# Sequencing of the MHC region defines *HLA-DQA1* as the major genetic risk for seropositive rheumatoid arthritis in Han Chinese population

Jianping Guo1,2#\*, Tao Zhang3,4#, Hongzhi Cao3,5,6#, Xiaowei Li3,4#, Hao Liang7#, Mengru Liu1, Yundong Zou1, Yuanwei Zhang 3,4, Yuxuan Wang1, Xiaolin Sun1,2, Fanlei Hu1,2, Yan Du1, Xiaodong Mo3,4, Xu Liu1, Yue Yang1, Huanjie Yang3,4, Xinyu Wu1, Xuewu Zhang1, Huijue Jia3,4, Hui Jiang3,4, Yong Hou3,4, Xin Liu3,4, Yin Su1,2, Mingrong Zhang3,4, Huanming Yang3,8, Jian Wang3,8, Liangdan Sun9, Liang Liu10, Leonid Padyukov11, Luhua Lai7, Kazuhiko Yamamoto12, Xuejun Zhang9,13\*, Lars Klareskog11\*, Xun Xu3,4\* and Zhanguo Li1,2,14,15\*

**1** Department of Rheumatology and Immunology, Peking University People's Hospital, Beijing 100044, China

**2** Beijing Key Laboratory for Rheumatism Mechanism and Immune Diagnosis (BZ0135), Beijing 100044, China

**3** BGI-Shenzhen, Shenzhen 518083, China

**4** China National GeneBank-Shenzhen, BGI-Shenzhen, Shenzhen 518120 China

**5** Shenzhen Digital Life Institute, Shenzhen 518083, China

**6** iCarbonX, Shenzhen 518083, China

**7** BNLMS, State Key Laboratory for Structural Chemistry of Unstable and Stable Species, Peking-Tsinghua Center for Life Sciences at College of Chemistry and Molecular Engineering, and Center for Quantitative Biology, Peking University, Beijing 100871, China.

**8** James D. Watson Institute of Genome Sciences, Hangzhou, 310008, China

**9** Institute of Dermatology and Department of Dermatology, the First Affiliated Hospital, Anhui Medical University, Hefei 230032, China

**10** State Key Laboratory of Quality Research in Chinese Medicine, Macau Institute For Applied Research in Medicine and Health, Macau University of Science and Technology, Macau, SAR, China

**11** Department of Medicine, Rheumatology Unit, Karolinska Institutet, Stockholm,

Sweden

**12** Laboratory for Autoimmune Diseases, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

**13** Institute of Dermatology and Department of Dermatology, Huashan Hospital, Fudan University, Shanghai, China.

**14** Peking-Tsinghua Center for Life Sciences, Beijing 100871, China

**15** State Key Laboratory of Natural and Biomimetic Drugs, School of Pharmaceutical

Sciences, Peking University, Beijing, China

**#** These authors contributed equally to this work

\*Correspondence should be addressed to Jianping Guo (jianping.guo@bjmu.edu.cn), Xuejun Zhang (ayzxj@vip.sina.com), Lars Klareskog (Lars.Klareskog@ki.se), Xun Xu (xunxu@genomics.cn) or Zhanguo Li (li99@bjmu.edu.cn).

## SUPPLEMENTARY MATERIAL & METHODS

### Quantification of ACPA

ACPA were quantified using a second generation anti-CCP (anti-cyclic citrullinated peptides) antibodies ELISA kit, with a cut-off of 5 RU/mL (Euroimmun, Luebeck, Germany).

### Variant calling for the target sequencing

SAMtools (v0.1.17) was used to converse the file from SAM to BAM, then the sorted and indexed BAM files were generated and duplicates were marked. The BAM files were analyzed using the Genome Analysis Toolkit (GATK v1.4) to detect single nucleotide variants and indels. To build a genotype matrix as input for the subsequent analysis, the genotypes for each detected variant position were extracted from all samples.

### Quality control

Sequencing data were evaluated against a quality control metrics for all the samples.

We restricted each individual as follows: (i) average sequencing depth ≥4X; (ii) 90% of the target region covered by 4X; (iii) GC content within 42%-48%. According to the criteria, a total of 58 samples from the discovery stage were filtered and removed for further analysis (**Supplementary Table 1**).

After initial sample quality control for the MHC capture sequencing, we performed further filtering to identify the high-confidence SNPs and Indels in targeted region. Following criteria were applied: i) pass ratio ≥ 0.9 (Q100 and Q500 were defined as pass for SNPs and Indels, respectively); ii) missing rate ≤ 0.1 (a depth of ≥ 5X was considered as high quality; individuals failed to meet the criteria were considered as missing); iii) minor allele frequency (MAF) ≥ 0.01; iv) Hardy-Weinberg test *P-* value ≥ 1.0 x 10–6 (**Supplementary Table 2**). For HLA types and amino acids, the same process was performed except for the pass ratio criteria (**Supplementary Table 2**).

