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

Saeidis Saevarsdottir,1,2,3,4 Lilja Stefansdottir,1 Patrick Sulem,1 Gudmar Thorleifsson,1 Egil Ferkingstad,1 Gudrun Rutsdottir,1 Bente Glintborg1,5,6 Helga Westerlind1,2, Gerdur Grondal,3,4,7 Isabella C Loft,8 Signe Bek Sorensen,9 Benedicte A Lie,10,11 Mikael Brink12, Lisbeth Arlestig,12 Asgeir Orn Arnthorsson,1 Eva Baecklund,13 Karina Banasik,14 Steffen Bank,9 Lena I Bjorkman,15 Torkell Ellingsen16,17 Christian Erikstrup,18 Oleksandr Frej19,20,21 Inger Gjertsson22, Daniel F Gudbjartsson,1,23 Sigurjon A Gudjonsson,1 Gisli H Halldorsson1,23 Oliver Hendricks24,25 Jan Hillert,26 Estrid Hогdall,27 Soren Jacobsen6,28 Dorte Vendelbo Jensen,29 Helgi Jonsson,3,4, Alf Kastbom,30 Ingrid Kockum,26 Salome Kristensen31,32 Helga Kristjansdottir,7 Margit H Larsen,33 Asta Linauskas34,35,36 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 Munk6,16,17 Kaspar R Nielsen,38 Gudmundur L Norddahl,1 Asmundur Oddsson,1,3 Thorunn A Olafsdottir,1,3 Pall I Olason,1 Tomas Olsson26 Sisse Rye Ostrowski,6,33 Kim Hoyerslev-Petersen,24 Solvi Rognvaldsdottir,1 Helga Sanner39,40,41 Gilad N Silberberg,41 Hreinn Stefansson,1 Erik Sorensen,33 Inge J Sorensen28, Carl Turesson42, Thomas Bergman,2 Lars Alfredsson,26,43 Tore K Kvien,44,45 Soren Brunak,46 Kristjan Steinsson,7 Vibeke Andersen9,16,46 Ole A Andreassen,19,20 Solbritt Rantapää-Dahlqvist12, Merete Lund Hetland5,6 Lars Klareskog41, Johan Askling,2 Leonid Padyukov,41 Ole BV Pedersen,8 Unnur Thorsteinsdottir1,3 Ingileif Jonsdottir,1,3,7,37 Kari Stefansson,1 Members of the DBDS Genomic Consortium, The Danish RA Genetics Working Group, The Swedish Rheumatology Quality Register Biobank Study Group (SRQb)

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 31313 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 an additional 2 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 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.

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.

variants, increases the risk of seropositive RA 2.27-fold (p=2.1×10−9), more than the rs2476601-A missense variant in PTPN22 (OR=1.59, p=1.3×10−160). STAT4 rs140675301-A replaces hydrophilic glutamic acid with hydrophobic valine (Glu128Val) in a conserved, surface-exposed loop. A stop-mutation (rs76428106-C) in FLT3 increases seropositive RA risk (OR=1.35, p=6.6×10−11). Independent missense variants in TYK2 (rs34536443-C,
INTRODUCTION

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 RA, where around two-thirds of RA patients are in the seropositive subset, based on autoantibodies (rheumatoid factor (RF) and/or antibodies against citrullinated peptide antigens (ACPA)). Although many risk loci have been identified in previous genome-wide association studies (GWAS), most reported RA associations are with common non-coding variants that confer low risk and lack a compelling candidate gene mediating the effect on RA.

The main exceptions are instances where RA patients are in the seropositive subset, based on autoantibodies (rheumatoid factor (RF) and/or antibodies against citrullinated peptide antigens (ACPA)). Although many risk loci have been identified in previous genome-wide association studies (GWAS), most reported RA associations are with common non-coding variants that confer low risk and lack a compelling candidate gene mediating the effect on RA.

ACPA-positive 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 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 314 788 controls), UK Biobank (5798 cases and 402 767 controls of self-reported white British ancestry, confirmed by genetic analysis) and FinnGen (https://www.finnngen.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.eirasedwen.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/imms-epidemiologisk-undersokning-av-riskfaktorer-for-multi-pel-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.srq.nu).

