# Cell-specific gene networks and drivers in rheumatoid arthritis synovial tissues

Aurelien Pelissier[1,2,3]*, Teresina Laragione[4], Percio S. Gulko[4]*† and María Rodríguez Martínez[2,5]*†

[1]Institute of Computational Life Sciences, Zürich University of Applied Sciences (ZHAW), Wädenswil, Switzerland, [2]AI for Scientific Discovery, IBM Research Europe, Rüschlikon, Switzerland, [3]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland, [4]Division of Rheumatology, Icahn School of Medicine at Mount Sinai, New York, NY, United States, [5]Department of Biomedical Informatics & Data Science, Yale School of Medicine, New Haven, CT, United States

Rheumatoid arthritis (RA) is a common autoimmune and inflammatory disease characterized by inflammation and hyperplasia of the synovial tissues. RA pathogenesis involves multiple cell types, genes, transcription factors (TFs) and networks. Yet, little is known about the TFs, and key drivers and networks regulating cell function and disease at the synovial tissue level, which is the site of disease. In the present study, we used available RNA-seq databases generated from synovial tissues and developed a novel approach to elucidate cell type-specific regulatory networks on synovial tissue genes in RA. We leverage established computational methodologies to infer sample-specific gene regulatory networks and applied statistical methods to compare network properties across phenotypic groups (RA versus osteoarthritis). We developed computational approaches to rank TFs based on their contribution to the observed phenotypic differences between RA and controls across different cell types. We identified 18 (fibroblast-like synoviocyte), 16 (T cells), 19 (B cells) and 11 (monocyte) key regulators in RA synovial tissues. Interestingly, fibroblast-like synoviocyte (FLS) and B cells were driven by multiple independent co-regulatory TF clusters that included MITF, HLX, BACH1 (FLS) and KLF13, FOSB, FOSL1 (B cells). However, monocytes were collectively governed by a single cluster of TF drivers, responsible for the main phenotypic differences between RA and controls, which included RFX5, IRF9, CREB5. Among several cell subset and pathway changes, we also detected reduced presence of Natural killer T (NKT) cells and eosinophils in RA synovial tissues. Overall, our novel approach identified new and previously unsuspected Key driver genes (KDG), TF and networks and should help better understanding individual cell regulation and co-regulatory networks in RA pathogenesis, as well as potentially generate new targets for treatment.

# 1 Introduction

Rheumatoid arthritis (RA) is a systemic autoimmune disorder characterized by synovial inflammation and hyperplasia that may lead to joint destruction (1, 2). With a prevalence estimated between 0.5 and 1% (3), it is one of the most common chronic inflammatory diseases. The risk of developing RA peaks around age 50 (3), and similarly to most autoimmune diseases, females are affected 2 to 3 times more than males (3, 4). The development of biologic disease-modifying anti-rheumatic drugs (bDMARDs) and JAK inhibitors targeting various inflammatory pathways has significantly improved disease control and outcomes (5, 6), yet a considerable number of RA patients still have an inadequate response to therapy (7). As the development and progression of RA involve dynamic interactions between multiple genetic and environmental factors, understanding the heterogeneous pathophysiological processes remains a major challenge (8).

Genome-Wide Association Studies (GWAS) have significantly improved the understanding of the disease's genetic underpinnings and identified multiple genetic loci associated with susceptibility (9, 10). However, these loci only explain a fraction of the overall heritability and phenotypic variance of RA (11). While new whole-genome and whole-exome sequencing studies are likely to identify additional rare variants previously undetectable with GWAS techniques, our ability to translate these results into disease understanding and novel therapeutics remains limited. At the transcriptomic level, Differential Gene Expression (DGE) studies have compared gene expression profiles between RA patients and healthy controls and identified numerous pathways implicated in inflammation, antigen presentation, hypoxia, and apoptosis during RA (12, 13). These data have not only deepened the comprehension of RA's pathogenesis but have also offered promising targets for therapeutic intervention. However, although DEG studies are effective for discovery, the large number of detected genes often obscures the identification of key regulatory or therapeutically actionable genes. Moreover, DEG studies often highlight pathways that are already well-characterized, underscoring the need for new approaches to integrate established knowledge and data-driven computational approaches.

Further complicating the analysis, most studies still rely on bulk data, wherein the cellular composition of tissues significantly confounds the molecular findings (14, 15). In other words, the observed differential expression between RA and controls may primarily arise from disparities in cell composition rather than differences in cellular gene expression. For instance, RA synovial tissues exhibit significantly more leukocytes than control groups (15, 16), and a meta-analysis has revealed variable DEGs across datasets, with approximately half of the DEGs up-regulated in synovial tissues being down-regulated in blood samples (12). Recent single-cell RNA sequencing (scRNA-Seq) sequencing studies are offering valuable insights into disease traits at the single-cell level (15, 16), however, the exploitation of these data is still challenged by limited patient numbers, batch effects, and sparse data (17). Notably, the inference of gene regulatory networks (GRNs) from scRNA-Seq has proven to be particularly challenging, in part because of the difficulty of identifying cell type-specific regulatory interactions from heterogeneous samples (18–20).

Addressing the cell composition challenge requires the development of novel approaches for identifying transcription factors (TFs) and their associated regulatory signatures in a cell type-specific manner. TFs play a pivotal role in regulating gene expression in a tissue-specific manner (21), and with an estimated count of 1,000 TFs in humans (22), identifying those that govern the phenotypic traits of RA synovial cells may offer opportunities for discovering novel therapeutic targets.
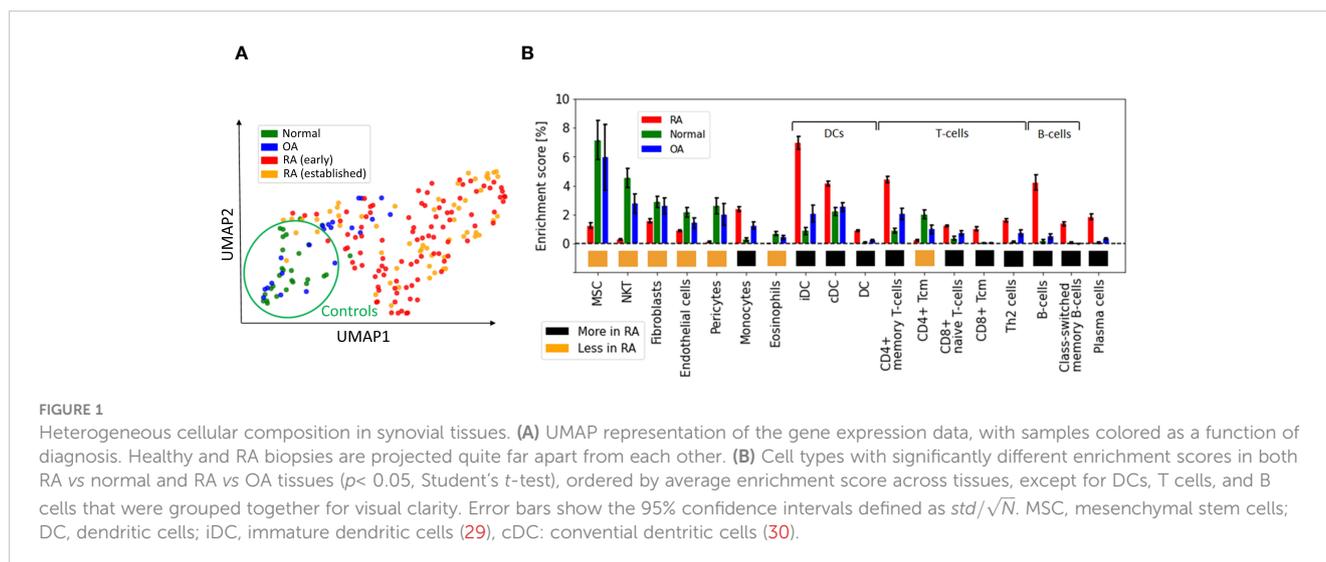
In this study, we provide a comprehensive analysis of gene regulation of RA in synovial tissues. Unlike previous studies in RA, which relied on the inference of cohort-averaged GRNs (23–25), our approach enabled us to gain new insights into sample-specific regulatory mechanisms by leveraging bulk gene expression data (15, 26) and to identify TFs driving gene expression in RA-associated cell types. Namely, we identify key RA regulators in a cell type-specific manner, such as IRF8 in monocytes, STAT5B in B cells, ELF4 in T cells, and MITF in fibroblast-like synoviocytes (FLS). We next construct a co-regulation network in each cell type by evaluating the correlation among the target genes that are shared by the identified TFs. In FLS and B cells, we observe a common regulatory pattern characterized by a strong correlation among all TFs, while distinct co-regulation clusters emerge independently in the other cell types. Our computational modeling and findings indicate new targets for cell-specific treatment strategies in RA and provide novel insights into the cell-specific regulation of RA pathogenesis.

# 2 Results

## 2.1 Heterogeneous cellular composition accounts for most of the gene expression variability across RA biopsies

We exploited analyzed public RNA-Seq data of synovial biopsies from 28 healthy control samples, 152 individuals with RA, and 22 patients with osteoarthritis (OA) (26). Each sample contained the gene expression profile of 25,000 genes. As healthy biopsies are rare and difficult to obtain (27), we compared RA synovial biopsies to both OA and healthy samples.

We first used UMAP (28) to visualize the samples in two dimensions, and observed significant distributional differences between the gene expression profiles of RA biopsies and controls, the latter comprising OA and healthy samples (Figure 1A). A DEG analysis revealed that more than half of the genes (~15k) were significantly differentially expressed [p< 0.05; Student's t-test with Benjamini-Hochberg correction (31)]. This uncommonly high number of DEGs, which is typically on the order of a few hundred in studies of blood samples from patients with other diseases, such as coronary artery disease (32), obesity (33, 34), diabetes (35, 36) or kidney (37), led us to hypothesize that the observed variability might arise from cellular heterogeneity across synovial biopsies, rather than from intra-cellular gene expression differences between the two tested groups. To test this hypothesis, we used xCell (38) to estimate the relative proportions of the various cell populations present in our samples. xCell is a signature-based method that employs single-

FIGURE 1
Heterogeneous cellular composition in synovial tissues. **(A)** UMAP representation of the gene expression data, with samples colored as a function of diagnosis. Healthy and RA biopsies are projected quite apart from each other. **(B)** Cell types with significantly different enrichment scores in both RA *vs* normal and RA *vs* OA tissues ($p < 0.05$, Student's *t*-test), ordered by average enrichment score across tissues, except for DCs, T cells, and B cells that were grouped together for visual clarity. Error bars show the 95% confidence intervals defined as $std/\sqrt{N}$. MSC, mesenchymal stem cells; DC, dendritic cells; iDC, immature dendritic cells (29); cDC: convential dentritic cells (30).

sample gene set enrichment analysis to compute an enrichment score per sample. This score is indicative of the relative proportion of each cell type within a sample (Methods Section 2.2). xCell detected significantly increased enrichment scores for several cell populations in both RA vs normal and RA vs OA tissue comparisons (Figure 1B; *p*< 0.05, Student's *t*-test). Interestingly, biopsies from early and established RA had similar gene expression signatures and cellularity, suggesting that similar cell types and processes regulate disease through the different stages of progression (Supplementary Section A.1, Supplementary Figure S1).

Our strategy to deconvolute synovial cellularity was able to identify populations known to be expanded in RA, such as T cells, B cells, plasma cells, dendritic cells (DC), and monocytes (15, 16, 39). To further validate these results, we used CIBERSORT (40), an alternative deconvolution method for bulk RNA expression. Compared to xCell, CIBERSORT focuses exclusively on immune cells and employs a different algorithmic approach, namely, a linear regression based on a predefined gene expression matrix of known cell types (Supplementary Section A.2, Supplementary Figure S2). CIBERSORT and xCell findings were similar regarding the enrichment of plasma cells, memory B cells, CD4+ memory T cells, and dentritic cells (DC), thus confirming the different synovial cellularity between RA and controls (Supplementary Table S1).

