Systemic sclerosis (SSc or scleroderma) is a chronic and often severe connective tissue disease of unknown etiology that results in significant morbidity and mortality. No cures or effective treatments are currently available. We conducted a study of twins with SSc. Patients had the diffuse cutaneous or limited cutaneous form of SSc. Our findings show disease concordance in both monozygotic (MZ) and dizygotic (DZ) twins is low (4.7%), suggesting that SSc is a multifactorial disease with genetic predisposition, acquired genetic changes, and environmental factors contributing to disease pathogenesis. In order to identify differentially expressed (DE) genes in our cohort of twins discordant for SSc, we conducted mRNA-seq analysis using total RNA from dermal fibroblasts of the twins. Our analysis identified (DE) genes in twins discordant for diffuse cutaneous SSc and twins discordant for limited cutaneous SSc. We now propose to conduct microRNA-seq (miRNA-seq) to identify differentially expressed miRNAs, the small non-coding RNAs that mediate post-transcriptional regulation of gene expression. We also propose to use novel computational tools to identify hub miRNAs, key miRNA partners, and key biological pathways. Our hypothesis is that a comprehensive expression profile of miRNAs can be identified that differentiates SSc patient dermal fibroblasts from those of healthy twins. We propose to: 1) generate a comprehensive miRNA-seq profile of miRNAs whose expression is deregulated in dermal fibroblasts from twins discordant for SSc and 2) identify hub miRNAs, miRNA partners, and key pathways using a novel computational tool for the miRNA-mRNA network analysis. Our approach will yield novel findings on miRNA-driven regulation of dermal fibrosis in SSc by identifying miRNA-mRNA networks underlying fibrosis in a unique cohort of twins discordant for SSc. Our twin cohort remains the only study of twins with SSc and constitutes a exclusive resource. Findings using this cohort can potentially propel progress in the field and provide new avenues for research and therapies.

Narrative

Systemic sclerosis/scleroderma is a connective tissue disease of unknown cause. We propose to identify networks of messenger RNAs and non-coding micro RNAs that promote the dermal fibrosis characteristic of scleroderma.

SPECIFIC AIMS

Systemic sclerosis (SSc or scleroderma) is a chronic and often severe connective tissue disease of unknown etiology that results in significant morbidity and mortality. No cures or effective treatments are currently available. We recently conducted a study of twins with SSc (1).Our findings show disease concordance in both monozygotic (MZ) and dizygotic (DZ) twins is low (4.7%), suggesting that SSc is a multifactorial disease with genetic predisposition, acquired genetic changes, and environmental factors contributing to disease pathogenesis. Although the gene expression profile of SSc dermal fibroblasts, peripheral blood mononuclear cells, and whole skin biopsies have been reported by us and other groups mainly using traditional microarrays , less is known about the deregulation of the miRNome that affects the fibrotic phenotype of the effector cell in fibrosis, the fibroblast (2). Several studies have examined the expression level of miRNAs in diseased versus healthy patients using miRNA microarray or qPCR (3-7) and determined that individual miRNAs are differently expressed (DE), however miRNA-seq has never been performed in a twin cohort discordant twin for SSc that includes both MZ and DZ twins.

Our preliminary data in early passage dermal fibroblasts cultured from skin punch biopsies of MZ and DZ twins discordant for SSc suggest that several miRNAs are differently expressed in diseased versus healthy twins, emphasizing that perturbation of the miRNome could help differentiate twins discordant for SSc. At the systems biology level, several biological pathways and gene ontology (GO) terms were significantly enriched in dermal fibroblasts of diseased versus healthy twins in both MZ and DZ groups (preliminary studies), and we identified potential key miRNA–mRNA networks using miRmapper to perform a comprehensive network analysis of DE genes and DE miRNAs obtained from total RNA-seq and miRNA qPCR (8). miRmapper is a novel tool for interpretation of miRNA–mRNA interaction networks that measures centrality (for genes and miRNAs) as well as structural equivalence (for miRNAs) to determine potentially critical hubs, miRNAs and genes involved in SSc. We therefore propose to conduct miRNA-seq analysis to comprehensively assess the miRNA expression profile in dermal fibroblasts used for the total RNA-seq studies shown in our preliminary results and identify hub miRNAs, key miRNA partners, and key biological pathways using novel computational tools. Our approach is outlined in **Fig. 1**.

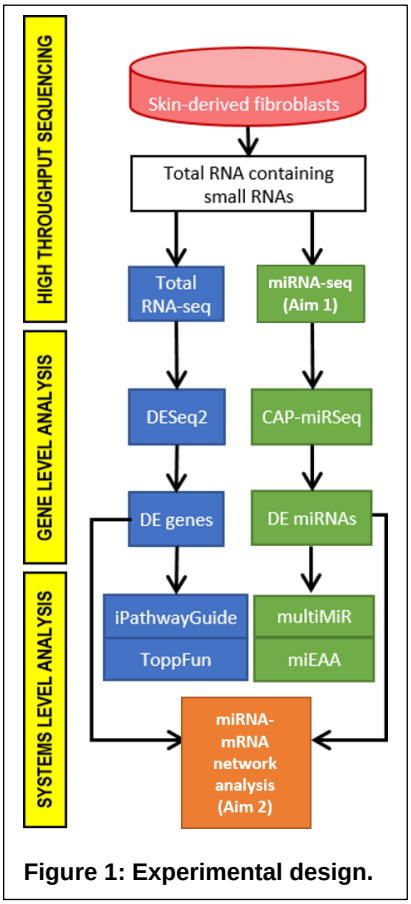

Figure 1: Experimental design.

