Multiomics analysis of rheumatoid arthritis yields sequence variants that have large effects on risk of the seropositive subset

Saedis Saevardsdottir,1,2,3,4 Lilja Stefansdottir,1 Patrick Sulem,1 Gudmar Thorleifsson,1 Egil Ferkistingstad,1 Gudrun Rutsdottir,1 Bente Glintborg2,5,6 Helga Westerlind2,6,7 Gerdur Grondal,3,4,7 Isabella C Loft,8 Signe Bek Sorensen,9 Benedicte A Lie,1,10 Mikael Brink,12 Lisbeth Åresteg,12 Asgeir Orn Arnthorsson,1 Eva Baeklund,13 Karina Banasik,14 Steffen Bank,9 Lena I Bjorkman,15 Torkell Ellingsen,1,16,17 Christian Erikstrup,18 Oleksandr Frej,19,20,21 Inger Gjertsson2,22 Daniel F Gudbjartsson,1,23 Sigurjon A Gudjonsson,1 Gisli H Halldorsson,1,23 Oliver Hendricks,24,25 Jan Hillert,26 Estrid Hodgall,27 Søren Jacobsen2,6,28 Dorte Vendelbo Jensen,29 Helgi Jonsson,3,4,30 Alf Kastbom3,30 Ingrid Kockum,26 Salome Kristensen1,31,32 Helga Kristjansdottir,7 Margit H Larsen,33 Asta Linauskas,3,34 Ellen-Margrethe Hauge,35,36 Anne G Loft,35,36 Bjorn R Ludviksson,3,37 Sigrun H Lund,1 Thorstein Markussen,1,3 Gisli Masson,1 Pall Melsted,1,23 Kristjan H S Moore,1 Heidi Munk1,16,17 Kaspar R Nielsen,38 Gudmundur L Norddahl,1 Asmundur Oddsson,1 Thorunn A Olafsdottir,1,3 Pall I Olason,1 Tomas Olsson,26 Sisse Rye Ostrowski,6,33 Kim Hørslev-Petersen,24 Solvi Rognavaldsson,1 Helga Sanner,39,40 Gilad N Silberberg,41 Hreinn Stefansson,1 Erik Sørensen,33 Inge J Sørensen,28 Carl Turesson1,42 Thomas Bergman,2 Lars Alfredsson,26,43 Tore K Kvien,43,44 Sören Brunak,14 Kristján Steinsson,7 Vibeke Andersen1,9,16,46 Ole A Andreassen,19,20 Solbritt Rantapää-Dahlqvist1,12 Merete Lund Hetland5,6 Lars Klareskog1,41 Johan Asling,12 Leonid Padyukov,41 Ole BV Pedersen,8 Unnur Thorsteinsdottir,1,3 Ingileif Jonsdottir,1,3,37 Kari Stefansson,1,3 Members of the DBDS Genomic Consortium, The Danish RA Genetics Working Group, The Swedish Rheumatology Quality Register Biobank Study Group (SRQb)

Handling editor Josef S Smolen

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/annrheumdis-2021-221754).

For numbered affiliations see end of article.

Correspondence to Professor Saedis Saevardsdottir, deCODE Genetics/Amgen, Reykjavik, Iceland; saedis.saevardsdottir@decode.is and Professor Kari Stefansson, deCODE genetics/Amgen, Reykjavik, Iceland; kstefans@decode.is

Received 26 October 2021
Accepted 4 April 2022


ABSTRACT

Objectives To find causal genes for rheumatoid arthritis (RA) and its seropositive (RF and/or ACPA positive) and seronegative subsets.

Methods We performed a genome-wide association study (GWAS) of 31 313 RA cases (68% seropositive) and ~1 million controls from Northwestern Europe. We searched for causal genes outside the HLA-locus through effect on coding, mRNA expression in several tissues and/or levels of plasma proteins (SomaScan) and did network analysis (Qiagen).

Results We found 25 sequence variants for RA overall, 33 for seropositive and 2 for seronegative RA, altogether 37 sequence variants at 34 non-HLA loci, of which 15 are novel. Genomic, transcriptomic and proteomic analysis of these yielded 25 causal genes in seropositive RA and additional two overall. Most encode proteins in the network of interferon-alpha/beta and IL-12/23 that signal through the JAK/STAT-pathway. Highlighting those with largest effect on seropositive RA, a rare missense variant in STAT4 (rs140675301-A) that is independent of reported non-coding STAT4-variants, increases the risk of seropositive RA 2.27-fold \( (p=2.1\times 10^{-8}) \), more than the rs2476601-A missense variant in PTPN22 \( (OR=1.59, p=1.3\times 10^{-160}) \). STAT4 rs140675301-A replaces hydrophilic glutamic acid with hydrophobic valine (Glu128Val) in a conserved, surface-exposed loop. A stop-mutation (rs76428106-G) in FLT3 increases seropositive RA risk \( (OR=1.35, p=6.6\times 10^{-11}) \).

Key messages

What is already known about this subject?

► Although many genetic risk loci have been identified in rheumatoid arthritis (RA) overall, there are limited data available on the seropositive and seronegative subsets. Furthermore, most reported RA associations outside the HLA-locus are with common non-coding variants with low risk, which lack a compelling candidate gene mediating the effect on RA.
rs12720356-C, rs53018800-A, latter two novel) associate with decreased risk of seropositive RA (ORs=0.63–0.87, p=10^{-9}–10^{-27}) and decreased plasma levels of interferon-alpha/beta receptor 1 that signals through TYK2/JAK1/STAT4.

Conclusion Sequence variants pointing to causal genes in the JAK/STAT pathway have largest effect on seropositive RA, while associations with seronegative RA remain scarce.

