Supplementary Online Content

Viatte S, Plant D, Han B, et al. Association of HLA-DRB1 haplotypes with rheumatoid arthritis severity, mortality, and treatment response. *JAMA*. doi:10.1001/jama.2015.3435.

This supplementary material has been provided by the authors to give readers additional information about their work.

eMethods

Cohorts and patients:

NOAR: Patients recruited between 1990-1994 remain under long-term follow-up (up to 20 years) except for (1) patients who do not cumulatively satisfy the ACR criteria for RA at 5 years and have been given a consultant diagnosis other than RA, undifferentiated inflammatory polyarthritis, psoriatic arthritis or post-viral arthritis to explain their symptoms; (2) patients whose disease has gone into spontaneous long-term remission (no inflamed joints at the 3rd or 5th anniversary and not on DMARDs or steroids). Patients who satisfied the 1987 ACR criteria applied cumulatively over the entire duration of the follow-up were classified as RA patients. All patients, including RA patients, were called inflammatory polyarthritis patients. Radiographic schedule: radiographs of hands and feet were performed at baseline, years 1 and 2 if required to establish the diagnosis of RA, systematically at year 5 for all patients, and at year 10 for those patients who were erosive at year 5. The definition of "early RA" varies [1], but a disease duration of < 2 years is commonly accepted as early disease. 90% NOAR patients have disease duration < 2 years.

ERAS: X-rays of hand and feet were taken at baseline, 6 months, one year and yearly thereafter, for 15 years. All radiographs were assessed for the presence of erosions, but only radiographs at baseline and yearly thereafter, to year 9, were scored using the Larsen technique.

Genotyping:

HLA genotyping was performed using a semiautomated reverse dot-blot method [7]. Amino acids at position 11, 71 and 74 of HLA-DRB1 were assigned according to the sequence information provided in the IMGT/HLA Database at the European Bioinformatic Institute http://www.ebi.ac.uk/ipd/imgt/hla/, using the Sequence Alignment Tool, Release 3.13.0 [8]. Where only 2-digit information was available from genotyping and several different amino acids were theoretically possible, decisions have been made based on the allele frequency in the UK, as derived from the HLA Allele Frequency Database: http://www.allelefrequencies.net/[9]. In unclear situations, the amino acid was set to missing.

Assigning amino acids at positions 11, 71, 74 based on 2-digit HLA-DRB1 typing:

1) Step one, exhaustive list: eTable 1

2) Step two: the frequency of every allele in eTable 1 was determined in Caucasian populations available at the HLA Allele Frequency Database. All alternative alleles, when found at all in the database, had a frequency of < 1 % for all populations, apart from 01:03 (2-3% in Caucasians), 08:05 (1% in Scotland Orkney, 1,8% Spain Catalonia Girona). 08:05 is a rare occurrence in NOAR (only one individual in the whole dataset), so that we can assume that the 08 are very unlikely to be 08:05. It was therefore decided to assign an “L” at position 74 for patients with a DRB1 *08 typing. Since 01:03 is not so rare, and there are several other alleles with non-R at 71 (for example, 01:09 is present in NOAR), position 71 and SE status are set to missing for 01. aa: amino acid. Single-letter amino acid code: V, Valine; L, Leucine; D, Aspartic Acid; P, Proline; G, Glycine; S, Serine; K, Lysine; R, Arginine; A, Alanine; E, Glutamic Acid; Q, Glutamine.

Therefore, the following table can be produced to assign aa at positions 11, 71 and 74 from the 2-digit typing: eTable 2

Susceptibility study:

We recalculated for the present study the univariable ORs of individual amino acids at positions 11, 71 and 74 of HLA-DRB1 for their association with ACPA-positive rheumatoid arthritis (susceptibility ORs), using 7,279 ACPA-positive RA cases and 15,870 controls from the ImmunoChip study[11] and presented the recalculated susceptibility ORs in eTable 4. The samples from the ImmunoChip study were more densely genotyped (>7,000 SNPs within the MHC) compared to the platforms used in Raychaudhuri et al.[10] (<1,000 SNPs within the MHC), which provided better imputation accuracy[12]. We performed a univariable logistic regression for every amino acid at position 11, 71 and 74 on 9,585 cases and 33,742 controls from 6 different populations and corrected for population stratification by adjusting for the 10 first principal components.

