Comparison of methods to construct a genetic risk score for prediction of rheumatoid arthritis in the population-based HUNT study, Norway

S. Rostami, M. Hoff, M. A. Brown, K. Hveem, V. Videm

Author information:

Sina Rostami (SR), MSc, Department of Clinical and Molecular Medicine, NTNU - Norwegian University of Science and Technology, Trondheim, Norway. *Email*: sina.rostami@ntnu.no

Mari Hoff (MH), MD PhD, Associate Professor, Department of Rheumatology, St. Olavs University Hospital, Trondheim, Norway; Department of Neuromedicine and Movement Science & Department of Public Health and Nursing, NTNU - Norwegian University of Science and Technology, Trondheim, Norway. *Email*: mari.hoff@ntnu.no

Mattew A. Brown (MAB), MBBS MD FRACP FAHMS FAA, Director of Genomics, Institute of Health and Biomedical Innovation, Translational Research Institute, Princess Alexandra Hospital, Queensland University of Technology, Brisbane, Australia. *Email*: matt.brown@qut.edu.au

Kristian Hveem (KV), MD PhD, Professor, KG Jebsen Center for Genetic Epidemiology, NTNU - Norwegian University of Science and Technology, Trondheim, Norway. *Email*: kristian.hveem@ntnu.no

Vibeke Videm (VV), MD PhD, Professor, Department of Clinical and Molecular Medicine, NTNU - Norwegian University of Science and Technology & Department of Immunology and Transfusion Medicine, St. Olavs University Hospital, Trondheim, Norway. *Email*: vibeke.videm@ntnu.no

Corresponding author:

Vibeke Videm,

Department Clinical and Molecular Medicine, Lab Center 3 East, St. Olavs Hospital, NO-7006 Trondheim, Norway

E-mail: vibeke.videm@ntnu.no

https://orcid.org/0000-0002-1042-3889

Abstract

Objectives: Evaluate selection methods among published SNPs associated with rheumatoid arthritis (RA) to construct predictive genetic risk scores (GRS) in a population-based setting.

Methods: The Nord-Trøndelag Health Study (HUNT) is a prospective cohort study among the whole adult population of northern Trøndelag, Norway. Participants in HUNT2(1995-1997) and HUNT3(2006-2008) were included (489 RA cases, 61,584 controls). The initial SNP selection from relevant genome-wide studies included 269 SNPs from 30 studies. Following different selection criteria, SNPs were weighted by published odds ratios (OR). The sum of each person's carriage of all weighted susceptibility variants was calculated for each GRS.

Results: The best-fitting risk score included 27 SNPs (wGRS27) and was identified using p-value selection criterion $<=5\times10^{-8}$, the largest possible SNP selection without high linkage disequilibrium ($r^2<0.8$), and Lasso regression to select for positive coefficients. In a logistic regression model adjusted for gender, age, and ever smoking, wGRS27 was associated with RA (OR=1.86 (95% CI:1.71-2.04) for each standard deviation increase, p<0.001). The AUC was 0.76 (95% CI:0.74-0.78). The positive (PPV) and negative predictive values were 1.6% and 99.7%, respectively, and the PPV was not improved in sensitivity analyses sub-selecting participants to illustrate settings with increased RA prevalences. Other schemes selected more SNPs but resulted in GRSs with lower predictive ability.

Conclusion: Constructing a wGRS based on a smaller selection of informative SNPs improved predictive ability. Even with a relatively high AUC, the low PPV illustrates that there was a large overlap in risk variants among RA patients and controls, precluding clinical usefulness.

Keywords: Rheumatoid arthritis; Genetics; Epidemiology

Key messages

- A selection of informative SNPs rather than the complete list showed the best predictive ability.
- Despite relatively good discrimination, positive predictive value was low given the low RA prevalence.
- Novel methods are needed for genetic risk prediction to become useful on an individual level.

Introduction

RA is the most common autoimmune inflammatory rheumatic disease. Without proper treatment, RA may lead to synovial inflammation, cartilage and bone destruction, and severe joint destruction (1). In addition to affecting the patient's quality of life, the disease has economic consequences for society due to healthcare needs, treatment, and patients' reduced work participation (2, 3).

Genomic prediction medicine attempts to use a person's genomic profile to predict her/his susceptibility to develop certain diseases more accurately and at earlier stages. Targeting those at extreme risk of developing a disease could potentially guide healthcare services in adopting more effective preventive and/or interventive measures, as well as motivate high-risk individuals to make suitable lifestyle changes. RA is a disease where such prediction is of considerable interest.

The heritability of RA is ~50% for anti-citrullinated protein antibody (ACPA)-positive RA and ~20% for ACPA-negative RA, based on a nation-wide prospective Swedish study using a familial aggregation approach (4). The currently known RA susceptibility loci explain 39% of the known (50%) heritability (5), and at least 60% of this explained heritability is due to HLA (6). Thus, people expressing the "shared epitope" (SE) HLA alleles have a significantly increased RA susceptibility (7).

Due to the complex pathogenesis for development of RA, prediction models have been developed including genetic, clinical, and serological factors, using different statistical/modelling approaches (8-10). Although there have been improvements in the discriminative and predictive abilities of such models, they are not currently predictive at an individual level. Genome-wide association studies and meta-analyses have identified a more comprehensive selection of susceptibility genetic variants for RA, mostly single-nucleotide polymorphisms (SNPs). However, there is no general agreement on the best strategy to select relevant variations to include in a genetic risk score (GRS).

In general, a predictive score should be parsimonious, give good discrimination (efficiently classify individuals as cases or controls) and be well calibrated (correctly predict the observed event rates). The variables need not necessarily be causative for the endpoint as long as they contribute to prediction. Local models often perform better than general models because they

are fine-tuned to the relevant population. Explanatory models, on the other hand, often include many factors and preferably causative ones if possible, to explain the reasons for the variability in the data.

Our hypothesis was that when genome-wide genetic data are not available, it is preferable to develop a predictive GRS for RA based on a smaller selection of informative SNPs rather than a combination of all SNPs previously reported from association studies on RA. All reported SNPs will not necessarily contribute significantly to prediction, and the proportion of missing data is usually smaller when fewer variables are needed for the complete score. Inclusion of non-informative SNPs will also tend to decrease the signal-to-noise ratio due to larger variation among individuals. Moreover, we assumed that inclusion of non-genetic risk factors such as smoking, gender, and age could further improve the predictive ability of the model. We finally hypothesized that the best-fitting model would be associated with RA in a population-based setting.

The aims of the present study were twofold, using data from the HUNT population-based study: 1) To compare different approaches for SNP selection when developing a GRS and 2) To evaluate the predictive ability of the best GRS when also including non-genetic factors.

Patients and Methods

The Nord-Trøndelag Health Study (HUNT) is a prospective cohort study conducted in the northern region of Trøndelag County, Norway, as previously described (11). All adults (≥20 years) were invited to take part. The present study used data from participants in HUNT2 (1995-1997) and/or HUNT3 (2006-2008). Several approaches including questionnaires, interviews, clinical examinations, and blood sampling were used to collect data in HUNT (11).

The present study was part of the HuLARS study (**HU**NT **L**ongitudinal **A**nkylosing spondylitis and **R**heumatoid arthritis **S**tudy) (12). RA cases were validated by evaluation of hospital case files at the three hospitals in the region using the American College of Rheumatology (ACR)/European League Against Rheumatism (EULAR) 2010 criteria or for some cases diagnosed before 2010, the ACR criteria from 1987 (13, 14). The RA diagnosis was given at any time before participation in HUNT or until the end of diagnosis ascertainment, which lasted from May till December 2015. Thus, the cases had different disease duration, and 32 cases

were diagnosed after participation in HUNT3. Baseline variables were registered at inclusion in HUNT, i.e. either at HUNT2 or HUNT3. We excluded participants with self-reported RA and missing case files or in a few instances an uncertain RA diagnosis, as well as those with ankylosing spondylitis, psoriatic arthritis, juvenile idiopathic arthritis, and other inflammatory arthritis. Thus, these cases were not included in the control group. The final dataset included 578 RA cases and 76,462 controls (Figure 1). After exclusion of participants due to other missing data, the main analyses were performed among 489 RA cases and 61,584 controls (Figure 1).

The HUNT study was approved by the Regional Committee for Medical and Health Research Ethics (REK), the Norwegian Data Inspectorate, and the National Directorate of Health. All participants gave written informed consent, and the study was performed in accordance with the Helsinki declaration. The HuLARS study was approved by REK (REK Midt 2009/661), and the Norwegian Data Inspectorate.