In addition to these well-documented cell populations, xCell also identified other cell-specific traits associated with RA, such as an increased representation of both immature dendritic cells (iDC) (29) and conventional dendritic cells (cDC) (30, 41), as well as a previously unrecognized reduction of Natural killer T cells (NKT), pericytes and eosinophils, compared with controls. These analyses revealed the high variability in cellular composition within synovial tissues, which may explain a significant fraction of the gene expression variability observed among RA patients (14). To quantify this effect, we calculated the expression variance explained by cellular composition using the R-squared score. This statistical measure represents the proportion of total gene expression variance explained by the enrichment scores estimated

with xCell. xCell scores were used as inputs in a linear regression model to predict gene expression values. The R-squared score was then derived by comparing the model's predictions with the original observations. Significantly, we found that the 18 cell types shown in Figure 1B account for 61% of the variance in gene expression (Supplementary Section A.3, Supplementary Figure S3).

## 2.2 Gene regulation in RA *vs* control samples differs widely across synovial cell types

The measured gene expression in synovial tissues is a mixture of gene expression profiles from different cell types, which complicates the task of extracting cell type-specific gene regulatory information. To address this challenge, we adjusted the gene expression values to correct for cellular composition biases using a linear model (Methods Section 2.3). The corrected gene expression data served as a basis for constructing a gene regulatory network, unbiased by cell type heterogeneity. In addition, we also leveraged cell type-specific bulk RNA-Seq data from RA and OA synovial fibroblasts, monocytes, B cells, and T cells (15) to complement our analyses of the synovial tissue biopsies, resulting in 5 independent gene expression datasets from synovial tissues (Table 1).

Focusing on the cell type-specific datasets, we used a Student's t-test to identify the genes that were differentially expressed between RA and OA samples. Here, we considered OA samples as the control group due to the unavailability of biopsies from healthy patients. From this analysis, we obtained a differential expression score for each gene (denoted $t_{expr}$). Next, we examined the correlation of these scores $t_{expr}$ across each pair of cell type-specific datasets (Methods Section 2.6). The correlation among the genes' differential expression scores was notably low (< 0.1) across all considered datasets, suggesting disparate regulatory mechanisms for each RA-associated cell type (Supplementary Figure S7).

TABLE 1 List of datasets of RNA-Seq data from synovial tissues and bipartite networks computed from them.

| Dataset/net-works name | Reference | Method | Number of patients/networks |
|---|---|---|---|
| PANDA_Synovial_Tissue | (26) | PANDA, LIONESS | 152 RA, 22 OA |
| PANDA_FLS | (15) | PANDA, LIONESS | 18 RA, 12 OA |
| PANDA_Synovial_Monocyte | (15) | PANDA, LIONESS | 17 RA, 13 OA |
| PANDA_Synovial_B cell | (15) | PANDA, LIONESS | 8 RA, 7 OA |
| PANDA_Synovial_T cell | (15) | PANDA, LIONESS | 17 RA, 13 OA |

An individual network is computed for each RA and OA sample.

Nonetheless, this approach failed to provide insights into the regulatory connections among the DEGs. It also did not reveal whether these connections involved TF-mediated inhibition, activation, or co-regulation. To investigate this question, we assembled gene regulatory networks of the synovial tissue and its constituent cell types by combining RA gene expression data with information about TF-binding motifs [CIS-BP (42)] and protein-protein interactions [StringDB (43)]. We leveraged PANDA (44), a computational strategy designed to optimize the alignment between gene expression data and pre-existing knowledge (Methods Section 2.9), and used LIONESS (45) to estimate individual gene regulatory networks for each sample in our cohort. In these sample-specific networks, each edge connecting a TF and a target gene (TG) has an associated weight that represents the likelihood of a regulatory interaction between the TF and the TG. We leveraged this collection of networks to evaluate whether the TF-TG interaction edge weights differ significantly between the RA and control samples and to identify potential TFs that might regulate the regulatory differences (Figure 2A). More precisely, we compared the differences in edge weights between RA and OA samples using a Student's t-test, and

obtained a score $t_{edge}$ for each edge TF → TG. For each cell type, we assembled all $t_{edge}$ scores in a *differential* $_{GRN}$ (dGRN) network, which highlights the edges that are differentially regulated between RA and OA (Figure 2A). Then, to quantify TFs regulatory function, we define a TF regulatory score ($t_{reg}$) defined as the mean of the absolute values of the edge scores ($|t_{edge}|$) between the TF and all its TGs (Method 2.10). As a positive (respectively negative) $t_{edge}$ indicates a strong upregulation (respectively downregulation) TF-TG interaction in RA with respect to OA (control) tissues, TFs with a high $t_{reg}$ are potential key regulators to explain the differential gene expression between RA and OA.

Table 2 lists the top 10 TFs ranked by their regulatory scores for each synovial tissue cell type. Interestingly, several TFs, including RFX5, CEBPZ, SCRT1, and MXI1, are in the top 10 in more than one cell type, suggesting that these TF are broader key regulators in synovial tissues. We examined the correlation between the TF scores ($t_{reg}$) on each pair of networks and obtained similar results to those previously observed with the cell-specific gene expression signatures, i.e. a low overall correlation (< 0.1) indicating distinct regulatory mechanisms across cell types (Figure 2B).

For additional insight into the pathways involved in RA in each cell type, we ran a pathway enrichment analysis on the major TF regulators using a collection of pathways compiled from the Gene Ontology (GO) (46), KEGG (47), and Reactome (48) databases. First, we selected the TFs with a $t_{reg}$ score higher than one standard deviation above the mean (Z-statistics > 1). This led to the selection of between 40 and 90 TFs per tissue and cell type (Supplementary Figure S8A). Interestingly, the overlap between the selected TFs was low, and only 7 TFs were shared by 3 cell types or more (Supplementary Figure S8B). Because the enrichments were run with TFs exclusively, we removed any terms associated with RNA and DNA transcription, as these are ubiquitous processes and not likely to be RA-specific. After this filtering, we obtained between 60 and 120 significantly enriched pathways (i.e. $p_{adj}$ < 0.05) for each cell type (Supplementary Figure S9A). In Figure 3, we show the 10 most significant pathways for each cell type. As expected, well-established RA pathways are evident across multiple cell types, such as TNF (49), IRF (50) and ATF6 (51), along with pathways



FIGURE 2
(A) Networks were inferred from the gene expression profiles of RA and OA biopsies in different cell types. Network edges are endowed with weights representing the probability of regulatory interactions between a transcription factor (TF) and a target gene (TG). The analysis of differences in edge weights between RA and OA facilitated the construction of a differential GRN (dGRN) for each cell type. The dGRN was used to compute a regulatory score for each TF in each cell type. (B) Heatmap of the Pearson correlation between the TF regulatory scores in each tissue type, i.e. synovial tissue, monocyte, FLS, B cell, and T cell).
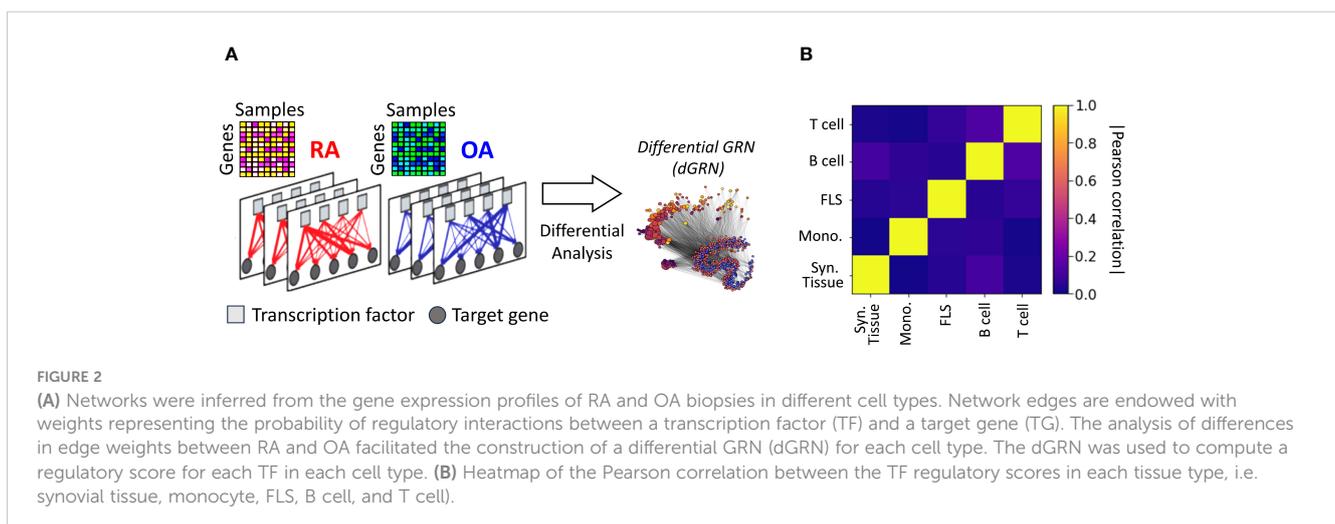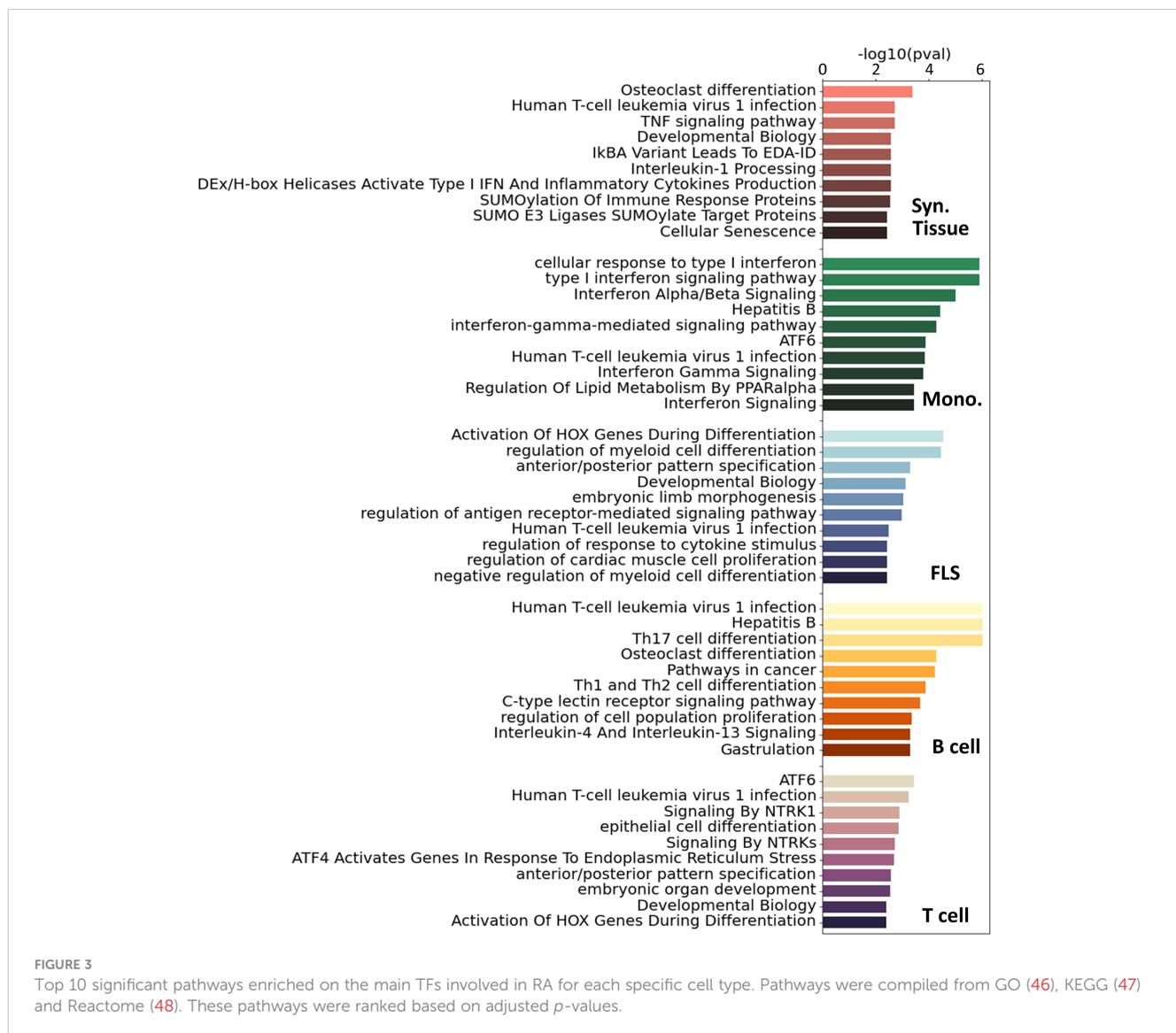
TABLE 2  Top 10 ranked TF regulators in each synovial cell type, along with their Z-scores in parentheses.

