Copy Number Variation of *HLA-DQA1* and *APOBEC3A/3B* Contribute to the Susceptibility of Systemic Sclerosis in the Chinese Han Population

Shicheng Guo, Yuan Li, Yi Wang, Haiyan Chu, Yulin Chen, Qingmei Liu, Gang Guo, Wenzhen Tu, Wenyu Wu, Hejian Zou, Li Yang, Rong Xiao, Yanyun Ma, Feng Zhang, Momiao Xiong, Li Jin, Xiaodong Zhou, and Jiucun Wang

ABSTRACT. ***Objective.*** Systemic sclerosis (SSc) is a systemic connective tissue disease caused by a genetic aberrant. The involvement of the copy number variations (CNV) in the pathogenesis of SSc is unclear. We tried to identify some CNV that are involved with the susceptibility to SSc.

Methods. A genome-wide CNV screening was performed in 20 patients with SSc. Five SSc-associated common CNV that included *HLA-DRB5*, *HLA-DQA1*, *IRGM*, *CDC42EP3*, and *APOBEC3A/3B* were identified from the screening and were then validated in 365 patients with SSc and 369 matched healthy controls.

Results. Three hundred forty-four CNV (140 gains and 204 losses) and 2 CNV hotspots (6q21.3 and 22q11.2) were found in the SSc genomes (covering 24.2 megabases), suggesting that CNV were ubiquitous in the SSc genome and played important roles in the pathogenesis of SSc. The high copy number of *HLA-DQA1* was a significantly protective factor for SSc (OR 0.07, $p = 2.99 \times 10^{-17}$), while the high copy number of *APOBEC3A/B* was a significant risk factor (OR 3.45, $p = 6.4 \times 10^{-18}$), adjusted with sex and age. The risk prediction model based on genetic factors in logistic regression showed moderate prediction ability, with area under the curve = 0.80 (95% CI 0.77–0.83), which demonstrated that *APOBEC3A/B* and *HLA-DQA1* were powerful biomarkers for SSc risk evaluation and contributed to the susceptibility to SSc.

Conclusion. CNV of *HLA-DQA1* and *APOBEC3A/B* contribute to the susceptibility to SSc in a Chinese Han population. (J Rheumatol First Release April 1 2016; doi:10.3899/jrheum.150945)

Key Indexing Terms:
SYSTEMIC SCLEROSIS GENETIC PREDISPOSITION TO DISEASE HLA ANTIGENS
COPY NUMBER VARIATION CASE-CONTROL STUDIES

From the State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University; Institute of Rheumatology, Immunology and Allergy, Fudan University; Shanghai Traditional Chinese Medicine (TCM)-Integrated Hospital; Division of Dermatology, and Division of Rheumatology, Huashan Hospital, Fudan University, Shanghai; Yiling Hospital, Shijiazhuang; Division of Rheumatology, Teaching Hospital of Chengdu University of TCM, Chengdu; Department of Dermatology, Second Xiangya Hospital, Central South University, Changsha, China; School of Public Health, and Medical School at Houston, University of Texas, Houston, Texas, USA.

Supported by research grants from the National Basic Research Program (2012CB944604), National Science Foundation of China (81270120, 81470254), International S&T Cooperation Program of China (2013DFA30870), the 111 Project (B13016), and the US National Institutes of Health (NIH) NIAID UO1, 1U01AI09090. The computations involved in this study were supported by Fudan University High-End Computing Center. Mr. Xiong was supported by Grant 1R01AR057120–01 and 1R01HL106034-01, from the NIH and the US National Heart, Lung, and Blood Institute.

S. Guo, PhD, State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University, and Institute of Rheumatology, Immunology and Allergy, Fudan University; Y. Li, MS, State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University; Y. Wang, PhD, State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University; H. Chu, PhD, State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University; Y. Chen, PhD, State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University; Q. Liu, PhD, State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University; G. Guo, MD, Yiling Hospital; W. Tu, MD, Shanghai TCM-Integrated Hospital; W. Wu, MD, Institute of Rheumatology, Immunology and Allergy, Fudan University, and Division of Dermatology, Huashan Hospital, Fudan University; H. Zou, PhD, MD, Institute of Rheumatology, Immunology and Allergy, Fudan University, and Division of Rheumatology, Huashan Hospital, Fudan University; L. Yang, MD, Division of Rheumatology, Teaching

Hospital of Chengdu University of TCM; R. Xiao, MD, Department of Dermatology, Second Xiangya Hospital, Central South University; Y. Ma, MS, State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University; F. Zhang, PhD, State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University; M. Xiong, PhD, School of Public Health, University of Texas; L. Jin, PhD, State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University; X. Zhou, MD, Medical School at Houston, University of Texas; J. Wang, PhD, State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University, and Institute of Rheumatology, Immunology and Allergy, Fudan University.

