nature communications

Article

https://doi.org/10.1038/s41467-024-53332-z

Widespread mutagenesis and chromosomal instability shape somatic genomes in systemic sclerosis

Received: 6 April 2024

Accepted: 9 October 2024

Published online: 15 October 2024

Check for updates

Sriram Vijayraghavan[1], Thomas Blouin[1], James McCollum[1], Latarsha Porcher[1], François Virard[2], Jiri Zavadil[3], Carol Feghali-Bostwick[4] & Natalie Saini[1]

Systemic sclerosis is a connective tissue disorder characterized by excessive fibrosis that primarily affects women, and can present as a multisystem pathology. Roughly 4-22% of patients with systemic sclerosis develop cancer, which drastically worsens prognosis. However, the mechanisms underlying systemic sclerosis initiation, propagation, and cancer development are poorly understood. We hypothesize that the inflammation and immune response associated with systemic sclerosis can trigger DNA damage, leading to elevated somatic mutagenesis, a hallmark of pre-cancerous tissues. To test our hypothesis, we culture clonal lineages of fibroblasts from the lung tissues of controls and systemic sclerosis patients and compare their mutation burdens and spectra. We find an overall increase in all major mutation types in systemic sclerosis samples compared to control lung samples, from small-scale events such as single base substitutions and insertions/deletions, to chromosome-level changes, including copy-number changes and structural variants. In the genomes of patients with systemic sclerosis, we find evidence of somatic hypermutation or kategis (typically only seen in cancer genomes), we identify mutation signatures closely resembling the error-prone translesion polymerase Polη activity, and observe an activation-induced deaminase-like mutation signature, which overlaps with genomic regions displaying kataegis.

Systemic sclerosis (SSc) or Scleroderma is a multisystem autoimmune disorder that affects roughly 1 in 5000 people in the United States, affecting thrice as many women as men[1–5]. The disease is characterized by vasculopathy and fibrosis of the skin and internal organs[6,7]. While scleroderma is a leading cause of SSc-associated disability, the disease often spreads to other organs such as the cardiopulmonary system. As such, heart, and lung abnormalities, such as interstitial lung disease (ILD) and pulmonary arterial hypertension (PAH) constitute the leading cause of mortality in SSc patients[8–12].

The underlying molecular mechanisms leading to SSc and associated downstream illnesses are only vaguely understood. Based on prior studies, it is likely that the disease manifests from a combination of shared environmental and genetic risk factors[13]. For example, polymorphisms in genes encoding immune factors such as STAT4 and IRF5[14–17], genes involved in the TGF-ß pathway[18,19], or those encoding collagenase enzymes[20] are associated with SSc onset. Other reports suggest workplace exposures to environmental pollutants like vinyl chloride or silica dust can trigger SSc-like disease symptoms[21–24].

[1]Department of Biochemistry and Molecular Biology, Medical University of South Carolina, Charleston, SC, USA. [2]University Claude Bernard Lyon 1, INSERM U1052–CNRS UMR5286, Cancer Research Center, Centre Léon Bérard, Lyon, France. [3]International Agency for Research on Cancer WHO, Epigenomics and Mechanisms Branch, Lyon, France. [4]Department of Medicine, Division of Rheumatology, Medical University of South Carolina, Charleston, SC, USA.
e-mail: sainina@musc.edu

In general, activation of fibroblasts is central to the disease process[25]. Remarkably, roughly 4-22% of SSc patients develop cancer, which drastically worsens prognosis for patients[26–28]. Studies show that the age- and sex-adjusted incidence of cancer in SSc is higher than what is expected within a general population[29–33].

SSc patients experience chronic lung and esophageal inflammation[34], which frequently precipitates genome instability[35–37]. Multiple lines of evidence show that tissues from SSc patients accumulate DNA damage, and display signs of a loss of genomic integrity, including autoantibodies to nuclear proteins[38], telomere attrition[39], formation of 8-oxoG adducts, and DNA single-and-double-strand breaks[40]. A recent study further demonstrated the role of inflammation-, and DNA damage-associated cGAS-STING response in the development of centromeric defects in skin fibroblasts derived from SSc samples[41]. Whole exome sequencing of skin biopsies from 8 patients with early progressive SSc revealed mutations in genes related to DNA damage response and epigenetic modifications and showed a clock-like mutation signature[42]. Based on these observations, we hypothesized that recurrent SSc-associated inflammation and autoimmune responses trigger genome-wide DNA damage, which could propagate systemic widespread mutagenesis across various tissue types. Because mutation accumulation is a hallmark feature of precancerous lesions, it can then be reasonably assumed that SSc-associated mutagenesis could drive the progression from inflammation to carcinogenesis.

Here, we report the mutational profile of single cell-derived clonal lung fibroblasts from a healthy individual and SSc patients. Using whole-genome sequencing of single cell-derived clonal lineages from healthy and SSc lung samples, we demonstrate that SSc samples have a heavy mutational burden. We further reveal enrichment of distinct signatures of mutational processes related to the translesion polymerase Polη (POLH) and activation-induced cytidine deaminase (AID)-like activities in SSc samples, and present evidence of localized hypermutation clusters, oncogenic driver mutations, somatic structural variants, and copy number alterations in SSc samples. Our work proposes somatic mutagenesis as a key step in illuminating the nexus between inflammation, SSc, and cancer.

Results

Clonal lineages from SSc samples carry a higher SBS burden than healthy cells

Because genome-wide somatic mutagenesis is often a precursor to cancer development[43], we hypothesized that inflammation/fibrosis likely underlies mutational accumulation in SSc patients, thereby increasing their future risk for developing cancer. For our analysis, we relied on lung fibroblasts, as these are effector cells in SSc initiation and development, and are therefore an ideal cell type for SSc research[25,44]. We obtained primary fibroblasts from explanted lungs of 5 healthy individuals (denoted as NL ("normal lung") throughout the text, including Figures) and 6 SSc patients (Supplementary Data 1). From the initial population of bulk fibroblasts, we expanded single-cell clones, isolated genomic DNA from ~10^6 cells, and simultaneously performed whole-genome sequencing (WGS) on bulk samples and their corresponding clones (Fig. 1, Supplementary Data 2). For the subset of clone samples that exhibited poor growth during clonal amplification, we harvested genomic DNA from ~10^4–10^5 cells and performed whole-genome amplification (WGA, Methods[45]) to obtain WGS-competent DNA. Traditional single-cell WGA protocols use limited propagation of the single cell in culture (~4–5 cell divisions) to obtain DNA for sequencing, and thereby carry inherent amplification errors, potentially resulting in erroneous mutation calls[46]. However, since our modified protocol uses kindred clones consisting of ≥ 10^4 cells, we expect minimal artifactual mutation calls arising from amplification biases[47]. In this manner, we were able to obtain a single clone for 9/11 lung fibroblasts, and >=2 independent clones for 2/11 lung fibroblasts (SSc-14, SSc-15, Supplementary Data 2). Median coverage for WGS was 57X (Supplementary Data 2). Since most mutations identified upon WGS of bulk cells represent pre-existing changes present in the vast majority of cells in the sample, we designated these mutations as a proxy for germline events. These mutations were subsequently removed from the list of mutations identified in the clonal samples.

Finally, we compiled consensus single base substitutions (SBS) data from three independent somatic mutation callers (see Methods) and applied stringent variant allele frequency (VAF) filters to avoid subclonal mutations that may have arisen either during tissue culture or from sequencing errors (Fig. 1A). Specifically, any VAF values within a 40-60% range for heterozygous calls and >90% for homozygous calls were considered clonal, whereas values outside this range were designated sub-clonal. In this manner, we were able to estimate accurate, sample-specific somatic mutation loads while removing any mutations that may be generated during cell culture.

Within this parsimonious list of mutation calls, we observed an increase in mutation loads in SSc samples compared to healthy samples (Median no. of SBS 1440 (SSc) vs 732 (NL), Fig. 1B, Supplementary Data 3a). Barring two samples from healthy donors who were heavy smokers (NL-127, NL-131, 1ppd > 20 years, Fig. 1 B-open circles, Supplementary Data 3), healthy samples had an overall low number of SNVs, whereas nearly all the SSc samples displayed elevated SNV levels. We additionally performed PCR amplification and Sanger sequencing on bulk and clone samples from three independent donors, and verified a subset of SBS calls (Supplementary Data 3(b)). We found that all tested SBS were true positives demonstrating the high accuracy of our method. We further noted that the rate of SBS accumulation as compared to the age of donors was higher in SSc than NL samples (Median no. of SBS per year, NL = 12.84, $n = 5$; SSc = 28.74, $n = 9$, Fig. 1C). Lastly, we asked if the observed mutations exhibit a non-uniform distribution across the genome, either demonstrating strand specificity of individual base changes, and/or a correlation with regions defined by replication timing[48–50]. However, we did not observe any statistically significant association of mutations with specific genomic features, including transcriptional strand bias, replicative strand bias, or replication timing (Supplementary Data 3c, d).

We conclude that compared to healthy tissues, SSc displays increased somatic single base substitutions.

SSc samples are enriched in the SBS93 mutation signature

The overall higher burden of mutations in SSc was also reflected in the spectrum, with increased cumulative C→T/G→A and T→C/A→G mutation loads (Supplementary Data 4).