Statistical analysis:

Longitudinal modelling in NOAR and ERAS: All time points with available x-ray data were incorporated in the analysis for NOAR, while we censored the data in ERAS at year < 7 for the analysis of erosions and at year < 9 for the Larsen score, similarly to analyses previously reported in ERAS[2,3]. Indeed, we have shown previously in ERAS (Viatte S et al. Investigation of rheumatoid arthritis genetic susceptibility markers in the early rheumatoid arthritis study further replicates the TRAF1 association with radiological damage. J Rheumatol 2013;40:144-56) that a) radiographic damage at different time points are highly correlated and that b) the variance in radiographic damage increases during the first years of follow-up, to decrease thereafter (for example, 90% of ERAS patients are erosive at year 13). As a consequence of those 2 points, longitudinal modelling needs to be censored at a time point dependent on the outcome used. We performed sensitivity analysis to determine the cut-off for censoring: at year < 7 for the analysis of erosions and at year < 9 for the Larsen score, similarly to analyses previously reported in ERAS.

Correlation between effect sizes for severity and susceptibility or treatment response and susceptibility was calculated using linear regression of beta coefficients from GEE models or (ordinal) logistic regressions (i.e. natural logarithms of ORs), hence the logarithmic scale of the plots, when axes are labelled with ORs.

Model construction:

For *univariable (predictor) analysis of amino acids (aa)*, adjustment with non-genetic covariates were performed as described above, but no adjustment was performed for any other aa at the same or at other positions. An effect size with confidence interval (CI) and p-value is obtained for every aa. For the *univariable (predictor) analysis of a position* (for example HLA-DRB1 position 11), all numerical variables for aa at this position were incorporated in the model, but none for the other positions. This corresponds therefore to a multivariable analysis with respect to the aa at the position considered, but to a univariable predictor analysis with respect to the position alone. Since a position is multi-allelic, no effect size can be obtained for a position, only a p-value. For the *multivariable (predictor) analysis of positions,* all numerical variables for all aa at all positions were incorporated in the same model and a p-value for the position was obtained. For *multivariable (predictor) analyses of haplotypes*, all 16 haplotypes, constructed from the 3 independent HLA-DRB1 positions 11, 71 and 74, are incorporated in the same model. An effect size for the independent effect of every haplotype is obtained, and a p-value for the entire model (the 16 haplotype model). The SE corresponds therefore to a two haplotype groups model.

Generalized Linear Latent and Mixed Model (GLLAMM):

GLLAMMs are a general class of multilevel latent variables models for multivariable response of mixed type including continuous, count, dichotomous and categorical responses[4-6]. We developed a linear hierarchical regression model for the Larsen score by fitting a GLLAMM with discrete random effects and three latent classes, which corresponds to a three-component mixture of regression models with common parameters for each covariate (age, disease duration, shared epitope or HLA-DRB1 markers), but different parameters for the latent trajectories across latent classes. This three-component mixture model has the advantage of capturing both low scores (by one component) and high skewness of Larsen score distribution (by the other two components). We fit the three-component mixture of regression for studying the association between the longitudinal Larsen scores outcomes and a genotype variable by including time-invariant covariates (age at onset and its quadratic term) and time-varying covariates (polynomial functions of time including an intercept, linear term and quadratic term of years since the disease onset) as follows:

$$LARSEN_{i,t} = \beta_1 GENOTYPE_i + \beta_2 DURATION_{i,t} + \beta_3 DURATION_{i,t}^2 + \beta_4 AGE_i + \beta_5 AGE_i^2 + \sum_{k=1}^{3} I(C_i = k)[\gamma_{1,k} + \gamma_{2,k} DURATION_{i,t} + \gamma_{2,k} DURATION_{i,t}^2 + \varepsilon_{i,t}]$$

where each subject i belongs to a "latent class k" (if $C_i = k$) in terms of Larsen score trajectory pattern over time (modelled by class-specific effects γ_k of disease duration and its quadratic term). This three-component latent class model has the advantage to capture common characteristics of Larsen trajectories within a subpopulation through latent classes and hence to capture extra heterogeneity in the longitudinal data to improve the model fitting. Furthermore, the three-component mixture distribution (of the error term ε) offers a good fit for non-normal distributed outcome such as Larsen score with extra low scores and high skewness. By comparison, an alternative approach, zero-inflated negative binomial model, is a two-component mixture of regression model, which is useful to capture low scores but its negative binomial part may not adequately address the high skewness in Larsen score data. To determine the number of latent classes, we have checked the model-fit statistics such as BIC. The best-fitting model (three-class) was then identified, leaning towards to parsimony in the number of model parameters.