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 multimetics 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
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 ($<500$ kbps from the primary signal, excluding HLA locus). We tested whether primary and secondary signals were in strong linkage disequilibrium ($R^2 > 0.8$) with top cis-variants for genes expressed in various tissues (online supplemental tables 5 and 6), and/or with levels of 4789 proteins in plasma (pQTL, SomaScan, SomaLogic) 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 8515 (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-analysis 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 $= 2.03 \times 10^{-5}$). 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 h2) were higher for seropositive RA (0.19 (0.022)) than for seronegative RA (0.099 (0.019)). For a substantial proportion of the RA-associated sequence variants, their effect was greater on seropositive RA than on seronegative RA risk (table 2, figure 1). However, the genetic correlation between seropositive and seronegative RA was high ($r_g = 0.87$, 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-A, 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 (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 $STAT4$ at DNA binding sites is necessary for full transcriptional activation of many of its target genes, and $STAT4$ 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 symbol</th>
<th>Annotation</th>
<th>OR (95% CI)</th>
<th>P (GWAS)</th>
<th>OR (95% CI)</th>
<th>P (GWAS)</th>
<th>OR (95% CI)</th>
<th>P (GWAS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>2 800 059</td>
<td>TTCG4</td>
<td>Missense</td>
<td>0.93 (0.81-1.08)</td>
<td>0.18</td>
<td>0.14</td>
<td>1.05 (0.90-1.24)</td>
<td>0.67</td>
<td>0.06</td>
<td>1.09 (0.96-1.23)</td>
</tr>
<tr>
<td>chr1</td>
<td>5 399 827</td>
<td>PTM22</td>
<td>Missense</td>
<td>1.03 (0.97-1.09)</td>
<td>0.08</td>
<td>0.06</td>
<td>1.04 (0.98-1.11)</td>
<td>0.10</td>
<td>0.68</td>
<td>1.06 (0.92-1.23)</td>
</tr>
<tr>
<td>chr1</td>
<td>16 506 411</td>
<td>TCGA8</td>
<td>Missense</td>
<td>1.11 (1.06-1.16)</td>
<td>0.05</td>
<td>0.03</td>
<td>1.14 (1.08-1.21)</td>
<td>0.005</td>
<td>0.50</td>
<td>1.20 (1.04-1.39)</td>
</tr>
<tr>
<td>chr1</td>
<td>20 380 280</td>
<td>CTIA4</td>
<td>Regulatory</td>
<td>0.97 (0.90-1.05)</td>
<td>0.07</td>
<td>0.05</td>
<td>0.96 (0.89-1.03)</td>
<td>0.10</td>
<td>0.95</td>
<td>0.95 (0.88-1.02)</td>
</tr>
<tr>
<td>chr7</td>
<td>30 315 968</td>
<td>IFITM1</td>
<td>Intergenic</td>
<td>0.91 (0.85-0.97)</td>
<td>0.04</td>
<td>0.02</td>
<td>1.02 (0.95-1.09)</td>
<td>0.28</td>
<td>0.90</td>
<td>0.90 (0.82-0.98)</td>
</tr>
</tbody>
</table>

*Sequence variants that are significant after adjustment for previously reported sequence variants (online supplemental table 1). Bold indicates candidate causal genes (summarised 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 (Iceland) or the first 20 principal components (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 are: 1.3x10^-05 for class 1 (intronic, intergenic, regulatory sites), 1.5x10^-05 for class 2 (DNase I hypersensitivity sites), 3.0x10^-06 for class 3 (TF binding sites), 4.0x10^-06 for class 4 (coding sites) and 5.0x10^-06 for class 5 (coding sites). For variants with moderate impact in vitro, a p value threshold of 1.0x10^-07 was used. For genome-wide association study (GWAS) analysis, we used the same threshold as in the lead variant discovery step for each individual (secondary analysis). When both low and moderate impact variants were found in one RA overall and seropositive RA, the seropositive RA signal was used as the final RA overall signal. When both low and moderate impact variants were found in one RA overall and seropositive RA, the seropositive RA signal was used as the final RA overall signal. When both low and moderate impact variants were found in and seropositive RA, the seropositive RA signal was used as the final RA overall signal. When both low and moderate impact variants were found in one RA overall and seropositive RA, the seropositive RA signal was used as the final RA overall signal.