Genotyping using HumanCoreExome arrays from Illumina (HumanCoreExome12 v1.0, HumanCoreExome12 v1.1 and UM HUNT Biobank v1.0) and imputation was performed as previously described (15). SNPs were selected based on a thorough literature review of association studies on RA in PubMed until 29.12.2018. Initially, 48 articles published in English including results from large population-based studies based on recent European Caucasian ancestry were found. We listed 335 SNPs and excluded 51 SNPs due to lack of reported p-value, missing information on the risk allele and/or odds ratio (OR) of the risk allele, or missing replication. We lacked genetic data for 15 SNPs, and finally included 269 SNPs from 30 papers (Table S1). Two SNPs related to the SE (HLA-DRB1*04 and *0401) were available, whereas 5 other genetic variants related to the SE were not. We also lacked data for 2 SNPs in HLA-DP and 1 in HLA-DOA. Due to the low coverage of genotyped SNPs in the HLA region as compared to the Illumina Immunochip which has been used in various other studies (16), imputation did not give sufficiently accurate results for use in the study. The remaining missing SNPs were not HLA-related.

Alternative approaches for SNP selection for the GRS were as follows: **A)** Starting with SNPs previously reported to have a p-value $<=5\times10^{-6}$ (174 SNPs); **B)** Starting with SNPs reported to have a p-value $<=5\times10^{-8}$ (120 SNPs); **C)** Starting with SNPs selected in A), and using the product of the risk allele frequency and risk allele OR reported in the literature. This product

was used as "selection weight" to rank and select risk SNPs of larger effect sizes which are more common, as previously suggested for coronary heart disease (17). For each of A) and B), a series of three GRS were then developed. **1)** after selection of SNPs on each chromosome having low linkage disequilibrium (LD, i.e. $r^2<0.8$) using LDlink (18) to choose the largest number of non-linked SNPs showing the highest ORs; **2**) After selection based on 1) additionally using Lasso regression (n=50-fold cross-validation due to the large dataset, using random sampling) to select the SNPs showing non-zero coefficients; 3) similar to 2) but choosing only SNPs showing positive coefficients in Lasso regression, i.e. removing SNPs for which the risk allele in our population was different from that of the previously published studies. As a sensitivity analysis, a GRS using Lasso regression starting with all 269 possible SNPs without preselection was also developed. The SNPs included in each GRS are listed in Table S1. After weighting the risk variants by the natural logarithm of their OR from previous studies on RA (references in Table S1), we constructed each final weighted genetic risk score (wGRS) for each participant as the sum of the weighted risk variants. The h2-index for variance explained by different SNP selections was estimated using the software Genome-wide Complex Trait Analysis (GCTA) (19), transforming the estimate to the liability scale assuming an RA prevalence in northern Trøndelag of 0.8% (12).

To be able to compare OR of different wGRSs, each wGRS was standardized (hereafter called swGRS) by dividing by its standard deviation in the HUNT population. Logistic regression was used to develop models including each of the swGRS as a continuous variable. In parallel, models were developed including adjustment variables (gender, age, and ever smoking defined as self-reporting present or previous smoking). Given that the baseline data were recorded at either HUNT2 or HUNT3, we used an additional indicator variable to account for potential differences.

Linearity of logits were checked using plots. Calibration (goodness-of-fit) was evaluated using the Hosmer-Lemeshow test, and model discrimination was evaluated by AUC (area under the receiver operating characteristics curve). The best model was internally validated using bootstrapping (n=1000 repetitions). Relative model fit among models was compared using the Akaike (AIC) and Bayesian (BIC) information criteria.

For a subset of participants, information on the presence or absence of the SE (20) was available (Figure 1). In these participants, models including the best swGRS from the study, the best

swGRS after removal of the only included SE-related SNP (rs6457617, which tags HLA-DRB1*0401), and SE carrier state were evaluated.

Potential clinical usefulness of the best-fitting model was evaluated using the Youden index to define the cut-point with the highest sensitivity and specificity, and to determine PPV and NPV (positive and negative predictive value) (21). This was first evaluated in the entire study population. To investigate how the model would behave in a population with a higher prevalence of RA, two subsets of HUNT participants were selected (Figure 1). First, participants were sub-selected based on self-report of chronic pain located in hands, knees, ankles, or feet (n=414 RA cases and 19,300 controls, Figure 1). Second, the sub-set of control participants with self-reported osteoarthritis (OA) in HUNT3 without RA were compared to the RA patients (n=489 RA cases and 3,275 controls). We finally also evaluated the model in 61,584 controls and the RA seropositive cases (anti-CCP and/or rheumatoid factor positive, n=350, i.e. 72% of all RA, Figure 1).

Statistical analysis was performed using Stata (v.15, StataCorp, College Station, Texas, USA). Data are presented as mean±SD or OR (95% CI) unless otherwise stated. P-values <0.05 (without adjustment for multiple-hypothesis testing due to different modelling schemes) were regarded statistically significant.

Results

Baseline characteristics of study participants for the main analysis (Figure 1) are given in Table 1. Table 2 summarizes associations for swGRS and other adjustment variables in the six models each including one of the swGRS. All models showed acceptable fit by Hosmer-Lemeshow tests.

The swGRS with the highest OR (1.86 (1.71-2.04) for one unit increase in SD, p<0.001) was found in Model B3. This was the swGRS with the most limited number of SNPs (27 SNPs), hereafter called "swGRS27" (Table 2). The numerical value of swGRS27 ranged from 2.37 to 10.26, with a median of 6.55 (interquartile range (IQR): 5.85-7.20) in the RA cases, and 5.85 (IQR: 5.18-6.55) in the controls (Figure 2). This implies that even for the best-fitting model developed, there was considerable overlap in the distribution of swGRS27 between RA cases and controls (Figure 2). Among models including adjustment variables, model B3 had the best fit (lowest AIC and BIC), and significantly better discrimination (AUC=0.76 (0.74-0.78)), than

the other models except A3 (p=0.38) (Table 2 and Figure 3). Thus, including adjustment variables with swGRS27 resulted in a model with significantly improved discrimination (p<0.001, unadjusted model: AUC=0.67(0.65-0.70); Figure 3). Internal validation of the adjusted model with swGRS27 demonstrated bootstrapped CI very close to the original ones, indicating that the results were unbiased.

Overall, ORs for swGRS and model fit were better in corresponding models when SNPs were initially selected by p-value<=5×10^{-8} rather than p-value<=5×10^{-6} (Model series B vs A, Tables 2 and S2). Table S2 summarizes the parallel unadjusted models. The swGRS based on Lasso selection from all available SNPs (n=269) included 50 SNPs and had no better predictive ability than the model with swGRS27 (p=0.82, data not shown).

The h2-index for the complete selection of 269 SNPs was 5.3 (95% CI: 3.1-7.5)%, 4.5(2.5-6.6)% for the selection of 174 SNPs associated with RA with a published p-value <=5×10^{-6}, and 5.2(2.0-8.3)% for the 27 SNPs in swGRS27.

The models based on both risk allele frequency and OR ("selection weights") were inferior to those from the other selection approaches. A histogram of the selection weights is shown in Figure S1. Different selections of SNPs with selection weights higher than arbitrary thresholds were investigated, showing no specific trend towards better fit or discrimination (Table S3). The best models either included 99 SNPs or all 115 possible SNPs.

The analysis in participants where SE carrier state was available showed that the SE alone was a statistically significant predictor (standardized OR 1.48(1.31-1.67), table S4). swGRS26, constructed by removal of the only SE-related SNP in swGRS27, had comparable predictive ability alone (OR 1.42(1.26-1.59)). The bivariable model including swGRS26 and SE carrier state had better fit (Table S4) and discrimination (p<0.001) than the univariable models with either swGRS26 or SE, and these variables were both significant, independent predictors of RA (SE: OR 1.50(1.32-1.69), swGRS26: OR 1.43(1.27-1.60)).

The subgroup analysis using prior selection of participants for chronic pain or using self-reported OA to restrict the number of controls while keeping all RA cases increased the prevalence of RA from 0.8% in the main analysis to 2.1% and 13%, respectively. Table 3 summarizes sensitivity, specificity, PPV, and NPV at the cut-point defined by the Youden

index for the best-fitting model (with swGRS27). The PPV increased from 1.6% in the main adjusted model to 3.3% and 18.3% in the two subsets, respectively. However, the PPV decreased to 1.2% when the analysis included only the seropositive RA cases (Table 3).

Discussion

In this general population-based study using replicated risk SNPs for RA, the best-fitting swGRS included 27 susceptibility SNPs (swGRS27) out of 269 possible. It was identified using a p-value selection criterion $<=5\times10^{-8}$ and Lasso regression to remove SNPs that did not contribute to prediction or had a different risk allele in our population than previously published. Including knowledge on population risk allele frequency was of no benefit. Thus, a parsimonious model was better than those including all available information. In accordance with the non-genetic component of RA development, the addition of gender, age, and ever smoking to the swGRS27 model further improved discrimination. The AUC for the adjusted model including swGRS27 was relatively high (0.76).

