Title: Rheumatoid factor production is genetically and molecularly distinct from rheumatoid arthritis.

Authors: Mehmet Hocaoğlu, MD[1,2], Amr H Sawalha, MD[1,2,3,4]

Affiliations:

[1]Division of Rheumatology and Clinical Immunology, Department of Medicine, University of Pittsburgh, Pittsburgh, PA, USA

[2]Lupus Center of Excellence, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

[3]Division of Rheumatology, Department of Pediatrics, University of Pittsburgh, Pittsburgh, PA, USA

[4]Department of Immunology, University of Pittsburgh, Pittsburgh, PA, USA.

Please address correspondence to Amr H. Sawalha, MD. Address: 7123 Rangos Research Center, 4401 Penn Avenue, Pittsburgh, PA 15224, USA. Phone: (412) 692-8140. Fax: (412) 692-5054. Email: asawalha@pitt.edu

Conflicts of Interest: The authors have declared that no conflict of interest exists.

Funding: This study was supported by the Rheumatoid Arthritis Research Program Grant from the Arthritis Foundation.

Role of Funding Source: The funding source had no role in the study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the article for publication

Abstract

Background: Rheumatoid factor (RF) autoantibodies are highly prevalent, yet the molecular determinants of RF development and its progression to rheumatoid arthritis (RA) remain poorly understood. Here, we define the genetic, phenotypic, and molecular architecture of RF and its progression to RA.

Methods: 469,036 UK Biobank participants with RF testing and 76 ALTRA cohort individuals were studied. Phenome-wide (PheWAS), genome-wide (GWAS), and proteome-wide association studies compared RF-positive individuals without autoimmune disease to RF-negative controls. Single-cell RNA sequencing enabled pseudobulk differential expression and cytokine signature enrichment analyses.

Results: RF seroprevalence was 9.3% and longitudinally stable in 94.5% of individuals. PheWAS identified 48 significant associations, led by chronic viral hepatitis (OR 4.8), hypersensitivity pneumonitis (OR 3.6), bronchiectasis (OR 1.9), and COPD (OR 1.4). GWAS of 24,216 RF-positive individuals revealed 29 independent loci; the strongest signal was in the extended HLA region (OR 1.45, *P*-value=5.4×10^{-221}). Non-HLA loci converged on B cell homeostasis genes (*ETS1*, *BACH2*, *PAX5*, *TNFRSF13B*, *FCGR2A*). RF-positive individuals did not carry elevated RA polygenic risk. Proteomic profiling identified 153 differentially abundant proteins enriched for humoral immunity and interferon-induced chemokines, with 79% showing dose-response relationships across titers. Progression to RA involved a shift toward activating tissue-damaging inflammatory pathways rather than amplification of the RF signature. Single-cell transcriptomics of RF-positive individuals without RA localized dysregulation to memory B cells, with downregulation of inhibitory genes (*FCGR2B*, *BACH2*, *FOXP1*) and upregulation of activation markers.

Conclusion: RF production is governed by HLA class II and B cell regulatory loci, associated with mucosal inflammation, and is genetically and molecularly distinct from RA.

Introduction

Rheumatoid factor (RF), an antibody against the Fc portion of IgG, is one of the most common autoantibodies in the general population (1, 2). RF has a strong association with rheumatoid arthritis (RA) and may precede the disease onset by several years (3). However, RF shows limited specificity for RA and can be observed in other autoimmune and infectious diseases, such as viral hepatitis (4-6). Furthermore, most individuals with RF do not develop RA or other autoimmune diseases, and RF positivity may be transient during infections, possibly reflecting physiological roles in clearing immune complexes (7).

Emerging data show that individuals with RA related autoantibodies, such as anti-cyclic citrullinated peptide antibodies (ACPA), harbor substantial immune dysregulation prior to the clinical onset of RA (8). ACPA is highly specific for RA and is much less common in the general population compared to RF (1, 9, 10). Whether individuals with RF positivity, especially those who are otherwise asymptomatic for autoimmune disease, also harbor substantial immune dysregulation remains incompletely understood. Moreover, while the genetic risk factors of RA are well studied (11), there is limited understanding of the genetic basis of RF production and how it relates to RA.

In this study, we comprehensively characterized the molecular signature of RF by performing genome-wide, phenome-wide, proteome-wide, and transcriptome-wide association studies (**Fig. 1**). We uncover the genetic basis of RF production and show that RF exhibits a genetically and molecularly distinct architecture from RA, marked by selective immune activation that occurs without tissue-destructive inflammatory responses.

Results

RF production is predominantly low titer and prevalent in the general population

The seroprevalence of RF positivity was 9.3% among 469,036 participants with baseline RF testing in the UK Biobank. RF production was predominantly low titer: 6.8% of UK Biobank participants had titers less than 3×ULN, whereas higher titers were progressively less common (2.5% >3×ULN; 1.1% >6×ULN; 0.5% >12×ULN). There were only negligible differences across RF titer subgroups in demographic characteristics (**Table 1)**. We observed markedly increased RA prevalence with rising RF titers, while the prevalence of overall clinical autoimmunity excluding RA was comparable (**Table 1**).

RF serostatus shows longitudinal stability with seroconversions concentrating at low titers

We next investigated whether RF positivity was transient or stable over time. Repeat RF testing was available for 16,660 individuals (mean interval 4.3 ± 0.92 years). Overall serostatus was stable in 94.5% of the repeat tests (n=15,736). 26% of baseline seropositive individuals (n=387) converted to seronegative and 3.5% of baseline seronegative individuals (n=537) converted to seropositive. Seroconversions were concentrated in the lowest titer range: 93% of the seropositive converters transitioned into less than 3×ULN category while 97% of seronegative converters had baseline titers less than 3×ULN (**Extended Data Fig. 1**).

RF is associated with conditions characterized by chronic infections and mucosal inflammation

To identify conditions associated with RF production beyond clinical autoimmunity, we performed a phenome-wide association study (PheWAS) after excluding individuals with autoimmune diseases. For this analysis, there were 36,049 RF-positive individuals and 383,422 RF-negative individuals with available phenotypic data. We identified 48 phenotypes with a significant RF association at an FDR of 0.1 (**Fig. 2**). The most significant association was chronic viral hepatitis (OR 4.8, Adjusted *P*-value 3.4×10^{-23}) with several associated viral hepatitis related signals (**Supplementary Table 1**). Beyond hepatic signals, the most prominent cluster of associations involved several respiratory disease phenotypes. Hypersensitivity pneumonitis

(OR 3.6, Adjusted *P*-value 6.6×10^{-4}) had the largest effect size across these disorders followed by bronchiectasis (OR 1.9, Adjusted *P*-value 2×10^{-12}), interstitial lung disease (OR 1.7, Adjusted *P*-value 5.8×10^{-3}) and COPD (OR 1.4, Adjusted *P*-value 5.1×10^{-14}). Other prominent signals included sepsis, anogenital infections, pregnancy complications and lung neoplasms (**Supplementary Table 1**).

Genetic architecture of asymptomatic RF production implicates the HLA region and B cell regulation

After quality controls, we analyzed 7,225,682 genetic variants in 24,216 asymptomatic RF-positive individuals and 244,786 RF-negative controls for a genome-wide association study (GWAS). There were 17,188 genetic variants across 29 genetic loci associated with asymptomatic RF production at the genome-wide significance level (**Fig. 3**). 16,058 of these variants were localized within the extended HLA region (chr 6, 25Mb-34Mb) (**Extended Data Fig. 2**). The top hit in the HLA region was the SNP rs3129959 (OR 1.45, 95% CI 1.42-1.48, *P*-value 5.4×10^{-221}) while the classical allele with the strongest association was HLA-DQA1:05 (OR 1.3, 95% CI 1.28-1.33, *P*-value 2.3×10^{-135}). We observed statistically independent signals between the rs3129959 and HLA-DQA1:05 in the pairwise conditional analysis (**Supplementary Table 2**). A protective missense association at *TNFRSF8* (rs2230624; OR 0.76, *P*-value=2.6×10^{-10}) locus showed the largest effect size among non-HLA variants (**Supplementary Table 3**). Other non-HLA risk loci mapped to genes responsible for B-cell differentiation and identity (*ETS1*, *BACH2*, *PAX5*, *FOXP1),* BAFF/TACI axis (*TNFRSF13B, TNFSF13B),* Fc receptor and immunoglobulin biology (*FCGR2A*, *FCRL5)*, immune cell development (*TBX21, ID2, GFI1, JAK1)*, and co-stimulation (*TNFSF4, CD70*) (**Supplementary Table 3**). Further, the RF-associated genetic loci showed a significant degree of protein-protein interactions (28 nodes/52 edges; expected number of edges, 6; *P*-value < 1×10^{-16}, average

node degree, 3.71) and was enriched for genetic pathways affecting the regulation of B cell activation (**Extended Data Fig. 3**).

Asymptomatic RF production is genetically distinct from RA

We next assessed whether RF-positive individuals without RA carried increased RA genetic risk. Our analysis showed that RF-positive individuals did not have a significant increase in RA genetic risk compared to RF-negative controls, whereas RF-positive RA cases expectedly had significantly higher RA genetic risk (**Fig. 4A**). Consistent with this observation, RA-PRS demonstrated good discrimination performance for identifying RF-positive RA patients among RF-positive individuals (ROC-AUC 0.76, 95% CI 0.75-0.77) (**Fig. 4B**). We observed comparable findings in sensitivity analyses restricted to individuals with higher RF titers (>3×ULN and >12×ULN) (**Extended Data Fig. 4**).