Ann Rheum Dis: first published as 10.1136/annrheumdis-2021-221754 on 25 April 2022. Downloaded from

http://ard.bmj.com/
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^{-29}), 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 ligand18 (online supplemental table 7). 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 causal 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=1.44×10^{-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 tetra-tricopeptide 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).

Rheumatoid arthritis

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 table 7). ANKRDS5 rs7731626-A associated with a decreased risk of RA and its subsets and a decreased mRNA expression in whole blood of two neighbouring genes at the locus: ANKRDS5 and IL6ST.

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

---

Figure 2 Identification of sequence variants that associate with seropositive RA and the multiomics approaches used to recognise candidate causal genes. (A) schematic overview of the experimental approach used to identify sequence variants that associate with seropositive RA and their systematic annotation, applying multiomics approach to identify candidate causal genes, that is, based on whether lead variants or correlated variants (R^2 > 0.8) affect protein coding (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)). (B) Out of 33 lead variant associations outside the HLA-locus (online supplemental table 3), 25 candidate causal genes were identified as listed, ranked by effect (OR). All effects are shown for the risk increasing allele based on GWAS in RA study populations from Northwestern Europe (table 1). Associations that are previously unreported in RA are marked with *. Grey boxes highlight where data point to a candidate causal gene. GWAS, genome-wide association study; RA, rheumatoid arthritis.

Figure 3 STA4 missense variant rs140675301 is associated with seropositive RA (18 019 cases), is not correlated with previously reported variants at the locus and leads to an amino acid change in a highly conserved area of the protein. (A) Locus plot for the association of variants at the STA4 locus with seropositive RA. The upper graph illustrates that the intronic variant rs4853458, that is the lead variant at the locus, is not correlated (r^2 < 0.2) with the missense variant rs140675301, that is coloured in purple. The missense variant rs140675301 is only highly correlated (r^2 > 0.8) with one variant, the intronic variant rs189948717 (coloured in red), that has less effect (seropositive RA: OR=1.81, p=3.69×10^-6). Neither of these variants have previously been reported in any disease. The lower graph highlights that the lead variant at the locus (rs140675301, coloured in purple) has many correlated variants, coloured by degree of correlation (r^2) with rs4853458. (B) Secondary structure of STA4 (viewed from two angles) based on a structural model with STA1 crystal structure (PDB code: 1yvl.1.A (Mao et al, Molecular Cell 2005;17:761–71) as template. Glu128Val (red) is located in a loop connecting the N-terminal domain (blue), important for tetramer formation of STATs and nuclear translocation, and the coiled coil domain (green), which provides a carbonised hydrophilic surface that binds to regulatory factors. 24 α-Helices are drawn as cylinders. Invariant residues are marked with asterix. (C) multiple sequence alignment of the conserved STA4 loop between the N-terminal domain (α8) and the coiled coil domain (α9) domain, performed with Clustal omega (https://www.ebi.ac.uk/Tools/msa/clustalo/). RA, rheumatoid arthritis.
exceptions: missense variants in DNASE1/L3 (rs35677470-A) and RIN3 (rs117068593-T) (online supplemental table 2). Both these missense variants are predicted to damage the function of the encoded protein (online supplemental table 4). DNASE1/L3 rs35677470-A is a known signal in RA, but the RIN3 locus has to our knowledge not been reported to associate with any disease before. It encodes Ras and Rab interactor 3 that functions as a guanine nucleotide exchange factor of unknown relevance in RA.

**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 (rs140673301-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 reported 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, R²=0.99 to lead intron variant rs4853458-A)33 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.34,35,36,37 Furthermore, reduced expression of STAT4 has been observed in RA patients that have responded well to disease-modifying treatment.32 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 IFNAR1.35

TYK2 also mediates the signalling of IL-6, IL-10 and IL-4/IL-13.36 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.37 The intronic variant rs7731626-A in ANKRD53 associated with a reduced risk of both seropositive and seronegative RA and also reduced expression of ANKRD53 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.38

The FLT3 receptor is another activator of the JAK/STAT pathway that signals through STAT59 (figure 4), and an intronic variant in the FLT3 gene (rs76428106-C) conferred 3.53% 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.39 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 phosphorolylates (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 rs76428106-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 ancestries.

