Multi-lineage transcriptional and cell communication signatures define evolving personalized mechanisms that initiate and perpetuate rheumatoid arthritis

Cong Liu[1], E. Barton Prideaux[1], Peiyao Wu[1], David L Boyle[4], Amy Westermann[4], Katherine Nguyen[5], Vlad Tsaltskan[4], Leander Lazaro[5], Andrea Ochoa[5], Kevin D. Deane[6], Marie L. Feser[6], M. Kristen Demoruelle[6], Kristine A. Kuhn[6], V. Michael Holers[6], Fan Zhang[6,7], Laura Kay Moss[6], Megan Criley[6], Brian Hattel[6], Marguerite Siedschlag[6], Lauren Okada[8], Mark A. Gillespie[8], Palak Genge[8], Morgan Weiss[8], Veronica Hernandez[8], Julian Reading[8], Lynne Becker[8], Jane H. Buckner[9], Cate Speake[10], Thomas F. Bumol[8], Peter Skene[8], Gary S. Firestein[4#*], and Wei Wang[1,2,3#*]

[1] Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, CA 92093-0359, USA

[2] Bioinformatics and Systems Biology Graduate Program, University of California San Diego, La Jolla, CA 92093-0359, USA

[3] Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, CA 92093-0359, USA

[4] Division of Rheumatology, Autoimmunity, and Inflammation, University of California San Diego, La Jolla, CA 92093-0359, USA

[5] Altman Clinical & Translational Research Institute, University of California San Diego, La Jolla, CA 92093-0359, USA

[6] Division of Rheumatology, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA

[7] Center for Health Artificial Intelligence, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA

[8] Allen Institute for Immunology, Seattle, WA 98109, USA

[9] Center for Translational Research, Benaroya Research Institute, Seattle, WA 98101, USA

[10] Center for Interventional Immunology, Benaroya Research Institute, Seattle, WA 98101, USA

[#]Equal contribution * Corresponding authors: gfirestein@health.ucsd.edu, wei-wang@ucsd.edu

Abstract

Elevated anti-citrullinated protein antibodies (ACPA) levels in the peripheral blood are associated with an increased risk for developing rheumatoid arthritis (RA) and are hallmarks of established disease. Currently, no treatments are available that prevent progression to RA in at-risk individuals. In addition, diverse pathogenic mechanisms underlying a common clinical phenotype in RA complicate therapy because no single agent is universally effective in established disease. We propose that a unifying set of transcription factor and their downstream pathways regulate pro-inflammatory cell communication networks, and that these networks allow multiple cell types to serve as pathogenic drivers in at-risk individuals and RA. To test this hypothesis, we identified ACPA-positive at-risk individuals, patients with early and established ACPA-positive RA and matched healthy controls. Single cell chromatin accessibility and transcriptomic profiles from peripheral blood mononuclear cells were integrated to define key TFs, TF-regulated targets and pathways. A distinctive TF signature was enriched in early RA, established RA and at-risk individuals that involved key pathogenic mechanisms in RA, especially SUMOylation, RUNX2, YAP1, NOTCH3, and β-Catenin Pathways. Interestingly, this signature was identified in multiple cell types and the pattern of cell type involvement varied among the at-risk and RA participants, supporting our hypothesis. Similar patterns of individualized gene expression patterns in varying cell types were confirmed in single cell studies of RA synovium. Cell communication analysis provided biological validation that diverse lineages can deliver the same core set of pro-inflammatory mediators to receiver cells *in vivo* that subsequently orchestrate rheumatoid inflammation. Longitudinal analysis showed that the signature cell type can evolve in individual at-risk participants but their core inflammatory mediator profile was stable. Cell-type-specific signature pathways could explain the personalized pathogenesis of RA and contribute to the diversity of clinical responses to targeted therapies. Overall, this study supports a new paradigm to understand how a common clinical phenotype could arise from diverse pathogenic mechanisms and demonstrates the relevance of peripheral blood cells to synovial disease.

Introduction

Rheumatoid arthritis (RA) is a systemic immune-mediated disease marked by synovial inflammation and joint destruction[1]. Recent advances understanding its pathogenic mechanisms led to novel treatments that markedly improve clinical outcomes, including targeted therapies that block cytokines or cell types such as B cells or T cells. Interestingly, responses to these agents are highly variable; a lack of response to one drug does not preclude a response to another with a different mechanism of action. These observations led us to propose that diverse mechanisms in individuals at-risk for developing RA and with RA converge to produce a common clinical phenotype[1]. However, the divergent pathogenic pathways are poorly understood, and we lack reliable tests that predict benefit of targeted therapeutics for individual patients.

Current models suggest that seropositive RA begins with mucosal inflammation and loss of self-tolerance in individuals that carry genetic risk alleles and are exposed to risk-elevating environmental factors[2]. During a prolonged asymptomatic phase, circulating autoantibody levels increase, most notably anti-citrullinated protein antibodies (ACPAs) that are strongly associated with the future development of RA in up to 60% of at-risk individuals[3]. Several clinical trials have attempted to prevent onset of synovitis including treatment with atorvastatin, rituximab, methotrexate, hydroxychloroquine, and abatacept[4–8]. Some of these agents delay, but none prevent, RA.

These observations pose a challenging question: how do the heterogeneous mechanisms in at-risk individuals or clinical RA lead to a common phenotype? To address this, we formulated a hypothesis proposing that a unifying set of transcription factors and their downstream pathways regulate a pro-inflammatory cell communication network, and that this network enables multiple cell types to serve as pathogenic drivers in at-risk individuals or RA[1]. Thus, clinical progression and the RA phenotype would be defined by a specific transcriptional program that orchestrates pro-inflammatory signals driving synovitis. Importantly, rheumatoid inflammation can arise from diverse cell types and inflammatory mediators in this model. This study is distinct from previous reports because it focused on defining transcription factors and pathways prior to onset of RA as opposed to longstanding established RA[9–11]. This approach requires peripheral blood cell analysis because synovial tissue is not accessible in at-risk individuals.