RF production has a titer-dependent proteomic signature enriched for humoral immunity, antigen presentation, and interferon-induced chemokines

After quality controls and exclusion of individuals with preexisting autoimmune diseases, there were 3,368 RF-positive individuals and 34,155 RF-negative individuals with available proteomic data in the UK Biobank. We identified 153 differentially abundant proteins between RF-positive and RF-negative individuals at an FDR of 1% (**Fig. 5**). Across the serum proteome, asymptomatic RF production was associated with an immune activation signature involving markers of antibody production and immune complex regulation (JCHAIN, MZB1, FCRL2, FCRL5, FCAMR, TRIM21), upregulation of interferon-induced chemokines (CXCL9, CXCL10, CXCL11, CXCL13, IFNG) and enhanced antigen presentation (HLA-E, CD74, CD80, CD27, TNFRSF9, LAG3, PDCD1) (**Supplementary Table 4**). In subgroup analyses, we identified 536 proteins with differential expression across RF titers, 79% of which showed monotonic relationships with rising titers underscoring the titer-dependent nature of the RF proteomic

signature (**Supplementary Table 5**). Notably, we observed a higher abundance in the serum for proteins which were associated with RF at the genetic level such as TNFRSF8, TNFRSF13B, TNFSF13B, and FCRL5 (**Supplementary Table 5**). Further, a sensitivity analysis excluding individuals with incident autoimmune disease development after cohort entry showed similar findings which further underscore that RF production exhibits an independent proteomic signature pointing to a distinct immune dysregulation state (**Supplementary Table 6**).

Progression to RA from RF production is marked by expansion of inflammatory and tissue-damaging pathways

We next compared the serum protein abundance in RF-positive individuals without RA at the time of sample collection to individuals with RF-positive RA and identified 747 differentially expressed proteins between the two groups (**Supplementary Table 7**). This showed that progression to RF-positive RA was characterized by a broad expansion in both the number and magnitude of altered proteins as asymptomatic RF production was characterized by comparatively limited proteomic perturbation with smaller effect sizes (n=153) (**Fig. 5a**). Further, the RA specific proteomic signature we identified had limited overlap with the proteomic alterations associated with RF production (n=46). This suggests that proteomic dysregulation of RA accelerates at or near disease onset rather than increasing uniformly during the asymptomatic rheumatoid factor production phase. Characterizing the RA-specific proteomic signature (**Fig. 5b)** revealed a dominant proteomic signature comprised of key cytokines and synovial inflammation markers such as TNF (12), CCL7 (13), RRM2 (14), MMP3 (15) that were largely unchanged with asymptomatic RF production but became strongly induced in RA while other key cytokines such as IL-6, IL-10, IL-12 (12, 16) increased progressively in abundance across both steps, suggesting early immune priming in RF that is further amplified in RA. Supporting these observations, enrichment analyses showed that pathways related to antigen presentation and adaptive immune system activation were upregulated in RF production, while

inflammatory pathways such as neutrophil degranulation, extracellular matrix organization and TNF-induced inflammation are more enriched in RF-positive RA compared to RF production without RA (**Extended Data Fig. 5**).

RF production is associated with selective reprogramming of the memory B cell compartment

To determine the immune cells and transcriptomic programs that are altered in asymptomatic RF production, we analyzed the single-cell RNA-sequencing data from the ALTRA cohort. For this analysis, there were 36 RF-positive and 40 RF-negative individuals without preexisting RA with available single cell sequencing data (**Supplementary Table 8**). Pseudobulk differential expression analyses showed that RF production is associated with limited transcriptional changes in contrast to the global dysregulation observed in ACPA. There were 25 differentially expressed genes (DEG) across 7 immune cell types in RF compared to 27,111 DEGs across 60 immune cell types in ACPA at FDR 0.1 (**Extended Data Fig. 6**). In subtype specific analysis, we observed a greater degree of transcriptional dysregulation in IgA RF-positive compared to IgM RF-positive individuals. Further, majority of immune dysregulation observed in RF production was localized to the core memory B cell compartment. Specifically, in IgA-RF, there were 180 DEGs across 7 cell subtypes, 165 of which were in core memory B cells; compared to IgM-RF 36 DEGs across 9 cell subtypes, 13 of which were in core memory B cells (**Supplementary Table 9**). Core memory B cells which represent the largest subpopulation of memory B cells characterized by markers such as *CD27* and *AIM2* (17), showed a coordinated transcriptional shift in B-cell signaling with downregulation of inhibitory genes and upregulation activation markers (**Supplementary Table 9**). Downregulated genes included canonical components of B-cell receptor and inhibitory signaling programs, including *BLNK*, *FCMR*, *FCGR2B*, *FCRL2*, and *CD72*. We also observed downregulation of multiple B-cell regulatory genes, including *FOXP1*, *BACH2*, *BCL11A*, and *AFF3*. In contrast, genes linked to activation, co-stimulation and cellular remodeling were upregulated, including *CD86*, *PRKCD*, *TNIP2*, *SCIMP*, *ITGB1, ANXA2*, and

S100A10 (**Supplementary Table 9**). The immunoglobulin gene expression showed downregulation of *IGHM* and *IGHD* with upregulation of *IGHA1* which suggests a global shift towards IgA in the core memory B cell compartment (**Supplementary Table 9**). Further supporting the memory B cell reprogramming in RF production, differential abundance analysis of immune cell types showed increased levels of CD95+ memory B cells in the peripheral blood (Beta 0.55, Adjusted *P*-value 1.1×10^{-2}), an effect observed for both IgA and IgM subtypes **(Supplementary Table 10)**. Notably, we also observed a significant upregulation of *PPP1R14A* (log2FC = +1.08, q-value 3×10^{-5}) within the CD95+ memory B cell compartment **(Supplementary Table 9)**. *PPP1R14A* inhibits protein phosphatase 1, and its upregulation could reflect a reprogramming of activation thresholds in CD95+ memory B cells. Our observation on the downregulation of *FOXP1* and *BACH2* genes in core memory B cells converges with the observation that germline variants in the same loci are associated with a protective effect on RF development **(Supplementary Table 3)**. This suggests that the genetic variations in B cell activation thresholds could confer a propensity to develop RF antibodies.

IgA RF and ACPA production exhibit shared and distinct transcriptional cytokine signatures

To identify global cytokine programs associated with RF and ACPA status, we performed cytokine signature enrichment analysis across 23 immune cell types utilizing the Human Cytokine Atlas. As the transcriptomic changes in RF was predominantly observed with the IgA subtype, this analysis was restricted to IgA subtype of RF. We identified 18 qualifying cytokine signatures for ACPA and 16 for IgA RF with 12 shared between both autoantibodies (**Supplementary Table 11**). IL-10 was the most broadly up-regulated cytokine signature across both autoantibodies, with positive enrichment in 14 cell types for ACPA and in 13 for IgA RF converging with the serum proteomic observations (**Fig. 6**). IL-4 and IL-21 were other cytokines concordantly upregulated in both autoantibody contexts. In contrast, type I interferon suppression signature was evident in ACPA-positive individuals, across IFN-β, IFN-ω, and IFN-

γ. This interferon suppression was also partially present in IgA RF, where IFN-γ and IFN-β were similarly downregulated but with heterogeneity across cell types, while IFN-α1 exhibited a positive enrichment in dendritic cells (**Fig. 6**). Several T cell-associated cytokines, including IL-2 and IL-15 showed positive enrichment across multiple cell types in ACPA, whereas IgA RF displayed an inverted pattern with predominantly negative enrichment. IL-1β also displayed a discordance with positive enrichment in ACPA but predominantly negative enrichment in IgA RF **(Fig. 6**). Collectively, these results reveal both convergent and divergent cytokine-driven transcriptional programs underlying ACPA and IgA RF seropositivity.

Discussion

In this study, we uncover the molecular determinants of asymptomatic RF production and its subsequent progression to clinical RA. We show that RF production and progression to RA involve distinct molecular pathways by integrating genome-wide and phenome-wide association studies with serum proteomics, polygenic risk assessment, and single-cell transcriptomics of immune cells. Our findings raise the potential of prioritizing the immune pathways for early prevention of autoimmunity by preventing RF development and subsequent RF to RA progression and enabling molecular risk stratification for RA progression across RF-positive individuals.

The classical allele with the strongest RF association in our study, HLA-DQA1:05, is a risk factor for the development of adalimumab antibodies (18, 19). The independent association of this classical allele with RF production potentially suggests a preferential binding of this HLA allele to immunoglobulin epitopes which warrants further study. This HLA signal is different than the robust shared epitope association seen in RA (12). Prior literature showed that the genetic association with shared epitope in RA is primarily seen in individuals with ACPA antibodies (20).

Differential HLA associations primarily driven by the specific autoantibodies regardless of the clinical status were also described previously (21). This suggests that divergent genetic signals in the HLA region might be the reason for antigen specific susceptibility for RF, ACPA, and other specific autoantibody related epitopes. Beyond the HLA region, individuals with asymptomatic RF positivity exhibited limited genetic overlap with RA. This suggests that genetic risk for the development of the asymptomatic phase of RA with the emergence of RF and subsequent development into clinical RA might be conferred in an independent and additive fashion. As RF itself confers marked future RA risk (3), assessment of RA genetic risk among individuals with RF positivity might be a feasible strategy to stratify individuals for future RA risk.

The dynamic nature of the serum proteome between RF production and established RA suggests that there are potentially separate mechanisms associated with early and late stages of the RA pathology. Supporting this, prior research showed that RA patients treated with immunosuppressive medications such as TNF inhibitors and methotrexate had comparable rates of autoreactive B cell clones despite improving clinical symptoms (22). Hence, RA protein biomarkers that we identify could potentially be grouped into early autoimmune markers associated with RF development and effector phase markers associated with driving clinical symptoms. Accordingly, it is possible that early autoimmune markers be used to predict future RA risk and be prioritized as drug candidates for early prevention efforts, while effector phase markers might be higher yield drug candidates for RA treatment after the disease onset. As illustrative examples, two significantly upregulated proteins in RA when compared to asymptomatic RF production: TNF and IL-6 are targets for highly effective RA medications (23). On the other hand, drugs targeting proteins such as interferon-gamma and TNFRSF13B which are upregulated in RF production but do not significantly change in RA (as compared to RF) failed in previous clinical trials (24, 25).

