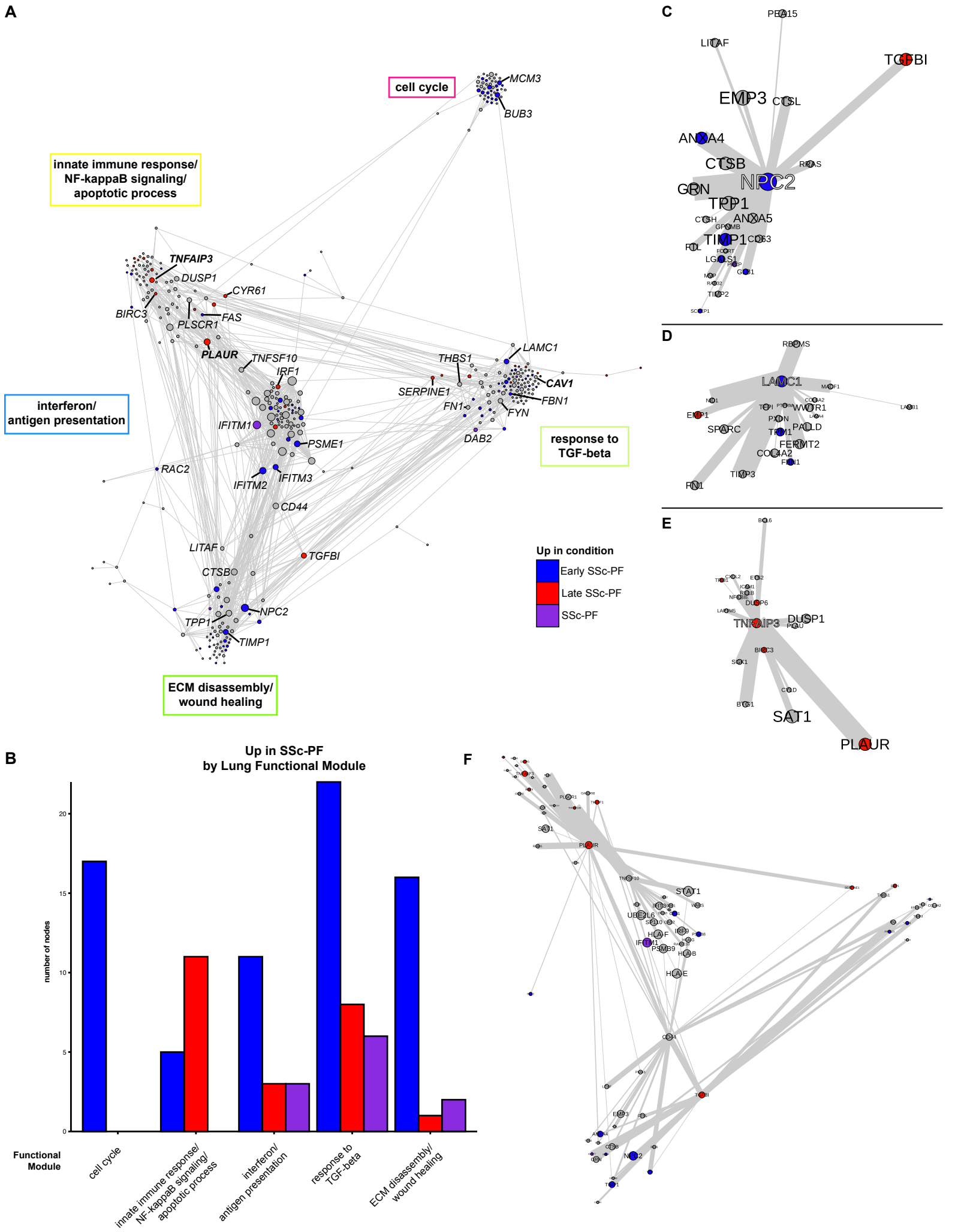


Fig 3. Genes that are overexpressed in late and early SSc-PF are distributed throughout the consensus lung network.

A. The SSc lung network shows functional connections between inflammatory and fibrotic processes. Genes in the giant component were clustered into functional modules using community detection. Biological processes associated with the functional modules are in boxes next to the modules. Genes are colored by whether they are over-expressed in late SSc-PF (red), early SSc-PF (blue), both ("SSc-PF", purple), or neither if they are in grey. Gene symbols in bold have putative SSc risk polymorphisms. Node (gene) size is determined by degree (number of functional interactions) and edge width is determined by the weight (probability of interaction between pairs of genes). The layout is determined by community membership, the strength of connections between communities, and finally the interactions between individual genes in the network. A fully labeled network is supplied as a supplemental figure intended to be viewed digitally (S3).