To test this hypothesis and identify the pathways and cell types that predispose to RA, the Allen Institute for Immunology-UCSD-CU Transition to Rheumatoid Arthritis Project (ALTRA) identified at-risk individuals with elevated ACPAs. Along with early RA patients and controls, we evaluated peripheral blood mononuclear cells (PBMCs) using single cell technologies to define the transcriptome and chromatin accessibility. Our initial analysis of this population using individual omics data identified broad-based evolving immune activation of PBMCs as participants progress to clinical RA[12]. However, we could not address the individualized mechanisms that account for the diversity of responses to targeted agents.

To answer this critical question and test our hypothesis, we used a novel integrative approach to determine if there is a common set of drivers that induce aberrant immunity. We discovered a distinctive TF signature enriched in peripheral blood immune cells of at-risk individuals as well as early RA and established disease. These signature TFs regulate key pathogenic processes in RA, including SUMOylation, RUNX2, YAP1, NOTCH3, and β-Catenin Pathways. Unexpectedly, this signature was identified in multiple cell types, and the pattern of cell type varied between participants. We then tested these predictions through cell-cell communication (CCC) analysis along with biologic validation. We found that lineages displaying this RA TF signature deliver overlapping sets of inflammatory mediators to receiver cells *in vivo*. Their potential for orchestrating synovial inflammation was supported by demonstrating that the same genes are expressed *in vivo* by RA synovial cells. Interestingly, the signature cell types could vary over time in an individual even though the inflammatory mediator profile remained relatively stable as at-risk participants progressed to clinical RA. This diversity might contribute to highly variable clinical responses to targeted therapeutics in RA patients.

Results

Integrative single cell analysis reveals cell types in At-Risk/ERA and CON individuals

Peripheral blood mononuclear cells (PBMCs) were obtained from 26 ACPA positive (At-Risk) and 6 early RA (ERA) and 35 age and sex-matched controls (CON) and subjected to scATAC-seq and scRNA-seq (**Fig. 1A, Supplementary Table S1**). These data were used to assign each cell to a cell type with Latent Semantic Indexing (LSI) and Principal Component Analysis (PCA) to reduce the dimensionality of the scATAC-seq and scRNA-seq count matrices, respectively. Nearest neighbor graphs in reduced dimensions were built to identify clusters of cells. Uniform Manifold Approximation and Projection (UMAP) was then used to visualize the single cells in reduced dimension space (**Fig. 1B**). Both scRNA-seq and scATAC-seq cells were diffused evenly across the sample space, demonstrating a good integration across samples without batch effect (**Supplementary Fig. S1B, D**).

To integrate scRNA-seq and scATAC-seq, each cell in the scATAC-seq space was assigned a predicted gene expression profile from the cell in the scRNA-seq that was most similar. Cells from scRNA-seq and scATAC-seq were then clustered in the same co-embedding space for each sample (**Fig. 1C**). Each co-embedded cluster was treated as a pseudo-bulk cluster by summing gene counts from all the scRNA-seq cells and aggregating the raw scATAC-seq peaks. The annotation was defined by the cell type that occurs most frequently in the cluster. In total, 1610 pseudo-bulk clusters were retained in the final dataset, which included 703,701 scRNA-seq cells and 932,986 scATAC-seq cells, or 1,636,687 cells from 67 samples (median: 25,194 cells/sample, 767 cells/cluster) after quality control (**Supplementary Table S2**).

The cells were assigned to 22 fine-grain transcriptional cell type for each sample (**Supplementary Table S3**). Thirteen major cell types, including B memory cells, B intermediate cells, B naive cells, CD14 monocytes (CD14 Mono), CD16 monocytes (CD16 Mono), CD4 naive T cells (CD4 T Naive), central memory CD4 T cells (CD4 TCM), CD8 naive T cells (CD8 T Naive), effector memory CD8 T cells (CD8 TEM), mucosal-associated invariant T cells (MAIT cells), natural killer cells (NK), CD56 bright natural killer cells (NK_CD56bright) and regulatory T cells (Treg), accounted for > 99% of total cells and had a sufficient number of cells for subsequent analysis (**Fig. 1D**). Two subtypes of CD4 T cells (CD4 T Naive and CD4 TCM) were the most abundant cell type among all 3 cohorts of PBMC samples with >20% of total cells on average. B intermediate cells, B memory cells, CD16 Mono, NK_CD56bright, and Treg cells were relatively rare cell subsets with each comprising <2% of total cells. The cell types showed similar distribution across At-Risk, ERA and CON groups except for B intermediate, B memory, and NK_CD56bright, which were modestly higher in At-Risk compared to two other groups (Centered Log-Ratio transformation followed by Kruskal-Wallis H test, p-value = 0.1, 0.04, and 0.08 respectively) (**Fig. 1E**). We then calculated the cluster purity as the percentage of the cells of most abundant cell type for all the 1610 clusters (**Supplementary Table S4**), which was 0.72 ± 0.19 across all clusters. The cluster purity showed minor different distributions across cell types (**Supplementary Fig. S1E**). B naive, CD14 Mono, CD16 Mono, MAIT, and NK displayed the highest purity scores (mean: 0.87 ± 0.13) while purity scores for T cell subsets were more diverse across clusters and relatively lower (mean: 0.68 ± 0.18). T cell subsets were sometimes included with other T cells. For instance, CD4 TCM cluster showed some other T cells like CD4 T Naive, CD8 T Naive, and CD8 TEM.

