Integrative studies of genetic and environmental factors with fibroblasts identify polymorphisms of TNFAIP3 in association with MMP expression

Journal:	*Arthritis & Rheumatology*
Manuscript ID:	ar-15-0325.R1
Wiley - Manuscript type:	Full Length
Date Submitted by the Author:	02-Jul-2015
Complete List of Authors:	Wei, Peng; University of Texas Health Science Center at Houston - School of Public Healt, Human Genetics Center and Department of Biostatistics Yang, Yang; University of Texas Health Science Center at Houston - School of Public Healt, Human Genetics Center and Department of Biostatistics Guo, Xinjian; University of Texas Health Sciences Center, Internal Medicine Hei, Nainan; University of Texas Health Science Center at Houston - School of Public Healt, Human Genetics Center and Department of Biostatistics Lai, Syeling; Baylor College of Medicine, Pathology Assassi, Shervin; University of Texas Health Sciences Center, Internal Medicine Liu, Mengyuan; University of Texas Health Sciences Center, Internal Medicine Tan, Filemon; University of Texas-Houston, Department of Internal Medicine Zhou, Xiaodong; University of Texas Health Sciences Center, Internal Medicine
Keywords:	Biostatistics, Fibroblasts, Dermal, Collagen, Metalloproteinase, Scleroderma
<B>Disease Category: Please select the category from the list below that best describes the content of your manuscript.:	Systemic Sclerosis

Integrative studies of scleroderma-associated genetic and environmental factors with fibroblasts identify polymorphisms of TNFAIP3 in association with MMP expression

Peng Wei[a*], Yang Yang[a,b*], Xinjian Guo[b], Nainan Hei[a], Syeling Lai[c], Shervin Assassi[b], Mengyuan Liu[b], Filemon Tan[b] and Xiaodong Zhou[b¶]

* These authors contributed equally to this work

[a] Human Genetics Center and Department of Biostatistics, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX 77030

[b] Division of Rheumatology, Department of Internal Medicine, University of Texas Health Science Center at Houston, Houston, TX 77030

[c] Department of Pathology, Baylor College of Medicine, Houston, TX 77030

[¶]Corresponding author:

Xiaodong Zhou, Professor

Division of Rheumatology

Department of Internal Medicine

University of Texas Health Science Center at Houston

Houston, TX 77030

Email: xiaodong.zhou@uth.tmc.edu

Phone: 713-500-6088

There is no financial support or other benefits from commercial sources for the work reported on in the manuscript

Abstract:

Objectives:

Systemic Sclerosis (SSc) is a fibrotic disease attributed to both genetic susceptibility and environmental factors. Our studies tried to demonstrate how human fibroblasts with SSc associated genetic variants respond to time-course and dose-response expression of the extracellular matrix (ECM) genes with silica particle stimulation.

Methods:

A total of 200 fibroblast strains were examined for ECM gene expression after stimulation by silica particles. The fibroblasts were genetically profiled with Immunochip assays, and followed by whole-genome genotype imputation. Associations of genotypes and gene expressions were first analyzed in a Caucasian cohort, and then validated by a meta-analysis which combines the results from Caucasian, Blacks and Hispanics. We applied the linear mixed model for longitudinal data analysis to identify genetic variants associated with ECM genes' expressions; we implemented haplotype-based longitudinal association test on identified loci region as a validation approach.

Results:

SNP rs58905141 of *TNFAIP3* was consistently associated with time-course and/or dose-response expressions of *MMP3* gene and *MMP1* gene of the fibroblasts stimulated with silica particles in both Caucasian only and meta-analysis. The haplotype-based analysis validated the association signals.

Conclusions:

A genetic variant of *TNFAIP3* is strongly associated with the silica-induced profibrotic response of the fibroblasts. *In silico* functional analysis based on ENCODE revealed that rs58905141 might affect binding activities of the transcription factors for *TNFAIP3*. This is the first genome-wide study of interaction between genetic and environmental factors in a complex SSc fibroblast model.

Keywords: Systemic Sclerosis, extracellular matrix gene, longitudinal study, *TNFAIP3*

Introduction

Systemic Sclerosis (SSc) is a multi-system disorder of connective tissue characterized by extensive cutaneous and visceral fibrosis. Although the etiopathogenesis of SSc is still unclear, both genetic susceptibility and environmental triggers are widely believed as two major contributors. Genetic association studies have reported a number of SSc-associated genes including HLA genes (e.g. *HLA-DQB1*, *-DPB1* and *-DRB1*), *STAT4, IRF5, CD247, TBX21, BANK1, C8orf13, PTPN22, TNFSF4, FAS, TNFAIP3, CD226, IRAK1, MECP2, MIF, ITGAM, PLD4, TLR-2, CAV1, IL2RA, NLRP1, OPN, PXK, JAZF, KIAA0319, IL-6, IL-21, CXCL8, CSK, PSD3, NFKB1, XRCC1, XRCC4, CCR6, IRF8, GRB10, SOX5, NOTCH4, TNIP1, PSOR1C1, RHOB, DNASEIL3, SCHIP1/IL12A and ATG5* (1-27). On the other hand, studies of environmental hazards contributing to SSc were mainly suggested by occupations with high incidence of SSc. Silica particles have been implicated as an environmental trigger of SSc in several epidemiological studies. Particularly, stonemasons and gold miners who were often exposed to silica particles were more likely to develop SSc (28, 29).

Fibrosis in connective tissues is the most prominent feature of SSc. A major component of connective tissues is fibroblasts. Fibroblasts synthesize, secrete, and maintain all major

extracellular matrix (ECM) components that are significantly increased in fibrotic tissues, which directly contribute to fibrosis. Perturbed expression of some major ECM genes have been observed in fibroblasts of SSc patients (30, 31). In particular, *COL1A2* and *COL3A1* are two major genes encoding structure proteins collagen type I and III, respectively. These collagens are frequently over-expressed in fibroblasts, and are accumulated in fibrotic tissue of SSc patients. Connective tissue growth factor (*CTGF*) is a profibrotic cytokine that is also over-expressed in SSc patients, and it induces ECM production. Matrix metalloproteinases (*MMP*) are antifibrotic molecules that induce the degradation of the collagens and other ECM components. Unbalanced expressions of MMP and tissue inhibitor of metalloproteinases (*TIMP*) are involved in fibrotic process in SSc (30, 31). Our hypothesis is that inherited genetic variations modulate cellular responses to the environment, and in some situations in a pathologic manner. The purpose of this paper is to use a systems-based approach in a fibroblast model that integrates genome-wide genetic variants and the expression of some fibrosis-associated ECM genes to investigate their interactive responses to environmental perturbation.

Methods

Fibroblasts

We examined 200 fibroblast strains obtained from skin biopsies (upper arm of non-lesional skin) of 96 SSc cases and 104 controls. The primary cultures were maintained in Dulbecco's Modified Essential Media (DMEM) with 10% FBS and supplemented with antibiotic and antimycotic as described previously (32). The 5th passage of fibroblast strains were plated at a density of 2.5 x 10^5 cells in a 35 mm dish and grown until 80% confluence. The culture medium was replaced with DMEM without FBS before stimulation assays.

All patients fulfilled American College of Rheumatology criteria for SSc (33). Among them, there were 37 limited and 53 diffused form of SSc (six patients were not reported in skin types). Autoantibodies positivity showed twenty patients with anti-topoisomerase I, twelve with anti-centromere, seventy with anti-RNA polymerase III. For disease duration, forty-nine patients were less than 5 years, eighteen were between 5 and 10 years, seventeen were 11 to 20 years, twelve were unknown. All normal controls were individuals with no history of autoimmune diseases. The study was approved by the Committee for the Protection of Human Subjects at University of Texas Health Science Center at Houston.

Silica stimulation on fibroblasts

Culture media then were replaced with FCS-free DMEM containing different doses (1, 5, 10, 25 and 50 μg) of silica particles obtained from Sigma-Aldrich. The cultures also were grouped into different time-courses including 24-, 48-, 72-, 96-, 120-hours stimulation with 10 μg silica. After stimulation, the fibroblasts were harvested for extraction of RNA. The RNAs were examined with RT-PCR for gene expression of *COL1A2*, *COL3A1*, *CTGF*, *MMP1*, *MMP3* and *TIMP3*. We used the "AQchange" (absolute quantification change) as the phenotype variable in the genetic association analysis, defined as $log2(\frac{\text{gene expression after stimulation at time } t}{\text{gene expression without stimulation at time } t})$. AQChange = 0 means gene expression remained the same after stimulation, while AQChange = 1 means gene expression was doubled after stimulation and AQChange = -1 means gene expression dropped by half after stimulation. AQchange can be any continuous value.

Genotyping of fibroblasts from each subject

Samples were genotyped using the Immunochip, an Illumina Infinium platform designed to densely genotype 186 immune-mediated disease loci identified in prior GWAS of autoimmune

diseases (26), according to the manufacturer's recommendations. Bead intensity data was processed and normalized for each sample in Genome Studio; data of successfully genotyped samples was extracted and genotypes were called within collections using optiCall (34). We obtained 196,517 SNPs before quality control (QC) as described below. NCBI build 36 (hg18) mapping was used in factory assembly file (Illumina manifest file Immuno_BeadChip_11419691_B.bpm); we then re-mapped all the SNPs to NCBI build 37 (hg19) locations.

Measurement of the ECM gene expression

Quantitative real time RT-PCR was performed using an ABI 7900 sequence detector (Applied Biosystems) for gene expression of human *COL1A2*, *COL3A1*, *CTGF*, *MMP1*, *MMP3* and *TIMP3* using standard protocol described previously (32, 35). The specific primers and probes for each gene were purchased through Assays-on-Demand from Applied Biosystems.

Linear mixed model to detect time-course and dose-response loci

We used the linear mixed model (LMM) to test genome-wide associations between SNPs and time-course/elevated-dosage responded mRNA levels, as measured by the quantitative real time RT-PCR, for each of the six ECM genes. The LMM takes into account the correlations between repeated measurements within the same subject by introducing population-level fixed effects and subject-specific random effects (random intercept and slope), and enjoys the parsimonious modeling of correlated outcomes and the ease of accommodating missing values (36). To correct for multiple hypotheses testing, we used the Bonferroni procedure with a genome-wide significance cut-off of 5.72×10^{-8} adjusting for a total of 874,949 genotyped and 1000 Genomes-imputed SNPs with non-missing p-values from the LMM. After we performed the race-specific association analyses by the LMM, we performed fixed-effect meta-analysis as

well as random-effect meta-analysis if determined necessary, combining results from the Caucasians, Hispanics and African Africans. Details on the statistical methods are provided in the Supplemental Methods, including QC procedures, 1000 Genomes-based imputation, LMM model specification, meta-analysis, haplotype-based association analysis, functional annotation and power calculation.

Results

Initial analysis

After stringent quality control measures, 183 subjects including 85 cases and 98 controls were included in the final association analyses; see Supplemental Methods for details. Demographic information of the study subjects is shown Table 1. We first analyzed all Caucasians among the subjects (50 SSc cases and 65 controls) as a discovery cohort (Table S1). Using LMM model assuming an additive genetic model, we found several genetic loci with SNPs significantly associated with expression of the ECM genes in this cohort (Table 2). Among them, three previously reported SSc loci including multiple SNPs of *TNFAIP3* (6q23) (10, 37), *IL2RA* (10p15-p14) (15) and *ITGAM* (16p11.2) (12, 38) showed strong associations with the expression of *MMP1* and/or *MMP3* of the fibroblasts in responses to dosage-dependent and/or time-course stimulations of silica particles (Table 2). SNPs of *CASC9* (8q21.11), intergenic region between *LINC00284* and *SMIM2-AS1* (13q14.11), *FGFR1OP* (6q27) and *THADA* (2p21) also showed associations with the expression of specific ECM genes of the fibroblasts (Table 2), but these genes have not been reported as SSc-associated.

Rs79411652 in the upstream of *TNFAIP3* was associated with *MMP1* dose-response, *MMP3* dose-response and time-course response. The minor allele frequency (MAF) of rs79411652 was 3.9% and 3.3%, respectively, in the Caucasian and African American samples.

However, the effect size of this SNP was quite large as shown in Figure 4A, B and C, with 3.6 to 12.7 fold change of gene expression between heterozygous alleles and homozygous major allele. Thus, we had high enough power (over 80%) to detect this association. In addition, regional plots, as shown in Figure 3A, B and C, indicate that there was a cluster of SNPs in high LD with rs79411652 showing association signals, which strengthen the findings about the potential pleiotropic effects of *TNFAIP3* on *MMP1* and *MMP3*.

Figures 3, 4 and S3 also show regional plots and genotype-specific expression trajectories for other previously identified SSc loci. SNP rs41290329 located in the 3' untranslated region (UTR) of *IL2RA* was significantly associated with *MMP1* dose-response, *MMP3* dose-response and time-course expression profiles (Figure 3D, E and F) and with 9.5 to 39 fold change of gene expression of *MMP1*/*MMP3* between heterozygous alleles and homozygous major allele (Figure 4D, E and F). Of note, rs41290329 is a low frequency SNP with an MAF of 1.3% in Caucasians and monomorphic in African Americans and Hispanics in our samples as well as in the 1000 Genomes Project. In addition, SNP rs12926702 was significantly associated with *MMP3* dose-response expression profile (Figure S3 H) with 34 fold change of gene expression of *MMP3* between heterozygous alleles and homozygous major alleles. Rs12926702 is located at the intergenic region of *TRIM72* and *ITGAM*. It is also a low frequency SNP with a MAF of 1.4% and 1.6% in Caucasians and African Americans, respectively, and monomorphic in our Hispanics samples. Given the low MAF of the peak SNPs and the current moderate sample size, we caution that future studies with larger sample size are warranted to confirm the signals identified here and obtain more reliable effect size estimates.

To evaluate whether the association signals of *TNFAIP3* and *IL2RA* could be better explained by multiple associated SNPs in LD or single peak SNPs, we performed haplotype-

based analysis of the chromosomal regions defined by 200 SNPs upstream and downstream of the peak SNPs, rs79411652 of *TNFAIP3* and rs41290329 of *IL2RA*. We defined haplotypes based on 3, 5 or 7 consecutive SNPs in sliding windows. The peak signals of haplotype-based analyses agreed well with those of single SNP-based association analyses, although the former yielded no as strong signals as the latter (Figure S2).

In addition to the signals within or near SSc-associated genes, we also identified several genetic loci that have not been reported to be associated with SSc (Table 2), including rs7823944 of *CASC9* associated with *COL3A1* dose-response expression (Figure S3 A), rs79365263 in the intergenic region of *LINC00284* and *SMIM2-AS1* associated with MMP1 and MMP3 dose-response expression (Figure S3 D and G), rs78409037 of *THADA* and rs75701002 of *FGFR1OP* associated with MMP3 time-course expression (Figure S3 I and K). In particular, rs7823944 of *CASC9* was supported by a cluster of imputed SNPs in high LD (r^2 around 0.80) (Table 2 and Figure S3 A), while rs79365263 of *LINC00284|SMIM2-AS1* was clustered with a number of genotyped SNPs in high LD (Figure S3 D and G).

HLA class II genes are well-documented in association with SSc. The nominal association signals ($5 \times 10^{-8} < P < 5 \times 10^{-2}$) of HLA class II genes were observed in the Caucasian cohort. Other SSc-loci did not show significant association with gene expression of the fibroblasts.

As elaborated in Supplemental Methods, the standard normal distribution-based p-values for the SNP effect in the LMM can be too liberal, i.e., not controlling the Type I error at the nominal level. We therefore calculated the more accurate t-distribution-based p-values using the computationally demanding Satterthwaite approximation for the peak SNPs in Table 2. As

shown in Table S2, although all of the p-values became larger as expected, the peak SNPs in the SSc-associated loci of *TNFAIP3*, *IL2RA* and *ITGAM*, remained significant.

Meta-analysis of Caucasian, African American and Hispanic cohorts

We conducted meta-analysis of the Caucasian (50 SSc patients and 65 controls), African American (13 SSc patients and 23 controls) and Hispanics (22 SSc patients and 10 controls) cohorts via the fixed-effect inverse-variance weighting method as well as random-effect meta-analysis if determined necessary; see the Supplemental Methods for details. Of note, in the presence of potential heterogeneous genetic effects across the three populations, fixed-effect meta-analysis may not be appropriate. To investigate this, we calculated the heterogeneity measure I^2 and Cochran's Q test p-value (Table 2), which indicated that most of genome-wide significant SNPs in the Caucasian cohort, might have heterogeneous genetic effects across different populations ($I^2 > 56\%$) (39). We, therefore, also performed random-effect meta-analysis for these SNPs. Noticeably, SNP rs58905141 of *TNFAIP3* remained genome-wide significantly associated with *MMP3* dose-response expression by both fixed-effect and random-effect meta-analysis, even though I^2 and Cochran's test results suggested that the random-effect model was not necessary for this SNP. Other SNPs did not show significance with random-effect meta-analysis. The Manhattan plots for fixed-effect and random-effect meta-analysis are, respectively, shown in Figure 1 and Figure S1. Meta-analysis results for all the genome-wide significant SNPs in the discovery analysis of the Caucasian samples are shown in Table S3. The non-significant random-effect meta-analysis results for most of the peak SNPs in Caucasians may be attributed to the low MAF of the SNPs, smaller sample sizes in the African American and Hispanic cohorts and disparate LD patterns across populations, leading to diminished statistical power.

Discussion

SSc is a fibrotic disease with complex genetic traits, to which environmental factors may trigger a pathological process toward systemic fibrosis. Functional association studies of genetic polymorphisms in connection with environmental triggers are challenge. It is unknown whether the SSc susceptibility loci contribute to specific fibrotic process in the presence of an environmental hazard. In comparison with conventional gene by environment (GxE) analysis based on observational studies (40, 41), the studies herein are novel in that we examined profibrotic effects of silica particles in a complex cell model of human fibroblasts that were genetically profiled with Immuno-chip SNPs.

A previously reported SSc locus at 6q23.3 with SNPs of *TNFAIP3* (10, 37) appeared to be persistently in association with expression of *MMP1* and/or *MMP3* genes of the fibroblasts in responses to dosage-dependent and time-course stimulations of silica particles in both Caucasian only analysis and meta-analysis of diverse populations (Table 2). *TNFAIP3* stands for tumor necrosis factor-α-induced-protein 3. It encodes A20, an ubiquitin-modifying enzyme that inhibits NF-kB activity and is a key regulator of TNF-mediated immune responses and inflammation (42, 43). Inflammation has been reported as a major consequence induced by silica particles in *in vivo* and *in vitro* studies (35, 44). Our previous studies indicated that silica might activate fibroblasts to overexpress the ECM genes through proinflammatory process (32, 35). Therefore, the results suggest that an interaction between genetic factors of *TNFAIP3* and environmental factor of silica stimulation may control fibroblasts to express *MMP1*/*MMP3* genes, which may be associated with development of SSc through an inflammatory and immune-control mechanism. Of note, intronic SNP rs5029939 in *TNFAIP3* was reported to be associated with SSc in a previous case-control study (10). LD analysis by the SNAP tool (45) indicated that this SNP was in perfect LD ($r^2 = 1$ and D'=1) with our top hit SNP rs79411652 in the discovery analysis of

Caucasian samples, and in high LD with our meta-analysis top hit SNP rs58905141 (r^2 = 0.80 and D'=1) associated with *MMP3* dose-response and time-course expression as well as *MMP1* dose-response expression. Regional plots (Figure 3A, C and E) also suggest that these SNPs were likely to tag the same functional locus/loci. We annotated the top SNPs of *TNFAIP3* with RegulomeDB (46) according to the ENCODE database (47). While rs79411652 was annotated to category 7, i.e., no known biological function, rs58905141 was annotated to the second highest functional category "2b", which is likely to affect binding of transcription factors FOXA1 and SMARCA4 by ChIP-seq experiments, binding of STAT1 by motif analysis, and is in DNase footprint by DNase-seq and in DNase peak by ChIP-seq experiments. This suggests that rs58905141 may play a functional role in regulating *TNFAIP3* expression and directly or indirectly regulate *MMP1*/*MMP3* expression in fibroblasts. Further replication and functional studies of the signals in TNFAIP3 are warranted to confirm this hypothesis.