/*** STATA code ****/

```
gen constant=1

eq fac1:constant

eq fac2:disease_duration

eq fac3:disease_duration2

gllamm LARSEN_SCORE GENOTYPE DISEASE_DURATION DISEASE_DURATION2 AGE AGE2, i(patient_id) nrf(3) eqs(fac1 fac2 fac3) ip(f) nip(3)
```

eResults

The role of anticyclic citrullinated peptide (anti-CCP) antibodies: only a partial mediator of radiographic outcome:

In statistics, mediation analysis consists of evaluating if a third variable (variable C or mediator variable) can be considered as a path variable in the association between a dependent variable A and an independent variable B. When anti-CCP antibody status (variable C) was considered as a time dependent variable (measured at baseline and year 5) in NOAR, the 16 haplotype (variable B) model showed a very strong association with anti-CCP positivity (p= 8.17E-58) and the observed hierarchy for susceptibility or severity was almost identical. Anti-CCP status itself (ever positivity) was strongly associated with the presence of erosive disease (variable A) (OR: 6.75, 95% CI 5.47;8.33, p=1.52E-71). However, when testing the association of the 16 haplotype model with Larsen score, controlling for anti-CCP did not lead to a complete abrogation of the signal, which remains significantly associated with radiographic damage (p=2.52E-02). The 16 HLA-DRB1 haplotype model, therefore, had a significant direct association (statistical nomenclature for mediation analysis: “direct effect”) on radiographic outcome, independently of the anti-CCP status. Anti-CCP status was therefore only a partial mediator of the association of the 16 haplotypes with radiographic damage.

Analysis of goodness of fit:

We show the incremental explanatory power of the 16 haplotype model over the SE (shared epitope) model by two ways:

1) Comparison of the log likelihood between the two models; the log likelihood for GLLAMM (Larsen score) in RA is -8483 for the 16 haplotype model, which is greater than that of the SE: -9016 (the higher the log likelihood, the better the model fit). We also apply a BIC-type penalized log-likelihood for model comparison (the lower the BIC, the better the model fit):

BIC= -2 Log-likelihood + ln(N) with N= number of parameters
BIC for the 16 haplotype model = 17148
BIC for the SE model = 18148

By comparing BIC, we can show that the 16 haplotype model represents an additional improvement over the SE model, even with a penalty for the increase in predictors included in the model.

Since GEE is not a likelihood-based approach, we used a mixed-effect logistic regression with a random intercept for the binary erosion outcome that is similar to the GEE model we fitted. The goodness of fit is also better for modelling the presence of erosion in RA with the 16 haplotype model (log likelihood= -1211; BIC = 2576) than for the SE model (log likelihood= -1307; BIC = 2660).

To check the significance of the incremental explanatory power, we apply a likelihood ratio test to compare the 16 haplotype model against the first-3-haplotype model (VKA, VRA, LRA, which define the SE) to test whether the effects of the other 13 haplotypes are zero.

For the Larsen score model: LR=2*(8495-8483)=23.2 **p-value=0.026**
For the erosions model: LR=2*(1223-1211)=23 **p-value=0.028**

We show therefore a significant overall incremental explanatory power of the other 13-haplotypes over the first three (VKA, VRA, LRA), which fully determine the SE.

2) Beside the first three haplotypes (VKA, VRA, LRA) which fully determine the SE, we obtain some significant p-values in both GEE and GLLAMM models for the effects of the other haplotypes. This also shows that the inclusion of additional markers in the model will increase the explanatory power. As a further example, in SE-negative patients, the VRE-haplotype (corresponding to HLA*04:03 and HLA*04:07, which are associated with increased RA susceptibility), is significantly associated with erosions (OR: 2.08, 95% CI 1.28;3.39, p=3.18E-03).

eTable 1. Exhaustive list of possible amino acids at DRB1 positions 11, 71 or 74 for a given 2-digit HLA-type, based on the IMGT/HLA Database at the EBI, Release 3.13.0