Taiji analysis reveals distinctive TF patterns

Integrated analysis. Single cells within the same cluster were treated as one "pseudo-bulk" sample with the annotation as the cell type occurring most frequently in the cluster. The gene counts of scRNA-seq were combined and the fragments of scATAC-seq were combined to generate the RNA-seq input and ATAC-seq input for the pseudo-bulk samples respectively. We then applied the Taiji pipeline[13] to each individual cluster in each patient to evaluate the PageRank scores of TFs, which represents the importance of the TFs. Taiji has been experimentally validated in multiple biological contexts and demonstrated the robustness and reliability in revealing unappreciated roles of TFs in cell fate specification[14–16]. To characterize the global influences of all 1047 TFs across different pseudo-bulk clusters, we grouped the clusters based on the normalized PageRank across TFs. First, PCA was performed for dimension reduction of the TF score matrix with the first 500 principal components (PCs) retained for further analysis based on the "elbow" method, which explained 85% variance (**Supplementary Fig. S2A**). The first several PCs are primarily related to cell type rather than the disease state or the specific cohorts (**Supplementary Fig. S2B**). To determine the optimal number of groups and similarity metrics, Silhouette method was used to evaluate the clustering quality using five distance metrics: Euclidean distance, Manhattan distance, Kendall correlation, Pearson correlation, and Spearman correlation (**Supplementary Fig. S2C**). Pearson correlation was the most appropriate distance metric since the average Silhouette width is the highest among the five distance metrics.

Kmeans clustering. We identified 5 Kmeans groups by unsupervised clustering, denoted G1 through G5, each of which showed distinct patterns of TF activity (**Supplementary Table S4**). The row-wise comparison demonstrates that some TFs have high PageRank scores in one or several Kmeans groups and suggests high TF activity in specific clusters (**Fig. 2A; Supplementary Fig. S2D**). In total, 640 TFs were identified as Kmeans group-specific TFs by comparing their PageRank scores between a specific group and the background groups (**Supplementary Table S5; Fig. 2A**). These TFs correlated with the function of the assigned cell types. For instance, *KLF4*, which regulates monocyte differentiation[17], was G1-specific. G1 was enriched with two subsets of monocytes, including 59.5% CD14 Mono and 31.3% CD16 Mono. T-bet (encoded by *TBX21*) and *EOMES* displayed high activities in G3 where CD8 TEM and NK were the most abundant cell types with 37.9% and 40.3%, respectively. Those two genes are responsible for the cell fates of memory $CD8^+$ T cells and natural killer cells[18] (see **Fig. 2A-B; Supplementary Table S4** for lineage and group specific TFs that define each Kmeans group). Interestingly, more than half (409/640) of the TFs were G2-specific and their z scores were significantly higher in G2 compared to other groups. More than 80% (531/640) of the TFs were identified as key TFs for only one Kmeans group, suggesting the Kmeans groups had unique active TF patterns (**Supplementary Fig. S2E**).

G2 is a multi-lineage group enriched with At-Risk and ERA and reveals an RA TF signature

The 5 Kmeans groups generally showed distinct compositions of cell types and disease states (**Supplementary Table S6-7**). As noted above, 4 of the 5 Kmeans groups had their own predominant cell types that accounted for more than 70% of their total clusters. G1, G3, G4, and G5 were enriched in monocytes; CD8 TEM and NK cells; CD4 T cells; B cells, respectively. However, G2 was unique in that it was mixed and displayed a cell type distribution similar to the overall PBMC distribution and included all 13 major cell types (**Fig. 2B**). Because At-Risk and ERA were indistinguishable, they were combined for the subsequent analysis to increase statistical power.

We noted that G2 was significantly enriched in At-Risk and ERA clusters compared with CON (58% higher in At-Risk and ERA vs. CON, adjusted by the null distribution, p-value < 0.0001; Chi-squared test) and G4 was modestly enriched in CON clusters (24% higher in CON, p-value < 0.001; Chi-squared test) (**Fig. 2C**). Many interesting TFs were G2-specific, including zinc finger family members like *ZNF304*, *SP7*, *GLIS1*, *ZNF254*. For the subsequent analysis, we combined At-Risk and ERA (i.e., At-Risk/ERA) because their TF activity profiles and cell type distributions in G2 were nearly identical (p-value > 0.2; Wilcoxon rank-sum test). Moreover, the identified G2-specific TFs along with the enriched pathways for ERA and At-Risk respectively showed almost complete overlap (p-value < 10^{-5}) (**Supplementary Fig. S2F-G**).

Multiple immunity-related TFs and the downstream genes regulated by those TFs conformed to pathways implicated in the pathogenesis of RA (**Fig. 2D; Supplementary notes**). This was most prominent true for G2, where 5 relevant and significant pathways were identified, namely *SUMOylation of Intracellular Receptors*[19], *Transcriptional regulation by RUNX2*[20], *YAP1 and WWTR1-stimulated Gene Expression*[21], *NOTCH3 Intracellular Domain Regulates*

Transcription[22], and *Deactivation of the β-Catenin Transactivating Complex*[23] Reactome pathways. The TFs and the representative target genes identified by our analysis are shown in **Supplementary Table S8**. These TFs and their downstream regulated genes are referred to as the *RA TF signature*. These TFs were significantly important in the signature pathways and the representative genes were among the top regulated genes by the corresponding TFs predicted by Taiji (**Methods**).