Address correspondence to J. Wang, School of Life Sciences, Fudan University Jiangwan Campus, 2005 Songhu Road, Shanghai 200438, China. E-mail: jcwang@fudan.edu.cn

Accepted for publication January 26, 2016.

Systemic sclerosis (SSc), also called scleroderma, is an immune-mediated disease characterized by extensive fibrosis of the skin and associated with various degrees of chronic inflammatory infiltration and significant microangiopathy, and changes in humoral or cellular immune system[1,2]. According to the degree and extent of fibrosis, there are 2 major clinical subtypes: limited cutaneous SSc (lcSSc) and diffuse cutaneous SSc (dcSSc).

Large numbers of the epidemiological characteristics were significantly different between the populations with different genetic architecture. For example, the incidence of SSc ranged from 3.7–23 per 100,000 population in different ethnic groups[3,4]; the prevalence of SSc for women is 4-fold to 5-fold higher than that for the male population[5,6]. Arnett, *et al*[7] and Mayes, *et al*[5] found that siblings had about a 15-fold higher risk of SSc, while first-degree relatives had about a 13-fold higher risk of SSc. In addition, analysis of SSc in twins reveals low concordance for the disease. However, high concordance for the presence of antinuclear antibodies (ANA) was observed[8]. This evidence shows that SSc is a complex disease caused by specific genetic and genomic variants[9]. Several genome-wide association followup studies[10,11,12,13] and case-control studies have shown multiple susceptible single-nucleotide polymorphisms (SNP) associated with susceptibility to SSc, such as *PPARG*[13], *IRF5*[14], transforming growth factor-β receptor[15], *TNIP1*[16], *STAT4*[17], *RHOB*, and more. However, we demonstrated previously that even for high familial risk diseases such as thyroid cancer, a few of the significant SNP could just have limited prediction power[18]. Some other genetic or epigenetic variation should be discovered, such as copy number variation (CNV)[19]. CNV is one of most important sources of genetic structural diversity in the human genome; it can cause gene structure and accordingly gene expression change. Evidence showed that at least 8.75% to 17.7% of the variation in gene expression could be explained by CNV[20]. Some CNV have been demonstrated to be widely associated with susceptibility to immune-mediated diseases such as ankylosing spondylitis (AS)[21], rheumatoid arthritis[22,23,24], systemic lupus erythematosus[25,26], and others. Some evidence has demonstrated that specific CNV were associated with SSc[27].

In our present study, we analyzed potential SSc-associated CNV with Agilent array comparative genomic hybridization (aCGH) and AccuCopy CNV genotyping technologies[28]. Genome-wide aCGH microarrays were conducted to detect CNV in 20 patients with SSc, and then 5 candidate CNV that were suspected of being related to the pathology of SSc including *HLA-DRB5*, *HLA-DQA1*, *IRGM*, *CDC42EP3*, and *APOBEC3A/3B*. These CNV were validated in a large Chinese Han population.

MATERIALS AND METHODS

Patients and controls. SSc mainly includes 2 subtypes: lcSSc and dcSSc. In the discovery stage, we covered both of them in 5 male and 5 female samples so that we could get some unbiased candidate CNV associated with SSc and adjusted for sex and subtype of SSc in a Chinese Han population (4 categories with sex and subtypes) for the aCGH array screening. In the validation stage, a total of 734 subjects were enrolled, including 365 cases with SSc and 369 ethnically matched healthy controls. All patients were recruited from a multicenter study that included hospitals and outpatient clinics in Shanghai, Hebei province, Sichuan province, and Hunan province in China. Patients either met the American College of Rheumatology classification criteria for SSc[29] or had at least 3 out of 5 CREST features (calcinosis, Raynaud phenomenon, esophageal dysmotility, sclerodactyly, and telangiectasia) with sclerodactyly being mandatory[30].

DNA extraction, autoantibody test, and organ involvement assessment. Peripheral blood was collected from all subjects. Genomic DNA was isolated from whole blood and stored at –30°C until used in our previous study[17]. The following autoantibodies were detected in patient sera: ANA, anti-DNA topoisomerase I (ATA), anticentromere (ACA), anti-U1RNP (ARA), anti-RNA polymerase 3 (anti-RNAP 3), anti-Sm, anti-SSA, anti-SSB, anti-PM-1, anti-Jo1 antibodies, and rheumatoid factor. The status of the autoantibodies was measured as binary data: positive and negative, except anti-RNAP 3, which had continuous data. In addition, autoantibody detection details could be found in our previous work[6]. Pulmonary fibrosis was assessed with radiograph and/or computed tomography. Organ involvement was defined as Steen and Medsger suggested[31].

Genome-wide CNV analysis. As in our previous study[21], Agilent SurePrint G3 Human CGH 1×1M Oligo Microarray was performed for genome-wide CNV detection and genotyping. DNA was isolated from 20 patients with SSc. For each sample, 2.2 μg input genomic DNA was restriction-digested and labeled with ULS-Cy5 and ULS-Cy3 in accordance with the manufacturer's protocol (Agilent 2010). The labeled product was then hybridized to the array and scanned on the Agilent Microarray Scanner. The data were extracted by Agilent Feature Extraction 10.7.3.1 and analyzed by Agilent Workbench 7.0 with default variables. A human genome coordinate was conversed in hg19 uniformly in our present study.