Our hypothesis is that a comprehensive expression profile of miRNAs can be identified that differentiates SSc patient dermal fibroblasts from those of their healthy twins. To test our hypothesis, we propose the following specific aims:

Specific aim 1: Generate a comprehensive miRNA-seq profile of miRNAs whose expression is deregulated in dermal fibroblasts from twins discordant for SSc.

Specific aim 2: Identify hub miRNAs, miRNA partners, and key biological pathways using a novel computational tool for the miRNA-mRNA network analysis.

Our approach will yield novel findings on miRNA-driven regulation of dermal fibrosis in SSc by identifying miRNA-mRNA networks underlying fibrosis in a unique cohort of twins discordant for SSc. Our approach will also contribute to better understanding of miRNA-mRNA networks in the pathophysiology of the fibrosis characteristic of SSc with relevance for fibrosis characteristic of other organs in patients with SSc and other fibrosing diseases. Our twin cohort remains **the only study of twins with SSc** and constitutes a unique resource. Findings using this cohort can potentially propel progress in the field and provide new avenues for research and identify new targets for therapies. Identification of pathways mediating fibrosis in twins discordant for SSc is paramount to understanding disease etiology and developing preventative and/or therapeutic strategies. **Thus, this project will respond to an unmet need for identification of miRNAs underpinning dermal fibrosis characteristic of SSc in twins discordant for the disease.**

SIGNIFICANCE

In addition to the significance of our findings for SSc, the data generated may also be applicable to other fibrotic diseases that have similar manifestations and will provide insights into mechanisms mediating fibrosis in other organs as well. Insights gained upon completion of our goals will allow us to more accurately assess the role of miRNA-mRNA networks in the development of SSc fibrosis and will provide new avenues for research through the identification of miRNAs and pathways they regulate in unique samples from twins discordant for disease.

INNOVATION

The innovation of the approach lies in the uniqueness of the twin cohort and the fact that our study remains the only one in the world examining disease concordance in twins. We are in a unique position to successfully complete this project due to the availability of the twin cohort. In addition, we will also develop and apply a technically innovative miRNA-mRNA network analysis algorithm. Utilization of this novel tool will facilitate deeper understanding of mRNA-miRNA networks associated with the development of SSc fibrosis.

BACKGROUND

Systemic sclerosis: Systemic sclerosis (SSc or scleroderma) is a chronic and often severe connective tissue disease characterized by immune cell dysregulation, vasculopathy, cutaneous and visceral fibrosis. Patients are commonly classified into two main clinical subsets on the basis of the extent of skin thickening: limited cutaneous SSc (lcSSc) and diffuse cutaneous SSc (dcSSc). The etiology of SSc remains largely unknown. SSc is associated with significant morbidity and mortality with the diffuse form of the disease having a worse prognosis than the limited form (as high as 50% mortality at 10 years) (9). SSc is a multi-factorial disease involving an interplay of multiple genetic, epigenetic, cellular and environmental risk factors (10).

The role of environmental factors and genetics: Environmental factors have been suggested to play a role in SSc development, however, no single trigger has yet been identified. Familial cases are rare (11), suggesting that the disease is not inherited. The gold standard for studying the relative roles of genetics and environment in the development of a disease is the study of twins who are genetically identical (monozygotic, MZ) or fraternal (dizygotic, DZ). Such twins are naturally matched for age and gender. In addition, twins share a common family background and usually share exposure to environmental factors, as well as social and medical variables during childhood and adolescence. Twins provide a unique tool since MZ twins are genetically identical and any differences in phenotype must be caused by environmental factors or somatic mutations occurring after division into two embryos (12). Twin studies in systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), and ankylosing spondylitis (AS) have shown increased concordance for disease in MZ compared with DZ twins (13). Overall, concordance for autoimmune disorders in MZ twins is approximately 25-30% (14). This suggests that autoimmune diseases are probably multigenic with environmental/acquired factors playing a role in disease development.

Disease concordance in twins with SSc: We have reported the results of a cross-sectional analysis of SSc concordance in a cohort of 42 twins, 24 MZ and 18 DZ (1). Participants were examined by an internationally renowned expert on Scleroderma, Thomas A. Medsger, Jr., M.D. for the confirmation of the presence of SSc (index case) and the absence of SSc (healthy co-twin) based on history and physical examination as well as nailfold capillary microscopy. Two twin pairs, a MZ and a DZ pair, have been found to be concordant for disease. Thus, the overall concordance rate for SSc in these twins is 4.7% and is low compared to concordance rates observed SLE and RA twins (13), and in comparison with other diseases in which strong genetic influences are believed to operate, suggesting that environmental factors and/or acquired genetic changes may play an important role in the development of SSc. Thus, our findings suggest that SSc develops in individuals with a susceptible genetic background following environmental triggers/acquired changes. Our preliminary findings thus establish the scientific premise of this proposal. The twins discordant for SSc are a valuable resource for this study and future studies on SSc.