To investigate the predominant mutation signatures associated with SSc-identified mutations, we asked if published signatures from a catalog of sequenced cancers were enriched in our datasets (Catalog Of Somatic Mutations In Cancer (COSMIC), https://cancer.sanger.ac.uk/signatures/ and[51,52]). As a primary approach, we used non-negative matrix factorization (NMF) to deconvolute the identified de novo signatures based on both algorithmic processes and experimental data (Supplementary Data 4)[52]. However, NMF-derived signatures often display broad profiles that can significantly overlap with each other, and/or lead to signature overfitting, which hampers signature attribution and confounds downstream analysis[53]. To circumvent this issue, and as recently reported[54], we used the predominant signatures identified by NMF and analyzed our datasets using Mutation Signature Analysis (MSA) tool for optimized signature refitting, which is based on simulations and parametric bootstrapping[53]. Using MSA, we noticed a recurrent SBS93 signature in SSc samples, which predominantly consists of mutations in the nCw and nTw motifs (Fig. 2A, B). In addition, we observed that the magnitude of difference in Signature 93 mutation burden was higher in SSc fibroblasts compared to healthy samples in our MSA analysis (Fig. 2B, Supplementary Fig. 1, Supplementary Data 4).

https://doi.org/10.1038/s41467-024-53332-z

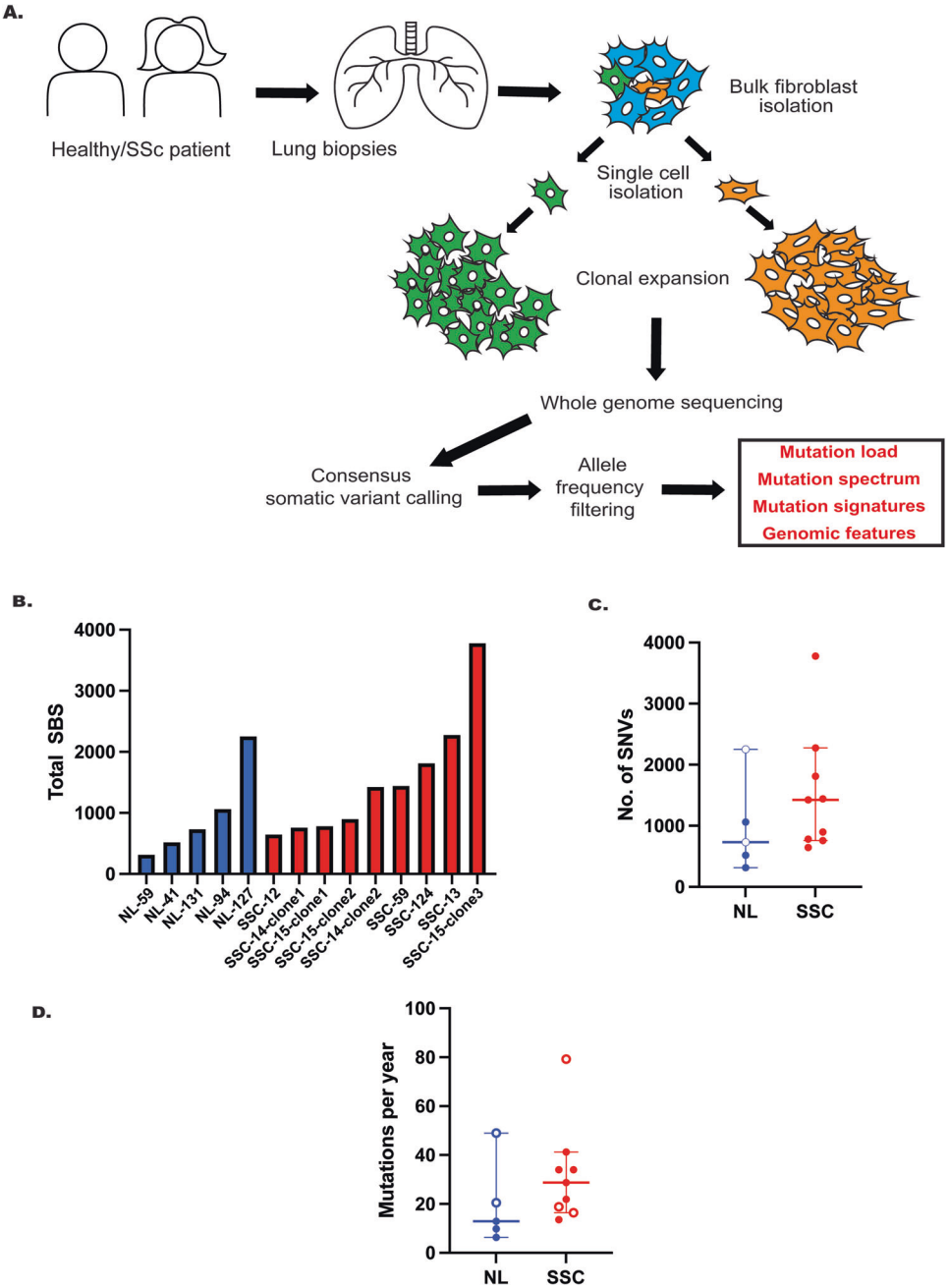


Fig. 1 | SBS levels are elevated in SSc samples. A Schematic of somatic mutation analysis of healthy v SSc samples, starting with single-cell clone isolation from patient lung fibroblasts, whole genome sequencing, and variant analysis. **B** Mutation load of SBS in NL and SSc samples. Per sample mutations and aggregate mutations (healthy v SSc, accounting for the sample size) are shown for all samples. **C** Median NL vs SSc SNVs. Values based on 5 healthy (NL) and 9 SSc samples. **D** NL v SSc SNVs accounting for smoking status and age of donors. Usage of 1ppd (pack-per-day) for >20 years was considered heavy smoking status (See Supplementary Data 1). For both **B** and **C**, heavy smokers are indicated by open circles. Color schematic- Healthy (NL)-blue, SSc-red. Error bars represent 95% confidence intervals. Source data are provided as a Source Data file.

We identified other common COSMIC signatures in our samples; these included SBS1 (2/5 NL samples, 4/9 SSc samples) and SBS5 (all NL and SSc samples), which have been shown to be age-associated signatures universally present in all tissues (Fig. 2A, Supplementary Fig. 1, Supplementary Data 4), and SBS40 which is a relatively flat signature with an unknown etiology, but was prevalent among two SSc samples (Supplementary Fig. 1, SSc-13, SSc-124). Notably, current signature extractors often over-assign SBS5 and SBS40 to nearly every sample in the analysis of cancer datasets, likely because of their overlap with other similar signatures[55,56]. Therefore, we cannot accurately determine their actual contribution to SSc-specific mutations. MSA also detected SBS18 in healthy and SSc samples- a signature frequently associated with DNA damage specific to reactive oxygen species, leading to G→T (C→A) substitutions[57]. However, we did not observe any notable difference within our healthy and SSc samples for SBS18 mutation loads (Supplementary Fig. 1). Of note, SBS18 has been previously reported to arise as an artifact *of* in vitro cell culture[58–61]. However, since our stringent allele frequency filtering strategy described above largely negates culture-associated mutations, we propose that normal respiration-associated oxidative damage might be a potential source for SBS18 in all lung samples.

https://doi.org/10.1038/s41467-024-53332-z

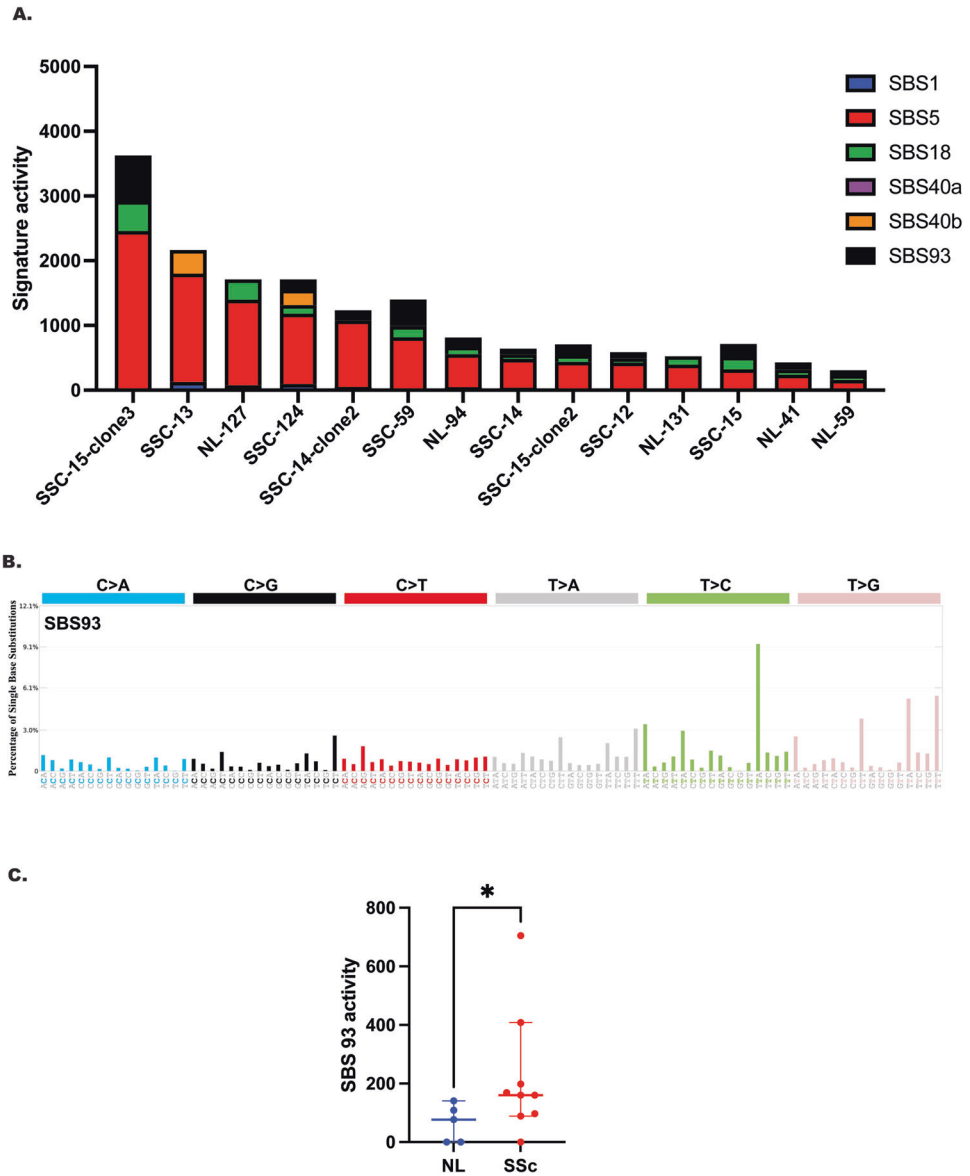


Fig. 2 | MSA-derived mutational signatures identified in SSc. Signature assignments are based on COSMIC-derived signatures from the analysis of human whole genome sequenced cancers. **A** SBS signature assignment per sample **B**. Mutational profile of COSMIC SBS93. **C** Median SBS93 signature activity in cumulative healthy v SSc samples. Values based on 5 healthy (NL) and 9 SSc samples. Error bars represent 95% confidence intervals. Asterisk represents the significant statistical difference in median values of NL vs SSc mutation loads based on a *p*-value of 0.0270 from a one-tailed Mann-Whitney T-test. Color schematic- Healthy (NL)-blue, SSc-red. Source data are provided as a Source Data file.