AccuCopy technology for CNV validation. Five common CNV resulting from aCGH were validated with the AccuCopy assay[28] (a multiple competitive real-time PCR) by Genesky Bio-Tech. Briefly, the genomic DNA of each subject was mixed with fluorescence-labeled specific primers (Supplementary Table 1, available from the authors on request), PCR Master Mix, and a competitive DNA with known copy number for a multiple

Table 1. Proportion of antibody and organ involvement. Values are n (%) unless otherwise specified.

Variables	Positive	Negative	Total, n
Antibodies			
ANA	278 (76.2)	18 (4.9)	296
ATA	164 (44.9)	151 (41.4)	315
ACA	32 (8.8)	227 (62.2)	259
ARA	45 (12.3)	251 (68.8)	296
Anti-Sm	11 (3)	250 (68.5)	261
Anti-SSA	84 (23)	181 (49.6)	265
Anti-SSB	22 (6)	239 (65.5)	261
Anti-PM-1	7 (1.9)	163 (44.7)	170
Anti-Jo1	9 (2.5)	241 (66)	250
RF	44 (12.1)	139 (38.1)	183
Organ involvement			
Lung	77 (21.1)	211 (57.8)	288
Heart	194 (53.2)	86 (23.6)	280

ANA: antinuclear; ATA: anti-DNA topoisomerase I; ACA: anticentromere; ARA: anti-U1RNP; RF: rheumatoid factor.

competitive real-time PCR reaction. The PCR products were diluted and loaded onto an ABI 3730XL sequencer for quantification analysis. Raw data were analyzed by Gene Mapper 4.0, and Hg19 was used for the genome build for the genomic coordinates. The peak ratio between sample DNA and the corresponding competitive DNA (S/C) was calculated and then normalized to the median of the 4 preset 2-copy reference genes, respectively. Two normalized S/C ratios were further normalized to the median value in all samples for each reference gene and then averaged. The copy number of each target fragment was determined by the average S/C ratio × 2. Cases and controls were examined and read at the same time to minimize nonrandom errors.

Statistical analysis. In the validation stage, binary logistic regression was applied to discover association between CNV and SSc, and adjusted with sex and age to estimate the marginal effects of candidate CNV (marginal effect model). The partial effect for a specific CNV was also estimated and adjusted with other remaining CNV, sex, and age (partial effect model). In the partial effect model, for each of the 5 CNV selected for validation, the remaining 4 were adjusted so that the partial effect for each CNV could be estimated simultaneously. OR and 95% CI were calculated with the R code. Association between CNV and organ involvement was conducted with binary logistic model, while association between CNV and anti-RNAP 3 (continuous variable) was conducted with linear regression models. The individual with missing age, sex, or some other information would be omitted in the regression model. Chi-square or the Fisher's exact test was applied for an independent test between autoantibody and organ involvement. The effect size of the CNV in subgroup analyses was conducted to certain samples with specific clinical characteristics compared with normal. R packages[32] "PredictABEL"[33] and "pROC"[34] were applied for the receiver-operating characteristic (ROC) plot.

RESULTS

CNV discovery in patients with SSc. Genome-wide CNV analysis was conducted in 20 patients with SSc with aCGH array. The demographic characteristics of the 20 samples are shown in Supplementary Table 2 (available from the authors on request). There were 344 CNV (average 57.3, SD 8.3) found in the 20 individuals, including 140 gains and 204 losses, and they covered a 24.2-megabase of the whole genome (Figure 1). CNV hotspot was observed in the 6q21.3 region, which indicated that CNV in the HLA region might be associated with SSc. In the 22q11.2 region, a CNV hotspot was also detected with a large amount of CNV loss. After filtering the common CNV and our previously reported Chinese common CNV (1440 CNV regions)[35], there were a total of 31 CNV remaining (8 gains, 23 loss; average = 5, SD 2.9), some of which may be the causal/risk CNV for SSc. There was no significant difference in the length of the common and novel CNV (Wilcoxon test, $p = 0.26$). There were 23 genes and correspondingly 119 RNA involved with 31 novel CNV (Supplementary Table 3, available from the authors on request). There was no significant difference between the loss/gain ratio between common and new CNV (chi-square test, $p = 0.11$). In our SSc CNV map, the number of deletion counts was much higher than that of duplication counts (Student t test, $p < 0.004$), while the size of duplication was much larger than that of deletion (Student t test, $p = 0.03$). These results suggested that common or novel CNV may be ubiquitous in the SSc genome and be involved in the pathogenesis of SSc.