**Statistical analyses**

To evaluate the overall significance of the multiallelic HLA amino acid polymorphisms for respective positions, the omnibus association test was conducted by using logistic regression model in R statistics program. For the multiallelic HLA amino acid polymorphisms of *m* alleles at each site, the most frequent allele in the controls was selected as the reference allele, and the *m-1* alleles were included as independent variables in the logistic regression model. The significance of the improvement of the model fit (defined as -2× the log likelihood) against the null model was calculated to assess the overall P value of the locus, which followed a chi-squared distribution with *m-1* degrees of freedom.

To assess the independent effects of the candidates identified by logistic analysis, the stepwise conditional regression analysis was performed by additionally including the most significant loci as covariates in logistic regression, If additional independent risk factors were identified, we further consecutively included them as covariates in the subsequent multivariate analyses in a forward conditional stepwise manner until none of loci met the cut off *P*-value[1](#_ENREF_1),[2](#_ENREF_2). Assuming that there may be linkage disequilibrium (LD) between the intergenic regulatory variants and specific genes, thus the independence of intergenic variants from its surrounding genes should be tested[3](#_ENREF_3). If the *P*-value of the intergenic variants is no longer less than the significant threshold after conditioning on the polymorphism sites (including SNPs, INDELs, amino acids, HLA type) of the nearby genes, we should perform condition analysis on its tagged genes, otherwise the intergenic variant is considered a real independent association locus.

The unpaired T-test was applied to assess the significance of differences in radiological scores between two groups and was conducted in R statistics program.

### Comparative modeling

For each comparative modeling, ten models were generated and the structure with the lowest probability density function total energy was selected for structural refinement. Energy minimization was performed using the Amber14 package with the Amber ff14SB force field[4](#_ENREF_4). Each structure model was solvated in an octahedron TIP3P water box and neutralized by adding proper counter ions. The distance of box boundary and structure model was set to 10 Å. The particle-mesh-Ewald (PME) method[5](#_ENREF_5) was used for the treatment of long-range electrostatic interactions. The non-bond interaction cutoff was set to 8.0 Å. Each simulation system was subjected to three stages of energy minimization, including (1) 5000 steps of steepest descent (SD) and 2000 steps of conjugate gradient (CG) minimization with harmonic restraints (10 kcal/Å) applied on all structural atoms; (2) 5000 steps of SD minimization and 2000 steps of CG minimization with reduced harmonic restraints (2 kcal/Å) on backbone atoms; (3) 10,000 steps of steepest descent and 5000 steps of conjugate gradient minimization with all restraints removed.

## SUPPLEMENTARY RESULTS

**Exclusive dissection of *HLA-DRB1* indicates that DRβ1 variants could be strong risks for ACPA-positive RA, if *HLA-DQA1* is not included.**

Since *HLA-DRB1* was recognized as the strongest RA risk in previous studies, especially for ACPA-positive RA. The variants *DRB1\*0405*, DRβ1:11, 13, 57, 74, and 71 have been shown to confer risks for RA in Asian patients[6](#_ENREF_6). We next investigated RA association at *HLA-DRB1* separately. As shown in Supplementary Table 9, multiple alleles at *HLA-DRB1* showed strong associations, with DRβ1:120N (*P* = 6.46 x 10−28, OR=2.27, 95% CI 1.96-2.63), *DRB1\*0405* (*P* = 6.55 x 10−28, OR=3.40, 95% CI 2.73-4.23), and DRβ1:11V (*P* = 2.43 x 10−27, OR=2.45, 95% CI 1.94-2.60) being the top three risks. When conditioning on either DRβ1:11V or :120N, the second signal was seen for DRβ1:31I (*P* = 2.23 x 10-18, OR=1.93, 95% CI 1.67-2.24) and DRβ1:13F (*P* = 2.90 x 10−17, OR=1.82, 95% CI 1.58-2.09) (Supplementary Fig2). After further conditioning to either DRβ1:13F or :31I, strong association signals were observed for *DRB1\*04:05* (*P* = 6.26 x 10−11, OR=2.53, 95% CI 1.92-3.35) and DRβ1:57S (*P* = 3.14 x 10−9, OR=1.81, 95% CI 1.49-2.20), respectively. Then we continued conditioning on DRβ1:57S, DRβ1:74A showed a suggestive association with ACPA-positive RA (*P* = 5.09 x 10-7, OR = 0.72, 95% CI 0.63-0.82). When conditioning on DRβ1:74A, DRβ1:71E also showed a suggestive association (*P* = 2.95 x 10−6, OR=0.37, 95% CI 0.24-0.56). Our results indicated that if the effect of *DQA1* is ignored, *DRB1\*0405*, amino acid variants at DRβ1:11, 13, 57, 74, and 71 can come up and be very strong risk factors for ACPA-positive RA.