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 pathway. Together, these data thus shed light on the molecular mechanism affected by most non-HLA sequence variants derived from over 64 million sequence variants at the majority of RA risk loci and highlight candidate causal genes at the majority of RA-risk loci and highlight candidate causal genes encoding proteins in the network of interferon alpha/beta and IL-12/IL-23 that signal through the JAK/STAT-pathway.

In summary, through a large genome, transcriptome and proteome analysis of RA and its subsets, we identified new RA sequence variants that predispose to seropositive RA. In contrast, the genetic background of seronegative RA remains largely unexplained.

Author affiliations
1deCODE genetics, Amgen, Reykjavik, Iceland
2Division of Clinical Epidemiology, Department of Medicine, Solna, Karolinska Institutet, Stockholm, Sweden
3Faculty of Medicine, School of Health Sciences, University of Iceland, Reykjavik, Iceland
4Department of Medicine, Landskally, the National University Hospital of Iceland, Reykjavik, Iceland
5The DANBIO registry, the Danish Rheumatologic Biobank and Copenhagen Center for Arthritis Research (COFCARE), Centre for Rheumatology and Spine Diseases, Centre of Head and Orthopaedics, Copenhagen University Hospital - Rigshospitalet, Glostrup, Denmark
6Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
7Center for Rheumatology Research, Landskally, the National University Hospital of Iceland, Reykjavik, Iceland
8Department of Clinical Immunology, Zealand University Hospital, Køge, Denmark
9Molecular Diagnostics and Clinical Research Unit, IRS-Center Sonderjylland, University Hospital of Southern Denmark, Aabenraa, Denmark
10Department of Medical Genetics, University of Oslo, Oslo, Norway
11Oslo University Hospital, Oslo, Norway
12Department of Public Health and Clinical Medicine, Rheumatology, Umeå University, Umeå, Sweden
13Department of Medical Sciences, Section of Rheumatology, Uppsala University, Uppsala, Sweden
14Department of Rheumatology and Inflammation Research, University of Gothenburg, Gothenburg, Sweden
15OPEN Exploratory Network, University of Southern Denmark, Odense, Denmark
16Rheumatology Research Unit, Odense University Hospital and University of Southern Denmark, Odense, Denmark
17Department of Clinical Immunology, Aarhus University Hospital, Aarhus, Denmark
18NORMENT Centre, Institute of Clinical Medicine, University of Oslo, Oslo, Norway
19Section of Rheumatology, Oslo University Hospital, Oslo, Norway
20Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway
21Center for Bioinformatics, Department of Informatics, University of Oslo, Oslo, Norway
22Department of Rheumatology and Inflammation Research, Gothenburg University, Gothenburg, Sweden
23School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland
24Danish Hospital for Rheumatic Diseases, University Hospital of Southern Denmark, Sønderborg, Denmark
25Department of Regional Health Research, University of Southern Denmark, Odense, Denmark
26Department of Clinical Neurosciences, Karolinska Institutet, Stockholm, Sweden
27Department of Pathology, Herlev Hospital, University of Copenhagen, Copenhagen, Denmark
28Copenhagen Lupus and Vasculitis Clinic, Center for Rheumatology and Spine Diseases, Rigshospitalet, Copenhagen, Denmark
29Department of Rheumatology, Center for Rheumatology and Spine Diseases, Gentofte and Herlev Hospital, Rosne, Denmark
30Department of Biomedical and Clinical Sciences, Linköping University, Linköping, Sweden
31Department of Rheumatology, Aalborg University Hospital, Aalborg, Denmark
32Department of Clinical Medicine, Aalborg University Hospital, Aalborg, Denmark
33Department of Clinical Immunology, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark
34Department of Rheumatology, North Denmark Regional Hospital, Hjørring, Denmark
35Department of Rheumatology, Aarhus University Hospital, Aarhus, Denmark
36Department of Clinical Medicine, Aarhus University, Aarhus, Denmark
37Department of Immunology, Landspitali, the National University Hospital of Iceland, Reykjavik, Iceland
38Department of Clinical Immunology, Aalborg University Hospital, Aalborg, Denmark
39Section of Rheumatology, Oslo University Hospital, Oslo, Norway
40Oslo New University College, Oslo, Norway
41Division of Rheumatology, Department of Medicine, Solna, Karolinska Institutet, Stockholm, Sweden
42Rheumatology, Department of Clinical Sciences, Malmö, Lund University, Malmö, Sweden
43Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden
44University of Oslo, Oslo, Norway
45Diakonhjemmet Hospital, Oslo, Norway
46Institute of Molecular Medicine, University of Southern Denmark, Odense, Denmark