Association between CNV of APOBEC3A/3B, HLA-DQA1, and SSc susceptibility. To confirm what we found in the discovery stage, 5 common CNV regions (*HLA-DRB5*, *HLA-DQA1*, *IRGM*, *CDC42EP3*, and *APOBEC3A/3B*), which were found in at least 2 of the above samples, were validated in the validation stage. A total of 365 patients with

Table 2. Association between serological profile and clinical characteristics in SSc. Chi-square or 1-way ANOVA was conducted to make association analysis between categorical variables or categorical continuous variables. P value < 0.05 was considered a significant association. Corresponding OR and 95% CI are shown in Supplementary Table 5 (available from the authors on request).

Characteristics	ANA	ATA	ACA	Anti-RNAP 3	ARA	Anti-Sm	SSA	SSB	Anti-PM-1	RF
Sex	0.3223	0.6592	0.1294	0.8778	0.003	0.1394	0.8701	0.5802	0.6897	0.5517
lcSSc/dcSSc	1	0.025	0.0165	0.0902	0.0105	0.2184	0.061	0.017	0.0585	0.1714
CREST	1	1	1	0.0198	0.0145	0.2599	0.2024	1	0.1979	1
Lung	0.7411	0.0205	0.026	0.8646	0.4398	0.1709	0.8726	0.08	1	0.3983
Heart	0.5132	1	0.8106	0.2396	0.7251	0.7676	0.3678	0.4733	1	0.8406

SSc: systemic sclerosis; ANA: antinuclear antibodies; ATA: anti-DNA topoisomerase I antibody; ACA: anticentromere antobodies; anti-RNAP 3: anti-RNA polymerase 3 antibodies; ARA: anti-U1RNP; RF: rheumatoid factor; lcSSc: limited cutaneous SSc; dcSSc: diffuse cutaneous SSc; CREST (syndrome): calcinosis, Raynaud phenomenon, esophageal dysmotility, sclerodactyly, and telangiectasias.

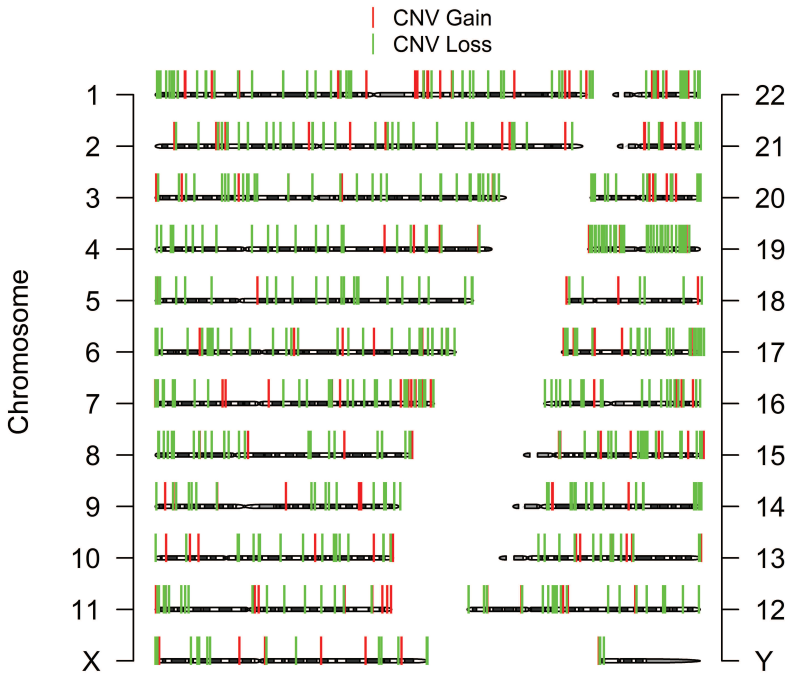

Figure 1. Genome-wide aCGH analysis reveals copy number variation (CNV) in the SSc population. Significant SSc-associated CNV identified by aCGH array were noted in the human genome. Red bars represents high copy number in SSc, while green bars represent low copy number in SSc. aCGH: array comparative genomic hybridization; SSc: systemic sclerosis.