Rheumatoid arthritis (RA) is a heterogeneous clinical syndrome that affects around 0.5%–1% of the general population. It is characterised by inflammatory polyarthritis and progressive joint damage if insufficiently treated. RA is divided into seropositive and seronegative subsets of RA, using the largest GWAS study population to date in RA (31 313 cases and ~1 million controls) from six countries in Northwestern Europe and searched for candidate causal genes through a genomic, transcriptomic and proteomic analysis.

METHODS

Study populations

Cases with RA were diagnosed by rheumatologists and/or captured through the nationwide Scandinavian rheumatology quality registries and/or the 10th revision of the International Statistical Classification of Diseases (ICD-10) code-based registration of all inpatient and outpatient healthcare visits (see four-digit based ICD-10 codes in table 1). If available, RF and anti-CCP measurement were used to define the seropositive/seronegative RA subsets, according to classification criteria. An overview of the study populations is provided in table 1. In the study populations from Iceland (3613 cases and 341 788 controls), UK Biobank (5798 cases and 402 767 controls of self-reported white British ancestry, confirmed by genetic analysis) and FinnGen (https://www.finngen.fi/en/access_results version R4: 4701 cases and 125 923 controls), RA cases were compared with the remaining non-RA individuals, with the Icelandic study covering a large part of the Icelandic population and the latter two being nationwide genetic cohort studies. From Sweden, we included: (1) the population-based EIRA case-control study (www.eirawedsten.se) with 3436 newly diagnosed cases and 3058 controls matched for age, sex and geographical area from mid and Southern parts of Sweden. In addition, we included 7488 controls from the parallel Swedish EIMS study (ki.se/imm/eims-epidemiologisk-undersokning-av-riskfaktorer-for-multipel-skleros); (2) the RA cohort from Umea (n=1933) and 1156 controls from Umea biobank, matched for age and sex (www.umu.se/en/biobank-research-unit); and (3) the Swedish Rheumatology Quality Register Biobank (n=3287, www.srnu.web). From Denmark, RA cases were identified in four study populations: (1) Danish Biomarker Protocol (n=2544 with samples in the Danish Rheumatological Biobank and clinical data in the Danish Rheumatology Quality Register, DANBIO) (2) the Copenhagen Hospital Biobank (n=3282), (3) the TARCID cohort (n=1826) and (4) the nationwide Danish Blood Donor Study (DBDS; 10 RA cases). Controls for these 7662 cases were age-matched and sex-matched non-RA individuals from DBDS (n=86 964).

From Norway, 881 RA cases from the Oslo RA cohort and 28 517 population-based controls from the Norwegian Mother, Father and Child Cohort Study were included. Patients were involved in the design and conduct of several of the studies that are included in this report.

Genotyping and multiomics analyses

For a detailed methodological description, see online supplemental information 2. In short, genotyping of all cohorts except UK Biobank and FinnGen was performed at deCODE genetics using the Illumina technology, and the sequence variants for imputation were identified through whole-genome sequencing of 67 645 individuals.

We used logistic regression to test the association of ~64 million sequence variants with RA overall, the seropositive and seronegative RA subtypes with ACPA-negative RA only (n=1922) that identified two genome-wide significant signals. Here, we searched for sequence variants outside the HLA-locus affecting the risk of RA overall, the seropositive and/or seronegative subsets of RA, using the largest GWAS study population to date in RA (31 313 cases and ~1 million controls) from six countries in Northwestern Europe and searched for candidate causal genes through a genomic, transcriptomic and proteomic analysis.
the seronegative subset. Sequence variants were split into five classes based on their genome annotation, and the significance threshold for each class was based on the number of variants in that class, thereby adjusting for all ~64 million variants tested, maintaining an unadjusted significance threshold of \(8 \times 10^{-10}\). The primary signal at each genomic locus has the lowest Bonferroni-adjusted \(p\) value. Conditional analysis was used to search for possible secondary signals (<500 kB from the primary signal, excluding HLA-loci). We tested whether primary and secondary signals were in strong linkage disequilibrium (\(R^2 > 0.8\)) with top cis-eQTL variants for genes expressed in various tissues (online supplemental tables 5 and 6), and/or with levels of 4789 proteins in plasma (pQTL, SomaScan, Soma-logic) in 35 559 Icelanders (online supplemental table 7).

We used the Ingenuity Pathway Analysis software (QIAGEN Inc) to evaluate whether there is experimental evidence for direct or indirect interaction between the proteins coded by candidate causal genes, supporting biological connection.

**RESULTS**

**Genome-wide association study**

Of the 31 313 RA cases, 26 534 (84.7%) had information on serological status. Of these, 18 019 (67.9%) were seropositive and 8 515 (32.1%) seronegative (table 1).

In separate meta-analyses of RA overall and the seropositive and seronegative RA subsets, we found in total 37 sequence variants at 34 non-HLA loci (online supplemental figure 1a–c), as summarised in table 2. Thus, we identified 25 lead signals for RA overall (online supplemental table 2), 33 for seropositive and 2 for seronegative RA (online supplemental table 3). When we searched for novel sequence variants, we adjusted for 82 independent sequence variants previously reported to associate with RA (\(p<5 \times 10^{-8}\) in the largest meta-analyses to date), and 15 of the 37 sequence variants are previously unreported. The 15 novel associations are at 12 loci and six of those loci are previously unreported. Little heterogeneity was observed between the study populations (see online supplemental tables 2 and 3 (\(P_{Het}\)) and online supplemental figure 4 (average effect)).