| Rank | Syn. Tissue | Syn. Monocyte | Syn. Fibroblast (FLS) | Syn. B cell | Syn. T cell |
|---|---|---|---|---|---|
| 1 | ZNF282 (7.01) | NFE2 (5.71) | RORC (3.39) | JUN (6.59) | HAND1 (5.85) |
| 2 | SCRT2 (6.30) | **CEBPZ** (5.63) | NKX2-1 (3.37) | STAT5B (4.89) | REST (4.97) |
| 3 | NR6A1 (4.06) | IRF8 (5.59) | HOXA1 (3.20) | CTCF (4.04) | ELF4 (4.59) |
| 4 | **SCRT1** (3.70) | PBX3 (5.60) | ETV2 (2.67) | TCF3 (3.45) | EHF (4.21) |
| 5 | FOSL2 (3.11) | NFYA (5.49) | MITF (2.64) | JUND (3.37) | NR4A1 (3.36) |
| 6 | **RFX5** (2.52) | **RFX5** (4.62) | SIX5 (2.61) | NKX2-3 (3.20) | **SCRT1** (3.32) |
| 7 | **MXI1** (2.41) | IRF9 (3.97) | ELF5 (2.46) | KLF13 (3.03) | NRF1 (3.19) |
| 8 | ZBTB33 (2.29) | **MXI1** (3.32) | ZBTB1 (2.42) | PBX3 (2.69) | EPAS1 (2.94) |
| 9 | JUNB (2.07) | IRF4 (3.21) | FOXC2 (2.27) | TWIST1 (2.69) | NFYB (2.81) |
| 10 | RELA (2.05) | ARID2 (2.88) | TBX4 (2.23) | FOSB (2.52) | **CEBPZ** (2.77) |

TFs found in the top 10 in more than one network are highlighted in bold. Z-statistics for all TFs in each network are available in Supplementary Data 1).



FIGURE 3
Top 10 significant pathways enriched on the main TFs involved in RA for each specific cell type. Pathways were compiled from GO (46), KEGG (47) and Reactome (48). These pathways were ranked based on adjusted *p*-values.

associated with osteoclast differentiation (52) and T helper (Th) cell differentiation (53). Our analysis also identified less commonly documented pathways in the context of RA, such as those associated with RUNX1, found in monocytes and fibroblasts, and HOX, predominantly in fibroblasts and T cells.

To get a more complete overview of the ~400 significant pathways that were identified across the three pathway databases [GO (46), KEGG (47), Reactome (48)], we split each term by words and counted the accumulated word count found in each cell type (Supplementary Figure S9B). We identified, among others, T cells, alpha-beta, myeloids, and B cells, as words that appear consistently across cell types, specifically T cells had 24 occurrences, "$\alpha_{/\beta}$ signaling" 9 occurrences, myeloid cells 8 occurrences and B cells 7 occurrences.

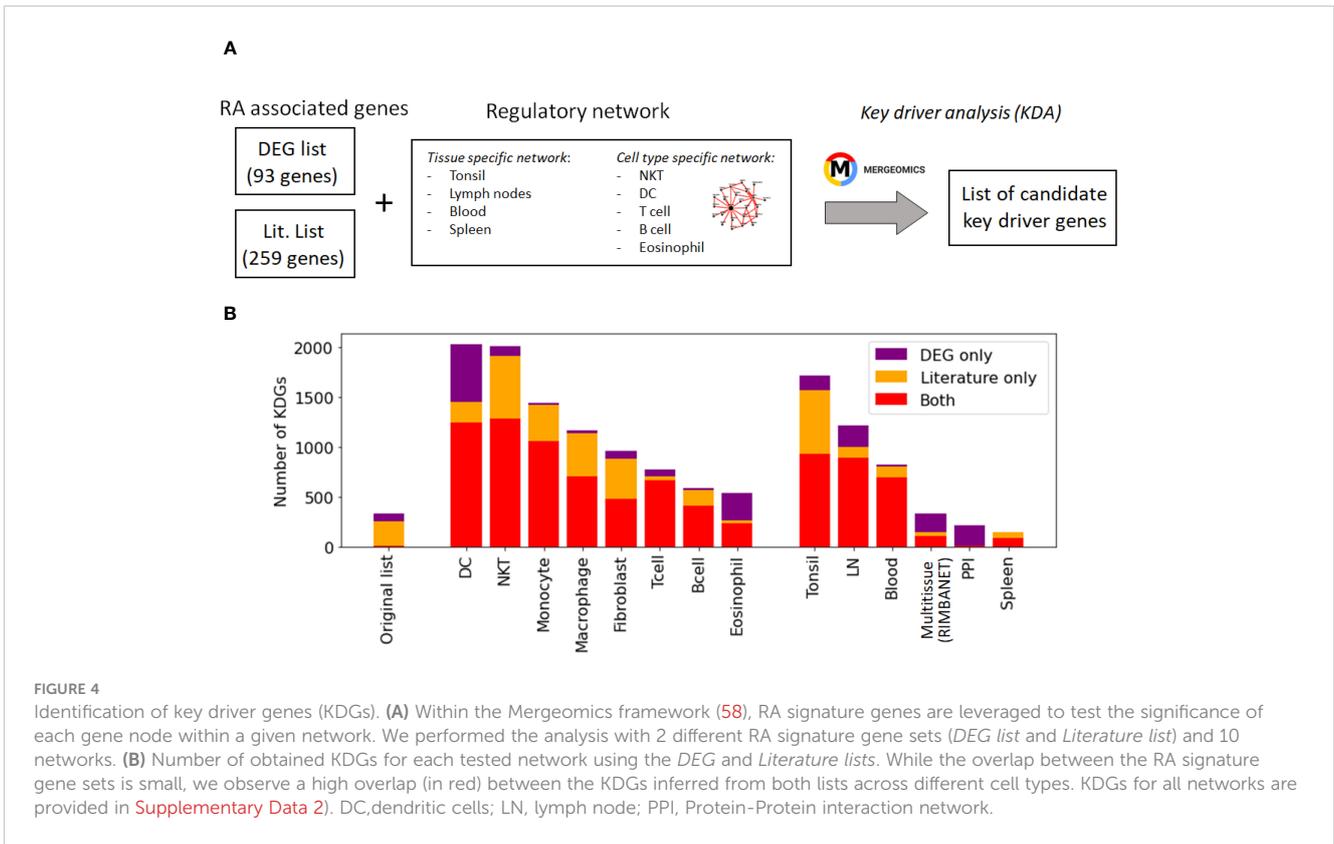## 2.3 A subset of RA key driver genes are consistently identified across cell types and tissues

The analysis of the sample-specific networks identified a list of candidate master regulator TFs in RA, and generated detailed statistics about the regulatory function of these TFs and their TGs within each cell type Supplementary Data 1) However, different network inference methods exhibit considerable variability in their inferred networks (54), typically due to the varying algorithmic assumptions and limited sample sizes. In our study, sample sizes were small in all considered cases, ranging from 15 to 40 samples for the cell type-specific networks. Hence, relying on the predictions of a single computational method might lack the robustness required to identify promising therapeutic targets. To increase our confidence in the identified RA regulators, we augmented our study by incorporating a selection of pre-existing literature-derived networks, which also included edge weights as a metric for assessing the confidence of the interactions between nodes. These include (i) RIMBANET (55), a probabilistic causal network reconstruction approach that integrates multiple data types, including metabolite concentration, RNA expression, DNA variation, DNA–protein binding, protein–metabolite interaction, and protein–protein interaction data; (ii) StringDB (43), a database of known and predicted protein–protein interactions from numerous sources, including experimental data, computational prediction methods, and public text collections; (iii) GIANT (56), a collection of networks that accurately capture tissue-specific and cell type-specific functional interactions. As RA is an autoimmune disease, we selected networks computed from immune-related tissues (including lymph nodes, spleen, tonsils, and blood). Additionally, when available, we extracted networks associated with cell types present in different proportions in RA vs control patients (Section 1.1). 14 additional networks were collected for our analysis, as detailed in the Supplementary Table S2.

While these networks recapitulate general immune knowledge derived from various data types, they are not specific to synovial tissues. Therefore, they are unable to discern RA-specific relationships between TFs and TGs as effectively as the PANDA framework does. We hence designed a different approach based on the *key driver analysis* (KDA) (57), a computational pipeline to uncover major disease-associated regulators or causative hubs in a biological network (Methods Section 2.8). Briefly, genes exhibiting more connections to RA-associated genes than expected by random chance were considered potential drivers (Figure 4A). Using a list of RA-associated signatures as a starting point, we identified potential *key drivers genes* (KDGs) linked to these signatures via network edges. To mitigate potential network size bias in the identification of KDGs, we only considered the top 1 million edges in each network. Note that a fully connected network of ~40k genes contains more than 1 billion edges, and hence, the selected edges roughly represent the top 0.1% network interactions.

KDA analysis requires the definition of RA-associated signatures, i.e. lists of genes associated with the disease. A common practice is to create gene signatures from DEGs (23, 25). However, these signatures are likely to be biased by heterogeneity in cellular composition and might include bystander genes that are not directly linked to the pathogenesis of the disease. Hence, we constructed two independent lists for KDA analysis. The first list exploited prior meta-studies and datasets (12, 13, 59–61) from which a list of DEGs was extracted (Supplementary Section B.1). The second signature combined known RA-associated genes from the literature, including GWAS (10, 62, 63), knowledge-based datasets (64–66), and known drug targets (67, 68) (Methods Section 2.7). To summarize, we performed KDA on 14 different networks (Supplementary Table S1), with two RA-associated gene signatures, which we refer to as the *DEG list* (93 genes) and the *gene literature list* (259 genes). Interestingly, the overlap between these databases was moderate (~2000 genes in total after combining all databases). Additional information is provided in Table 3; Supplementary Section B.2.

While the overlap between the *literature list* and the *DEG list* was quite small, i.e. $overlap = |A \cap B|/min(|A|,|B|) = 9/93 \, genes \sim 10\%$, they converge to a similar set of driver genes after KDA in all analyzed networks (average overlap of 90%, Table 3 & red bars in Figure 4B). This suggests that there are common drivers behind the set of DEGs and the set of known RA-associated genes reported in the literature. The number of identified KDGs, even when derived from the same lists of RA-associated genes, varied significantly across the different cell type-specific networks used in the analysis. The highest number of KDGs was found in DC and NKT, and tonsils & lymph nodes, among the various tissue types (all identified KDGs in each network are available in Supplementary Data 2). This suggests that crucial RA regulation occurs in these cell types as well as within these tissues. We also found that several genes were consistently identified as key drivers across most networks. More precisely, more than 500 genes were found in more than half of the tested networks (Supplementary Section C and Supplementary Figure S6), and the top 20 genes were found in 75% of the networks (Table 4). Among them, there are several that were already included in our *DEG* or *Literature* lists (HLA and IL2 variants, CCL5, PSMB8, CTSH), but also some that were not (PTPN6, SRGN, GBP1, LCP2, GLIPR1, CTSS, CTSH, CASP1, CD44). Importantly, the majority of genes in our top 20 list had been previously characterized in the context of RA (Table 4), thus providing further validation for our discovery approach.

**FIGURE 4**
Identification of key driver genes (KDGs). **(A)** Within the Mergeomics framework (58), RA signature genes are leveraged to test the significance of each gene node within a given network. We performed the analysis with 2 different RA signature gene sets (*DEG list* and *Literature list*) and 10 networks. **(B)** Number of obtained KDGs for each tested network using the *DEG* and *Literature lists*. While the overlap between the RA signature gene sets is small, we observe a high overlap (in red) between the KDGs inferred from both lists across different cell types. KDGs for all networks are provided in Supplementary Data 2). DC, dendritic cells; LN, lymph node; PPI, Protein-Protein interaction network.