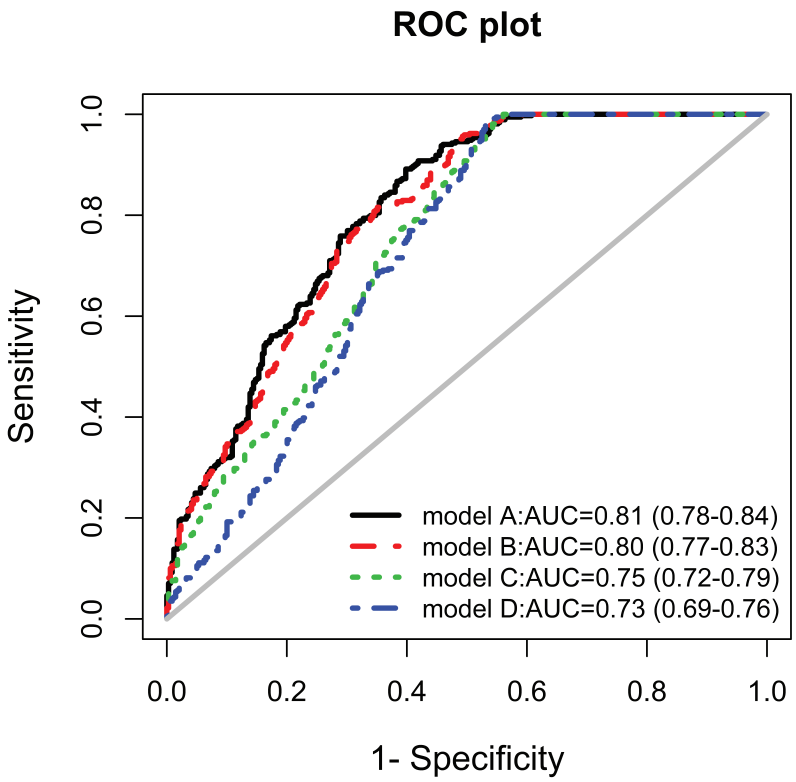

Figure 2. Performance of 4 prediction models based on the ROC curve. Model A included sex, age, and these CNV: *HLA-DRB5*, *HLA-DQA1*, *IRGM*, *CDC42EP3*, and *APOBEC3A/3B*. Model B included the sex, age, and copy number of *HLA-DQA1* and *APOBEC3A/3B*, which were selected with a forward conditional stepwise method from all the covariates. Model C included sex and copy number of *HLA-DQA1* and *APOBEC3A/3B*. Model D included 2 SSc-susceptibility genetic factors and the copy number of *HLA-DQA1* and *APOBEC3A/3B*. ROC: receiver-operation characteristic; CNV: copy number variations; SSc: systemic sclerosis; AUC: area under the curve.

Table 3. Association between CNV and SSc in marginal and partial logistic model.

Variables	OR (95% CI)*	p*	OR (95% CI)**	p**
HLA-DRB5	2.54 (2.09–3.08)	**7.40×10^{-21}**	1.25 (0.98–1.59)	0.073
HLA-DQA1	0.07 (0.03–0.12)	**2.99×10^{-17}**	0.09 (0.04–0.18)	**2.59×10^{-11}**
Near *IRGM*	1.19 (0.95–1.49)	0.1319	1.25 (0.96–1.63)	0.093
Near *CDC42EP3*	1.17 (0.96–1.44)	0.1209	1.33 (1.05–1.68)	0.02
APOBEC3A/3B	3.4 (2.58–4.5)	**6.40×10^{-18}**	1.6 (1.18–2.18)	**0.002**

* Values from partial association based on multivariate logistic regression. ** Values from marginal association based on univariate logistic regression. Both of these logistic regression tests were adjusted for age and sex. Neither p value was adjusted by multiple comparison correction. Significant data are in bold face. CNV: copy number variations; SSc: systemic sclerosis.

SSc and 369 ethnically matched healthy individuals were enrolled in our study. There were no significant differences for the proportion of men/women in cases and controls (Supplementary Table 4, available from the authors on request). The patients with SSc included 52% with lcSSc and 41% with dcSSc, while the remaining 7% were not clearly subtyped and therefore could not be assigned to either group. The autoantibody status is shown in Table 1. Four autoantibodies were observed to be significantly differently distributed between lcSSc and dcSSc subtypes: ACA (OR 2.58, 95% CI 1.2–5.7, p = 0.024), ATA (OR 0.57, 95% CI 0.35–0.91, p = 0.025), ARA (OR 2.37, 95% CI 1.22–4.62, p = 0.014), and anti-SSB (OR 0.27, 95% CI 0.09–0.83, p = 0.016). The ARA-positive proportion in women was significantly higher than in men (OR 11.5, Fisher's exact test, p = 0.003). Associations among other clinical characteristics are shown in Table 2 and Supplementary Table 5 (available from the authors on request).

The copy numbers of 5 genomic regions located in or near *HLA-DRB5*, *HLA-DQA1*, *IRGM*, *CDC42EP3*, and *APOBEC3A/3B* were examined in 365 patients with SSc and 369 healthy individuals (Supplementary Table 3, available from the authors on request). Among all the samples, CNV genotyping rates were greater than 80%. Both marginal effect and partial effect from the logistic regressions showed that CNV of *HLA-DQA1* and *APOBEC3A/B* were significantly associated with the risk of SSc (Table 3). Low copy number of *HLA-DQA1* was a significantly protective factor for SSc (OR 0.07, $p = 2.99 \times 10^{-17}$) while high *APOBEC3A/B* was a significant risk factor (OR 3.45, $p = 6.4 \times 10^{-18}$). Inconsistent conclusions were found for *HLA-DRB5* and *CDC42EP3* between the marginal and partial effect models. High *HLA-DRB5* was significantly associated with SSc in the marginal effect model (OR 2.54, $p = 7.4 \times 10^{-21}$), while high *CDC42EP3* was significantly associated with SSc only in the partial effect model (OR 1.33, p = 0.02). Because no significant association was found between CNV of *IRGM* and SSc, subgroup univariate logistic regressions were conducted to validate whether CNV of *IRGM* was associated with certain subtypes of SSc. Case individuals from the subtypes by the status of 10 autoimmune antibodies or organ involvement were compared with controls using binary logistic regression. However, no significant association was found between *IRGM* and susceptibility to any subtype of SSc (Supplementary Tables 6–11, available from the authors on request).