Acknowledgements We would like to thank the individuals who participated in this study and the staff at the Icelandic Patient Recruitment Center, the deCODE genetics core facilities, the Swedish EIRA and EIMS study groups, the Swedish Rheumatology Quality Register Biobank study group (https://srr.nuibiobank-vderijaw/), KI Biobank at Karolinska Institutet, the Biobank Research Unit, Umeå University (https://www.umu.se/en/biobank-research-unit/), Västerbotten Intervention Programme, the Northern Sweden MONICA study and the County Council of Västerbotten for providing data and samples in Sweden; the Danish
DANBIO registry and the Danish Rheumatologic Biobank for supplying data from Danish RA patients, including Niels Steen Krogh, Zitelab ApS, Denmark for database management. Further thanks to all our colleagues who contributed to the data collection and phenotypic characterisation of clinical samples, including Arni J Geirsson, Gudrun B Reynisdottir, Thorunn Jonsdottir and Gunnar Tomasson from Iceland; Arni J Geirsson, Gudrun B Reynisdottir, Thorunn Jonsdottir and Gunnar Tomasson from Iceland; Arni J Geirsson, Gudrun B Reynisdottir, Thorunn Jonsdottir and Gunnar Tomasson from Iceland; Arni J Geirsson, Gudrun B Reynisdottir, Thorunn Jonsdottir and Gunnar Tomasson from Iceland; and also acknowledge colleagues working with the genotyping and analysis of the whole-genome association data. We would also like to thank Vibeke Østergaard Thomsen, International Reference Laboratory of Mycobioculture, Statens Serum Institut and Marianne Krågh Thomsen Department of Clinical Microbiology, Aarhus University Hospital, Aarhus, Denmark, for collecting blood samples, as well as Elvira Chapka, Ewa Kogutowska and Mette Ebreo Ranne, Statens Serum Institut, for laboratory support. We also like to thank the Danish Institute of Public Health for access to genomic data and the families in Norway who take part in the ongoing Norwegian Mother, Father and Child Cohort Study. Last but not least, we want to acknowledge the participants and investigators of the FinnGen study and the UK Biobank.