Two other reported SSc-associated genetic loci with SNPs of *IL2RA* (10p15-p14) and *ITGAM* (16p11.2) genes were significantly associated with the expression of *MMP1* and/or *MMP3* in silica response assays in Caucasians (Table 2). *IL2RA* stands for interleukin 2 receptor α. IL2RA is involved in various pathways that help control the differentiation of effector cells, T-cell proliferation, and immune tolerance (48). A meta-analysis study with a large Caucasian European cohort revealed an association between rs2104286 of *IL2RA* and ACA positive SSc patients (15). LD analysis using the SNAP software showed that *MMP1/MMP3* expression associated rs41290329 was in complete LD with rs2104286 (D' = 1), but in low LD as measured by r^2 (=0.06), due to the disparate MAFs of the two SNPs (0.013 and 0.225 in the 1000 Genomes CEU population, respectively). Functional analysis by RegulomeDB indicated that biological function for rs41290329 was unknown. As for *ITGAM* that encodes the α subunit of the αMβ2-

integrin, it has been recently identified as an autoimmune disease risk gene (49). It is expressed on the surface of leukocytes, and it regulates adhesion of neutrophils and monocytes, cell activation, which is important for innate immunity (50). Rs1143679 of *ITGAM* was associated with SSc in European cohorts (12) and a large-scale meta-analysis (38). LD analysis suggested that *MMP3* dose-response expression associated rs12926702 was in complete LD with rs1143679 (D' = 1), but in low LD as measured by r^2 (=0.003), due to the low MAFs of both SNPs (0.033 and 0.083 in the 1000 Genomes CEU population, respectively). Considering the low MAFs of rs41290329 and rs12926702, in addition to our moderate sample size, especially for non-Caucasian samples, we caution that the findings on *IL2RA* and *ITGAM* need to be confirmed in future studies with larger sample size.

In addition, several other genetic loci and genes unrelated to SSc susceptibility were also strongly associated with expression of the ECM genes of silica-stimulated fibroblasts in Caucasian-only analysis and/or meta-analysis (Table 2). Whether these associations represent a general impact of specific genetic loci on immune-mediated diseases is worth further investigation.

It is worth noting that many of those known SSc-loci did not show significant impact on gene expression of the fibroblasts in this study. For instance, the HLA class II genes that confer the strongest susceptibility to SSc (1) showed only nominal association signals (see Figure 2), which suggest that potential genetic impact of these genes to profibrotic responses of silica-stimulated fibroblasts may be less significant in this study model. It is likely that genetic impact of specific SSc-loci may be determined by interactions of specific types of cells and environmental triggers, as well as specific responding genes.

The current study employed the cost-effective Immunochip targeted at 186 gene regions which were identified in previous association studies of autoimmune diseases. Although we successfully performed 1000 Genomes-based imputation to boost the 130,000 directly genotyped SNPs to around 800,000 total SNPs, the coverage was likely to be very low beyond the 186 gene regions. As a result, we cannot exclude the possibility that other genetic variants that were not investigated in the current study might be associated with the ECM gene expression in the fibroblasts. In addition, heterogeneity of patients' clinical status is another concern. This warrants further investigation in the future.

In summary, this is the first attempt to study interactions between genome-wide genetic variants and environmental factors of SSc in a complex fibroblast model. The results indicated that a previously identified SSc locus of *TNFAIP3* was strongly and persistently associated with silica-induced profibrotic responses of the fibroblasts in both Caucasians and meta-analysis of mixed populations. Similar associations were observed in two other reported SSc loci of *IL2RA* and *ITGAM* in Caucasians. In addition, some non-SSc loci may also impact a general response of fibroblasts to silica particles. This notion may be further verified in other cohorts with larger sample size and in functional studies beyond the ECM genes.

Acknowledgments

This work was supported by the grants from the Department of the Army, Medical Research Acquisition Activity [PR064803 to XD.Z]; Scleroderma Foundation [2012 to XD.Z]; and the National Institutes of Health [NIAID UO1, 1U01AI09090 to XD.Z]. P.W. was partially supported by NIH grants R01CA169122 and R01HL116720. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing

HPC resources that have contributed to the research results reported within this paper. The authors indicated that there is no conflict of interest.

References

1. Arnett FC, Gourh P, Shete S, Ahn CW, Honey RE, Agarwal SK, et al. Major histocompatibility complex (MHC) class II alleles, haplotypes and epitopes which confer susceptibility or protection in systemic sclerosis: analyses in 1300 Caucasian, African-American and Hispanic cases and 1000 controls. Ann Rheum Dis. 2010;69(5):822-7.
2. Rueda B, Broen J, Simeon C, Hesselstrand R, Diaz B, Suarez H, et al. The STAT4 gene influences the genetic predisposition to systemic sclerosis phenotype. Hum Mol Genet. 2009;18(11):2071-7.
3. Dieude P, Guedj M, Wipff J, Avouac J, Fajardy I, Diot E, et al. Association between the IRF5 rs2004640 functional polymorphism and systemic sclerosis: a new perspective for pulmonary fibrosis. Arthritis Rheum. 2009;60(1):225-33.
4. Radstake TR, Gorlova O, Rueda B, Martin JE, Alizadeh BZ, Palomino-Morales R, et al. Genome-wide association study of systemic sclerosis identifies CD247 as a new susceptibility locus. Nature genetics. 2010;42(5):426-9.
5. Gourh P, Agarwal SK, Divecha D, Assassi S, Paz G, Arora-Singh RK, et al. Polymorphisms in TBX21 and STAT4 increase the risk of systemic sclerosis: evidence of possible gene-gene interaction and alterations in Th1/Th2 cytokines. Arthritis Rheum. 2009;60(12):3794-806.
6. Dieude P, Wipff J, Guedj M, Ruiz B, Melchers I, Hachulla E, et al. BANK1 is a genetic risk factor for diffuse cutaneous systemic sclerosis and has additive effects with IRF5 and STAT4. Arthritis Rheum. 2009;60(11):3447-54.
7. Gourh P, Agarwal SK, Martin E, Divecha D, Rueda B, Bunting H, et al. Association of the C8orf13-BLK region with systemic sclerosis in North-American and European populations. Journal of autoimmunity. 2010;34(2):155-62.
8. Gourh P, Arnett FC, Tan FK, Assassi S, Divecha D, Paz G, et al. Association of TNFSF4 (OX40L) polymorphisms with susceptibility to systemic sclerosis. Ann Rheum Dis. 2010;69(3):550-5.
9. Broen J, Gourh P, Rueda B, Coenen M, Mayes M, Martin J, et al. The FAS -670A>G polymorphism influences susceptibility to systemic sclerosis phenotypes. Arthritis Rheum. 2009;60(12):3815-20.
10. Dieude P, Guedj M, Wipff J, Ruiz B, Riemekasten G, Matucci-Cerinic M, et al. Association of the TNFAIP3 rs5029939 variant with systemic sclerosis in the European Caucasian population. Ann Rheum Dis. 2010;69(11):1958-64.
11. Dieude P, Bouaziz M, Guedj M, Riemekasten G, Airo P, Muller M, et al. Evidence of the contribution of the X chromosome to systemic sclerosis susceptibility: association with the functional IRAK1 196Phe/532Ser haplotype. Arthritis Rheum. 2011;63(12):3979-87.
12. Carmona FD, Simeon CP, Beretta L, Carreira P, Vonk MC, Rios-Fernandez R, et al. Association of a non-synonymous functional variant of the ITGAM gene with systemic sclerosis. Annals of the rheumatic diseases. 2011;70(11):2050-2.

13. Terao C, Ohmura K, Kawaguchi Y, Nishimoto T, Kawasaki A, Takehara K, et al. PLD4 as a novel susceptibility gene for systemic sclerosis in a Japanese population. Arthritis Rheum. 2013;65(2):472-80.
14. Manetti M, Allanore Y, Saad M, Fatini C, Cohignac V, Guiducci S, et al. Evidence for caveolin-1 as a new susceptibility gene regulating tissue fibrosis in systemic sclerosis. Ann Rheum Dis. 2012;71(6):1034-41.
15. Martin JE, Carmona FD, Broen JC, Simeon CP, Vonk MC, Carreira P, et al. The autoimmune disease-associated IL2RA locus is involved in the clinical manifestations of systemic sclerosis. Genes Immun. 2012;13(2):191-6.
16. Dieude P, Guedj M, Wipff J, Ruiz B, Riemekasten G, Airo P, et al. NLRP1 influences the systemic sclerosis phenotype: a new clue for the contribution of innate immunity in systemic sclerosis-related fibrosing alveolitis pathogenesis. Ann Rheum Dis. 2011;70(4):668-74.
17. Barizzone N, Marchini M, Cappiello F, Chiocchetti A, Orilieri E, Ferrante D, et al. Association of osteopontin regulatory polymorphisms with systemic sclerosis. Human immunology. 2011;72(10):930-4.
18. Martin JE, Assassi S, Diaz-Gallo LM, Broen JC, Simeon CP, Castellvi I, et al. A systemic sclerosis and systemic lupus erythematosus pan-meta-GWAS reveals new shared susceptibility loci. Hum Mol Genet. 2013;22(19):4021-9.
19. Diaz-Gallo LM, Simeon CP, Broen JC, Ortego-Centeno N, Beretta L, Vonk MC, et al. Implication of IL-2/IL-21 region in systemic sclerosis genetic susceptibility. Ann Rheum Dis. 2013;72(7):1233-8.
20. Salim PH, Jobim M, Bredemeier M, Chies JA, Brenol JC, Jobim LF, et al. Combined effects of CXCL8 and CXCR2 gene polymorphisms on susceptibility to systemic sclerosis. Cytokine. 2012;60(2):473-7.
21. Martin JE, Broen JC, Carmona FD, Teruel M, Simeon CP, Vonk MC, et al. Identification of CSK as a systemic sclerosis genetic risk factor through Genome Wide Association Study follow-up. Hum Mol Genet. 2012;21(12):2825-35.
22. Palomino GM, Bassi CL, Wastowski IJ, Xavier DJ, Lucisano-Valim YM, Crispim JC, et al. Patients with systemic sclerosis present increased DNA damage differentially associated with DNA repair gene polymorphisms. The Journal of rheumatology. 2014;41(3):458-65.
23. Koumakis E, Bouaziz M, Dieude P, Ruiz B, Riemekasten G, Airo P, et al. A regulatory variant in CCR6 is associated with susceptibility to antitopoisomerase-positive systemic sclerosis. Arthritis Rheum. 2013;65(12):3202-8.
24. Allanore Y, Saad M, Dieude P, Avouac J, Distler JH, Amouyel P, et al. Genome-wide scan identifies TNIP1, PSORS1C1, and RHOB as novel risk loci for systemic sclerosis. PLoS genetics. 2011;7(7):e1002091.
25. Gorlova O, Martin JE, Rueda B, Koeleman BP, Ying J, Teruel M, et al. Identification of novel genetic markers associated with clinical phenotypes of systemic sclerosis through a genome-wide association strategy. PLoS genetics. 2011;7(7):e1002178.
26. Mayes MD, Bossini-Castillo L, Gorlova O, Martin JE, Zhou X, Chen WV, et al. Immunochip analysis identifies multiple susceptibility loci for systemic sclerosis. American journal of human genetics. 2014;94(1):47-61.
27. Jin J, Chou YC, Lima M, Zhou Z, Zhou X. Systemic sclerosis is a complex genetic disease associated mainly with immune regulatory and inflammatory genes. Manuscript under review. 2014.

28. Slimani S, Ben Ammar A, Ladjouze-Rezig A. Connective tissue diseases after heavy exposure to silica: a report of nine cases in stonemasons. Clin Rheumatol. 2010;29(5):531-3.
29. Cowie RL. Silica-dust-exposed mine workers with scleroderma (systemic sclerosis). Chest. 1987;92(2):260-2.
30. Trojanowska M. Molecular aspects of scleroderma. Frontiers in bioscience : a journal and virtual library. 2002;7:d608-18.
31. Kuroda K, Shinkai H. Gene expression of types I and III collagen, decorin, matrix metalloproteinases and tissue inhibitors of metalloproteinases in skin fibroblasts from patients with systemic sclerosis. Archives of dermatological research. 1997;289(10):567-72.
32. Xiong M, Arnett FC, Guo X, Xiong H, Zhou X. Differential dynamic properties of scleroderma fibroblasts in response to perturbation of environmental stimuli. PLoS One. 2008;3(2):e1693.
33. Subcommittee for scleroderma criteria of the American Rheumatism Association Diagnostic and Therapeutic Criteria Committee. Preliminary criteria for the classification of systemic sclerosis (scleroderma). Subcommittee for scleroderma criteria of the American Rheumatism Association Diagnostic and Therapeutic Criteria Committee. Arthritis Rheum. 1980;23(5):581-90.
34. Shah TS, Liu JZ, Floyd JA, Morris JA, Wirth N, Barrett JC, et al. optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants. Bioinformatics. 2012;28(12):1598-603.
35. Guo X, Jagannath C, Espitia MG, Zhou X. Uptake of silica and carbon nanotubes by human macrophages/monocytes induces activation of fibroblasts in vitro -- potential implication for pathogenesis of inflammation and fibrotic diseases. Int J Immunopathol Pharmacol. 2012;25(3):713-9.
36. Fitzmaurice G, Laird N, Ware J. Applied Longitudinal Analysis. 2nd ed. Hoboken, New Jersey: John Wiley & Sons; 2011.
37. Koumakis E, Giraud M, Dieude P, Cohignac V, Cuomo G, Airo P, et al. Brief report: candidate gene study in systemic sclerosis identifies a rare and functional variant of the TNFAIP3 locus as a risk factor for polyautoimmunity. Arthritis Rheum. 2012;64(8):2746-52.
38. Anaya JM, Kim-Howard X, Prahalad S, Chernavsky A, Canas C, Rojas-Villarraga A, et al. Evaluation of genetic association between an ITGAM non-synonymous SNP (rs1143679) and multiple autoimmune diseases. Autoimmun Rev. 2012;11(4):276-80.
39. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21(11):1539-58.
40. Thomas D. Gene--environment-wide association studies: emerging approaches. Nat Rev Genet. 2010;11(4):259-72.
41. Hsu L, Jiao S, Dai JY, Hutter C, Peters U, Kooperberg C. Powerful cocktail methods for detecting genome-wide gene-environment interaction. Genet Epidemiol. 2012;36(3):183-94.
42. Ma A, Malynn BA. A20: linking a complex regulator of ubiquitylation to immunity and human disease. Nat Rev Immunol. 2012;12(11):774-85.
43. Shembade N, Harhaj NS, Parvatiyar K, Copeland NG, Jenkins NA, Matesic LE, et al. The E3 ligase Itch negatively regulates inflammatory signaling pathways by controlling the function of the ubiquitin-editing enzyme A20. Nat Immunol. 2008;9(3):254-62.
44. Park EJ, Park K. Oxidative stress and pro-inflammatory responses induced by silica nanoparticles in vivo and in vitro. Toxicol Lett. 2009;184(1):18-25.

45. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics. 2008;24(24):2938-9.
46. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012;22(9):1790-7.
47. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A. 2014;111(17):6131-8.
48. Shevach EM. Certified professionals: CD4(+)CD25(+) suppressor T cells. J Exp Med. 2001;193(11):F41-6.
49. Nath SK, Han S, Kim-Howard X, Kelly JA, Viswanathan P, Gilkeson GS, et al. A nonsynonymous functional variant in integrin-alpha(M) (encoded by ITGAM) is associated with systemic lupus erythematosus. Nature genetics. 2008;40(2):152-4.
50. Solovjov DA, Pluskota E, Plow EF. Distinct roles for the alpha and beta subunits in the functions of integrin alphaMbeta2. J Biol Chem. 2005;280(2):1336-45.

Figure Legends

Figure 1. Manhattan plots of fixed-effect meta-analysis association p-values for dose-response and time-course ECM gene expressions in fibroblasts in response to silica stimulation. A. *MMP1* gene expression in silica dose response showed significant signals at 6q23.3 (*TNFAIP3*), 10p15.1 (*IL2RA*) and 13q14.11 (*LINC00284*). B. *MMP3* gene expression in silica dose response showed significant signals at 6q23.3 (*TNFAIP3*), 10p15.1 (*IL2RA*), 13q14.11 (*LINC00284*) and 16p11.2 (*TRIM72|ITGAM*). C. *MMP3* gene expression in silica time course assays showed significant signals at 2p21(*THADA*), 6q23.3 (*TNFAIP3*) and 6q27(*FGFR10P|CCR6*). Red line corresponds to the p-value cutoff for genome-wide significance (5×10^{-8}); blue line corresponds to the suggestive significance cutoff (1×10^{-5}). P-values from Caucasian-only analysis are displayed if fixed-effect meta-analysis was not performed due to missing association results in African Americans and Hispanics.

Figure 2. Chromosome 6 Manhattan plots of fixed-effect meta-analysis association p-values for dose-response and time-course ECM gene expressions in fibroblasts in response to silica stimulation. Panels A, B and C correspond to Figure 1 A, B and C with a zoom in for chromosome 6 only. Green dots show loci on or near known SSc risk genes. Red line

corresponds to the p-value cutoff for genome-wide significance (5×10^{-8}); blue line corresponds to the suggestive significance cutoff (1×10^{-5}). P-values from Caucasian-only analysis are displayed if fixed-effect meta-analysis was not performed due to missing association results in African Americans and Hispanics.

Figure 3. Regional plots of association results between previously identified SSc risk loci and gene expression of fibroblasts in the Caucasian cohort. 500kb upstream and downstream of the peak SNPs are plotted. Panels A to C: genetic locus at 6q23.3 containing SNPs of *TNFAIP3* in association with expression of (A) *MMP1* in dose-response assays, (B) *MMP3* in dose-response assays, and (C) *MMP3* in time-course assay. Panels D to F: genetic locus at 10p15.1 containing SNPs of *IL2RA* in association with expression of (D) *MMP1* in dose-response assays, (E) *MMP3* in dose-response assays, and (F) *MMP3* in time-course assays. Round dots represent genotyped SNPs while square dots represent imputed SNPs. Purple dots correspond to the peak SNPs. Linkage disequilibrium (r^2) with the peak SNP (rs79411652 of *TNFAIP3* and rs41290329 of *IL2RA*, respectively) is shown by different colors.