MicroRNAs: microRNAs (miRNA) are small non-coding RNAs that contain 21-22 nucleotides and are evolutionarily conserved, suggesting they have important biological functions. By base-pairing with complementary sequences within mRNAs, miRNAs mediate silencing and post-transcriptional regulation of gene expression and regulate diverse biological processes in health and disease. miRNAs can also cause epigenetic changes including histone modification and DNA methylation of promoter regions (15, 16). Each miRNA targets multiple mRNAs. Deregulation of miRNAs has been implicated in multiple malignant and non-malignant

diseases. **It is our goal to examine the miRNome profile of MZ and DZ twins discordant for SSc to identify DE miRNAs and miRNA-mRNA interaction networks that may contribute to disease pathogenesis and provide new markers for early disease detection. We also propose to identify a causal relationship between risk factors and miRNAs.**

PRELIMINARY STUDIES

We conducted the only study to date of MZ and DZ discordant twins with SSc. Our findings show that disease concordance was low in both fraternal and identical twins (1), suggesting that SSc is likely a multifactorial disease and acquired changes and/or environmental changes may play a role in disease development. In previous studies, we assessed the gene expression profile of dermal fibroblasts from twins discordant for SSc using early microarrays with limited gene representation (17). We recently conducted total RNA-seq to obtain a comprehensive analysis of gene expression changes in early passage dermal fibroblasts from that unique twin cohort. Gene level analysis of the RNA-seq data (q-value $\leq$0.1) revealed several differently expressed (DE) genes between diseased and healthy twins in each category (**Table 1**).

Table 1: Gene and pathway level analysis of RNA-seq data.

Comparisons	Biological pathways	Hit in query list
MZ diffuse versus MZ control	Insulin-like Growth Factor-2 mRNA Binding Proteins (IGF2BPs/IMPs/VICKZs) bind RNA	IGF2BP1,IGF2BP3
	Malaria	COMP,ITGB2,IL6
	Cells and Molecules involved in local acute inflammatory response	ITGB2,IL6
	Signaling by Interleukins	SAA1,IL1RL1,CRLF1,ITGB2,IL6
	Interleukin-6 family signaling	CRLF1,IL6
	Ensemble of genes encoding extracellular matrix and extracellular matrix-associated proteins	HAPLN3,EGLN3,CRLF1,COMP,IL6,EGFL6
	amb2 Integrin signaling	ITGB2,IL6
DZ diffuse versus DZ control	Cell adhesion molecules (CAMs)	CADM3,CNTN1,L1CAM,JAM2
	Drug metabolism - cytochrome P450	GSTT1,ALDH1A3,ADH1B
	Metabolism of xenobiotics by cytochrome P450	GSTT1,ALDH1A3,ADH1B
	Chemical carcinogenesis	GSTT1,ALDH1A3,ADH1B
MZ limited versus MZ control	none	
DZ limited versus DZ control	RA biosynthesis pathway	ALDH1A3,ADH1A
	Triacylglycerol biosynthesis	PLPPR4,GPAT2
	O-glycan biosynthesis, mucin type core	GALNT14,GALNT18
	MAP00350 Tyrosine metabolism	ALDH1A3,ADH1A
	Cargo concentration in the ER	GRIA1,COL7A1
	Tyrosine metabolism	ALDH1A3,ADH1A
	Signaling by FGFR2 fusions	FGFR2

Volcano plots of the DE genes demonstrate that differential gene expression profiles can be identified in all 4 experimental groups (**Fig. 2**). Our preliminary findings were further visualized using heatmaps (**Fig. 3 & 4**) which show that diseased patients cluster together while healthy twin controls cluster in a separate group. This demonstrates that DE genes can be identified in fibroblasts of twins discordant for SSc. It also demonstrates that differences in gene expression can be identified in MZ and DZ twins as well as in twins discordant for the two different forms of systemic disease, limited and diffuse cutaneous SSc.

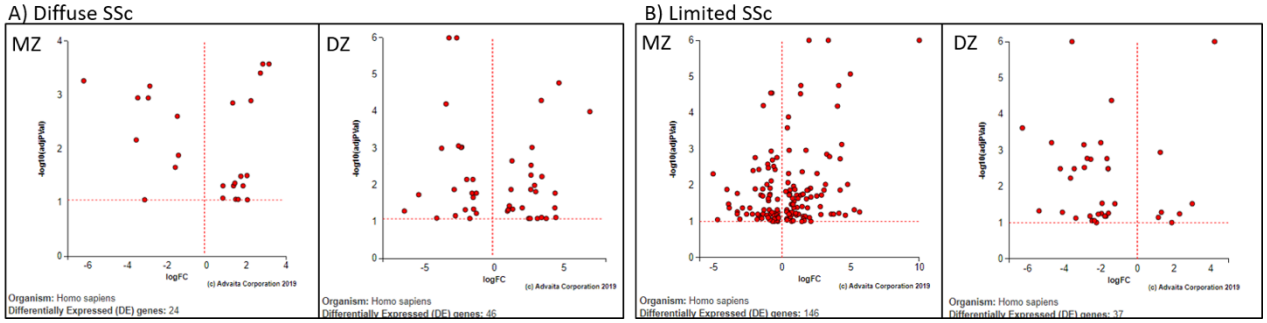

Figure 2: Volcano plots of DE Genes – RNA seq

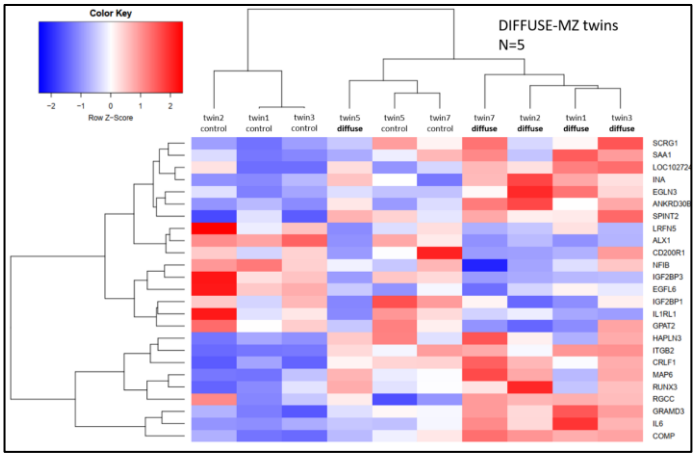

Figure 3: Heatmap of genes DE in fibroblasts from MZ twins discordant for the diffuse cutaneous form of SSc.