Collaborators Collaborators from the DBDS Genomic Consortium, the Danish RA Genetics Working Group and the Swedish Rheumatology Quality Register Biobank Study Group are listed in online supplemental information 1. Members of the DBDS Genomic Consortium: Steffan Andersen (Department of Finance Copenhagen Business School Copenhagen Denmark); Karina Banasik (Novo Nordisk Foundation Center for Protein Research Faculty of Health and Medical Sciences University of Copenhagen Copenhagen Denmark); Soren Bruun (Novo Nordisk Foundation Center for Protein Research Faculty of Health and Medical Sciences University of Copenhagen Copenhagen Denmark); Kristoffer Burgdorf (Department of Clinical Immunology University Hospital Copenhagen Copenhagen Denmark); Christian Erkers (Department of Clinical Immunology University Hospital Copenhagen Copenhagen Denmark); Thomas Follmann Hansen (Danish Headache Center Department of Neurology Righospitalet Glostrup Denmark); Henrik Hyrglim (Department of Epidemiology Research Statens Serum Institut Copenhagen Denmark); Gregor Jemec (Department of Clinical Medicine Zealand University Hospital Roskilde Denmark); Poul Jenum (Department of Clinical Neurophysiology at University of Copenhagen Copenhagen Denmark); Pär Ingemar Johansson (Department of Clinical Immunology Copenhagen University Hospital Copenhagen Denmark); Kasper Rene Nielsen (Department of Clinical Immunology Aalborg University Hospital Aalborg Denmark); Mette Ngaeyard (Department of Biomedical Aarhus University Denmark); Mie Topholm Brun (Department of Clinical Immunology Odense University Hospital Odense Denmark); Ole Birger Pedersen (Department of Clinical Immunology Zealand University Hospital, Køge Denmark); Susan Mikkelsen (Department of Clinical Immunology Aarhus University Hospital Aarhus Denmark); Erik Sørensen (Department of Clinical Immunology Copenhagen University Hospital Copenhagen Denmark); Henrik Ulum (Department of Clinical Immunology Copenhagen University Hospital Copenhagen Denmark); Sisse Rye Ottesen (Department of Clinical Immunology Copenhagen University Hospital Copenhagen Denmark); Thomas Werge (Institute of Biological Psychiatry Mental Health Centre St. Hans Copenhagen University Hospital Roskilde Denmark); Daniel Gudbjartsson (deCODE genetics Reykjavik Iceland); Kari Stefansson (deCODE genetics Reykjavik Iceland); Hreinn Stefansson (deCODE genetics Reykjavik Iceland); Unnur Thorsteinsdottir (deCODE genetics Reykjavik Iceland); Margit Anita Laeyens (Department of Clinical Immunology Copenhagen University Hospital Copenhagen Denmark); Khoa Manh Dinh (Department of Clinical Immunology Aarhus University Hospital Aarhus Denmark); Erik Sørensen (Department of Clinical Immunology Copenhagen University Hospital Copenhagen Denmark); Henrik Ulum (Department of Clinical Immunology Copenhagen University Hospital Copenhagen Denmark); Sisse Rye Ottesen (Department of Clinical Immunology Copenhagen University Hospital Copenhagen Denmark); Thomas Werge (Institute of Biological Psychiatry Mental Health Centre St. Hans Copenhagen University Hospital Roskilde Denmark); Daniel Gudbjartsson (deCODE genetics Reykjavik Iceland); Kari Stefansson (deCODE genetics Reykjavik Iceland); Hreinn Stefansson (deCODE genetics Reykjavik Iceland); Unnur Thorsteinsdottir (deCODE genetics Reykjavik Iceland); Margit Anita Harup Larsen(Department of Clinical Immunology Copenhagen University Hospital Copenhagen Denmark); Maria Didriksen (Department of Clinical Immunology Copenhagen University Hospital Copenhagen Denmark); Susanne Sækmose (Department of Clinical Immunology, Zealand University Hospital Køge Denmark); Alba Álvaro Arnaud (Department of Clinical Immunology, University Hospital Reykjavik Iceland); Oddur FMisaksson (Department of Clinical Immunology, University Hospital Reykjavik Iceland); Hreiðmar Jonsdóttir (Department of Clinical Immunology, University Hospital Reykjavik Iceland); Gudbjorg Hauksdottir (Department of Clinical Immunology, University Hospital Reykjavik Iceland); Karin Flosdottir (Department of Clinical Immunology, University Hospital Reykjavik Iceland); and Arni J Geirsson, Gudrun B Reynisdottir, Thorunn Jonsdottir and Gunnar Tomasson from Iceland.