**Replication of previously reported signals**

We replicated 53 of the 82 previously reported variants (online supplemental table 1, correcting for multiple testing, \(p\) value threshold=0.05/82 variants /3 phenotypes=2.03×10^{-4}). However, only 36 of the 82 variants were previously reported to be genome-wide significant in Europeans, and we replicated 34 of these 36 variants (94%).

**Comparison of RA subsets**

The heritability estimates (total observed scale \(h^2\)) were higher for seropositive RA (0.19 (0.022)) than for seronegative RA (0.099 (0.019)). For a substantial proportion of the RA-associ-
ated sequence variants, their effect was greater on seropositive RA than seronegative RA risk (table 2, figure 1). However, the genetic correlation between seropositive and seronegative RA was high (rg 0.87, SE 0.13, \(p=4.5 \times 10^{-12}\) (online supplemental table 9).

**Genomic, transcriptomic and proteomic analysis of lead signals**

We searched for candidate causal genes with an omics approach (figure 2A) and evaluated the effect of lead signals (or correlated variants, \(R^2 > 0.8\)) on amino acid sequence (online supplemental tables 2–4), mRNA expression (cis-eQTL (online supplemental tables 5 and 6) and/or plasma levels of proteins (pQTL (online supplemental table 7). This yielded a total of 27 candidate causal genes in RA overall and/or its subsets.

**Seropositive RA**

Twenty-four of the 33 lead signals in seropositive RA pointed to 25 candidate causal genes, as shown in figure 2B ranked by effect. The one with the largest effect is a rare (MAF=0.14%) missense variant in the STAT4 gene (rs140675301-CA, Glu128Val) that associates with 2.27-fold increased risk (\(p=2.1 \times 10^{-8}\), table 2 and figure 2B). Rs140675301-A is the first coding variant identified at the STAT4 locus that associates with RA and has not been reported in any disease before. This signal is independent (online supplemental table 8) of the common lead STAT4 intronic variant (rs4853458-A), which is strongly correlated (\(R^2 = 1\)) with other intronic variants in STAT4, previously reported to associate with RA (\(22, 23\) (figure 3A) and online supplemental table 1). STAT4 contains six domains that have different functions, and the rare missense rs140675301-A variant leads to an amino acid change from negatively charged, hydrophilic, glutamic acid to non-polar hydrophobic valine at position 128 (Glu128Val) in a loop on the surface of the protein (figure 3B), between the N-terminal domain and the helical coiled coil domain. The coiled coil domain provides a carbonised hydrophilic surface that binds to regulatory factors. The amino acid sequence and secondary structure of the loop is highly conserved between species (figure 3C) and within the family of STAT proteins, indicating its importance for the function of STAT4. Tetramer formation of STAT at DNA binding sites is necessary for full transcriptional activation of many of its target genes, and STAT without the N-terminal domain cannot form tetramers.

The second largest effect on the risk of seropositive RA had the well-known missense variant rs2476601-A in the PTPN22 gene, followed by a novel missense variant in the TYK2 gene (rs35018800-A, Ala928Val), encoding tyrosine kinase 2, which is a member of the JAK/STAT-pathway like STAT4. This rare (MAF=0.60%) missense variant in TYK2 conferred reduced risk
Table 2  Sequence variants outside the HLA locus that associate with RA overall, seropositive (rheumatoid factor and/or anti-CCP antibody positive) and/or seronegative RA in GWAS meta-analysis within six Northwestern-European countries (table 1). Association results are shown for the lead signals for all three RA groups, and the heterogeneity between the seropositive and seronegative subsets.* Effect alleles with novel associations are marked with.*