B. Quantification of differentially expressed genes in each of the five largest functional modules.

C-E. Hubs of the consensus lung network; only the first neighbors of the hub that are in the same functional module are shown. (C) *NPC2* is a hub of the ECM disassembly, wound healing module. (D) *LAMC1* is a hub of the response to TGF-beta module. (E) *TNFAIP3* is a hub of the innate immune response, NF-kappaB signaling, and apoptotic processes module.

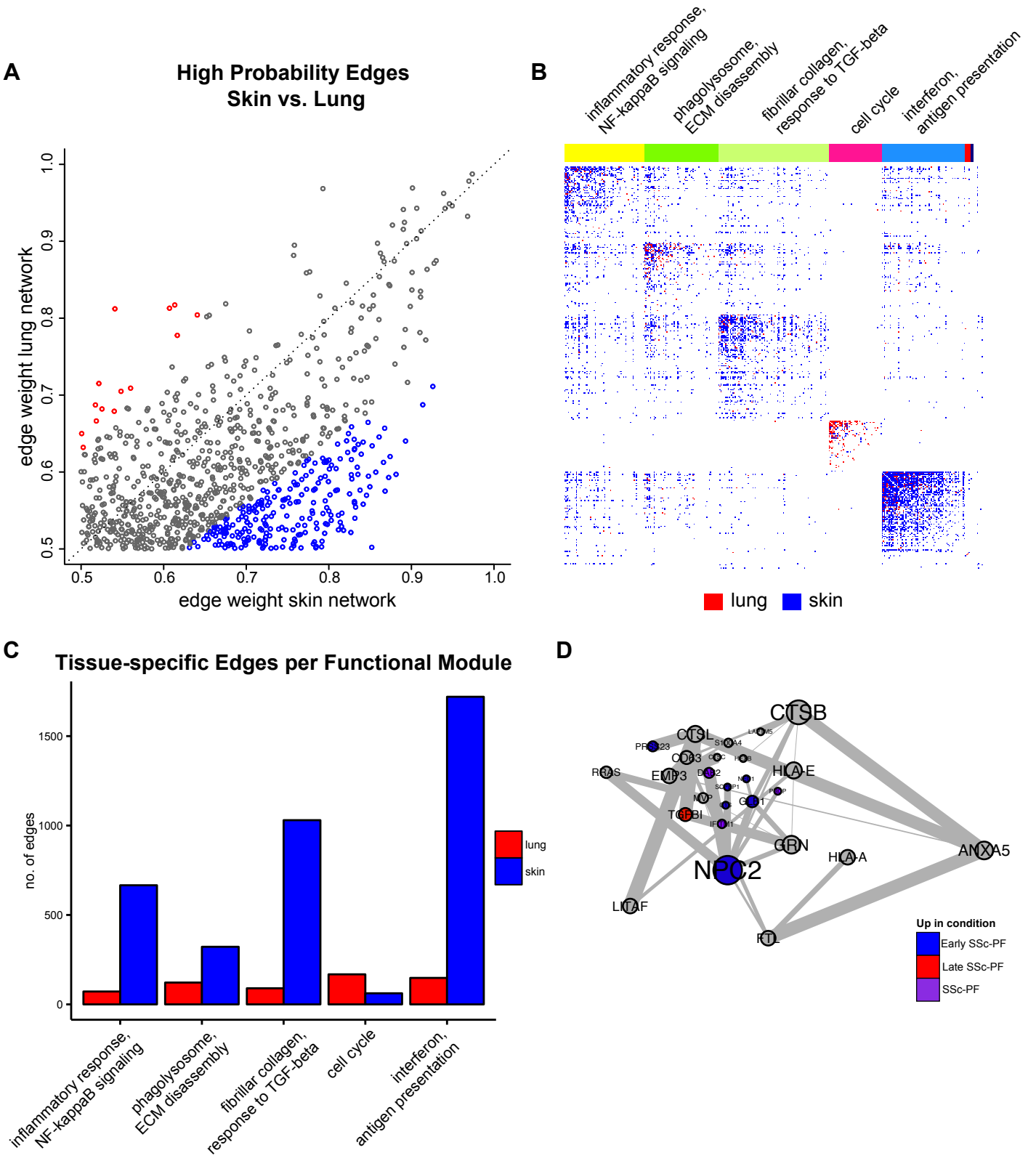


Fig 4. The lung and skin network structures are similar with key differences. The skin and lung networks were compared by first finding the giant component of the lung network and then collapsing to nodes only found in both the skin and lung networks (which are termed the common skin and common lung networks).