HLA-typing: 2-digit	Position 11 (DRB1)	Position 71 (DRB1)	Position 74 (DRB1)
01	L	R for all, but E for **01:03** and 01:42, A for 01:06, 01:09, 01:15, K for 01:10, 01:16	A for all, but R for 01:16 and 01:55, E for 01:17, V for 01:59
03	S for all alleles, but R for 03:42, A for 03:69, C for 03:86, L for 03:87	K, but R for 03:91	R, but G for 03:84, A for 03:76, Q for 03:11 and for 03:17 and for 03:24, 03:27, 03:35, 03:81
04	V for 161 alleles, but I for 04:49, A for 04:66	No overwhelming majority of alleles with a specific aa	No overwhelming majority of alleles with a specific aa
07	G	R	Q
08	S	R, but K for 08:40	L, but E for 08:29, A for **08:05**, 08:18, 08:24, 08:25, 08:31, 08:40, 08:41, 08:47
09	D	R	E
11	S for 147 alleles, but V for 11:22, L for 11:30, F for 11:104, C for 11:144	No overwhelming majority of alleles with a specific aa	A, but R for 11:07, 11:103, 11:105, 11:107, 11:125, E for 11:17, 11:52, 11:54, 11:89, L for 11:23, 11:25 and 11:45, 11:55, 11:64, 11:67, 11:119, V for 11:32, 11:70, Q for 11:53, 11:136
12	S	R	A
13	S, but L for 13:67	No overwhelming majority of alleles with a specific aa	A, but L for 13:13, 13:18, 13:47, 13:55, 13:119, 13:144, 13:146, 13:154, 13:156, 13:158, 13:164, E for 13:76, R for 13:113N
14	S, but V for 14:10, 14:57, P for 14:39, R for 14:46, L for 14:141	R, but E for 14:16, 14:57, K for 14:19, 14:21, 14:76, 14:79, 14:107, 14:109, 14:111, 14:137N,	No overwhelming majority of alleles with a specific aa

		14:141, A for 14:24, 14:37	
15	P	A, but E for 15:10, R for 15:17N, 15:21, 15:27, 15:34, 15:54, 15:66, K for 15:25, 15:84	A, but P for 15:17N, 15:91, L for 15:21 , R for 15:25, G for 15:50N, 15:80N
16	P	R	A, but L for 16:04, 16:18

eTable 2. Conversion table from 2-digit HLA-DRB1-typing to amino acids at positions 11, 71 and 74

HLA-typing: 2-digit	Position 11 (DRB1)	Position 71 (DRB1)	Position 74 (DRB1)
01	L	Missing	A
03	S	K	R
04	V	Missing	Missing
07	G	R	Q
08	S	R	L
09	D	R	E
11	S	Missing	A
12	S	R	A
13	S	Missing	A
14	S	R	Missing
15	P	A	A
16	P	R	A

Currently, DRB1*02 does not exist anymore, but has been split into *15 and *16. Similarly, *05 is now either *11 or *12. Single-letter amino acid code: V, Valine; L, Leucine; D, Aspartic Acid; P, Proline; G, Glycine; S, Serine; K, Lysine; R, Arginine; A, Alanine; E, Glutamic Acid; Q, Glutamine.

eTable 3. Concordance between the reverse dot-blot method and the imputation from the ImmunoChip in assigning HLA-DRB1 types or amino acids at positions 11, 71 or 74

HLA-DRB1	Concordance (%)	
	NOAR	BRAGGSS
4-digit typing	91.8	93.3
2-digit typing	96.5	98.8
Position 11	97.6	99.2
Position 71	95.7	96.9
Position 74	96.8	98.6

The percentages presented are based on 680 samples genotypes with both techniques in NOAR and 816 in BRAGGSS.

eTable 4. Unadjusted ORs of amino acids for their association with ACPA-positive rheumatoid arthritis

Gene	Position	Residue	OR	Lower 95%CI	Upper 95%CI
HLA-DRB1	11	Val	3.59	3.42	3.78
HLA-DRB1	11	Leu	1.34	1.26	1.42
HLA-DRB1	11	Asp	1.18	1.00	1.40
HLA-DRB1	11	Pro	0.65	0.61	0.69
HLA-DRB1	11	Gly	0.52	0.48	0.56
HLA-DRB1	11	Ser	0.39	0.37	0.41
HLA-DRB1	71	Lys	1.94	1.85	2.04
HLA-DRB1	71	Arg	0.99	0.95	1.03
HLA-DRB1	71	Ala	0.62	0.58	0.67
HLA-DRB1	71	Glu	0.35	0.33	0.39
HLA-DRB1	74	Ala	1.98	1.89	2.09
HLA-DRB1	74	Glu	0.75	0.68	0.83
HLA-DRB1	74	Arg	0.59	0.55	0.64
HLA-DRB1	74	Gln	0.52	0.48	0.56
HLA-DRB1	74	Leu	0.50	0.43	0.59