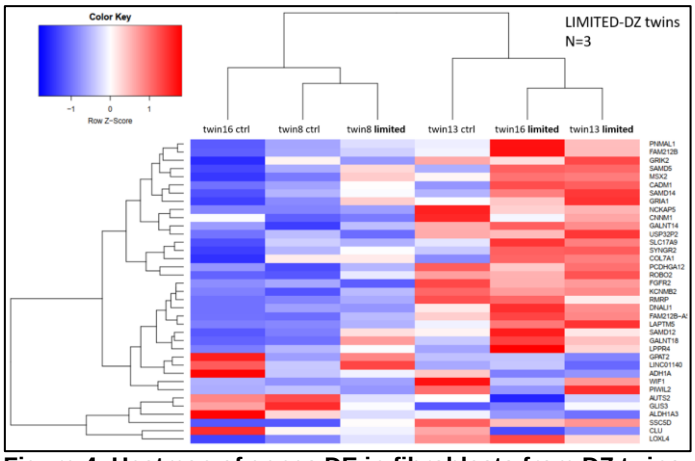

Figure 4: Heatmap of genes DE in fibroblasts from DZ twins discordant for the limited cutaneous form of SSc.

Experience and know-how of developing web interface for interactive visualization of network data. Our team has extensive experience in developing databases and web interfaces for network analyses. Examples include *GAIL* (http://chunglab.io/GAIL/), a database and web interface for gene-gene network analysis, and *DDNet* (http://chunglab.io/ddnet/), a database and web interface for disease-disease network analysis (**Fig. 5**). These web interfaces allow dynamic and interactive visualizations using the *Django* framework based on *D3.js* and *sigma.js* technologies. In addition, we utilize a graph database *Neo4j* as a backend database for these web interfaces and this allows computationally efficient query and search. Our experiences and know-hows obtained from these developments will be critically helpful for development of an interactive visualization software for miRNA-mRNA networks proposed in Aim 2.

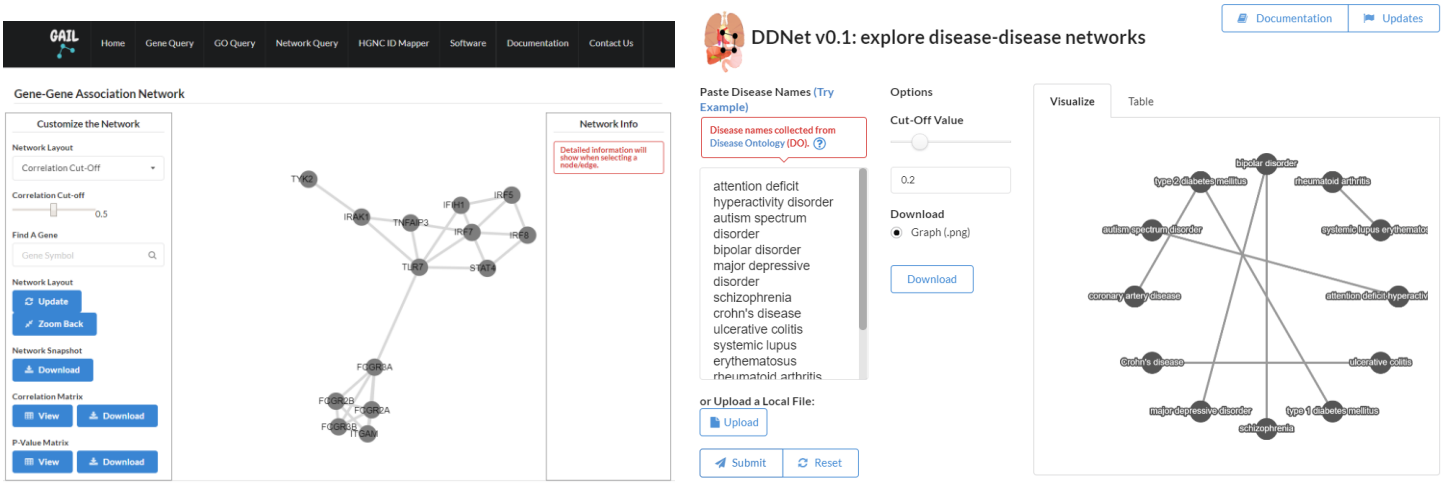

Figure 5. Screenshots of *GAIL* (http://chunglab.io/GAIL/) and *DDNet* (http://chunglab.io/ddnet/), dynamic and interactive web interfaces for inference of the gene-gene association and disease-disease association.

RESEARCH APPROACH

Specific aim 1: Generate a comprehensive miRNA-seq profile of miRNAs whose expression is deregulated in dermal fibroblasts from twins discordant for SSc.