A. A scatterplot of high probability edges (> 0.5 in both networks) illustrates that pairs of genes with a higher probability of interacting in skin than lung exist and vice versa. Edges are colored red if the weight (probability) is 1.25x higher in lung or blue if it is 1.25x higher in skin.

B. The differential adjacency matrix where a cell is colored if the edge weight in a given tissue is over and above the weight in the global average and tissue comparator networks. For instance, a cell is red if the edge weight was positive following the successive subtraction of the global average weight and skin weight. Community detection was performed on the common lung network to identify functional modules; common functional modules largely recapitulate modules from the full lung network. Representative processes that modules are annotated to are above the adjacency matrix. The annotation track indicates a genes functional module membership. Nodes (genes) are ordered within their community by common lung within community degree. A fully labeled heatmap is supplied as a supplemental figure intended to be viewed digitally (S4).

C. Quantification of tissue-specific interactions in each of the 5 largest functional modules.

D. The lung-resident macrophage module found in the differential lung network (consists only of edges in red in panel B).

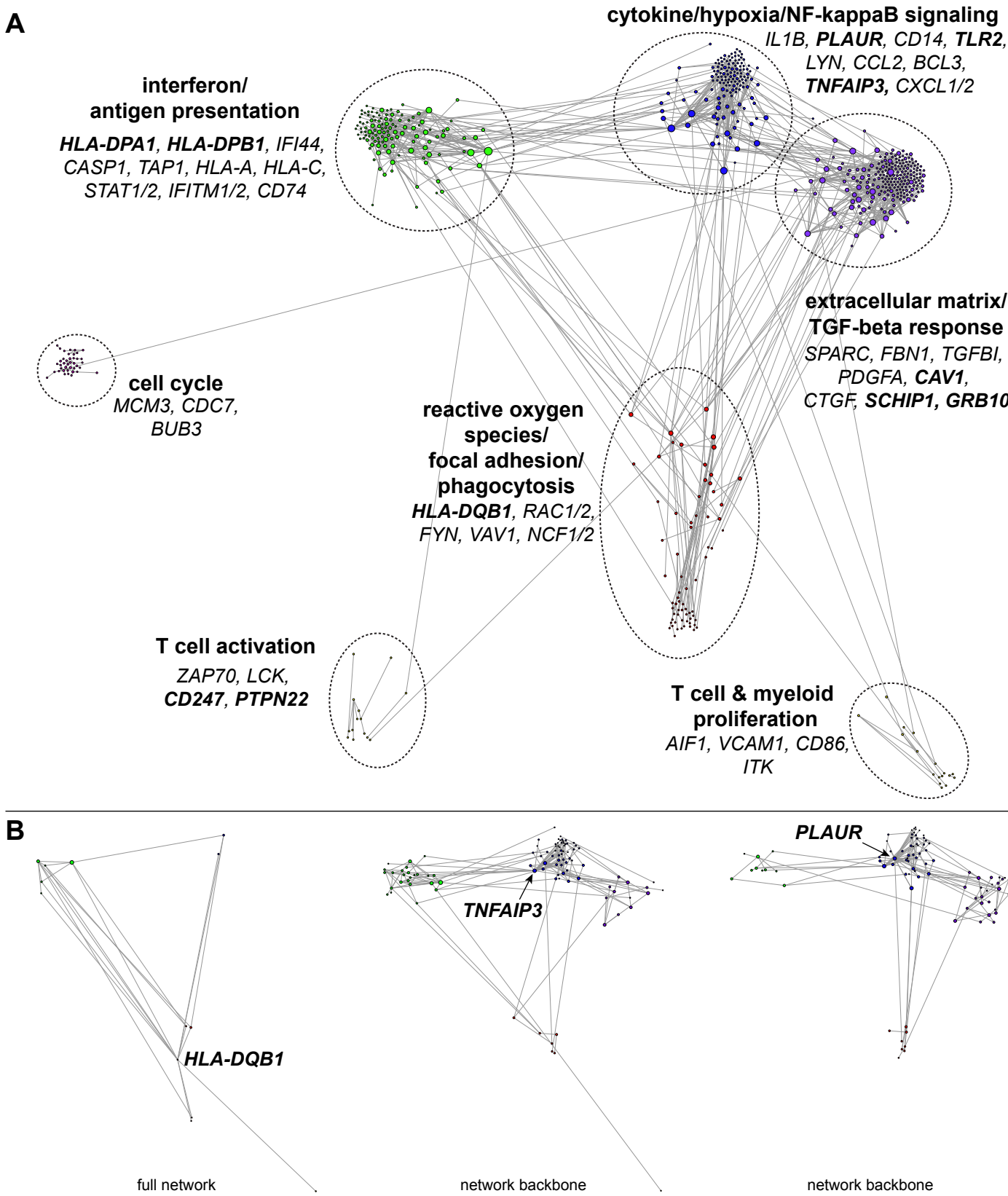


Fig 5. The consensus skin network shows innate and adaptive immunity processes are connected with fibrosis. A fully labeled, full network is supplied as a supplemental figure intended to be viewed digitally (S6).

A. Locally adaptive network sparsification (LANS, α = 0.05) backbone of the skin network. Genes are divided into 7 functional modules by spin-glass community detection. Nodes are sized by degree, colored by functional module (processes in bold), and selected genes are shown. Gene symbols in bold have putative SSc risk polymorphisms. The layout was generated in the same manner as in the lung network using the full network.

B. The first neighbors of genes with putative SSc risk polymorphisms. The genes selected are in the top 10% of nodes' bridgeness and are displayed either with first neighbors in the full network or the backbone as labeled below each network for visualization purposes.

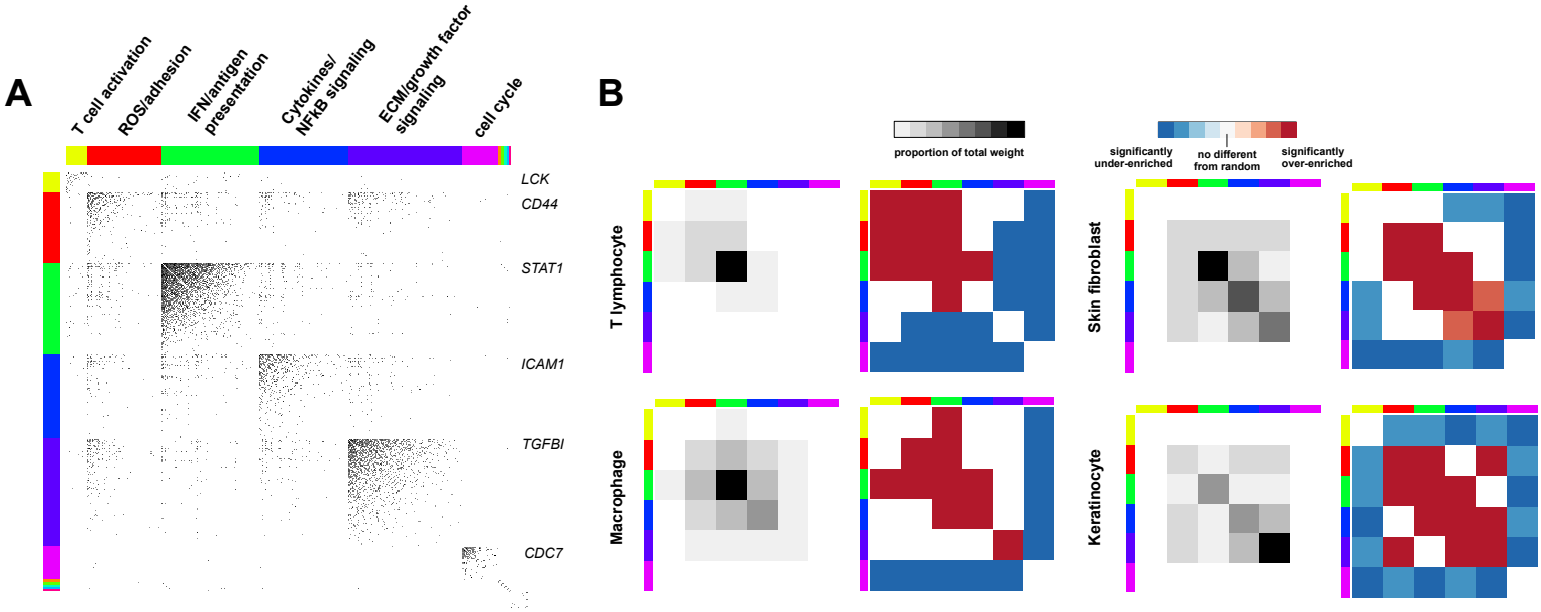


Fig 6. The skin network is composed of cell type-specific network functional modules. Networks were queried with the SSc disease consensus gene sets. Networks were collapsed to nodes that were present in all networks (were connected after pruning edges below 0.5, termed 'common' networks). For cell type-specific networks, the global average (all tissues) edge weights were subtracted off.