The G2 RA TF signature is enriched in multiple cell types

Interestingly, we observed that the At-Risk/ERA TFs identified in G2 were present across all the major cell types with sufficient numbers to analyze (**Fig. 2E**), thereby establishing them as a hallmark "RA TF signature" and their downstream pathways as "signature pathways" comprised of "signature genes". We calculated the percentage of G2 clusters per cell type of total global clusters for At-Risk/ERA and CON groups (**Fig. 3A; Supplementary Fig. S2H**). Notably, CD4 T Naive, CD4 TCM, and CD8 T Naive showed the greatest enrichment in At-Risk/ERA compared to CON (31% vs 18%, p-value < 0.01; 23% vs 12%, p-value < 0.01; 65% vs 26%, p-value < 0.01, respectively for At-Risk/ERA compared with CON; Chi-squared test). Of interest, MAIT cells with the TF profile were only found in CON clusters (0% vs 43% for At-Risk/ERA and CON, p-value < 0.1; Chi-squared test). Despite the negative correlation between MAIT cell abundance and age, the comparable age of the CON group with At-Risk/ERA (**Supplementary Table S1**) suggests that age does not account for these differences and MAIT cells might be protective of conversion/progression of RA. Overall, the top RA signature TFs determined by unsupervised clustering showed significantly higher PageRank scores in G2 compared to other groups across all cell types (**Fig. 3B**).

The major cell types were enriched in this common set of At-Risk/ERA signature pathways while some individual cell types demonstrated some specific enriched pathways (**Fig. 3C**). For example, activation of HOX genes was enriched in B cells, CD4 T cells, CD8 T Naive, and monocytes. RUNX3 regulation is more highly associated with CD8 TEM, NK, CD4 T Naive, and monocytes[24]. Despite individual variations described above, the general pattern of pathways associated with pathogenesis of RA is consistent and extends across the identified cell types.

Patterns of cell types with the G2 RA TF signature vary across individuals

At-Risk and Early RA signature cell types. We then determined which cell types display the TF signature in each member of the At-Risk and ERA cohorts. Multiple combinations of cell types were identified in individual participants (**Fig. 3D**). Twenty-five out of 26 At-Risk and all 6 ERA participants (as well as all 5 established RA [see below]) had the signature in at least one cluster and in at least one of the key cell types. We suspect that the one negative At-Risk individual likely had a similar pattern in a less common cell type that was beyond the resolution of this analysis. However, the distribution of cell types was highly variable among participants. In some cases, only one cell type was identified for an individual participant, while in others there were multiple cell types. For instance, participant 9 had clusters with the signature in all the cell types except NK and Treg, while participant 27 only had CD4 TCM clusters. Some patients displayed broader distribution across multiple cell types like participant 31 while others had predominant signature cell type like participant 3.

Among the involved cell types, the signature was most enriched in T cell types including CD4 T Naive, CD8 T Naive, CD4 TCM, and CD8 TEM (**Fig. 3D**). Different cell types also displayed diverse distribution patterns across patients. CD4 Naive and CD4 TCM had much wider appearances in many patients while Treg, B cell and monocytes were only found in a few participants. Therefore, the patterns displayed by various individuals were diverse with highly variable cell types. Some CONs also displayed these signatures although the number of clusters was significantly less than At-Risk/ERA, particularly for certain T cell subsets (p-value < 0.005; Wilcoxon rank-sum test) (**Supplementary Fig. S3A**).

Established RA signature cell types. To explore whether the At-Risk/ERA signature persists in longstanding RA, a separate cohort of 5 established RA PBMC samples were obtained from patients with established RA requiring arthroplasty. We applied the same analysis pipeline to obtain the PageRank scores of TFs across pseudo-bulk clusters from established RAs. Hierarchical clustering demonstrated that a group of TFs have high PageRank scores in specific clusters (**Supplementary Fig. S3B**). All RA participants had the RA TF signature in at least one cell type and the combinations of cell types were diverse despite being treated with a variety of anti-rheumatic agents (**Fig. 3E**). Comparison of signature TFs identified from At-Risk/ERA and established RA datasets showed a significant overlap (**Fig. 3F**) as did the enriched Reactome pathways (**Fig. 3G**). All the signature pathways identified in At-Risk/ERA dataset were also observed in the established RA dataset.

Enhanced cellular communication networks in At-Risk/ERA

Quantifying cell-cell communication. After demonstrating individualized patterns of signature cluster cell types in At-Risk/ERA, we then investigated how the signature cell inflammation signals are transmitted. Cell-cell communications (CCC) were analyzed by correlating expression levels of ligands, such as cytokines in the sender cells with expression of their corresponding receptor in the receiver cells for each individual using CellChat[25]. We first aggregated CCC between the same signature cells in At-Risk/ERA and CON across all the ligand-receptor pairs and all the individuals within the group. We observed distinct CCC patterns: At-Risk/ERA participants displayed significantly more interactions within signature clusters than did CON, particularly between T cells and NK cells. Cellular communications with signature monocytes were less common and only observed in the At-Risk/ERA group (**Fig. 4A**). The difference between the total number of CCC in the two groups approached statistical significance (p-value=0.06 using Wilcoxon rank-sum test).

We next evaluated the cellular communication strength. Notably, communication between CD8 T Naive and CD4 TCM was more pronounced in At-Risk/ERA group than in CON (**Fig. 4B**). The total communication strength in At-Risk/ERA was significantly higher than CON (p-value=0.04 using Wilcoxon rank-sum test). As a representative example, participant 53 from control group and participant 9 from At-Risk/ERA group had a diverse cell type distribution in signature clusters (**Supplementary Fig. S3A; Fig. 4C**), providing an overview of almost all the cell types. It is worth noting that the number and intensity of the total CCC aggregating all the clusters from all the Kmeans groups were comparable between the At-Risk/ERA and CON groups, highlighting the importance of the signature cells in G2 differentiating the two groups (**Supplementary Fig. S3D**).

Overlapping core sets of inflammatory mediators in the cell communication network. A diverse array of inflammatory cytokines, chemokines, proteases and growth factors contribute to a core set of inflammatory mediators that have been implicated in RA pathogenesis. We curated a representative list of these mediators for subsequent analysis (**Methods; Supplementary Table S9**). As with the diversity of signature cell types across individuals, the CCC pattern transmitting the inflammatory signals also varied from individual to individual. For example, major senders and receivers differed among individual participants (**Supplementary Fig. S3C**). Some individuals such as participant 5, 26, and 27 used only one cell type as major communicator while others like participant 9, 18, and 23 relied on multiple cell types. Among those with multiple cell types, some displayed consistent distributions of signals across cell types like participant 9 and 23 while others exhibited a predominant signature cell type (e.g., CD8 TEM in participant 18).