Figure 3 of Raychaudhuri S et al. [10]presents the univariable effect sizes of individual amino acids at positions 11, 71 and 74 of HLA-DRB1 on ACPA-positive RA susceptibility. While Raychaudhuri S et al. used 5,018 cases and 14,978 controls, we recalculated for the present study the univariable effect sizes of individual amino acids at positions 11, 71 and 74 of HLA-DRB1, using 7,279 ACPA-positive RA cases and 15,870 controls from the ImmunoChip study[11], as described in eMethods. We performed the univariable logistic regression on 9,585 cases and 33,742 controls. "Univariable" or "unadjusted" refers to the fact that ORs of amino acids have not been adjusted neither for amino acids at the same position, nor for amino acids at other positions. The recalculated ORs presented in eTable 4 were highly concordant with the ones reported in

Raychaudhuri et al.[10]. For each amino acid shown, the comparison group is all other amino acids at this position listed in the table. OR: odds ratio. 95%CI: 95% confidence interval.

eTable 5. Univariable analysis of amino acids at position 11 in NOAR, adjusted for the shared epitope

		Adjusted for the shared epitope							
		Inflammatory polyarthritis				Rheumatoid arthritis			
Pos.	Aa	OR erosions, 95% CI	p	Change in Larsen score, 95% CI	p	OR erosions, 95% CI	p	Change in Larsen score, 95% CI	p
11	Val	1.32 (1.09;1.62)	5.0E-03*	0.94 (0.10;1.78)	0.03*	1.30 (1.04;1.62)	0.02	0.99 (-0.01;1.98)	0.05
11	Leu	0.75 (0.61;0.93)	8.8E-03*	-1.29 (-2.23;-0.34)	7.8E-03*	0.77 (0.61;0.97)	0.03	-1.32 (-2.45;-0.19)	0.02*
11	Asp	1.48 (0.89;2.48)	0.14	2.67 (0.55;4.79)	0.01*	1.65 (0.93;2.94)	0.09	3.14 (0.68;5.61)	0.01*
11	Pro	1.04 (0.84;1.31)	0.7	0.02 (-0.97;1.01)	0.97	1.03 (0.81;1.31)	0.82	0.14 (-1.01;1.29)	0.81
11	Gly	1.12 (0.90;1.40)	0.32	1.04 (0.13;1.95)	0.02*	1.12 (0.88;1.42)	0.38	1.14 (0.08;2.21)	0.04*
11	Ser	0.84 (0.71;1.00)	0.05	-1.09 (-1.85;-0.33)	4.9E-03*	0.84 (0.70;1.02)	0.08	-1.18 (-2.07;-0.28)	9.7E-03*

Associations with the presence of erosions were tested with GEE models (generalised estimating equation model) and the effect size expressed as an OR with its 95% confidence interval (CI). Associations with the Larsen score were tested with GLLAMM (Generalized Latent and Linear Mixed Model) and the effect size is expressed as an increase of Larsen score with its 95% (CI). Many association signals with radiological outcome remains significant at a nominal level after adjustment for the SE. * remains significant with a false discovery rate (FDR) of 0.05 after adjustment with the Benjamini-Hochberg method (1 aa is collinear with the others, therefore the number of independent tests is 5). For each amino acid shown, the comparison group is all other amino acids at this position listed in the table.

eTable 6. Univariable analysis of HLA-DRB1 positions in NOAR

	Inflammatory polyarthritis		Rheumatoid arthritis	
	p-value erosions	**p-value Larsen score**	**p-value erosions**	**p-value Larsen score**
Position 11	2.40E-15	3.80E-08	7.76E-14	2.19E-07
Position 71	2.01E-04	2.09E-03	2.86E-05	2.52E-03
Position 74	6.83E-03	1.68E-02	1.71E-02	8.88E-03
Adjusted for the SE:				
Position 11	0.03	6.70E-05	0.06	2.74E-04