Risk prediction models established based on sex, age, and/or CNV. Compared with the OR to display the association between genetic variants and a phenotype, disease risk prediction models are more clinically useful. Herein, we report 4 prediction models of the absolute risk for patients with SSc; these models were based on a logistic regression model.

As Figure 2 shows, model A included sex, age, and all of the 5 identified CNV. Model B included the sex, age, and copy number of *HLA-DQA1* and *APOBEC3A/3B*, which were selected with the forward conditional stepwise method from all the covariates. Model C included sex and copy number of *HLA-DQA1* and *APOBEC3A/3B*, because these factors remain unchanged. Model D features included SSc-susceptibility genetic factor and copy number *of*

Table 4. Distinguishing the ability of 4 models to predict risk of SSc. The outpoint corresponding to maximum of the sum of sensitivity and specificity was considered the threshold.

Model	Sensitivity	Specificity	AUC (95% CI)	Variate
Model A	0.89	0.60	0.81 (0.78–0.84)	Sex, age, 5 CNV
Model B	0.82	0.65	0.80 (0.77–0.83)	Sex, age, *HLA-DQA1*, *APOBEC3A/3B*
Model C	0.99	0.45	0.75 (0.72–0.79)	Sex, *HLA-DQA1*, *APOBEC3A/3B*
Model D	0.99	0.46	0.73 (0.69–0.76)	*HLA-DQA1*, *APOBEC3A/3B*

SSc: systemic sclerosis; AUC: area under the curve; CNV: copy number variations.

HLA-DQA1 and *APOBEC3A/3B* (Table 4). The ROC curve analysis showed that *HLA-DQA1* and *APOBEC3A/3B* combined with sex and age had moderate prediction power [area under the curve (AUC) = 0.80, 95% CI 0.77–0.83]. Thus, the CNV of *HLA-DQA1* and *APOBEC3A/3B* would be a potentially independent genetic susceptibility factor to SSc.

DISCUSSION

Our results provided evidence that the low copy number of *HLA-DQA1* (OR 0.07, $p = 2.99 \times 10^{-17}$) and high copy number of *APOBEC3A/3B* (OR 3.45, $p = 6.4 \times 10^{-18}$) significantly contributed to the susceptibility to SSc in the Chinese Han population.

Common CNV represent an important source of genetic diversity, yet their influence on phenotypic variability and disease susceptibility remains poorly analyzed. We provided more evidence about the involvement of common CNV on complex disease, especially on immune system-related disease. Our genome-wide CNV profile in SSc and normal subjects was also supported by some previous studies, such as one finding that 22q11.2 deletion may result in variable clinical phenotypes and that 22q11.2 deletion individuals are at increased risk of a variety of autoimmune diseases[36]. In addition, the CNV pattern in our present study were also observed in our previous discovery in a healthy Chinese population[35].

The *HLA-DQA1* gene is one of the HLA complex genes encoding a protein that presents specific antigen peptides to T cell receptor to initiate immune response. Genetic allelic variation in *HLA-DQA1* has been reported to be associated with SSc[37]. The association of CNV of *HLA-DQA1* with SSc indicates dosage of *HLA-DQA1* and influences the susceptibility to SSc. Additionally, CNV of *HLA-DQA1* has been confirmed to be associated with many kinds of autoimmune diseases, such as AS[21].

APOBEC3 are a family of DNA-editing enzymes thought to be part of the innate immune system by restricting retroviruses, mobile genetic elements such as retrotransposons, and endogenous retroviruses. *APOBEC3A* is an important epigenetic-related regulation factor[38,39] that is highly expressed in monocytes and macrophages upon stimulation with interferon. It can activate the DNA damage response and cause cell-cycle arrest[40]; *APOBEC3A* and *APOBEC3B* also are potent inhibitors of long terminal repeat retrotransposon function in human cells[41]. In such situations, the copy number of *APOBEC3A/3B* will cause different reactions of the innate immune system and might have some effect on the pathogenesis of SSc.