Even if AUC provides information on discrimination, the NPV and especially PPV in the relevant study population are more important for practical clinical use. These measures are dependent on the disease prevalence, which was 0.8% for RA in the present study population. Even with a prevalence as high as 13% in the sub-selection using OA controls, the PPV was only 18.3%. Cut-offs could be defined at different thresholds, but the proportion of false-positives and false-negatives would still be high. Our results are in accordance with previous literature reporting that classification based on genetic markers with significant OR is not straightforward: For a prediction model with an AUC of 0.79 and disease prevalences of 15%, 5.5%, or 1.5%, cases among those at high risk are correctly classified in 30%, 12% and 3%, respectively (22).

A limitation of our study is that we had few data for the SE. The subgroup analysis indicated that the SE and the non-SE-associated SNPs each independently contributed to prediction with approximately the same OR. In the main analysis, where information on the SE was missing, the marker SNP rs6457617 of the common SE variant HLA-DRB1*0401 was selected in all GRS including swGRS27. Thus, rs6457617 may partly have acted as a proxy for the SE. For the unadjusted models, it is noteworthy that the variation in OR for different models was relatively small (swGRS27, i.e. best model: OR=1.85, swGRS including all 115 SNPs:

OR=1.64, swGRS including 3 SNPs using selection-weight-based selection: OR=1.51, SE alone in a subgroup of participants: OR=1.48, Tables S2, S3 and S4).

Furthermore, other SNPs in swGRS27 may have captured some of the same information as the SE, in accordance with the finding of many genetic interactions between the SE and variants outside of the HLA region (23). We cannot exclude that our results would have changed with access to more SE-related data. This would not necessarily influence the findings regarding identification of the best SNP selection method, which was the main aim of the present study. Furthermore, the AUC for our best model was comparable to 0.74 found for a model including 45 RA susceptibility SNPs, 5 HLA amino acids, and gender in a case-control study with 11,366 RA cases and 15,489 controls (24).

Many studies on RA genetics have been restricted to seropositive cases, who have high carriage of the SE. This may increase the relative importance of SE-related SNP. One of the highest AUCs (0.857) for RA prediction was reported based a computer simulation approach that used 15 four-digit/10 two-digit *HLA-DRB1* alleles, 31 SNPs, and smoking to discriminate ACPA-positive RA cases from controls (8).

We started with a comprehensive list of potential susceptibility SNPs and also included seronegative cases to avoid case loss, which may have influenced the results regarding the importance of HLA. Our best model had a slightly higher AUC when evaluated only in seropositive cases, but this did not translate to a useful increase in PPV (data not shown). Thus, the inclusion of seronegative cases did not seem to substantially have influenced the results for the seropositive cases. Furthermore, the distinction between seronegative and seropositive RA is becoming more blurred with the findings of autoantibodies with other specificities than ACPA (25). The relatively low h2 index may partly be explained by our inclusion of seronegative cases, and possibly also by our study capturing less of the heritability due to missing SNPs.

We excluded participants with ankylosing spondylitis, psoriatic arthritis and other forms of inflammatory arthritis. A previous genetic study has shown relatively small overlap in SNPs associated with ankylosing spondylitis/psoriasis and RA (26), but the exclusion may still have resulted in somewhat better performance of our risk score than without such exclusion.

Recently, several genome-wide polygenic risk scores (GPS) that evaluate thousands to millions of genetic variants have been tested for potentially better prediction of complex diseases. To our knowledge, no GPS for RA has yet been published. A GPS captures a greater proportion of the heritability. As an example, a model based on GPS for coronary artery disease (CAD) including 6,630,150 SNPs and adjustment variables identified a larger proportion of CAD events in high risk score percentiles than two previous wGRS based on 49,310 and 50 SNPs (27). Even with an AUC of 0.81, the risk score distributions for cases and controls showed a noticeable overlap, which is another way of illustrating that there would be many false-positives and false-negatives.

We may speculate that the overlapping scores may be a consequence of the genetic architecture of many common diseases, where familial risk through rare variants with moderate effect has some importance, whereas non-familial risk is caused by a very large number of common variants that each have a small and additive effect (28). Thus, the exact risk profile probably differs among individuals, which would lead to noise and overlap in the risk score distribution. For explanation of disease pathogenesis, it is important to capture as many risk variants as possible. For prediction, on the other hand, it is more useful to capture the largest differences among cases and controls; thus, more SNPs will not necessarily perform better. The use of selection weights where common risk variants were prioritized has previously been useful to select SNPs for prediction of CAD (17). In our study, the definition of a threshold for SNPs to include was not obvious and this approach led to worse models for prediction of RA.

In the present study, Lasso regression helped identify a smaller subset of SNPs that improved prediction. This approach fine-tuned the selection for our population, at the cost of lower generalizability due to differences in genotype frequencies and other characteristics among populations. It is well-known from other areas of risk prediction that local scores usually perform better. We are not suggesting that the swGRS27 is appropriate for clinical use, so lack of generalizability was not an issue when comparing the selection schemes. We did not perform internal validation of all the different risk models but bootstrapping of the model including swGRS27 indicated little bias.

The inclusion of non-genetic factors improved model performance, as expected because such factors play about the same role as heritability for disease development. Smoking had a high OR and is an important target for prevention (29). Inclusion of other environmental risk factors

would help improve the risk score, but there is a lack of general knowledge about the most useful factors and their importance.

Most previous prediction studies on RA were relatively small and did not have a prospective cohort design. However, such a design will lead to fewer RA cases than case-control studies unless the initial sample size is very large. A strength of our study was that population controls may give less selection bias. We did not have information regarding family history of RA, which is an important risk factor (30) that would have captured genetic risk specific to each individual. We cannot exclude false-positive and false-negative RA diagnoses.

Developing a wGRS based on fewer, but informative SNPs improved the predictive ability. Despite an AUC of 0.76, the PPV was low, which precludes clinical usefulness. Our study also confirmed previous observations that there was substantial overlap in risk variants among RA cases and controls, as expected when individual genetic risk depends on small, additive effects from a large number of potential risk variants.

References

1. McInnes IB, Schett G. The pathogenesis of rheumatoid arthritis. N Engl J Med 2011;365:2205-19.

2. Bala SV, Samuelson K, Hagell P, Fridlund B, Forslind K, Svensson B, et al. Living with persistent rheumatoid arthritis: a barfot study. Journal of Clinical Nursing 2017;26:2646-56.

3. Uhlig T, Moe RH, Kvien TK. The burden of disease in rheumatoid arthritis. Pharmacoeconomics 2014;32:841-51.

4. Frisell T, Holmqvist M, Kallberg H, Klareskog L, Alfredsson L, Askling J. Familial risks and heritability of rheumatoid arthritis: role of rheumatoid factor/anti-citrullinated protein antibody status, number and type of affected relatives, sex, and age. Arthritis Rheumatol 2013;65:2773-82.

5. Viatte S, Barton A. Genetics of rheumatoid arthritis susceptibility, severity, and treatment response. Semin Immunopathol 2017;39:395-408.

6. Kim K, Bang SY, Lee HS, Bae SC. Update on the genetic architecture of rheumatoid arthritis. Nat Rev Rheumatol 2017;13:13-24.

7. Viatte S, Plant D, Raychaudhuri S. Genetics and epigenetics of rheumatoid arthritis. Nat Rev Rheumatol 2013;9:141-53.

8. Scott IC, Seegobin SD, Steer S, Tan R, Forabosco P, Hinks A, et al. Predicting the risk of rheumatoid arthritis and its age of onset through modelling genetic risk variants with smoking. PLoS Genet 2013;9:e1003808.

9. Joo YB, Kim Y, Park Y, Kim K, Ryu JA, Lee S, et al. Biological function integrated prediction of severe radiographic progression in rheumatoid arthritis: a nested case control study. Arthritis Res Ther 2017;19:244.

10. Karlson EW, van Schaardenburg D, van der Helm-van Mil AH. Strategies to predict rheumatoid arthritis development in at-risk populations. Rheumatology (Oxford) 2016;55:6-15.

11. Krokstad S, Langhammer A, Hveem K, Holmen TL, Midthjell K, Stene TR, et al. Cohort profile: the hunt study, Norway. Int J Epidemiol 2013;42:968-77.

12. Videm V, Thomas R, Brown MA, Hoff M. Self-reported diagnosis of rheumatoid arthritis or ankylosing spondylitis has low accuracy: data from the Nord-Trøndelag health study. J Rheumatol 2017;44:1134-41.

13. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO, et al. 2010 rheumatoid arthritis classification criteria: an american college of rheumatology/european league against rheumatism collaborative initiative. Ann Rheum Dis 2010;69:1580-8.

14. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, et al. The american rheumatism association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315-24.

15. Rostami S, Hoff M, Brown MA, Hveem K, Holmen OL, Fritsche LG, et al. Prediction of ankylosing spondylitis in the population-based hunt study by a genetic risk score combining 110 snps of genome-wide significance. J Rheumatol 2019 published on 1 April 2019. doi:10.3899/jrheum.181209

16. Cortes A, Brown MA. Promise and pitfalls of the immunochip. Arthritis Res Ther 2011;13:101.

17. Beaney K, Drenos F, Humphries SE. How close are we to implementing a genetic risk score for coronary heart disease?. Expert Rev Mol Diag 2017;17:905-15.

18. Ldlink [Internet]. National Cancer Institute (USA); [cited 20.05.2019]; Available from: https://ldlink.nci.nih.gov/.

19. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet 2011;88:294-305.

20. du Montcel ST, Michou L, Petit-Teixeira E, Osorio J, Lemaire I, Lasbleiz S, et al. New classification of hla-drb1 alleles supports the shared epitope hypothesis of rheumatoid arthritis susceptibility. Arthritis Rheum 2005;52:1063-8.
21. Youden WJ. Index for rating diagnostic tests. Cancer 1950;3:32-5.
22. Jakobsdottir J, Gorin MB, Conley YP, Ferrell RE, Weeks DE. Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. PLoS Genet 2009;5:e1000337.
23. Diaz-Gallo LM, Ramskold D, Shchetynsky K, Folkersen L, Chemin K, Brynedal B, et al. Systematic approach demonstrates enrichment of multiple interactions between non-hla risk variants and hla-drb1 risk alleles in rheumatoid arthritis. Ann Rheum Dis 2018;77:1454-62.
24. Yarwood A, Han B, Raychaudhuri S, Bowes J, Lunt M, Pappas DA, et al. A weighted genetic risk score using all known susceptibility variants to estimate rheumatoid arthritis risk. Ann Rheum Dis 2015;74:170-6.
25. Trouw LA, Mahler M. Closing the serological gap: promising novel biomarkers for the early diagnosis of rheumatoid arthritis. Autoimmun Rev 2012;12:318-22.
26. Parkes M, Cortes A, van Heel DA, Brown MA. Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nat Rev Genet. 2013;14(9):661-73.
27. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet 2018;50:1219-24.
28. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet 2018;19:581-90.
29. Klareskog L, Gregersen PK, Huizinga TW. Prevention of autoimmune rheumatic disease: state of the art and future perspectives. Ann Rheum Dis 2010;69:2062-6.
30. Jiang X, Frisell T, Askling J, Karlson EW, Klareskog L, Alfredsson L, et al. To what extent is the familial risk of rheumatoid arthritis explained by established rheumatoid arthritis risk factors? Arthritis Rheumatol 2015;67:352-62.

Funding

This work was supported by the Liaison Committee between the Central Norway Regional Health Authority and NTNU [5056/46051000 to VV]; the Research Council of Norway [249944 to SR and VV]; and National Health and Medical Research Council Senior Principal Research Fellowship [APP1024879 to MAB].

Acknowledgements

The HUNT study is a collaboration between the HUNT Research Centre (Faculty of Medicine and Health sciences, NTNU – Norwegian University of Science and Technology), Nord-Trøndelag County Council, Central Norway Health Authority, and the Norwegian Institute of Public Health.

Conflict of interest

The authors declare no conflicts of interest.

Data sharing

Data from HUNT are available upon reasonable request from the HUNT Research Centre (https://www.ntnu.edu/hunt/data), following approval from the Regional Research Ethics Committee. However, restrictions apply to the availability of the data for the present paper, which were used under license for the current study and are not publicly available in accordance with Norwegian law.

Author contributions

Study conception and design: MH, MAB, VV. Acquisition and analysis of data: SR, VV. Interpretation of data: MH, MAB, KH, VV. Drafting the manuscript: SR, VV. Revising the manuscript critically for important intellectual content: MH, MAB, KH, VV. All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Figure legends

Figure 1: Inclusion and exclusion of study participants

HUNT2 denotes individuals who participated in HUNT2 only, HUNT2 & HUNT3 denotes individuals who participated both in HUNT2 and HUNT3, and HUNT3 denotes individuals who participated in HUNT3 only. Combined dataset denotes the sum of these three groups.

Figure 2: Genetic risk score (swGRS27) distributions

Distributions in RA cases and controls of the best-fitting weighted genetic risk score (swGRS27, Model B3 in Table 2), standardized using the population standard deviation.

Figure 3: ROC curve for unadjusted and adjusted models including swGRS27

Receiver operating characteristic curves (ROC) for models with the best-fitting weighted genetic risk score (swGRS27, Model B3 in Table 2), standardized using the population standard deviation. Adjusted model also included gender, age, and ever smoking.

Table 1: Baseline characteristics of study participants in the main analysis[a]

	HUNT2 (n=51,275)		HUNT3 (n=10,798)[b]	
	RA cases (n=443)	Controls (n=50,832)	RA cases (n=46)	Controls (n=10,752)
Seropositive RA[c], n(%)	315(73)	NA	35(76)	NA
Age years, mean ± SD	57±13	49±17	60±15	40±15
Women, n(%)	297(67)	26576(52)	29(63)	5715(53)
Ever smoker, n(%)	294(66)	28067(55)	30(65)	5137(48)
Previous cardiovascular disease[d], n(%)	52(12)	3605(7)	6(13)	400(4)
Diabetes, n(%)	22(5)	1463(3)	5(11)	282(3)
Hypertension[e], n(%)	215(49)	21101(42)	27(71)	2106(22)
Body mass index (kg/m^2), mean ± SD	27±4	26±4	28±4	27±5

RA: rheumatoid arthritis; HUNT: Nord-Trøndelag Health Study, waves 2 or 3

[a]The main analysis included 489 RA cases and 61,584 controls.

[b]Individuals who participated in HUNT3 only (and not HUNT2).

[c]Seropositive: Positive for rheumatoid factor and/or anti-citrullinated protein antibody

[d]Previous cardiovascular disease: Self-reported myocardial infarction, angina pectoris or stroke.

[e]Hypertension: Either a "yes" response to the question "Are you using medication for high blood pressure?", or measurement of systolic blood pressure ≥140 mmHg and/or diastolic blood pressure ≥90 mmHg

Table 2: Adjusted logistic regression models using different standardized weighted genetic risk scores for rheumatoid arthritis[a]

Model	Number of SNPs[b]	swGRS[c] OR(95%CI)	Female gender OR(95%CI)	Age OR(95%CI)	Ever smoker OR(95%CI)	AUC (95%CI)	AIC	BIC
A1[d]	115	1.65 (1.51-1.81)*	2.11 (1.74-2.57)*	1.04 (1.03-1.04)*	1.85 (1.53-2.25)*	0.74 (0.72-0.76)	5348.50	5402.72
A2[d]	36	1.79 (1.64-1.96)*	2.10 (1.73-2.55)*	1.04 (1.03-1.04)*	1.85 (1.52-2.24)*	0.75 (0.73-0.77)	5305.39	5359.61
A3[d]	30	1.85 (1.69-2.02)*	2.10 (1.73-2.55)*	1.04 (1.03-1.04)*	1.85 (1.52-2.24)*	0.76 (0.74-0.78)	5287.99	5342.20
B1[d]	88	1.69 (1.54-1.85)*	2.12 (1.75-2.57)*	1.04 (1.03-1.04)*	1.85 (1.53-2.25)*	0.75 (0.73-0.77)	5339.84	5394.06
B2[d]	29	1.84 (1.68-2.01)*	2.10 (1.73-2.55)*	1.04 (1.03-1.04)*	1.86 (1.53-2.26)*	0.76 (0.74-0.78)	5291.57	5345.79
B3[d]	27	1.86 (1.71-2.04)*	2.10 (1.73-2.55)*	1.04 (1.03-1.04)*	1.85 (1.53-2.25)*	0.76 (0.74-0.78)	5282.37	5336.59

[a]All models included 489 RA cases and 61,584 controls

[b]Number of SNPs in each swGRS

[c]Weighted genetic risk score standardized using corresponding population standard deviation

[d]**A** models were based on SNPs with reported p-value<=5×10^{-6} in association studies in European Caucasian populations. **B** models were based on SNPs with reported p-value <=5×10^{-8}. Model numbers indicate selection based on: 1) the largest selection of SNPs showing low linkage disequilibrium (r^2<0.8) on each chromosome; 2) SNPs additionally showing non-zero coefficients using Lasso regression; 3) SNPs additionally showing positive coefficients in Lasso regression.