A mucosal origin for RF and ACPA has long been suspected due to several observations (26), such as the detection of RA-related autoantibodies in the sputum (27), increased proportion of circulating IgA+ plasmablasts (28) and asymptomatic small airway inflammation observed in imaging in individuals with RA related antibodies (29). In line with these observations, lung diseases such as COPD, bronchiectasis, and asthma were previously shown to have an association with RA (30, 31). Our results provide additional support to the mucosal origin hypothesis by demonstrating that there is a strong association between RF production and these conditions independent of RA and other clinical autoimmunity. Further, we expand the link between mucosal inflammation and RF production by identifying hypersensitivity pneumonitis as one of the strongest associations with RF. A potential unifying mechanism for RF generation across these conditions could be the development of ectopic lymphoid tissues in the lung mucosa. Indeed, prior research demonstrated that inducible bronchus-associated lymphoid tissue (iBALT), an ectopic lymphoid lung tissue that is developed in response to various insults, produces RF antibodies in the lungs from RA patients (32). iBALT also develops in lung diseases such as hypersensitivity pneumonitis, COPD, and bronchiectasis, which might be the source of the RF production in these conditions (33-35). The association of RF with viral hepatitis, anogenital herpes infections, and pregnancy complications could also stem from inflammation in the biliary, genital, and anal mucosa within the framework of mucosal origin hypothesis (36). Supporting this interpretation, an association between pregnancy complications and RF-positive RA development was previously reported (37).

At the molecular level, multi-omics architecture of RF production converged primarily on the reprogramming of B cell activation thresholds. Notably, several RF risk loci we identified encode protein pairs with physical and/or functional interactions playing a role in B cell transcriptomic regulation: *PAX5-ETS1* (38)*, ID2-GFI1 (39), TNFRSF13B-TNFSF13B* (40)*, PAX5-BACH2* (41), *FOXP1-BACH2* (42). These factors collectively regulate the B cell activation thresholds and

differentiation into antibody-secreting cells, suggesting that altered control of these pathways may be a key mechanism underlying RF generation (42-44). Consistent with this interpretation, immune complex receptors that tune B cell activation thresholds such as FCGR2B and Fc-like receptors, were also implicated at the genetic, proteomic and transcriptomic levels with RF (45, 46). This framework might also help explain the expansion of CD95+ memory B cells with RF that we observed in the ALTRA cohort. Although the developmental origin and function of CD95+ memory B cells remain incompletely defined, they have been proposed to represent an effector-memory population enriched for recent germinal-center emigrants (47). CD95 (Fas) has an essential role in the elimination of autoreactive B cells during germinal center selection (48). Hence, genetic reprogramming of B cell activation thresholds that we observed might counteract the apoptotic effect of CD95 and favor the escape autoreactive B cell clones. The spontaneous emergence of RF antibodies in lupus-prone MRL/*lpr* mice with Fas deficiency further supports a potential link between aberrant CD95 signaling and RF production (49).

Progressively greater transcriptomic dysregulation observed in IgM RF, IgA RF and ACPA potentially suggests a stepwise model of disease evolution in at least a subset of patients. In this framework, autoreactive IgM RF-positive B cell clones may first escape deletion because of genetic variations in B cell activation thresholds and stimulation through mucosal inflammation, followed by class switching to IgA RF and subsequent emergence of ACPA before the onset of clinical disease. This framework is supported by the fact that IgA RF positivity in the absence of IgM RF is very rare (50), and IgA RF confers a greater RA risk compared to IgM RF (51). Further, median duration of RF positivity before clinical onset was longer compared to ACPA in a study of RA patients who underwent blood sampling prior to disease onset (52). Nevertheless, another study showed discrepant findings with ACPA having longer duration prior to disease onset which suggests it is also possible that the broad immune activation seen in ACPA might trigger RF production in certain contexts (53).

Interestingly, we observed an upregulation of IL-10 with RF, ACPA and established RA at both proteomic and transcriptomic levels. Given its immunosuppressive effects, it is possible that IL-10 upregulation is a compensatory mechanism for the increased immune activation seen in autoantibody production and established RA (54). However, it is worth noting that IL-10 was found to upregulate Fc gamma receptor expression in monocytes and increase TNF production in response to immune complex stimulation which may counteract its otherwise anti-inflammatory effects and worsen proinflammatory responses in RA (55). Hence, considering our data, the role of IL-10 in development of RF and progression to RA warrants further study.

Our findings challenge the view that low-titer RF production is necessarily transient or clinically benign (7). Indeed, about two thirds of individuals with low titer RF positivity (less than 3×ULN) remained positive after a mean follow-up of 4 years suggesting that low titer RF production can be persistent. Further, individuals with low-titer RF positivity had significantly higher rates of RA, indicating that low-titer RF production could have clinical implications. RF proteomic signature showed titer dependent associations including at low titers without a clear threshold effect. Hence, our data suggests that low grade RF production is a mild form of immune dysregulation linked with RF mechanisms rather than a transient and benign immunological phenomenon.

Our study has a few limitations. The PheWAS analysis was exploratory in nature and was performed using diagnostic codes from linked medical health records and self-reports which could exhibit misclassification bias. For UK Biobank cohort, we did not have data on ACPA serostatus, and the RF status was ascertained through immunoturbidimetry which could not separate between different immunoglobulin subtypes of RF. Although we utilized a broad definition of autoimmunity and had clinical data capture across diverse resources such as self-report, primary care and hospital records, it is possible that some individuals with undiagnosed autoimmune diseases were included in our RF-positive cohort. GWAS analysis was restricted to individuals from White populations due to limited sample size available for other populations.

Functional studies will be needed to understand which genes and cell types are modulated by the SNPs in genetic loci that showed an association with RF production. While we were able to demonstrate the selective reprogramming of memory B cell compartment in RF and identify key RF-related markers supported across omics layers, further characterization of the RF-related specific transcriptional programs will be an important research direction.

In conclusion, our findings reframe RF production as a genetically controlled trait anchored in memory B cell dysregulation with a mucosal inflammation link and uncover the molecular pathways of RF to RA progression. Further research will be needed to turn the molecular signatures that we identified into accurate prediction models and validating the RF specific and RA specific pathways for identifying potential drug targets for RA prevention.

Materials and Methods

Sex as a biological variable

Our study examined data from both male and female human participants. Biological sex was adjusted in the analytic models as reported below. No sex specific analysis was performed.

Study populations and design

UK Biobank is a prospective cohort study that enrolled around 500,000 individuals aged 40 to 69 between 2006 and 2010, with extensive molecular profiling and health record linkage (56). Individuals enrolled in the UK Biobank were tested for rheumatoid factor (RF) for research purposes at cohort entry as a part of biomarker panel with testing being repeated for a subset of the cohort (around 20,000) individuals several years later (56). The Allen Institute for Immunology–University of California, San Diego–University of Colorado Transition to Rheumatoid Arthritis (ALTRA) is a prospective longitudinal cohort study designed to characterize the immune alterations and molecular pathogenesis occurring during early,

preclinical phase of rheumatoid arthritis (8). ALTRA cohort includes individuals at risk for RA as defined by ACPA positivity and healthy controls without autoimmune disease. ALTRA participants underwent RF testing for research purposes. We performed an original reanalysis of the individual level single cell RNA sequencing data (8). The details regarding sample processing, assay methodology, and quality controls employed for RF testing in UK Biobank and ALTRA cohorts are described in **Supplementary Material 1**.

Case selection strategies for rheumatoid arthritis and autoimmune diseases in the UK Biobank

We defined asymptomatic RF production in the UK Biobank as individuals who tested positive for RF and did not have any autoimmune disease at cohort enrollment. To maximize sensitivity, we defined the presence of autoimmune disease as any report of autoimmune disease diagnosis in either nurse interviews (self-report), hospital admissions, or primary care records (**Supplementary List 1).** For subgroup analyses involving RF-positive rheumatoid arthritis, we defined RA as individuals who had an RA diagnosis code, RF positivity, and a self-reported DMARD medication use at cohort entry (**Supplementary List 2)**. The positive predictive value (PPV) of a similar case selection strategy was estimated as 97% in prior studies (57).

Phenome-wide association study

We performed a phenome-wide association study using the 3-digit level ICD codes mapped from clinical data obtained during the nurse interview at enrollment and linked hospital admission, primary care, and cancer registry records. Details of the data linking process are described in **Supplementary Material 2**. We defined prevalent cases with respect to the rheumatoid factor testing date as individuals who had a diagnosis before or within 6 months of sample collection. Individuals who had their first diagnosis at least 6 months after sample collection were considered incident cases and were excluded from the analyses. We tested the association between RF status and each prevalent phenotype independently using logistic

regression adjusted for age at sample collection, sex, self-reported ethnic background, and smoking status. Individuals with missing covariate data (~0.05% of the whole cohort) and phenotypes with less than 30 cases were excluded from the analyses. We controlled the false discovery rate (FDR) at 10% through Benjamini-Hochberg (BH) approach (58). These analyses were performed using PheWAS package in R.