To orthogonally evaluate mutation signatures in our samples, we utilized our knowledge-based pipeline, TriMS, to ask which predetermined somatic mutations in trinucleotide motifs were statistically enriched in the samples. TriMS allows us to evaluate the enrichment of mutations within a given trinucleotide motif in each sample and determines if mutation signature enrichment is statistically significant[62]. For samples with statistically significant enrichment, we further determine the minimal mutation load associated with the signature. The nTw→N mutations were amongst the predominant changes in SBS93. Using TriMS, we noticed a marked increase in the minimum mutation load of nTw→N (w = A or T, n/N = A/T/G/C) in the SSc samples as compared to NL Samples (Fig. 3A, Supplementary Data 5). SBS93 is also made up of a smaller number of nCw→N mutations. Since C→T changes can be attributed to a large number of mutagens, we further analyzed nCw→R (R = A or G) changes to avoid overlapping mutations. SSc samples showed an overall increase in the median mutational burden with the nCw→N and nCw→R compared to NL samples (Fig. 3A, Supplementary Data 5). Furthermore, we saw that the mutation loads for nTw→N correlated with nCw→N showed a positive, albeit small, overall correlation for SSc samples, but not healthy samples, hinting that perhaps they are components of the same mutation signature (Fig. 3B). The nTw→N mutations in SSc samples correlate with age (Fig. 3C, *p*-value 0.013), revealing clock-like properties for this signature in SSc lung fibroblasts. SSc-15 clone 3 represented an SSc sample with very high mutation loads for all the above signature analyses. Nevertheless, even when SSc-15 clone 3 was excluded from analysis, overall SSc mutation loads for the above signatures remained significantly higher compared to healthy samples (Supplementary Fig. 2). Furthermore, while exclusion of SSc-15 clone 3 abrogated the statistically significant correlation between the nCw→N and nTw→N signature, we nevertheless observed a positive slope (Supplementary Fig. 2), indicating common mutational processes operating across diverse SSc samples.

https://doi.org/10.1038/s41467-024-53332-z

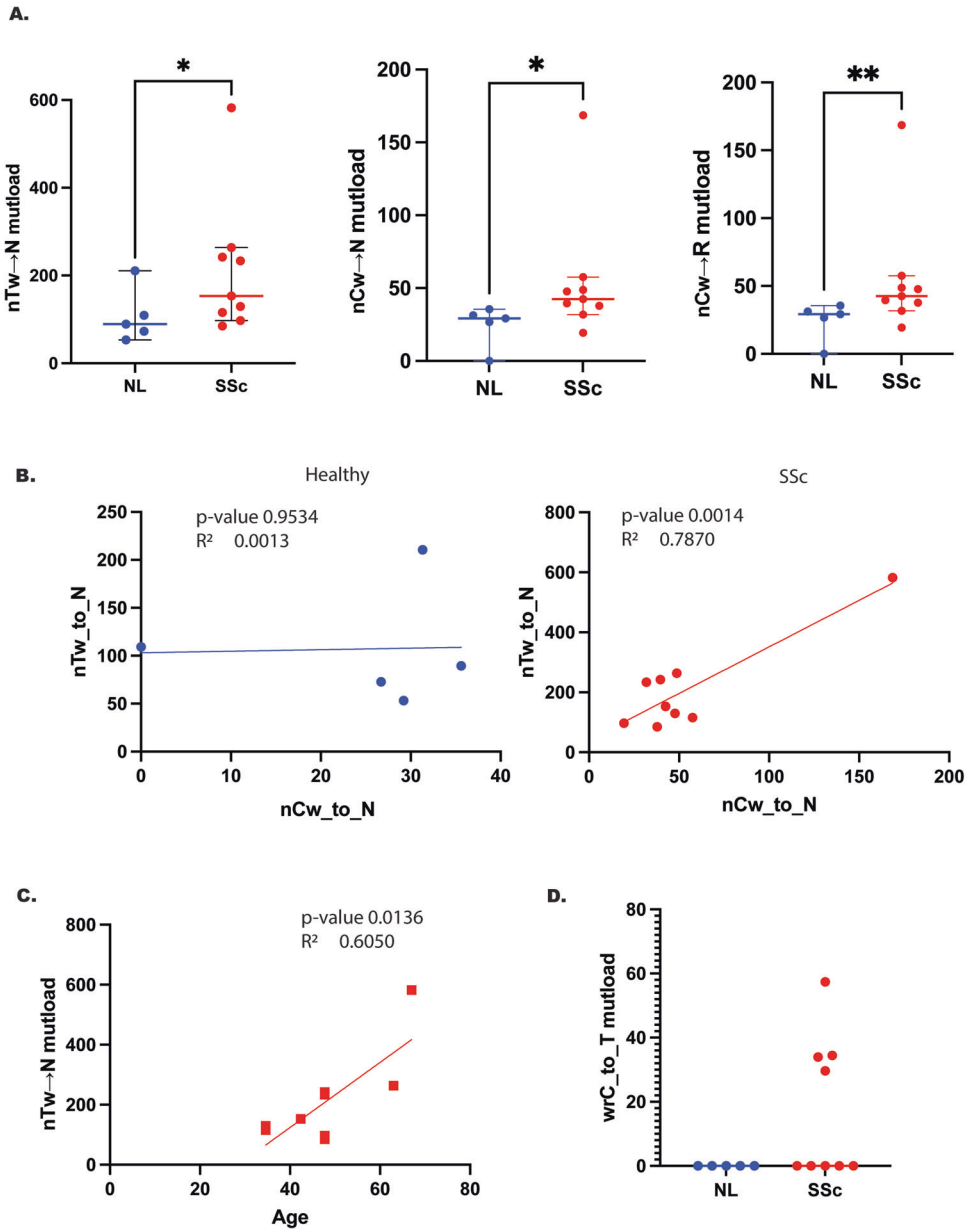


Fig. 3 | Discrete *POLH* and AID-like mutation signatures in SSc samples. **A** Comparisons of minimum mutation loads of *POLH* -associated mutation signatures between healthy and SSc samples from TriMS signature analysis. Error bars represent 95% confidence intervals. Values based on 5 healthy (NL) and 9 SSc samples. * represents statistically significant *p*-values based on a one-sided Mann-Whitney test comparing median mutation loads, with * denoting *p*-values < 0.05, ** denoting *p*-values < 0.005. *p*-values are as follows: nTw→N = 0.0466, nCw→N = 0.0148, nCw→R = 0.0095. **B** Correlation between the minimum mutation loads of nTw→N vs.nCw→N *POLH* signatures for SSc samples. n = A/T/G/C, w = A/T. **C** Correlation of nTw→N minimum mutation load with sample age in SSc samples. **D** Comparison of AID-like wrC→T mutation loads between NL ("healthy") and SSc samples. Mutated residue is at the 3rd position in the trinucleotide, w = A/T, r = A/G. For panels **B** and **C** *p*-values are based on a simple linear regression, whereby R-squared values represent the goodness-of-fit, and *p*-values < 0.05 are deemed statistically significant. For all panels, the color schematic is as follows: Healthy (NL)-blue, SSc-red. Source data are provided as a Source Data file.

We analyzed other signatures associated with C→T changes, which is the predominant mutation type associated with SSc samples (Supplementary Fig. 3, Supplementary Data 4). Similar to the NMF analysis, we saw that SBS1 (nCg→T) changes were enriched in all samples analyzed (Supplementary Data 5). Pro-inflammatory cytokines have also been shown to upregulate APOBEC3 expression, as such, we also analyzed the APOBEC3-specific tCw→T mutation signature in our samples[51,63]. None of the NL or SSc samples showed tCw→T mutation enrichment (Supplementary Data 5). To our surprise, we detected an enrichment of wrC→T mutations (w = A/T, r = A/G, C = mutated cytosine) in clones from 4- SSc patients (SSc-13, SSc-14 (both clones), SSc-15 (clone 3), Fig. 3D, Supplementary Data 5). Importantly, none of the NL samples carried this mutation signature (Supplementary Data 5). This mutation signature closely resembles activation-induced cytosine deaminase (AID) activity, encoded by the *AICDA* gene, belonging to the APOBEC family of base editing enzymes that is primarily expressed in B-lymphocytes[64,65]. However, because of our small dataset, we are limited in predicting the frequency and extent of association between AID mutation signatures and SSc. Overall, even within our small sample set, we see consistent evidence of enrichment for discrete mutational signatures in SSc samples.

https://doi.org/10.1038/s41467-024-53332-z

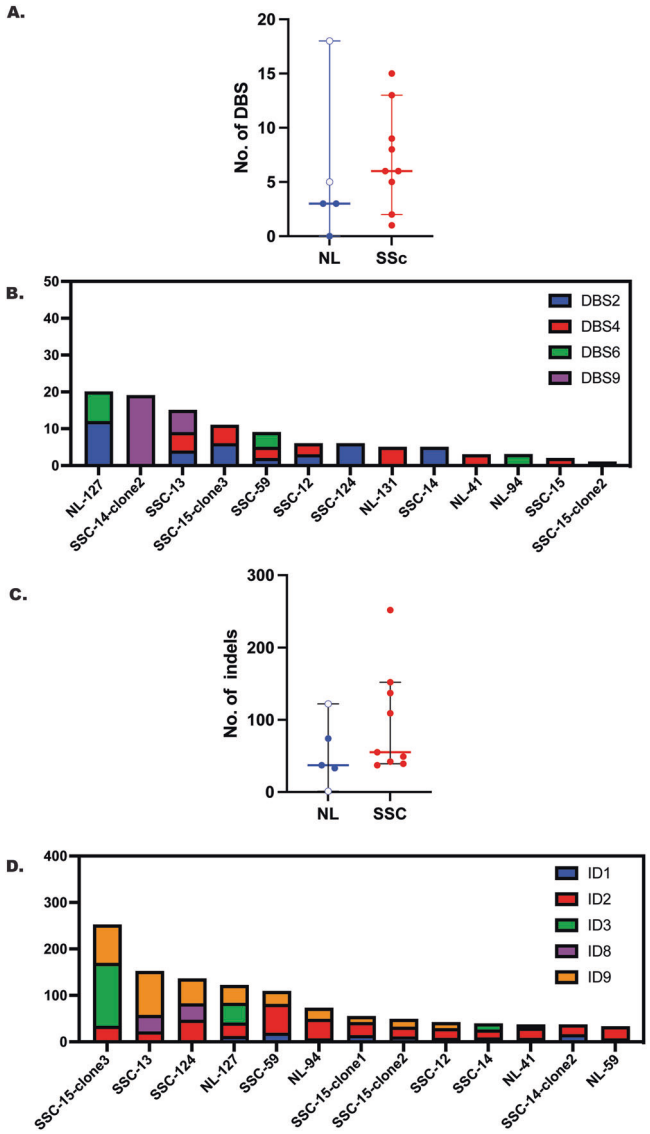


Fig. 4 | Other small mutational events in SSc. A Comparison of overall DBS loads between healthy and SSc samples, heavy smokers are identified by open circles. Median mutation loads plotted and p-value = 0.1983 calculated based on a one-sided Mann Whitney test. **B** Comparisons of overall COSMIC DBS mutation signature activity in cumulative healthy v SSc samples. **C** Comparisons of overall COSMIC INDEL mutation signature activity in cumulative healthy v SSc samples. Median mutation loads plotted and p-value = 0.0789 calculated based on a one-sided Mann Whitney test. **D** Comparison of overall INDEL loads between healthy and SSc samples. Heavy smokers are identified by open circles. For **A** and **C**, error bars represent 95% confidence intervals. Values based on 5 healthy (NL) and 9 SSc samples. Color schematic- Healthy (NL)-blue, SSc-red. Source data are provided as a Source Data file.