Of the identified significant ligand-receptor pairs in each participant, twelve ligand-receptor pairs were related to this inflammatory mediator gene set. We ranked the important pathways based on the difference in total information flow within signature clusters when comparing At-Risk/ERA to CON samples. The IL16 - CD4, CD160 - TNFRSF14, TGF-β1 - (TGFBR1+TGFBR2), and BTLA - TNFRSF14 were the most prominent ligand-receptor pairs enriched in At-Risk/ERA based on the difference and absolute information flow values (**Fig. 4D**).

Classification model for At-Risk/ERA using the RA TF signature

Classification model. To characterize pathogenic genes in At-Risk/ERA participants, we developed a random forest classification model to distinguish CON and At-Risk/ERA participants with gene expression as features. Sixty-three genes were identified as candidate predictors, which were active across each At-Risk/ERA participant in signature group G2 (**Methods**). The test accuracy was monotonically increasing with more predictors, reaching a plateau of 0.93 (**Supplementary Fig. S3F**). Top predictors included *MMP23B, TGFB1, IFNL1, CCL5,* and *IL15* (**Supplementary Fig. S3G**).

Gene expression was greater for the top 30 predictors in At-Risk/ERA participants compared with CON, including *CCL4*, *IL12A*, *TNFSF14*, *IL15*, *NOTCH1*, and *CCL5* (**Fig. 5A**). *MMP23B*, which emerged as a top predictor in classification model, regulates the Kv1.3 potassium channel, which has been implicated in autoimmunity[26]. Although a common set of pathogenic genes were shared across At-Risk/ERA participants, the cell types and individual mediators that were most likely to produce the specific gene were variable (**Supplementary Fig. S3E**).

Biologic validation of computational predictions

Biologic validation of mediator predictions. To validate our predictions related to the inflammatory mediator profile, we measured serum protein expression levels of 6 genes (CCL3, CCL4, IFN-λ1, IL-15, TGF-β1, and TNFSF14; see **Methods**). Mediators predicted to be elevated based on gene expression predictions displayed increased protein expression level in the serum of At-Risk/ERA group compared to CON (**Fig. 5B**).

Biological validation of signature genes in RA synovium. We then evaluated the Accelerating Medicine Partnerships (AMP) synovial scRNA-seq data to determine if the RA gene expression signature observed in PBMCs was also reflected in inflamed synovium and, more importantly, whether a diversity of cell types was also present[11]. **Fig. 5D** shows that the top PBMC mediators were also expressed by rheumatoid synovial cells and were marked by the same the broad diversity of the cell types observed in blood cells. Like PBMCs, the number of genes expressed across all cell types displayed distinct patterns across samples (**Fig. 5E**). This heterogeneity suggests that the synovial tissue of each RA patient, like At-Risk and early RA PBMCs, have a similar molecular signature. We then determined which cell types display the regulatory signature for each patient. As with PBMCs, distribution of cell types that expressed signature genes was highly variable among RA samples (**Fig. 5F**).

Biologic validation of cell communication networks. To validate the communication network we performed a more detailed analysis of several individual sender genes. *TGFB1* displayed an overall elevated expression in At-Risk/ERA, but the sender source was individualized and could include multiple cell types. For example, *TGFB1* was highly expressed in CD8 TEM and NK cells in many participants while it was more highly expressed in B cells in participant 13 and monocytes in participant 7 (**Fig. 5C**). It also showed much denser and stronger intercellular communications (**Fig. 4D-E**). Biological validation for these predicted interactions came from the receiver cells themselves. When the transcriptome of the predicted receiver cell was evaluated, TGFß-mediated gene induction including *TGFBR1*, *TGFBR2*, and *TNFAIP8* was confirmed in the expected receiver cell type even though the sender cell types were diverse (**Fig. 4F-G**). Additionally, we explored the relationship between the G2 RA signature TFs identified above and *TGFB1* as their target gene. The signature regulators of *TGFB1* included some well-known RA-related TFs like *RORC, TFAP2A,* and *KLF1*.

We also confirmed the predicted receiver cell response for other sender signals in At-Risk/ERA individuals. For instance, IL16 - CD4 signaling pathway, which has been implicated in RA[27], showed significantly stronger signals in At-Risk/ERA group than control group (**Fig. 4D; Supplementary Fig. S4A**). Multiple cell types send signals of IL16, including B cells and monocytes that are unique senders in At-Risk/ERA and CD8 TEM and monocytes are unique receivers, responding to IL16 signals regardless of their source. CD4 T cells are most widely used as communicators across participants while B cells and NK cells only act as senders in IL16 signaling pathway (**Supplementary Fig. S4B**). We confirmed significantly elevated expression of key downstream target genes, including *CDKN1B*, *IL32*, and *TNFAIP8*, in the predicted receiver cell types (**Supplementary Fig. S4C**).

Two other pathways that are elevated in At-Risk/ERA are CD160 and BTLA signaling pathways (**Fig. 4D; Supplementary Fig. S4D, G**). While NK cells were the most common senders for CD160 signaling, B cells acted as the exclusive senders for BTLA signaling, and both pathways targeting a diverse range of receiver cell types (**Supplementary Fig. S4E, H**). Like TGFß, we confirmed significantly elevated expression of key downstream genes induced by the sender mediator in the predicted receiver cell types, including *TNFRSF14* and *TNFAIP8* (**Supplementary Fig. S4F, I**).