## 2.4 Combining the KDA and PANDA analysis

For each cell type, we compiled a list of key regulators by retaining TFs that met the following criteria: (i) Their regulatory scores ($t_{reg}$), computed with the PANDA network specific to each cell type, were at least one standard deviation above the mean score of all TFs (Z-statistics > 1); and (ii) they were identified as KDGs by both the *DEG* and *Literature lists* in at least one of the 14 networks. With this approach, we select TFs that have been identified by multiple independent methods, thereby reducing the likelihood of false positives without excessively eliminating potential candidates.

We obtained between 10 and 18 TFs for the different cell types. The full list is available in Table 5. Interestingly, while the majority of key regulators were specific to individual cell types, we found that several TFs, such as RFX5, RELA, FOS, HIVEP1, IRF9, MITF,

ETV7, FOSL1, FOSB, KLF2, and ELF4, were identified as key regulators in two or more cell types.

To evaluate the agreement between the two methods we used, namely KDA and PANDA, we examined whether TFs identified as key drivers in one of the 14 networks exhibited higher PANDA scores ($t_{reg}$) than other TFs. Interestingly, we found that, on average, the TFs identified by KDA in any of the tested networks had a significantly higher regulatory score in the PANDA networks than the genes not identified by KDA ($p = 1 \times 10^{-5}$, Wilcoxon Signed-Rank Test). Figure 5 illustrates the positive relationship between the KDA and PANDA scores, where KDG TFs identified with KDA typically exhibit higher PANDA scores in at least one cell type.

## 2.5 Comparing the TF-TF co-regulation network across cell types

While we have identified a list of key TF regulators, it remained uncertain whether these regulators collectively controlled the same genes and phenotypes, or whether they independently regulate distinct targets. Clarifying the potential co-regulatory role of these TFs might open the door to combined therapeutic strategies targeting multiple TFs simultaneously. To investigate key RA driver co-regulation within each cell type, we computed the Pearson correlation between the differential edge weight $t_{edge}$ of the common TGs between two TFs.

To maintain consistency across all cell types, we selected the top 30 key driver TFs ranked by their Z-statistics per tissues type, and computed pairwise correlations between all key driver TF-TF pairs in

**TABLE 3  Genes associated with RA and those identified with key driver gene (KDG) analyses.**

| A. RA associated genes | | | B. Obtained KDGs (averg. over networks) | | |
|---|---|---|---|---|---|
| Lit. list | DEG list | Overlap | Lit. KDG | DEG KDG | Overlap |
| 259 | 93 | 10% | 741 | 827 | 90% |

(A) Number of genes utilized for KDA analysis in two different lists. The intersection is calculated as $|A \cap B|/min(|A|,|B|)$. (B) Mean number of genes designated as KDG following KDA in various networks from the two lists.

TABLE 4 The top 20 KDGs identified using the KDA with both the *DEG* and *Literature lists* are presented below, and references are provided for genes with documented literature in the context of RA.

| Gene | KDA† (Lit-DEG) | Reference | Gene | KDA† (Lit-DEG) | Reference |
|------|----------------|-----------|------|----------------|-----------|
| PTPN6 | 12 - 12 | (69) | CASP1 | 11 - 11 | (70) |
| HLA-E | 12 - 11 | (71) | **HLA-B** | 12 - 10 | (71) |
| HLA-F | 12 - 11 | (71) | **HLA-C** | 12 - 10 | (71) |
| GBP1 | 12 - 11 | (72) | CD44 | 12 - 10 | (73) |
| LCP2 | 12 - 11 | (74) | IFI30 | 12 - 10 | None |
| GLIPR1 | 11 - 11 | None | TNFAIP8 | 12 - 10 | (75) |
| **HLA-A** | 11 - 11 | (71) | CCL5 | 12 - 10 | (76) |
| CTSS | 11 - 11 | (77) | **PSMB8** | 12 - 10 | (78) |
| SRGN | 11 - 11 | None | TAP1 | 12 - 10 | (79) |
| CTSH | 11 - 11 | None | ICAM1 | 12 - 10 | (80) |

Genes also identified in GWAS are highlighted in bold. The complete list of the top 100 KDGs is available in Supplementary Section C. †The left (right) number corresponds to the number of networks the TF was identified as a KDG with the literature (DEG) list (adjusted $p_{val} < 0.05$).

these subsets. Then, in each cell type, we employed a hierarchical clustering algorithm with a correlation threshold of 0.5 to cluster these 30 TFs based on their pairwise co-regulation patterns (Figure 6A).

Significantly, monocytes exhibited a notably strong overall correlation of their TF drivers, with an average pairwise correlation coefficient of 0.77. This suggests a coordinated regulation of the observed transcriptomic distinctions between RA and OA FLS. Conversely, FLS displayed an average pairwise correlation coefficient of 0.46, and their regulators were divided into 7 clusters with minimal correlation among them. This suggests the co-existence of multiple independent FLS regulation clusters targeting different genes contributing to RA pathogenesis. T cell and B cell co-regulation clusters had an average correlation coefficient of 0.69 and 0.56, respectively. These findings are illustrated in Figure 6B, where distinct cluster populations across

cell types can be observed. To quantify these disparities, we computed various diversity indices, such as dominance and Shannon entropy (81, 82) (Methods Section 2.12), and observed substantial differences across cell types in all metrics. This divergence is also evident in Figures 6C, D, where the generated TF-TF networks exhibit distinct co-regulated clusters in the case of FLS, in contrast to monocytes, which are predominantly regulated by a single co-regulatory cluster.

# 3 Discussion

RA is a common autoimmune and inflammatory disease that affects nearly 1% of the population (3). Despite significant advances in treatments targeting different aspects of the immune response over the last two decades, achieving sustained disease remission remains uncommon (83). A small percentage of patients may achieve complete disease control on one type of therapy, yet predicting treatment responses remains challenging. This variability in treatment efficacy could be partly due to the diverse genetic factors influencing the disease, with certain genes playing more prominent roles in some patients. Additionally, variations in the disease itself and differences in individual patients' synovial tissue cell composition and activation states may contribute to this variability. A third and important missing component is the role of central genes in regulating transcriptional networks in different cells, along with their interaction and co-regulation. In this study, we combined recently published RNA-Seq data with innovative bioinformatics and analytical techniques. We identified previously uncharacterized transcriptional networks, along with their essential driver genes and TFs in synovial tissues and synovial cells from RA patients.

Extensive new RNA-seq databases and studies have significantly deepened our understanding of disease processes in RA. Nevertheless, conventional bioinformatics workflows for data-

TABLE 5 Key transcription factors (TFs) implicated in the regulation of RA identified in our analyses (Z-statistics > 1), ordered from highest to lowest Z-statistics for the different cell types.

| Cell type | (# of TFs) | List of TF drivers (Ranked from highest to lowest Z-statistics) |
|-----------|------------|------------------------------------------------------------------|
| Syn. Tissue | (10) | FOSL2, **RFX5**, **RELA**, JUNB, **FOS**, MYBL1, NFKB2, ZNF274, **HIVEP1**, HIVEP2 |
| Syn. Monocyte | (11) | IRF8, **RFX5**, **IRF9**, IRF2, CREB5, IRF3, XBP1, STAT2, EOMES, BCL11A, SPI1 |
| Syn. Fibroblast | (18) | **MITF**, CBFB, IKZF1, HLX, BACH1, **FOS**, **ETV7**, **RFX5**, HIF1A, TGIF1, **ELF4**, RUNX1, NFATC1, IRF7, CREM, **FOSL1**, FLI1, ELF1 |
| Syn. B cell | (19) | STAT5B, JUND, KLF13, **FOSB**, **FOSL1**, **KLF2**, PLAGL1, EGR3, CEBPD, TCF7, RELB, HBP1, RUNX3, BATF3, RORA, STAT6, EGR2, BHLHE40, CIC |
| Syn. T cell | (15) | **ELF4**, NR4A1, ETV6, **MITF**, ETS2, **ETV7**, MSC, **FOSB**, **RELA**, **RFX5**, **KLF2**, REL, ATF4, **HIVEP1**, VENTX, **IRF9** |

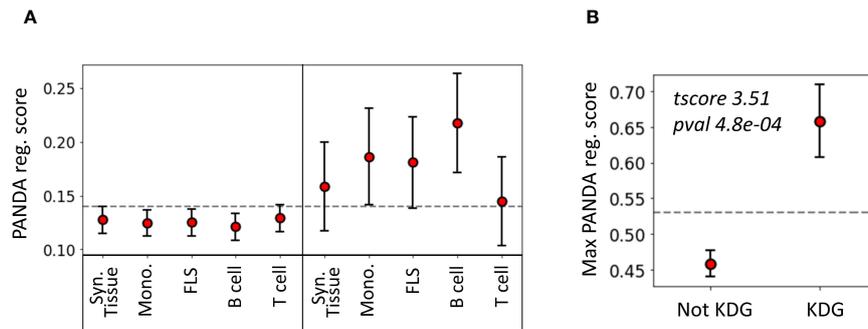TFs identified in more than one cell type are highlighted in bold.

**FIGURE 5**

Relationship between key driver analysis (KDA) and PANDA network-based TF analysis. **(A)** Average PANDA score of TFs identified as key driver genes and not identified by the KDA analysis for each tissue type. **(B)** Max PANDA score (defined as the maximum PANDA score across all cell types) for both key driver TFs and non-key drivers. The gray line represents the expected score if all TFs were randomly scored, and the error bars correspond to the 95% confidence intervals, defined as $std\sqrt{N}$.
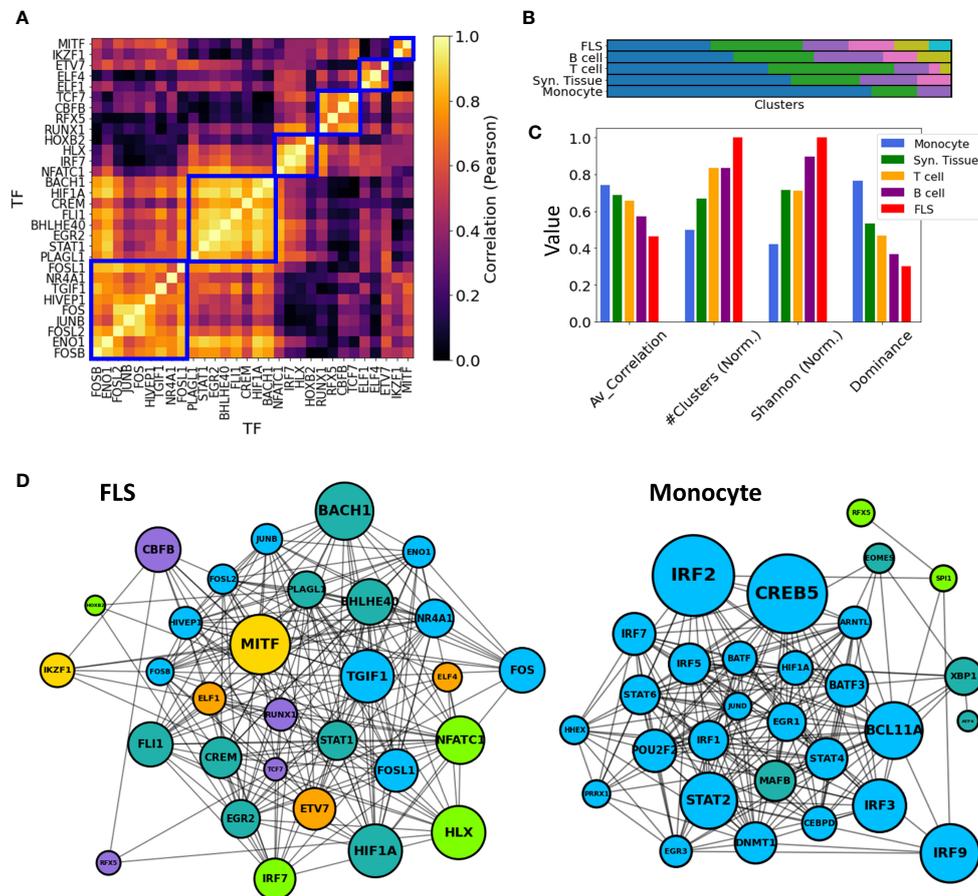