Genome-wide association study (GWAS)

We used genotyping data in the UK Biobank which includes approximately 850,000 genotyped variants through the UK BiLEVE Axiom array and UK Biobank Axiom array with TOPMed reference-based imputation. We performed quality controls by filtering SNPs with minor allele frequency (MAF) less than 1%, significant departures from Hardy-Weinberg (P-value<1×10^{-3}), high missingness rates (≥3%) and INFO scores (indicating imputation quality) less than 0.9. We also removed individuals with poor genotyping success rates (≤98%), who are outliers for heterozygosity, have genetic relatedness up to third-degree relatives, are of self-reported mixed ancestral backgrounds, have sex chromosome aneuploidies, or have mismatches between genetic and self-reported sex. Further details regarding the genetic data in the UK Biobank, imputation processes, and quality controls are described in **Supplementary Material 3**. From the asymptomatic RF-positive cohort at baseline, we also excluded individuals who received an incident autoimmune disease diagnosis after cohort enrollment. The analysis was restricted to individuals from self-reported White populations due to limited sample size from other populations. We tested the association between individual SNPs and binary rheumatoid factor status using logistic regression adjusted for age at recruitment, sex, and first 5 principal components (PC) to control for population stratification. The genome-wide significance was defined as a P-value less than or equal to 5×10^{-8}. HLA classical alleles were imputed using SNP2HLA module of the HLA-TAPAS analysis workflow and Beagle version 5.4 (59). HLA classical alleles with an imputation quality metric R2 greater than 0.9 were used in subsequent

analyses. We performed stepwise logistic regression to test for independent effects of the SNPs and classical HLA alleles in the extended HLA region. The GWAS analysis was conducted using PLINK version 2 software (60).

Lead genome-wide significant variants in each locus were mapped to the closest protein coding gene followed by protein–protein interaction (PPI) analyses using the STRING database (61). STRING was used to construct a PPI network among input proteins using curated and predicted functional associations (61). To assess whether the observed connectivity exceeded that expected by chance given a random protein set of comparable size, we used STRING's built-in PPI enrichment test (61). Functional enrichment was then evaluated within STRING for Gene Ontology using the whole genome as background and multiple-testing correction with terms considered significant at FDR ≤ 0.05 using BH approach (58).

We used the precalculated RA polygenic risk score (RA-PRS) provided by the UK Biobank (62) to compare the RA genetic risk between RF-positive individuals without RA and RF-negative healthy controls. Discrimination performance of the RA-PRS was assessed through Area Under the Receiver Operating Characteristic Curve (ROC-AUC) with 95% confidence intervals calculated through the Delong's method (63).

Proteome-wide association study

A subset of UK Biobank cohort (~50,000) underwent proteomic profiling through the Proximity Extension Assay (PEA) method of the Olink Explore platform (64). This platform provides normalized protein expression (NPX) value which is an arbitrary unit for relative quantification of serum proteins on log2 scale. For quality controls, data were preprocessed by the UK Biobank Pharma Proteomics consortium as described in **Supplementary Material 4**. We performed differential protein abundance analyses for each protein by fitting a multivariable linear model with NPX as the dependent variable and the binary asymptomatic RF status as the primary

independent variable. To further investigate the dose-response relationship between serum proteins and RF levels, a subgroup analysis was performed comparing the following RF-positive titer groups to RF-negative individuals: low (<3x ULN), medium (3-12 ULN), and high (>12 ULN). We performed an additional subgroup analysis by excluding individuals who received an incident autoimmune disease diagnosis after cohort enrollment to investigate if RF proteomic signature is influenced by the proteomic alterations of RA or other autoimmune diseases prior to symptom onset. We also compared RF-positive individuals without RA to individuals with RF-positive RA. All models were adjusted for age, sex, ethnic background, processing time, and batch to account for demographic and technical variation. Missing data for each protein were imputed with its respective median across all samples. The analysis employed linear models with empirical Bayes moderation of standard errors (65). Differential expression statistics were obtained in the form of log fold-changes (logFC) with their associated moderated t-statistics and p-values for each protein. Proteins with an absolute logFC of at least 0.05 were considered significant. False discovery rate (FDR) was controlled at 1% through the Benjamini-Hochberg (BH) approach (58). All analyses were conducted in R (version 4.4.2) using the limma package (version 3.60.6) (65). Gene set overrepresentation analysis was performed for proteins with a differential abundance via Reactome v7.2 using WebGestaltR (v0.4.6, method =ORA) with all measured proteins by Olink Explore panel used as background genes and FDR controlled at 5% through the BH approach (58, 66, 67).

Pseudobulk differential gene expression analysis

We utilized the preprocessed scRNA-seq data from the ALTRA cohort which underwent extensive quality controls. The quality controls employed by ALTRA are described in **Supplementary Material 5.** Using the first collected sample for each participant, we performed pseudobulk differential expression analysis for each immune cell type comparing the RF-positive to RF-negative individuals without RA at the time of sample collection adjusted for age

at sample collection, sex, BMI, ACPA and batch B182 status (as recommended by the authors of the original analysis due to significant batch effects observed with this particular batch) (8). Immune cell type labels were assigned using Celltypist model with AIFI Immune Cell Atlas used as reference (17). We also performed the same analysis comparing ACPA positive and ACPA negative individuals without RA adjusted for RF status using the same model structure. FDR was controlled at 10% through the Storey-Tibshirani procedure. All differential expression analyses were performed using PyDESeq2 package (68*,* 69).

Immune cell type population differential abundance analysis

We utilized the precalculated centered log ratios (CLR) for immune cell type populations from the ALTRA cohort. CLR was calculated by extracting counts and frequencies for each cell subset from the scRNA-seq data followed by adding a pseudocount of 1 to raw counts and CLR transformation. We performed linear regressions to compare CLR values of each immune cell type across RF-positive and RF-negative individuals without RA, adjusting for age at sample collection, sex, BMI, and ACPA status. FDR was controlled at 10% through the BH approach. Further details regarding data processing in the ALTRA cohort have been published elsewhere (8).

Gene set enrichment analysis of cytokine response signatures

We performed a signed preranked gene set enrichment analysis (GSEA) utilizing the Human Cytokine Atlas dataset which provides data on cytokine-induced transcriptomic signatures for 90 different cytokines in immune cells (70). For each mapped cell type and cytokine, cytokine-responsive genes were defined as those with adjusted *P*-value less than 0.01 and absolute logFC > 0.8 and were further partitioned into cytokine-upregulated and cytokine-downregulated subsets. To encode directionality within a single bidirectional gene set, the ranked query statistic was transformed on a per-cytokine basis such that genes upregulated by the cytokine retained

their original DESeq2 statistic, whereas genes downregulated by the cytokine were multiplied by -1 while all other genes were left unchanged and served as background. For each query cell type, genes were ranked using the DESeq2 test statistic from the pseudobulk differential expression analysis. Cytokine signatures containing fewer than 10 total responsive genes were excluded. The union of upregulated and downregulated cytokine-responsive genes was tested as a single gene set using pre-ranked GSEA with 1,000 permutations (71). A normalized enrichment score (NES) indicates concordant enrichment of the cytokine signature, with cytokine-induced genes enriched among pseudobulk upregulated genes and cytokine-repressed genes enriched among pseudobulk downregulated genes. We only considered cytokines with statistically significant enrichment in at least 2 different cell types with concordant effect directions.

Study approval: This research was conducted using the UK Biobank resource under application 145447. UK Biobank has received ethical approval from the North West Multi-Centre Research Ethics Committee (REC reference: 16/NW/0274). All participants provided informed consent at the time of enrollment.

Data availability: Data underlying this research including raw sequencing data are available from UK Biobank for authorized researchers. ALTRA cohort data are publicly available in the following resource: https://apps.allenimmunology.org/aifi/insights/ra-progression/downloads/. Raw sequencing data from ALTRA cohort are available at dbGaP (phs003944.v1.p1).

Author contributions:

MH: Conceptualization, Funding acquisition, Investigation, Methodology, Writing – original draft,

AHS: Conceptualization, Methodology, Supervision, Writing – review & editing

Funding support: This research was supported by the Arthritis Foundation through the 2025 Rheumatoid Arthritis Research Program Pilot Award to MH.

References:

1. Dillon CF, Weisman MH, and Miller FW. Population-based estimates of humoral autoimmunity from the U.S. National Health and Nutrition Examination Surveys, 1960-2014. *PLoS One.* 2020;15(1):e0226516.
2. Waaler E. On the occurrence of a factor in human serum activating the specific agglutintion of sheep blood corpuscles. 1939. *APMIS.* 2007;115(5):422-38; discussion 39.
3. Nielsen SF, Bojesen SE, Schnohr P, and Nordestgaard BG. Elevated rheumatoid factor and long term risk of rheumatoid arthritis: a prospective cohort study. *BMJ.* 2012;345:e5244.
4. Muller K, Manthorpe R, Permin H, Hoier-Madsen M, and Oxholm P. Circulating IgM rheumatoid factors in patients with primary Sjogren's syndrome. *Scand J Rheumatol Suppl.* 1988;75:265-8.
5. Shmerling RH, and Delbanco TL. How useful is the rheumatoid factor? An analysis of sensitivity, specificity, and predictive value. *Arch Intern Med.* 1992;152(12):2417-20.
6. Orge E, Cefle A, Yazici A, Gurel-Polat N, and Hulagu S. The positivity of rheumatoid factor and anti-cyclic citrullinated peptide antibody in nonarthritic patients with chronic hepatitis C infection. *Rheumatol Int.* 2010;30(4):485-8.
7. Newkirk MM. Rheumatoid factors: host resistance or autoimmunity? *Clin Immunol.* 2002;104(1):1-13.
8. He Z, Glass MC, Venkatesan P, Feser ML, Lazaro L, Okada LY, et al. Progression to rheumatoid arthritis in at-risk individuals is defined by systemic inflammation and by T and B cell dysregulation. *Sci Transl Med.* 2025;17(817):eadt7214.
9. Nishimura K, Sugiyama D, Kogata Y, Tsuji G, Nakazawa T, Kawano S, et al. Meta-analysis: diagnostic accuracy of anti-cyclic citrullinated peptide antibody and rheumatoid factor for rheumatoid arthritis. *Ann Intern Med.* 2007;146(11):797-808.
10. van Zanten A, Arends S, Roozendaal C, Limburg PC, Maas F, Trouw LA, et al. Presence of anticitrullinated protein antibodies in a large population-based cohort from the Netherlands. *Ann Rheum Dis.* 2017;76(7):1184-90.
11. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature.* 2014;506(7488):376-81.
12. Smolen JS, Aletaha D, Barton A, Burmester GR, Emery P, Firestein GS, et al. Rheumatoid arthritis. *Nat Rev Dis Primers.* 2018;4:18001.
13. Chen J, Shi S, Li X, Gao F, Zhu X, Feng R, et al. CCL7 promotes macrophage polarization and synovitis to exacerbate rheumatoid arthritis. *iScience.* 2025;28(4):112177.
14. Zhang L, Wang Z, Zheng J, and Zhong Q. RRM2 contributes to pathogenic phenotype of fibroblast-like synoviocytes by activating NF-kappaB signaling and inhibiting ferroptosis in rheumatoid arthritis. *J Orthop Surg Res.* 2025;20(1):976.
15. Sun S, Bay-Jensen AC, Karsdal MA, Siebuhr AS, Zheng Q, Maksymowych WP, et al. The active form of MMP-3 is a marker of synovial inflammation and cartilage turnover in inflammatory joint diseases. *BMC Musculoskelet Disord.* 2014;15:93.
16. Kim W, Min S, Cho M, Youn J, Min J, Lee S, et al. The role of IL-12 in inflammatory activity of patients with rheumatoid arthritis (RA). *Clin Exp Immunol.* 2000;119(1):175-81.