A. The common skin network adjacency matrix where nodes are sorted by community and then by within-community degree. The annotation track above the adjacency matrix represents node community membership. Hubs are shown on the right side of the matrix.

B. Compressed adjacency matrices (black and white) and significance matrices (red, white, blue) for cell type-specific networks. Compressed matrices were generated by calculating the total proportion of weight in the on- and off-diagonal blocks of the adjacency matrix in a given cell type-specific network. The darkness of the cell color indicates the total proportion of weight, where darker indicates a higher proportion of weight. Significance matrices represent whether or not the total proportion of weight is significantly different than what we would expect at random, obtained by performing 10,000 permutations of the adjacency matrix and finding the compressed matrix for each permutation (Bonferroni adj. $p < 0.05$). White indicates no different than expected, blue indicates an under-enrichment of weight, and red indicates an over-enrichment.

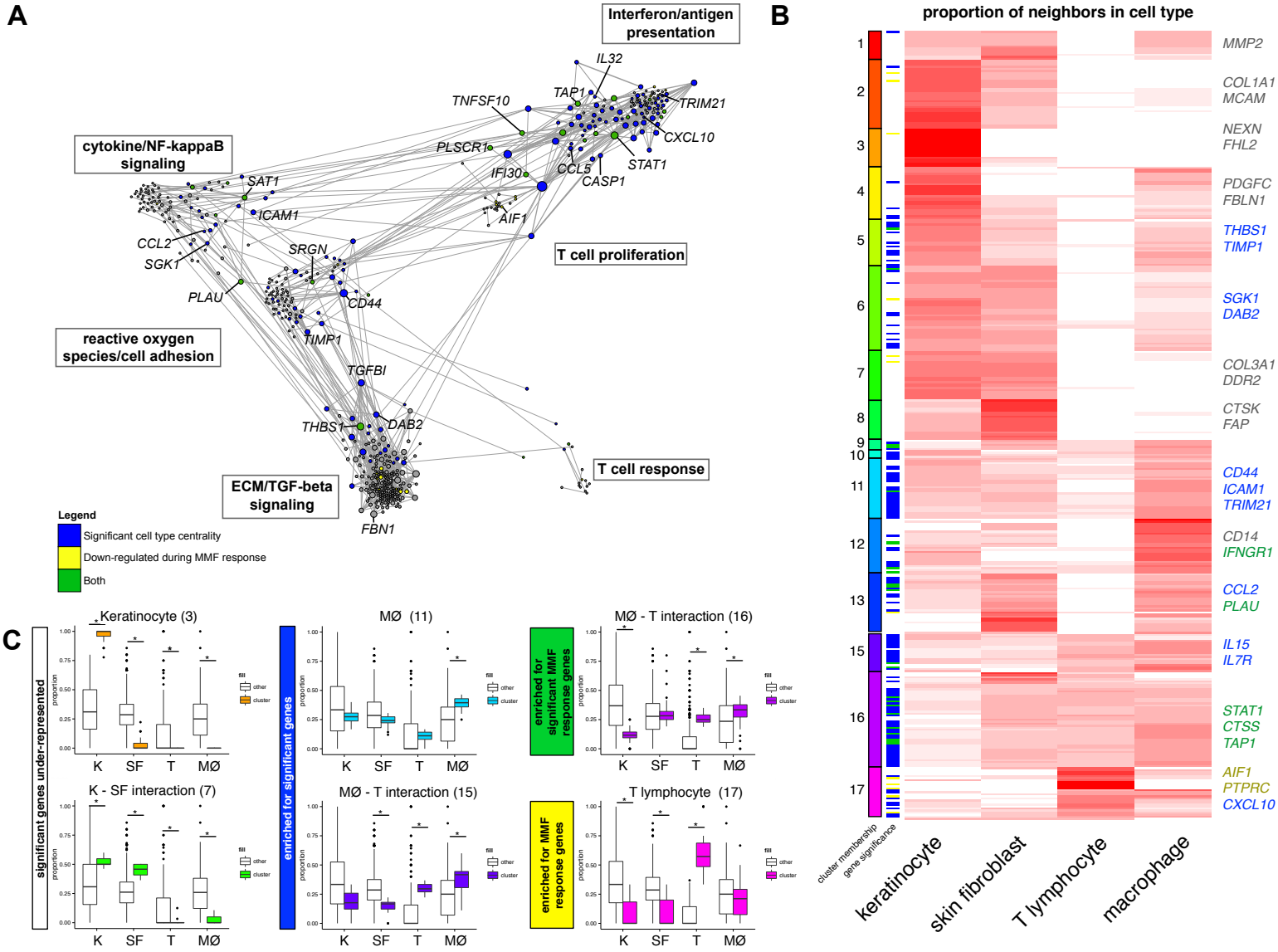


Fig 7. Genes in the skin cell type network with significant cell type centrality are downregulated during clinically significant reduction in skin disease.

A. LANS backbone ($p < 0.05$) of the network is shown for visualization purposes. Genes in the skin giant component were retained only if they were found in at least one of the cell type networks following filtering for high probability (> 0.5) edges. Genes