Doublet base substitutions, insertions, and deletions are elevated in SSc

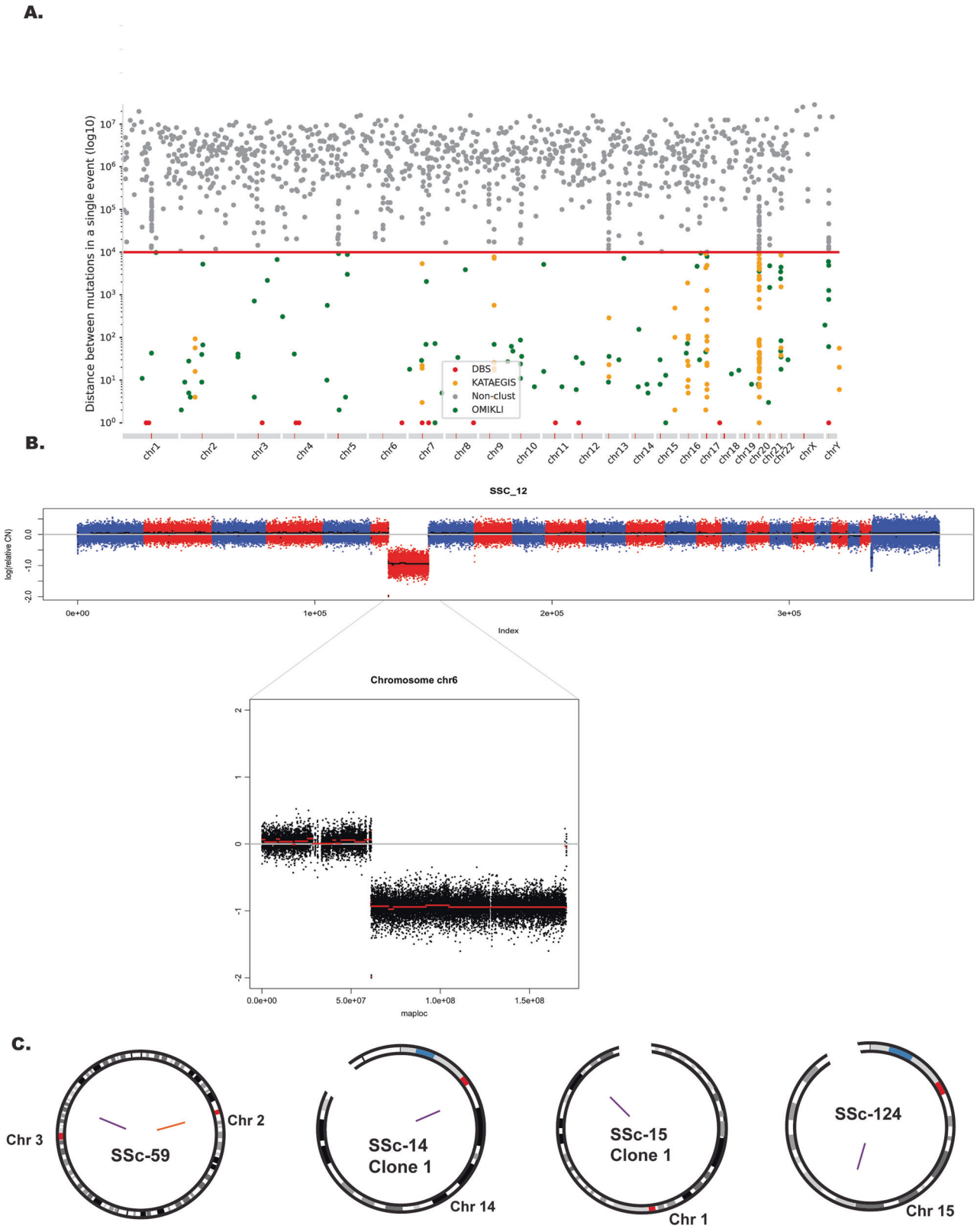
Among other mutation classes, we detected a small number of doublet base substitutions (DBS) within our samples. SSc samples carried a marginally higher median number of DBS mutations compared to healthy samples (Median no. of DBS = 3 (NL) vs 6 (SSc), Supplementary Data 3e, Fig. 4A). The largest number of DBS mutations within healthy samples coincided with heavy smoking (NL-127, NL-131, Fig. 4A open circles). We subsequently asked if any COSMIC DBS signatures are represented among the SSc sample. A DBS signature associated with tobacco smoking, DBS2, was prominent in one (NL-127) of the two samples from healthy people with a history of heavy smoking. This signature was present in almost all SSc samples (Supplementary Data 4). We further identified a low number of mutations associated with DBS4 (5/9 SSc samples, 2/5 NL samples) and DBS9 (2/9 SSc samples, 0/5 NL samples) in our samples, although their etiologies are currently unknown (<=5 for DBS4, <=20 for DBS9, Fig. 4B, Supplementary Data 4). Both these DBS signature involve changes in GC and TC motifs (Supplementary Data 4).

We further noted that INDEL (Insertions Deletions) levels were generally elevated across SSc samples (Median no. of INDELS = 37 (NL) vs 55 (SSc), Fig. 4C, Supplementary Data 3f-g), with INDEL sizes ranging from 1-2 bp (Fig. 4C). Among healthy samples, the largest number of INDEL mutations were found within NL-127, which was among the two heavy smokers in the study. We subsequently analyzed INDEL signatures in our datasets using SigProfiler[66]. INDEL signature ID3, was detected in two samples (SSc-15, NL-127, Fig. 4D). ID3 is characterized by CC→NN (GG→NN) changes, which overlaps with a well-known motif for acetaldehyde-induced DNA interstrand crosslinks[67,68]. Given that both the samples harboring these signatures were obtained from light (SSc-15) or heavy (NL-127) smokers, our analysis provides a proof-of-principle for the validity of ID signature detection in our samples. ID1 and ID2, associated with replication defects/fork slippage was detected across all samples in our analysis (Fig. 4D, Supplementary Data 4). Two SSc samples (SSc-13, SSc-124) displayed an ID8 signature (Fig. 4D, Supplementary Data 4). The latter is associated with non-homologous end joining (NHEJ)-mediated DSB repair or topoisomerase TOP2A -associated mutations[52,69]. Though we currently lack direct evidence of NHEJ/TOP2A-like signatures in our samples, the presence of ID8 hints at the underlying genome instability within SSc genomes. Finally, signature ID9 was found in 7 out of 9 SSc samples (Fig. 4D). Although the etiology of ID9 is currently unknown, we observed a positive correlation between the presence of this signature and the POLH-associated nTw→N signature across all samples (Supplementary Fig. 4). Overall, our data strongly suggests the involvement of multiple different mutational processes in SSc, with possible origins in inflammation and altered DNA metabolism.

Evidence of clustered mutations ("kataegis") in SSc samples

Cancer genomes often display regions of localized hypermutation, characterized by a burst of four or more SNVs within a narrow inter-mutational distance. Such instances of clustered mutagenesis, also termed "kataegis", were initially described in a cohort of 21 breast cancers[70], but have also been reported in other studies, where they occur in long stretches of single stranded DNA associated with DNA double strand breaks (DSBs)[71]. In addition, diffuse hypermutation, consisting of clusters of 2-3 mutations, termed "omikli" have also been observed in cancer[72]. Recently, samples derived from non-tumor associated intestinal crypts displayed APOBEC-associated clustered mutagenesis, demonstrating the presence of clustered mutations in normal aging tissues[73]. We used SigProfiler Clusters[72] to identify mutational clusters within our sample datasets. Only one out of 5 healthy samples analyzed in our study showed limited mutational clustering on Chromosome 1 (Supplementary Data 3h), whereas 5 out of 9 SSc samples showed evidence of kataegis across different chromosomes (Fig. 5A, Supplementary Fig. 5). Most clusters contained 4-6 SBS, however, one sample (SSc 14, clone 2) had multiple clusters, with one cluster on Chr 20 containing 47 SBS (Fig. 5A, Supplementary Data 3p). Signature analysis on only those SSc-associated SBS that were classified as clustered demonstrated an enrichment for nTw→N mutations (SSc-14-clone2, min. mut. Load =41.60568023; SSC15-clone 3, min. mut. Load = 7.020496224) and AID (SSC14, clone 1, min. mut load = 13.49594814) signatures (Supplementary Data 6). Finally, one SSc sample showed an enrichment for the tCw→T signature (w = A/T), which is characteristic of APOBEC induced mutagenesis (SSc-124, min. mut. Load = 5.59402242, Supplementary Data 6).

https://doi.org/10.1038/s41467-024-53332-z



Chromosomal levels genomic changes are enriched in SSc samples

We asked if SSc genomes also carried larger variations at the chromosomal level such as copy number changes. Using VarScan2 DNA copy number analysis[74], we noticed that three SSc samples had chromosome arm level amplification and deletion events (SSc-12, SSc-13, SSc-15 clone 3, Fig. 5B, Supplementary Fig. 6, Supplementary Data 3i). Specifically, the SSc-13 sample had a large amplification and deletion event on chr 10 (~88 Mb) and Chr 6 (~52 Mb), respectively, SSc-12 showed a large Chr 6 deletion (109 Mb) (Fig. 5B), whereas SSc-15 showed a ~4 Mb deletion on Chr 21 (Supplementary Data 3i). Importantly, we did not see any such megabase level copy number changes in normal fibroblasts.