17. Gong Q, Sharma M, Glass MC, Kuan EL, Chander A, Singh M, et al. Multi-omic profiling reveals age-related immune dynamics in healthy adults. *Nature.* 2025;648(8094):696-706.
18. Adler J, Galanko JA, Ammoury R, Benkov KJ, Bousvaros A, Boyle B, et al. HLA DQA1*05 and Risk of Antitumor Necrosis Factor Treatment Failure and Anti-Drug Antibody Development in Children With Crohn's Disease. *Am J Gastroenterol.* 2025;120(5):1076-86.
19. Reppell M, Zheng X, Dreher I, Blaes J, Regan E, Haslberger T, et al. HLA-DQA1*05 Associates With Anti-Tumor Necrosis Factor Immunogenicity and Low Adalimumab Trough Concentrations in Inflammatory Bowel Disease Patients From the SERENE Ulcerative Colitis and Crohn's Disease Studies. *J Crohns Colitis.* 2025;19(1).
20. van der Helm-van Mil AH, Verpoort KN, Breedveld FC, Huizinga TW, Toes RE, and de Vries RR. The HLA-DRB1 shared epitope alleles are primarily a risk factor for anti-cyclic citrullinated peptide antibodies and are not an independent risk factor for development of rheumatoid arthritis. *Arthritis Rheum.* 2006;54(4):1117-21.
21. Arnett FC, Hamilton RG, Reveille JD, Bias WB, Harley JB, and Reichlin M. Genetic studies of Ro (SS-A) and La (SS-B) autoantibodies in families with systemic lupus erythematosus and primary Sjogren's syndrome. *Arthritis Rheum.* 1989;32(4):413-9.
22. Menard L, Samuels J, Ng YS, and Meffre E. Inflammation-independent defective early B cell tolerance checkpoints in rheumatoid arthritis. *Arthritis Rheum.* 2011;63(5):1237-45.
23. Konzett V, and Aletaha D. Management strategies in rheumatoid arthritis. *Nature Reviews Rheumatology.* 2024;20(12):760-9.
24. . NCT00281294 Ctn. A phase 2 study to evaluate the safety, tolerability, and activity of fontolizumab in subjects with active rheumatoid arthritis. ClinicalTrialsgov. https://clinicaltrials.gov/study/NCT00281294.
25. Genovese MC, Kinnman N, de La Bourdonnaye G, Pena Rossi C, and Tak PP. Atacicept in patients with rheumatoid arthritis and an inadequate response to tumor necrosis factor antagonist therapy: results of a phase II, randomized, placebo-controlled, dose-finding trial. *Arthritis Rheum.* 2011;63(7):1793-803.
26. Holers VM, Demoruelle MK, Kuhn KA, Buckner JH, Robinson WH, Okamoto Y, et al. Rheumatoid arthritis and the mucosal origins hypothesis: protection turns to destruction. *Nat Rev Rheumatol.* 2018;14(9):542-57.
27. Willis VC, Demoruelle MK, Derber LA, Chartier-Logan CJ, Parish MC, Pedraza IF, et al. Sputum autoantibodies in patients with established rheumatoid arthritis and subjects at risk of future clinically apparent disease. *Arthritis Rheum.* 2013;65(10):2545-54.
28. Kinslow JD, Blum LK, Deane KD, Demoruelle MK, Okamoto Y, Parish MC, et al. Elevated IgA Plasmablast Levels in Subjects at Risk of Developing Rheumatoid Arthritis. *Arthritis Rheumatol.* 2016;68(10):2372-83.
29. Demoruelle MK, Weisman MH, Simonian PL, Lynch DA, Sachs PB, Pedraza IF, et al. Brief report: airways abnormalities and rheumatoid arthritis-related autoantibodies in subjects without arthritis: early injury or initiating site of autoimmunity? *Arthritis Rheum.* 2012;64(6):1756-61.
30. Sparks JA, Lin TC, Camargo CA, Jr., Barbhaiya M, Tedeschi SK, Costenbader KH, et al. Rheumatoid arthritis and risk of chronic obstructive pulmonary disease or asthma

among women: A marginal structural model analysis in the Nurses' Health Study. *Semin Arthritis Rheum.* 2018;47(5):639-48.
31. Choi H, Han K, Jung JH, Park J, Kim BG, Yang B, et al. Impact of Rheumatoid Arthritis and Seropositivity on the Risk of Non-Cystic Fibrosis Bronchiectasis. *Chest.* 2024;165(6):1330-40.
32. Rangel-Moreno J, Hartson L, Navarro C, Gaxiola M, Selman M, and Randall TD. Inducible bronchus-associated lymphoid tissue (iBALT) in patients with pulmonary complications of rheumatoid arthritis. *J Clin Invest.* 2006;116(12):3183-94.
33. Suda T, Chida K, Hayakawa H, Imokawa S, Iwata M, Nakamura H, et al. Development of bronchus-associated lymphoid tissue in chronic hypersensitivity pneumonitis. *Chest.* 1999;115(2):357-63.
34. Bracke KR, Verhamme FM, Seys LJ, Bantsimba-Malanda C, Cunoosamy DM, Herbst R, et al. Role of CXCL13 in cigarette smoke-induced lymphoid follicle formation and chronic obstructive pulmonary disease. *Am J Respir Crit Care Med.* 2013;188(3):343-55.
35. Frija-Masson J, Martin C, Regard L, Lothe MN, Touqui L, Durand A, et al. Bacteria-driven peribronchial lymphoid neogenesis in bronchiectasis and cystic fibrosis. *Eur Respir J.* 2017;49(4).
36. Haruna Y, Kanda T, Honda M, Takao T, and Hayashi N. Detection of hepatitis C virus in the bile and bile duct epithelial cells of hepatitis C virus-infected patients. *Hepatology.* 2001;33(4):977-80.
37. Ma KK, Nelson JL, Guthrie KA, Dugowson CE, and Gammill HS. Adverse pregnancy outcomes and risk of subsequent rheumatoid arthritis. *Arthritis Rheumatol.* 2014;66(3):508-12.
38. Garvie CW, Hagman J, and Wolberger C. Structural studies of Ets-1/Pax5 complex formation on DNA. *Mol Cell.* 2001;8(6):1267-76.
39. Li H, Ji M, Klarmann KD, and Keller JR. Repression of Id2 expression by Gfi-1 is required for B-cell and myeloid development. *Blood.* 2010;116(7):1060-9.
40. Mackay F, and Schneider P. Cracking the BAFF code. *Nat Rev Immunol.* 2009;9(7):491-502.
41. Casolari DA, Makri M, Yoshida C, Muto A, Igarashi K, and Melo JV. Transcriptional suppression of BACH2 by the Bcr-Abl oncoprotein is mediated by PAX5. *Leukemia.* 2013;27(2):409-15.
42. van Keimpema M, Gruneberg LJ, Mokry M, van Boxtel R, van Zelm MC, Coffer P, et al. The forkhead transcription factor FOXP1 represses human plasma cell differentiation. *Blood.* 2015;126(18):2098-109.
43. Horcher M, Souabni A, and Busslinger M. Pax5/BSAP maintains the identity of B cells in late B lymphopoiesis. *Immunity.* 2001;14(6):779-90.
44. John SA, Clements JL, Russell LM, and Garrett-Sinha LA. Ets-1 regulates plasma cell differentiation by interfering with the activity of the transcription factor Blimp-1. *J Biol Chem.* 2008;283(2):951-62.
45. Espeli M, Bashford-Rogers R, Sowerby JM, Alouche N, Wong L, Denton AE, et al. FcgammaRIIb differentially regulates pre-immune and germinal center B cell tolerance in mouse and human. *Nat Commun.* 2019;10(1):1970.