***DQA1\*0303*, an allele encoding DQα1:160D, confers increased risk of joint damage in early ACPA-positive RA**

DQα1:160D is encoded by two alleles, i.e. *DQA1\*0302* and *\*0303*. We next examined whether the DQα1:160D and its encoding alleles confer a risk for the severity of joint damage in ACPA-positive RA. A total of 557 patients with available SHS data were divided into three groups according to the disease durations (≤ 1 years, 1–10 years, or ≥ 10 years). As smoking is a well-established environmental factor contributing to ACPA-positive RA, we further stratified the patients by smoking status. As shown in Supplementary Fig. 3, in the early disease stage one of its coding allele *DQA1\*0303* displayed high impact on radiographic scores in smoking group (*P* = 3.02 x 10−5, SupplementaryFig. 3A). Similarly, in early disease stage *DQA1\*0303* carrierswith smoking had higher radiographic scores than *DQA1\*0303* carrierswithout smoking (*P* = 4.05 x 10−8). In the early disease stage *DRB1\*0405* alsoshowed a higher impact on radiographic score in smoking group (*P* = 3.02 x 10−5). *DRB1\*0405* carrierswith smoking had increased radiographic scores, compared to *DRB1\*0405* carrierswithout smoking (*P* = 6.96 x 10−6, SupplementaryFig. 3B).

## SUPPLEMENTARY REFERENCES

1. Sud A & Thomsen H. Genome-wide association study of classical Hodgkin lymphoma identifies key regulators of disease susceptibility. *Nat Commun* 2017;8:1892-903.

2. Zhou F, Cao H, Zuo X *et al.* Deep sequencing of the MHC region in the Chinese population contributes to studies of complex disease. *Nat Genet* 2016;48:740.

3. Patsopoulos NA, Barcellos LF, Hintzen RQ *et al.* Fine-Mapping the Genetic Association of the Major Histocompatibility Complex in Multiple Sclerosis: HLA and Non-HLA Effects. *PLoS Genet* 2013;9.

4. Maier JA, Martinez C, Kasavajhala K *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* 2015;11:3696-713.

5. Petersen HG. Accuracy and efficiency of the particle mesh Ewald method. *The Journal of Chemical Physics* 1995;103:3668-79.

6. Okada Y, Kim K, Han B *et al.* Risk for ACPA-positive rheumatoid arthritis is driven by shared HLA amino acid polymorphisms in Asian and European populations. *HUM MOL GENET* 2014;23:6916-26.

## SUPPLEMENTARY FIGURES

### Supplementary Figure 1

**Plots of stepwise conditional analysis to assess whether DQα1:160D and DRβ1:37N were independent each other.** Each diamond represents -log10(P) of the variants, including the SNPs, the classical HLA alleles and the amino acid polymorphisms of the HLA genes. The horizontal axis represents position information (unit: Mb) and vertical axis represents –log10(P). The dotted horizontal line represents the suggestive significance threshold of *P* = 1 x 10-6. (**a**) The major genetic determinants of ACPA–positive RA mapped to the HLA-DQα1 corresponding Asp-160. The blue diamond represents HLA-DRβ1 Asn-37. (**b**) Subsequent conditional analyses controlling for HLA-DRβ1 Asn-37 revealed an independent association at HLA-DQα1 corresponding Asp-160. (**c**) Upon controlling for HLA-DRβ1 Asn-37 and HLA-DQα1 Asp-160, no additional significant association signal was observed.

### Supplementary Figure 1



### Supplementary Figure 2

**Plots of stepwise conditional analysis for HLA-DRβ1 in combined cohort** (**a**) Amino acid position 120 displayed the strongest association with ACPA–positive RA (*P* < 10−27), followed by position 11 (*P* < 10−26). (**b**) Controlling for position 11 or 120, position 13 showed an independent risk for ACPA–positive RA (*P* = 2.90 × 10−17). (**c**) Controlling for positions 11 and 13, position 57 became an independent association (*P* = 3.14 × 10−9). (**d**) Controlling for positions 11, 13 and 57, position 74 showed a suggestive signal (*P* = 5.09 ×10-7) (**e**) Conditioning on positions 11,13, 57 and 74, a suggestive association was appeared for amino acid 71 (*P* = 2.95 ×10-6).