**FIGURE 6**

TF-TF RA co-regulation network in each cell type. **(A)** Pairwise TF-TF co-regulation heatmap in FLS, quantified in terms of the Pearson correlation between the differential edge weight $t_{edge}$ to their common target genes (Method 2.11). A hierarchical clustering approach is used to group them into clusters (depicted with a blue square). **(B)** The resulting clusters are visualized as a bar plot, with populations from each cluster depicted in distinct colors. **(C)** The networks are characterized by the average correlation of their edges, their number of clusters, and other diversity metrics (Shanon Entropy, dominance). These metrics are plotted next to each other for each cell type. For visual clarity, some metrics were normalized by their maximum values. **(D)** Networks show the main TFs involved in FLS and monocyte RA regulation. Edges indicate correlations exceeding the median co-regulatory scores (Methods Section 2.11). Node sizes are proportional to the node degree times TF's regulatory score, ($t_{reg}$). Node colors indicate the different co-regulatory clusters. Networks and pairwise correlation matrices are provided for all cell types in Supplementary Figure S10.

driven discovery, including GWAS, DEG analyses, and cohort-averaged GRNs, are not without challenges. For example, many GWAS loci are situated outside protein-coding regions, which complicates their functional interpretation (84). Additionally, DEGs can identify many genes that are not causally associated with the disease (85), and GRNs inferred from heterogeneous samples fail to discriminate cell type-specific regulatory mechanisms. Gene regulatory mechanisms vary significantly across cell types in both health and disease. Elucidating cell type-specific pathogenic mechanisms associated with major diseases can significantly facilitate the development of novel therapeutics, with the potential to target specific networks, pathways and cell types, with reduced risk for side effects. However, the identification of cell type-specific regulatory processes remains a challenge (84). In this context, we have developed a novel computational pipeline to identify key drivers underlying RA pathogenesis in synovial tissues. A key aspect of our analysis pipeline is the inference of sample-specific GRNs, as opposed to the more commonly used cohort-specific GRNs. This approach offers significant advantages: it reveals the regulatory mechanisms associated with individual samples and facilitates the use of statistical techniques to compare network properties across samples and phenotypic groups (86). Our approach also enabled us to rank TFs based on their contribution to phenotypic disease differences.

In the context of RA, our analysis revealed that conventional DEG analyses of synovial tissues were heavily confounded by the heterogeneous cellular composition across tissues. Indeed, we discovered that 60% of the variability in gene expression could be attributed to varying cell type proportions rather than actual differences in tissue gene regulation. Interestingly, biopsies from early and established RA had similar gene expression signatures and cellularity, suggesting that similar cell types and processes regulate disease all through the different stages of progression and, therefore, therapies can be effective throughout the disease course. Among the overrepresented cell types in RA tissues compared with control samples were DC, CD4+ memory T cells and B cells. Conversely, NKT cells emerged as the most statistically significant underrepresented cell type.

The observed reduced numbers of NKT cells in synovial tissues of RA patients suggested either impaired differentiation or impaired tissue migration or chemotaxis. Numbers of NKT cells were previously reported to be decreased in the peripheral blood of RA patients (87). NKT cells typically accumulate in the liver and move into tissues in response to chemotactic factors such as CCL5, CXCL16 and others, which are known to be expressed in the RA synovial tissues (88). NKT cells express an invariant TCR that recognizes glycolipids presented by CD1d (89). Synthetic versions of these glycolipids have been used to successfully treat arthritis in rodent models (90, 91). Additionally, recently discovered and naturally-occurring glycolipids produced by Bacteroides fragilis were shown to induce the differentiation and activation of NKT cells (92, 93). Interestingly, Bacteroides fragilis is a bacterial species commonly depleted in the intestinal microbiome of RA patients, raising the possibility of a connection between the intestinal microbiome and the reduced numbers of NKT cells in synovial tissues and blood of RA patients (94). Functionally, NKT cells can

produce IL-10 to suppress immune responses, and also can inhibit autoreactive B cells (95), which are expanded in RA synovial tissues and have central in disease. Therefore, the reduced NKT numbers could favor an expansion of the autoimmune and inflammatory response in the synovial tissues (96).

Another interesting cell population that emerged from our analysis were eosinophils, which were almost absent in RA tissues compared to OA and healthy controls. Interestingly, eosinophils with a regulatory phenotype were recently reported in the synovial tissues of RA patients in remission (97), suggesting that their presence may help control disease. Likewise, eosinophil activation can suppress inflammation in arthritis in rodent models (98).

To elucidate the potential role of the RA-associated candidate genes identified in our analyses, we utilized previously published cell type-specific gene regulatory networks (56). Namely, we identified genes associated with published RA signatures and used them in the key driver analysis (KDA) (57) framework. To increase the robustness of the method, we ran the analysis using two independent sets of RA-associated signatures with low overlap. The first gene set was compiled from online databases, including GWAS, knowledge-based and drug targets databases. The second gene set was developed with a DEG meta-analysis. Interestingly, numerous major regulators were consistently identified across diverse cell types and tissues. Several of these genes have been previously documented in the literature, highlighting the robustness of our methodology.

Our analyses identified new TF implicated in the regulation of RA synovial tissue gene expression, and more precisely implicating TF in cell specific gene regulatory networks (GRNs). As expected, regulatory interactions showed significant variability across cell types. Both the gene expression profiles and the regulatory edge weights of the cell type-specific GRNs showed minimal correlation. Certain cell types, such as FLS and B cells, were governed by multiple independent co-regulatory clusters, while the TF drivers in monocytes collectively controlled the regulatory distinctions between RA and OA. For example, CEBPZ was implicated in monocyte networks, SCRT1 in global synovial tissue and T cell networks, and RFX5 in global synovial tissue and monocyte cell GRNs. Most of the TF were specific to a cell type with IRF4 and IRF9 for monocyte, RORC and HOXA1 for FLS, JUN and STAT5B for B cells and ELF4 and HAND1 for T cell networks. Global synovial tissue and cell specific key drivers such as ELF4, FOSL1, FOSL2, HIVEP1, IRF9, KLF2, MITF, and RFX5 and were identified, several for the first time in RA. We also identified a major TF-TF co-regulation in the synovial tissue and synovial cells from RA patients, highlighting the complexities involved in cell regulation. It is conceivable that such networks and their dominant role in disease pathogenesis vary from patient to patient, which might help explain highly variable patient response to different treatments. But our analyses point to the likely relevant central target driving each network and may help point to new target for treatment.

Several of the KDG and TF have not been previously implicated in RA pathogenesis and their discovery opens new possibilities for studies and drug targeting. For example, SCRT1 (scratch family transcriptional repressor 1), is a recently discovered TF that has been implicated in pancreatic islet cell proliferation (99) and cancer

proliferation and metastasis (100). To our knowledge this is the first time that SCRT1 is implicated in the regulation of an inflammatory and autoimmune disease. RFX5 (regulatory factor X5) was only recently implicated in the regulation of synovial macrophage metabolism and survival (101). Our findings suggest that this TF has a major role not only on monocyte regulatory networks but also in the synovial tissues in general. We also detected homeobox genes (NKX2-1, NK2 homeobox 1, and HOXA1) among the top KDG in RA FLS, including an overrepresentation of HOX genes pathway genes. MITF (melanocyte inducing transcription factor) is another new FLS and T-cell KDG identified in this study. MITF was previously implicated in osteoclast differentiation and function (102) and also recently shown to mediate T-cell maturation (103). However, this is the first time that MITF is implicated in FLS and T-cell transcriptomic networks.

In conclusion, we used a robust and innovative combination of computational strategies to identify KDG, TF and GRNs in RA synovial tissues and synovial cells. While many have already been validated in previous publications, experimental validation of the newly identified KDGs and TFs would further strengthen the robustness and reliability of our findings. Incorporating experimental validation, such as functional assays or gene knockout studies, could provide more conclusive evidence of the roles of these genes in RA. These discoveries open new possibilities for experimental validation and discovery in cells or *in vivo* studies (104). For example, we identified BACH1 among the strongest TF regulators in RA FLS and we recently validated its role in RA (104).

Our KDG and TF discoveries generate new clues to understanding RA pathogenesis, and potential new targets towards developing different types of cells-specific treatment, such as targeting the FLS to maximize disease control in a patient with partial response to an anti-TNF and JAK inhibitor therapy. They may also facility the characterization of individual cell subset predominance in a patient and guide individual's therapy. Additionally, our findings should help understand the role that cellularity and the multiple pathways involved in cell regulation and co-regulations have in disease, and potentially in patient response to treatment. Unfortunately, RNA-Seq profiles of several other important cell types that significantly contribute to the heterogeneity of synovial tissues, such as NKT cells, DCs, eosinophils, or pericytes, were unavailable, and thus their GRNs could not be constructed. As such, integrating data from additional cell types could provide a more comprehensive understanding of the regulatory networks involved in RA pathogenesis, and further identify candidate genes for further analyses and consideration for therapeutic development (16).

# 4 Methods

## 4.1 Gene expression data and normalization

We used two public datasets. The first one, a bulk RNA-Seq study of synovial biopsies (GSE89408) (26, 105), comprises gene expression profiles spanning over 25k genes across 28 healthy samples, 152 RA, and 22 OA patients. The second dataset is a bulk RNA-Seq study of synovial tissues including 18 RA patients with RA 13 OA patients used as controls from the Accelerating Medicines Partnership (AMP) Phase I project (15). The first dataset is relatively bigger than the second and other publicly accessible synovial tissue studies. Moreover, it has the advantage of including both healthy and OA control groups. Also, focusing on a single big dataset, rather than combining several smaller ones, removes the batch effect bias.

All data underwent scaling normalization (106) to remove potential biases of other experimental artifacts across samples. The underlying assumption is that any sample-specific bias, such as variations in capture or amplification efficiency, uniformly scales the expected mean count for each gene. As the size factor for each sample represents the estimated relative bias in that sample, dividing its counts by its size factor should mitigate this bias.

## 4.2 Estimation of cellular compositions in synovial tissues

The cell compositions in synovial tissues were estimated with xCell (38), a machine learning framework trained using the profiles of 64 immune and stroma cell datasets. xCell takes the gene expression count as input and generates enrichment scores. Briefly, the xCell score measures the enrichment of genes specific to each cell type and is further adjusts for correlations among closely related cell types. The resulting enrichment scores are normalized to unity to enable consistent comparisons across samples.

As an alternative method, we used CIBERSORT (40), which deconvolute directly the cellular composition of the tissues from a *signature matrix* comprised of barcode genes that are enriched in each cell-type of interest (Supplementary Figure S2). To run CIBERSORT, We used the web-tool CIBERSORx (https://cibersortx.stanford.edu/runcibersortx.php) with a pre-loaded signature matrix comprising 22 immune cells.

## 4.3 Correction for cellular composition in synovial tissues

Our analysis revealed that cellular composition variation in synovial tissues accounted for a significant portion (61%) of gene expression variability. To differentiate gene expression variability arising from actual molecular state changes in cells from those due to compositional shifts, we adjusted the gene expression profiles for these covariates. To prevent over-correction, we corrected only the 18 cell types that exhibited significant differences in both RA *vs* normal and RA *vs* OA comparisons [Student's *t*-test with $p < 0.05$ after Benjamini-Hochberg correction (31)].

For each gene $k$ in sample $s$, we performed a linear regression analysis using the proportion of cell types $ci$ on that sample $s$ as covariates, i.e. $x_{ci,s}$. Mathematically, this is expressed as Equation 1:

$$Y_{ks}^{new} = Y_{ks} - \left[ \beta_{0,k} + \sum_{cells}^{ci} \beta_{ci,k} \ x_{ci,s} \right], \qquad (1)$$

where the $\beta$ s are regression coefficients derived using a least-square fit. We then used the residuals from this regression, $(Y_{ks})$, as the actual gene expression value in our analyses (107). Namely, we utilized these adjusted gene expressions for differential gene expression (DEG) analysis and to assemble the RA synovial network.

## 4.4 Differentially expressed genes

After correcting the gene expression data for cell composition, we defined differentially expressed genes (DEGs) as genes with a p-value below 0.05 in a t-test comparing the RA and control groups, after applying the Benjamini & Hochberg method (31) to control the False Discovery Rate (FDR) at 0.05. This approach ensures that, on average, only 5% of the identified DEGs are false positives, offering a robust balance against multiple testing errors.