46. Jackson TA, Haga CL, Ehrhardt GR, Davis RS, and Cooper MD. FcR-like 2 Inhibition of B cell receptor-mediated activation of B cells. *J Immunol.* 2010;185(12):7405-12.
47. Glass DR, Tsai AG, Oliveria JP, Hartmann FJ, Kimmey SC, Calderon AA, et al. An Integrated Multi-omic Single-Cell Atlas of Human B Cell Identity. *Immunity.* 2020;53(1):217-32 e5.
48. Koncz G, and Hueber AO. The Fas/CD95 Receptor Regulates the Death of Autoreactive B Cells and the Selection of Antigen-Specific B Cells. *Front Immunol.* 2012;3:207.
49. Jacobson BA, Rothstein TL, and Marshak-Rothstein A. Unique site of IgG2a and rheumatoid factor production in MRL/lpr mice. *Immunol Rev.* 1997;156:103-10.
50. Heutz JW, Looijen AEM, Kuijpers J, Schreurs MWJ, van der Helm-van Mil AHM, and de Jong PHP. The prognostic value of IgA anti-citrullinated protein antibodies and rheumatoid factor in an early arthritis population with a treat-to-target approach. *Immunol Res.* 2024;72(5):982-90.
51. Rantapaa-Dahlqvist S, de Jong BA, Berglin E, Hallmans G, Wadell G, Stenlund H, et al. Antibodies against cyclic citrullinated peptide and IgA rheumatoid factor predict the development of rheumatoid arthritis. *Arthritis Rheum.* 2003;48(10):2741-9.
52. Majka DS, Deane KD, Parrish LA, Lazar AA, Baron AE, Walker CW, et al. Duration of preclinical rheumatoid arthritis-related autoantibody positivity increases in subjects with older age at time of disease diagnosis. *Ann Rheum Dis.* 2008;67(6):801-7.
53. Nielen MM, van Schaardenburg D, Reesink HW, van de Stadt RJ, van der Horst-Bruinsma IE, de Koning MH, et al. Specific autoantibodies precede the symptoms of rheumatoid arthritis: a study of serial measurements in blood donors. *Arthritis Rheum.* 2004;50(2):380-6.
54. Saraiva M, Vieira P, and O'Garra A. Biology and therapeutic potential of interleukin-10. *J Exp Med.* 2020;217(1).
55. van Roon J, Wijngaarden S, Lafeber FP, Damen C, van de Winkel J, and Bijlsma JW. Interleukin 10 treatment of patients with rheumatoid arthritis enhances Fc gamma receptor expression on monocytes and responsiveness to immune complex stimulation. *J Rheumatol.* 2003;30(4):648-51.
56. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature.* 2018;562(7726):203-9.
57. Singh JA, Holmgren AR, and Noorbaloochi S. Accuracy of Veterans Administration databases for a diagnosis of rheumatoid arthritis. *Arthritis Rheum.* 2004;51(6):952-7.
58. Benjamini Y, and Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological).* 1995;57(1):289-300.
59. Luo Y, Kanai M, Choi W, Li X, Sakaue S, Yamamoto K, et al. A high-resolution HLA reference panel capturing global population diversity enables multi-ancestry fine-mapping in HIV host response. *Nat Genet.* 2021;53(10):1504-16.
60. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, and Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* 2015;4:7.
61. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* 2023;51(D1):D638-D46.

62. Thompson DJ, Wells D, Selzam S, Peneva I, Moore R, Sharp K, et al. A systematic evaluation of the performance and properties of the UK Biobank Polygenic Risk Score (PRS) Release. *PLoS One.* 2024;19(9):e0307270.
63. DeLong ER, DeLong DM, and Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* 1988;44(3):837-45.
64. Eldjarn GH, Ferkingstad E, Lund SH, Helgason H, Magnusson OT, Gunnarsdottir K, et al. Large-scale plasma proteomics comparisons through genetics and disease associations. *Nature.* 2023;622(7982):348-58.
65. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015;43(7):e47.
66. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, et al. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* 2005;33(Database issue):D428-32.
67. Liao Y, Wang J, Jaehnig EJ, Shi Z, and Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* 2019;47(W1):W199-W205.
68. Muzellec B, Telenczuk M, Cabeli V, and Andreux M. PyDESeq2: a python package for bulk RNA-seq differential expression analysis. *Bioinformatics.* 2023;39(9).
69. Love MI, Huber W, and Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15(12):550.
70. Oesinghaus L, Becker S, Vornholz L, Papalexi E, Pangallo J, Moinfar AA, et al. A single-cell cytokine dictionary of human peripheral blood. *bioRxiv.* 2025.
71. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A.* 2005;102(43):15545-50.

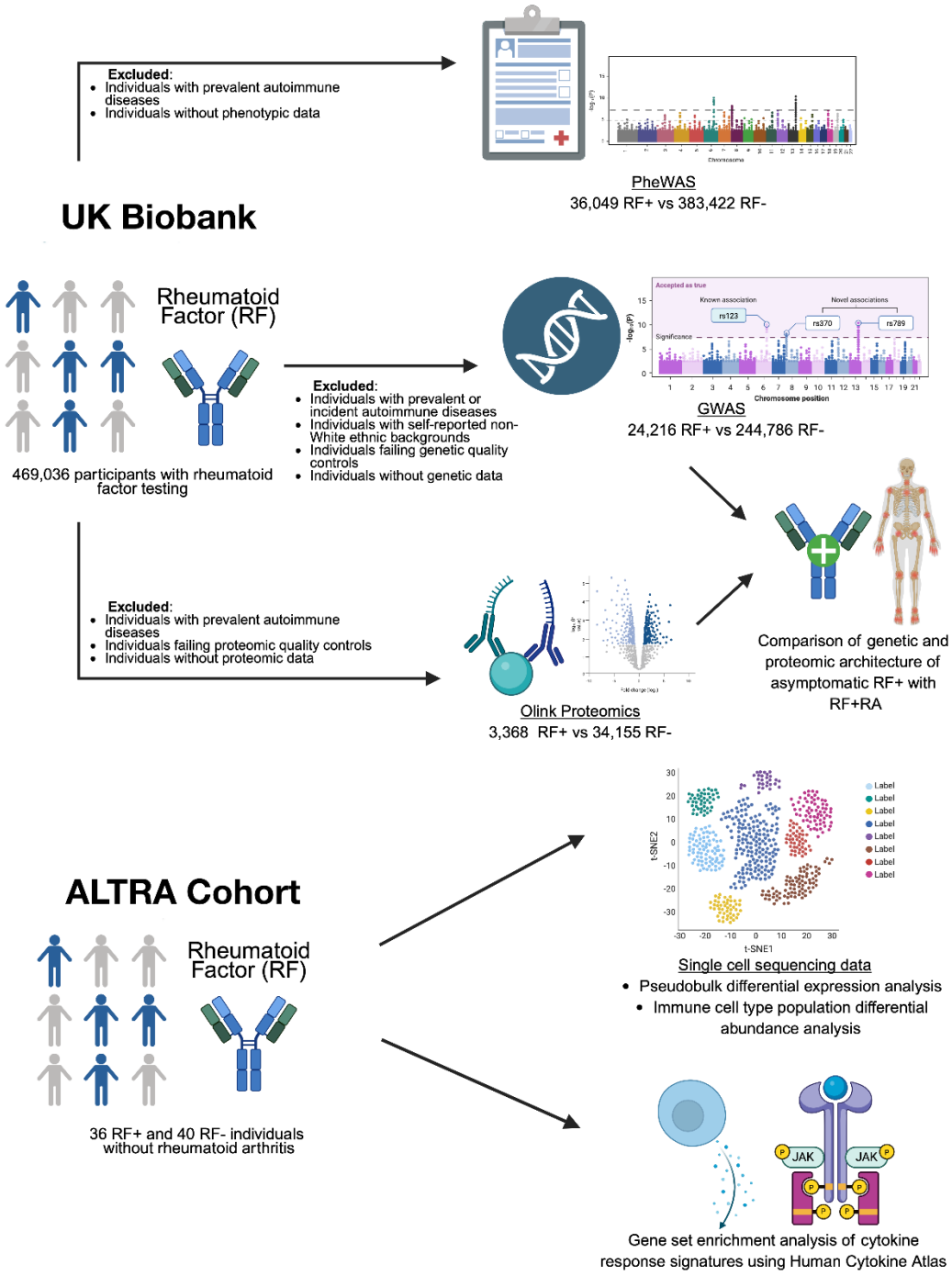

Figure 1. Diagram depicting the study design. GWAS, Genome-wide association study. PheWAS, Phenome-wide association study.

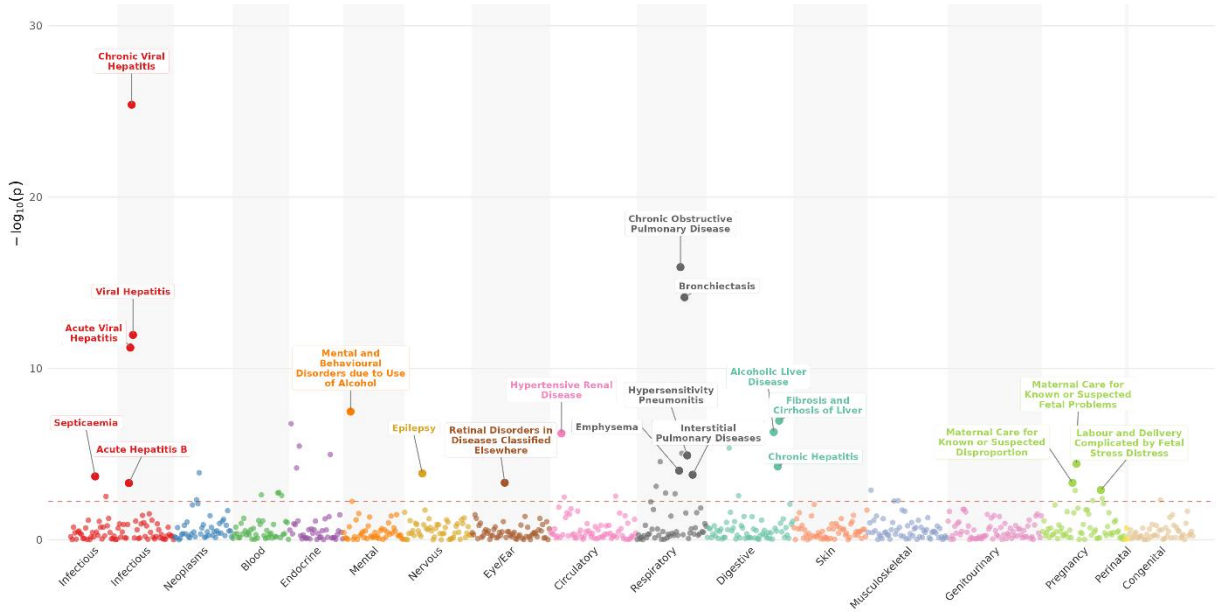

Figure 2. Manhattan plot for phenome-wide association study of rheumatoid factor positivity among individuals without autoimmune disease diagnosis in the UK Biobank.