Fig. 5 | Mutation clusters and chromosomal instability in SSc. A Representative rainfall plot shown for sample SSc-14-clone 2. Inter-mutational distance was calculated, and mutations assigned as doublet base substitutions (DBS), kataegis (long strand coordinated clustered hypermutation), or *omiklii* (diffuse hypermutation) using parameters described in ref. 72. **B** Copy number variation in SSc. Representative segmented chromosome plots for SSc-12, showing large scale deletion at Chr 6 (inset). Plots were generated after smoothening and segmenting of the raw output from Varscan copynumber[74]. $\log_2$ threshold was re-centered to 0.0 (neutral), with anything above the threshold counting as a "gain" (amplification), and anything below as a "loss" (deletion) event. Red and blue represent numerically-ordered bins of chromosomes. **C** Circular chromosome plot of structural variants (SV) in SSc samples, showing the following SVs −SSc-59- inversion (orange line) in the pericentromeric region (red square in chromosome track) on Chr 2 and a deletion in Chr3 (purple line); SSc-14- deletion in Chr 14 (purple line); SSc-15-clone 1-deletion in Chr 1 (purple line), SSc-124- deletion in Chr 15 (purple line). Plots were generated using RCircos[115]. Source data are provided as a Source Data file.

We also analyzed large structural variations (SV) within the samples using consensus calls from two independent SV callers SvaBa[75] and DELLY[76]. Both callers analyze breakpoint junctions allowing for more precise annotation of chromosomal rearrangements as compared to CNV analysis. SSc samples carried at least one deletion ranging from 2-100 kb in length (Fig. 5C, Supplementary Data 3j). SSc-59 additionally harbored a large inversion of >200 kb on Chr 2 (Fig. 5C). Using the Database of Genomic Variants (DGV)[77], we notice that the inversion maps to the centromeric region between 2p11.1-2q11.1 and is most likely a pericentric inversion which is a common chromosomal rearrangement in humans, occurring at a frequency of 1-2%[78]. We did not detect similar CNVs or SVs for any of the samples from healthy subjects, indicating that the observed chromosomal events are likely with SSc-specific genomic instability. It has previously been reported that DNA samples obtained through WGA can lead to erroneous SV calls arising resulting from amplification errors[47]. As such, we excluded WGA-derived genomic DNA samples from our SV analysis.

We assessed the frequency of Loss-of-heterozygosity (LOH) events in our samples. LOH is a frequent mutational event associated with carcinogenesis, and results primarily from mitotic recombination leading to deletions during cell division[74]. Overall, per sample LOH events were higher in SSc compared to healthy subjects (Supplementary Fig. 7, Supplementary Data 3k–m). We further noted that Chr6 in SSc12 and SSc13 carried long tracks of LOH events in the regions that coincided with a copy number loss by CNV analysis (Fig. 5B, Supplementary Fig. 7). Overall, our findings demonstrate that large structural and copy number variations are present in SSc lung fibroblasts.

For a better understanding of the molecular pathways that could explain the underlying genomic instability in our SSc samples, we annotated gene-associated SNVs using the Cancer Genome Interpreter[79], which classifies the list of annotated genes as cancer drivers or passengers, and identified two predicted oncogenic driver mutations in our SSc samples – SSc-124 (NF1, E572K) and SSc-15-clone3 (SEC31A, I573V) (Supplementary Data 3o, p). We did not observe an over-enrichment of specific genes or pathways, rather most mutations in both healthy and SSc samples identified as passenger mutations spread across a diverse set of genes spanning cell adhesion (*COL4A5*), Wnt signaling (*TANC*) to DNA metabolism (*CGAS*) (Supplementary Data 3o, p). Interestingly, one of our samples harbored a missense mutation in *BCOR* (SSc-15 clone 3, Supplementary Data 3o, p), which encodes a Bcl6 co-repressor and plays an important role in immune response, and was identified as a mutated driver in intestinal crypts from IBD patients[80], suggesting possible mechanistic commonalities between diverse inflammation associated diseases. Finally, in SSc-124, we noted a G→A at chr10 position 66009010 within the *POLH*-associated nCw→R/wGn→Y motif[81]. This mutation mapped to *CTNNA3*, which encodes an alpha-T-catenin, a pro-inflammatory factor that was previously reported to be hypomethylated and differentially overexpressed in SSc endothelial cells[82,83], perhaps suggesting a positive selection for pro-fibrotic mutations in SSc samples.

Discussion

The inherent molecular connection between SSc and cancer is poorly understood, which necessitates novel research strategies. In the present study, we explore the somatic mutational landscape of systemic sclerosis in patient-derived lung fibroblasts. While prior studies have shown an association between DNA damage and SSc, to our knowledge, our study specifically emphasizes the contrasting nucleotide-level somatic variation in single cells obtained from patients with systemic sclerosis vs healthy individuals. Our data highlights the range of somatic variation present in genomes of SSc patients, from increased SBS, INDELs, and LOH events to structural variants (SVs) and copy number changes (CNVs).

Recent studies have explored the somatic mutational burdens in other inflammation-associated diseases. Whole genome sequencing of colonic crypts from patients of inflammatory bowel disease (IBD) revealed a sharp increase in single base substitutions and INDELs compared to normal colon biopsies, likely through recurrent inflammation[80]. Many of the mutations mapped to potential driver genes and other canonical mutation hotspots, implying that clones harboring these mutations were under selection for better growth, either via faster division or via the acquisition of resistance to inflammation-associated cellular toxicity[80]. Similarly, whole-exome sequencing of non-cancer skin samples from SSc patients displayed numerous somatic mutations, including non-synonymous mutations in cancer-associated genes, and reported clock-like mutation signatures[42]. While these studies highlight the close association between fibrosis and genome instability, likely resulting from altered immune signaling, the lack of parallel data from healthy patient samples in this study largely obfuscates functional interpretations of the observed mutation load and spectra. In addition, analysis of mutations from heterogenous bulk cells (a mixture of germline and somatic variations), or via error prone single-cell sequencing obscures signals from other minor but significant ongoing processes involved in disease development and propagation, thereby further confounding mutational analysis.

In contrast, our clonal expansion-based methodology allowed us to accurately capture the somatic mutation loads and specific signatures from SSc DNA. Our mutation calling pipeline relied on consensus calls from multiple variant callers, and stringent filtering criteria (VAF between 40-60% for heterozygous calls and >=90% for homozygous calls), which enabled us to measure minimum mutation loads in healthy and SSc datasets with high accuracy. Conceivably, our high stringency mutation calling pipeline could generate false-negative calls, and as such this possibility should be considered when making orthogonal comparisons of our data to other mutational data sets. We found that somatic mutations generally tend to accumulate with age, and contributions from underlying disease process(es) could explain the observed SSc-specific increase. It is conceivable that SSc patients experience inflammation-associated DNA damage and/or DNA repair errors early in their lifespan, resulting in an accelerated rate of mutation accumulation with age. Indeed, prior studies have suggested that the incidence of telomere attrition, which is another indicator of DNA damage, tends to increase at a younger age in cells from SSc patients compared to healthy subjects[39].

Interestingly, analysis of mutation signatures showed that SBS93 was higher in SSc than healthy subjects. SBS93 is a minor COSMIC signature that was previously associated with gastric cancers[84]. The mutation spectrum SBS93 can be roughly divided into T→N and C→N mutations in the nTw and nCw trinucleotide motifs, respectively

(N = A, T, G, or, C, W = A/T). Remarkably, the nTw motif was previously found to be associated with activity of *POLH* which encodes Polη, a Y-family specialized DNA polymerase involved in the error-free bypass of DNA lesions[85,86]. *POLH* signatures were linked to somatic mutagenesis in cancer as well as non-cancer associated tissues experiencing DNA damage, such as skin[81,87–89]. A mutation signature for *POLH* was first established in vitro[88,90] and subsequently confirmed in vivo in follicular lymphomas, and consists of mutations in the WA/TW (where W = A or T)[81]. We see a remarkable correlation between specific *POLH*-associated trinucleotide signature nTw→N with the less mutated nCw→N signature in our samples (Fig. 3) indicating that both these mutation types likely originate from the same source. In COSMIC, SBS9 is associated with POLH activity. However, a closer inspection of SBS9 mutable motifs and SBS93 mutable motifs demonstrates that nCw→R and nTw→N form the major consensus peaks of both signatures, with minor differences in the individual peaks (Supplementary Fig. 8a, 96-channel cosine similarity = 0.7512). Additionally, SBS9 and SBS93 share similar patterns of transcriptional strand asymmetry (Supplementary Fig. 8b). The above observations further suggest that SBS93 may in fact represent a POLH mutation signature. Since POLH mutation signatures have been identified either in the context of UV-induced DNA damage or activity[89,91,92] during somatic hypermutation, it is possible that in samples with other types of DNA damage POLH mutation signature might vary in terms of mutations within individual trinucleotide motifs. As such, we hypothesize that the presence of SBS93 (nTw→N and nCw→R) mutations in our samples likely represent *POLH* signatures. How prevalent might these signatures be in other normal tissues? As a proof-of-principle, we analyzed mutational datasets from WGS of normal bronchial epithelia from tobacco smokers and non-smokers[93]. We observe a significant increase in the minimum mutational loads for nCw→N, nTw→N, and nCw→R in samples from smokers/former-smokers, but not in never-smokers (Supplementary Fig. 9, Supplementary Data 7). Furthermore, there is a strong positive correlation between nCw→N and nTw→N in smokers/former-smokers but a lack of correlation in never-smokers (Supplementary Fig. 9, Supplementary Data 7). This observation indicates that DNA damage from smoking might also lead to elevated translesion synthesis and POLH-associated mutagenesis. We further stratified samples from this dataset according to known pulmonary disease status. Remarkably, lung samples from patients with chronic obstructive pulmonary disorder (COPD), an inflammatory lung disease, show a significant increase in both nCw→N and nTw→N mutation loads compared to samples from disease-free subjects (Supplementary Fig. 9, Supplementary Data 7). However, in this study the samples for COPD were derived from only two patients, as such, it is not possible to determine if elevated PolH-like mutagenesis is a signature of other lung diseases.