Figure 4. Dose-response and time-course expression trajectories of the ECM genes by genotypes of the peak SNPs in *TNFAIP3* and *IL2RA* in the Caucasian cohort. Panels A-F correspond to panels A-F in Figure 3. Red solid line: sample average trajectory for the heterozygous alleles; blue solid line: sample average trajectory for the homozygous major alleles; gray solid lines: individual trajectories; dashed horizontal grey line at 0 of the y-axis: the reference line representing constant gene expression across time/doses. Y-axis corresponds to the absolute quantity change (AQchange); x-axis corresponds to silica dosage (in μg) or time course (in days).

Table 1. Distribution of demographics among SSc cases and controls

Variable	SSc (n = 85) n (%)	Controls (n = 98) n (%)	P-value (χ^2)
Age group			
≤30	7 (8.24)	25 (25.51)	
31−40	14 (16.5)	23 (23.47)	
41−50	24 (28.24)	22 (22.45)	
>50	40 (47.06)	28 (28.57)	0.0034
Race[a]			
Caucasians	50 (58.82)	65 (66.33)	
Hispanics	22 (25.88)	10 (10.20)	
African Americans	13 (15.29)	23 (23.47)	0.015
Sex			
Female	67 (78.82)	60 (61.22)	
Male	18 (21.18)	38 (38.78)	0.016

[a]Self-reported race and corrected by STRUCTURE analysis

Table 2. Associations between expression of the ECM genes and genetic backgrounds in Caucasians and mixed populations by meta-analysis.

Peak SNP	Chr	Position[a]	Type	Gene(s)[b]	RA	P[f]	P-FE	P-RE	HetISq (%)	HetP	Direction[c]	Estimates(W|B|H)[e]	StdErr(W|AA|H)[e]	MAF(W|AA|H)[e]	responseType
rs58417815*	8	76159689	ncRNA_intronic	*CASC9*	C	5.10E-08	-[g]	-	-	-	+??	1.77|NA|NA	0.29|NA|NA	0.074|NA|NA	COL3A1_dose
rs7823944	8	76182788	ncRNA_intronic	*CASC9*	T	3.47E-08	2.00E-07	7.10E-02	51.20	0.13	+-+	1.54|-0.03|1.00	0.25|0.73|1.05	0.052|0.033|0.021	COL3A1_dose
rs79411652	6	138138945	intergenic	*OLIG3*|***TNFAIP3***[d]	C	3.87E-11	-	5.69E-01	88.80	0.00	+-?	2.29|-0.80|NA	0.32|0.97|NA	0.039|0.033|NA	MMP1_dose
rs58905141	6	138132123	intergenic	*OLIG3*|***TNFAIP3***	G	3.97E-11	3.46E-11	0.402918	43.1	0.17	++-	2.29|62.41|-2.18	0.32|32.09|65.44	0.039|0.014|0.016	MMP1_dose
rs41290329	10	6054083	UTR3	***IL2RA***	C	1.51E-08	-	-	-	-	+??	3.43|NA|NA	0.57|NA|NA	0.013|NA|NA	MMP1_dose
rs79365263	13	44618508	intergenic	*LINC00284*|*SMIM2-AS1*	C	3.13E-09	-	-	-	-	+??	2.22|NA|NA	0.35|NA|NA	0.024|NA|NA	MMP1_dose
rs79411652	6	138138945	intergenic	*OLIG3*|***TNFAIP3***	C	5.55E-18	-	5.12E-01	93.30	0.00	+-?	3.69|-0.91|NA	0.41|0.99|NA	0.039|0.033|NA	MMP3_dose
rs58905141	6	138132123	intergenic	*OLIG3*|***TNFAIP3***	G	6.43E-18	6.00E-18	1.32E-17	0	0.81	+++	3.69|28.75|22.87	0.41|34.61|128.32	0.039|0.014|0.016	MMP3_dose
rs41290329	10	6054083	UTR3	***IL2RA***	C	1.02E-11	-	-	-	-	+??	5.29|NA|NA	0.74|NA|NA	0.013|NA|NA	MMP3_dose
rs79365263	13	44618508	intergenic	*LINC00284*|*SMIM2-AS1*	C	5.33E-08	-	-	-	-	+??	2.80|NA|NA	0.49|NA|NA	0.024|NA|NA	MMP3_dose
rs12926702	16	31240971	intergenic	*TRIM72*|***ITGAM***	C	1.33E-10	-	3.42E-01	87.90	0.00	+-?	5.09|-0.35|NA	0.76|1.52|NA	0.014|0.016|NA	MMP3_dose
rs78409037	2	43740411	intronic	*THADA*	C	4.51E-05	-	9.14E-02	94.10	0.00	+?+	0.97|NA|3.63	0.23|NA|0.55	0.069|NA|0.041	MMP3_timecourse
rs77533229	2	43479638	intronic	*THADA***	G	2.68E-08	-	-	-	-	+??	3.35|NA|NA	0.60|NA|NA	0.010|NA|NA	MMP3_timecourse
rs79411652	6	138138945	intergenic	*OLIG3*|***TNFAIP3***	C	3.52E-10	-	7.42E-01	93.80	0.00	+-?	1.86|-1.02|NA	0.29|0.64|NA	0.039|0.033|NA	MMP3_timecourse
rs58905141	6	138132123	intergenic	*OLIG3*|***TNFAIP3***	G	3.73E-10	3.23E-10	0.446757	29.1	0.24	++-	1.85|38.02|-11.59	0.29|21.66|53.94	0.039|0.014|0.016	MMP3_timecourse
rs75701002	6	167443918	intronic	*FGFR1OP*	C	2.18E-08	5.68E-08	5.17E-02	47.50	0.17	+?+	2.39|NA|0.32	0.42|NA|1.32	0.026|NA|0.021	MMP3_timecourse
rs41290329	10	6054083	UTR3	***IL2RA***	C	1.03E-10	-	-	-	-	+??	3.25|NA|NA	0.47|NA|NA	0.013|NA|NA	MMP3_timecourse

RA: risk allele; P: P-value in Caucasians; P-FE: P-value by fixed-effect meta-analysis; P-RE: P-value in random-effect meta-analysis; HetISq: heterogeneity measure I^2; HetP: Cochran's heterogeneity test p-valie; Estimates/StdErr/MAF: per risk allele effect estimates/standard error/minor allele frequency in different ethnic groups; W|AA|H: Caucasians|African Americans|Hispanics; responseType: the ECM gene expression in response to silica stimulation in elevated-dosage or time-course.

[a]GRCh37/hg19 assembly; [b]Gene in which the peak signal is located in, or genes flanking peak signal if intergenic; [c]Direction shown in W|AA|H order, where "?" represents not applicable for that specific ethnic group; [d]Bold-face genes are known SSc risk loci; [e]"NA": not applicable for that specific ethnic group; [f]P-value based on standard normal approximation; [g]"-": meta-analysis was not performed when only one ethnicity was present.

*Imputed, otherwise Peak SNP was genotyped; **The peak SNP in Caucasians was different from that in fixed-effect meta-analysis with both SNPs listed.

Integrative studies of scleroderma-associated genetic and environmental factors with fibroblasts identify polymorphisms of TNFAIP3 in association with MMP expression

Peng Wei[a*], Yang Yang[a,b*], Xinjian Guo[b], Nainan Hei[a], Syeling Lai[c], Shervin Assassi[b], Mengyuan Liu[b], Filemon Tan[b] and Xiaodong Zhou[b¶]

* These authors contributed equally to this work

[a] Human Genetics Center and Department of Biostatistics, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX 77030

[b] Division of Rheumatology, Department of Internal Medicine, University of Texas Health Science Center at Houston, Houston, TX 77030

[c] Department of Pathology, Baylor College of Medicine, Houston, TX 77030

[¶]Corresponding author:

Xiaodong Zhou, Professor

Division of Rheumatology

Department of Internal Medicine

University of Texas Health Science Center at Houston

Houston, TX 77030

Email: xiaodong.zhou@uth.tmc.edu

Phone: 713-500-6088

Abstract:

Objectives:

Systemic Sclerosis (SSc) is a fibrotic disease attributed to both genetic susceptibility and environmental factors. Our studies tried to demonstrate how human fibroblasts with SSc associated genetic variants respond to time-course and dose-response expression of the extracellular matrix (ECM) genes with silica particle stimulation.

Methods:

A total of 200 fibroblast strains were examined for ECM gene expression after stimulation by silica particles. The fibroblasts were genetically profiled with Immunochip assays, and followed by whole-genome genotype imputation. Associations of genotypes and gene expressions were first analyzed in a Caucasian cohort, and then validated by a meta-analysis which combines the results from Caucasian, Blacks and Hispanics. We applied the linear mixed model for longitudinal data analysis to identify genetic variants associated with ECM genes' expressions; we implemented haplotype-based longitudinal association test on identified loci region as a validation approach.

Results:

SNP rs58905141 of *TNFAIP3* was consistently associated with time-course and/or dose-response expressions of *MMP3* gene and *MMP1* gene of the fibroblasts stimulated with silica particles in both Caucasian only and meta-analysis. The haplotype-based analysis validated the association signals.

Conclusions:

A genetic variant of *TNFAIP3* is strongly associated with the silica-induced profibrotic response of the fibroblasts. *In silico* functional analysis based on ENCODE revealed that rs58905141 might affect binding activities of the transcription factors for *TNFAIP3*. This is the first genome-wide study of interaction between genetic and environmental factors in a complex SSc fibroblast model.

Keywords: Systemic Sclerosis, extracellular matrix gene, longitudinal study, *TNFAIP3*

Introduction

Systemic Sclerosis (SSc) is a multi-system disorder of connective tissue characterized by extensive cutaneous and visceral fibrosis. Although the etiopathogenesis of SSc is still unclear, both genetic susceptibility and environmental triggers are widely believed as two major contributors. Genetic association studies have reported a number of SSc-associated genes including HLA genes (e.g. *HLA-DQB1*, *-DPB1* and *-DRB1*), *STAT4, IRF5, CD247, TBX21, BANK1, C8orf13, PTPN22, TNFSF4, FAS, TNFAIP3, CD226, IRAK1, MECP2, MIF, ITGAM, PLD4, TLR-2, CAV1, IL2RA, NLRP1, OPN, PXK, JAZF, KIAA0319, IL-6, IL-21, CXCL8, CSK, PSD3, NFKB1, XRCC1, XRCC4, CCR6, IRF8, GRB10, SOX5, NOTCH4, TNIP1, PSOR1C1, RHOB, DNASEIL3, SCHIP1/IL12A and ATG5* (1-27). On the other hand, studies of environmental hazards contributing to SSc were mainly suggested by occupations with high incidence of SSc. Silica particles have been implicated as an environmental trigger of SSc in several epidemiological studies. Particularly, stonemasons and gold miners who were often exposed to silica particles were more likely to develop SSc (28, 29).

Fibrosis in connective tissues is the most prominent feature of SSc. A major component of connective tissues is fibroblasts. Fibroblasts synthesize, secrete, and maintain all major

extracellular matrix (ECM) components that are significantly increased in fibrotic tissues, which directly contribute to fibrosis. Perturbed expression of some major ECM genes have been observed in fibroblasts of SSc patients (30, 31). In particular, *COL1A2* and *COL3A1* are two major genes encoding structure proteins collagen type I and III, respectively. These collagens are frequently over-expressed in fibroblasts, and are accumulated in fibrotic tissue of SSc patients. Connective tissue growth factor (*CTGF*) is a profibrotic cytokine that is also over-expressed in SSc patients, and it induces ECM production. Matrix metalloproteinases (*MMP*) are antifibrotic molecules that induce the degradation of the collagens and other ECM components. Unbalanced expressions of MMP and tissue inhibitor of metalloproteinases (*TIMP*) are involved in fibrotic process in SSc (30, 31). Our hypothesis is that inherited genetic variations modulate cellular responses to the environment, and in some situations in a pathologic manner. The purpose of this paper is to use a systems-based approach in a fibroblast model that integrates genome-wide genetic variants and the expression of some fibrosis-associated ECM genes to investigate their interactive responses to environmental perturbation.

Methods

Fibroblasts

We examined 200 fibroblast strains obtained from skin biopsies (upper arm of non-lesional skin) of 96 SSc cases and 104 controls. The primary cultures were maintained in Dulbecco's Modified Essential Media (DMEM) with 10% FBS and supplemented with antibiotic and antimycotic as described previously (32). The 5th passage of fibroblast strains were plated at a density of 2.5 x 10^5 cells in a 35 mm dish and grown until 80% confluence. The culture medium was replaced with DMEM without FBS before stimulation assays.

All patients fulfilled American College of Rheumatology criteria for SSc (33). Among them, there were 37 limited and 53 diffused form of SSc (six patients were not reported in skin types). Autoantibodies positivity showed twenty patients with anti-topoisomerase I, twelve with anti-centromere, seventy with anti-RNA polymerase III. For disease duration, forty-nine patients were less than 5 years, eighteen were between 5 and 10 years, seventeen were 11 to 20 years, twelve were unknown. All normal controls were individuals with no history of autoimmune diseases. The study was approved by the Committee for the Protection of Human Subjects at University of Texas Health Science Center at Houston.

Silica stimulation on fibroblasts

Culture media then were replaced with FCS-free DMEM containing different doses (1, 5, 10, 25 and 50 μg) of silica particles obtained from Sigma-Aldrich. The cultures also were grouped into different time-courses including 24-, 48-, 72-, 96-, 120-hours stimulation with 10 μg silica. After stimulation, the fibroblasts were harvested for extraction of RNA. The RNAs were examined with RT-PCR for gene expression of *COL1A2*, *COL3A1*, *CTGF*, *MMP1*, *MMP3* and *TIMP3*. We used the "AQchange" (absolute quantification change) as the phenotype variable in the genetic association analysis, defined as $log2(\frac{\text{gene expression after stimulation at time } t}{\text{gene expression without stimulation at time } t})$. AQChange = 0 means gene expression remained the same after stimulation, while AQChange = 1 means gene expression was doubled after stimulation and AQChange = -1 means gene expression dropped by half after stimulation. AQchange can be any continuous value.

Genotyping of fibroblasts from each subject

Samples were genotyped using the Immunochip, an Illumina Infinium platform designed to densely genotype 186 immune-mediated disease loci identified in prior GWAS of autoimmune

diseases (26), according to the manufacturer's recommendations. Bead intensity data was processed and normalized for each sample in Genome Studio; data of successfully genotyped samples was extracted and genotypes were called within collections using optiCall (34). We obtained 196,517 SNPs before quality control (QC) as described below. NCBI build 36 (hg18) mapping was used in factory assembly file (Illumina manifest file Immuno_BeadChip_11419691_B.bpm); we then re-mapped all the SNPs to NCBI build 37 (hg19) locations.

Measurement of the ECM gene expression

Quantitative real time RT-PCR was performed using an ABI 7900 sequence detector (Applied Biosystems) for gene expression of human *COL1A2*, *COL3A1*, *CTGF*, *MMP1*, *MMP3* and *TIMP3* using standard protocol described previously (32, 35). The specific primers and probes for each gene were purchased through Assays-on-Demand from Applied Biosystems.

Linear mixed model to detect time-course and dose-response loci

We used the linear mixed model (LMM) to test genome-wide associations between SNPs and time-course/elevated-dosage responded mRNA levels, as measured by the quantitative real time RT-PCR, for each of the six ECM genes. The LMM takes into account the correlations between repeated measurements within the same subject by introducing population-level fixed effects and subject-specific random effects (random intercept and slope), and enjoys the parsimonious modeling of correlated outcomes and the ease of accommodating missing values (36). To correct for multiple hypotheses testing, we used the Bonferroni procedure with a genome-wide significance cut-off of 5.72×10^{-8} adjusting for a total of 874,949 genotyped and 1000 Genomes-imputed SNPs with non-missing p-values from the LMM. After we performed the race-specific association analyses by the LMM, we performed fixed-effect meta-analysis as

well as random-effect meta-analysis if determined necessary, combining results from the Caucasians, Hispanics and African Africans. Details on the statistical methods are provided in the Supplemental Methods, including QC procedures, 1000 Genomes-based imputation, LMM model specification, meta-analysis, haplotype-based association analysis, functional annotation and power calculation.

Results

Initial analysis

After stringent quality control measures, 183 subjects including 85 cases and 98 controls were included in the final association analyses; see Supplemental Methods for details. Demographic information of the study subjects is shown Table 1. We first analyzed all Caucasians among the subjects (50 SSc cases and 65 controls) as a discovery cohort (Table S1). Using LMM model assuming an additive genetic model, we found several genetic loci with SNPs significantly associated with expression of the ECM genes in this cohort (Table 2). Among them, three previously reported SSc loci including multiple SNPs of *TNFAIP3* (6q23) (10, 37), *IL2RA* (10p15-p14) (15) and *ITGAM* (16p11.2) (12, 38) showed strong associations with the expression of *MMP1* and/or *MMP3* of the fibroblasts in responses to dosage-dependent and/or time-course stimulations of silica particles (Table 2). SNPs of *CASC9* (8q21.11), intergenic region between *LINC00284* and *SMIM2-AS1* (13q14.11), *FGFR1OP* (6q27) and *THADA* (2p21) also showed associations with the expression of specific ECM genes of the fibroblasts (Table 2), but these genes have not been reported as SSc-associated.

Rs79411652 in the upstream of *TNFAIP3* was associated with *MMP1* dose-response, *MMP3* dose-response and time-course response. The minor allele frequency (MAF) of rs79411652 was 3.9% and 3.3%, respectively, in the Caucasian and African American samples.

However, the effect size of this SNP was quite large as shown in Figure 4A, B and C, with 3.6 to 12.7 fold change of gene expression between heterozygous alleles and homozygous major allele. Thus, we had high enough power (over 80%) to detect this association. In addition, regional plots, as shown in Figure 3A, B and C, indicate that there was a cluster of SNPs in high LD with rs79411652 showing association signals, which strengthen the findings about the potential pleiotropic effects of *TNFAIP3* on *MMP1* and *MMP3*.

Figures 3, 4 and S3 also show regional plots and genotype-specific expression trajectories for other previously identified SSc loci. SNP rs41290329 located in the 3' untranslated region (UTR) of *IL2RA* was significantly associated with *MMP1* dose-response, *MMP3* dose-response and time-course expression profiles (Figure 3D, E and F) and with 9.5 to 39 fold change of gene expression of *MMP1*/*MMP3* between heterozygous alleles and homozygous major allele (Figure 4D, E and F). Of note, rs41290329 is a low frequency SNP with an MAF of 1.3% in Caucasians and monomorphic in African Americans and Hispanics in our samples as well as in the 1000 Genomes Project. In addition, SNP rs12926702 was significantly associated with *MMP3* dose-response expression profile (Figure S3 H) with 34 fold change of gene expression of *MMP3* between heterozygous alleles and homozygous major alleles. Rs12926702 is located at the intergenic region of *TRIM72* and *ITGAM*. It is also a low frequency SNP with a MAF of 1.4% and 1.6% in Caucasians and African Americans, respectively, and monomorphic in our Hispanics samples. Given the low MAF of the peak SNPs and the current moderate sample size, we caution that future studies with larger sample size are warranted to confirm the signals identified here and obtain more reliable effect size estimates.