Rationale: We previously performed total RNA-seq (Illumina 2500 sequencer) on RNA extracted from dermal fibroblasts of 16 twin pairs discordant for SSc and more recently miRNA PCR array (Qiagen Human Fibrosis miScript miRNA PCR Array kit) on total RNA containing small RNA extracted from dermal fibroblasts of 3 of the 16 pairs of twins discordant for SSc, and obtained preliminary data using miRmapper as a proof of concept. However, a full miRNA expression profile obtained from miRNA-seq would allow us to conduct a comprehensive analysis of the miRNA-mRNA interaction networks out of the fibroblasts of our twin cohort. The miRNA PCR array we performed only allows specific miRNA quantification of 84 miRNAs known to play a role in fibrosis rather

than a full miRNA signature. In this aim, we will generate a comprehensive miRNA expression profile from miRNA-seq that will be used in aim 2 as input, and conduct a systems level analysis of the DE miRNAs identified.

The data generated from 3 twin pairs was analyzed using miRmapper (8), and two lists of significantly deregulated miRNAs were generated: 22 upregulated and 55 downregulated miRNAs. Using the DEseq2 output file from RNA-seq analysis, we generated 2 lists of genes based on fold change (FC): upregulated (FC>1) and downregulated genes (FC<1), total 2,152 perturbed genes. Predicted interactions (3,587 total) between upregulated miRNAs/downregulated genes and downregulated miRNAs/upregulated genes were obtained on multiMiR (p-value: 10% confidence) and entered in miRmapper along with the identified miRNAs. Amongst the 2,152 deregulated genes between SSc patients and healthy MZ twins, 886 genes are target for at least 1 identified miRNA, and certain genes are targeted by several "collaborative" miRNAs (**Fig. 6A**). These collaborative clusters reflect "miRNAs structural equivalence" based on the number of similar mRNA targets they share. A bar plot output showing the top 20 miRNAs with the largest predicted impact on target genes is shown in **Fig. 6B**. Based on the "centrality" postulate that DE miRNAs modulating a large number of mRNA transcripts ultimately have a greater influence in determining phenotypic outcomes and are more important in a global biological context than miRNAs that modulate just a few mRNA transcripts, our data suggest that hsa-miR-320a, hsa-miR-93-5p and hsa-miR-15b-5p play an important role in fibroblasts since they are predicted to target about 15% of of the 886 DE target genes.

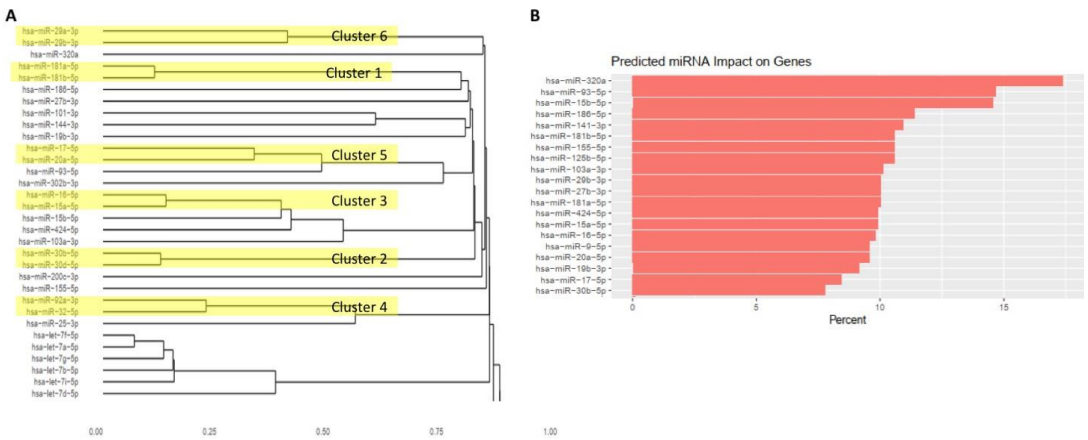

Figure 6: A. Dendrogram generated by miRmapper, showing miRNA collaboration patterns based on how many DE genes they commonly target for silencing. B. miRmapper bar plot output showing predicted miRNA impact on target genes. Percent: proportion of genes amongst the 886 target genes identified that each miRNA is predicted to target for silencing.

Experimental approach: 100-200 ng of total RNA will be used to prepare small RNA-Seq libraries using the QIAseq miRNA Library Kit following the protocol as described by the manufacturer (Qiagen, Germantown, MD). Following library preparation, libraries quality will be determined using Agilent's TapeStation (Agilent, Santa Clara, CA). Libraries will be clustered on the cBot as described by the manufacturer (Illumina, San Diego, CA). Clustered miRNA-seq libraries will be single read sequenced using version 4 with 1X50 cycles on an Illumina HiSeq2500. Demultiplexing will be performed utilizing bcl2fastq-1.8.4 to generate Fastq files for downstream analysis. Small RNA-sequencing will be performed on samples from 16 pairs of twins discordant for SSc.

Data analysis: Systems level analysis will be done using Rosalind (OnRamp), a genomics and bioinformatics analytic platform that streamline CAP-miRSeq pipeline (18), a comprehensive analysis pipeline for microRNA sequencing data, with multiMiR (19), a R package and database that provides integration of microRNA–target interactions along with their disease and drug associations. Functional analysis of DE miRNAs will be done using the miRNA Enrichment Analysis and Annotation Tool (miEAA) (20).