Biological validation of signature cell biology in an RA clinical trial. To evaluate the functional relevance of the sender cell type in personalized therapy, we analyzed data from a clinical trial in RA evaluating the T cell directed agent abatacept[28]. Only PBMC bulk RNA-seq data were available, so we deconvoluted the data to identify which cell types expressed signature genes. We then stratified patients based on their response to abatacept. As shown in **Fig. 5G**, the patients without clinical benefit to this therapy (“non-responders”) were characterized by signature gene expression in monocytes rather than T cells as predicted.

Longitudinal analysis: Evolving signature cell types and stable mediator expression

Evolving signature cell types in converters. In addition to obtaining PBMCs when participants enrolled, we also obtained additional samples at a later time point and/or at the time of transition to clinical RA. Longitudinal analysis of individuals that progressed to RA reveals temporal dynamics prior to and at the time of clinical symptoms. **Fig. 6A** shows that the signature cell types in an individual are not fixed but can evolve over time. This suggests a stochastic process, perhaps due to further environmental exposure at mucosal surfaces. Importantly, the specific cell types carrying the RA signature typically returns to a previous cell type when participants transition to RA (p = 0.0272).

Stable mediator profiles in converters. Despite signature cell type variability, the underlying inflammatory profile is more consistent over time. For example, key mediators implicated in RA pathogenesis, such as *CCL5* and *TGFB1*, were persistently expressed from baseline through conversion, regardless of which signature sender cell type was implicated (**Fig. 6B**). At the time of clinical onset, a closer look at the mediator expression patterns reveals distinct, cell-type-specific contributions to the inflammatory environment (**Fig. 6C**). For instance, NK cells are highly active producers of chemokines like *XCL1* and *XCL2* while CD8 TEM cells make the greatest contribution to *CCL5* and *TNFSF8* expression. We also observed a prominent activation signature in CD4 Naive T cells, a finding consistent with previous work linking this cellular profile to abatacept response in established RA[12].

Discussion

Our study provides evidence that individuals at risk for developing RA and as well as patients with clinical RA exhibit consistent TF signatures in peripheral blood immune cells. These signatures, which involve pathways implicated in disease pathogenesis, could contribute to disease onset and persist after transition to classifiable RA. The signature TFs regulate key genes in pathways implicated in RA pathogenesis like *SUMOylation*, *RUNX2*, *YAP1*, *NOTCH3*, and *β-Catenin* pathways[19–23]. The individual cell types that display the signatures are plastic over time, although inflammatory mediators that they produce is more stable. Furthermore, the signature cell type at conversion recapitulates one from pre-conversion time points. These data build upon our previous analysis of single omics data, which identified evolving generalized immune activation during the at-risk period[12]. By integrating chromatin accessibility and the transcriptome, we markedly increased the depth of our analysis and allowed us to identify individualized pathogenic cell types and mediators in at-risk participants, early RA and established RA. A summary of the proposed transition from at-risk to clinical RA is shown in Fig. 6D.

Our analysis and computational predictions were also biologically validated using a variety of experimental approaches. First, the inflammatory mediators that we predicted would be elevated *in vivo* were increased in serum as measured using Olink protein assay. Second, cell communication data were validated by showing that the predicted receiver cell types expressed genes *in vivo* that are regulated by their paired sender cell signal. Third, deconvolution of data from an RA clinical trial showed that a T cell targeted agent (abatacept) was not effective in individuals when the signature transcriptome was primarily present in monocytes rather than T cells. Finally, the inflammatory mediator repertoire and signature cell types that we defined in PBMCs was also identified in a separate analysis of the synovial tissue single cell transcriptome from patients with RA. Taken together, these data strongly support the notion that this RA signature is biologically relevant and reflects underlying pathogenesis.

One of our most intriguing observations is that the signatures occurred in multiple cell types, each of which could then have arthritogenic potential. All participants exhibit distinct combinations of cell types at baseline, along with unique overlapping subsets of signature pathways and pathogenic genes, suggesting a stochastic component to remodeling the epigenome. The signature TFs drive a defined set of pro-inflammatory genes that, in turn, could contribute to the onset and perpetuation of RA. Employing classification models, we defined candidate genes associated with transition to clinical synovitis in individual patients including *TGFB1*, *CCL4*, *IL15*, and *TNFSF14*, each of which were confirmed by protein expression data. Thus, common inflammatory drivers are regulated by this RA TF signature and transmitted to receiver cells that can trigger transition to clinical synovitis. Analysis of the cellular signaling network confirmed that the responding cells actively transcribe the appropriate genes and are agnostic about which cell provides the signal as long as the receiving pathogenic cell has access to them. Even though the cell types have the capacity to evolve over time, the mediator drivers are stable and suggest a final common individualized pathway over time.

Previous studies in at-risk individuals predominantly focused on the transcriptome of established RA synovium or peripheral blood[9–11,29,30]. The present study is unique in that it integrates transcriptome and chromatin accessibility data to reveal pathways that would have been missed by transcriptome-only analysis[31]. In addition, the RA TF signature was identified by clustering TF PageRank scores calculated from the integrated data from pseudo-bulk clusters in each individual participant. This method is distinct from traditional single cell analysis that typically aims to identify cell clusters unique to disease compared to control. Single cell analysis using single technologies would not enable discovery of the signature given the high variability of signature TFs and cell types in individual participants.

Many of the pathways and genes that we discovered in our integrated analysis in at-risk individuals were also observed in synovial tissue cells of patients with established RA, especially within certain T cell clusters. For instance, *CCL5*, identified as a key player in both communication pathway and a top pathogenic gene in our study, is also a top marker gene of CD8+ GZMK+ memory clusters in RA synovial tissues[9,11]. Our findings also corroborate previous research highlighting the importance of other chemokines like *CCL4*, *CCL4L2*, *CCL3*, *XCL1*, and *XCL2*. Furthermore, *TNFSF9* and *IFNG*, which emerged as top predictors in our model, are also noted in CD4 T cells isolated from established RA synovium[11]. The concordance between our pre-RA PBMC and established RA synovium data suggests that blood sampling is relevant to synovial mechanisms and that potential early biomarkers or patient stratification is feasible using PBMCs.