To evaluate whether the association signals of *TNFAIP3* and *IL2RA* could be better explained by multiple associated SNPs in LD or single peak SNPs, we performed haplotype-

based analysis of the chromosomal regions defined by 200 SNPs upstream and downstream of the peak SNPs, rs79411652 of *TNFAIP3* and rs41290329 of *IL2RA*. We defined haplotypes based on 3, 5 or 7 consecutive SNPs in sliding windows. The peak signals of haplotype-based analyses agreed well with those of single SNP-based association analyses, although the former yielded no as strong signals as the latter (Figure S2).

In addition to the signals within or near SSc-associated genes, we also identified several genetic loci that have not been reported to be associated with SSc (Table 2), including rs7823944 of *CASC9* associated with *COL3A1* dose-response expression (Figure S3 A), rs79365263 in the intergenic region of *LINC00284* and *SMIM2-AS1* associated with MMP1 and MMP3 dose-response expression (Figure S3 D and G), rs78409037 of *THADA* and rs75701002 of *FGFR1OP* associated with MMP3 time-course expression (Figure S3 I and K). In particular, rs7823944 of *CASC9* was supported by a cluster of imputed SNPs in high LD (r^2 around 0.80) (Table 2 and Figure S3 A), while rs79365263 of *LINC00284|SMIM2-AS1* was clustered with a number of genotyped SNPs in high LD (Figure S3 D and G).

HLA class II genes are well-documented in association with SSc. The nominal association signals ($5 \times 10^{-8} < P < 5 \times 10^{-2}$) of HLA class II genes were observed in the Caucasian cohort. Other SSc-loci did not show significant association with gene expression of the fibroblasts.

As elaborated in Supplemental Methods, the standard normal distribution-based p-values for the SNP effect in the LMM can be too liberal, i.e., not controlling the Type I error at the nominal level. We therefore calculated the more accurate t-distribution-based p-values using the computationally demanding Satterthwaite approximation for the peak SNPs in Table 2. As

shown in Table S2, although all of the p-values became larger as expected, the peak SNPs in the SSc-associated loci of *TNFAIP3*, *IL2RA* and *ITGAM*, remained significant.

Meta-analysis of Caucasian, African American and Hispanic cohorts

We conducted meta-analysis of the Caucasian (50 SSc patients and 65 controls), African American (13 SSc patients and 23 controls) and Hispanics (22 SSc patients and 10 controls) cohorts via the fixed-effect inverse-variance weighting method as well as random-effect meta-analysis if determined necessary; see the Supplemental Methods for details. Of note, in the presence of potential heterogeneous genetic effects across the three populations, fixed-effect meta-analysis may not be appropriate. To investigate this, we calculated the heterogeneity measure I^2 and Cochran's Q test p-value (Table 2), which indicated that most of genome-wide significant SNPs in the Caucasian cohort, might have heterogeneous genetic effects across different populations ($I^2 > 56\%$) (39). We, therefore, also performed random-effect meta-analysis for these SNPs. Noticeably, SNP rs58905141 of *TNFAIP3* remained genome-wide significantly associated with *MMP3* dose-response expression by both fixed-effect and random-effect meta-analysis, even though I^2 and Cochran's test results suggested that the random-effect model was not necessary for this SNP. Other SNPs did not show significance with random-effect meta-analysis. The Manhattan plots for fixed-effect and random-effect meta-analysis are, respectively, shown in Figure 1 and Figure S1. Meta-analysis results for all the genome-wide significant SNPs in the discovery analysis of the Caucasian samples are shown in Table S3. The non-significant random-effect meta-analysis results for most of the peak SNPs in Caucasians may be attributed to the low MAF of the SNPs, smaller sample sizes in the African American and Hispanic cohorts and disparate LD patterns across populations, leading to diminished statistical power.

Discussion

SSc is a fibrotic disease with complex genetic traits, to which environmental factors may trigger a pathological process toward systemic fibrosis. Functional association studies of genetic polymorphisms in connection with environmental triggers are challenge. It is unknown whether the SSc susceptibility loci contribute to specific fibrotic process in the presence of an environmental hazard. In comparison with conventional gene by environment (GxE) analysis based on observational studies (40, 41), the studies herein are novel in that we examined profibrotic effects of silica particles in a complex cell model of human fibroblasts that were genetically profiled with Immuno-chip SNPs.

A previously reported SSc locus at 6q23.3 with SNPs of *TNFAIP3* (10, 37) appeared to be persistently in association with expression of *MMP1* and/or *MMP3* genes of the fibroblasts in responses to dosage-dependent and time-course stimulations of silica particles in both Caucasian only analysis and meta-analysis of diverse populations (Table 2). *TNFAIP3* stands for tumor necrosis factor-α-induced-protein 3. It encodes A20, an ubiquitin-modifying enzyme that inhibits NF-kB activity and is a key regulator of TNF-mediated immune responses and inflammation (42, 43). Inflammation has been reported as a major consequence induced by silica particles in *in vivo* and *in vitro* studies (35, 44). Our previous studies indicated that silica might activate fibroblasts to overexpress the ECM genes through proinflammatory process (32, 35). Therefore, the results suggest that an interaction between genetic factors of *TNFAIP3* and environmental factor of silica stimulation may control fibroblasts to express *MMP1*/*MMP3* genes, which may be associated with development of SSc through an inflammatory and immune-control mechanism. Of note, intronic SNP rs5029939 in *TNFAIP3* was reported to be associated with SSc in a previous case-control study (10). LD analysis by the SNAP tool (45) indicated that this SNP was in perfect LD ($r^2 = 1$ and D'=1) with our top hit SNP rs79411652 in the discovery analysis of

Caucasian samples, and in high LD with our meta-analysis top hit SNP rs58905141 (r^2 = 0.80 and D'=1) associated with *MMP3* dose-response and time-course expression as well as *MMP1* dose-response expression. Regional plots (Figure 3A, C and E) also suggest that these SNPs were likely to tag the same functional locus/loci. We annotated the top SNPs of *TNFAIP3* with RegulomeDB (46) according to the ENCODE database (47). While rs79411652 was annotated to category 7, i.e., no known biological function, rs58905141 was annotated to the second highest functional category "2b", which is likely to affect binding of transcription factors FOXA1 and SMARCA4 by ChIP-seq experiments, binding of STAT1 by motif analysis, and is in DNase footprint by DNase-seq and in DNase peak by ChIP-seq experiments. This suggests that rs58905141 may play a functional role in regulating *TNFAIP3* expression and directly or indirectly regulate *MMP1*/*MMP3* expression in fibroblasts. Further replication and functional studies of the signals in *TNFAIP3* are warranted to confirm this hypothesis.

Two other reported SSc-associated genetic loci with SNPs of *IL2RA* (10p15-p14) and *ITGAM* (16p11.2) genes were significantly associated with the expression of *MMP1* and/or *MMP3* in silica response assays in Caucasians (Table 2). *IL2RA* stands for interleukin 2 receptor α. IL2RA is involved in various pathways that help control the differentiation of effector cells, T-cell proliferation, and immune tolerance (48). A meta-analysis study with a large Caucasian European cohort revealed an association between rs2104286 of *IL2RA* and ACA positive SSc patients (15). LD analysis using the SNAP software showed that *MMP1/MMP3* expression associated rs41290329 was in complete LD with rs2104286 (D' = 1), but in low LD as measured by r^2 (=0.06), due to the disparate MAFs of the two SNPs (0.013 and 0.225 in the 1000 Genomes CEU population, respectively). Functional analysis by RegulomeDB indicated that biological function for rs41290329 was unknown. As for *ITGAM* that encodes the α subunit of the αMβ2-

integrin, it has been recently identified as an autoimmune disease risk gene (49). It is expressed on the surface of leukocytes, and it regulates adhesion of neutrophils and monocytes, cell activation, which is important for innate immunity (50). Rs1143679 of *ITGAM* was associated with SSc in European cohorts (12) and a large-scale meta-analysis (38). LD analysis suggested that *MMP3* dose-response expression associated rs12926702 was in complete LD with rs1143679 (D' = 1), but in low LD as measured by r^2 (=0.003), due to the low MAFs of both SNPs (0.033 and 0.083 in the 1000 Genomes CEU population, respectively). Considering the low MAFs of rs41290329 and rs12926702, in addition to our moderate sample size, especially for non-Caucasian samples, we caution that the findings on *IL2RA* and *ITGAM* need to be confirmed in future studies with larger sample size.

In addition, several other genetic loci and genes unrelated to SSc susceptibility were also strongly associated with expression of the ECM genes of silica-stimulated fibroblasts in Caucasian-only analysis and/or meta-analysis (Table 2). Whether these associations represent a general impact of specific genetic loci on immune-mediated diseases is worth further investigation.

It is worth noting that many of those known SSc-loci did not show significant impact on gene expression of the fibroblasts in this study. For instance, the HLA class II genes that confer the strongest susceptibility to SSc (1) showed only nominal association signals (see Figure 2), which suggest that potential genetic impact of these genes to profibrotic responses of silica-stimulated fibroblasts may be less significant in this study model. It is likely that genetic impact of specific SSc-loci may be determined by interactions of specific types of cells and environmental triggers, as well as specific responding genes.

The current study employed the cost-effective Immunochip targeted at 186 gene regions which were identified in previous association studies of autoimmune diseases. Although we successfully performed 1000 Genomes-based imputation to boost the 130,000 directly genotyped SNPs to around 800,000 total SNPs, the coverage was likely to be very low beyond the 186 gene regions. As a result, we cannot exclude the possibility that other genetic variants that were not investigated in the current study might be associated with the ECM gene expression in the fibroblasts. In addition, heterogeneity of patients' clinical status is another concern. This warrants further investigation in the future.

In summary, this is the first attempt to study interactions between genome-wide genetic variants and environmental factors of SSc in a complex fibroblast model. The results indicated that a previously identified SSc locus of *TNFAIP3* was strongly and persistently associated with silica-induced profibrotic responses of the fibroblasts in both Caucasians and meta-analysis of mixed populations. Similar associations were observed in two other reported SSc loci of *IL2RA* and *ITGAM* in Caucasians. In addition, some non-SSc loci may also impact a general response of fibroblasts to silica particles. This notion may be further verified in other cohorts with larger sample size and in functional studies beyond the ECM genes.

Acknowledgments

This work was supported by the grants from the Department of the Army, Medical Research Acquisition Activity [PR064803 to XD.Z]; Scleroderma Foundation [2012 to XD.Z]; and the National Institutes of Health [NIAID UO1, 1U01AI09090 to XD.Z]. P.W. was partially supported by NIH grants R01CA169122 and R01HL116720. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing

HPC resources that have contributed to the research results reported within this paper. The authors indicated that there is no conflict of interest.

References

1. Arnett FC, Gourh P, Shete S, Ahn CW, Honey RE, Agarwal SK, et al. Major histocompatibility complex (MHC) class II alleles, haplotypes and epitopes which confer susceptibility or protection in systemic sclerosis: analyses in 1300 Caucasian, African-American and Hispanic cases and 1000 controls. Ann Rheum Dis. 2010;69(5):822-7.
2. Rueda B, Broen J, Simeon C, Hesselstrand R, Diaz B, Suarez H, et al. The STAT4 gene influences the genetic predisposition to systemic sclerosis phenotype. Hum Mol Genet. 2009;18(11):2071-7.
3. Dieude P, Guedj M, Wipff J, Avouac J, Fajardy I, Diot E, et al. Association between the IRF5 rs2004640 functional polymorphism and systemic sclerosis: a new perspective for pulmonary fibrosis. Arthritis Rheum. 2009;60(1):225-33.
4. Radstake TR, Gorlova O, Rueda B, Martin JE, Alizadeh BZ, Palomino-Morales R, et al. Genome-wide association study of systemic sclerosis identifies CD247 as a new susceptibility locus. Nature genetics. 2010;42(5):426-9.
5. Gourh P, Agarwal SK, Divecha D, Assassi S, Paz G, Arora-Singh RK, et al. Polymorphisms in TBX21 and STAT4 increase the risk of systemic sclerosis: evidence of possible gene-gene interaction and alterations in Th1/Th2 cytokines. Arthritis Rheum. 2009;60(12):3794-806.
6. Dieude P, Wipff J, Guedj M, Ruiz B, Melchers I, Hachulla E, et al. BANK1 is a genetic risk factor for diffuse cutaneous systemic sclerosis and has additive effects with IRF5 and STAT4. Arthritis Rheum. 2009;60(11):3447-54.
7. Gourh P, Agarwal SK, Martin E, Divecha D, Rueda B, Bunting H, et al. Association of the C8orf13-BLK region with systemic sclerosis in North-American and European populations. Journal of autoimmunity. 2010;34(2):155-62.
8. Gourh P, Arnett FC, Tan FK, Assassi S, Divecha D, Paz G, et al. Association of TNFSF4 (OX40L) polymorphisms with susceptibility to systemic sclerosis. Ann Rheum Dis. 2010;69(3):550-5.
9. Broen J, Gourh P, Rueda B, Coenen M, Mayes M, Martin J, et al. The FAS -670A>G polymorphism influences susceptibility to systemic sclerosis phenotypes. Arthritis Rheum. 2009;60(12):3815-20.
10. Dieude P, Guedj M, Wipff J, Ruiz B, Riemekasten G, Matucci-Cerinic M, et al. Association of the TNFAIP3 rs5029939 variant with systemic sclerosis in the European Caucasian population. Ann Rheum Dis. 2010;69(11):1958-64.
11. Dieude P, Bouaziz M, Guedj M, Riemekasten G, Airo P, Muller M, et al. Evidence of the contribution of the X chromosome to systemic sclerosis susceptibility: association with the functional IRAK1 196Phe/532Ser haplotype. Arthritis Rheum. 2011;63(12):3979-87.
12. Carmona FD, Simeon CP, Beretta L, Carreira P, Vonk MC, Rios-Fernandez R, et al. Association of a non-synonymous functional variant of the ITGAM gene with systemic sclerosis. Annals of the rheumatic diseases. 2011;70(11):2050-2.

13. Terao C, Ohmura K, Kawaguchi Y, Nishimoto T, Kawasaki A, Takehara K, et al. PLD4 as a novel susceptibility gene for systemic sclerosis in a Japanese population. Arthritis Rheum. 2013;65(2):472-80.
14. Manetti M, Allanore Y, Saad M, Fatini C, Cohignac V, Guiducci S, et al. Evidence for caveolin-1 as a new susceptibility gene regulating tissue fibrosis in systemic sclerosis. Ann Rheum Dis. 2012;71(6):1034-41.
15. Martin JE, Carmona FD, Broen JC, Simeon CP, Vonk MC, Carreira P, et al. The autoimmune disease-associated IL2RA locus is involved in the clinical manifestations of systemic sclerosis. Genes Immun. 2012;13(2):191-6.
16. Dieude P, Guedj M, Wipff J, Ruiz B, Riemekasten G, Airo P, et al. NLRP1 influences the systemic sclerosis phenotype: a new clue for the contribution of innate immunity in systemic sclerosis-related fibrosing alveolitis pathogenesis. Ann Rheum Dis. 2011;70(4):668-74.
17. Barizzone N, Marchini M, Cappiello F, Chiocchetti A, Orilieri E, Ferrante D, et al. Association of osteopontin regulatory polymorphisms with systemic sclerosis. Human immunology. 2011;72(10):930-4.
18. Martin JE, Assassi S, Diaz-Gallo LM, Broen JC, Simeon CP, Castellvi I, et al. A systemic sclerosis and systemic lupus erythematosus pan-meta-GWAS reveals new shared susceptibility loci. Hum Mol Genet. 2013;22(19):4021-9.
19. Diaz-Gallo LM, Simeon CP, Broen JC, Ortego-Centeno N, Beretta L, Vonk MC, et al. Implication of IL-2/IL-21 region in systemic sclerosis genetic susceptibility. Ann Rheum Dis. 2013;72(7):1233-8.
20. Salim PH, Jobim M, Bredemeier M, Chies JA, Brenol JC, Jobim LF, et al. Combined effects of CXCL8 and CXCR2 gene polymorphisms on susceptibility to systemic sclerosis. Cytokine. 2012;60(2):473-7.
21. Martin JE, Broen JC, Carmona FD, Teruel M, Simeon CP, Vonk MC, et al. Identification of CSK as a systemic sclerosis genetic risk factor through Genome Wide Association Study follow-up. Hum Mol Genet. 2012;21(12):2825-35.
22. Palomino GM, Bassi CL, Wastowski IJ, Xavier DJ, Lucisano-Valim YM, Crispim JC, et al. Patients with systemic sclerosis present increased DNA damage differentially associated with DNA repair gene polymorphisms. The Journal of rheumatology. 2014;41(3):458-65.
23. Koumakis E, Bouaziz M, Dieude P, Ruiz B, Riemekasten G, Airo P, et al. A regulatory variant in CCR6 is associated with susceptibility to antitopoisomerase-positive systemic sclerosis. Arthritis Rheum. 2013;65(12):3202-8.
24. Allanore Y, Saad M, Dieude P, Avouac J, Distler JH, Amouyel P, et al. Genome-wide scan identifies TNIP1, PSORS1C1, and RHOB as novel risk loci for systemic sclerosis. PLoS genetics. 2011;7(7):e1002091.
25. Gorlova O, Martin JE, Rueda B, Koeleman BP, Ying J, Teruel M, et al. Identification of novel genetic markers associated with clinical phenotypes of systemic sclerosis through a genome-wide association strategy. PLoS genetics. 2011;7(7):e1002178.
26. Mayes MD, Bossini-Castillo L, Gorlova O, Martin JE, Zhou X, Chen WV, et al. Immunochip analysis identifies multiple susceptibility loci for systemic sclerosis. American journal of human genetics. 2014;94(1):47-61.
27. Jin J, Chou YC, Lima M, Zhou Z, Zhou X. Systemic sclerosis is a complex genetic disease associated mainly with immune regulatory and inflammatory genes. Manuscript under review. 2014.