Data interpretation, potential pitfalls and alternative approaches: Both multiMiR and miRmapper are based on the classical paradigm that miRNAs directly target genes for silencing by inhibiting translation, but new evidence supports another hypothesis that the bulk of miRNA effects come from indirectly changing protein levels through large-scale transcriptional regulation (21). Additionally, we may be able to detect few long non-coding RNAs (lncRNAs) but our approach is not designed to specifically investigate lncRNAs and many will not be detected by total RNA-seq. Moreover, we are not addressing methylation here, and our study is limited to transcriptomics.

Specific aim 2: Identify hub miRNAs, miRNA partners, and key biological pathways using a novel computational tool for the miRNA-mRNA network analysis.

Rationale: We recently developed miRmapper, a novel tool to investigate miRNA–mRNA interaction networks using centrality (for genes and miRNAs) as well as structural equivalence (for miRNAs) (8) to determine potentially critical hub miRNAs and genes. Its application to miRNA and mRNA datasets from the human bladder cancer cell lines showed its utility in identification of hub miRNAs, miRNA partners, and key pathways. However, there are still some rooms for improvement in the miRmapper algorithm. First, currently miRmapper does not provide significance measures (e.g., *p*-values or confidence intervals) for its findings. Second, there is no guidance about how to select parameters to determine miRNA partners. Third, its software currently provides only static plots and tables, which do not allow interactive investigation of the results. In this aim, we will further improve the miRmapper algorithm by addressing these limitations, develop a web interface with an interactive visualization of miRNA-mRNA networks, and apply these improved computational tools to the miRNA and mRNA datasets generated from Aim 1, to identify hub miRNAs, miRNA partners, and key pathways.

Input: 1) Predicted targets of each miRNA; 2) a list of differentially expressed (DE) miRNAs; 3) a list of DE mRNAs. Predicted targets of miRNAs can be collected from databases such as microRNA.org, TargetScan, and the multiMiR R Package, while DE miRNAs and mRNAs will be obtained from Aim 1.

Improving target predictions of miRNAs: By considering that the input miRNA-target predictions might not be optimal for a given experiment, we will first improve the miRNA-target predictions by integrating the input miRNA-target predictions with the information from the miRNA and mRNA datasets. Specifically, we will first calculate Pearson correlation for each miRNA-mRNA pair and obtain corresponding *z*-value using Fisher transformation, an approximate variance-stabilizing transformation and follows a normal distribution. Then, we will model the input miRNA target predictions and these z values to be joint emissions of latent binary states of miRNA-mRNA relationships using Bernoulli and standard normal distributions, respectively. We will estimate latent binary states of miRNA-mRNA relationships using the Expectation-Maximization algorithm (22). Finally, we will consider only the miRNA-mRNA associations of which posterior probabilities are larger than 0.5 in the analysis below.

Identification of hub miRNAs: Based on the improved target predictions of miRNAs, along with lists of DE miRNAs and DE mRNAs, two statistics measuring influence of a miRNA will be calculated:

$$T_m = [\ \#\ \text{targets of m-th DE miRNA}\ ] / [\ \#\ \text{union of targets of all the miRNAs}\ ].$$

$$D_m = [\ \#\ \text{targets of m-th DE miRNA among DE genes}\ ] / [\ \#\ \text{all the DE genes}\ ].$$

Then, we will identify hub miRNAs based on $wT_m + (1\text{-}w)D_m$. By default, we will use $w = 0.5$ but also investigate other weighting schemes. Then, to calculate significance, we will implement a permutation approach by randomly assigning target genes for miRNAs and also randomly assigning labels to subjects. Then, we will calculate $wT_m + (1\text{-}w)D_m$ for each permuted data. Finally, we will estimate the *p*-value for each miRNA as the proportion of the permuted data having values larger than the $wT_m + (1\text{-}w)D_m$ value calculated from the original data. After applying the Benjamini-Hochberg correction (23) to these *p*-values, the miRNAs with adjusted *p*-values less than the nominal level will considered as hub miRNAs.

Identification of miRNA partners: We will first generate an adjacency matrix based on Jaccard distance to measure dissimilarity between miRNAs in the sense of shared gene targets, i.e., $J_{ij} = 1 - [\ \#\ \text{intersection of DE mRNAs targeted by i- and j-th DE miRNAs}\ ] / [\ \#\ \text{union of DE mRNAs targeted by i- and j-th DE miRNAs}\ ]$. To calculate significance, we will implement a permutation approach by randomly assigning target genes for miRNAs and also randomly assigning twin labels to subjects. Then, we will calculate J_{ij} for each permuted data. Finally, we estimate the *p*-value for each miRNA pair as the proportion of the permuted data having values smaller than the J_{ij} value calculated from the original data. After applying the Benjamini-Hochberg correction (23) to these *p*-values, the miRNA pairs with adjusted *p*-values less than the nominal level will be considered as miRNA partners.

Identification of hub genes and subnetworks: We will first generate an adjacency matrix based on Jaccard distance to measure dissimilarity between mRNAs in the sense of shared miRNA regulators, i.e., $G_{ij} = 1 - [\ \#\ \text{intersection of DE miRNAs targeting i- and j-th DE mRNAs}\ ] / [\ \#\ \text{union of DE miRNAs targeting i- and j-th DE mRNAs}\ ]$. To calculate significance, we will implement a permutation approach by randomly assigning target genes for miRNAs and also randomly assigning twin labels to subjects. Then, we will calculate G_{ij} for each permuted data. Finally, we estimate the *p*-value for each mRNA pair as the proportion of the permuted data having values smaller than the G_{ij} value calculated from the original data. After applying the Benjamini-Hochberg correction (23) to these *p*-values, we will construct a mRNA network by linking the mRNA pairs of which adjusted

p-values are less than the nominal level. Finally, hub genes and subnetworks will be identified using the Bayesian stochastic block model approach (24), which allows simultaneous identification of hubs and subnetworks.