We infer the observed enrichment of *POLH*-associated signatures to arise from SSc-associated inflammation. Chronic inflammation is considered a large risk factor in spawning pro-carcinogenic mutations[94], partly by producing replication fork blocking lesions, or through increased cellular proliferation. Possibly, inflammation-associated reactive oxygen and nitrogen species (RONS) could generate adducts that, when combined with defects in replicative polymerases, homologous recombination (HR), or tumor suppressor genes like *TP53* undergo mutagenic bypass by POLH, leading to a large increase in SSc-associated mutation burden. Whether the observed mutations result from aberrant regulation of POLH activity, its differential expression, or through other mechanisms necessitates additional validation in different in vivo models. In fact, we still observe a significant enrichment of nCw→N (median min. mutation load NL = 24.49, $n = 5$, SSc=38.83, $n = 9$, p-value = 0.0025) and nTw→N (median min. mutation load NL = 84.29, $n = 5$, SSc=126.9, $n = 9$, p-value- = 0.03) mutation loads in SSc samples when we only look at non-clustered mutations (Supplementary Data 6, Supplementary Fig. 1 Supplementary Fig. 10), further suggesting that the *POLH*-like mutagenesis is a general mutational mechanism operating within SSc samples.

Based on our context-based signature analysis, we also unexpectedly found an AID-like signature in a subset of our SSc samples, which colocalized with sites of clustered mutations as well with a *POLH*-associated signature (Fig. 3). Given that AID expression is largely limited to lymphoid geminal centers[95], why would an AID-associated mutation signature appear in SSc-associated fibroblasts? Multiple studies have demonstrated ectopic expression of AID outside of the canonical germinal center (GC) B-cell population[96–98]. Prior studies have shown an association between AID-associated hypermutation and autoimmunity, with constitutive ectopic expression leading to tumorigenesis[99], and translocations at a diverse set of genes[100]. The observed signature combination could result in mutations of driver genes that could promote tumorigenesis in SSc patients. Additionally, off-target AID activity has been previously reported in a variety of cancers ranging from gastric tumors to skin melanomas[101–103]. While we cannot rule out that the wrC→T changes were made by another mechanism, our data suggests that perhaps a combination of heightened immune response and inflammation creates an ideal microenvironment for non-canonical AID activation, leading to an AID-like mutation signature in SSc samples. Moreover, AID has been shown to lead to hypermutation in single-stranded DNA stretches leading to clustered mutations[64,104]. Finally, it formally remains possible that the wrC motif is mutable via unidentified mechanisms that do not involve AID activation. As such the role of AID in SSc-associated mutagenesis needs further testing using orthogonal approaches.

The presence of chromosome-level mutational events (CNVs, SVs) in SSc-associated lung fibroblasts is consistent with prior observations of similar genomic instability in non-cancer fibroblasts[89,105,106]. However, in contrast to earlier studies, our work specifically demonstrates elevated genomic instability in lung fibroblasts, which are central to SSc pathology. In combination with our mutational analysis, our CNV and SV analyses support a model where recurrent cycles of inflammation and DNA damage can cause SSc genomes to progressively acquire mutations across the genome. Such mutations could affect genes that are important for genome stability (e.g. centromeric proteins, replicative enzymes, DNA repair factors), leading to compromised genomic integrity. DNA damage can further trigger genome surveillance pathways such as cGAS-STING[41], which would promote immune activation and inflammation, thereby creating a feedback loop of inflammation, DNA damage, and immune response. Lastly, mutations that alter surface antigens could allow a subset of mutated cells to escape immune surveillance and serve as cancer neo-antigens to initiate tumorigenesis. Future work with larger cohorts of healthy and SSc-associated tissue samples would highlight the extent of the genomic changes prevalent in SSc.

In summary, our work offers a snapshot into the mutational landscape of SSc and highlights some of the mechanisms that could drive inflammation-associated genomic instability in SSc. Follow-up studies using larger datasets, candidate gene approaches, and/or different tissue types would shed further light on the central players involved in SSc-specific inflammation, immune response, and mutagenesis. Such approaches are key to explore the progression of disease in SSc, and possibly other similar diseases.

Methods

Sample collection

Lung tissues were obtained from patients with SSc undergoing lung transplantation and healthy donors whose lungs were not used for transplantation as previously described[107,108]. Briefly, approximately 2 cm^3 pieces of peripheral lung with pleural margins removed were minced, and the resulting fibroblasts were cultured in Dulbecco's modified Eagle's medium (Thermo Fisher Scientific) supplemented

with 10% fetal bovine serum (Millipore Sigma), and antibiotics (penicillin, streptomycin, and anti-mycotic agent, Thermo Fisher Scientific). Early passage (3-7) cells were used for further sub-culturing.

Cell culture and clonal expansion

Primary fibroblasts were cultured from lung tissues using the outgrowth method. Cells were grown in T25 or T75 flasks (Genesee Scientific) in Dulbecco's Modified Eagle Medium (DMEM, Gibco) supplemented with 10% Cosmic Calf Serum (Hyclone), 10% AmnioMax C-100 supplement (Gibco), and 100 µg/ml Primocin (Invivogen). Cells were incubated at 37 °C with 5% CO_2. After 3-5 passages, roughly 1×10^6 cells were harvested from bulk samples for genomic DNA isolation. From the same culture, 100 mm dishes were seeded with 500-1000 cells for clone isolation and grown for 3-5 weeks until colonies were visible. Clones were subsequently isolated into 48-well plates, grown to confluence, and passaged into increasingly larger volumes until cell density reached ~ 1×10^6 cells, at which point genomic DNA was harvested using the Blood and tissue DNeasy kit (Qiagen). Clones for which cell density was insufficient for obtaining WGS-compatible genomic DNA, were subjected to whole genome amplification via isothermal multi-displacement amplification[47] using the Repli-G kit (Qiagen) according to the manufacturer's protocol. At least 10,000 cells were used for whole genome amplification of each sample. All DNA quantification was performed using the Qubit dsDNA BR assay kit (ThermoFisher). DNA libraries for WGS were prepared and sequenced through Medgenome Inc. (CA, USA) at ~ 60X coverage using the NovaSeq 6000 platform (Illumina). The coverage profile of the aligned samples was analyzed with *indexcov*[109].

Somatic mutation calling

FASTQ reads for bulk and clone DNA files were aligned along the hg38 genome using GATK best practices[110]). Somatic SBS calling was performed on matched bulk (i.e. "normal") and clone (i.e. "tumor") samples using three independent callers- VarScan2[74], Strelka[111], and SomaticSniper[112], with only consensus calls reported in the final analysis. SNVs were filtered to only keep heterozygous calls with variant allele frequency (VAF) in the 40-60% range and homozygous calls in the 90-100% range. Calls falling outside these ranges were deemed subclonal, arising either from sequencing errors, or via propagation during cell culture. The SNV calling pipeline additionally filtered out any SNVs that have previously been identified in the dbSNP138 database. A subset of candidate SNVs were orthogonally verified using PCR and Sanger sequencing (Supplementary Data 3). Somatic INDELs were similarly identified using consensus calls from three independent callers- VarScan, Strelka, and SVaBA, and with the same VAF filers as above.

Analyzing patterns of mutagenesis

SigProfilerClusters was used for analyzing inter-mutational distances between SNVs and sub-classify the events accordingly as doublet-base substitutions (DBS), multi-base substitutions (MBS), kataegis or *omiklii*. Rainfall plots were generated using karyoploteR[113]. Transcriptional strand bias was estimated using bedtools intersect using the hg38 reference genome. Significance was calculated by performing a binomial test on the number of mutations in each class in transcribed vs non-transcribed strands, followed by applying a Benjamini-Hoecheberg correction of the estimated *p*-values. Replication strand bias was analyzed via MutationalPatterns[114] using default parameters, with significance estimated via a two-sided Poisson test followed by FDR correction of reported *p*-values. SNV and INDEL annotations were carried out using the Cancer Genome Interpreter (CGI) (https://www.cancergenomeinterpreter.org/analysis[79],), which maps mutations to cancer driver genes and accordingly classifies them as either passengers or oncogenic drivers. Pearson's correlation coefficients were calculated to analyze the relationship between variant data and patient age. For computing statistical differences between healthy and SSc variant data, unpaired non-parametric Mann-Whitney tests were performed with *p*-values < 0.05 indicating significant deviation from H_O (i.e. no difference).

Copy number variants (CNVs) and Loss-of-heterozygosity (LOH) events were identified using Varscan2. For CNV identification, the command Varscan copynumber was run on bulk and clone pairs for a given sample, and copyCaller was subsequently run on the mpileup output from the previous step to perform the initial CNV calls and adjust for GC content. Segmentation was assigned to the copy-called files using DNaCopy package in R, data points were re-centered to a baseline neutral value (0.0), and finally, adjacent segments were merged in VarScan2. Events were classified as focal or large-scale depending on the event size. Only large-scale deletions and amplifications that could be visually verified in the segmentation plots were retained in the final analysis as true CNV events.

Structural variants (SVs), which include deletions, duplication, translocations, and inversions, were called using the consensus of two independent callers- DELLY and SVaBA[75] using default parameters. SVs calls with >30% reads supporting the variant and being absent in the bulk sample were counted as true variants. Consensus calls that did not match the above criteria and/or were classified as "LOW QUAL" were removed from the analysis. The Integrative Genome Visualization browser (https://igv.org/doc/desktop/). was used for genome-scale visualizations of the SVs. RCircos[115] was used to generate circular plots showing the position of the identified SVs on specific chromosomal tracks.

Mutation signature analysis

Single base substitution signature extraction was conducted on all samples using SigProfilerExtractor. MSA tool (v2.0) was used to fit the COSMIC signature set proposed by SigProfilerExtractor to each sample individually with the optimal threshold suggested by the MSA tool[53]. Doublet base substitution and INDEL signatures were extracted from the filtered mutation files using SigProfilerMatrixGenerator[66] and SigProfilerExtractor[84]. MutationalPatterns [114]was used to estimate the cosine similarity of identified signatures to COSMIC DBS and INDEL signatures, as well as to measure the contribution of COSMIC signatures to the mutation profile of the samples. Mutation enrichment and mutation loads were calculated using Trinucleotide Mutation Signatures (TriMS), which is based on the previously developed P-MACD for mutation signature analysis[47]. Base substitutions within specific trinucleotide motifs are compared against the total number of the given substitution genome-wide, as well the incidence of the mutated residue surrounding the mutated residue (±20 nucleotides), using the following equation:

$$Enrichment_{gCn \to A} = \frac{Mutations_{gCn \to A}\, X\, Context_C}{Mutations_{C \to A}\, X\, Context_{gcn}}$$

Minimum mutation loads for a given signature were calculated with a minimum enrichment probability of >1, as follows, accounting for multiple hypothesis testing by the Benjamini-Hocheberg method of *p*-value correction:

$$Mutload_{gCn \to A} = \frac{Mutations_{gCn \to A}\, X\, (Enrichment_{gCn \to A} - 1)}{Enrichment_{gCn \to A}}$$

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

BAM files and the associated index files, and somatic mutation calls from sequencing from this study are deposited with dbGAP and can

be accessed under registration number phs003700.v1.p1 [http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs003700.v1.p1] (note on SSc-124- This participant has not consented to sharing data on a public repository. Request for data associated with this subject should be personally requested from N.Saini (sainina@musc.edu). Upon request the data will be made available within a week as long as the requestor has obtained dbGap access for the remaining dataset). Source data are provided with this paper. Data for Supplementary Fig. 9 is from Yoshida et al. (2020)[93], and is available on Mendeley Data [https://doi.org/10.17632/b53h2kwpyy.2]. Source data are provided with this paper.