Contributors SS, LS, PS, PT, UT, JI and KS designed the study and interpreted the results. SS, BG, HG, GG, ICL, SBS, BAL, MA, LA, KB, SB, LB, TE, CE, OF, IG, OH, JH, EH, E-MH, SJ, DVI, HJ, AK, IK, SK, HK, ML, AL, AG, TM, HM, TO, KH-F, HS, ES, IU, CT, LAI, KL, SJ, KS, VA, OAA, SR-D, ML, LH, JK, JA, OPB and UH carried out the subject ascertainment and recruitment. SS, BG, HG, GG, ICL, SBS, BAL, MA, LA, KB, CE, OF, IK, HK, BRL, TO, SRO, GNS, HS, ES, LA, TKK, SB, KS, VA, OAA, SR-D, MLH, BK, JA, LP and OPB managed the data processing of participating study populations/ biobanks. SS, LS, PS, FE, GR, AOA, DFG, SAG, GH, SHL, GM, KHS, ML, AL, GHN, TAO, PIO, SR, UT and UI performed the sequencing, genotyping, imputation, expression and epigenomics analyses. SS and JI performed the statistical and bioinformatics analyses. SS, PS, PT, UT, JI and KS 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 NORDFORSK (grant agreement no. 90825, project NORA), the Swedish Research Council (2018-02803), the Swedish innovation Agency (Vinnova), Innovationsfonden and The Research Council of Norway, Region Stockholm-Karolinska Institutet and Region Västerbotten (ALF), the Danish Research Association (R194-46956, A1923, A3037 and A3370 – www.giftfgenerationen.dk), Region of Southern Denmark’s PhD Fund, 12/7725 (www.regionsofdanmark.dk) and the Department of Rheumatology, Frederiksberg Hospital (www.frederiksberghospital.dk). The Rare Genetics has been approved by the National Bioethics Committee of Norway (2014/ 223273), South East and Western Norway Health Authorities, ERC AdG project SELECToN-PREDPOSED, Stiftelsen Kristian Gerhard Jebsen, Trond Mohn Foundation, the Novo Nordisk Foundation and the University of Bergen. KB and SB acknowledge the Novo Nordisk Foundation (grant NNF14CC0001).

Competing interests Authors affiliated with deCODE Genetics/Amgen declare competing financial interests as employees. OAA is a consultant to Healthlyx. The following coauthors report the following but unrelated to the current report: Karolinska Institutet, with JF as principal investigator, has entered into agreements with the following entities, mainly but not exclusively for safety monitoring of rheumatology immunomodulators: Abbvie, BMS, Eli Lilly, Janssen, MSD, Pfizer, Roche, Samsung Bioepis and Sanofi, unrelated to the present study. SB has ownerships in Intromics A/S, Hoba Therapeutics ApS, Novo Nordisk A/S, Lundbeck A/S and managing board memberships in Proscin A/S and Intromics A/S. BG has received research grants from AbbVie, Bristol Myers-Squibb and Pfizer, OH has received research grants from AbbVie, Novartis and Pfizer, DH has received speaker and consultation fees from AbbVie, Janssen, Lilly, MSD, Novartis, Pfizer, Roche and UCB, AOA 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 24898, REC Reference Number: 06/MRE08/65), and was approved by the National Bioethics Committees in Iceland (approval no. VSN-15-045 and VSN-16-042), Sweden (approval no. 96-174, 2006/476-31/4, 2007/889-31/2, 2012/2070-31/2, 2015/1746-31/4 and 04-2521/1-4), 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 (NKV: 1700407, NKV: 1803683 and H-2-2014-086)) and Norway (Regional Committees for Medical and Health Research Ethics, REC South-East C, 2019/ 268469, REK-13/04 and 2010/744). All data processing complies with the instructions of the Data Protection Authority in Iceland (P_2017060950ÞS) 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/summarystats/. Sequence variants passing GATK filters will be deposited in the European Variation Archive (https://www.ebi.ac.uk/ena/data/) where they are used publicly available data sets (UPD data) along with the algorithms in the sequencing processing pipeline (whole-genome sequencing association testing, RNA-sequence mapping and analysis, see methods description in Supplementary Information 2): BWA 0.7.10 mem (https://github.com/lb3/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 sole responsibility of the author(s) and are not endorsed by BMJ. BMJ disclaims all error and/or omissions arising from translation and 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 is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iDs
Saedsl Saevardsdottir http://orcid.org/0000-0001-9392-6184
Bente Glinborg http://orcid.org/0000-0002-8931-8482
Helga Westerlind http://orcid.org/0000-0003-3880-5242
Torkell Ellingsen http://orcid.org/0000-0003-0424-4962
Inger Gjertson http://orcid.org/0000-0002-9301-4844
Sølve Jacobsen http://orcid.org/0000-0002-5654-4993
Alf Kastbom http://orcid.org/0000-0001-7187-1477
Carl Turesson http://orcid.org/0000-0002-3805-2290
Torkell Ellingsen http://orcid.org/0000-0001-0101-2863

REFERENCES