were clustered into functional modules using community detection. Biological processes associated with the functional modules are in boxes next to the modules. Genes that have significant cell type centrality are in blue, genes that are downregulated during MMF response are in yellow, and genes that are in both categories are in green. Node size indicates node degree in the full skin network. A fully-labeled network is supplied as a supplemental figure intended to be viewed digitally (S7).

B. Clustering of genes' neighborhood composition or cell type proportions matrix. A more intense red in the data matrix indicates that the node has a higher proportion of neighbors with max degree in that cell type. The annotation bars on the left of the matrix indicate the cluster membership and whether or not the gene has significant CTC, is an MMF response gene, or both ("gene significance"). Representative genes are listed to the right of the proportion matrix and are colored by significant CTC, MMF response status, or both.

C. Box plots summarizing clusters neighborhood composition in clusters over- or under-enriched for genes with interesting features. A Fisher's exact test (2 x 4 contingency table – 2: in-cluster v. out-of-cluster; 4: non-significant & non-response genes, MMF response genes, significant CTC genes, significant CTC & MMF response response genes) was performed, Bonferroni correction applied; all clusters summarized with boxplots had a corrected $p < 0.05$.

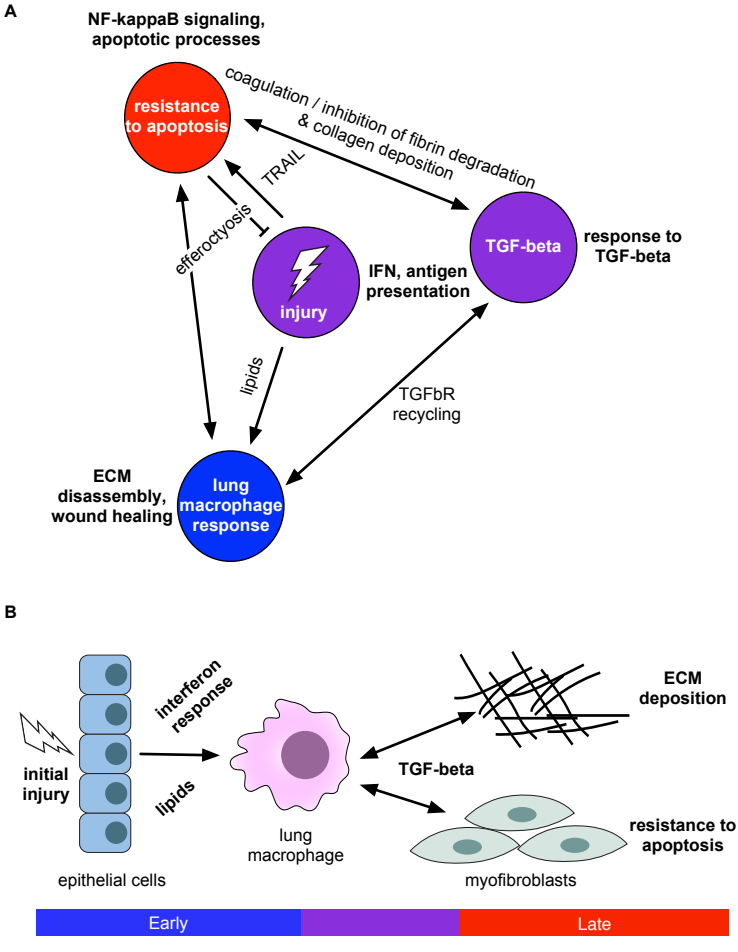


Fig 8. A model of SSc-PF progression.