*p<0.001

Table 3: Properties of the best-fitting model (with swGRS27)[a] in different participant selections

Model	Sensitivity%	Specificity%	Positive predictive value%	Negative predictive value%
All participants[b]				
Unadjusted model	67.3	59.8	1.3	99.6
Adjusted[c] model	78.9	61.3	1.6	99.7
Seropositive RA[d], all controls				
Unadjusted model	71.7	59.8	1.0	99.7
Adjusted[c] model	82.9	61.3	1.2	99.8
Participants with chronic pain[e]				
Unadjusted model	68.6	59.4	3.5	98.9
Adjusted[c] model	79.7	50.4	3.3	99.1
All RA, controls with osteoarthritis[f]				
Unadjusted model	67.3	60.8	20.4	92.6
Adjusted[c] model	78.9	47.4	18.3	93.8

[a]swGRS27: best-fitting weighted genetic risk score (Model B3 in Table 2), standardized using the population standard deviation

[b]489 RA cases and 61,584 controls, using cut-point defined by the Youden index

[c]Adjusted for gender, age, and ever smoking

[d]350 seropositive RA cases and 61,584 controls

[e]Participants having self-reported pain located in hands, knees, ankles, or feet and lasting for three months or more during last year before HUNT participation, 414 RA cases and 19,300 controls

[f]All available RA cases (n=489) and 3,275 controls with self-reported osteoarthritis

Figure 1

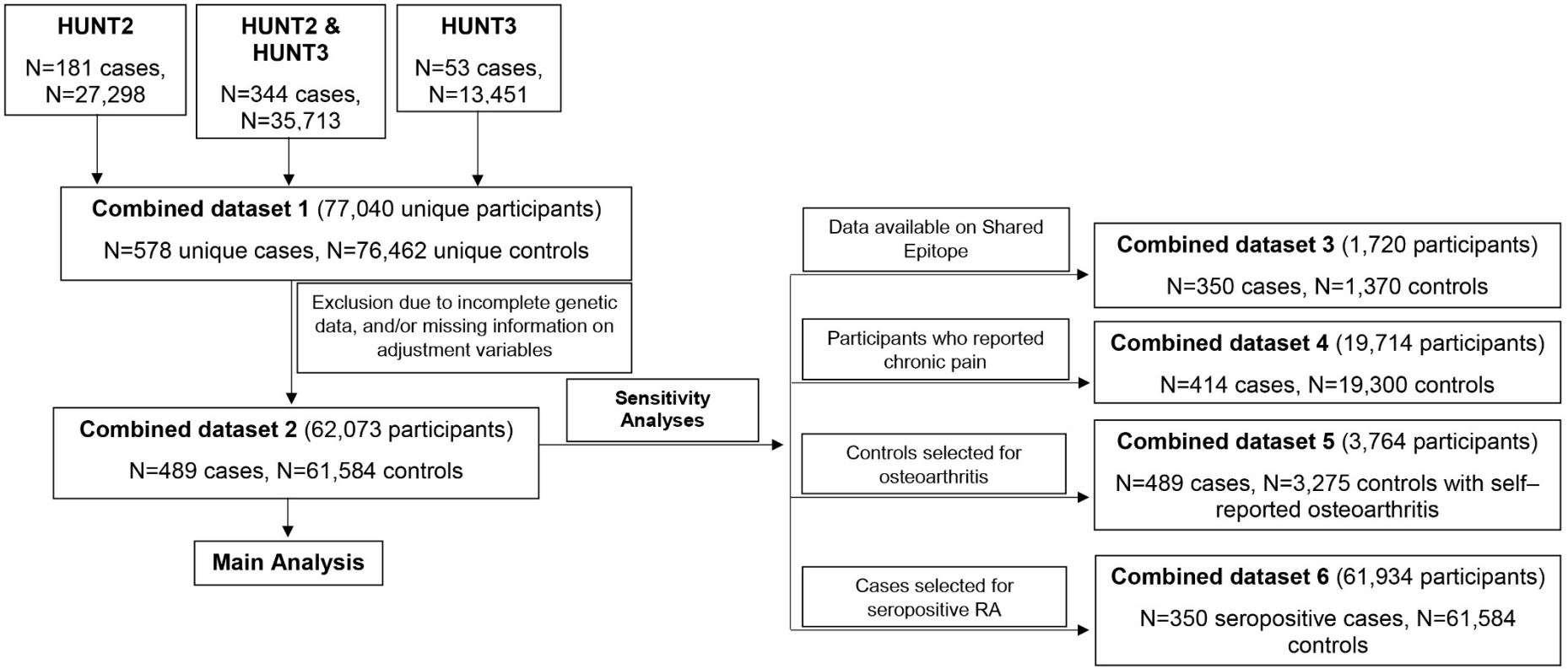

HUNT2
N=181 cases, N=27,298
HUNT2 & HUNT3
N=344 cases, N=35,713
HUNT3
N=53 cases, N=13,451
Combined dataset 1 (77,040 unique participants)
N=578 unique cases, N=76,462 unique controls
Exclusion due to incomplete genetic data, and/or missing information on adjustment variables
Combined dataset 2 (62,073 participants)
N=489 cases, N=61,584 controls
Main Analysis
Sensitivity Analyses
Data available on Shared Epitope
Combined dataset 3 (1,720 participants)
N=350 cases, N=1,370 controls
Participants who reported chronic pain
Combined dataset 4 (19,714 participants)
N=414 cases, N=19,300 controls
Controls selected for osteoarthritis
Combined dataset 5 (3,764 participants)
N=489 cases, N=3,275 controls with self–reported osteoarthritis
Cases selected for seropositive RA
Combined dataset 6 (61,934 participants)
N=350 seropositive cases, N=61,584 controls

Figure 2

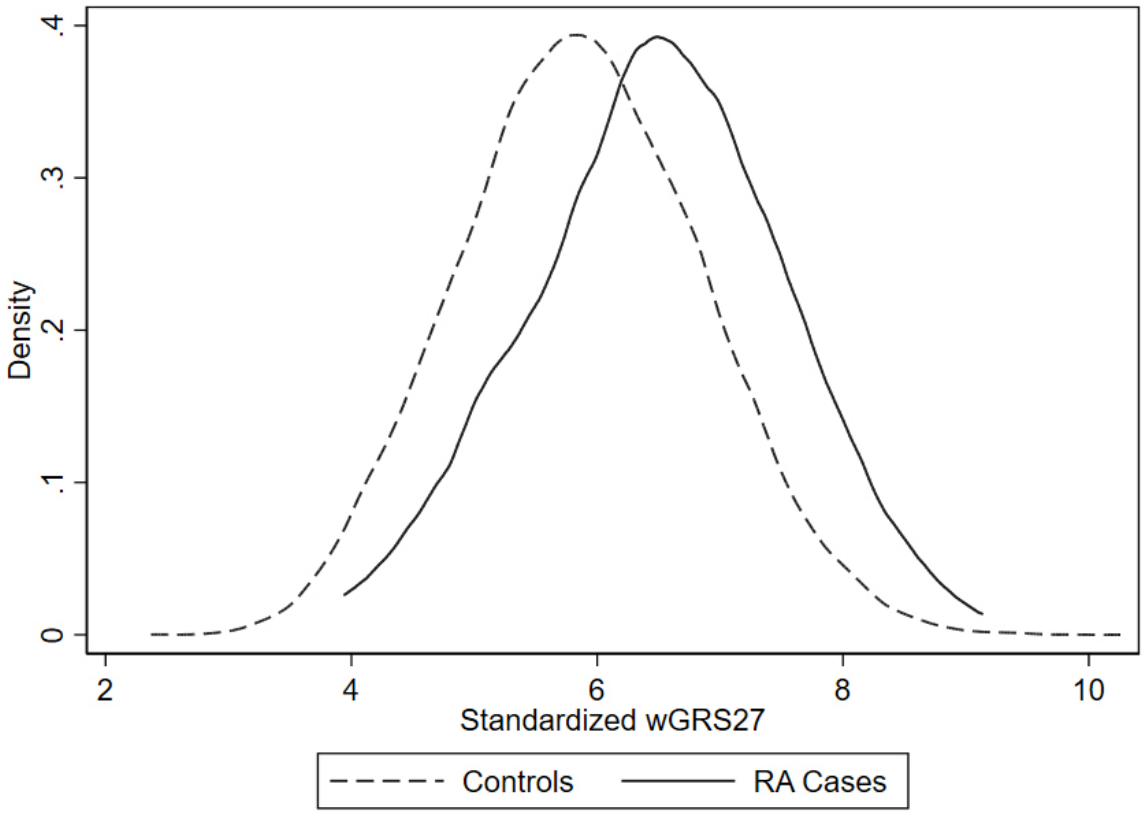
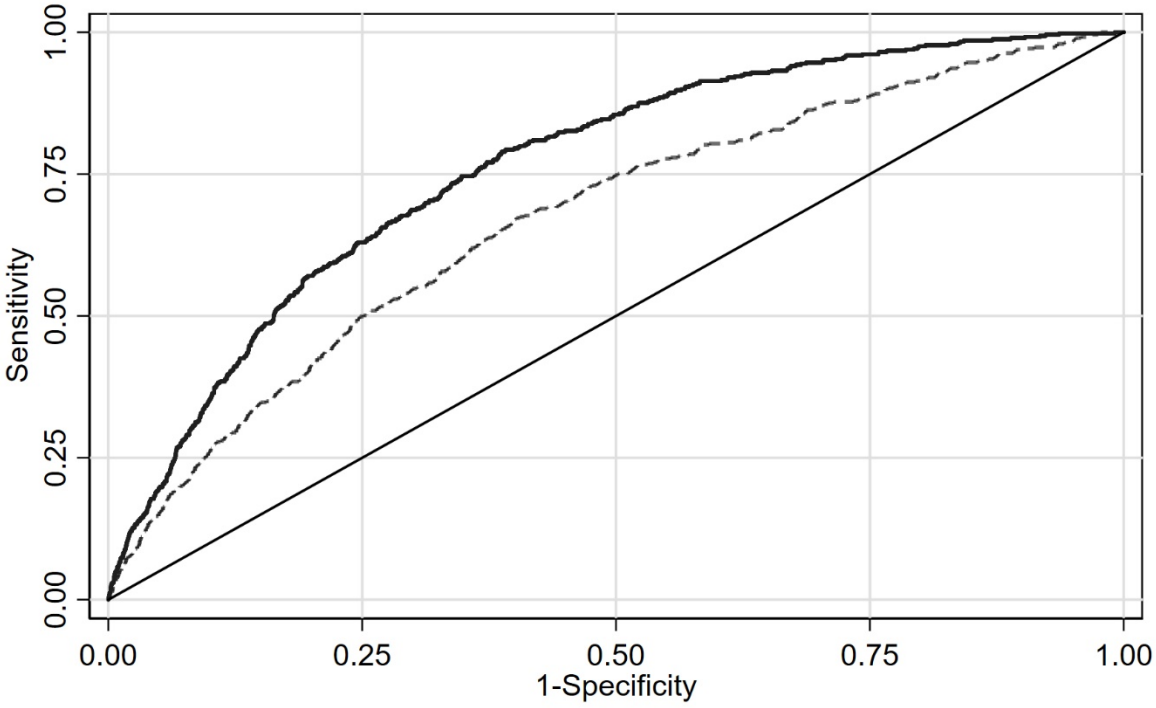