In the prediction model section, we show the prediction performance based on CNV and some other confounders quantitatively. In addition, AUC could be used to compare with other prediction models, which were established from SNP or epidemiological factors. However, we also admit that the eventual prediction model might include all the true explanatory variables such as SNP, CNV, and even some epigenetic variations in the clinical or epidemiological application in the population or hospital scenario. Moreover, the perfect prediction model should be applied 5-fold or 10-fold in cross-validation or independent dataset validation, while the sample size in our present study is limited. We would test and reevaluate our model in our future samples. We also could not detect all the significantly associated variations such as SNP, methylation, and others; thus our present study is preliminary, and we tried to identify some CNV variations associated with SSc. In the future, we could build more accurate and credible prediction models for SSc risk evaluation.

More and more aberrant CNV were found in SSc, such as *FCGR3B*[27]; therefore, genome-wide association between CNV and SSc should be studied to identify more susceptibility factors to SSc. To our knowledge, ours is the first SSc CNV map in the Chinese population, although the sample size is limited. More samples are being tested, to discover more SSc-specific or -associated CNV. In addition, we designed several probes to detect the CNV in our regions of interest. However, some CNV might be located outside the target region, which would provide underestimates for the effect size of the CNV to SSc susceptibility. Also, CNV can be limited to a single gene or include a contiguous set of genes. In the latter situation, CNV will greatly influence the human genome and so may be responsible for a substantial amount of human phenotypic variability, complex behavioral traits, and disease susceptibility. Therefore, our present study provided important evidence to identify more and more missing heritability for complex diseases based on common CNV.

Low copy of *HLA-DQA1* and high copy of *APOBEC3A/3B* regions are significantly associated with the susceptibility to SSc.

REFERENCES

1. Bunn CC, Black CM. Systemic sclerosis: an autoantibody mosaic. Clin Exp Immunol 1999;117:207-8.
2. Tamby MC, Chanseaud Y, Guillevin L, Mouthon L. New insights into the pathogenesis of systemic sclerosis. Autoimmun Rev 2003;2:152-7.
3. Silman A, Jannini S, Symmons D, Bacon P. An epidemiological study of scleroderma in the West Midlands. Br J Rheumatol 1988;27:286-90.
4. Arias-Nuñez MC, Llorca J, Vazquez-Rodriguez TR, Gomez-Acebo I, Miranda-Filloy JA, Martin J, et al. Systemic sclerosis in northwestern Spain: a 19-year epidemiologic study. Medicine 2008;87:272-80.
5. Mayes MD, Lacey JV Jr, Beebe-Dimmer J, Gillespie BW, Cooper B, Laing TJ, et al. Prevalence, incidence, survival, and disease characteristics of systemic sclerosis in a large US population. Arthritis Rheum 2003;48:2246-55.
6. Wang J, Assassi S, Guo G, Tu W, Wu W, Yang L, et al. Clinical and serological features of systemic sclerosis in a Chinese cohort. Clin Rheumatol 2013;32:617-21.
7. Arnett FC, Cho M, Chatterjee S, Aguilar MB, Reveille JD, Mayes MD. Familial occurrence frequencies and relative risks for systemic sclerosis (scleroderma) in three United States cohorts. Arthritis Rheum 2001;44:1359-62.