Web interface for interactive investigation and visualization of miRNA-mRNA network. Although the proposed computational tool improves analysis of the miRNA-mRNA network, results from this analysis still remain challenging to visualize and interpret due to the overwhelming amount of information included within a plot and crowded visualization. In order to address this challenge, we will develop a dynamic and interactive visualization interface based on the powerful *Shiny* technology (https://shiny.rstudio.com/), which provides a web-based interface between *R* and *D3.js*. This interface will allow users to 1) easily change the viewpoint of visualization (i.e., visualize mRNAs from the perspective of miRNAs, and vice versa); 2) control the complexity of visualization (i.e., the number of "dots" (miRNAs or mRNAs) presented in the plot); and 3) quickly check additional information with pop-up information boxes. In addition, this visualization tool will be integrated into a graphical user interface, which will allow non-expert users to upload input data, fit the proposed approach, generate visualization results, control visualization options (e.g., font size, line thickness, and so on), and download visualization results that can be incorporated into a grant proposal or a research article. All the statistical and computational methods developed in Aim 2 will be implemented in software packages in R, the commonly used mathematical and statistical platform, with a user manual with examples of how to use the methods. The *Comprehensive R Archive Network (CRAN)* publishes software packages making them publicly available to download (25). We will validate the R software package by applying standard software validation steps.

Computational validation: We will first evaluate this aim with computational experiments, by simulating miRNA-mRNA network data and evaluating how well the proposed method can recover true hub miRNAs, miRNA partners, hub genes, and gene subnetworks. We will further assess performance of the proposed method using multiple public miRNA and mRNA datasets. Specifically, we will split data by replicate in each setting, apply the proposed method to each subset, and then evaluate reproducibility between subsets.

Data interpretation, potential pitfalls and alternative approaches: First, confidence about miRNA-mRNA relationships can vary across different pairs of miRNA-mRNA. We will address this issue by using confidence measures for interactions between miRNA-mRNA as an ordinal variable rather than a binary variable. Second, computational burdens can increase significantly when large interaction data sets are studied. We will investigate other estimation techniques and parallel computing techniques to improve computational efficiency.

Rigor and reproducibility: Zygosity of the twins included in this study was confirmed using two different methods: DNA fingerprint analysis (1) and analysis of 26 short tandem repeat (STR) autosomal markers. Throughout the stages of study design and data processing, we will conform to reproducible research guidelines to ensure full replication of our results, including:
i) Ensuring all *in vitro*, *in vivo* and public data sets utilized adhere to rigorous design principles, including issues related to randomization, error rate control, bias and confounding, and effect modification;
ii) Considering sex (*in vivo* studies) and gender (human studies) as biological variables with appropriate consideration with respect to study design and data analysis;
iii) Saving all raw and intermediate data files;
iv) Evaluating data QC (e.g., quality of primary data, appropriateness of replicates, batch effects);
v) Avoiding error-prone manual data manipulations (e.g., 'cut and paste');
v) Using literate programming tools such as *Markdown* to promote transparency;
vi) Archiving custom scripts and exact versions of external programs or R libraries used (26);
vii) Recording all initial seeds used in algorithms that rely upon randomness;
viii) Documenting data analysis pipelines using workbench tools such as the *IPython Notebook*, *Roxygen2*, *Revolution Analytics* and *Taverna Workbench Bioinformatics Management System* (27, 28).

Sex as a biological variable: All dermal fibroblasts available are from women with SSc or their healthy twin. This reflects the female predominance in SSc. This project will thus focus on primary fibroblasts from women. Future studies will be designed to extend our findings to dermal fibroblasts from males with SSc. Although such fibroblasts are not available from male twins discordant for SSc, they are available from non-twin male patients with the disease.

Timeline:

Literature cited:

1. Feghali-Bostwick C, Medsger TA, Jr., Wright TM. Analysis of systemic sclerosis in twins reveals low concordance for disease and high concordance for the presence of antinuclear antibodies. Arthritis Rheum. 2003;48(7):1956-63. Epub 2003/07/09. doi: 10.1002/art.11173. PubMed PMID: 12847690.
2. Garrett SM, Frost DB, Feghali-Bostwick C. The mighty fibroblast and its utility in scleroderma research. J Scleroderma Relat Disord. 2017;2(2):69-134. Epub 2017/12/23. doi: 10.5301/jsrd.5000240. PubMed PMID: 29270465; PMCID: PMC5736140.
3. Li H, Yang R, Fan X, Gu T, Zhao Z, Chang D, Wang W, Wang C. MicroRNA array analysis of microRNAs related to systemic scleroderma. Rheumatology international. 2012;32(2):307-13.
4. Maurer B, Stanczyk J, Jüngel A, Akhmetshina A, Trenkmann M, Brock M, Kowal-Bielecka O, Gay RE, Michel BA, Distler JH. MicroRNA-29, a key regulator of collagen expression in systemic sclerosis. Arthritis & Rheumatism. 2010;62(6):1733-43.
5. Zhu H, Li Y, Qu S, Luo H, Zhou Y, Wang Y, Zhao H, You Y, Xiao X, Zuo X. MicroRNA expression abnormalities in limited cutaneous scleroderma and diffuse cutaneous scleroderma. Journal of clinical immunology. 2012;32(3):514-22.
6. Honda N, Jinnin M, Kajihara I, Makino T, Makino K, Masuguchi S, Fukushima S, Okamoto Y, Hasegawa M, Fujimoto M. TGF-β–mediated downregulation of microRNA-196a contributes to the constitutive upregulated type I collagen expression in scleroderma dermal fibroblasts. The Journal of Immunology. 2012:1100876.
7. Makino K, Jinnin M, Hirano A, Yamane K, Eto M, Kusano T, Honda N, Kajihara I, Makino T, Sakai K. The downregulation of microRNA let-7a contributes to the excessive expression of type I collagen in systemic and localized scleroderma. The Journal of Immunology. 2013:1200822.
8. da Silveira W, Renaud L, Simpson J, Glen W, Hazard E, Chung D, Hardiman G. miRmapper: A Tool for Interpretation of miRNA–mRNA Interaction Networks. Genes. 2018;9(9):458.
9. Silman AJ. Epidemiology of scleroderma. Current opinion in rheumatology. 1991;3(6):967-72. Epub 1991/12/01. PubMed PMID: 1772752.
10. Ramos PS, Silver RM, Feghali-Bostwick CA. Genetics of systemic sclerosis: recent advances. Current opinion in rheumatology. 2015;27(6):521-9. doi: 10.1097/BOR.0000000000000214. PubMed PMID: 26317679; PMCID: 4608482.
11. Arnett FC, Cho M, Chatterjee S, Aguilar MB, Reveille JD, Mayes MD. Familial occurrence frequencies and relative risks for systemic sclerosis (scleroderma) in three United States cohorts. Arthritis Rheum. 2001;44(6):1359-62. Epub 2001/06/16. doi: 10.1002/1529-0131(200106)44:6<1359::AID-ART228>3.0.CO;2-S. PubMed PMID: 11407695.
12. Vogel F, Motulsky AG. Vogel and Motulsky's Human Genetics: Problems and Approaches: Springer Science & Business Media; 2013.
13. Jarvinen P, Aho K. Twin studies in rheumatic diseases. Semin Arthritis Rheum. 1994;24(1):19-28. Epub 1994/08/01. PubMed PMID: 7985034.
14. Gregersen PK. Discordance for autoimmunity in monozygotic twins. Are "identical" twins really identical? Arthritis Rheum. 1993;36(9):1185-92. Epub 1993/09/01. PubMed PMID: 8216411.
15. Hawkins PG, Morris KV. RNA and transcriptional modulation of gene expression. Cell Cycle. 2008;7(5):602-7. Epub 2008/02/08. doi: 10.4161/cc.7.5.5522. PubMed PMID: 18256543; PMCID: PMC2877389.
16. Tan Y, Zhang B, Wu T, Skogerbo G, Zhu X, Guo X, He S, Chen R. Transcriptional inhibiton of Hoxd4 expression by miRNA-10a in human breast cancer cells. BMC Mol Biol. 2009;10:12. Epub 2009/02/24. doi: 10.1186/1471-2199-10-12. PubMed PMID: 19232136; PMCID: PMC2680403.
17. Zhou X, Tan FK, Xiong M, Arnett FC, Feghali-Bostwick CA. Monozygotic twins clinically discordant for scleroderma show concordance for fibroblast gene expression profiles. Arthritis Rheum. 2005;52(10):3305-14. Epub 2005/10/04. doi: 10.1002/art.21355. PubMed PMID: 16200604.

18. Sun Z, Evans J, Bhagwate A, Middha S, Bockol M, Yan H, Kocher JP. CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data. BMC Genomics. 2014;15:423. doi: 10.1186/1471-2164-15-423. PubMed PMID: 24894665; PMCID: PMC4070549.
19. Ru Y, Kechris KJ, Tabakoff B, Hoffman P, Radcliffe RA, Bowler R, Mahaffey S, Rossi S, Calin GA, Bemis L, Theodorescu D. The multiMiR R package and database: integration of microRNA-target interactions along with their disease and drug associations. Nucleic acids research. 2014;42(17):e133-e. Epub 2014/07/25. doi: 10.1093/nar/gku631. PubMed PMID: 25063298.
20. Backes C, Khaleeq QT, Meese E, Keller A. miEAA: microRNA enrichment analysis and annotation. Nucleic Acids Research. 2016;44(W1):W110-W6. doi: 10.1093/nar/gkw345.
21. Cloonan N. Re-thinking miRNA-mRNA interactions: Intertwining issues confound target discovery. Bioessays. 2015;37(4):379-88.
22. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society Series B (methodological). 1977:1-38.
23. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (methodological). 1995(Jan 1):289-300.
24. van der Pas S, van der Vaart A. Bayesian community detection. Bayesian Analysis. 2018;13(3):767-96.
25. R Development Core Team. R. 2016.
26. Team R. Development Core (2007) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www. R-project. org, 2005.
27. Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P, Oinn T. Taverna: a tool for building and running workflows of services. Nucleic Acids Research. 2006;34(Web Server issue):W729-W32. doi: 10.1093/nar/gkl320. PubMed PMID: PMC1538887.
28. Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A, Li P. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics (Oxford, England). 2004;20(17):3045-54. Epub 2004/06/18. doi: 10.1093/bioinformatics/bth361. PubMed PMID: 15201187.