<table>
<thead>
<tr>
<th>Chr</th>
<th>Position</th>
<th>Effect allele*</th>
<th>Gene</th>
<th>Annotation</th>
<th>OR (95% CI)</th>
<th>OR (95% CI)</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>2 800 059</td>
<td>rs897628-T*</td>
<td>ITCO4</td>
<td>Missense</td>
<td>0.90 (3.3E-16)</td>
<td>0.98 (0.18)</td>
<td>0.94 (1.9E-10)</td>
</tr>
<tr>
<td>chr1</td>
<td>113 834 946</td>
<td>rs476601-A</td>
<td>PTPN22</td>
<td>Missense</td>
<td>1.59 (1.3E-16)</td>
<td>1.29 (2.9E-27)</td>
<td>1.41 (3.9E-144)</td>
</tr>
<tr>
<td>chr1</td>
<td>161 506 414</td>
<td>rs427 3973*T</td>
<td>FCGR2A</td>
<td></td>
<td>1.11 (2.0E-0)</td>
<td>1.02 (0.55)</td>
<td>1.07 (3.3E-06)</td>
</tr>
<tr>
<td>chr2</td>
<td>60 881 694</td>
<td>rs675 7466-A</td>
<td>RELP</td>
<td>UTR</td>
<td>1.08 (6.2E-0)</td>
<td>1.01 (0.57)</td>
<td>1.05 (3.4E-07)</td>
</tr>
<tr>
<td>chr2</td>
<td>111 119 036</td>
<td>rs283 1646-C*</td>
<td>BCL11</td>
<td>Upstream gene</td>
<td>1.14 (2.5E-10)</td>
<td>1.01 (0.75)</td>
<td>1.10 (7.5E-09)</td>
</tr>
<tr>
<td>chr2</td>
<td>191 073 180</td>
<td>rs406 7530-A*</td>
<td>STAT4</td>
<td>Missense</td>
<td>2.27 (2.1E-09)</td>
<td>1.23 (3.4E-01)</td>
<td>1.63 (3.9E-06)</td>
</tr>
<tr>
<td>chr2</td>
<td>191 073 180</td>
<td>rs406 7530-A*</td>
<td>STAT4</td>
<td>Missense</td>
<td>1.11 (5.2E-0)</td>
<td>1.10 (1.1E-06)</td>
<td>1.10 (2.1E-19)</td>
</tr>
<tr>
<td>chr2</td>
<td>20 380 280</td>
<td>rs157 1219-C</td>
<td>CTLA4</td>
<td>Regulatory</td>
<td>0.89 (2.2E-20)</td>
<td>0.95 (2.2E-03)</td>
<td>0.92 (4.4E-19)</td>
</tr>
<tr>
<td>chr2</td>
<td>20 380 280</td>
<td>rs157 1219-C</td>
<td>CTLA4</td>
<td>Regulatory</td>
<td>0.89 (2.2E-20)</td>
<td>0.95 (2.2E-03)</td>
<td>0.92 (4.4E-19)</td>
</tr>
<tr>
<td>chr3</td>
<td>38 196 381</td>
<td>rs174 0146-A</td>
<td>IL18R1</td>
<td>Intergenic</td>
<td>0.67 (2.6E-16)</td>
<td>1.03 (0.75)</td>
<td>0.96 (1.2E-06)</td>
</tr>
<tr>
<td>chr5</td>
<td>84 048 156</td>
<td>rs228 972-A</td>
<td>ANKRD55</td>
<td></td>
<td>0.87 (2.6E-0)</td>
<td>0.87 (2.4E-0)</td>
<td>0.88 (1.1E-0)</td>
</tr>
<tr>
<td>chr6</td>
<td>137 678 425</td>
<td>rs159 668-G</td>
<td>SIX1</td>
<td>Regulatory</td>
<td>1.12 (4.3E-10)</td>
<td>1.02 (0.24)</td>
<td>0.99 (1.5E-10)</td>
</tr>
<tr>
<td>chr6</td>
<td>159 085 568</td>
<td>rs240 25-C</td>
<td>CCR6</td>
<td>Regulatory</td>
<td>0.91 (1.6E-12)</td>
<td>0.97 (0.75)</td>
<td>0.96 (1.2E-10)</td>
</tr>
<tr>
<td>chr6</td>
<td>167 127 770</td>
<td>rs909 137-C</td>
<td>CCR6</td>
<td>Regulatory</td>
<td>1.11 (1.8E-18)</td>
<td>1.04 (0.03)</td>
<td>1.07 (7.6E-15)</td>
</tr>
<tr>
<td>chr7</td>
<td>50 313 596</td>
<td>rs303 1719-G*</td>
<td>KLF1</td>
<td>Regulatory</td>
<td>1.07 (6.9E-07)</td>
<td>1.04 (0.04)</td>
<td>1.07 (3.6E-12)</td>
</tr>
<tr>
<td>chr7</td>
<td>128 938 247</td>
<td>rs200 46440-G*</td>
<td>IRF5</td>
<td>Splice donor</td>
<td>0.92 (1.4E-11)</td>
<td>0.94 (1.9E-14)</td>
<td>0.94 (5.1E-13)</td>
</tr>
<tr>
<td>chr8</td>
<td>11 480 078</td>
<td>rs71 871-C</td>
<td>BGLB1</td>
<td>Regulatory</td>
<td>1.09 (1.1E-09)</td>
<td>1.05 (9.1E-03)</td>
<td>1.08 (3.1E-12)</td>
</tr>
<tr>
<td>chr9</td>
<td>33 953 906</td>
<td>rs140 6601-A</td>
<td>AGS2</td>
<td>Regulatory</td>
<td>1.08 (7.4E-10)</td>
<td>1.04 (3.4E-03)</td>
<td>1.05 (9.1E-08)</td>
</tr>
<tr>
<td>chr10</td>
<td>726 556</td>
<td>rs199 7640-C</td>
<td>LOC15 2350-540</td>
<td>Intergenic</td>
<td>1.08 (2.7E-10)</td>
<td>1.04 (0.26)</td>
<td>0.96 (1.7E-14)</td>
</tr>
<tr>
<td>chr11</td>
<td>112 466 904</td>
<td>rs188 504-T</td>
<td>SPDE3</td>
<td>Missense</td>
<td>1.10 (7.6E-16)</td>
<td>1.07 (1.6E-06)</td>
<td>1.08 (1.6E-17)</td>
</tr>
<tr>
<td>chr11</td>
<td>28 029 870</td>
<td>rs642 016-C</td>
<td>IL7R</td>
<td>Intergeric</td>
<td>1.09 (1.2E-11)</td>
<td>1.15 (0.03)</td>
<td>1.23 (1.7E-08)</td>
</tr>
<tr>
<td>chr13</td>
<td>39 768 092</td>
<td>rs100 271-C</td>
<td>COG6</td>
<td>Intergeric</td>
<td>0.92 (3.5E-10)</td>
<td>0.94 (2.1E-04)</td>
<td>0.93 (1.1E-14)</td>
</tr>
<tr>
<td>chr13</td>
<td>69 551 888</td>
<td>rs130 8951-T*</td>
<td>BMD1</td>
<td>Missense</td>
<td>0.92 (3.2E-11)</td>
<td>0.97 (1.9E-03)</td>
<td>0.93 (4.3E-13)</td>
</tr>
<tr>
<td>chr15</td>
<td>85 082 415</td>
<td>rs220 619-C</td>
<td>IL18R1</td>
<td>Intergeric</td>
<td>1.10 (5.2E-11)</td>
<td>1.06 (4.6E-03)</td>
<td>1.07 (9.1E-10)</td>
</tr>
<tr>
<td>chr16</td>
<td>88 981 246</td>
<td>rs204 181-C*</td>
<td>CBFAL3</td>
<td>Upstream gene</td>
<td>0.93 (8.0E-10)</td>
<td>1.00 (9.3E-01)</td>
<td>0.96 (3.1E-15)</td>
</tr>
<tr>
<td>chr17</td>
<td>39 309 216</td>
<td>rs107 8293-C</td>
<td>GSDMB</td>
<td>Splice acceptor</td>
<td>1.07 (1.3E-07)</td>
<td>1.05 (1.3E-07)</td>
<td>1.04 (1.9E-05)</td>
</tr>
<tr>
<td>chr17</td>
<td>10 354 422</td>
<td>rs237 644-C</td>
<td>TYK2</td>
<td>Missense</td>
<td>0.69 (2.7E-27)</td>
<td>0.81 (1.6E-26)</td>
<td>0.75 (2.5E-29)</td>
</tr>
<tr>
<td>chr17</td>
<td>35 340 290</td>
<td>rs129 010-T</td>
<td>TYK2</td>
<td>Missense</td>
<td>0.67 (2.9E-10)</td>
<td>0.90 (7.5E-04)</td>
<td>0.90 (3.8E-10)</td>
</tr>
<tr>
<td>chr21</td>
<td>44 226 819</td>
<td>rs155 819-T*</td>
<td>ILECL2</td>
<td>Regulatory</td>
<td>0.91 (1.6E-09)</td>
<td>0.97 (0.25)</td>
<td>0.95 (1.2E-05)</td>
</tr>
</tbody>
</table>