A primary finding in previous analyses of peripheral blood cells in at-risk individuals, identified individual cell types such as CD4+ T naïve cells or CCR2+ CD4+ T cells[12,30]. However, preponderance of a single pathogenic cell type would not explain the diversity of benefit or lack of responses to targeted T cell agents like abatacept or even anti-CD4 antibodies[32]. Although the same cell types are identified in our analysis, many other lineages were also present based on the RA TF signature. Our ability to discover other potentially pathogenic cells is likely due to the greater resolution afforded by integrating transcriptome and chromatin accessibility and discovering the most relevant TFs. This method also highlighted distinct patterns of pathogenic cell types for each participant. This improved resolution confirms our previous observation that combining both technologies markedly increases the ability to distinguish between cell populations and pathways[31]. CD4+ T cells account for many of the clusters in our analysis, but B cells, CD8+ T cells, monocytes and NK cells also exhibit the signature and produce the same pathogenic mediators as CD4+ T cells in some participants.

Signature cell type plasticity over time was surprising and probably reflects stochastic events leading to immune activation, perhaps due to repeated stimulation at mucosal surfaces by circulating cells. Despite individual variations over time, the same TF signature was present in all phases of disease, including at-risk, early RA and established RA. Interestingly, the final “conversion” cell type for an individual patient was usually present in a previous pre-transition sample. Thus, a second “hit”, such as DNA methylation[33] might push someone from “at-risk” to clinical autoimmunity. While it is tempting to use signature cell type information to target an individual pathogenic cell type, this should be tempered by the observation that it can vary over time, and many individuals have multiple signature cell types. Interestingly, the inflammatory mediator profile is more stable and might be more useful as a biomarker for stratification and developing personalized approaches to treatment.

The surprising overlap of the RA TF signature, genes, and pathways across multiple cell types suggests that there might be common mechanisms that shape the RA-associated transcriptome and epigenome. The nature of these influences is not yet known, but its consistency across the spectrum of cell types, cell location (blood and synovium) and serum protein mediator levels implies that they are shared. Environmental and mucosal stresses, especially in the airway due to its critical role in the RA, are possible influences because all circulating cell types can be exposed to irritants at these sites. For example, cigarette smoke is a known risk factor for RA and can induce stress throughout the airway. Smoking is also associated with alterations in the epigenome of peripheral blood cells[34]. We also previously described DNA methylation abnormalities in circulating B cells and memory and naive CD4 T cells in the at-risk population[35], which supports this concept. It is also possible that multiple cell types in G2 are influenced by similar inflammatory signals, but the impact could be divergent depending on where they are imprinted (e.g., gut, lung, or synovium).

Our study primarily focused on pre-RA, but we also observed similar patterns in early RA and, surprisingly, established RA even though the latter were treated with a variety of anti-rheumatic drugs. However, it remains uncertain whether the signature is specific to RA due to the absence of comparable datasets for other “at-risk” populations. It is plausible that this signature represents a general phenomenon occurring during the “at-risk” period across various immune-mediated diseases. If so, the ultimate manifestation of a particular autoimmune disease might be determined by other factors, such as genetic predisposition and environmental influences. This phenomenon could provide insight into the variability in therapeutic responses observed across different immune-mediated diseases. Nevertheless, in certain diseases, this scenario seems less likely. For instance, the vast majority of psoriasis patients respond favorably to Th17-directed therapies[36], suggesting a more limited cellular repertoire driving disease pathology compared to RA. Thus, while the immune signature identified in pre-RA may have broader relevance, its specificity and cellular distribution likely vary across autoimmune diseases, warranting further investigation.

In conclusion, our study defined distinctive RA TF signatures and genes enriched in the peripheral blood mononuclear cells at-risk individuals and RA. These TFs are implicated in the known pathogenic pathways, offering new insights into the molecular events that lead to RA. Analysis of cell-cell communication shows that the signature-bearing cells deliver shared pro-inflammatory signal to receiver cells. Notably, the signatures and mediators are present in diverse cell types from different individuals, providing a potential explanation for the diverse clinical responses with targeted therapeutics. We propose that multiple cell types and their respective inflammatory mediators can be responsible for the transition to clinical arthritis, and that the receiver cells do not discriminate based on the source of the signal. These individualized signature patterns potentially open avenues for prognostic tests and personalized treatments. Overall, our findings represent a novel paradigm for understanding how a common clinical phenotype arises from diverse mechanisms. Similar processes might account for variable therapeutic responses in other immune-mediated diseases.

Materials and Methods

Clinical cohorts

Four groups of participants were recruited for this study. The demographics and baseline characteristics of the ALTRA cohorts are provided in **Supplementary Table S1**. Additional information on the criteria and cohorts is available in He et al[12]. The first cohort (At-Risk) included individuals who were at-risk for future clinical RA as indicated by serum ACPA positivity >2x the upper limit of normal[2] using the assay anti-cyclic citrullinated peptide-3 anti-CCP3, IgG ELISA (Werfen, San Diego, CA USA). The second cohort (ERA) was comprised of patients who were anti-CCP3 positive and had early RA meeting the 2010 American College of Rheumatology/European Alliance of Associations for Rheumatology (ACR/EULAR) classification criteria for RA and were diagnosed <1 year from study enrollment[37]. The At-Risk and ERA participants were identified and recruited at the University of Colorado Anschutz and UC San Diego. The third cohort (CON) was comprised of control participants without inflammatory arthritis who were recruited at the Benaroya Research Institute and the University of Colorado Anschutz. The fourth cohort was not part of ALTRA and included five individuals with established RA identified at the time of arthroplasty. These individuals were treated with a variety of agents at the time of enrollment, including tofacitinib, TNF blockers, low dose prednisone and hydroxychloroquine. The studies were approved by ethical review boards at the University of Colorado Anschutz, UC San Diego and the Benaroya Research Institute, and all participants gave informed consent. Metadata for the established RA cohort was limited because the samples were de-identified.