****

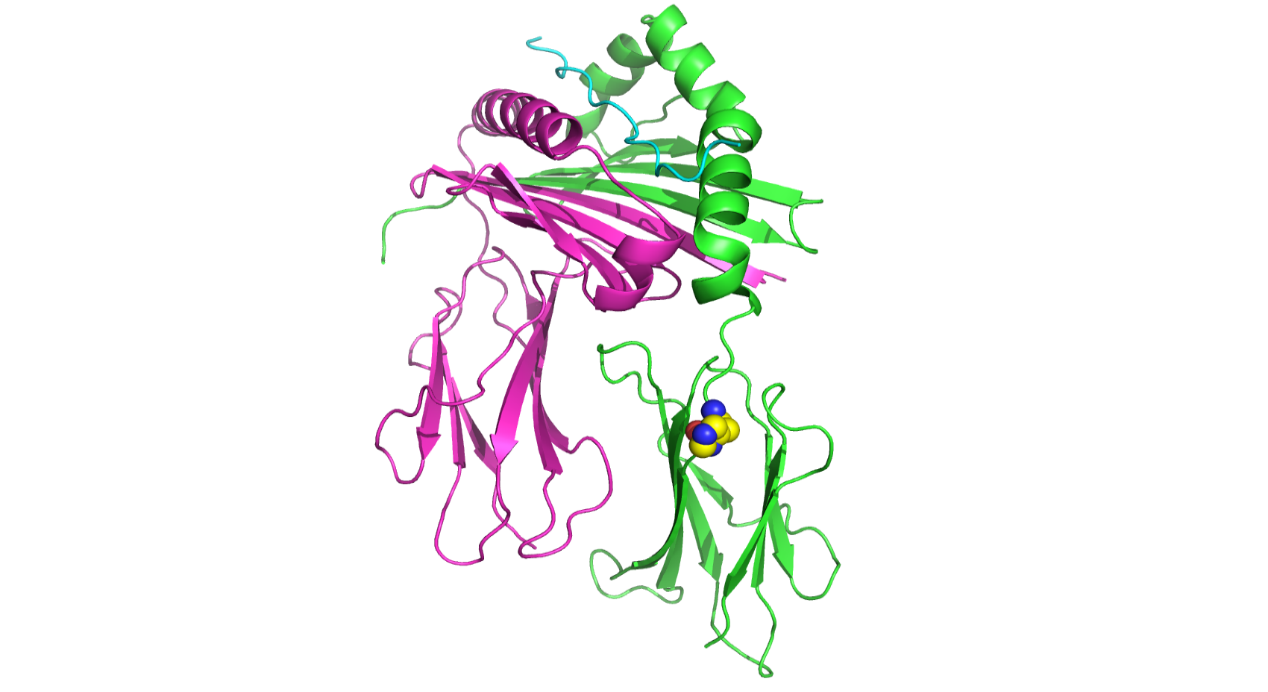
### Supplementary Figure 3

**Impact of *DQA1\*0303* on risk of joint damage in ACPA-positive RA**.(a) In early disease stage *DQA1\*0303* showed high impact on radiographic scores in smoking group (*P* = 3.02 x 10−5). Similarly, *DQA1\*0303* carriers with smoking had higher radiographic scores than DQA1\*0303 carriers without smoking (*P* = 4.05 x 10−8). (b) In early disease stage *DRB1\*0405* also showed a high impact on radiographic score in smoking group (*P* = 3.02 x 10−5). *DRB1\*0405* carriers with smoking had increased radiographic scores, compared to *DRB1\*0405* carriers without smoking (*P* = 6.96 x 10−6).

****

**Supplementary Figure 4**

**Crystal structure of HLA-DR11 (PDB code: 6CPL).** H96 displayed as colored spheres



## SUPPLEMENTARY TABLES

### Supplementary Table 1 Sample quality control in discovery cohort

|  |  |  |
| --- | --- | --- |
| **Filters** | **Criteria** | **Removed (n)** |
| Mean sequencing depth | Fold coverage < 4X | 0 |
| Coverage | 4X coverage < 90% | 58 |
| GC content | < 42%, or > 48% | 0 |
| All above filters | Any one | 58 |

Supplementary Table 2Quality control of variants, HLA type and amino acids in the study cohort

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Filters** | **Criteria** | **Variants remained** | | **HLA types**  **remained** | | **Amino acids remained** | |
| **1#** | **2#** | **1#** | **2#** | **1#** | **2#** |
| Pass ratio | Pass ratio < 0.9 | 86,971 | - | 384 | 220 | 7,289 | 1,712 |
| MAF | MAF < 0.01 | 26,728 | - | 170 | 89 | 1,373 | 653 |
| HWE | HWE < 10-6 | 24,450 | - | 166 | 89 | 1,283 | 653 |
| Missing rate | Missing rate > 0.1 | 24,177 | - | 166 | 89 | 1,283 | 653 |
| All above filters | Any one | 24,177 | - | 166 | 89 | 1,283 | 653 |

1#: the discovery stage

2#: the validation stage

MAF: Minor allele frequency

HWE: Hardy-Weinberg Equilibrium