*Effect variants that are significant after adjustment for previously reported sequence variants (online supplemental table 1). Bold indicates candidate causal genes (summarized in figure 2).

We performed a meta-analysis using logistic regression analysis assuming a multiplicative model, reporting OR and two-sided p values adjusted for year of birth, sex and origin (local or the first 20 principal components for other countries). Variants were split into five classes based on their genome annotation and significance threshold based on the number of variants in each class. The adjusted genome wide thresholds for the five classes were 1.3E-15 for common variants (minor allele frequency (MAF) > 0.5), 2.5E-14 for low frequency variants (0.01 < MAF < 0.5), 1.3E-12 for rare variants (MAF < 0.01), and 1.3E-10 for other low frequency variants (MAF < 0.005). GWAS, genome-wide association study; Phet, a p value for test of heterogeneity between the effects in seropositive and seronegative RA subsets; RA, rheumatoid arthritis.
of seropositive RA (OR=0.63, p=1.4×10^{-11}), independently of a known missense variant in TYK2 (rs34536443-C, Pro1104Ala, MAF 4.3%), which we also found to decrease the risk of RA overall (OR=0.75, p=2.5×10^{-22}), and here, we extend this association to the seropositive RA subset (OR=0.69, p=2.7×10^{-27}; table 2, online supplemental table 3 and online supplemental figure 2). In addition, we identified a common missense variant in TYK2 that independently associated with reduced risk of seropositive RA (rs12720356-C, Ile684Ser, MAF=8.82%, OR=0.87, p=2.3×10^{-9}). Analysis of the plasma proteome (online supplemental table 7) showed that the minor alleles of the variants encoding both Ile684Ser and Pro1104Ala in TYK2 are the only sequence variants that associate in trans with plasma levels of interferon alpha/beta receptor 1 (IFNAR1, Ile684Ser: effect=−0.19 SD, p=7×10^{-25}; Pro1104Ala, effect=−0.13 SD, p=6×10^{-10}). These variants did not associate with levels of any other plasma protein measured. Notably, both the missense variants in TYK2 and STAT4 are predicted to damage the function of the encoded protein (online supplemental table 4).

An intronic variant (rs76428106-C) in the FLT3 gene, encoding another tyrosine kinase receptor that signals through the JAK/STAT-pathway, conferred 35% increase in risk of seropositive RA (p=6.6×10^{-11}). This is in accordance with our previous report, where we discovered this variant in a GWAS on autoimmune thyroid disease and found that it also associated nominally with the risk of seropositive RA (OR=1.41, p=4.3×10^{-9}) and with increased levels of 22 proteins in plasma (trans-pQTL), including the FLT3 ligand (rs76428106-C associated with increased mRNA expression of FLT3 in lung tissue (beta=0.82 SD, p=1.3×10^{-10}, online supplemental table 6).

We performed a network analysis of the 25 seropositive RA candidate genes and found that 18 of them encode proteins that are linked in the same network (online supplemental figure 3), either through direct protein–protein interaction (eg, STAT4-TYK2, PTPN22-IRF5 and FLT3-SH2B3) or indirectly (eg, one affecting the level of another). Other molecules that are central in this network, and directly interact with proteins encoded by the candidate genes, are interferon alpha beta and IL12/IL-23.

Among the other candidate causal genes, we also identified novel loss-of-function variants in genes encoding molecules in this network, although with more modest effect on seropositive RA risk (table 2 and figure 2B). This includes a splice-donor variant in the IRF5 gene (rs2004640-G, OR=0.92, p=1.4×10^{-11}) that encodes interferon regulatory factor 5. IRF5 rs2004640-G association with decreased risk of seropositive RA was independent from previously reported non-coding variants at the IRF5 locus (online supplemental table 1) and rs2004640-G is also associated with decreased mRNA expression of IRF5 in several tissues (online supplemental table 6). Other novel coding variants pointing to putative causal genes were missense variants in ICOSLG (rs11558819-T, OR=0.91, p=1.56×10^{-9}) encoding ICOS ligand and TTC34 (rs897628-T, OR=0.90, p=3.28×10^{-16}). TTC34 encodes tetrameric repeat protein 34 that has an unknown role in the pathogenesis of RA and belongs to another network that includes the remaining seven candidate causal genes for seropositive RA (online supplemental figure 3).