28. Slimani S, Ben Ammar A, Ladjouze-Rezig A. Connective tissue diseases after heavy exposure to silica: a report of nine cases in stonemasons. Clin Rheumatol. 2010;29(5):531-3.
29. Cowie RL. Silica-dust-exposed mine workers with scleroderma (systemic sclerosis). Chest. 1987;92(2):260-2.
30. Trojanowska M. Molecular aspects of scleroderma. Frontiers in bioscience : a journal and virtual library. 2002;7:d608-18.
31. Kuroda K, Shinkai H. Gene expression of types I and III collagen, decorin, matrix metalloproteinases and tissue inhibitors of metalloproteinases in skin fibroblasts from patients with systemic sclerosis. Archives of dermatological research. 1997;289(10):567-72.
32. Xiong M, Arnett FC, Guo X, Xiong H, Zhou X. Differential dynamic properties of scleroderma fibroblasts in response to perturbation of environmental stimuli. PLoS One. 2008;3(2):e1693.
33. Subcommittee for scleroderma criteria of the American Rheumatism Association Diagnostic and Therapeutic Criteria Committee. Preliminary criteria for the classification of systemic sclerosis (scleroderma). Subcommittee for scleroderma criteria of the American Rheumatism Association Diagnostic and Therapeutic Criteria Committee. Arthritis Rheum. 1980;23(5):581-90.
34. Shah TS, Liu JZ, Floyd JA, Morris JA, Wirth N, Barrett JC, et al. optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants. Bioinformatics. 2012;28(12):1598-603.
35. Guo X, Jagannath C, Espitia MG, Zhou X. Uptake of silica and carbon nanotubes by human macrophages/monocytes induces activation of fibroblasts in vitro -- potential implication for pathogenesis of inflammation and fibrotic diseases. Int J Immunopathol Pharmacol. 2012;25(3):713-9.
36. Fitzmaurice G, Laird N, Ware J. Applied Longitudinal Analysis. 2nd ed. Hoboken, New Jersey: John Wiley & Sons; 2011.
37. Koumakis E, Giraud M, Dieude P, Cohignac V, Cuomo G, Airo P, et al. Brief report: candidate gene study in systemic sclerosis identifies a rare and functional variant of the TNFAIP3 locus as a risk factor for polyautoimmunity. Arthritis Rheum. 2012;64(8):2746-52.
38. Anaya JM, Kim-Howard X, Prahalad S, Chernavsky A, Canas C, Rojas-Villarraga A, et al. Evaluation of genetic association between an ITGAM non-synonymous SNP (rs1143679) and multiple autoimmune diseases. Autoimmun Rev. 2012;11(4):276-80.
39. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21(11):1539-58.
40. Thomas D. Gene--environment-wide association studies: emerging approaches. Nat Rev Genet. 2010;11(4):259-72.
41. Hsu L, Jiao S, Dai JY, Hutter C, Peters U, Kooperberg C. Powerful cocktail methods for detecting genome-wide gene-environment interaction. Genet Epidemiol. 2012;36(3):183-94.
42. Ma A, Malynn BA. A20: linking a complex regulator of ubiquitylation to immunity and human disease. Nat Rev Immunol. 2012;12(11):774-85.
43. Shembade N, Harhaj NS, Parvatiyar K, Copeland NG, Jenkins NA, Matesic LE, et al. The E3 ligase Itch negatively regulates inflammatory signaling pathways by controlling the function of the ubiquitin-editing enzyme A20. Nat Immunol. 2008;9(3):254-62.
44. Park EJ, Park K. Oxidative stress and pro-inflammatory responses induced by silica nanoparticles in vivo and in vitro. Toxicol Lett. 2009;184(1):18-25.

45. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics. 2008;24(24):2938-9.
46. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012;22(9):1790-7.
47. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A. 2014;111(17):6131-8.
48. Shevach EM. Certified professionals: CD4(+)CD25(+) suppressor T cells. J Exp Med. 2001;193(11):F41-6.
49. Nath SK, Han S, Kim-Howard X, Kelly JA, Viswanathan P, Gilkeson GS, et al. A nonsynonymous functional variant in integrin-alpha(M) (encoded by ITGAM) is associated with systemic lupus erythematosus. Nature genetics. 2008;40(2):152-4.
50. Solovjov DA, Pluskota E, Plow EF. Distinct roles for the alpha and beta subunits in the functions of integrin alphaMbeta2. J Biol Chem. 2005;280(2):1336-45.

Figure Legends

Figure 1. Manhattan plots of fixed-effect meta-analysis association p-values for dose-response and time-course ECM gene expressions in fibroblasts in response to silica stimulation. A. *MMP1* gene expression in silica dose response showed significant signals at 6q23.3 (*TNFAIP3*), 10p15.1 (*IL2RA*) and 13q14.11 (*LINC00284*). B. *MMP3* gene expression in silica dose response showed significant signals at 6q23.3 (*TNFAIP3*), 10p15.1 (*IL2RA*), 13q14.11 (*LINC00284*) and 16p11.2 (*TRIM72|ITGAM*). C. *MMP3* gene expression in silica time course assays showed significant signals at 2p21(*THADA*), 6q23.3 (*TNFAIP3*) and 6q27(*FGFR10P|CCR6*). Red line corresponds to the p-value cutoff for genome-wide significance (5×10^{-8}); blue line corresponds to the suggestive significance cutoff (1×10^{-5}). P-values from Caucasian-only analysis are displayed if fixed-effect meta-analysis was not performed due to missing association results in African Americans and Hispanics.

Figure 2. Chromosome 6 Manhattan plots of fixed-effect meta-analysis association p-values for dose-response and time-course ECM gene expressions in fibroblasts in response to silica stimulation. Panels A, B and C correspond to Figure 1 A, B and C with a zoom in for chromosome 6 only. Green dots show loci on or near known SSc risk genes. Red line

corresponds to the p-value cutoff for genome-wide significance (5×10^{-8}); blue line corresponds to the suggestive significance cutoff (1×10^{-5}). P-values from Caucasian-only analysis are displayed if fixed-effect meta-analysis was not performed due to missing association results in African Americans and Hispanics.

Figure 3. Regional plots of association results between previously identified SSc risk loci and gene expression of fibroblasts in the Caucasian cohort. 500kb upstream and downstream of the peak SNPs are plotted. Panels A to C: genetic locus at 6q23.3 containing SNPs of *TNFAIP3* in association with expression of (A) *MMP1* in dose-response assays, (B) *MMP3* in dose-response assays, and (C) *MMP3* in time-course assay. Panels D to F: genetic locus at 10p15.1 containing SNPs of *IL2RA* in association with expression of (D) *MMP1* in dose-response assays, (E) *MMP3* in dose-response assays, and (F) *MMP3* in time-course assays. Round dots represent genotyped SNPs while square dots represent imputed SNPs. Purple dots correspond to the peak SNPs. Linkage disequilibrium (r^2) with the peak SNP (rs79411652 of *TNFAIP3* and rs41290329 of *IL2RA*, respectively) is shown by different colors.

Figure 4. Dose-response and time-course expression trajectories of the ECM genes by genotypes of the peak SNPs in *TNFAIP3* and *IL2RA* in the Caucasian cohort. Panels A-F correspond to panels A-F in Figure 3. Red solid line: sample average trajectory for the heterozygous alleles; blue solid line: sample average trajectory for the homozygous major alleles; gray solid lines: individual trajectories; dashed horizontal grey line at 0 of the y-axis: the reference line representing constant gene expression across time/doses. Y-axis corresponds to the absolute quantity change (AQchange); x-axis corresponds to silica dosage (in μg) or time course (in days).

Table 1. Distribution of demographics among SSc cases and controls

Variable	SSc (n = 85) n (%)	Controls (n = 98) n (%)	P-value (χ^2)
Age group			
≤30	7 (8.24)	25 (25.51)	
31−40	14 (16.5)	23 (23.47)	
41−50	24 (28.24)	22 (22.45)	
>50	40 (47.06)	28 (28.57)	0.0034
Race[a]			
Caucasians	50 (58.82)	65 (66.33)	
Hispanics	22 (25.88)	10 (10.20)	
African Americans	13 (15.29)	23 (23.47)	0.015
Sex			
Female	67 (78.82)	60 (61.22)	
Male	18 (21.18)	38 (38.78)	0.016

[a]Self-reported race and corrected by STRUCTURE analysis

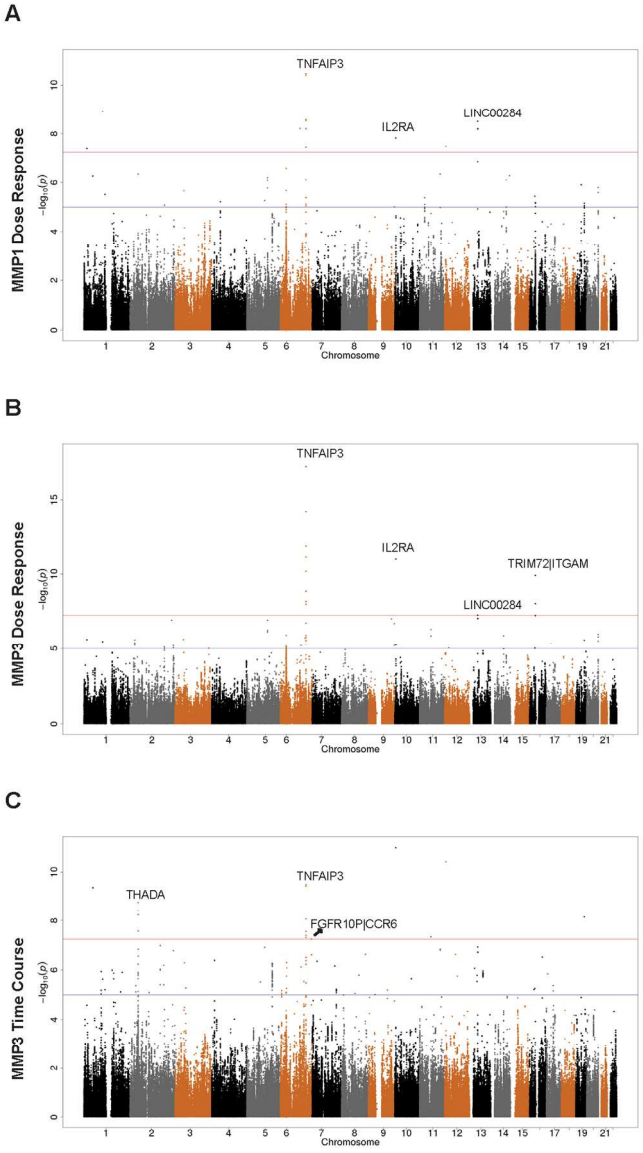
Table 2. Associations between expression of the ECM genes and genetic backgrounds in Caucasians and mixed populations by meta-analysis.

Peak SNP	Chr	Position[a]	Type	Gene(s)[b]	RA	P[f]	P-FE	P-RE	HetISq (%)	HetP	Direction[c]	Estimates(W|B|H)[e]	StdErr(W|AA|H)[e]	MAF(W|AA|H)[e]	responseType
rs58417815*	8	76159689	ncRNA_intronic	*CASC9*	C	5.10E-08	-[g]	-	-	-	+??	1.77|NA|NA	0.29|NA|NA	0.074|NA|NA	COL3A1_dose
rs7823944	8	76182788	ncRNA_intronic	*CASC9*	T	3.47E-08	2.00E-07	7.10E-02	51.20	0.13	+-+	1.54|-0.03|1.00	0.25|0.73|1.05	0.052|0.033|0.021	COL3A1_dose
rs79411652	6	138138945	intergenic	*OLIG3*|***TNFAIP3***[d]	C	3.87E-11	-	5.69E-01	88.80	0.00	+-?	2.29|-0.80|NA	0.32|0.97|NA	0.039|0.033|NA	MMP1_dose
rs58905141	6	138132123	intergenic	*OLIG3*|***TNFAIP3***	G	3.97E-11	3.46E-11	0.402918	43.1	0.17	++-	2.29|62.41|-2.18	0.32|32.09|65.44	0.039|0.014|0.016	MMP1_dose
rs41290329	10	6054083	UTR3	***IL2RA***	C	1.51E-08	-	-	-	-	+??	3.43|NA|NA	0.57|NA|NA	0.013|NA|NA	MMP1_dose
rs79365263	13	44618508	intergenic	*LINC00284*|*SMIM2-AS1*	C	3.13E-09	-	-	-	-	+??	2.22|NA|NA	0.35|NA|NA	0.024|NA|NA	MMP1_dose
rs79411652	6	138138945	intergenic	*OLIG3*|***TNFAIP3***	C	5.55E-18	-	5.12E-01	93.30	0.00	+-?	3.69|-0.91|NA	0.41|0.99|NA	0.039|0.033|NA	MMP3_dose
rs58905141	6	138132123	intergenic	*OLIG3*|***TNFAIP3***	G	6.43E-18	6.00E-18	1.32E-17	0	0.81	+++	3.69|28.75|22.87	0.41|34.61|128.32	0.039|0.014|0.016	MMP3_dose
rs41290329	10	6054083	UTR3	***IL2RA***	C	1.02E-11	-	-	-	-	+??	5.29|NA|NA	0.74|NA|NA	0.013|NA|NA	MMP3_dose
rs79365263	13	44618508	intergenic	*LINC00284*|*SMIM2-AS1*	C	5.33E-08	-	-	-	-	+??	2.80|NA|NA	0.49|NA|NA	0.024|NA|NA	MMP3_dose
rs12926702	16	31240971	intergenic	*TRIM72*|***ITGAM***	C	1.33E-10	-	3.42E-01	87.90	0.00	+-?	5.09|-0.35|NA	0.76|1.52|NA	0.014|0.016|NA	MMP3_dose
rs78409037	2	43740411	intronic	*THADA*	C	4.51E-05	-	9.14E-02	94.10	0.00	+?+	0.97|NA|3.63	0.23|NA|0.55	0.069|NA|0.041	MMP3_timecourse
rs77533229	2	43479638	intronic	*THADA***	G	2.68E-08	-	-	-	-	+??	3.35|NA|NA	0.60|NA|NA	0.010|NA|NA	MMP3_timecourse
rs79411652	6	138138945	intergenic	*OLIG3*|***TNFAIP3***	C	3.52E-10	-	7.42E-01	93.80	0.00	+-?	1.86|-1.02|NA	0.29|0.64|NA	0.039|0.033|NA	MMP3_timecourse
rs58905141	6	138132123	intergenic	*OLIG3*|***TNFAIP3***	G	3.73E-10	3.23E-10	0.446757	29.1	0.24	++-	1.85|38.02|-11.59	0.29|21.66|53.94	0.039|0.014|0.016	MMP3_timecourse
rs75701002	6	167443918	intronic	*FGFR1OP*	C	2.18E-08	5.68E-08	5.17E-02	47.50	0.17	+?+	2.39|NA|0.32	0.42|NA|1.32	0.026|NA|0.021	MMP3_timecourse
rs41290329	10	6054083	UTR3	***IL2RA***	C	1.03E-10	-	-	-	-	+??	3.25|NA|NA	0.47|NA|NA	0.013|NA|NA	MMP3_timecourse

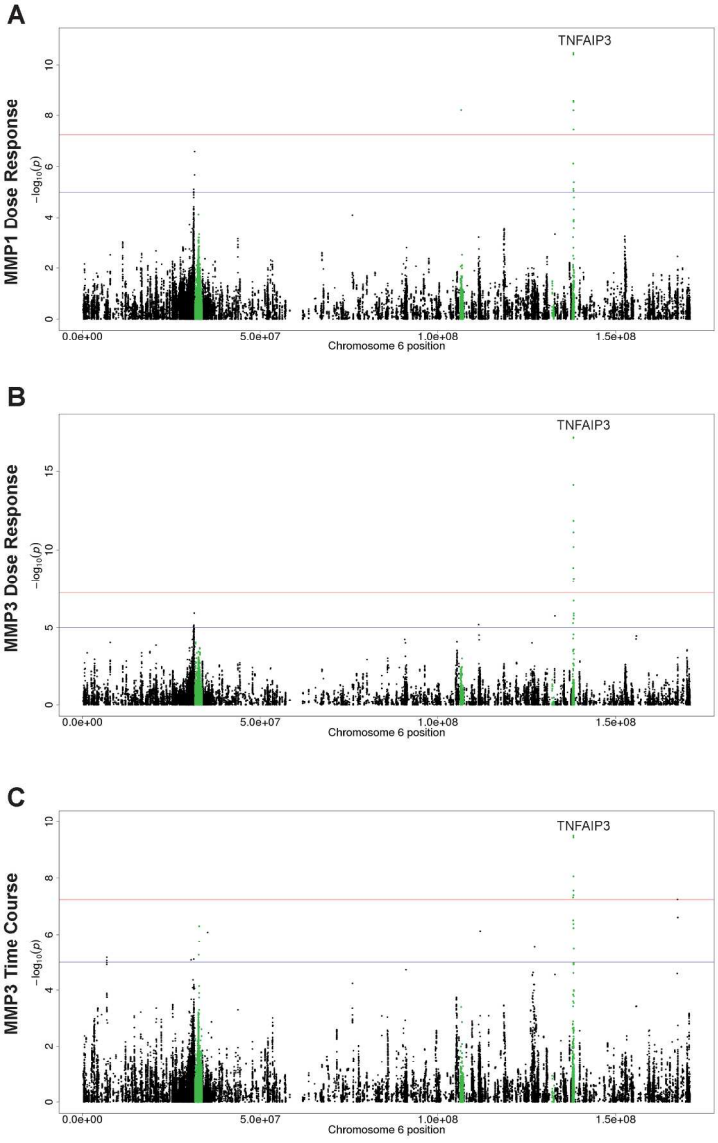
RA: risk allele; P: P-value in Caucasians; P-FE: P-value by fixed-effect meta-analysis; P-RE: P-value in random-effect meta-analysis; HetISq: heterogeneity measure I^2; HetP: Cochran's heterogeneity test p-valie; Estimates/StdErr/MAF: per risk allele effect estimates/standard error/minor allele frequency in different ethnic groups; W|AA|H: Caucasians|African Americans|Hispanics; responseType: the ECM gene expression in response to silica stimulation in elevated-dosage or time-course.

[a]GRCh37/hg19 assembly; [b]Gene in which the peak signal is located in, or genes flanking peak signal if intergenic; [c]Direction shown in W|AA|H order, where "?" represents not applicable for that specific ethnic group; [d]Bold-face genes are known SSc risk loci; [e]"NA": not applicable for that specific ethnic group; [f]P-value based on standard normal approximation; [g]"-": meta-analysis was not performed when only one ethnicity was present.

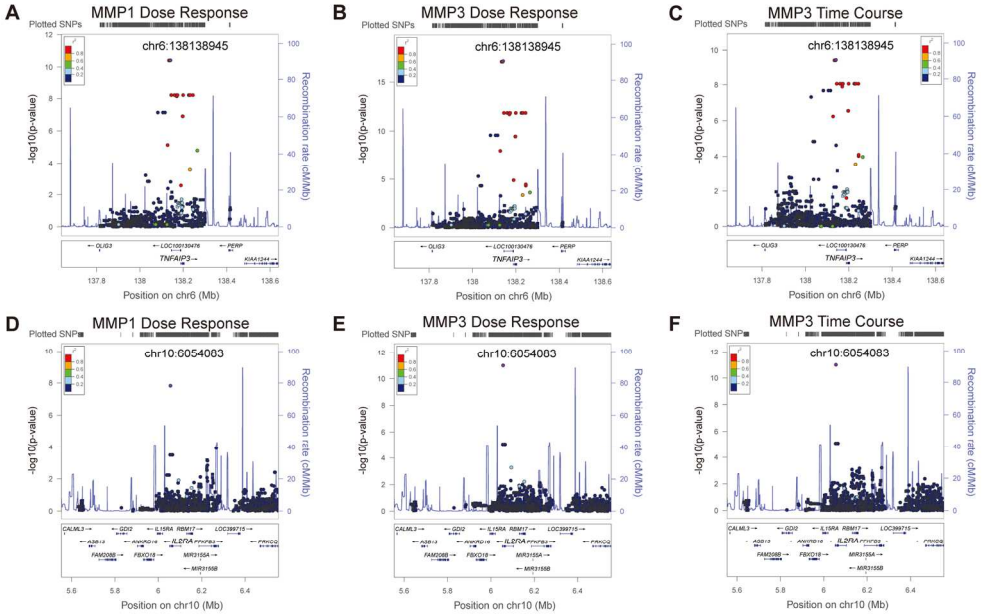
*Imputed, otherwise Peak SNP was genotyped; **The peak SNP in Caucasians was different from that in fixed-effect meta-analysis with both SNPs listed.