The x-axis displays individual phenotypes grouped by clinical category, and the y-axis represents $-\log_{10}$ P values. Each semi-transparent point represents a distinct phenotypic trait, color-coded by its respective clinical domain. Alternating background shading delineates these broader categories. The horizontal dashed red line indicates the Benjamini-Hochberg-corrected false discovery rate (FDR) of 10% significance threshold. To emphasize the most robust findings, the top 20 most significant associations that also possess a clinically notable effect size (Odds Ratio > 1.2 or < 0.83) are highlighted with larger, outlined markers and explicitly labeled with their condition descriptions.

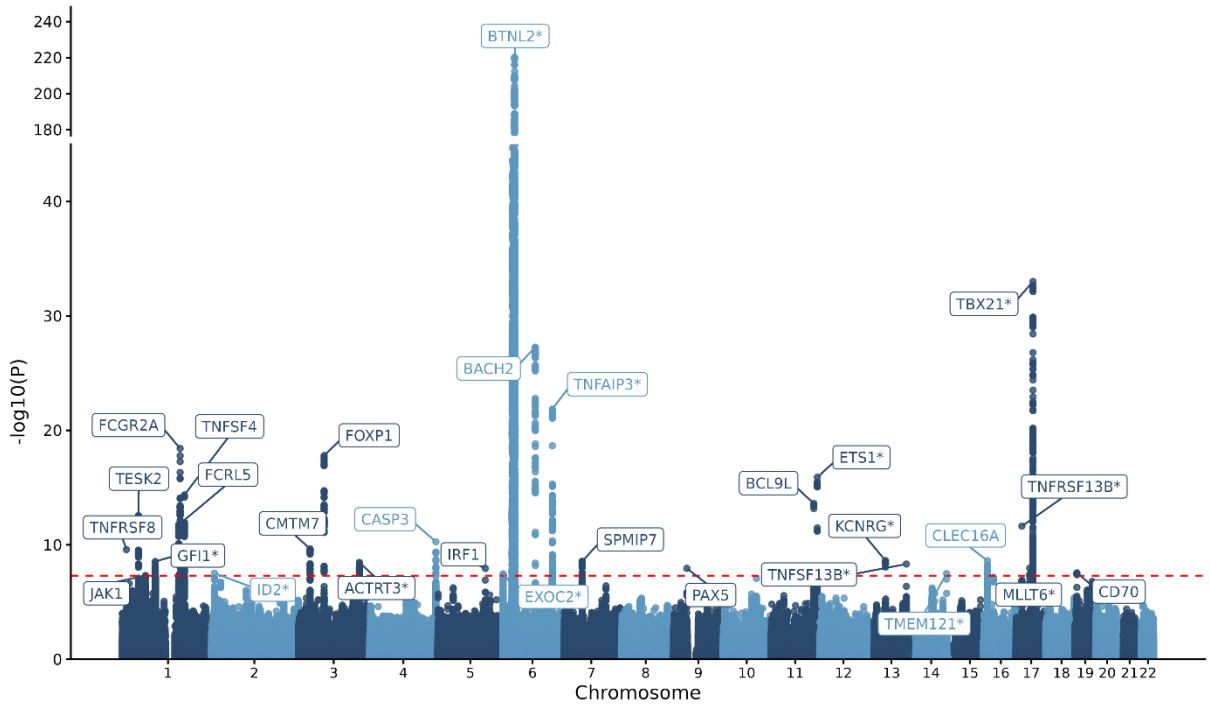

Figure 3. Manhattan plot for genome-wide association study (GWAS) of rheumatoid factor positivity among individuals without autoimmune disease diagnosis in the UK Biobank

The Y and X axes refer to the $-\log_{10}$ P values and chromosome positions, respectively. The red horizontal line represents the genome-wide association threshold (*P*-value < 5×10^{-8}). Blue and dark blue dots represent individual genetic variants. Variants with a significant association at the genome-wide level were annotated to the closest protein coding gene based on hg38 build. *Labels loci with intergenic variants mapped to the closest protein coding gene, while the remaining variants are located within the annotated gene.

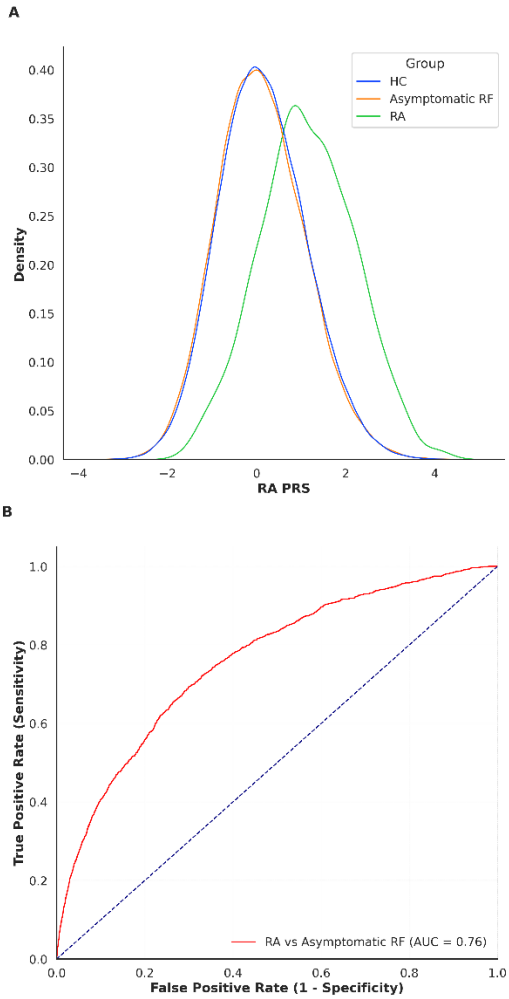

Figure 4. Distribution and discriminative performance of the rheumatoid arthritis polygenic risk score (RA-PRS) among individuals with asymptomatic rheumatoid factor production.

(A) Density plot illustrating the distribution of the RA-PRS across the study groups. **(B)** Receiver Operating Characteristic (ROC) curve evaluating the ability of the RA PRS to discriminate between patients with RF-positive RA and asymptomatic RF-positive individuals. The red solid line indicates the model's performance, with the Area Under the Curve (AUC) reported in the legend. The dashed navy line represents the baseline of a random classifier (AUC = 0.50). RA, rheumatoid arthritis; PRS, polygenic risk score; RF, rheumatoid factor; ROC, receiver operating characteristic; AUC, area under the curve; FPR, false positive rate; TPR, true positive rate.

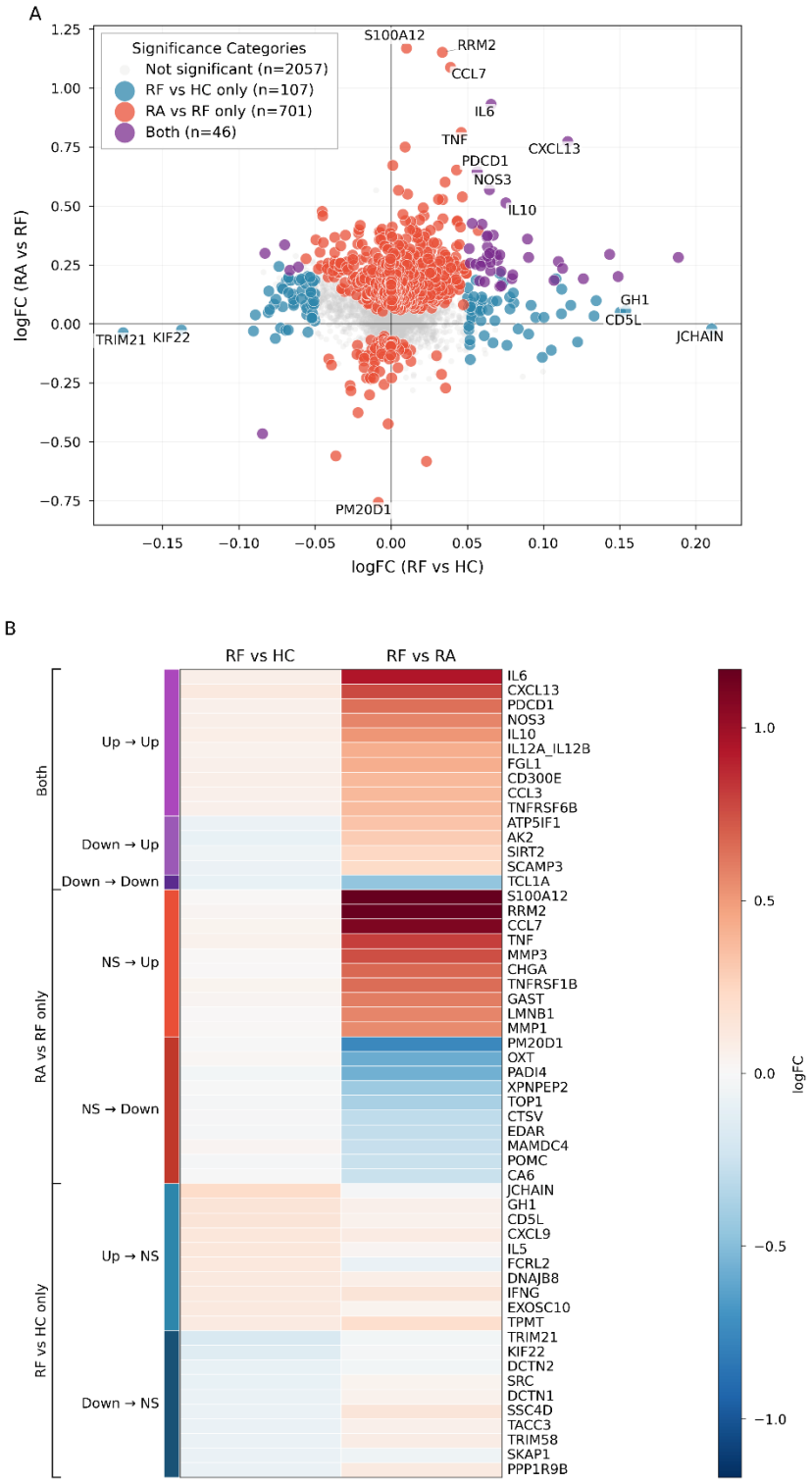

Figure 5. Comparative proteomic profiling of asymptomatic rheumatoid factor (RF) production and RF-positive rheumatoid arthritis.