Seronegative RA
Both signals in seronegative RA were also found in seropositive RA and pointed to causal genes: a missense variant rs2476601-A in PTPN22 and intronic variant rs7731626-A in ANKRD55 (table 2 and online supplemental tables 2; 3).
PTPN22 rs2476601-A associated with plasma levels of several proteins (trans-pQTL), and it was the only variant in the genome to affect the levels of these proteins (online supplemental tables 2–4), mRNA expression (cis-eQTL (online supplemental tables 5 and 6)) or levels of proteins in plasma (pQTL (online supplemental table 7)).

RA overall
The lead signals pointing to causal genes in RA overall were also identified in the seropositive subset (table 2), with two
DISCUSSION

In this largest GWAS study on RA to date, we studied both RA overall and the seropositive and seronegative RA subsets and found 37 sequence variants of which 15 were previously unreported. Several of these have large effect on seropositive RA risk, while only two signals were identified in the seronegative subset, both previously reported in RA overall. Through a multomics approach, we identified candidate causal genes for most signals and show that the majority of those associated with seropositive RA are in the interferon alpha/beta and IL-12/23 signalling networks, with largest risk associated with sequence variants in genes encoding proteins in the JAK/STAT pathway.

Novel missense variant in the STAT4 gene (rs140675301-A) confers 2.27-fold increased risk that is higher risk than any previously reported RA association, including the well-known HLA-DRB1 shared epitope and the lead missense variant at the PTPN22 locus. Although the STAT4 locus has been found in genome-wide studies, this is the first STAT4 coding variant found to associate with RA. This coding variant points directly to STAT4 as the causal gene at the locus. It has not been reported for any other disease before, and we found that it leads to an amino acid change in a surface loop of the protein that is highly conserved, thereby underscoring its importance for STAT4 function. STAT4 encodes STAT4, a cytoplasmic transcription factor that regulates gene expression through the JAK/STAT-pathway. It is phosphorylated in response to various cytokines and displacement of the N-terminal and coiled coil domains within the protein structure could interfere with DNA binding, transcriptional activation and/or target selectivity. As highlighted in the network analysis and illustrated in figure 4, both interferon alpha, IL-12 and IL-23, signal through STAT4 via TYK2/JAK1 and TYK2/JAK2. Another RA-associated variant in STAT4 (rs7574865-T, R^2=0.99 to lead intron variant rs4853458-A) increases IL-12-induced IFN-γ production in T cells. STAT4 is expressed at inflammatory sites in activated peripheral blood monocytes, fibroblasts, dendritic cells and macrophages and also in synovial macrophages and dendritic cells from patients with seropositive RA. Furthermore, reduced expression of STAT4 has been observed in RA patients that have responded well to disease-modifying treatment. Thus, STAT4 may have a central role in the inflammatory cascade in joints of RA patients.

Tyrosine kinase 2, encoded by the TYK2 gene, is another key molecule in the JAK/STAT pathway that regulates signal transduction pathways downstream of the receptors for several cytokines, including interferon alpha/beta and IL-23/IL12 as described previously. We found that three independent coding variants in TYK2 associated with 25%–37% reduced risk of seropositive RA, and they associated with lower plasma levels of the IFNAR1 receptor for interferon-alpha/beta. Accordingly, one of the missense variants (Pro1104Ala) is located in the catalytic kinase domain of TYK2 and has previously been shown to reduce signalling through IFNAR1.

TYK2 also mediates the signalling of IL-6, IL-10 and IL-4/IL-13. IL-6 signals through the IL-6 receptor (IL-6R), thereby inducing IL6ST homodimerisation and activation of TYK2/JAK1/2 and STAT3 signalling pathway (figure 4), known to play a role in RA. The intronic variant rs7731626-A in ANKRD55 associated with a reduced risk of both seropositive and seronegative RA and also reduced expression of ANKRD55 and IL6ST. The effect on IL6ST expression and its biological function points to IL6ST as a candidate causal gene at that locus. Accordingly, drugs inhibiting IL-6R are effective in RA.

The FLT3 receptor is another activator of the JAK/STAT pathway that signals through STAT5 (figure 4), and an intronic variant in the FLT3 gene (rs76428106-C) conferred 35% increase in risk of seropositive RA. This confirms a non-genome-wide significant signal in our previous report, in which we identified this variant as a strong risk factor for autoimmune thyroid disease and found that it generates a cryptic splice site, introducing a stop codon in 30% of transcripts that are predicted to encode a truncated protein, lacking its tyrosine kinase domains. FLT3 encodes fms-related tyrosine kinase 3 receptor, a key regulator in the development of monocytes and dendritic cells. The cell-surface receptor is expressed on common dendritic cells and lymphoid/myeloid progenitors that give rise to both classical and plasmacytoid dendritic cells, which produce large amount

---

**Figure 4** The JAK-STAT pathway. The figure and table shows which receptors, JAK and STAT subtypes certain cytokines bind to, highlighting proteins encoded by and/or affected by causal genes in seropositive RA, based on the multomics analysis of sequence variants associated with risk of seropositive RA (shown in bold). Binding of a cytokine to its receptor activates the associated Janus kinases (JAK). The JAK in turn phosphorylates (P) the receptor, which provides a docking for signal transducers and activators of transcription (STATs) and other signalling molecules to bind to the receptor. STATs also become phosphorylated and translocate to the nucleus, where they regulate gene expression.