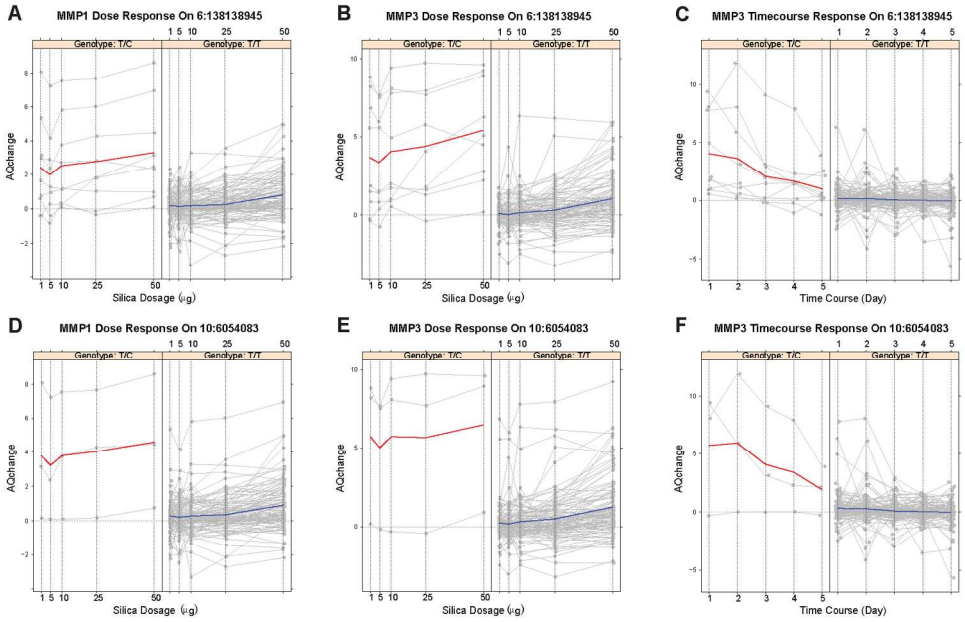
123x197mm (300 x 300 DPI)



185x275mm (300 x 300 DPI)



166x109mm (300 x 300 DPI)



165x108mm (300 x 300 DPI)

Supplemental Methods

Quality control (QC)

We used PLINK v1.07 (1) to perform QC and data manipulation on the 196,517 SNP genotype data. We first removed 18,516 monomorphic SNPs. We then used three standard QC filter criteria for each ethnic group we have (Caucasians, African Americans and Hispanics), including (a) minor allele frequency (MAF) < 0.01, (b) Hardy-Weinberg exact P-value < 1×10^{-5}, and (c) SNPs with genotyping missing rate > 5%. After removing all those SNPs that didn't pass at least one of the filters, we had 134,072, 135,939 and 132,870 SNPs for each of the three ethnic groups. We also performed sample-level QC by examining subjects with genotype missing rate > 10% or sex mismatch. Seventeen out of 200 samples were removed. Since self-reported ethnicity may not be accurate enough for genetic association studies (2), we used the STRUCTURE software (3) to infer and refine each subject's ethnicity. We first identified the common SNPs surviving the QC steps among the three ethnic groups. We then applied linkage disequilibrium (LD)-based pruning to this set of common SNPs using PLINK. Using the self-reported ethnicity as prior and the pruned SNP set as input, we classified the 183 samples into their best posterior ethnic groups. As expected, there were several race switches between Caucasians and Hispanics/African Americans, while majority of the 183 samples held consistent results with the self-reported ethnicity (Table 1). After the race group reconsolidation, we repeated the three standard QC filters stratified by races as before. After removing all those SNPs that failed any one of the filters, we had 131,994, 133,040 and 124,294 SNPs for Caucasians, African Americans and Hispanics, respectively.

Phasing and 1000 Genomes-based imputation

We used MACH 1.0 (4) to pre-phase the 183 individuals stratified by the three ethnic groups. We then used the minimac software (downloaded on July 17th, 2013) (5), with reference haplotypes from the 1000 Genomes Phase I Integrated Release Version 3 (combined 1,092 samples), to impute genome-wide SNPs un-typed on the Immunochip stratified by ethnic groups. As post-imputation QC criteria, we removed imputed SNPs with quality score $R^2 < 0.5$ or MAF < 0.05. After imputation and post-imputation QC, we were able to increase the number of SNPs from around 130,000 to 882,916, 846,862 and 784,903 for Caucasians, African Americans and Hispanics, respectively. We extracted the genotype dosage and phase information for the three ethnic groups for the following SNP- and haplotype-based association analyses.

Linear mixed model to detect time-course and dose-response loci

We used the linear mixed model (LMM) to test genome-wide associations between SNPs and time-course/elevated-dosage responded mRNA levels, as measured by the quantitative real time RT-PCR, for each of the six ECM genes. The LMM takes into account the correlations between repeated measurements within the same subject by introducing population-level fixed effects and subject-specific random effects (random intercept and slope), and enjoys the parsimonious modeling of correlated outcomes and the ease of accommodating missing values (6). We specify the LMM for time-course gene expression as follows: $y_{ij} = \beta_0 + X_i\beta + b_{0i} + b_{1i}t_{ij} + \varepsilon_{ij}$, where y_{ij} is a given gene's expression level in subject i at the *j*th measurement (i = 1,…, n and j = 1,…, 5). Let $X_i = (A_i, Z_i)$ denote the *i*th subject's covariates, where A_i contains all the subject level covariates and Z_i contains all of the observation level covariates for the fixed effects

β, including continuous time t_{ij}, continuous t_{ij}^2 if tested necessary, disease status (binary), the first two principle components (PCs) accounting for population stratification, as well as the genotype of a SNP, coded as 0 (homozygous major alleles), 1 (heterozygotes), or 2 (homozygous minor alleles) assuming an additive genetic model for a genotyped SNP. For imputed SNPs, the allele dosage is a continuous value between 0 and 2. b_{0i} and b_{1i} are the random intercept and slope of time effect for subject i, jointly following a bivariate normal distribution, while ε_{ij} is a random error following a normal distribution. To investigate potential "disease status x time" and "disease status x time^2" interaction effects, we compared the LMM with the linear time interaction term or both linear and quadratic interaction terms as fixed effects with the model without interaction terms (null model) based on the Akaike Information Criterion (AIC). Except for *TIMP3* in Caucasians, the other five genes were better fit under the LMM without the interaction terms. For *TIMP3*, the "disease status x time" interaction term was included in the final LMM as a fixed effect. For the dose-response expression data, we used the same LMM except that t_{ij} represented the *j*th dosage. Model selection analysis suggested that all the dose-response data were better fit with the main effects model. Stratified by the three ethnic groups, we performed LMM analyses for the six ECM genes of two response types (time-course and dose-response) for each of the SNPs we obtained from the 1000 Genomes-based imputation. We conducted the LMM analyses using the R package "lme4" (http://cran.r-project.org/web/packages/lme4/)(7). To correct for multiple hypotheses testing, we used the Bonferroni procedure with a genome-wide significance cut-off of 5.72×10^{-8} adjusting for 874,949 SNPs with non-missing p-values from the LMM. We adopted 1×10^{-5} as the suggestive significance cut-off by convention. We obtained both

the Wald test and likelihood ratio test (LRT) statistics for the SNP effect in the LMM and found that they agreed with each other well as expected (Figure S4). To facilitate the meta-analysis combining race-specific analysis results, we employed the Wald test throughout our analyses since the current major meta-analysis software, such as METAL (8) and GWAMA (9), use the Wald test exclusively.

Meta-Analysis

After we performed the race-specific association analyses by the LMM, we used METAL (8) and GWAMA v.2.1 (9) to perform fixed-effect meta-analysis as well as random-effect meta-analysis if determined necessary, combining results from the Caucasians, Hispanics and African Africans. For the fixed-effect meta-analysis, race-specific beta coefficients and standard errors were combined based on inverse variance weighting in METAL and GWAMA. We also calculated the Cochran's Q test p-value and I^2 statistic for each SNP to evaluate possible heterogeneity between ethnic groups. It has been suggested that $I^2 > 56\%$ indicates severe heterogeneity and thus random-effect meta-analysis is more appropriate for such SNPs (10).

Genomic control and principle component analysis

As the Immunochip is a platform with SNPs densely covering 186 known autoimmune disease risk loci, rather than a genome-wide array, global deviation of the association p-values from the null distribution was reported in previous studies due to shared susceptibility factors among autoimmune diseases and the polygenic effects (2, 11). We observed similar inflation of p-values in the quantile-quantile (QQ) plots (Figure S5). To exclude the possibility of undue confounding of population stratification, we included the top two PCs, obtained from the EIGENSOFT 4.2 software (12, 13) based on

high-quality SNPs after QC procedures, in the race-stratified LMM (Figure S6). To further adjust for potential population stratification, we used the genomic control (GC) approach (14) to correct the global inflation of p-values after removing known SSc risk loci in both race-stratified analysis and meta-analysis. We also performed the traditional case-control analysis of SSc using the logistic regression on the 882,881 genome-wide imputed and genotyped SNPs in Caucasians (50 SSc cases versus 65 controls). Although, as expected, we did not have adequate statistical power to validate previously reported SSc risk loci (the minimum p-value was 8.42×10^{-6}), we found that the QQ-plot was well behaved (See Figures S7 and S8), suggesting that population stratification was at most moderate in our samples and the correction by PCs and the GC method should be effective. Finally, we noted that the commonly used standard normal approximation to the Wald test statistics null distribution in the LMM could also lead to liberal p-values (see page 100 of (6)). Although the exact null distribution of the Wald test statistics is a t-distribution with finite degrees of freedom, the standard normal distribution, based on which the p-values can be calculated much faster, approximates the t-distribution well when the sample size is very large. However, as the sample size here was only moderate, the standard normal p-values, which are default in the R function for the LMM, might be misleadingly too small. The Satterthwaite approximation, a computationally demanding procedure, has been proposed in the literature to calculate the exact t-distribution-based p-values (15). We computed these more accurate p-values for those genome-wide significant SNPs by the standard normal-based p-values.

Haplotype-based Analysis

To evaluate whether the peak association signals could be better explained by multiple associated SNPs in LD or single peak SNPs, we performed haplotype-based analysis of the chromosomal regions defined by 200 SNPs upstream and downstream of the peak SNPs. We extracted the haplotypes of each individual in the imputation step to define short haplotypes of 3, 5 or 7 consecutive SNPs in sliding windows with a step size of one SNP, and did LMM-based association tests with the ECM gene expression as follows. To avoid numerical instability and loss of power, we combined rare haplotypes (frequency < 1%) with their most similar haplotypes based on the hierarchical clustering method of (16). Given K unique haplotypes in a sliding window, we included K regression coefficients with the design matrix coded according to (17) in the LMM replacing the SNP genotype. We then performed a K degrees of freedom Wald test for the overall haplotype effect.

Functional annotation

We used ANNOVAR (18), SNPnexus (www.snp-nexus.org) (19, 20), Genetic Association Database (http://geneticassociationdb.nih.gov/) (21), GeneCards (www.genecards.org) (22) and RegulomeDB based on the ENCODE project (23) to perform functional annotations on the association results. We used the software LocusZoom (http://csg.sph.umich.edu/locuszoom/) (24) to create regional plots for zoomed in examination of signal peaks in the Caucasians. Pairwise LD and recombination rate shown in the regional plots were based on hg19/1000 Genomes Mar 2012 EUR reference population.

Power calculation

We used the method of (25) to calculate the power for the LMM with 5 repeated measures (5 time points after simulation or 5 doses), as implemented in the "longpower" package (http://cran.r-project.org/web/packages/longpower/index.html) in the statistical software R (26). Power calculation was based on (i) the within-individual correlation structure and random error variance estimated from the fitted model without SNP covariate for the time-course *MMP3* expression in the 115 Caucasian samples, (ii) effect size as fold change of gene expression ($2^{AQchange}$) per minor allele averaged across time/dosage, and (iii) significance level $\alpha = 5x10^{-8}$. Because of the use of repeated measures (5 time points after simulation/5 doses), we had adequate statistical power to detect moderate to strong genetic effect size in terms of fold change of gene expression ($2^{AQchange}$) per minor allele averaged across time/dosage. Assuming homogenous genetic effects, at a significance level $\alpha = 5 \times 10^{-8}$, we had over 80% power to detect a fold change of at least 1.3, 1.4, 1.6, 2.6 and 6.5 with 183 individuals for a SNP with MAF > 40%, 30%, 20%, 10% and 5%, respectively.

References

1. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559-75.
2. Mayes MD, Bossini-Castillo L, Gorlova O, Martin JE, Zhou X, Chen WV, et al. Immunochip analysis identifies multiple susceptibility loci for systemic sclerosis. American journal of human genetics. 2014;94(1):47-61.
3. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2):945-59.
4. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34(8):816-34.
5. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature genetics. 2012;44(8):955-9.
6. Fitzmaurice G, Laird N, Ware J. Applied Longitudinal Analysis. 2nd ed. Hoboken, New Jersey: John Wiley & Sons; 2011.
7. Bates D. Linear mixed model implementation in lme4. 2012 [cited; http://cran.r-project.org/web/packages/lme4/vignettes/Implementation.pdf]. Available from: http://cran.r-project.org/web/packages/lme4/vignettes/Implementation.pdf

8. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190-1.
9. Magi R, Morris AP. GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics. 2010;11:288.
10. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21(11):1539-58.
11. Juran BD, Hirschfield GM, Invernizzi P, Atkinson EJ, Li Y, Xie G, et al. Immunochip analyses identify a novel risk locus for primary biliary cirrhosis at 13q14, multiple independent associations at four established risk loci and epistasis between 1p31 and 7q32 risk variants. Hum Mol Genet. 2012;21(23):5209-21.
12. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2(12):e190.
13. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006;38(8):904-9.
14. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999;55(4):997–1004.
15. Pan W, Wall MM. Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. Stat Med. 2002;21(10):1429-41.
16. Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, Morris AP. Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. American journal of human genetics. 2004;75(1):35-43.
17. Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered. 2002;53(2):79-91.
18. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research. 2010;38(16):e164.
19. Chelala C, Khan A, Lemoine NR. SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. Bioinformatics. 2009;25(5):655-61.
20. Dayem Ullah AZ, Lemoine NR, Chelala C. SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). Nucleic acids research. 2012;40(Web Server issue):W65-70.
21. Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. Nature genetics. 2004;36(5):431-2.
22. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about genes, proteins and diseases. Trends Genet. 1997;13(4):163.
23. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012;22(9):1790-7.
24. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010;26(18):2336-7.
25. Liu G, Liang KY. Sample size calculations for studies with correlated observations. Biometrics. 1997;53(3):937-47.
26. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012. p. http://www.R-project.org/.

Supplemental Figures and Tables

Figure S1. Manhattan plots of random-effect meta-analysis association p-values for dose-response and time-course ECM gene expressions in fibroblasts in response to silica stimulation. Red line corresponds to the p-value cutoff for genome-wide significance (5×10^{-8}); blue line corresponds to the suggestive significance cutoff (1×10^{-5}). P-values are displayed only if association results were non-missing in Caucasians and at least one other ethnic group (African Americans or Hispanics).

Figure S1

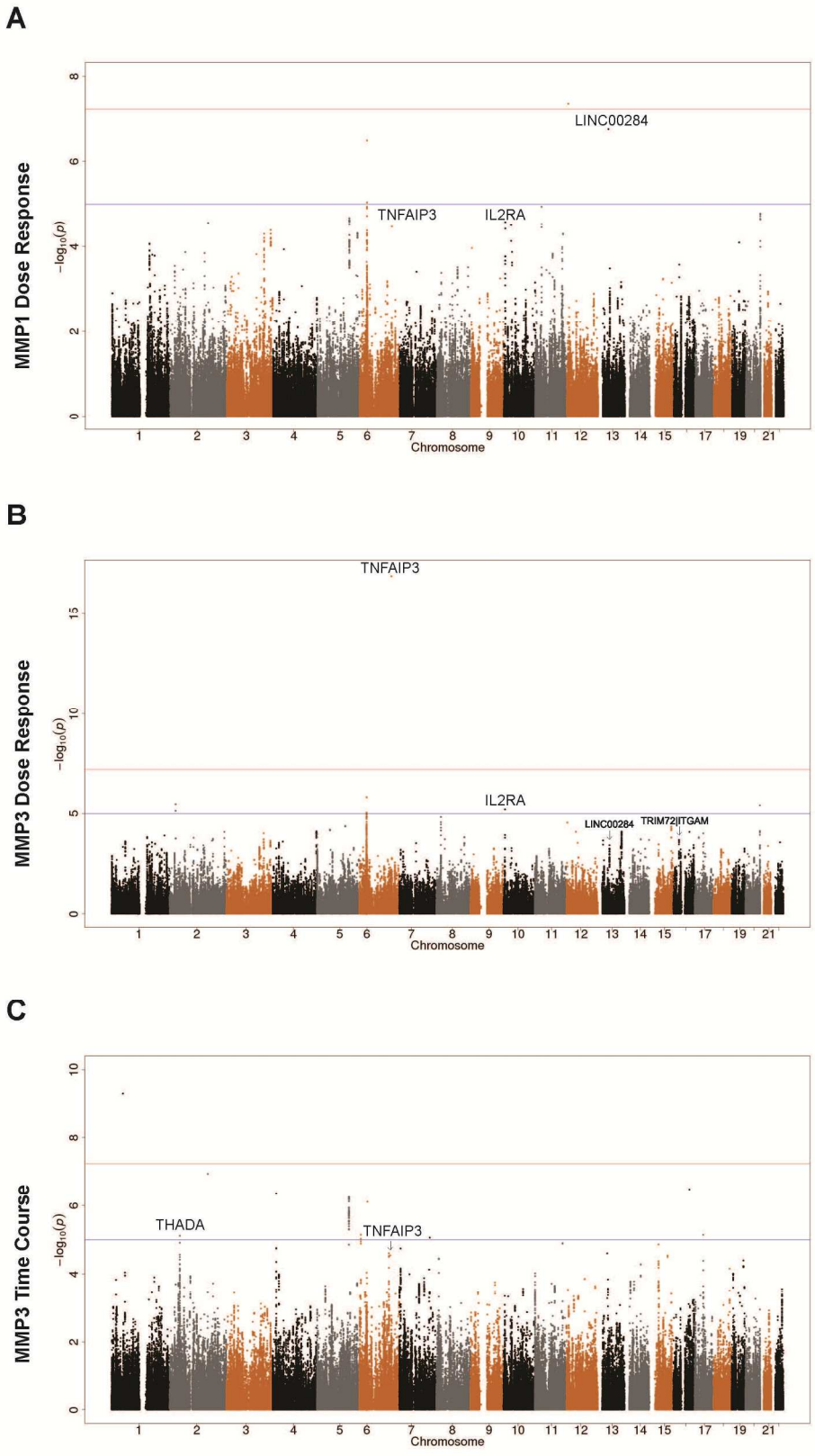

A
MMP1 Dose Response
LINC00284
TNFAIP3
IL2RA
B
MMP3 Dose Response
TNFAIP3
IL2RA
LINC00284
TRIM72|ITGAM
C
MMP3 Time Course
THADA
TNFAIP3
Chromosome

Figure S2. Haplotype-based LMM association analysis of genetic loci with peak SNPs Each alphabetical column (A-F) contains three haplotype-based association regional plots with varying haplotype lengths: top (three consecutive SNPs); middle (five SNPs); bottom (seven SNPs). Red dots represent the peak SNPs (rs79411652 of *TNFAIP3* and rs41290329 of *IL2RA*, respectively) identified in single-SNP based analysis. Haplotype-based analysis was performed in the genetic region defined by 200 SNPs downstream/upstream from the peak SNP. Panel A-F corresponds to the order of panels shown in Figure 3 (regional plots). Panels A to C: genetic locus at 6q23.3 containing SNPs of *TNFAIP3* in association with expression of (A) *MMP1* in dose-response assays, (B) *MMP3* in dose-response assays, and (C) *MMP3* in time-course assay. Panels D to F: genetic locus at 10p15.1 containing SNPs of *IL2RA* in association with expression of (D) *MMP1* in dose-response assays, (E) *MMP3* in dose-response assays, and (F) *MMP3* in time-course assays.