Associations with the presence of erosions were tested with GEE models (generalised estimating equation model). Associations with the Larsen score were tested with GLLAMM (Generalized Latent and Linear Mixed Model).Cumulative association of all aa at one position was tested, without adjustment for the other positions. The p-value is the model p-value and represents a measure of association for all amino acids considered together at one position (strictly, this is a multivariable analysis with regards to the amino acids at one postion, but a univariable analysis with regards to the position). The association of position 11 of HLA-DRB1 is independent of the shared epitope (SE). The second strongest signal is position 71. Position 71 remained significantly associated with Larsen score in inflammatory polyarthritis after adjustment for the SE (p=4.87E-02), which shows that the haplotype group represented by the SE does not capture the entire risk conferred by aa at position 71, though this position is part of, and used to define the SE. Apart from the only p-value > 0.05, all tests remain significant with a false discovery rate (FDR) of 0.05 after adjustment with the Benjamini-Hochberg method (3 independent tests).

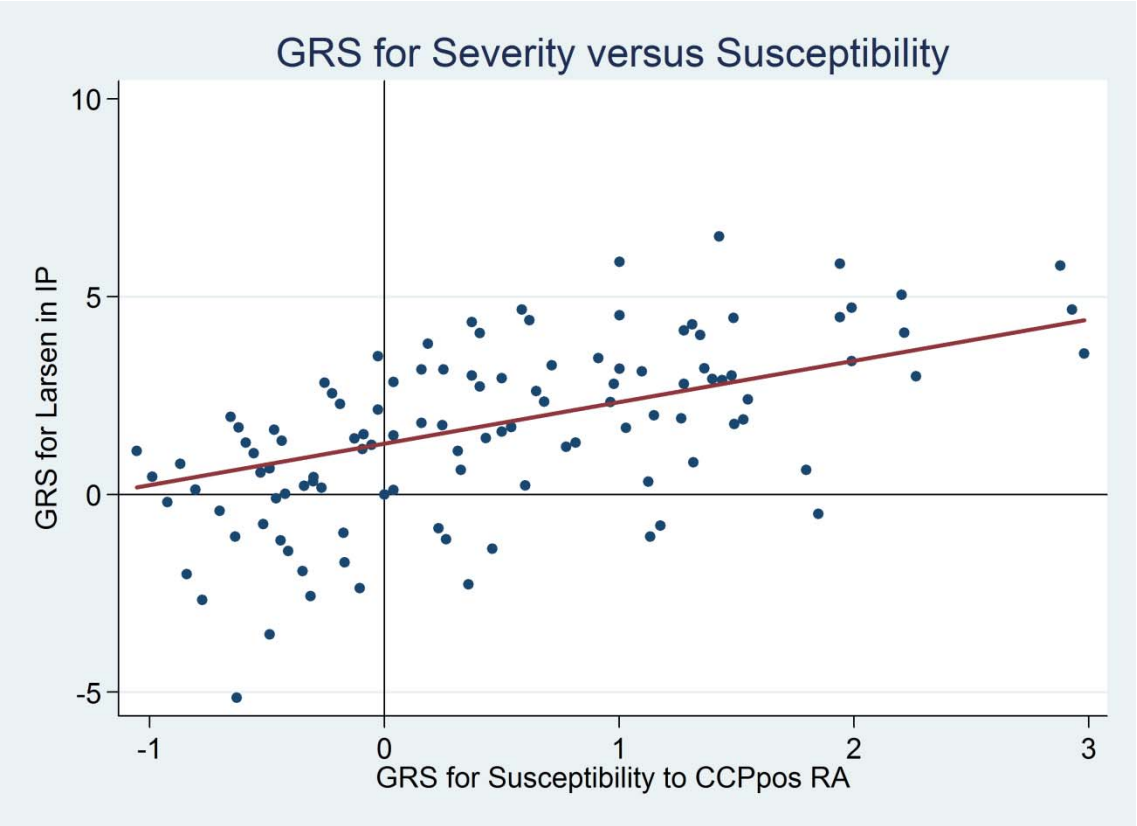
eTable 7. Multivariable analysis of haplotypes (the 16 haplotype model): presentation of the 16 haplotypes defined by HLA-DRB1 positions 11, 71 and 74 and their association with radiographic outcome in inflammatory polyarthritis (IP) in NOAR