Because the DEG analysis relies on an error-prone cell composition correction of the synovial tissues, we combined our DEGs with several meta-analyses (12, 13, 60, 61) from synovial tissues to increase the DEG analysis robustness. Genes that were identified in at least two of the lists above (either our DEGs or from one of the meta-studies) were kept as the *gene DEG list* (93 genes).

## 4.5 Pathway enrichment analysis

Pathway enrichment analysis was conducted using the Python library GSEApy (108). The pathways were sourced from GO (46), KEGG (47), and Reactome (48), and ranked by adjusted p-value. Given that all TFs are DNA-binding proteins, we removed any terms associated with RNA and DNA transcription, as these processes are ubiquitous and therefore not likely to be specific to RA.

## 4.6 Correlation of gene expression across cell types

For a given gene $g \in \mathcal{G}$, we performed a Student t-test to compare the difference in expression values between RA and the control group within a specific cell type $C$. From this test, we obtained its t-statistic, denoted as $t_{\text{expr}(g,C)}$. Next, we calculate the correlation of this score across cell type pairs $(C1, C2)$, as follows (Equation 2):

$$\text{Correlation}(C1, C2)$$
$$= Pearson \left( \left\{ t_{\text{expr}}(g, C1) \right\}_{g \in \mathcal{G}}, \left\{ t_{\text{expr}}(g, C2) \right\}_{g \in \mathcal{G}} \right) \qquad (2)$$

## 4.7 Genes associated with the susceptibility to RA

We performed an extensive literature review to aggregate known genes associated with RA from different contexts. Recent

GWAS data, highlighting genetic risk factors for RA, were collected from publicly available studies (10, 62) and the GWASdb SNP-Disease Associations database (63). Various genes associated with RA were obtained from publicly-available databases DISEASES (65), DisGeNet (64) and the Comparative Toxicogenomics Database (CTD) (66). Briefly, the CTD score, ranging from 0 to 100, measures the deviation of a gene's connectivity in the CTD chemical-RA network from that of a random network. (see http://ctdbase.org/help/diseaseGeneDetailHelp.jsp for additional details). Among the ~25k genes in the database, only 175 (less than 1%) had a CTD score higher than 50. We collected drug targets either already on the market or undergoing clinical trial (67) as well as from the DrugBank database (68) (https://go.drugbank.com/categories/DBCAT003604). Genes that were identified in at least two of the lists or databases were kept as the *gene literature list* (259 genes).

## 4.8 Networks and key driver analysis

The networks of different tissues & cell types were downloaded from the GIANT database (56) https://hb.flatironinstitute.org/download. In addition, the precomputed networks *Bayesian_Multitissue* and *PPI* were used directly on the Mergeomics (109) web service (http://mergeomics.research.idre.ucla.edu/samplefiles.php. To remove potential biases associated with network sizes, we only considered the top 1.5 million edges in terms of regulatory scores for each network.

Each of these networks was used to run a key driver analysis (KDA) associated with a list of RA-associated genes. We performed KDA with the Mergeomics R library (57) with a search depth and edge weight set to 1 and 0.5, respectively. In brief, each node (gene) in the network was tested independently. Mergeomics computed the number of edges connecting the node to any gene listed as RA-associated. A node was designated as a key driver if its linkage count exceeded the average number of links to the RA list by more than one standard deviation. In practice, Mergeomics adjusts this number accounting for the regulatory weight associated with each of these links.

## 4.9 Gene regulatory networks in synovial tissues and cell types

We inferred GRNs with PANDA (44). PANDA combines gene expression profiles of synovial tissues (and cell types) with prior knowledge about TF binding motifs (a list of target genes for each TF) and TF-TF interactions (86, 110). TF-TF interactions and TF motifs were inferred from the StringDB (43) and CIS-BP database (42), respectively. They were downloaded directly from the GRAND database (111) (https://grand.networkmedicine.org/).

PANDA employs message passing to merge a prior network (derived from mapping TF motifs onto the genome) with protein-protein interaction and gene expression datasets, iteratively refining edge weights in the networks. Applied to our data, PANDA produced directed networks of TFs to their target genes (TGs), comprising 644 TFs and 18992 genes, resulting in 12230848 edges.

Here, each edge between a TF and its TG is associated with a weight, which represents the probability of a regulatory interaction between the TF and the TG. The weight values, after undergoing a Z-transformation, typically range between -4 and 4. These indicate the number of standard deviations below (for negative Z-scores) or above (for positive Z-scores) the average weight of the network.

Then, we used LIONESS (45) to estimate an individual gene regulatory network for each individual sample in our RNA-Seq data (see Supplementary Table S2 for the sample count per cell type). LIONESS estimates sample-specific networks by sequentially leaving each sample out, calculating a network (with PANDA) with and without that sample, and using linear interpolation to estimate the network for the left-out sample. All networks were inferred with the python library netZooPy (https://github.com/netZoo/netZooPy).

## 4.10 Analysis of TFs RA regulatory activity in gene regulatory networks

We leveraged this collection of networks to test whether the weights of these regulatory edges differ significantly between RA and control samples, and to identify the TFs that may potentially be driving these regulatory differences. A Student's $t$-test was used to estimate: (i) the differential gene expression between RA and the control group, denoted as $t_{expr}$; and (ii) the differential weight of the regulatory edges between RA and the control group, denoted as $t_{edge}$. We define RA differentially expressed genes (DEGs) as the ones having a $|t_{expr}| > 1$.

Note that the $t$-score represents the difference between the mean values of the two groups being compared, divided by the standard error of the difference. A positive (respectively negative) score indicates situations where the mean of the RA group is larger (respectively smaller) than the mean of the control group. The larger the absolute value of the $t$-score, the more statistically significant the difference is relative to the variability of the data. Hence, we quantified the regulatory importance of TFs as the absolute values of the differential weights of the regulatory edges, $|t_{edge}|$ (Equation 3), averaged over RA DEGs. The hypothesis behind this choice is that TFs whose edge weights for DEGs are also differentiated between RA and control should be the main drivers of these observed regulatory differences. To prevent evaluating TFs on genes they typically do not target, we only considered the RA-associated TGs listed in the prior knowledge about TF regulon [CIS-BP database (42)]. Defining $\mathcal{G}$ as the network's gene set, the above definition can be formalized as follows (Equation 3):

$$
\begin{aligned}
\{\text{RA DEGs}\} &= \left\{ |t_{expr}(gene)| > 1, \ gene \ \in \ \mathcal{G} \right\}, \\
\{\text{TF targets}\} &= \{\text{is\_regulon}\ (TF, gene) = \text{true}, \ gene \ \in \ \mathcal{G}\}, \\
\mathcal{T} &= \{\text{RA DEGs}\} \cap \{\text{TF targets}\}, \\
TF_{score} &= \frac{1}{|\mathcal{T}|} \sum_{gene \ \in \ \mathcal{T}} |t_{edge}(TF, gene)| .
\end{aligned}
$$
(3)

We expect that TFs with the highest scores are the most likely to contribute to RA regulation.

## 4.11 TF-TF co-regulation network

We quantified the co-regulation between TFs by evaluating the correlation of their common TGs' differential edge weights. Let G be the set of all genes in the network, then Equation 4:

$$
\begin{aligned}
&\text{Co} - \text{regulation}(TF_i, TF_j) \\
&= Correlation\ \left( \{ t_{edge}(TF_i, gene), t_{edge}(TF_j, gene), gene \in \mathcal{G} \} \right)
\end{aligned}
$$
(4)

## 4.12 Clustering and diversity analysis

First, we computed for each cell type a TF-TF distance matrix, defined as one minus the absolute value of the pairwise correlation matrix. Then, in each TF-TF co-regulation network, we clustered TFs into co-regulation groups with a hierarchical agglomerative clustering (HAC) (112). The clustering criterion was defined as the Ward's minimum variance method (113), with a distance threshold of 0.5. Ward's method minimizes the total within-cluster variance, which means at every step, the algorithm finds the pair of clusters that leads to a minimum increase in total within-cluster variance after merging, until the intra-cluster distance is above the threshold.

Then, we characterized these clusters with diversity metrics such as richness (the number of clusters), Shannon entropy and dominance (82). Dominance is defined as the fractional abundance of the most abundant cluster, while the Shannon entropy ($H$) provides a measure of the overall diversity within the system by considering the frequencies of all clusters, each weighted by the logarithm of its frequency. Denoting $S$ the number of clusters, and $p_i$ the fraction of population of cluster $i$, $H$ is defined as Equation 5 (114).

$$
H = \sum_{i=1}^{S} -p_i \log (p_i) .
$$
(5)

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding authors. All the gene lists obtained in this study, along with the data and the code to reproduce all figures presented in this article are made available publicly on Github at https://github.com/AI-SysBio/RA-drug-discovery.

## Author contributions

AP: Data curation, Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review & editing. TL: Writing – original draft, Writing – review & editing. PG: Writing – original draft, Writing – review & editing. MR: Writing – original draft, Writing – review & editing.

## Funding

## Conflict of interest

Authors AP and MR were employed by the company IBM Research Europe.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2024.1428773/full#supplementary-material

## References

1. Smolen JS, Aletaha D, McInnes IB. Rheumatoid arthritis. *Lancet*. (2016) 388:2023–38. doi: 10.1016/S0140-6736(16)30173-8

2. Smolen JS, Aletaha D, Barton A, Burmester GR, Emery P, Firestein GS, et al. Rheumatoid arthritis. *Nat Rev Dis Primers*. (2018) 4. doi: 10.1038/nrdp.2018.1

3. Alamanos Y, Drosos AA. Epidemiology of adult rheumatoid arthritis. *Autoimmun Rev*. (2005) 4:130-6. doi: 10.1016/j.autrev.2004.09.002

4. Gabriel SE. The epidemiology of rheumatoid arthritis. *Rheumatic Dis Clinics North America*. (2001) 27:269–81. doi: 10.1016/S0889-857X(05)70201-5

5. Smolen JS, Aletaha D. Rheumatoid arthritis therapy reappraisal: strategies, opportunities and challenges. *Nat Rev Rheumatol*. (2015) 11:276. doi: 10.1038/nrrheum.2015.8

6. Nam JL, Ramiro S, Gaujoux-Viala C, Takase K, Leon-Garcia M, Emery P, et al. Efficacy of biological disease-modifying antirheumatic drugs: a systematic literature review informing the 2013 update of the EULAR recommendations for the management of rheumatoid arthritis. *Ann Rheumatic Dis*. (2014) 73:516–28. doi: 10.1136/annrheumdis-2013-204577

7. Aterido A, Cañete JD, Tornero J, Blanco F, Fernández-Gutierrez B, Pérez C, et al. A combined transcriptomic and genomic analysis identifies a gene signature associated with the response to anti-TNF therapy in rheumatoid arthritis. *Front Immunol*. (2019) 10:1459. doi: 10.3389/fimmu.2019.01459

8. Jung SM, Park K-S, Kim K-J. Deep phenotyping of synovial molecular signatures by integrative systems analysis in rheumatoid arthritis. *Rheumatology*. (2020). doi: 10.1093/rheumatology/keaa751

9. Okada Y, Eyre S, Suzuki A, Kochi Y, Yamamoto K. Genetics of rheumatoid arthritis: 2018 status. *Ann Rheumatic Dis*. (2019) 78:446–53. doi: 10.1136/annrheumdis-2018-213678

10. Ha E, Bae S-C, Kim K. Large-scale meta-analysis across East Asian and European populations updated genetic architecture and variant-driven biology of rheumatoid arthritis, identifying 11 novel susceptibility loci. *Ann rheumatic Dis*. (2021) 80:558–65. doi: 10.1136/annrheumdis-2020-219065

11. Viatte S, Plant D, Raychaudhuri S. Genetics and epigenetics of rheumatoid arthritis. *Nat Rev Rheumatol*. (2013) 9:141–53. doi: 10.1038/nrrheum.2012.237

12. Afroz S, Giddaluru J, Vishwakarma S, Naz S, Khan AA, Khan N. A comprehensive gene expression meta-analysis identifies novel immune signatures in rheumatoid arthritis patients. *Front Immunol*. (2017) 8:74. doi: 10.3389/fimmu.2017.00074