Figure S2

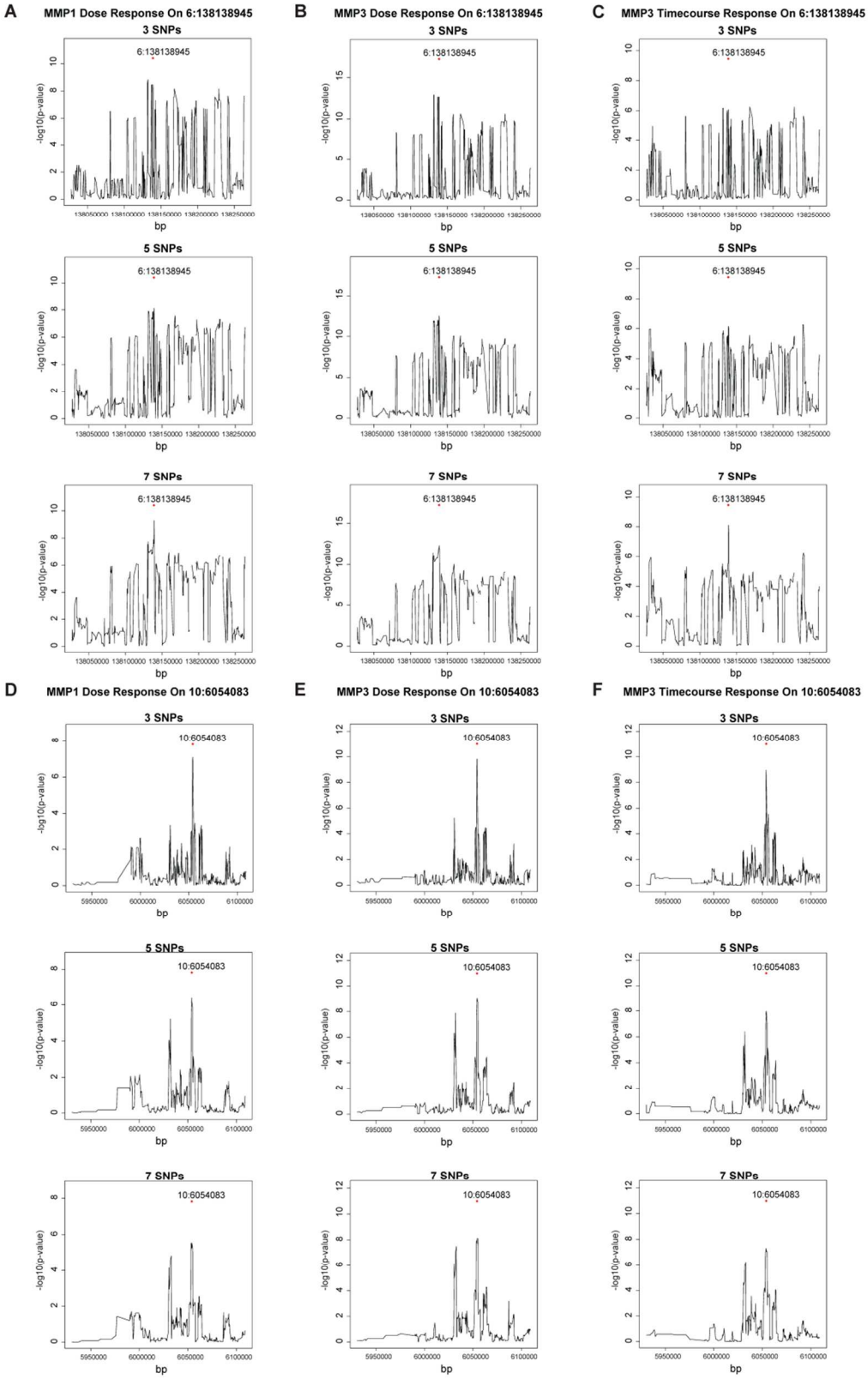
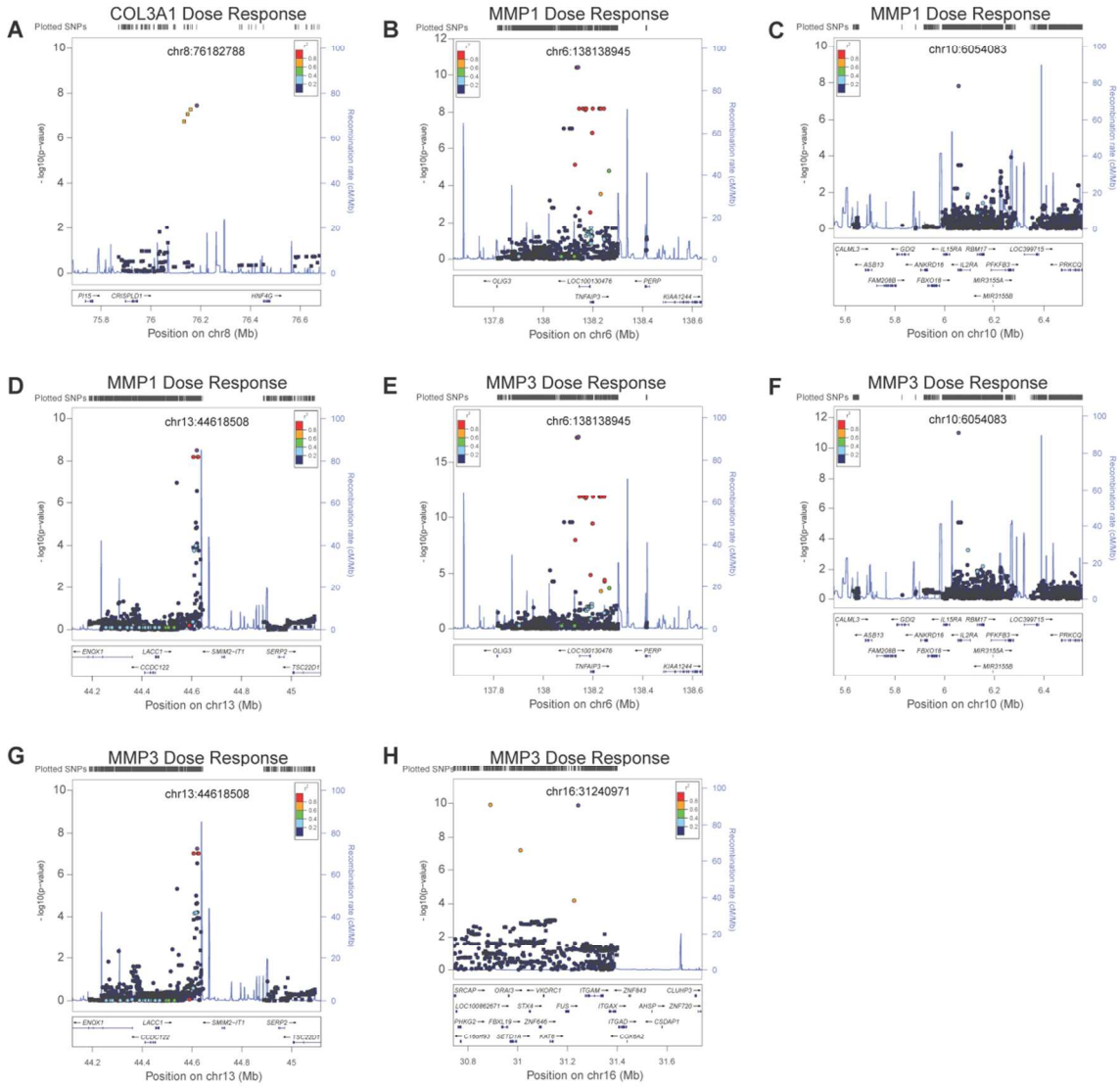

A
MMP1 Dose Response On 6:138138945
B
MMP3 Dose Response On 6:138138945
C
MMP3 Timecourse Response On 6:138138945
D
MMP1 Dose Response On 10:6054083
E
MMP3 Dose Response On 10:6054083
F
MMP3 Timecourse Response On 10:6054083
3 SNPs
5 SNPs
7 SNPs
6:138138945
10:6054083
-log10(p-value)
bp

Figure S3. Regional plots of all associations identified in the Caucasian cohort as listed in Table 2. 500kb upstream and downstream of the peak SNP are plotted. A. SNPs in the intergenic region of *CASC9* in association with *COL3A1* dose-response expression; B. SNPs of *TNFAIP3* in association with *MMP1* dose-response expression; C. SNPs of *IL2RA* in association with *MMP1* dose-response expression; D. SNPs in the intergenic region of *LINC00284|SMIM2-AS1* in association with *MMP1* dose-response expression; E. SNPs of *TNFAIP3* in association with *MMP3* dose-response expression; F. SNPs of *IL2RA* in association with *MMP3* dose-response expression; G. SNPs in the intergenic region of *LINC00284|SMIM2-AS1* in association with *MMP3* dose-response expression; H. SNPs of *ITGAM* in association with *MMP3* dose-response expression; I. SNPs of *THADA* in association with *MMP3* time-course expression; J. SNPs of *TNFAIP3* in association with *MMP3* time-course expression; K. SNPs of *FGFR1OP* in association with *MMP3* time-course expression; L. SNPs of *IL2RA* in association with *MMP3* time-course expression. Round dot represents the array genotyped SNP while square dot represents the imputed SNP; dot filled color shows the measurement of linkage disequilibrium (r^2) between a specific SNP and the peak SNP.

Figure S3


A
COL3A1 Dose Response
chr8:76182788
Position on chr8 (Mb)
B
MMP1 Dose Response
chr6:138138945
Position on chr6 (Mb)
C
MMP1 Dose Response
chr10:6054083
Position on chr10 (Mb)
D
MMP1 Dose Response
chr13:44618508
Position on chr13 (Mb)
E
MMP3 Dose Response
chr6:138138945
Position on chr6 (Mb)
F
MMP3 Dose Response
chr10:6054083
Position on chr10 (Mb)
G
MMP3 Dose Response
chr13:44618508
Position on chr13 (Mb)
H
MMP3 Dose Response
chr16:31240971
Position on chr16 (Mb)

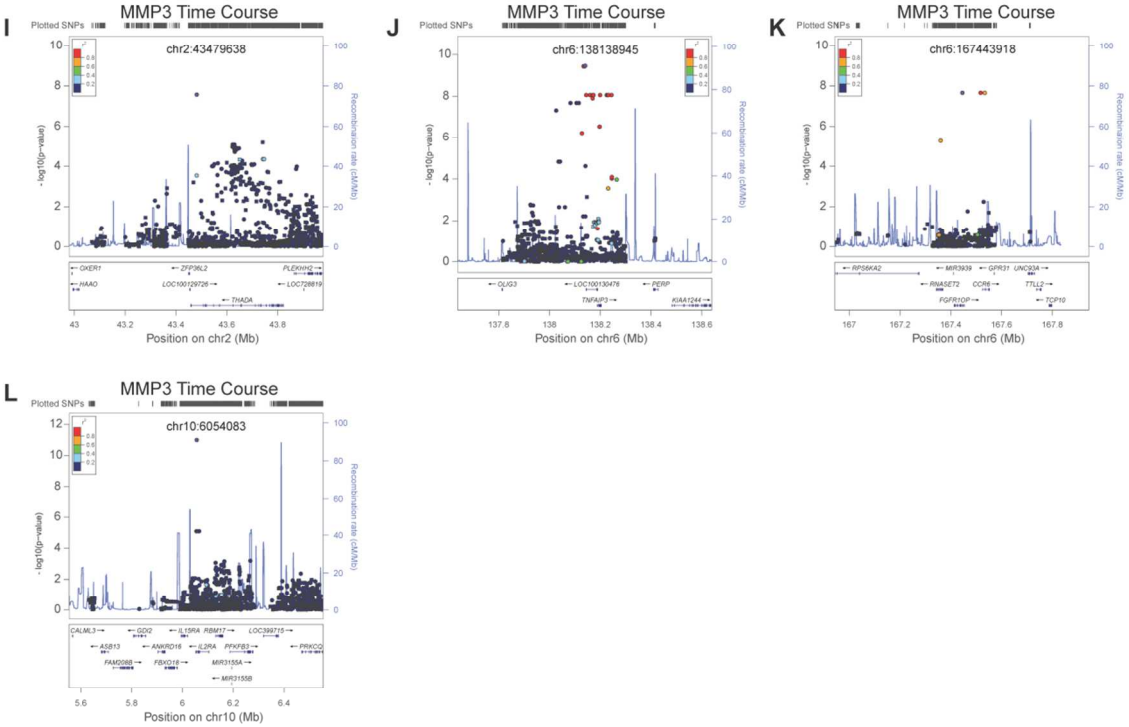

I
MMP3 Time Course
chr2:43479638
Position on chr2 (Mb)
J
MMP3 Time Course
chr6:138138945
Position on chr6 (Mb)
K
MMP3 Time Course
chr6:167443918
Position on chr6 (Mb)
L
MMP3 Time Course
chr10:6054083
Position on chr10 (Mb)

Figure S4. QQ-plots for comparison between Wald test and likelihood-ratio test in LMM association study. Left column corresponds to Wald test, and right column corresponds to LRT test. A. The response variable is *MMP1* dose-response expression; B. the response variable is *MMP3* in dose-response expression; C. The response variable is *MMP3* time-course expression.

Figure S4

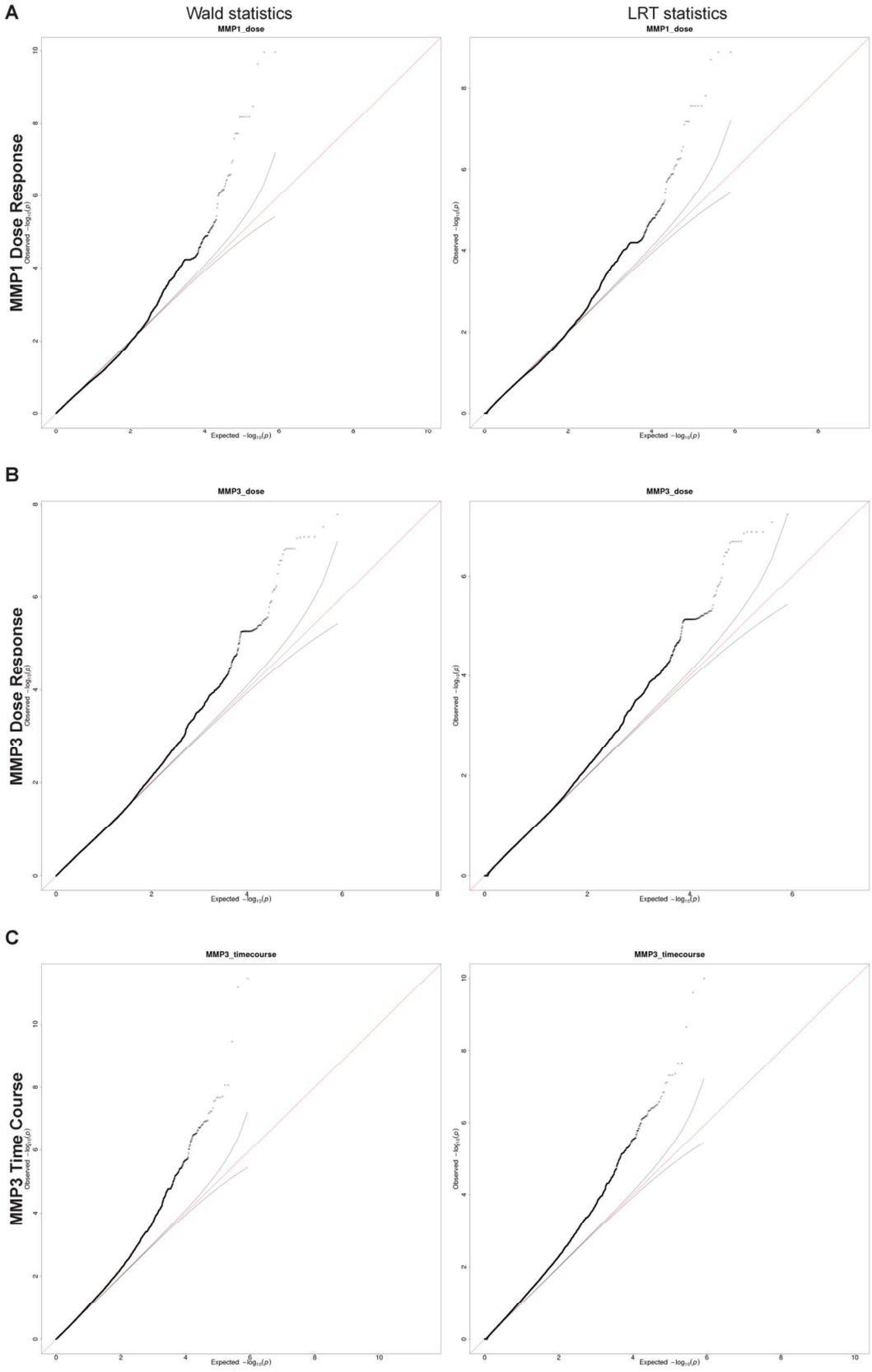

A
Wald statistics
LRT statistics
MMP1_dose
MMP1_dose
MMP1 Dose Response
B
MMP3_dose
MMP3_dose
MMP3 Dose Response
C
MMP3_timecourse
MMP3_timecourse
MMP3 Time Course
Expected $-\log_{10}(p)$
Observed $-\log_{10}(p)$

Figure S5. QQ-plots of LMM-based association p-values of time-course and dose-response gene expression for all three ethnic groups. Each alphabetical row (A-C) contains three plots (left to right) for Caucasians, African Americans and Hispanics, respectively. A. *MMP1* dose-response expression; B. *MMP3* dose-response expression; C. *MMP3* time-course expression. Genomic control lambda is shown in each panel.