Code availability

The code for TriMS is publicly available on GitHub [https://github.com/SainiLabMUSC/TriMS] and is deposited with Zenodo [https://doi.org/10.5281/zenodo.13862689].

References

1. Peoples, C., Medsger, T. A. Jr., Lucas, M., Rosario, B. L. & Feghali-Bostwick, C. A. Gender differences in systemic sclerosis: relationship to clinical features, serologic status and outcomes. *J. Scleroderma Relat. Disord.* **1**, 177–240 (2016).
2. Fan, Y., Bender, S., Shi, W. & Zoz, D. Incidence and prevalence of systemic sclerosis and systemic sclerosis with interstitial lung disease in the United States. *J. Manag Care Spec. Pharm.* **26**, 1539–1547 (2020).
3. Ferri, C. et al. Systemic sclerosis: demographic, clinical, and serologic features and survival in 1,012 Italian patients. *Medicine* **81**, 139–153 (2002).
4. Scussel-Lonzetti, L. et al. Predicting mortality in systemic sclerosis: analysis of a cohort of 309 French Canadian patients with emphasis on features at diagnosis as predictive factors for survival. *Med. (Baltim.)* **81**, 154–167 (2002).
5. Meier, F. M. et al. Update on the profile of the EUSTAR cohort: an analysis of the EULAR Scleroderma Trials and Research group database. *Ann. Rheum. Dis.* **71**, 1355–1360 (2012).
6. Varga, J. & Abraham, D. Systemic sclerosis: a prototypic multisystem fibrotic disorder. *J. Clin. Invest.* **117**, 557–567 (2007).
7. Barsotti, S. et al. One year in review 2019: systemic sclerosis. *Clin. Exp. Rheumatol.* **37**, 3–14 (2019).
8. Volkmann, E. R. & Fischer, A. Update on morbidity and mortality in systemic sclerosis-related interstitial lung disease. *J. Scleroderma Relat. Disord.* **6**, 11–20 (2021).
9. Tyndall, A. J. et al. Causes and risk factors for death in systemic sclerosis: a study from the EULAR Scleroderma Trials and Research (EUSTAR) database. *Ann. Rheum. Dis.* **69**, 1809–1815 (2010).
10. Rubio-Rivas, M., Royo, C., Simeon, C. P., Corbella, X. & Fonollosa, V. Mortality and survival in systemic sclerosis: systematic review and meta-analysis. *Semin Arthritis Rheum.* **44**, 208–219 (2014).
11. McNearney, T. A. et al. Pulmonary involvement in systemic sclerosis: associations with genetic, serologic, sociodemographic, and behavioral factors. *Arthritis Rheum.* **57**, 318–326 (2007).
12. Steen, V. D. & Medsger, T. A. Changes in causes of death in systemic sclerosis, 1972-2002. *Ann. Rheum. Dis.* **66**, 940–944 (2007).
13. Mouawad, J. E. & Feghali-Bostwick, C. The molecular mechanisms of systemic sclerosis-associated lung fibrosis. *Int. J. Mol. Sci.* **24** (2023).
14. Rueda, B. et al. The STAT4 gene influences the genetic predisposition to systemic sclerosis phenotype. *Hum. Mol. Genet.* **18**, 2071–2077 (2009).
15. Tsuchiya, N. et al. Association of STAT4 polymorphism with systemic sclerosis in a Japanese population. *Ann. Rheum. Dis.* **68**, 1375–1376 (2009).
16. Xu, Y., Wang, W., Tian, Y., Liu, J. & Yang, R. Polymorphisms in STAT4 and IRF5 increase the risk of systemic sclerosis: a meta-analysis. *Int J. Dermatol.* **55**, 408–416 (2016).
17. Dieude, P. et al. Phenotype-haplotype correlation of IRF5 in systemic sclerosis: role of 2 haplotypes in disease severity. *J. Rheumatol.* **37**, 987–992 (2010).
18. Lafyatis, R. Transforming growth factor beta-at the centre of systemic sclerosis. *Nat. Rev. Rheumatol.* **10**, 706–719 (2014).
19. Morris, E. et al. Endoglin promotes TGF-beta/Smad1 signaling in scleroderma fibroblasts. *J. Cell Physiol.* **226**, 3340–3348 (2011).
20. Herrmann, K., Heckmann, M., Kulozik, M., Haustein, U. F. & Krieg, T. Steady-state mRNA levels of collagens I, III, fibronectin, and collagenase in skin biopsies of systemic sclerosis patients. *J. Invest. Dermatol.* **97**, 219–222 (1991).
21. Garabrant, D. H. et al. Scleroderma and solvent exposure among women. *Am. J. Epidemiol.* **157**, 493–500 (2003).
22. Muntyanu, A. et al. Exposure to silica and systemic sclerosis: A retrospective cohort study based on the Canadian Scleroderma Research Group. *Front Med.* **9**, 984907 (2022).
23. Shivakumar, D. S., Kamath, N. S. & Naik, A. Silica associated systemic sclerosis: an occupational health hazard. *BMJ Case Rep.* **16** (2023).
24. Lescoat, A. et al. Silica exposure and scleroderma: more bridges and collaboration between disciplines are needed. *Am. J. Respir. Crit. Care Med.* **201**, 880–882 (2020).
25. Garrett, S. M., Baker Frost, D. & Feghali-Bostwick, C. The mighty fibroblast and its utility in scleroderma research. *J. Scleroderma Relat. Disord.* **2**, 69–134 (2017).
26. Morrisroe, K. & Nikpour, M. Cancer and scleroderma: recent insights. *Curr. Opin. Rheumatol.* **32**, 479–487 (2020).
27. Zhang, J. Q. et al. The risk of cancer development in systemic sclerosis: a meta-analysis. *Cancer Epidemiol.* **37**, 523–527 (2013).
28. Onishi, A., Sugiyama, D., Kumagai, S. & Morinobu, A. Cancer incidence in systemic sclerosis: meta-analysis of population-based cohort studies. *Arthritis Rheum.* **65**, 1913–1921 (2013).
29. Christenson, L. J. et al. Incidence of basal cell and squamous cell carcinomas in a population younger than 40 years. *JAMA* **294**, 681–690 (2005).
30. Weeding, E., Casciola-Rosen, L. & Shah, A. A. Cancer and Scleroderma. *Rheum. Dis. Clin. North Am.* **46**, 551–564 (2020).
31. Bonifazi, M. et al. Systemic sclerosis (scleroderma) and cancer risk: systematic review and meta-analysis of observational studies. *Rheumatology* **52**, 143–154 (2013).
32. Lepri, G. et al. Systemic Sclerosis Association with Malignancy. *Clin. Rev. Allergy Immunol.* **63**, 398–416 (2022).
33. Mecoli, C. A., Rosen, A., Casciola-Rosen, L. & Shah, A. A. Advances at the interface of cancer and systemic sclerosis. *J. Scleroderma Relat. Disord.* **6**, 50–57 (2021).
34. Hoffmann-Vold, A. M. et al. Tracking impact of interstitial lung disease in systemic sclerosis in a complete nationwide cohort. *Am. J. Respir. Crit. Care Med.* **200**, 1258–1266 (2019).
35. Pezone, A. et al. Inflammation and DNA damage: cause, effect or both. *Nat. Rev. Rheumatol.* **19**, 200–211 (2023).
36. Kawanishi, S., Ohnishi, S., Ma, N., Hiraku, Y. & Murata, M. Crosstalk between DNA damage and inflammation in the multiple steps of carcinogenesis. *Int. J. Mol. Sci.* **18** (2017).
37. Kay, J., Thadhani, E., Samson, L. & Engelward, B. Inflammation-induced DNA damage, mutations and cancer. *DNA Repair.* **83**, 102673 (2019).
38. Igusa, T. et al. Autoantibodies and scleroderma phenotype define subgroups at high-risk and low-risk for cancer. *Ann. Rheum. Dis.* **77**, 1179–1186 (2018).
39. Usategui, A. et al. Evidence of telomere attrition and a potential role for DNA damage in systemic sclerosis. *Immun. Ageing* **19**, 7 (2022).