Density
0
.1
.2
.3
.4
2
4
6
8
10
Standardized wGRS27
Controls
RA Cases

Figure 3



	N	AUC	Standard Error	95% Confidence interval	
swGRS27 Unadjusted	62,073	0.671	0.012	0.647	0.695
swGRS27 Adjusted	62,073	0.762	0.010	0.742	0.782

Supplement

SNPs were initially selected based on the reported p-value<5×10^{-6} or <5×10^{-8} in previous associations studies among European Caucasian populations. "Non-LD" refer to largest SNP selection not showing high linkage disequilibrium (defined as $r^2>0.8$) with other SNPs on the same chromosome. The non-LD selected SNPs were then evaluated in Lasso regression to identify SNPs showing non-zero coefficients or positive coefficients.

SNPs indicated with "*" were included to develop the respective risk scores, and SNPs with empty fields were not included.

Table S1: Selection of SNPs from previous association studies to develop weighted genetic risk scores for rheumatoid arthritis in HUNT

			Based on p<5×10^{-6}				Based on p<5×10^{-8}				
Chromosome	**Position**	**SNP ID**	**p-value <5×10^{-6} (n=174)**	**Non-LD (n=115)**	**Lasso showing non-zero coefficient (n=36)**	**Lasso showing positive coefficient (n=30)**	**P-value <5×10^{-8} (n=120)**	**Non-LD (n=88)**	**Lasso non-zero (n=29)**	**Lasso pos (n=27)**	**Reference**
1	2528133	rs2843401	*	*			*	*	*	*	1
1	2553624	rs3890745	*								2
1	17672730	rs2301888	*	*			*	*			3
1	17674402	rs2240336	*	*	*	*	*	*	*	*	1
1	38278579	rs28411352	*	*			*	*			3
1	38616871	rs883220	*	*			*	*			1
1	38633879	rs12140275	*				*				3
1	80572058	rs11162922	*	*							4
1	114303808	rs6679677	*	*	*	*	*	*	*	*	4
1	114377568	rs2476601	*				*				3
1	117263138	rs11586238	*	*			*	*			5
1	117263790	rs624988	*	*	*	*	*	*	*	*	3
1	154426970	rs2228145	*	*			*	*			1
1	161405053	rs72717009	*	*	*	*					3
1	161479745	rs1801274	*	*							6
1	167408670	rs840016	*	*							7
1	167411384	rs864537	*								8
1	173349725	rs2105325	*	*	*	*	*	*	*	*	3
1	198700442	rs10919563	*	*							5
1	198704294	rs7540378	*								6
2	30449594	rs10175798	*	*			*	*			3
2	42080624	rs4305317	*	*	*	*					9
2	61124850	rs34695944	*				*				3
2	61136129	rs13031237	*	*			*	*			10

2	61164331	rs13017599	*				*				10
2	65556324	rs6546146	*	*							1
2	65595586	rs934734	*	*			*	*			7
2	65598300	rs1858037	*				*	*			3
2	100672692	rs10209110	*	*			*	*			1
2	100806940	rs11676922	*				*				7
2	100825367	rs9653442	*	*			*	*			3
2	100835734	rs10865035	*								11
2	111607832	rs6732565	*	*	*	*	*	*	*	*	3
2	191933254	rs13426947	*	*			*	*			1
2	191943742	rs11889341	*				*				3
2	191964633	rs7574865	*	*	*	*	*	*	*	*	12
2	191969879	rs10181656	*								13
2	202154397	rs6715284	*	*			*	*			3
2	204610396	rs1980422	*	*	*	*	*	*	*	*	3
2	204693876	rs231735	*	*			*	*			10
2	204732714	rs231775	*	*							6
2	204738919	rs3087243	*	*	*	*	*	*	*	*	3
2	204742934	rs11571302	*				*				1
3	17047032	rs4452313	*	*			*	*	*	*	3
3	17072997	rs4535211	*	*							11
3	27764623	rs3806624	*	*			*	*	*	*	3
3	58183636	rs35677470	*	*							1
3	58302935	rs73081554	*	*			*	*			3
3	58556841	rs13315591	*	*			*	*	*	*	11
3	136402060	rs9826828	*	*			*	*			3
4	10727357	rs13142500	*	*							3
4	26085511	rs10517086	*	*							14
4	26090862	rs932036	*				*				1
4	26108197	rs874040	*				*	*			11
4	26120001	rs11933540	*				*				3
4	48220839	rs2664035	*	*			*	*			3
4	123218313	rs13119723	*	*							15
4	123399491	rs45475795	*	*							3
5	55438580	rs6859219	*				*				11
5	55440730	rs71624119	*	*	*	*	*	*	*	*	1
5	55444683	rs7731626	*	*			*	*			3

5	102596720	rs26232	*	*	*	*	*	*	*	*	7
5	102597292	rs39984	*								1
5	102608924	rs2561477	*				*				3
6	426155	rs9378815	*	*							3
6	434364	rs9328192	*	*	*	*					16
6	28169241	rs13195291	*				*				17
6	28191288	rs35656932	*				*				17
6	28201531	rs13204012	*								17
6	28214698	rs17720293	*				*				17
6	28225311	rs13208096	*				*				17
6	28238059	rs67998226	*	*	*	*	*	*	*	*	17
6	32663851	rs6457617	*	*	*	*	*	*	*	*	4
6	32663999	rs6457620	*				*				18
6	36355654	rs2234067	*	*	*		*	*			3
6	44233921	rs2233424	*	*	*		*	*			3
6	90976768	rs72928038	*	*	*	*	*	*	*	*	19
6	106568034	rs548234	*	*			*	*			5
6	106667535	rs9372120	*	*	*		*	*	*		3
6	137973068	rs2327832	*	*	*	*	*	*	*	*	8
6	138002637	rs10499194	*	*	*	*	*	*	*	*	20
6	138005515	rs17264332	*				*				3
6	138006504	rs6920220	*				*				1
6	138195151	rs5029937	*								7
6	138196066	rs2230926	*				*				6
6	138227364	rs7752903	*	*	*	*	*	*	*	*	3
6	159482521	rs394581	*	*	*		*	*			11
6	159489791	rs212389	*	*	*	*					8
6	159496713	rs629326	*								1
6	159506600	rs2451258	*				*	*	*	*	3
6	167534290	rs3093023	*				*				11
6	167537754	rs59466457	*	*			*	*			1
6	167540842	rs1571878	*	*	*	*	*	*			3
7	28174986	rs67250450	*	*			*	*			3
7	45899359	rs6972219	*								9
7	45901549	rs2173035	*								9
7	45903807	rs6956740	*	*							9
7	92236829	rs4272	*	*			*	*			3