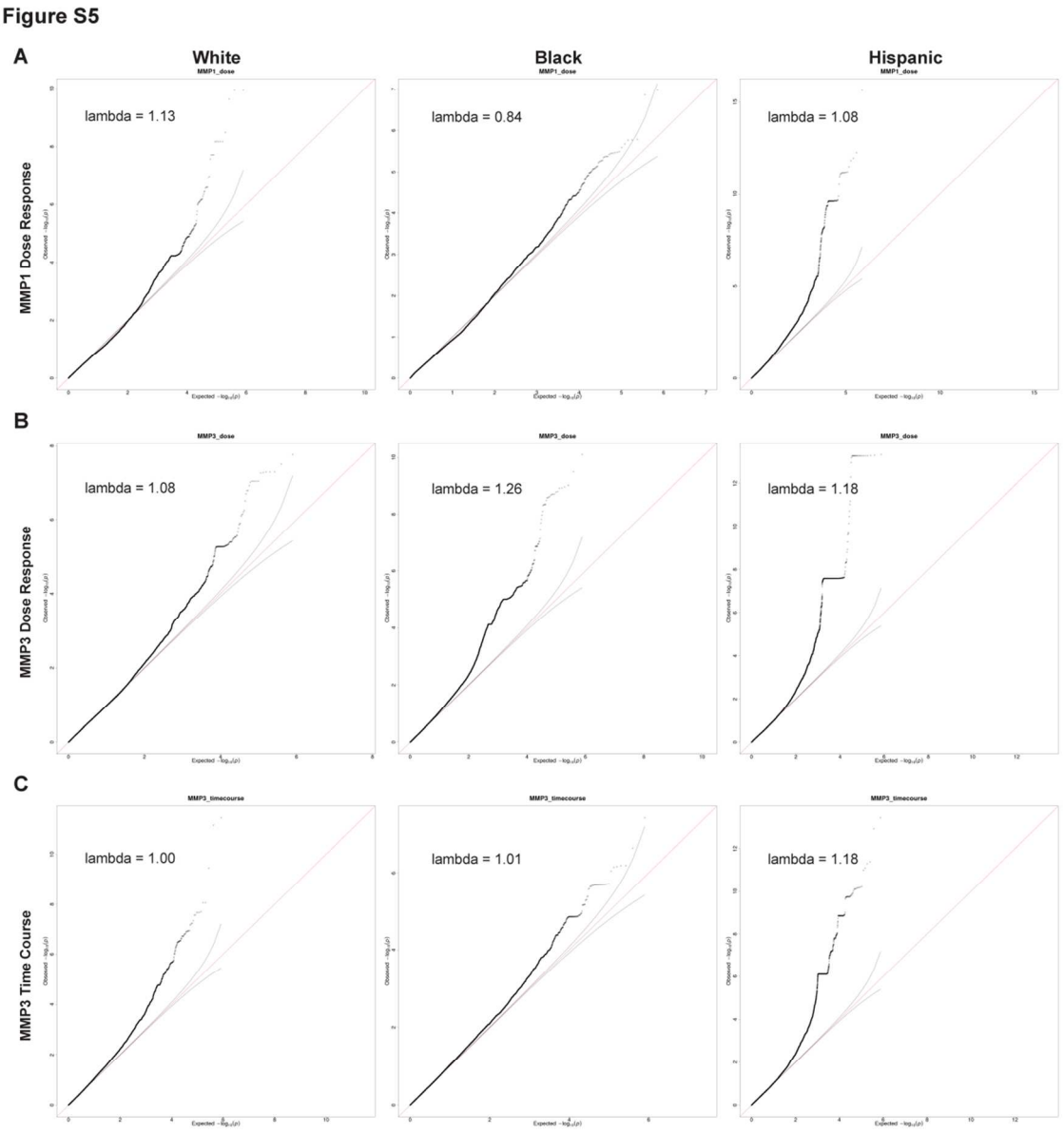


Figure S6. Scatter Plots of the first two principle components for all three ethnic groups. A. Caucasians; B. African Americans; C. Hispanics.

Figure S6

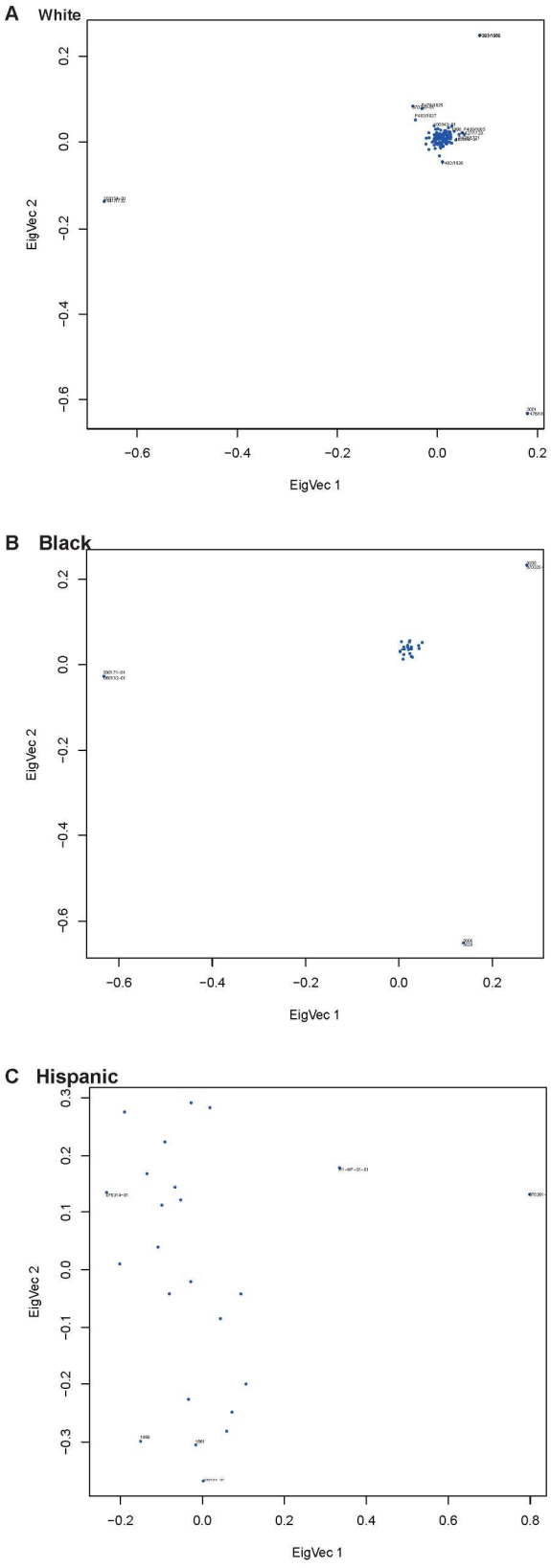

A White
EigVec 2
EigVec 1
B Black
EigVec 2
EigVec 1
C Hispanic
EigVec 2
EigVec 1

Figure S7. Case-control GWAS analysis of SSc in Caucasian samples. A. The overall Manhattan Plot; B. QQ-plot of p-values.

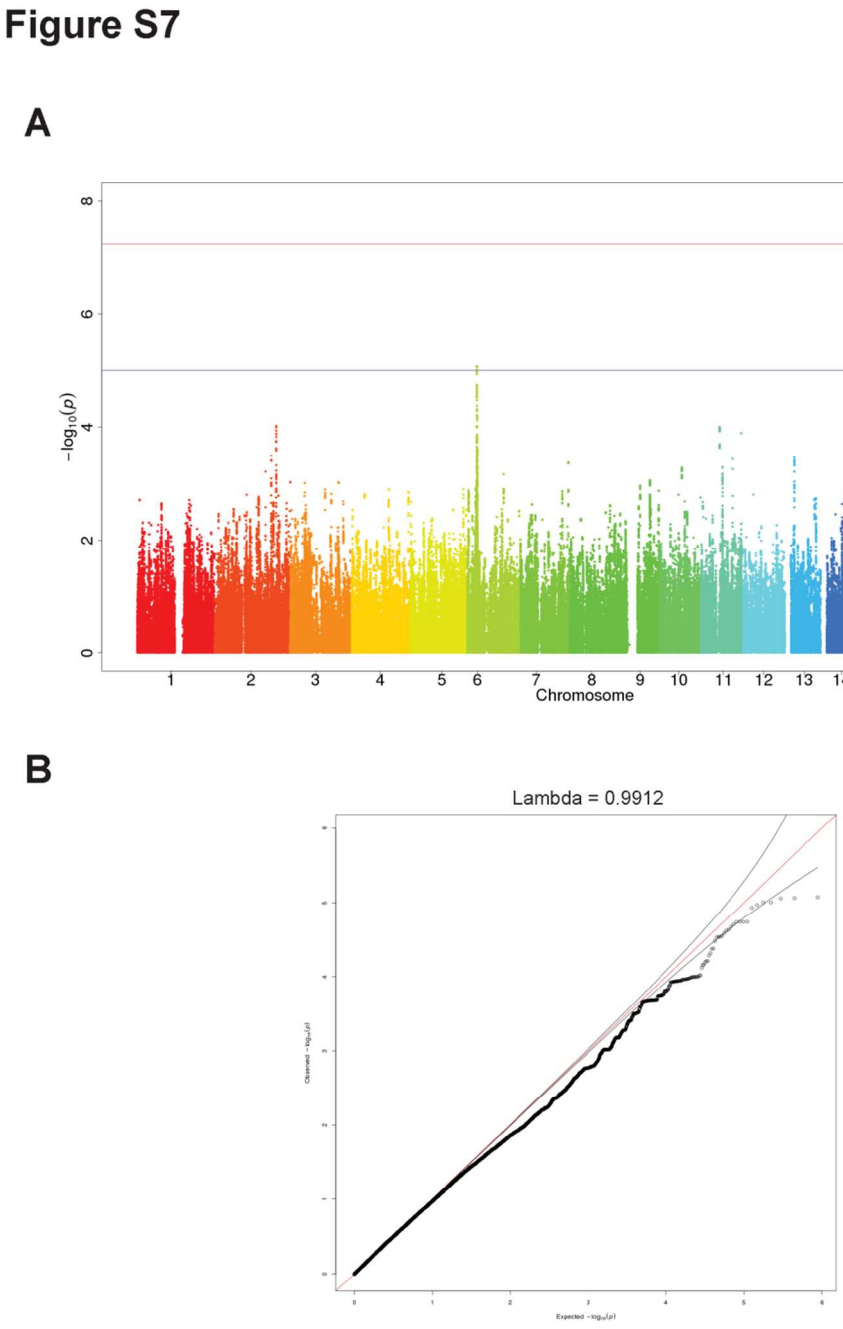


Figure S8. Zoom-in Manhattan plots for case-control GWAS analysis of SSc in Caucasian samples. Chromosomes with known SSc risk genes are displayed. Chromosome 6 is plotted separately for better visual details. Green dots show genetic loci on or near known SSc risk genes. Red line corresponds to the p-value cutoff for genome-wide significance (5×10^{-8}); blue line corresponds to the suggestive significance cutoff (1×10^{-5}).

Figure S8

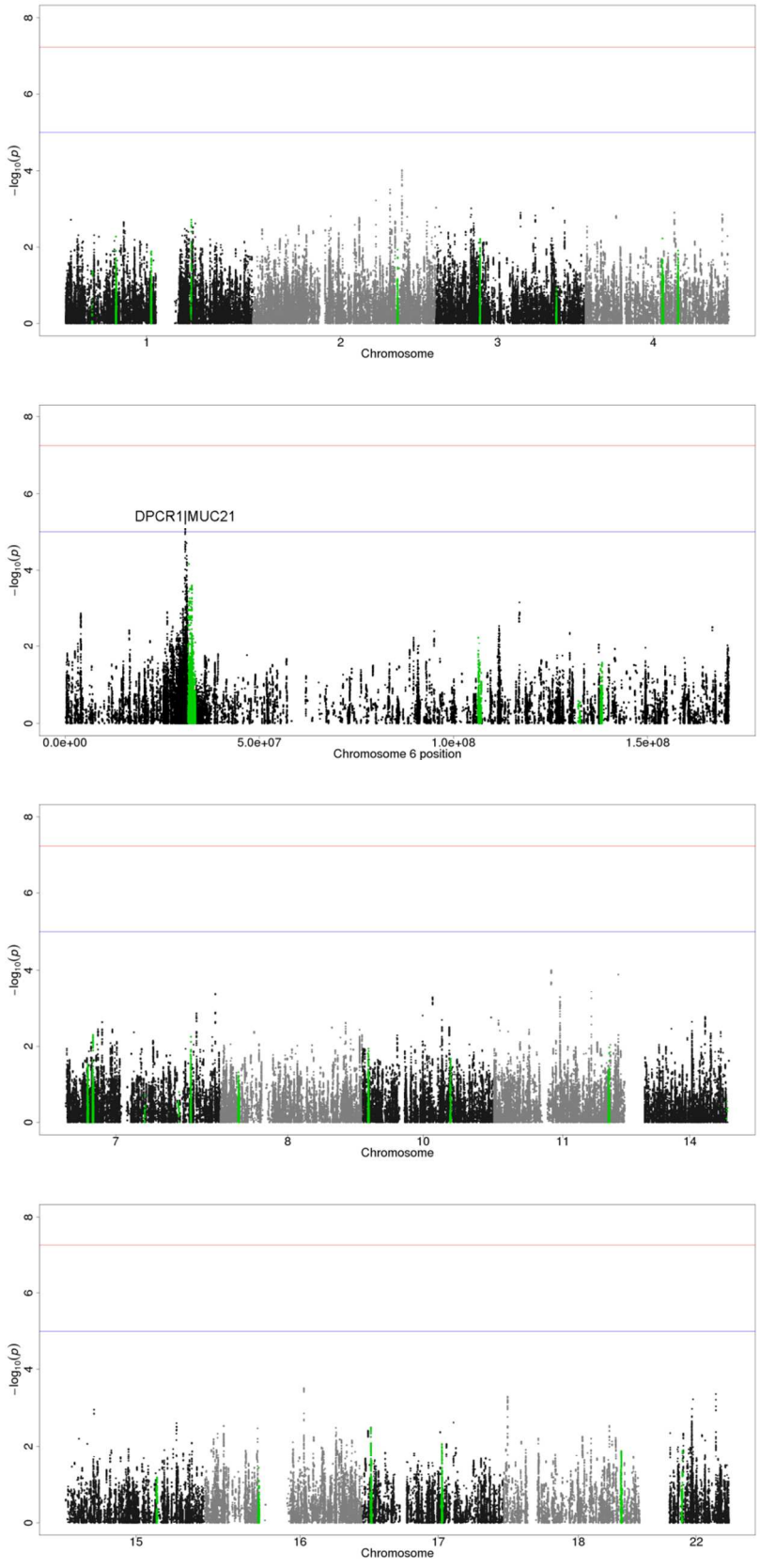

DPCR1|MUC21
−log10(p)
Chromosome
Chromosome 6 position
0.0e+00
5.0e+07
1.0e+08
1.5e+08

Table S1. Distribution of demographics among SSc cases and controls in Caucasians.

Variable	SSc (n = 50) n (%)	Controls (n = 65) n (%)	P-value (χ^2)
Age group			
≤30	2 (4.00)	19 (29.23)	
31−40	8 (16.00)	15 (23.08)	
41−50	16 (32.00)	10 (15.38)	
>50	24 (48.00)	21 (32.31)	0.001253
Sex			
Female	35 (70.00)	37 (56.92)	
Male	15 (30.00)	28 (43.08)	0.2141

Table S2. P-values based on the Satterthwaite approximation to the null distribution of the Wald test statistics for the peak SNPs.

Peak SNP	Chr	Position[a]	Type	Gene(s)[b]	RA	P-Normal	P-Satterthwaite	P-Satterthwaite with Age adjusted	responseType
rs58417815*	8	76159689	ncRNA_intronic	*CASC9*	C	5.10E-08	2.06E-08	1.68E-08	COL3A1_dose
rs7823944	8	76182788	ncRNA_intronic	*CASC9*	T	3.47E-08	1.46E-08	1.32E-08	COL3A1_dose
rs79411652	6	138138945	intergenic	*OLIG3*|***TNFAIP3***[c]	C	3.87E-11	1.95E-10	2.09E-10	MMP1_dose
rs58905141	6	138132123	intergenic	*OLIG3*|***TNFAIP3***	G	3.97E-11	1.98E-10	2.13E-10	MMP1_dose
rs41290329	10	6054083	UTR3	***IL2RA***	C	1.51E-08	2.45E-08	2.72E-08	MMP1_dose
rs79365263	13	44618508	intergenic	*LINC00284*|*SMIM2-AS1*	C	3.13E-09	6.56E-09	7.36E-09	MMP1_dose
rs79411652	6	138138945	intergenic	*OLIG3*|***TNFAIP3***	C	5.55E-18	8.44E-15	5.33E-15	MMP3_dose
rs58905141	6	138132123	intergenic	*OLIG3*|***TNFAIP3***	G	6.43E-18	9.33E-15	5.77E-15	MMP3_dose
rs41290329	10	6054083	UTR3	***IL2RA***	C	1.02E-11	1.49E-10	1.39E-10	MMP3_dose
rs79365263	13	44618508	intergenic	*LINC00284*|*SMIM2-AS1*	C	5.33E-08	1.25E-07	1.17E-07	MMP3_dose
rs12926702	16	31240971	intergenic	*TRIM72*|***ITGAM***	C	1.33E-10	1.03E-09	1.81E-09	MMP3_dose
rs78409037	2	43740411	intronic	*THADA*	C	4.51E-05	8.57E-05	1.03E-04	MMP3_timecourse
rs77533229	2	43479638	intronic	*THADA***	G	2.68E-08	1.91E-07	7.72E-08	MMP3_timecourse
rs79411652	6	138138945	intergenic	*OLIG3*|***TNFAIP3***	C	3.52E-10	7.14E-09	3.97E-09	MMP3_timecourse
rs58905141	6	138132123	intergenic	*OLIG3*|***TNFAIP3***	G	3.73E-10	7.45E-09	4.14E-09	MMP3_timecourse
rs75701002	6	167443918	intronic	*FGFR1OP*	C	2.18E-08	1.62E-07	7.59E-07	MMP3_timecourse
rs41290329	10	6054083	UTR3	***IL2RA***	C	1.03E-10	5.64E-10	3.48E-10	MMP3_timecourse

RA, risk allele; P-Normal, P-value using Normal Approximation in Non-Hispanic White; P-Satterthwaite, P-value using Satterthwaite Degree of Freedom Approximation in Non-Hispanic White; P-Satterthwaite with Age adjusted, linear mixed model additionally adjust for sample's age as a fixed effect.

[a]GRCh37/hg19 assembly. [b]Gene in which the peak signal is located in, or genes flanking peak signal if intergenic. [c]Bold-face Genes are known SSc risk loci.
*Imputed, otherwise Peak SNP was genotyped.

**Peak SNP in Caucasians was different from that in fixed-effect meta-analysis with both SNPs listed.

Table S3. Meta-analysis results for all the genome-wide significant SNPs in the discovery analysis of the Caucasian samples.

In a separate excel file

END of Supplemental File