A. A functional module-based model of SSc-PF progression.

B. A cell type-centric timeline of SSc-PF progression.

Table I. Datasets included in this study.

Dataset label	Tissue	Phenotypes of interest	Citation(s)	GEO Accession
Milano	Diffuse skin	Inflammatory subset, Proliferative subset	Milano et al. (11)	GSE9285
Pendergrass	Diffuse skin	Inflammatory subset, Proliferative subset	Pendergrass et al. (12)	GSE32413
Hinchcliff	Diffuse skin	Inflammatory subset, Proliferative subset	Hinchcliff et al. (13) Mahoney et al. (24)	GSE45485, GSE59785
LSSc	Limited skin	N/A	Present study	GSE76806
UCL	Limited skin	N/A	Present study	GSE76807
Christmann	Lung	SSc-PF	Christmann et al. (18)	GSE76808
Bostwick	Lung	SSc-PF, IPF, IPAH, SSc-PAH	Hsu et al. (17)	GSE48149
Esophagus	Esophagus	Inflammatory subset, Proliferative subset, SSc-PAH	Taroni et al. (16)	GSE68698
PBMC	PBMC	SSc-PAH	Pendergrass et al. (14)	GSE19617
Risbano	PBMC	IPAH, SSc-PAH	Risbano et al. (15)	GSE22356

Abbreviations: PBMC – Peripheral blood mononuclear cells, PAH – Pulmonary arterial hypertension, PF – Pulmonary fibrosis

Table II. Number of arrays and WGCNA coexpression modules in each of the datasets included in this study.

Datasets	Number of Arrays	Number of Coexpression Modules
Milano	75	39
Pendergrass	89	38
Hinchcliff	165	62
LSSc	24	39
UCL	15	98
Christmann	18	56
Bostwick	62	54
Esophagus	33	71
PBMC	54	38
Risbano	38	54

Table III. Selected pathways that are similar to overlapping coexpression patterns in consensus clusters in the information graph. We calculated the Jaccard similarity index between edges in the information graph and canonical pathways and used a Mann-Whitney U test to assess whether a particular pathway was more similar to edges within a consensus cluster than outside the consensus cluster.

Consensus cluster	Summary of selected pathways
1A	DNA repair Cell cycle RNA metabolism Transcription
2	Cell-cell junction organization Aquaporin mediated transport Tight junctions
3A	Endocytosis mRNA processing Metabolism of proteins
4A	T cytotoxic & helper pathway Antigen processing and presentation Allograft rejection
4B	ECM receptor interaction Collagen formation ECM organization TGF-beta signaling Signaling by PDGF
5	G2 M checkpoint Unwinding of DNA Cell cycle
6	Notch signaling Nuclear receptors in lipid metabolism and toxicity
7	Steroid biosynthesis Fatty acid metabolism PPAR signaling pathway
8	Keratin metabolism FGFR ligand binding and activation

Table IV. Selected genes in the consensus lung network.