7	92246744	rs42041	*	*							2
7	128580042	rs3778753	*	*			*	*			3
7	128580680	rs3807306	*								1
7	128594183	rs10488631	*	*			*	*			11
8	11341880	rs2736337	*								3
8	11343973	rs2736340	*	*	*	*	*	*	*	*	10
8	11345545	rs4840565	*								1
8	81095395	rs998731	*	*			*	*			3
8	102463602	rs678347	*	*			*	*			3
8	129542100	rs1516971	*	*	*	*	*	*	*	*	3
8	129567181	rs6651252	*								19
9	34710260	rs2812378	*	*			*	*			1
9	34710338	rs11574914	*				*				3
9	34743681	rs951005	*	*			*	*			11
9	123636121	rs10985070	*				*				3
9	123640500	rs1953126	*	*			*	*			8
9	123671520	rs2239657	*								6
9	123690239	rs3761847	*	*			*	*			21
9	123695282	rs10739580	*								1
10	6098949	rs706778	*	*			*	*			3
10	6108340	rs10795791	*								1
10	8095340	rs2275806	*	*			*	*			1
10	8104722	rs3824660	*	*			*	*			3
10	9049253	rs12413578	*	*							3
10	31415106	rs793108	*	*							3
10	63779871	rs71508903	*	*			*	*			3
10	63800004	rs12764378	*	*			*	*			1
11	36501787	rs331463	*	*	*	*					3
11	36525293	rs540386	*								5
11	60906450	rs508970	*								3
11	60909581	rs595158	*	*			*	*			1
11	61595564	rs968567	*	*			*	*			3
11	95311422	rs4409785	*	*	*	*	*	*			3
11	107967350	rs138193887	*	*			*	*			3
11	118611781	rs10892279	*				*				8
11	118729391	rs10790268	*	*	*	*	*	*			3
11	118741842	rs4938573	*								1

11	128496952	rs73013527	*	*							3
12	56394954	rs773125	*	*	*	*	*	*	*	*	3
12	58108052	rs1633360	*	*	*						3
12	111833788	rs10774624	*	*	*	*	*	*	*	*	3
12	111884608	rs3184504	*								15
13	40334852	rs9603612	*				*				22
13	40350912	rs7993214	*								19
13	40368069	rs9603616	*	*			*	*			3
14	68753593	rs911263	*				*				19
14	68760141	rs1950897	*	*			*	*			3
14	70541026	rs17175346	*	*			*	*			14
14	75960536	rs7155603	*	*							7
15	38828140	rs8043085	*	*			*	*			1
15	38834033	rs8032939	*				*				3
15	69991417	rs8026898	*	*	*	*	*	*	*	*	3
16	11839326	rs4780401	*	*			*	*			3
16	86019087	rs13330176	*	*			*	*			3
17	38031857	rs59716545	*				*				3
17	38040763	rs2872507	*	*			*	*			1
17	38043649	rs12936409	*				*				1
18	12857758	rs62097857	*	*							19
18	12877060	rs7234029	*	*			*	*			11
18	12881361	rs8083786	*				*				3
19	10463118	rs34536443	*	*	*	*	*	*	*	*	3
19	10771941	rs147622113	*	*	*		*	*	*		3
20	44734310	rs6032662	*				*				1
20	44747947	rs4810485	*	*			*	*			7
20	44749251	rs4239702	*				*				3
21	34764288	rs73194058	*	*			*	*			3
21	35911599	rs2834512	*	*			*	*			1
21	35928240	rs147868091	*								3
21	36715761	rs9979383	*	*			*	*			1
21	36738242	rs8133843	*				*				3
21	42511918	rs2837960	*	*							4
21	43836186	rs11203203	*	*							15
21	43855067	rs1893592	*	*			*	*			3
22	21979096	rs11089637	*	*							3

22	37551607	rs743777	*	*							4
22	39747671	rs909685	*	*			*	*			3

References

1. Eyre S, Bowes J, Diogo D, Lee A, Barton A, Martin P, et al. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. Nat. Genet. 2012;44:1336-40.
2. Raychaudhuri S, Remmers EF, Lee AT, Hackett R, Guiducci C, Burtt NP, et al. Common variants at cd40 and other loci confer risk of rheumatoid arthritis. Nat. Genet. 2008;40:1216-23.
3. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 2014;506:376-81.
4. Wellcome Trust Case Control C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661-78.
5. Raychaudhuri S, Thomson BP, Remmers EF, Eyre S, Hinks A, Guiducci C, et al. Genetic variants at cd28, prdm1 and cd2/cd58 are associated with rheumatoid arthritis risk. Nat. Genet. 2009;41:1313-8.
6. Diogo D, Kurreeman F, Stahl EA, Liao KP, Gupta N, Greenberg JD, et al. Rare, low-frequency, and common variants in the protein-coding sequence of biological candidate genes from gwass contribute to risk of rheumatoid arthritis. Am J Hum Genet 2013;92:15-27.
7. Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat. Genet. 2010;42:508-14.
8. Zhernakova A, Stahl EA, Trynka G, Raychaudhuri S, Festen EA, Franke L, et al. Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-hla shared loci. PLoS Genet 2011;7:e1002004.
9. Padyukov L, Seielstad M, Ong RT, Ding B, Ronnelid J, Seddighzadeh M, et al. A genome-wide association study suggests contrasting associations in acpa-positive versus acpa-negative rheumatoid arthritis. Ann Rheum Dis 2011;70:259-65.
10. Gregersen PK, Amos CI, Lee AT, Lu Y, Remmers EF, Kastner DL, et al. Rel, encoding a member of the nf-kappab family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. Nat. Genet. 2009;41:820-3.
11. Cobb JE, Plant D, Flynn E, Tadjeddine M, Dieude P, Cornelis F, et al. Identification of the tyrosine-protein phosphatase non-receptor type 2 as a rheumatoid arthritis susceptibility locus in europeans. PloS one 2013;8:e66456.
12. Remmers EF, Plenge RM, Lee AT, Graham RR, Hom G, Behrens TW, et al. Stat4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. N Engl J Med 2007;357:977-86.
13. Viatte S, Massey J, Bowes J, Duffus K, arc OC, Eyre S, et al. Replication of associations of genetic loci outside the hla region with susceptibility to anti-cyclic citrullinated peptide-negative rheumatoid arthritis. Arthritis Rheumatol 2016;68:1603-13.
14. Julia A, Gonzalez I, Fernandez-Nebro A, Blanco F, Rodriguez L, Gonzalez A, et al. A genome-wide association study identifies slc8a3 as a susceptibility locus for acpa-positive rheumatoid arthritis. Rheumatology (Oxford) 2016;55:1106-11.

15. Kurreeman FA, Stahl EA, Okada Y, Liao K, Diogo D, Raychaudhuri S, et al. Use of a multiethnic approach to identify rheumatoid-arthritis-susceptibility loci, 1p36 and 17q12. Am J Hum Genet 2012;90:524-32.
16. Lopez-Isac E, Martin JE, Assassi S, Simeon CP, Carreira P, Ortego-Centeno N, et al. Brief report: Irf4 newly identified as a common susceptibility locus for systemic sclerosis and rheumatoid arthritis in a cross-disease meta-analysis of genome-wide association studies. Arthritis Rheumatol 2016;68:2338-44.
17. Xie G, Lu Y, Sun Y, Zhang SS, Keystone EC, Gregersen PK, et al. Identification of the nf-kappab activating protein-like locus as a risk locus for rheumatoid arthritis. Ann Rheum Dis 2013;72:1249-54.
18. Kurreeman F, Liao K, Chibnik L, Hickey B, Stahl E, Gainer V, et al. Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. Am J Hum Genet 2011;88:57-69.
19. McAllister K, Yarwood A, Bowes J, Orozco G, Viatte S, Diogo D, et al. Identification of bach2 and rad51b as rheumatoid arthritis susceptibility loci in a meta-analysis of genome-wide data. Arthritis Rheum 2013;65:3058-62.
20. Plenge RM, Cotsapas C, Davies L, Price AL, de Bakker PI, Maller J, et al. Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. Nat. Genet 2007;39:1477-82.
21. Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, Ding B, et al. Traf1–c5 as a risk locus for rheumatoid arthritis — a genomewide study. N Engl J Med 2007;357:1199-209.
22. Marquez A, Vidal-Bralo L, Rodriguez-Rodriguez L, Gonzalez-Gay MA, Balsa A, Gonzalez-Alvaro I, et al. A combined large-scale meta-analysis identifies cog6 as a novel shared risk locus for rheumatoid arthritis and systemic lupus erythematosus. Ann Rheum Dis 2017;76:289-94.

