Supplementary Materials and Methods

Specimen procurement and pre-processing

All participants underwent bowel preparation with Miralax prior to the colonoscopy. Patients did not take any other laxatives, probiotics or proton pump inhibitors for at least one week prior to the colonoscopy (verified by EV and NB on three separate occasions prior to the colonoscopy). Study personnel collected the colonic lavage samples (EV, MT, NB). For each sample region, approximately 30ml of sterile water was endoscopically flushed on the mucosal surface and recollected via aspiration. Samples were obtained from the cecum and the sigmoid colon regions. Samples were kept on ice for the duration of the pre-processing immediately following their collection. Samples were subsequently centrifuged at 4,000 x g for 10 minutes at 4°C. The supernatant was aliquoted into two 50-ml tubes with equal volumes and frozen at -80°C for future proteomic analyses. The pellets were resuspended in 2 ml of RNAprotect Bacteria Reagent (Qiagen, Valenica, CA, USA), aliquoted into 2 separate 15-ml conical tubes, centrifuged at 4,000 x g for 10 minutes at 4°C, separated from the supernatant and frozen at -80°C.

16S rRNA gene sequencing and microbial composition analysis

High-throughput sequencing analysis of bacterial rRNA genes was performed using extracted genomic DNA as templates. The PCR primers targeted the portion of the 16S rRNA gene containing the hypervariable V4 region. De-multiplexing, quality control, and operational taxonomic unit (OTU) binning were performed using quantitative insights into microbial ecology (MACQIIME v1.9.0).[1] Low quality

sequences were removed using the following parameters: (*i*) Q20, minimum number of consecutive high-quality base calls = 100 bp; (*ii*) maximum number of N characters allowed = 1; (*iii*) maximum number of consecutive low quality base calls allowed before truncating a read = 3. The remaining reads were subsequently used to select OTUs from the GreenGenes reference database (May 20, 2013), which automatically bins OTUs at 97% identify. Prior to the analysis, the species level OTUs that were observed fewer than two times were discarded. 16S rRNA sequence data will be deposited in Database of Genotypes and Phenotypes (dBGaP).

Statistical and bioinformatics analyses

Rarefraction and diversity analysis

After selection of the OTUs from GreenGenes database, microbal OTUs were rarified down to 15,000 reads per sample using MACQIIME. To compare the microbial communities of SSc versus control samples, alpha and beta diversity were analyzed. Alpha diversity, which represents the complexity of composition within members of a group, was calculated using the metrics of phylogenetic diversity, Chao1, observed species, and Shannon index. The comparison of alpha diversity between the two groups was performed using the two-sided student *t*-test at a depth of 15000.

Beta diversity represents the between-subject similarity of microbial composition and enables the identification of differences between samples within a group.[2] Beta diversity was performed in MACQIIME and utilized both unweighted and weighted UniFrac distances to estimate sample distributions. Analysis of variance using distance matrices (Adonis) significance analysis was performed for each pairwise comparison of

sample groups using the Adonis function from the R package. Principal coordinate analysis (PCoA) was performed to visualize the resulting UniFrac distance matrix.

Microbial composition analysis

Association between SSc disease state and bacterial phylotypes

To identify differentially abundant bacterial phylotypes in SSc and control samples, we performed a Kruskal-Wallis analysis and subsequently performed a Linear Discriminant Analysis Effect Size (LEfSe) analysis [3] to determine the effect size for the association of differentially abundant taxa with disease status. We were mainly interested in genus level differences because although alterations in the abundance of broad bacterial groups, such as entire phyla, can be associated with GIT dysfunction, small introductions of pathogenic strains at lower taxonomic levels can may have a greater impact on diseases.[4] Per convention, the log Linear Discriminant Analysis (LDA) score threshold was set at 2.

Relationship between microbes and SSc GIT symptoms

Differential expression analysis for sequence count data (DESeq2) [5] was used to compare differential abundance of bacterial species between patients with none to mild GIT 2.0 symptoms versus patients with moderate to severe GIT 2.0 symptoms as previously defined by Khanna et al. [6] We performed these analyses using the total GIT 2.0 scores, as well as GIT 2.0 scores for the following individual domains reflecting lower GIT dysfunction: Distention/Bloating, Diarrhea, and Constipation.

Imputation of microbial gene content and metagenomes

Phylogenetic investigation of communities by reconstruction of unobserved states (PICRUSt) was used to predict the functional composition of microbial communities.[7] Specifically, this computational approach combines 16S data and the gene content of reference genomes to model the abundance of metagenes in samples based on microbial composition. Metagene abundances for the individual healthy and SSc samples were imputed using the KEGG database. Metagenes were also grouped by KEGG pathway to yield pathway abundances. Variation in the metagenome across samples was calculated using Bray-Curtis distances and visualized by PCoA. Pair-wise comparisons between healthy and SSc subjects were calculated using the Kruskal-Wallis test (with adjustment for multiple hypothesis testing [q<0.1]), to identify imputed KEGG pathways and metagenes with differential abundance in SSc. LefSe analysis was used to compute the effect sizes of each pathway and genes. The significant threshold was set at q-values lower than 0.1 to identify imputed KEGG pathways and genes with differential abundances in SSc.

References for Supplementary Materials and Methods

1. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. Nature methods 2010;7(5):335-36.

2. Lozupone C, Lladser ME, Knights D, et al. UniFrac: an effective distance metric for microbial community comparison. The ISME journal 2011;5(2):169.

3. Segata N, Izard J, Waldron L, et al. Metagenomic biomarker discovery and explanation. Genome Biol 2011;12(6):R60.

4. Hamady M, Knight R. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. Genome Res 2009;19(7):1141-52.

5. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):550.

6. Khanna D, Hays RD, Maranian P, et al. Reliability and validity of the University of California, Los Angeles scleroderma clinical trial consortium gastrointestinal tract instrument. Arthritis Care & Research 2009;61(9):1257-63.

7. Langille MG, Zaneveld J, Caporaso JG, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol 2013;31(9):814-21.