*Protein targeted by drugs that are registered for RA. **Proteins targeted by drugs registered or in pipeline for other diseases. RA, rheumatoid arthritis.
of interferons when activated. As previously reported, FLT3 rs764281106-C increases plasma levels of the FLT3 ligand, and RA patients have increased levels of FLT3 ligand both in serum and synovial fluid of inflamed joints. FLT3 ligand deficient mice are protected against collagen-induced arthritis, and in a mouse model of collagen-induced arthritis, an oral inhibitor of FLT3/JAK2/c-Fms was found to block signalling through TYK2 and STAT4 and decrease both inflammation and bone resorption.

Yet another variant affecting interferon signalling is a splice-donor variant in the IRF5 (rs2004640-G) gene that encodes interferon regulatory factor 5 and reduced both RA risk and IRF5 expression. IRF5-rs2004640-G has not been reported in GWAS on RA before, although the locus is known, and a tentative association was reported in a meta-analysis of candidate gene studies (4818 cases, p=0.003).

The size and homogeneous background of the study populations, with −64 million sequence variants derived from over 67 thousand whole-genome sequenced individuals, increases the likelihood to detect rare and low-frequency sequence variants that associate with disease. Furthermore, we were able to test their functional relevance through analysis of RNA sequence and plasma proteome. However, it remains to be seen whether the sequence variants associate with RA in populations of another ancestry.

The SNP-based heritability estimate for seropositive RA was the same as in a previous study (0.19), while lower for seronegative RA (0.099) where previous findings are scarce.

In addition to the causal genes highlighted previously, the network analysis illustrated how majority of all candidate causal genes encode proteins in the interferon alpha/beta and IL-12/IL-23 signalling network. Furthermore, we observed a consistent direction of the effect on seropositive RA risk, gene expression and protein levels in plasma, indicating that increased signalling through the JAK/STAT-pathway is central in the inflammatory cascade in seropositive RA. Our findings are in line with the documented effectiveness of IL-6 receptor and JAK inhibitors (baricitinib, tofacitinib, filgotinib and upadacitinib) as well as CRS receptor and cytokine inhibition in seropositive RA.

In summary, through a large genome, transcriptome and proteome analysis of RA and its subsets, we identified new RA risk loci and highlight candidate causal genes at the majority of RA-associated loci. Most sequence variants have larger effect on the risk of seropositive than seronegative RA. Majority of those with largest effect on RA risk have not been reported before and point to candidate causal genes encoding proteins in the network of interferon alpha/beta and IL-12/IL-23 that signal through the JAK/STAT pathway. Together, these data thus shed light on the molecular mechanism affected by most non-HLA sequence variants derived from over 67 thousand whole-genome sequenced individuals, increases the likelihood to detect rare and low-frequency sequence variants that associate with disease. Furthermore, we were able to test their functional relevance through analysis of RNA sequence and plasma proteome. However, it remains to be seen whether the sequence variants associate with RA in populations of another ancestry.

The SNP-based heritability estimate for seropositive RA was the same as in a previous study (0.19), while lower for seronegative RA (0.099) where previous findings are scarce.

In addition to the causal genes highlighted previously, the network analysis illustrated how majority of all candidate causal genes encode proteins in the interferon alpha/beta and IL-12/IL-23 signalling network. Furthermore, we observed a consistent direction of the effect on seropositive RA risk, gene expression and protein levels in plasma, indicating that increased signalling through the JAK/STAT-pathway is central in the inflammatory cascade in seropositive RA. Our findings are in line with the documented effectiveness of IL-6 receptor and JAK inhibitors (baricitinib, tofacitinib, filgotinib and upadacitinib) as well as CTLA4-Ig in RA. Furthermore, there are inhibitors of other proteins in this pathway that are in development or already marketed for other diseases but have to our knowledge not been tested for treatment of RA, including FLT3 inhibitors used to treat acute myeloid leukaemia and other cancer forms. TYK2 inhibitors that show promising results in clinical trials for psoriatic arthritis and IFNAR1 inhibitors in systemic lupus erythematosus.

In summary, through a large genome, transcriptome and proteome analysis of RA and its subsets, we identified new RA risk loci and highlight candidate causal genes at the majority of RA-associated loci. Most sequence variants have larger effect on the risk of seropositive than seronegative RA. Majority of those with largest effect on RA risk have not been reported before and point to candidate causal genes encoding proteins in the network of interferon alpha/beta and IL-12/IL-23 that signal through the JAK/STAT pathway. Together, these data thus shed light on the molecular mechanism affected by most non-HLA sequence variants derived from over 67 thousand whole-genome sequenced individuals, increases the likelihood to detect rare and low-frequency sequence variants that associate with disease. Furthermore, we were able to test their functional relevance through analysis of RNA sequence and plasma proteome. However, it remains to be seen whether the sequence variants associate with RA in populations of another ancestry.
Contributors SS, LS, PS, GT, UT, and KS designed the study and interpreted the results. SS, BG, HW, GG, ICL, SBS, BAL, MB, LA, KA, SB, CE, OF, IK, HK, BRL, TO, SRO, GNS, HS, ES, LA, TKK, SB, KS, VA, OAA, SR-D, MLH, KL, JA, LP, and OBP managed the data processing of participating study populations/biobanks. SS, LS, PS, EF, GR, AOA, DFG, SAG, GHG, SHL, GM, KHS, GM, GLN, TAO, PIO, SR, UT, and IJ performed the sequencing, genotyping, imputation, expression and genomics analysis. SS, LS, PS, EF, GR, and OBP performed the statistical and bioinformatic analyses. SS, PS, GT, UT, and IJ drafted the manuscript. SS and KS accept full responsibility for the work, had access to the data and controlled the decision to publish. All authors contributed to the final version of the paper.