Functional Module	Gene symbol	Description	Network Position	Up in	Function/Potential Role in Disease
cell cycle	*BUB3*	BUB3 Mitotic Checkpoint Protein	-	Early SSc-PF	Encodes a mitotic cell cycle checkpoint protein that regulates the onset of anaphase.
	CDC7	Cell Division Cycle 7	-	-	Regulates MCM complex.
	MCM3	Minichromosome Maintenance Complex Component 3	-	Early SSc-PF	Subunit of minichromosome maintenance (MCM) complex
	MSH6	MutS Homolog 6	-	Early SSc-PF	Participates in DNA mismatch repair.
ECM disassembly/wound healing	*CD44*	CD44 Molecule (Indian Blood Group)	Bridge	-	A hyaluronic acid receptor that can interact with many other ligands found in the ECM. Primary idiopathic PF fibroblasts exhibit an invasive phenotype that was abrogated with treatment with anti-CD44 (33).
	CD63	CD63 Molecule	-	-	Has been observed to interact with TIMP1 (76)
	CTSB	Cathepsin B	-	-	Regulates NPC2 secretion, TNF-alpha production, and cholesterol trafficking genes in an animal model of obesity (32)
	CTSL	Cathepsin L	-	-	Regulates NPC2 secretion, TNF-alpha production, and cholesterol trafficking genes in an animal model of obesity (32)
	GLB1	Galactosidase, beta 1	-	Early SSc-PF	Mutations in this gene can lead to GM1-gangliosidosis, a manifestation of which includes foam cell accumulation in the lungs (77).
	NPC2	Niemann-Pick disease, type C2	Hub	Early SSc-PF	Mutations in this gene result in a lipid storage disorder. Functions in the regulation of cholesterol trafficking through the lysosome by binding to cholesterol released from low density lipoproteins taken up by cells.
	TGFBI	Transforming Growth Factor, Beta-Induced	Bridge	Late SSc-PF	Induced by phagocytosis of apoptotic debris in monocyte-derived MØs and regulates collagen turnover (78)
	TIMP1	TIMP Metallopeptidase Inhibitor 1	-	Early SSc-PF	Has been observed to interact with CD63 and overexpression has been noted to inhibit apoptosis in a CD63-dependent manner (76)
innate immune response/NFkB signaling/apoptotic process	*BIRC3*	Baculoviral IAP repeat-containing protein 3	-	Late SSc-PF	Has antiapoptotic activity through interactions with caspases as well as the TNF superfamily members TRAF1 and TRAF2 (79,80).
	CYR61	Cysteine-Rich, Angiogenic Inducer, 61		Late SSc-PF	Also known as CCN1. Implicated in apoptosis in fibroblasts (81). Has been shown to play a role in Fas-mediated and TRAIL-induced apoptosis (82,83).
	DUSP6	Dual Specificity Phosphatase 6	-	Late SSc-PF	Plays a role in the positive regulation of apoptosis (84)
	FAS	Fas Cell Surface Death Receptor	-	Early SSc-PF	Cell surface death receptor.
	NFKBIE	Nuclear Factor Of Kappa Light Polypeptide Gene Enhancer In B-Cells Inhibitor, Epsilon	-	-	Negative regulator of NFkB signaling
	PLAUR	Plasminogen Activator, Urokinase Receptor	Bridge	Late SSc-PF	Also known as uPAR. Contains an SSc risk SNP. Pulmonary fibroblasts from patients with idiopathic PF over express uPAR and that uPAR ligation results in a hypermotile phenotype (34).
	PLSCR1	Phospholipid Scramblase 1	-	-	Regulates phospholipid membrane asymmetry.
	TNFAIP3	Tumor Necrosis Factor, Alpha-Induced Protein 3	Hub		Also known as A20. Contains an SSc risk SNP (also associated with other autoimmune conditions). Negative regulator of NFkB signaling.
	TNFSF10	Tumor Necrosis Factor (Ligand) Superfamily, Member 10	Bridge	-	Also known as TRAIL. Elevated in serum of SSc patients (85)
	TNFRSF10B	Tumor Necrosis Factor Receptor Superfamily, Member 10b	-	Late SSc-PF	Also known as TRAILR2.
interferon/antigen presentation	*HLA-E*	Major Histocompatibility Complex, Class I, E	-	-	Class I MHC molecule.
	HLA-F	Major Histocompatibility Complex, Class I, F	-	-	Class I MHC molecule.
	IFITM1	Interferon Induced Transmembrane Protein 1	-	SSc-PF	Interferon signaling.
	IFITM2	Interferon Induced Transmembrane Protein 2	-	Early SSc-PF	Interferon signaling.
	IFITM3	Interferon Induced Transmembrane Protein 3	-	Early SSc-PF	Interferon signaling.
	IRF1	Interferon Regulatory Factor 1	-	Late SSc-PF	Activator of type I interferon signaling.
	OAS1	2'-5'-Oligoadenylate Synthetase 1, 40/46kDa	-	Early SSc-PF	Involved in innate immune response to viral infection.
response to TGF-beta	*CAV1*	Caveolin 1	-	-	Contains an SSc risk SNP.
	CTGF	Connective tissue growth factor	-	-	Also known as CCN2. Has been shown to play a role in Fas-mediated and TRAIL-induced apoptosis (82,83).
	DAB2	Dab, Mitogen-Responsive Phosphoprotein, Homolog 2 (Drosophila)	-	SSc-PF	Required for the epithelial to mesenchymal transition induced by TGF-beta in mouse and for type II TGFbR recycling (86,87)
	FN1	Fibronectin 1	-	-	Extracellular matrix protein.
	LAMC1	Laminin gamma1 chain	Hub	Early SSc-PF	Expression of this gene is essential for the development of basement membranes (88).
	THBS1	Thrombospondin 1	-	-	Mediates cell-to-cell and cell-to-matrix interactions. Putative biomarker of modified Rodnan skin score (89).