Haplotype name	aa11	aa71	aa74	N of heterozygote RA patients (total: 2534)	N of homozygote RA patients (total: 2534)	Haplotype freq. in IP (%)	GEE (OR for erosions, CI), N=1528	GEE p-val.	GLLAMM (Change in Larsen score, CI), N=1528	GLLAMM p-val.	Haplotype group p-value
VKA-haplotype	**Val**	**Lys**	**Ala**	**690**	**63**	**16.1**	**1.79 (1.35;2.36)**	**3.96E-05**	**1.78 (0.62;2.94)**	**2.70E-03**	**GEE: 2.83E-12 GLLAMM: 3.73E-06**
VRA-haplotype	**Val**	**Arg**	**Ala**	**406**	**18**	**8.7**	**1.88 (1.38;2.59)**	**7.19E-05**	**2.89 (1.52;4.26)**	**3.63E-05**	
LRA-haplotype	**Leu**	**Arg**	**Ala**	**630**	**43**	**14.1**	**1.43 (1.05;1.93)**	**0.02**	**1.20 (-0.10;2.49)**	**0.07**	
PRA-haplotype	Pro	Arg	Ala	43	1	0.9	1.16 (0.64;2.12)	0.61	3.26 (0.22;6.30)	0.04	
VRE-haplotype	Val	Arg	Glu	109	6	2.4	1.62 (1.01;2.61)	0.05	1.59 (-0.37;3.55)	0.11	
DRE-haplotype	Asp	Arg	Glu	84	2	1.7	1.22 (0.70;2.14)	0.49	2.94 (0.51;5.36)	0.02	
VEA-haplotype	Val	Glu	Ala	17	0	0.3	0.24 (0.05;1.07)	0.06	-2.27 (-6.29;1.76)	0.27	
SKA-haplotype	Ser	Lys	Ala	32	0	0.6	0.86 (0.35;2.12)	0.75	0.11 (-3.01;3.24)	0.94	
PAA-haplotype	**Pro**	**Ala**	**Ala**	**466**	**43**	**10.9**	**1.00 (reference)**	**-**	**0 (reference)**	**-**	
GRQ-haplotype	**Gly**	**Arg**	**Gln**	**528**	**41**	**12.0**	**1.07 (0.79;1.46)**	**0.63**	**1.14 (-0.12;2.39)**	**0.08**	
SRA-haplotype	**Ser**	**Arg**	**Ala**	**259**	**11**	**5.5**	**1.05 (0.72;1.54)**	**0.81**	**1.41 (-0.12;2.94)**	**0.07**	
SRE-haplotype	Ser	Arg	Glu	91	2	1.9	0.44 (0.25;0.77)	4.24E-03	-0.97 (-3.00;1.06)	0.35	
LEA-haplotype	Leu	Glu	Ala	47	1	1.0	0.63 (0.31;1.31)	0.22	-2.57 (-5.95;0.81)	0.14	
SRL-haplotype	Ser	Arg	Leu	118	0	2.3	1.14 (0.65;1.99)	0.65	0.22 (-1.71;2.15)	0.82	
SKR-haplotype	**Ser**	**Lys**	**Arg**	**592**	**50**	**13.7**	**0.90 (0.67;1.22)**	**0.51**	**-0.10 (-1.32;1.12)**	**0.87**	
SEA-haplotype	**Ser**	**Glu**	**Ala**	**348**	**23**	**7.8**	**0.86 (0.61;1.21)**	**0.40**	**0.55 (-0.81;1.91)**	**0.43**	

Haplotype frequency was assessed in the entire NOAR cohort with available haplotype (2534 patients), including individuals without available radiological assessment. The association study with erosions or Larsen score was performed on samples with available haplotype and radiological data (1528 RA patients). This number is smaller than the 1691 inflammatory polyarthritis patients described in Table 1, because some of them have missing genotypes at one or two of the 3 positions used to construct haplotypes (see Table 2). Haplotype name has been derived from table 1 of Raychaudhuri et al.[10]. aa11, aa71 and aa74: amino acid (three letter code) at positions 11, 71 and 74. Associations with the presence of erosions were tested with GEE models (generalised estimating equation model) and the effect size expressed as an OR with its