Funding The study was funded by FORDORSK (grant number 90825), the Swedish Research Council (2018-02803), the Swedish Innovation Agency (Vinnova), Innovationsfonden and The Research Council of Norway, Region Stockholm-Karolinska Institutet and Region Västerbottens ALF, the Danish Rheumatism Association (RFT-042), the Swedish Foundation for Strategic Research (SSF-2019/00629), and the Swedish Foundation for Strategic Research (SSF-2019/00629). The study was also funded by the Novo Nordisk Foundation (655172), Section of rheumatology, The University of Copenhagen (ID 1067863), and the Novo Nordisk Foundation (580843). The study was also funded by the Novo Nordisk Foundation (NNF15OC0016923), the Danish Research Council for Health (143225-00371), Aage and tartışma's legat, the Psoriasis Association, the Novo Nordisk Foundation (456397/2012), the University of Copenhagen, the University of Aarhus, the University of Lyngby, and the Novo Nordisk Foundation (909771). The study was also funded by the Novo Nordisk Foundation (550316), the Danish Research Council for Health (143225-00371), and the University of Copenhagen (ID 1067863). The study was also funded by the Novo Nordisk Foundation (NNF15OC0016923), the Danish Research Council for Health (143225-00371), Aage anddiscussion's legat, the Psoriasis Association, the Novo Nordisk Foundation (456397/2012), the University of Copenhagen, the University of Aarhus, the University of Lyngby, and the Novo Nordisk Foundation (909771).

Competing interests Authors affiliated with deCODE Genetics/Angen declare competing financial interests as employees. OAA is a consultant to Healthlytics. The following coauthors report the following but unrelated to the current report: Karolinska Institutet, with Ja as principal investigator, has entered into agreements with the following entities, mainly but not exclusively for safety monitoring of rhematology immunomodulators: Abbvie, BMS, Eli Lilly, Janssen, MSD, Pfizer, Roche, Samsung Bioepis and Sanofi, unrelated to the present study. SS has ownerships in Intomics A/S, Hoba Therapeutics ApS, Novo Nordisk A/S, Lundbeck A/S and managing board memberships in Proscion A/S and Intomics A/S. BG has received research grants from AbbVie, Bristol Myers-Squibb and Pfizer, OH has received research grants from AbbVie, Novartis and Pfizer. DHV has received speaker and/or consulting fees from AbbVie, Janssen, Lilly, MSD, Novartis, Pfizer, Roche and UCB. AGL has received speaking and/or consulting fees from AbbVie, Janssen, Lilly, MSD, Novartis, Pfizer, Roche and UCB; and CT has received consulting fees from Roche, speaker fees from AbbVie, Bristol Myers-Squibb, Nordic Drugs, Pfizer and Roche, and an unrestricted grant from Bristol Myers-Squibb.

Patient and public involvement Patients and/or the public were involved in the design, conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not applicable.

Ethics approval This research has been conducted using the UK Biobank Resource (application licence number 24838, REC Reference Number: 06/MRE08/65), and the ethics study was approved by the Data Protection National Committees in Iceland (approval no. VSN-15-045 and VSN-16-042), Sweden (approval no. 2016/476:314, 2007/889-312, 2012/2070-312, 2015/176-34 and 04-2521-1), Denmark (Danish Data Protection Agency (general approval number 2012-58-0004 and local number RH-2007-30-4129 I-suite 00678) and the National Committee on Health Research Ethics (NVK-1700407, NVK-1803863 and H-2-2014-086)) and Norway (Regional Committees for Medical and Health Research Ethics, REC South-East C, 2019 B-63469, REK 13-013 and 2010/764). All data processing complies with the instructions of the Data Protection Authority in Iceland (PV_20170609595) and the Norwegian Data Inspectorate. Patients were involved in the design and conduct of several of the studies that are included in this report. Participants gave informed consent to participate in the study before taking part wherever applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. All data relevant to the study are included in the article or uploaded as supplementary information. The GWAS summary statistics are available at https://www.decode.com/summarystatdata. Sequence variants passing GATK filters will be deposited in the European Variation Archive (https://www.ebi.ac.uk/ena-data/ where the used publicly available software (UPA) was used with the algorithms in the sequencing processing pipeline (whole-genome sequencing association testing, RNA-seq mapping and analysis, see methods description in Supplementary Information 2): BWA 0.7.10 (https://github.com/broadinstitute/bwa);
Rheumatoid arthritis


Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use and license their derivative works on different terms, provided the original work is

ORCID iDs
Saelis Sævardsdottir http://orcid.org/0000-0001-9392-6184
Bente Glintborg http://orcid.org/0000-0002-8931-8482
Helga Westerlind http://orcid.org/0000-0003-3380-5242
Torkell Ellingsen http://orcid.org/0000-0003-0426-4962
Inger Gjertsson http://orcid.org/0000-0002-9301-4844
Søren Jacobsen http://orcid.org/0000-0002-5654-4993
Ach KalfasMoth http://orcid.org/0000-0001-7187-1477
Sanne Koldrup http://orcid.org/0000-0002-5812-5234
Heidi Munk http://orcid.org/0000-0002-2212-6283
Carl Turesson http://orcid.org/0000-0002-3805-2290
Vibeke Andersen http://orcid.org/0000-0002-0127-2863
Sølbritt Rantapää-Dahlqvist http://orcid.org/0000-0001-8259-3863
Merete Lund Hetland http://orcid.org/0000-0003-4229-6818
Lars Klareskog http://orcid.org/0000-0001-9601-6186
Johan Åskling http://orcid.org/0000-0003-0433-0616

REFERENCES