40. Vlachogiannis, N. I. et al. Association between DNA damage response, fibrosis and Type I Interferon signature in systemic sclerosis. *Front. Immunol.* **11**, 582401 (2020).
41. Paul, S. et al. Centromere defects, chromosome instability, and cGAS-STING activation in systemic sclerosis. *Nat. Commun.* **13**, 7074 (2022).
42. Gniadecki, R. et al. Genomic instability in early systemic sclerosis. *J. Autoimmun.* **131**, 102847 (2022).
43. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
44. Kendall, R. T. & Feghali-Bostwick, C. A. Fibroblasts in fibrosis: novel roles and mediators. *Front. Pharm.* **5**, 123 (2014).
45. Huang, L., Ma, F., Chapman, A., Lu, S. & Xie, X. S. Single-cell whole-genome amplification and sequencing: methodology and applications. *Annu. Rev. Genomics Hum. Genet.* **16**, 79–102 (2015).
46. Dong, X. et al. Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nat. Methods* **14**, 491–493 (2017).
47. Saini, N. et al. UV-exposure, endogenous DNA damage, and DNA replication errors shape the spectra of genome changes in human skin. *PLoS Genet.* **17**, e1009302 (2021).
48. Shinbrot, E. et al. Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. *Genome Res.* **24**, 1740–1750 (2014).
49. Lujan, S. A. et al. Mismatch repair balances leading and lagging strand DNA replication fidelity. *PLoS Genet.* **8**, e1003016 (2012).
50. Haradhvala, N. J. et al. Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell* **164**, 538–549 (2016).
51. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
52. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
53. Senkin, S. MSA: reproducible mutational signature attribution with confidence based on simulations. *BMC Bioinforma.* **22**, 540 (2021).
54. Deneuve, S. et al. Molecular landscapes of oral cancers of unknown etiology. *medRxiv* (2023).
55. Wu, A. J., Perera, A., Kularatnarajah, L., Korsakova, A. & Pitt, J. J. Mutational signature assignment heterogeneity is widespread and can be addressed by ensemble approaches. *Brief Bioinform.* **24** (2023).
56. Huang, X., Wojtowicz, D. & Przytycka, T. M. Detecting presence of mutational signatures in cancer with confidence. *Bioinformatics* **34**, 330–337 (2018).
57. Koh, G., Degasperi, A., Zou, X., Momen, S. & Nik-Zainal, S. Mutational signatures: emerging concepts, caveats and clinical applications. *Nat. Rev. Cancer* **21**, 619–637 (2021).
58. Petljak, M. et al. Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC Mutagenesis. *Cell* **176**, 1282–1294.e20 (2019).
59. Rouhani, F. J. et al. Mutational History of a Human Cell Lineage from Somatic to Induced Pluripotent Stem Cells. *PLoS Genet* **12**, e1005932 (2016).
60. Kuijk, E. et al. The mutational impact of culturing human pluripotent and adult stem cells. *Nat. Commun.* **11**, 2493 (2020).
61. Milholland, B. et al. Differences between germline and somatic mutation rates in humans and mice. *Nat. Commun.* **8**, 15183 (2017).
62. Vijayraghavan, S., Porcher, L., Mieczkowski, P. A. & Saini, N. Acetaldehyde makes a distinct mutation signature in single-stranded DNA. *Nucleic Acids Res.* **50**, 7451–7464 (2022).
63. Roberts, S. A. et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet* **45**, 970–976 (2013).
64. Pham, P., Bransteitter, R., Petruska, J. & Goodman, M. F. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature* **424**, 103–107 (2003).
65. Rogozin, I. B. et al. Activation induced deaminase mutational signature overlaps with CpG methylation sites in follicular lymphoma and other cancers. *Sci. Rep.* **6**, 38133 (2016).
66. Bergstrom, E. N. et al. SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics* **20**, 685 (2019).
67. Matsuda, T., Kawanishi, M., Yagi, T., Matsui, S. & Takebe, H. Specific tandem GG to TT base substitutions induced by acetaldehyde are due to intra-strand crosslinks between adjacent guanine bases. *Nucleic Acids Res.* **26**, 1769–1774 (1998).
68. Sonohara, Y. et al. Acetaldehyde forms covalent GG intrastrand crosslinks in DNA. *Sci. Rep.* **9**, 660 (2019).
69. Otlu, B. et al. Topography of mutational signatures in human cancer. *Cell Rep.* **42**, 112930 (2023).
70. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
71. Roberts, S. A. et al. Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol. Cell* **46**, 424–435 (2012).
72. Bergstrom, E. N., Kundu, M., Tbeileh, N. & Alexandrov, L. B. Examining clustered somatic mutations with SigProfilerClusters. *Bioinformatics* **38**, 3470–3473 (2022).
73. Wang, Y. et al. APOBEC mutagenesis is a common process in normal human small intestine. *Nat. Genet* **55**, 246–254 (2023).
74. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
75. Wala, J. A. et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28**, 581–591 (2018).
76. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
77. MacDonald, J. R., Ziman, R., Yuen, R. K., Feuk, L. & Scherer, S. W. The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* **42**, D986–D992 (2014).
78. Fickelscher, I. et al. The variant inv(2)(p11.2q13) is a genuinely recurrent rearrangement but displays some breakpoint heterogeneity. *Am. J. Hum. Genet* **81**, 847–856 (2007).
79. Tamborero, D. et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* **10**, 25 (2018).
80. Olafsson, S. et al. Somatic Evolution in Non-neoplastic IBD-Affected Colon. *Cell* **182**, 672–684.e11 (2020).
81. Rogozin, I. B. et al. DNA polymerase eta mutational signatures are found in a variety of different types of cancer. *Cell Cycle* **17**, 348–355 (2018).
82. Nada, S., Kahaleh, B. & Altorok, N. Genome-wide DNA methylation pattern in systemic sclerosis microvascular endothelial cells: Identification of epigenetically affected key genes and pathways. *J. Scleroderma Relat. Disord.* **7**, 71–81 (2022).
83. Folmsbee, S. S., Budinger, G. R. S., Bryce, P. J. & Gottardi, C. J. The cardiomyocyte protein alphaT-catenin contributes to asthma through regulating pulmonary vein inflammation. *J. Allergy Clin. Immunol.* **138**, 123–129.e2 (2016).
84. Islam, S. M. A. et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genom.* **2** (2022).
85. Burgers, P. M. et al. Eukaryotic DNA polymerases: proposal for a revised nomenclature. *J. Biol. Chem.* **276**, 43487–43490 (2001).

86. McCulloch, S. D. & Kunkel, T. A. The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Res.* **18**, 148–161 (2008).
87. Washington, M. T., Johnson, R. E., Prakash, L. & Prakash, S. Accuracy of lesion bypass by yeast and human DNA polymerase eta. *Proc. Natl Acad. Sci. USA* **98**, 8355–8360 (2001).
88. Matsuda, T., Bebenek, K., Masutani, C., Hanaoka, F. & Kunkel, T. A. Low fidelity DNA synthesis by human DNA polymerase-eta. *Nature* **404**, 1011–1013 (2000).
89. Saini, N. et al. The impact of environmental and endogenous damage on somatic mutation load in human skin fibroblasts. *PLoS Genet.* **12**, e1006385 (2016).
90. Matsuda, T. et al. Error rate and specificity of human and murine DNA polymerase eta. *J. Mol. Biol.* **312**, 335–346 (2001).
91. Rogozin, I. B. et al. Mutational signatures and mutable motifs in cancer genomes. *Brief. Bioinform.* **19**, 1085–1101 (2018).
92. Martincorena, I. et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
93. Yoshida, K. et al. Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).
94. Kiraly, O., Gong, G., Olipitz, W., Muthupalani, S. & Engelward, B. P. Inflammation-induced cell proliferation potentiates DNA damage-induced mutations in vivo. *PLoS Genet.* **11**, e1004901 (2015).
95. Muramatsu, M. et al. Specific expression of activation-induced cytidine deaminase (AID), a novel member of the RNA-editing deaminase family in germinal center B cells. *J. Biol. Chem.* **274**, 18470–18476 (1999).
96. Mao, C. et al. T cell-independent somatic hypermutation in murine B cells with an immature phenotype. *Immunity* **20**, 133–144 (2004).
97. William, J., Euler, C., Christensen, S. & Shlomchik, M. J. Evolution of autoantibody responses via somatic hypermutation outside of germinal centers. *Science* **297**, 2066–2070 (2002).
98. Schroder, A. E., Greiner, A., Seyfert, C. & Berek, C. Differentiation of B cells in the nonlymphoid tissue of the synovial membrane of patients with rheumatoid arthritis. *Proc. Natl Acad. Sci. USA* **93**, 221–225 (1996).
99. Okazaki, I. M. et al. Constitutive expression of AID leads to tumorigenesis. *J. Exp. Med.* **197**, 1173–1181 (2003).
100. Casellas, R. et al. Mutations, kataegis and translocations in B cells: understanding AID promiscuous activity. *Nat. Rev. Immunol.* **16**, 164–176 (2016).
101. Li, L. et al. Activation-induced cytidine deaminase expression in colorectal cancer. *Int J. Clin. Exp. Pathol.* **12**, 4119–4124 (2019).
102. Nonaka, T. et al. Involvement of activation-induced cytidine deaminase in skin cancer development. *J. Clin. Invest.* **126**, 1367–1382 (2016).
103. Sawai, Y. et al. Activation-induced cytidine deaminase contributes to pancreatic tumorigenesis by inducing tumor-related gene mutations. *Cancer Res.* **75**, 3292–3301 (2015).
104. Taylor, B. J. et al. DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *Elife* **2**, e00534 (2013).
105. Abyzov, A. et al. Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. *Nature* **492**, 438–442 (2012).
106. Zhou, Y. et al. Single-cell multiomics sequencing reveals prevalent genomic alterations in tumor stromal cells of human colorectal cancer. *Cancer Cell* **38**, 818–828.e5 (2020).
107. Hsu, E. et al. Lung tissues in patients with systemic sclerosis have gene expression patterns unique to pulmonary fibrosis and pulmonary hypertension. *Arthritis Rheum.* **63**, 783–794 (2011).
108. Renaud, L., da Silveira, W. A., Takamura, N., Hardiman, G. & Feghali-Bostwick, C. Prominence of IL6, IGF, TLR, and bioenergetics pathway perturbation in lung tissues of scleroderma patients with pulmonary fibrosis. *Front. Immunol.* **11**, 383 (2020).
109. Pedersen, B. S., Collins, R. L., Talkowski, M. E. & Quinlan, A. R. Indexcov: fast coverage quality control for whole-genome sequencing. *Gigascience* **6**, 1–6 (2017).
110. Koboldt, D. C. Best practices for variant calling in clinical sequencing. *Genome Med.* **12**, 91 (2020).
111. Saunders, C. T. et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).
112. Larson, D. E. et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311–317 (2012).
113. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017).
114. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, 33 (2018).
115. Zhang, H., Meltzer, P. & Davis, S. RCircos: an R package for Circos 2D track plots. *BMC Bioinforma.* **14**, 244 (2013).

Acknowledgements

We would like to thank all members of the Saini lab, Zavadil lab, and Feghali-Bostwick lab for their helpful comments and suggestions. This work has been supported by a National Scleroderma Foundation grant awarded to N.S. and NIH/NIAMS K24AR060297 and R01 HL153195 awarded to C.F.-B.

Author contributions

N.S., C.F.B., S.V. were involved in the study conceptualization, design, implementation. C.F.B., N.S., J.Z. provided resources and materials. S.V., N.S., J.M., T.B., L.P., F.V., J.Z. performed data generation, analysis, and interpretation. S.V., N.S. wrote the original draft of the manuscript. C.F.B., S.V., N.S., J.Z. performed editing and revisions.

Competing interests

The authors declare no competing interests.

Ethics statement

Lung tissues were obtained under a protocol approved by the University of Pittsburgh Institutional Review Board. Written informed consent statements were provided by the study participants/patients. The study was designed and conducted in compliance with all relevant regulations regarding the use of human study participants, and in accordance to the criteria set by the Declaration of Helsinki.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-53332-z.

Correspondence and requests for materials should be addressed to Natalie Saini.

Peer review information *Nature Communications* thanks Fran Supek, Rémi Buisson, and the other, anonymous, reviewer for their contribution to the peer review of this work. A peer review file is available.