(A) Scatter plot of protein-level log fold changes (logFC) in asymptomatic RF-positive vs RF-negative healthy controls (x-axis) versus RF-positive RA vs RF-positive without RA (y-axis). Proteins were classified using a combined significance criterion requiring both adjusted $P < 0.01$ and $|logFC| > 0.05$ in each comparison. Points are colored by significance category: RF vs HC only (significant only in RF vs healthy controls), RA vs RF only (significant only in RA vs RF), Both (significant in both comparisons), and not significant. Dotted gray lines mark the $|logFC| = 0.05$ threshold. Proteins which were top hits defined by absolute logFC within single-comparison groups and by Euclidean distance for proteins significant in both comparisons are annotated in the plot. **(B)** Heatmap of proteins showing logFC values for RF vs healthy controls and RA vs RF comparisons. Proteins are grouped by direction of change across comparisons and are ranked within direction groups by their logFC magnitude. Up to 10 proteins with most statistically significant change in each direction group was depicted. “Up” and “Down” represents upregulation and downregulation in the respective analyses while “NS” represents no statistically significant change. Direction group names such as “Up to Up” or “Up to Down” represent the direction of change for the depicted protein in RF vs healthy controls and RA vs RF comparisons, respectively. Cell colors represent logFC magnitude and direction (centered at 0).

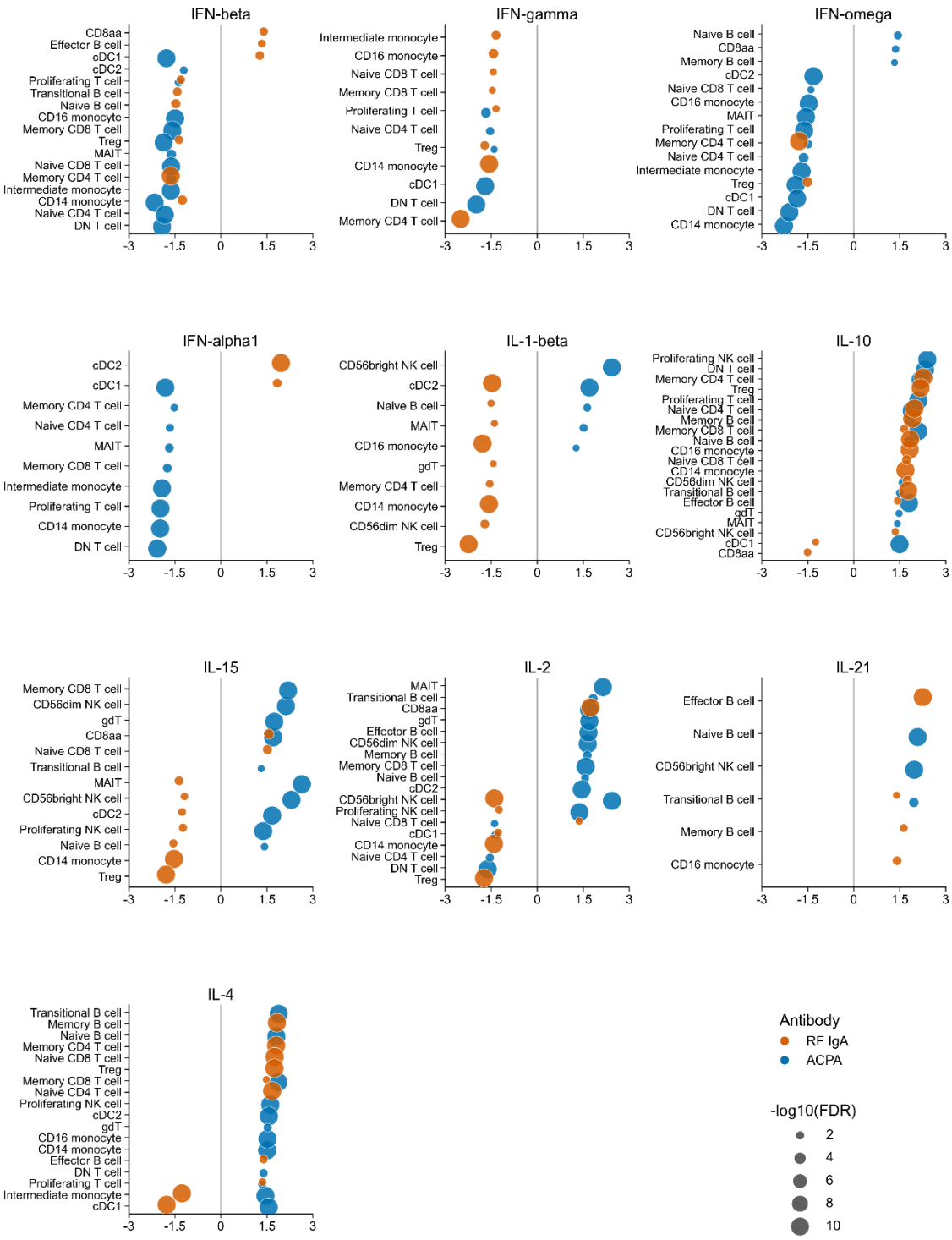

Figure 6. Cell-type specific enrichment of cytokine signatures associated with IgA RF and ACPA.

Dot plot displaying the normalized enrichment scores (NES) for targeted cytokine gene signatures across immune cell populations. Gene set enrichment analysis (GSEA) was performed independently for IgA rheumatoid factor (RF) and anti-citrullinated protein antibodies (ACPA) pseudobulk differential expression analysis results. The x-axis represents the NES, where positive and negative values indicate up- and downregulation of the respective cytokine signatures. Dot color denotes the associated autoantibody (orange for IgA RF; blue for ACPA). The size of each dot is proportional to the statistical significance of the enrichment.

Table 1. Demographics and clinical characteristics of rheumatoid factor positivity at enrollment in the UK Biobank

	RF-negative (n=425,446)	**RF (1-3xULN)** (n=31,955)	**RF (3-6xULN)** (n=6,479)	**RF (6-12xULN)** (n=2,833)	**RF (>12xULN)** (n=2,323)
Age (mean ± SD)	56.9 ± 8.1	58.2 ± 7.7	58.7 ± 7.7	58.9 ± 7.6	59.6 ± 7.3
Female sex (n, %)	229,690 (54)	17,887 (56)	3,575 (55.2)	1,659 (58.6)	1,462 (62.9)
White (n, %)	400,749 (94.2)	30,377 (95.1)	6,105 (94.2)	2,666 (94.1)	2,167 (93.3)
Asian* (n, %)	8,267 (1.9)	556 (1.7)	120 (1.9)	61 (2.2)	62 (2.7)
Black (n, %)	6,690 (1.6)	402 (1.3)	90 (1.4)	43 (1.5)	41 (1.8)
Other Populations (n, %)	9,740 (2.3)	620 (1.9)	164 (2.5)	63 (2.2)	53 (2.3)
RA, all, (n, %)	3768 (0.9)	948 (3)	564 (8.7)	469 (16.6)	662 (28.5)
RA, DMARD, (n, %)	1074 (0.25)	524 (1.64)	380 (5.9)	333 (11.8)	487 (21)
Autoimmune disease excluding RA (n, %)	17,484 (4.1)	1,442 (4.5)	297 (4.6)	129 (4.6)	96 (4.1)

Age reported at sample collection. N, case numbers; RA, rheumatoid arthritis; RF, rheumatoid factor; SD, standard deviation; DMARD, disease-modifying antirheumatic drug. Other Populations: Chinese, Other ethnic group, “Do not know”, “Prefer not to answer”.

*UK Biobank survey had a separate category for self-reported Chinese ethnic background. RA, all, at least one RA diagnosis code. RA, DMARD, at least one diagnosis RA code and a self-reported DMARD medication use.

Statistical comparisons across RF titer groups and RF-negative were conducted by one-way ANOVA (age) and chi-squared tests (rest of the variables). All comparisons were statistically significant ($p < 0.05$) but the effect sizes were negligible for demographic variables ($\eta^2 = 0.003$ for age; Cramér's $V < 0.02$ for sex, ethnic background, and autoimmune diseases excluding RA), indicating that differences are attributable to very large sample size rather than clinically meaningful variation. Meaningful effect sizes were only observed for RA, all ($V = 0.21$) and RA, DMARD ($V = 0.24$) prevalence across RF titer groups.