95% confidence interval (CI). Associations with the Larsen score were tested with GLLAMM (Generalized Latent and Linear Mixed Model) and the effect size is expressed as an increase of Larsen score with its 95% (CI). The Larsen score ranges from 0 to 200; a higher score indicates a more severe damage. PAA-haplotype: previously reported to be the most frequent haplotype in control samples [10], and used as reference. N: number. Freq: frequency. The haplotype frequency is calculated as follows: (N of heterozygote RA patients + 2x N of homozygote RA patients)/(2x1832). OR: odds ratio. CI: confidence interval. p-val: p-value. Three letter amino acid code: Val, Valine; Leu, Leucine; Asp, Aspartic Acid; Pro, Proline; Gly, Glycine; Ser, Serine; Lys, Lysine; Arg, Arginine; Ala, Alanine; Glu, Glutamic Acid; Gln, Glutamine. **Bold: haplotype frequency > 5%; cumulatively, those haplotypes represent 89% of the inflammatory polyarthritis population.**

eFigure. Correlation between genetic risk scores for susceptibility and severity in NOAR



The genetic-risk score (GRS) is calculated as the sum of the risk-allele counts, weighted by the natural logarithm of their OR (which correspond to the regression coefficients), according to the method of Karlson and colleagues [13]. The GRS was calculated solely with the 16 HLA-DRB1 haplotypes described here. A GRS for susceptibility has been calculated using OR for the risk of developing anti-citrullinated protein antibodies positive RA (X axis). For the same patients, another GRS has been calculated using the coefficients from the GLLAMM regression in inflammatory polyarthritis (as taken from Supplementary table 7), in order to obtain the GRS for the Larsen score (y axis). For haplotypes associated with a decreased risk of disease susceptibility or severity, ln(OR) and regression coefficients will be < 1. Therefore, a negative GRS can for example occur when a patient carries two haplotypes associated with a decreased risk of disease susceptibility or severity. The GRS was calculated for all NOAR patients with available haplotype data (N=2534). One dot can represent several patients with the same GRS. The red line was fitted by linear regression (r^2: 0.57, p<1.0E-300).

eReferences

(1) Hwang YG, Moreland LW. Induction therapy with combination TNF inhibitor and methotrexate in early rheumatoid arthritis. *Curr Rheumatol Rep.* 2014;16:417.

(2) Viatte S, Plant D, Lunt M et al. Investigation of rheumatoid arthritis genetic susceptibility markers in the early rheumatoid arthritis study further replicates the TRAF1 association with radiological damage. *J Rheumatol.* 2013;40:144-156.

(3) James D, Young A, Kulinskaya E et al. Orthopaedic intervention in early rheumatoid arthritis. Occurrence and predictive factors in an inception cohort of 1064 patients followed for 5 years. *Rheumatology (Oxford).* 2004;43:369-376.

(4) Rabe-Hesketh S, Skrondal A, Pickles A. Generalized multilevel structural equation modelling. *Psychometrika.* 2004;69:167-190.

(5) Rabe-Hesketh S, Skrondal 528 A, Pickles A. Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics.* 2005;128:301-323.

(6) Rabe-Hesketh S, Skrondal A. Classical latent variable models for medical research. *Stat Methods Med Res.* 2008;17:5-32.

(7) Thomson W, Harrison B, Ollier B et al. Quantifying the exact role of HLA-DRB1 alleles in susceptibility to inflammatory polyarthritis: results from a large, population-based study. *Arthritis Rheum.* 1999;42:757-762.

(8) Robinson J, Halliwell JA, McWilliam H, Lopez R, Parham P, Marsh SG. The IMGT/HLA database. *Nucleic Acids Res.* 2013;41:D1222-D1227.

(9) Gonzalez-Galarza FF, Christmas S, Middleton D, Jones AR. Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. *Nucleic Acids Res.* 2011;39:D913-D919.

(10) Raychaudhuri S, Sandor C, Stahl EA et al. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet.* 483 2012;44:291-296.

(11) Eyre S, Bowes J, Diogo D et al. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat Genet.* 2012;44:1336-1340.

(12) Jia X, Han B, Onengut-Gumuscu S et al. Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One.* 2013;8:e64683.

(13) Karlson EW, Chibnik LB, Kraft P et al. Cumulative association of 22 genetic variants with seropositive rheumatoid arthritis risk. *Ann Rheum Dis.* 2010;69:1077-1085.