Table S2. Unadjusted logistic regression models using different standardized weighted genetic risk scores for rheumatoid arthritis[a]

Model	Number of SNPs[b]	swGRS[c] OR(95%CI)	AUC (95%CI)	AIC	BIC
A1[d]	115	1.64 (1.50-1.79)*	0.63 (0.61-0.66)	5594.77	5612.85
A2[d]	36	1.78 (1.63-1.95)*	0.66 (0.63-0.68)	5550.86	5568.93
A3[d]	30	1.83 (1.68-2.00)*	0.67 (0.64-0.69)	5533.17	5551.24
B1[d]	88	1.67 (1.53-1.83)*	0.64 (0.61-0.66)	5586.98	5605.06
B2[d]	29	1.82 (1.67-1.99)*	0.67 (0.64-0.69)	5538.93	5557.00
B3[d]	27	1.85 (1.69-2.02)*	0.67 (0.65-0.70)	5529.20	5547.27

[a]All models included 489 RA cases and 61,584 controls

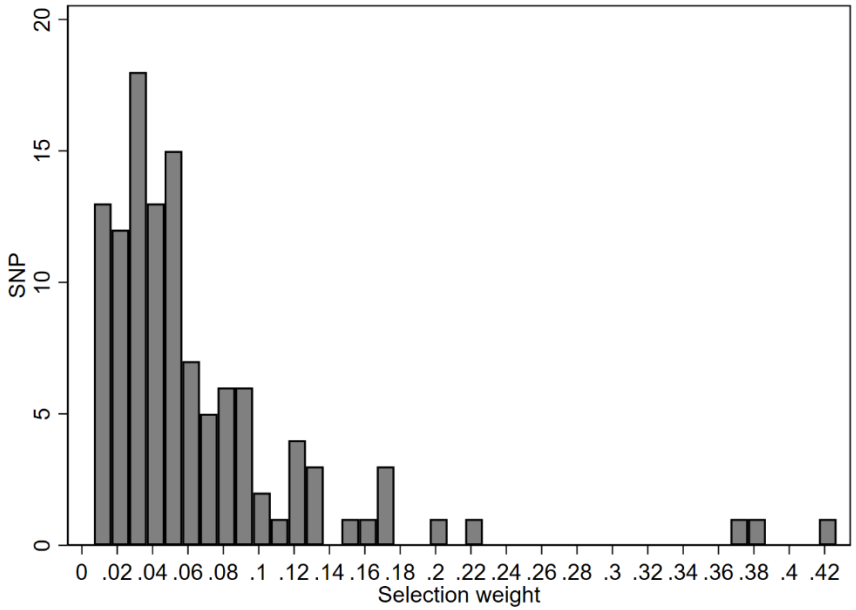
[b]Indicates number of SNPs included in each swGRS

[c]Weighted genetic risk scores standardized using the corresponding population standard deviation

[d]**A** models were based on selection of SNPs with reported $p <= 5\times10^{-6}$. **B** models were based on $p <= 5\times10^{-8}$ in association studies on recent European Caucasian populations. Model numbers indicate selection based on: 1) the largest selection of SNPs showing low linkage disequilibrium ($r^2<0.8$) on each chromosome; 2) additionally showing non-zero coefficients using Lasso regression; 3) additionally showing positive coefficients in Lasso regression.

*$p<0.001$

Figure S1: SNP selection weights



Selection weight for each SNP (n=115) refers to the product of the natural logarithm of the risk allele odds ratio and risk allele frequency reported in previously published association studies. SNPs were selected based on the largest number of non-linked SNPs (r^2 for linkage disequilibrium<0.8) reported with p-value<5×10^{-6} in previous studies (listed in Table S1).

Table S3: Logistic regression models for standardized weighted genetic risk scores constructed based on selection weights[a]

Model	Selection Weight threshold[a]	Number of SNPs[a]	OR(95%CI) Unadjusted[b]	AUC(95%CI) Unadjusted[b]	AIC Unadjusted[b]	BIC Unadjusted[b]	OR(95%CI) Adjusted[b]	AUC(95%CI) Adjusted[b]	AIC Adjusted[b]	BIC Adjusted[b]
C1	>0.36	3	1.51 (1.38- 1.65)	0.61 (0.59-0.63)	5635.77	5653.84	1.52 (1.39- 1.66)	0.73 (0.71-0.75)	5389.63	5443.85
C2	>0.18	5	1.52 (1.39- 1.66)	0.62 (0.59-0.64)	5632.59	5650.66	1.53 (1.40- 1.68)	0.73 (0.71-0.75)	5386.84	5441.05
C3	>0.14	10	1.51 (1.37- 1.65)	0.62 (0.59-0.64)	5635.80	5653.87	1.51 (1.38- 1.66)	0.73 (0.71-0.75)	5390.95	5445.16
C4	>0.1	19	1.50 (1.37- 1.64)	0.62 (0.59-0.64)	5638.26	5656.33	1.51 (1.38- 1.650)	0.73 (0.71-0.75)	5392.83	5447.05
C5	>0.07	34	1.54 (1.41- 1.69)	0.62 (0.60-0.65)	5626.16	5644.23	1.55 (1.42- 1.70)	0.73 (0.71-0.75)	5380.91	5435.12
C6	>0.06	42	1.63 (1.50- 1.79)	0.64 (0.61-0.66)	5597.78	5615.85	1.64 (1.50- 1.79)	0.74 (0.72-0.76)	5352.85	5407.06
C7	>0.055	47	1.60 (1.46- 1.75)	0.63 (0.61-0.66)	5608.36	5626.43	1.61 (1.470- 1.76)	0.74 (0.72-0.76)	5363.04	5417.26
C8	>0.05	55	1.60 (1.47- 1.75)	0.63 (0.61-0.66)	5607.15	5625.22	1.61 (1.47- 1.76)	0.74 (0.72-0.76)	5362.23	5416.44
C9	>0.03	87	1.63 (1.49- 1.78)	0.63 (0.61-0.66)	5598.20	5616.27	1.64 (1.50- 1.80)	0.74 (0.72-0.76)	5351.64	5405.85

C10	>0.02	99	1.65 (1.51- 1.80)	0.64 (0.61-0.66)	5593.67	5611.75	1.66 (1.52- 1.81)	0.74 (0.72-0.76)	5347.38	5401.59
C11	>0	115	1.64 (1.50- 1.79)	0.63 (0.61-0.66)	5594.77	5612.85	1.65 (1.51- 1.81)	0.74 (0.72-0.76)	5348.50	5402.72

[a]All models included 489 RA cases and 61,584 controls. Selection weight for each SNP was calculated using the product of risk allele frequency and risk allele odds ratio reported in previously published association studies. After ranking the SNPs by their selection weights, arbitrary thresholds were used to select for SNPs, thus combining information on higher effect sizes and prevalence (Figure S1). Weighted genetic risk scores were standardized using the corresponding population standard deviation.
[b]Unadjusted models include only the swGRS; adjusted models included additional variables gender, age, and ever smoking.

Table S4: Addition of Shared Epitope variable to the best-fitting logistic regression model[a]

Model	Number of participants	SNP number	OR (95%CI)	AUC	AIC	BIC
D1 swGRS27[b]	1720	27	1.74 (1.54-1.96)	0.66 (0.62-0.69)	1657.26	1668.16
D2 swGRS26[c]	1720	26	1.42 (1.26-1.59)	0.60 (0.57-0.63)	1705.90	1716.80
D3 SE only	1720	1	1.48 (1.31-1.67)	0.60 (0.57-0.63)	1699.80	1710.70
D4 swGRS26[c] & SE	1720	26+1	swGRS26: 1.43 (1.27-1.60) SE: 1.50 (1.32-1.69)	0.65 (0.62-0.68)	1664.76	1618.11

[a]**D** models include 350 RA cases and 1,370 controls, where information regarding the Shared Epitope (SE) was available. The SE variable (carrier vs. non-carrier) was standardized using the population standard deviation to permit direct comparison among models. The models were not adjusted for non-genetic variables.

[b]swGRS27: best-fitting weighted genetic risk score (Model B3 in Table 2), standardized using population standard deviation.

[c]swGRS26 is similar to swGRS27 except that rs6457617 that marks the common SE variant HLA-DRB1*0401 was removed.