13. Rychkov D, Neely J, Oskotsky T, Yu S, Perlmutter N, Nititham J, et al. Cross-tissue transcriptomic analysis leveraging machine learning approaches identifies new biomarkers for rheumatoid arthritis. *Front Immunol*. (2021) 12:2104. doi: 10.3389/fimmu.2021.638066

14. Wang W, Wang L, Gulko PS, Zhu J. Computational deconvolution of synovial tissue cellular composition: presence of adipocytes in synovial tissue decreased during arthritis pathogenesis and progression. *Physiol Genomics*. (2019) 51:241–53. doi: 10.1152/physiolgenomics.00009.2019

15. Zhang F, Wei K, Slowikowski K, Fonseka CY, Rao DA, Kelly S, et al. Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. *Nat Immunol*. (2019) 20:928–42. doi: 10.1038/s41590-019-0378-1

16. Zhang F, Jonsson AH, Nathan A, Millard N, Curtis M, Xiao Q, et al. Deconstruction of rheumatoid arthritis synovium defines inflammatory subtypes. *Nature*. (2023), 1–9. doi: 10.1038/s41586-023-06708-y

17. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet*. (2015) 16:133–45. doi: 10.1038/nrg3833

18. Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali T. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods*. (2020) 17:147–54. doi: 10.1038/s41592-019-0690-6

19. Badia-i-Mompel P, Wessels L, Müller-Dott S, Trimbour R, Flores Ramirez RO, Argelaguet R, et al. Gene regulatory network inference in the era of single-cell multi-omics. *Nat Rev Genet*. (2023), 1–16. doi: 10.1038/s41576-023-00618-5

20. Keyl P, Bischoff P, Dernbach G, Bockmayr M, Fritz R, Horst D, et al. Single-cell gene regulatory network prediction by explainable AI. *Nucleic Acids Res*. (2023) 51:e20–0. doi: 10.1093/nar/gkac1212

21. Zerrouk N, Miagoux Q, Dispot A, Elati M, Niarakis A. Identification of putative master regulators in rheumatoid arthritis synovial fibroblasts using gene expression data and network inference. *Sci Rep*. (2020) 10:1–13. doi: 10.1038/s41598-020-73147-4

22. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The human transcription factors. *Cell*. (2018) 172:650–65. doi: 10.1016/j.cell.2018.01.029

23. Jung SM, Park K-S, Kim K-J. Deep phenotyping of synovial molecular signatures by integrative systems analysis in rheumatoid arthritis. *Rheumatology*. (2021) 60:3420–31:7. doi: 10.1093/rheumatology/keaa751

24. You S, Yoo S.-A, Choi S, Kim J.-Y, Park S.-J, Ji JD, et al. Identification of key regulators for the migration and invasion of rheumatoid synoviocytes through a systems approach. *Proc Natl Acad Sci*. (2014) 111:550–5. doi: 10.1073/pnas.1311239111

25. Kim K-J, Kim M, Adamopoulos IE, Tagkopoulos I. Compendium of synovial signatures identifies pathologic characteristics for predicting treatment response in rheumatoid arthritis patients. *Clin Immunol*. (2019) 202:1–10. doi: 10.1016/j.clim.2019.03.002

26. Guo Y, Walsh AM, Fearon U, Smith MD, Wechalekar MD, Yin X, et al. CD40L-dependent pathway is active at various stages of rheumatoid arthritis disease progression. *J Immunol*. (2017) 198:4490–501. doi: 10.4049/jimmunol.1601988

27. Ingegnoli F, Coletto LA, Scotti I, Compagnoni R, Randelli PS, Caporali R. The crucial questions on synovial biopsy: when, why, who, what, where, and how? *Front Med*. (2021), 1232. doi: 10.3389/fmed.2021.705382

28. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*. (2018). doi: 10.21105/joss.00861

29. Lutz MB, Schuler G. Immature, semi-mature and fully mature dendritic cells: which signals induce tolerance or immunity? *Trends Immunol*. (2002) 23:445–449. doi: 10.1016/S1471-4906(02)02281-0

30. Patente TA, Pinho MP, Oliveira AA, Evangelista GC, Bergami-Santos PC, Barbuto JA. Human dendritic cells: their heterogeneity and clinical application potential in cancer immunotherapy. *Front Immunol*. (2019) 9:3176. doi: 10.3389/fimmu.2018.03176

31. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat society: Ser B (Methodological)*. (2017) 57:289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

32. Zheng Y, He J-Q. Common differentially expressed genes and pathways correlating both coronary artery disease and atrial fibrillation. *EXCLI J*. (2021) 20:126. doi: 10.17179/excli2020-3262

33. Li L, Wang G, Li N, Yu H, Si J, Wang J. Identification of key genes and pathways associated with obesity in children. *Exp Ther Med*. (2017) 14:1065–73. doi: 10.3892/etm.2017.4597

34. Lu Z, Meng L, Sun Z, Shi X, Shao W, Zheng Y, et al. Differentially expressed genes and enriched signaling pathways in the adipose tissue of obese people. *Front Genet*. (2021) 12:620740. doi: 10.3389/fgene.2021.620740

35. Pujar M, Vastrad B, Kavatagimath S, Vastrad C, Kotturshetti S. Identification of candidate biomarkers and pathways associated with type 1 diabetes mellitus using bioinformatics analysis. *Sci Rep*. (2022) 12:9157. doi: 10.1038/s41598-022-13291-1

36. Che X, Zhao X, Xu H, Liu X, Zhao S, Ma H. Differently expressed genes (DEGs) relevant to type 2 diabetes mellitus identification and pathway analysis via integrated bioinformatics analysis. *Med Sci Monitor: Int Med J Exp Clin Res*. (2019) 25:9237. doi: 10.12659/MSM.918407

37. Cao H, Rao X, Jia J, Yan T, Li D. Exploring the pathogenesis of diabetic kidney disease by microarray data analysis. *Front Pharmacol*. (2022) 13:932205. doi: 10.3389/fphar.2022.932205

38. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol*. (2017) 18:1–14. doi: 10.1186/s13059-017-1349-1

39. Marzaioli V, Canavan M, Floudas A, Flynn K, Mullan R, Veale DJ, et al. CD209/CD14+ dendritic cells characterization in rheumatoid and psoriatic arthritis patients: activation, synovial infiltration, and therapeutic targeting. *Front Immunol*. (2022) 12:5792. doi: 10.3389/fimmu.2021.722349

40. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*. (2015) 12:453–7. doi: 10.1038/nmeth.3337

41. Collin M, Bigley V. Human dendritic cell subsets: an update. *Immunology*. (2018) 154:3–20. doi: 10.1111/imm.12888

42. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*. (2014) 158:1431–43. doi: 10.1016/j.cell.2014.08.009

43. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*. (2015) 43:D447–52. doi: 10.1093/nar/gku1003

44. Glass K, Huttenhower C, Quackenbush J, Yuan G.-C. Passing messages between biological networks to refine predicted interactions. *PloS One*. (2013) 8:e64832. doi: 10.1371/journal.pone.0064832

45. Kuijjer ML, Tung MG, Yuan G, Quackenbush J, Glass K. Estimating sample-specific regulatory networks. *Iscience*. (2019) 14:226–40. doi: 10.1016/j.isci.2019.03.021

46. Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, Ebert D, et al. The gene ontology knowledgebase in 2023. *Genetics*. (2023) 224:iyad031. doi: 10.1093/genetics/iyad031

47. Kanehisa M. The KEGG database. *'In silico'simulation Biol processes: Novartis Foundation Symposium*. (2002) 247:91–103. doi: 10.1002/0470857897.ch8

48. Croft D, O'kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res*. (2010) 39:D691–7. doi: 10.1093/nar/gkq1018

49. Friday SC, Fox DA. Phospholipase D enzymes facilitate IL-17-and TNFα-induced expression of proinflammatory genes in rheumatoid arthritis synovial fibroblasts (RASF). *Immunol Lett*. (2016) 174:9–18. doi: 10.1016/j.imlet.2016.04.001

50. Duffau P, Menn-Josephy H, Cuda CM, Dominguez S, Aprahamian TR, Watkins AA, et al. Interferon regulatory factor 5 promotes inflammatory arthritis. *Arthritis Rheumatol (Hoboken NJ)*. (2015) 67:3146. doi: 10.1002/art.39321

51. Ge L, Wang T, Shi D, Geng Y, Fan H, Zhang R, et al. ATF6α contributes to rheumatoid arthritis by inducing inflammatory cytokine production and apoptosis resistance. *Front Immunol*. (2022) 13:965708. doi: 10.3389/fimmu.2022.965708

52. Sato K, Takayanagi H. Osteoclasts, rheumatoid arthritis, and osteoimmunology. *Curr Opin Rheumatol*. (2006) 18:419–26. doi: 10.1097/01.bor.0000231912.24740.a5

53. Chen J, Li J, Gao H, Wang C, Luo J, Lv Z, et al. Comprehensive evaluation of different T-helper cell subsets differentiation and function in rheumatoid arthritis. *BioMed Res Int*. (2012) 2012. doi: 10.1155/2012/535361

54. Manica M, Bunne C, Mathis R, Cadow J, Ahsen ME, Stolovitzky GA, et al. COSIFER: a Python package for the consensus inference of molecular interaction networks. *Bioinformatics*. (2020). doi: 10.1093/bioinformatics/btaa942

55. Zhu J, Wiener MC, Zhang C, Fridman A, Minch E, Lum PY, et al. Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PloS Comput Biol*. (2007) 3:e69. doi: 10.1371/journal.pcbi.0030069

56. Wong AK, Krishnan A, Troyanskaya OG. GIANT 2.0: genome-scale integrated analysis of gene networks in tissues. *Nucleic Acids Res*. (2018) 46:W65–70. doi: 10.1093/nar/gky408

57. Shu L, Zhao Y, Kurt Z, Byars SG, Tukiainen T, Kettunen J, et al. Mergeomics: integrative network analysis of omics data. (2017).

58. Shu L, Zhao Y, Kurt Z, Byars SG, Tukianen T, Kettunen J, et al. Mergeomics: multidimensional data integration to identify pathogenic perturbations to biological systems. *BMC Genomics*. (2016) 17:1–16. doi: 10.1186/s12864-016-3198-9

59. Li Z-C, Xiao J, Peng J.-L, Chen J.-W, Ma T, Cheng G.-Q, et al. Functional annotation of rheumatoid arthritis and osteoarthritis associated genes by integrative genome-wide gene expression profiling analysis. *PloS One*. (2014) 9:e85784. doi: 10.1371/journal.pone.0085784

60. Zhang D, Li Z, Zhang R, Yang X, Zhang D, Li Q, et al. Identification of differentially expressed and methylated genes associated with rheumatoid arthritis based on network. *Autoimmunity*. (2020) 53:303–13. doi: 10.1080/08916934.2020.1786069

61. Long NP, Park S, Anh NH, Min JE, Yoon SJ, Kim HM, et al. Efficacy of integrating a novel 16-gene biomarker panel and intelligence classifiers for differential diagnosis of rheumatoid arthritis and osteoarthritis. *J Clin Med*. (2019) 8:50. doi: 10.3390/jcm8010050

62. Yamamoto K, Okada Y, Suzuki A, Kochi Y. Genetics of rheumatoid arthritis in Asia—present and future. *Nat Rev Rheumatol*. (2015) 11:375–9. doi: 10.1038/nrrheum.2015.7

63. Li MJ, Wang P, Liu X, Lim EL, Wang Z, Yeager M, et al. GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res*. (2012) 40:D1047–54. doi: 10.1093/nar/gkr1182

64. Piñero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, Baron M, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*. (2015) 2015. doi: 10.1093/database/bav028

65. Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ. DISEASES: Text mining and data integration of disease–gene associations. *Methods*. (2015) 74:83–9. doi: 10.1016/j.ymeth.2014.11.020

66. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, McMorran R, Wiegers J, et al. The comparative toxicogenomics database: update 2019. *Nucleic Acids Res*. (2019) 47:D948–54. doi: 10.1093/nar/gky868

67. Huang J, Fu X, Chen X, Li Z, Huang Y, Liang C. Promising therapeutic targets for treatment of rheumatoid arthritis. *Front Immunol*. (2021) 12:2716. doi: 10.3389/fimmu.2021.686155

68. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res*. (2018) 46:D1074–82. doi: 10.1093/nar/gkx1037

69. Kiratikanon S, Chattipakorn SC, Chattipakorn N, Kumfu S. The regulatory effects of PTPN6 on inflammatory process: Reports from mice to men. *Arch Biochem Biophysics*. (2022) 721:109189. doi: 10.1016/j.abb.2022.109189

70. Cascão R, Polido-Pereira J, Canhão H, Rodrigues AM, Navalho M, Raquel H, et al. Caspase-1 is active since the early phase of rheumatoid arthritis. *Ann Rheumatic Dis*. (2011) 70.Suppl:A2–2. doi: 10.1136/ard.2010.149096.4

71. Furukawa H, Oka S, Shimada K, Hashimoto A, Tohma S. Human leukocyte antigen polymorphisms and personalized medicine for rheumatoid arthritis. *J Hum Genet*. (2015) 60:691–6. doi: 10.1038/jhg.2015.36

72. Haque M, Singh AK, Ouseph MM, Ahmed S. Regulation of synovial inflammation and tissue destruction by guanylate binding protein 5 in synovial fibroblasts from patients with rheumatoid arthritis and rats with adjuvant-induced arthritis. *Arthritis Rheumatol*. (2021) 73:943–54. doi: 10.1002/art.41611

73. Naor D. and Shlomo Nedvetzki. "CD44 in rheumatoid arthritis. *Arthritis Res Ther*. (2003) 5:1–11. doi: 10.1186/ar746

74. Iyer VS, Boddul SV, Johnsson A.-K, Raposo B, Sharma RK, Shen Y, et al. Modulating T-cell activation with antisense oligonucleotides targeting lymphocyte cytosolic protein 2. *J Autoimmun*. (2022) 131:102857. doi: 10.1016/j.jaut.2022.102857

75. Zhang H-G, Hyde K, Page GP, Brand JP, Zhou J, Yu S, et al. Novel tumor necrosis factor α–regulated genes in rheumatoid arthritis. *Arthritis Rheumatism: Off J Am Coll Rheumatol*. (2004) 50:420–31. doi: 10.1002/art.20037

76. Agere SA, Akhtar N, Watson JM, Ahmed S. RANTES/CCL5 induces collagen degradation by activating MMP-1 and MMP-13 expression in human rheumatoid arthritis synovial fibroblasts. *Front Immunol*. (2017) 8:1341. doi: 10.3389/fimmu.2017.01341

77. Smyth P, Sasiwachirangkul J, Williams R, Scott CJ. Cathepsin S (CTSS) activity in health and disease-A treasure trove of untapped clinical potential. *Mol Aspects Med*. (2022) 88:101106. doi: 10.1016/j.mam.2022.101106

78. Verbrugge SE, Scheper RJ, Lems WF, Gruijl TD, Jansen G. Proteasome inhibitors as experimental therapeutics of autoimmune diseases. *Arthritis Res Ther*. (2015) 17:1–10. doi: 10.1186/s13075-015-0529-1

79. Zhang S-L, Chabod J, Penfornis A, Reviron D, Tiberghien P, Wendling D, et al. TAP1 and TAP2 gene polymorphism in rheumatoid arthritis in a population in eastern France. *Eur J immunogenetics*. (2002) 29:241–9. doi: 10.1046/j.1365-2370.2002.00307.x

80. Davies ME, Sharma H, Pigott R. ICAM-1 expression on chondrocytes in rheumatoid arthritis: induction by synovial cytokines. *Med Inflamm*. (1992) 1:71–4. doi: 10.1155/S0962935192000140

81. Spellerberg IF, Fedor PJ. A tribute to Claude Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon–Wiener'Index. *Global Ecol biogeography*. (2003) 12:177–9. doi: 10.1046/j.1466-822X.2003.00015.x

82. Pelissier A, Luo S, Stratigopoulou M, Guikema JE, Martinez Rodriguez M. Exploring the impact of clonal definition on B-cell diversity: implications for the analysis of immune repertoires. *Front Immunol*. (2023) 14:1123968. doi: 10.3389/fimmu.2023.1123968

83. Aletaha D, Smolen JS. Remission in rheumatoid arthritis: missing objectives by using inadequate DAS28 targets. *Nat Rev Rheumatol*. (2019) 15:633–4. doi: 10.1038/s41584-019-0279-6

84. Walsh AM, Whitaker JW, Huang CC, Cherkas Y, Lamberth SL, Brodmerkel C, et al. Integrative genomic deconvolution of rheumatoid arthritis GWAS loci into gene and cell type associations. *Genome Biol*. (2016) 17:1–16. doi: 10.1186/s13059-016-0948-6

85. Porcu E, Sadler MC, Lepik K, Auwerx C, Wood AR, Weihs A, et al. Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. *Nat Commun*. (2021) 12:1–9. doi: 10.1038/s41467-021-25805-y

86. Lopes-Ramos CM, Chen C.-Y, Kuijjer ML, Paulson JN, Sonawane AR, Fagny M, et al. Sex differences in gene expression and regulatory networks across 29 human tissues. *Cell Rep*. (2020) 31:107795. doi: 10.1016/j.celrep.2020.107795

87. Yanagihara Y, Shiozawa K, Takai M, Kyogoku M, Shiozawa S. Natural killer (NK) T cells are significantly decreased in the peripheral blood of patients with rheumatoid arthritis (RA). *Clin Exp Immunol*. (1999) 118:131–6. doi: 10.1046/j.1365-2249.1999.01018.x

88. Slauenwhite D, Johnston B. Regulation of NKT cell localization in homeostasis and infection. *Front Immunol*. (2015) 6:255. doi: 10.3389/fimmu.2015.00255

89. Exley MA, Dellabona P, Casorati G. Exploiting CD1-restricted T cells for clinical benefit. *Mol Immunol*. (2021) 132:126–31. doi: 10.1016/j.molimm.2020.12.015

90. Horikoshi M, Goto D, Segawa S, Yoshiga Y, Iwanami K, Inoue A, et al. Activation of Invariant NKT cells with glycolipid ligand $\alpha$-galactosylceramide ameliorates glucose-6-phosphate isomerase peptide-induced arthritis. *PloS One*. (2012) 7:e51215. doi: 10.1371/journal.pone.0051215

91. Yoshiga Y, Goto D, Segawa S, Horikoshi M, Hayashi T, Matsumoto I, et al. Activation of natural killer T cells by $\alpha$-carba-GalCer (RCAI-56), a novel synthetic glycolipid ligand, suppresses murine collagen-induced arthritis. *Clin Exp Immunol*. (2011) 164:236–47. doi: 10.1111/j.1365-2249.2011.04369.x

92. Cameron G, Nguyen T, Ciula M, Williams SJ, Godfrey DI. Glycolipids from the gut symbiont Bacteroides fragilis are agonists for natural killer T cells and induce their regulatory differentiation. *Chem Sci*. (2023) 14:7887–96. doi: 10.1039/D3SC02124F

93. Oh SF, Praveena T, Song H, Yoo J.-S, Jung D.-J, Erturk-Hasdemir D, et al. Host immunomodulatory lipids created by symbionts from dietary amino acids. *Nature*. (2021) 600:302–7. doi: 10.1038/s41586-021-04083-0

94. Scher JU, Sczesnak A, Longman RS, Segata N, Ubeda C, Bielski C, et al. Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. *elife*. (2013) 2:e01202. doi: 10.7554/eLife.01202

95. Yang J-Q, Wen X, Kim PJ, Singh RR. Invariant NKT cells inhibit autoreactive B cells in a contact-and CD1d-dependent manner. *J Immunol*. (2011) 186:1512–20:3. doi: 10.4049/jimmunol.1002373

96. Miellot A, Zhu R, Diem S, Boissier M.-C, Herbelin A, Bessis N. Activation of invariant NK T cells protects against experimental rheumatoid arthritis by an IL-10-dependent pathway. *Eur J Immunol*. (2005) 35:3704–13. doi: 10.1002/eji.200535235

97. Andreev D, Liu M, Kachler K, Perez ML, Kirchner P, Kölle J, et al. Regulatory eosinophils induce the resolution of experimental arthritis and appear in remission

state of human rheumatoid arthritis. *Ann Rheumatic Dis*. (2021) 80:451–68. doi: 10.1136/annrheumdis-2020-218902

98. Chen Z. Suppression of inflammatory arthritis by Th2 responses and eosinophil activation. Erlangen, Germany: Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) (2015). PhD thesis.

99. Sobel J, Guay C, Elhanani O, Rodriguez-Trejo A, Stoll L, Menoud V, et al. Scrt1, a transcriptional regulator of $\beta$-cell proliferation identified by differential chromatin accessibility during islet maturation. *Sci Rep*. (2021) 11:8800. doi: 10.1038/s41598-021-88003-2

100. Chen Y, Wang F, Li J, Wang W, Ge L, Ge L. Long non-coding RNA TCL6 induced by SCRT1 promotes proliferation and metastasis of non-small cell lung cancer through PDK1/AKT signaling. *Pathology-Research Pract*. (2023) 246:154491. doi: 10.1016/j.prp.2023.154491

101. Hu Z, Zhao TV, Huang T, Ohtsuki S, Jin K, Goronzy IN, et al. The transcription factor RFX5 coordinates antigen-presenting function and resistance to nutrient stress in synovial macrophages. *Nat Metab*. (2022) 4:759–74. doi: 10.1038/s42255-022-00585-x

102. Miyazaki M, Fujikawa Y, Takita C, Tsumura H. Tacrolimus and cyclosporine A inhibit human osteoclast formation via targeting the calcineurin-dependent NFAT pathway and an activation pathway for c-Jun or MITF in rheumatoid arthritis. *Clin Rheumatol*. (2007) 26:231–9. doi: 10.1007/s10067-006-0287-1

103. Karigane D, Haraguchi M, Toyama-Sorimachi N, Nishimura EK, Takubo K. Mitf is required for T cell maturation by regulating dendritic cell homing to the thymus. *Biochem Biophys Res Commun*. (2022) 596:29–35. doi: 10.1016/j.bbrc.2022.01.091

104. Pelissier A, Laragione T, Harris C, Martinez MR, Gulko PS. Gene network analyses identify co-regulated transcription factors and BACH1 as a key driver in rheumatoid arthritis fibroblast-like synoviocytes. *bioRxiv*. (2023), 2023–12. doi: 10.1101/2023.12.28.573506

105. Walsh AM, Wechalekar MD, Guo Y, Yin X, Weedon H, Proudman SM, et al. Triple DMARD treatment in early rheumatoid arthritis modulates synovial T cell activation and plasmablast/plasma cell differentiation pathways. *PloS One*. (2017) 12: e0183928. doi: 10.1371/journal.pone.0183928

106. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. (2010) 11:1–9. doi: 10.1186/gb-2010-11-3-r25

107. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*. (2010) 464:768–72. doi: 10.1038/nature08872

108. Fang Z, Liu X, Peltz G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics*. (2023) 39:btac757. doi: 10.1093/bioinformatics/btac757

109. Ding J, Blencowe M, Nghiem T, Ha S.-m, Chen Y.-W, Li G, et al. Mergeomics 2.0: a web server for multi-omics data integration to elucidate disease networks and predict therapeutics. *Nucleic Acids Res*. (2021). doi: 10.1093/nar/gkab405

110. Sonawane AR, Platig J, Fagny M, Chen C.-Y, Paulson JN, Lopes-Ramos CM, et al. Understanding tissue-specific gene regulation. *Cell Rep*. (2017) 21:1077–88. doi: 10.1016/j.celrep.2017.10.001

111. Guebila MB, Lopes-Ramos CM, Weighill D, Sonawane AR, Burkholz R, Shamsaei B, et al. GRAND: a database of gene regulatory network models across human conditions. *Nucleic Acids Res*. (2021). doi: 10.1101/2021.06.18.448997

112. Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview. *Wiley Interdisc Reviews: Data Min Knowledge Discovery*. (2012) 2:86–97. doi: 10.1002/widm.53

113. Szekely GJ, Rizzo ML. Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. *J classification*. (2005) 22:151–184. doi: 10.1007/s00357-005-0012-9

114. Jost L. Entropy and diversity. *Oikos*. (2006) 113:363–75. doi: 10.1111/j.2006.0030-1299.14714.x