8. Feghali-Bostwick C, Medsger TA Jr, Wright TM. Analysis of systemic sclerosis in twins reveals low concordance for disease and high concordance for the presence of antinuclear antibodies. Arthritis Rheum 2003;48:1956-63.
9. Jüngel A, Distler JH, Gay S, Distler O. Epigenetic modifications: novel therapeutic strategies for systemic sclerosis? Expert Rev Clin Immunol 2011;7:475-80.
10. 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;21:2825-35.
11. 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 Genet 2011;7:e1002178.
12. 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 Genet 2011;7:e1002091.
13. López-Isac E, Bossini-Castillo L, Simeon CP, Egurbide MV, Alegre-Sancho JJ, Callejas JL, et al. A genome-wide association study follow-up suggests a possible role for PPARG in systemic sclerosis susceptibility. Arthritis Res Ther 2014;16:R6.
14. Sharif R, Mayes MD, Tan FK, Gorlova OY, Hummers LK, Shah AA, et al. IRF5 polymorphism predicts prognosis in patients with systemic sclerosis. Ann Rheum Dis 2012;71:1197-202.
15. Koumakis E, Wipff J, Dieude P, Ruiz B, Bouaziz M, Revillod L, et al. TGFβ receptor gene variants in systemic sclerosis-related pulmonary arterial hypertension: results from a multicentre EUSTAR study of European Caucasian patients. Ann Rheum Dis 2012;71:1900-3.
16. Bossini-Castillo L, Martin JE, Broen J, Simeon CP, Beretta L, Gorlova OY, et al. Confirmation of TNIP1 but not RHOB and PSORS1C1 as systemic sclerosis risk factors in a large independent replication study. Ann Rheum Dis 2013;72:602-7.
17. Yi L, Wang JC, Guo XJ, Gu YH, Tu WZ, Guo G, et al. STAT4 is a genetic risk factor for systemic sclerosis in a Chinese population. Int J Immunopathol Pharmacol 2013;26:473-8.
18. Guo S, Wang YL, Li Y, Jin L, Xiong M, Ji QH, et al. Significant SNPs have limited prediction ability for thyroid cancer. Cancer Med 2014;3:731-5.
19. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461:747-53.
20. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 2007;315:848-53.
21. Wang J, Yang Y, Guo S, Chen Y, Yang C, Ji H, et al. Association between copy number variations of HLA-DQA1 and ankylosing spondylitis in the Chinese Han population. Genes Immun 2013;14:500-3.
22. McKinney C, Fanciulli M, Merriman ME, Phipps-Green A, Alizadeh BZ, Koeleman BP, et al. Association of variation in Fcgamma receptor 3B gene copy number with rheumatoid arthritis in Caucasian samples. Ann Rheum Dis 2010;69:1711-6.
23. Thabet MM, Huizinga TW, Marques RB, Stoeken-Rijsbergen G, Bakker AM, Kurreeman FA, et al. Contribution of Fcgamma receptor IIIA gene 158V/F polymorphism and copy number variation to the risk of ACPA-positive rheumatoid arthritis. Ann Rheum Dis 2009;68:1775-80.
24. McKinney C, Merriman ME, Chapman PT, Gow PJ, Harrison AA, Highton J, et al. Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis. Ann Rheum Dis 2008;67:409-13.
25. García-Ortiz H, Velázquez-Cruz R, Espinosa-Rosales F, Jiménez-Morales S, Baca V, Orozco L. Association of TLR7 copy number variation with susceptibility to childhood-onset systemic lupus erythematosus in Mexican population. Ann Rheum Dis 2010;69:1861-5.
26. Wu L, Guo S, Yang D, Ma Y, Ji H, Chen Y, et al. Copy number variations of HLA-DRB5 is associated with systemic lupus erythematosus risk in Chinese Han population. Acta Biochim Biophys Sin 2014;46:155-60.
27. McKinney C, Broen JC, Vonk MC, Beretta L, Hesselstrand R, Hunzelmann N, et al. Evidence that deletion at FCGR3B is a risk factor for systemic sclerosis. Genes Immun 2012;13:458-60.
28. Du R, Lu C, Jiang Z, Li S, Ma R, An H, et al. Efficient typing of copy number variations in a segmental duplication-mediated rearrangement hotspot using multiplex competitive amplification. J Hum Genet 2012;57:545-51.
29. Lonzetti LS, Joyal F, Raynauld JP, Roussin A, Goulet JR, Rich E, et al. Updating the American College of Rheumatology preliminary classification criteria for systemic sclerosis: Addition of severe nailfold capillaroscopy abnormalities markedly increases the sensitivity for limited scleroderma. Arthritis Rheum 2001;44:735-6.
30. LeRoy EC, Black C, Fleischmajer R, Jablonska S, Krieg T, Medsger TA Jr, et al. Scleroderma (systemic sclerosis): classification, subsets and pathogenesis. J Rheumatol 1988;15:202-5.
31. Steen VD, Medsger TA Jr. Severe organ involvement in systemic sclerosis with diffuse scleroderma. Arthritis Rheum 2000; 43:2437-44.
32. Dessau RB, Pipper CB. ["R"—project for statistical computing]. [Article in Danish] Ugeskr Laeger 2008;170:328-30.
33. Kundu S, Aulchenko YS, van Duijn CM, Janssens AC. PredictABEL: an R package for the assessment of risk prediction models. Eur J Epidemiol 2011;26:261-4.
34. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12:77.
35. Lou H, Li S, Yang Y, Kang L, Zhang X, Jin W, et al. A map of copy number variations in Chinese populations. PLoS One 2011;6:e27341.
36. McLean-Tooke A, Spickett GP, Gennery AR. Immunodeficiency and autoimmunity in 22q11.2 deletion syndrome. Scand J Immunol 2007;66:1-7.
37. Lambert NC, Distler O, Müller-Ladner U, Tylee TS, Furst DE, Nelson JL. HLA-DQA1*0501 is associated with diffuse systemic sclerosis in Caucasian men. Arthritis Rheum 2000;43:2005-10.
38. Berger G, Durand S, Fargier G, Nguyen XN, Cordeil S, Bouaziz S, et al. APOBEC3A is a specific inhibitor of the early phases of HIV-1 infection in myeloid cells. PLoS Pathog 2011;7:e1002221.
39. Stenglein MD, Burns MB, Li M, Lengyel J, Harris RS. APOBEC3 proteins mediate the clearance of foreign DNA from human cells. Nat Struct Mol Biol 2010;17:222-9.
40. Landry S, Narvaiza I, Linfesty DC, Weitzman MD. APOBEC3A can activate the DNA damage response and cause cell-cycle arrest. EMBO Rep 2011;12:444-50.
41. Bogerd HP, Wiegand HL, Doehle BP, Lueders KK, Cullen BR. APOBEC3A and APOBEC3B are potent inhibitors of LTR-retrotransposon function in human cells. Nucleic Acids Res 2006;34:89-95.