Genomic data acquisition

Sample preparation

Blood was drawn into BD NaHeparin vacutainer tubes (for PBMC; BD #367874) or K2-EDTA vacutainer tubes (for plasma; BD #367863). PBMC isolation and plasma processing were started within 2 hours post draw. For PBMC isolation, the samples in NaHeparin tubes for each donor were pooled into one common pool and combined with an equivalent volume of room temperature PBS (ThermoFisher #14190235). PBMCs were isolated using Leucosep tubes (Greiner Bio-One #227290) with 15 ml of Ficoll Premium (GE Healthcare #17-5442-03). After centrifugation, the PBMCs were recovered and resuspended with 15 ml cold PBS+0.2% BSA

(Sigma #A9576; “PBS+BSA”). The cells were pelleted, resuspended in 1 ml cold PBS+BSA per 15 ml whole blood processed and counted with a Cellometer Spectrum (Nexcelom) using Acridine Orange/Propidium Iodide solution. PBMCs were cryopreserved in 90% FBS (ThermoFisher #10438026) / 10% DMSO (Fisher Scientific #D12345) at a target of 5 x 10^6 cells/ml by slow freezing in a Coolcell LX (VWR #75779-720) overnight in a -80°C freezer followed by transfer to liquid nitrogen.

For genomics assays PBMCs were removed from liquid nitrogen storage and immediately thawed in a 37°C water bath. Cells were diluted dropwise into 40 mL AIM V media (Thermo Fisher Scientific #12055091) pre-warmed to 37°C. Cells were pelleted at 400 x g, resuspended in 5 mL cold AIM V media, and recounted using a Cellometer Spectrum. 30 mL cold AIM V media was added to the cells, which were re-pelleted and resuspended to appropriate concentration for the assays.

scRNA-seq

scRNA-seq was performed on PBMCs as previously described[38] *(P. C. Genge, STAR Protoc 2, 100900 (2021))*. In brief, scRNA-seq libraries were generated using a modified 10x genomics chromium 3′ single cell gene expression assay with Cell Hashing. Sample libraries were constructed across different batches, with the addition of a common control donor leukopak sample in each library as batch control. Libraries were sequenced on the Illumina Novaseq platform. Hashed 10x Genomics scRNA-seq data processing was carried out using BarWare[39] to generate sample-specific output files.

scATAC-seq

FACS neutrophil depletion

To remove dead cells, debris, and neutrophils prior to scATAC-seq, PBMC samples were sorted by fluorescence-activated cell sorting (FACS) following established protocols[38]. Cells were incubated with Fixable Viability Stain 510 (BD, 564406) for 15 minutes at room temperature and washed with AIM V medium (Gibco, 12055091) before incubating with TruStain FcX (BioLegend, 422302) for 5 minutes on ice, followed by staining with mouse anti-human CD45 FITC (BioLegend, 304038) and mouse anti-human CD15 PE (BD, 562371) antibodies for 20 minutes on ice. After washing, cells were then sorted on a BD FACSAria Fusion with a standard viable CD45+ cell gating scheme. Neutrophils were then excluded in the final sort gate. An aliquot of each post-sort population was used to collect 50,000 events to assess post-sort purity.

Sample processing

Permeabilized-cell scATAC-seq was performed as described previously[38]. A 5% w/v digitonin stock was prepared stored at −20°C. To permeabilize, 1×10^6 cells were centrifuged and resuspended in cold isotonic Permeabilization Buffer. Then they were diluted with 1 mL of isotonic Wash Buffer and centrifuged, and the supernatant was slowly removed. Cells were resuspended in chilled TD1 buffer (Illumina, 15027866) to a target concentration of 2,300-10,000 cells per µL. Cells were filtered through 35 µm Falcon Cell Strainers (Corning, 352235) before counting on a Cellometer Spectrum Cell Counter (Nexcelom) using ViaStain acridine orange/propidium iodide solution (Nexcelom, C52-0106-5).

Sequencing library preparation

scATAC-seq libraries were prepared following established protocol[38]. In brief, 15,000 cells were combined with TD1 buffer (Illumina, 15027866) and Illumina TDE1 Tn5 transposase (Illumina, 15027916) and incubated at 37°C for 60 minutes. A Chromium NextGEM Chip H (10x Genomics, 2000180) was loaded and a master mix was then added to each sample well. Chromium Single Cell ATAC Gel Beads v1.1 (10x Genomics, 2000210) were loaded into the chip, along with Partitioning Oil. The chip was loaded into a Chromium Single Cell Controller instrument (10x Genomics, 120270) for GEM generation. After the run, GEMs were collected and linear amplification was performed on a C1000 Touch thermal cycler.

GEMs were separated into a biphasic mixture with Recovery Agent (10x Genomics, 220016), and the aqueous phase was retained and removed of barcoding reagents using Dynabead MyOne SILANE and SPRIselect reagent bead clean-ups. Sequencing libraries were constructed as described in the 10x scATAC User Guide. Amplification was performed in a C1000 Touch thermal cycler. Final libraries were prepared using a dual-sided SPRIselect size-selection cleanup.

Quantification and sequencing

Final libraries were quantified using a Quant-iT PicoGreen dsDNA Assay Kit (Thermo Fisher Scientific, P7589) on a SpectraMax iD3 (Molecular Devices). Library quality and average fragment size were assessed using a Bioanalyzer (Agilent, G2939A) High Sensitivity DNA chip (Agilent, 5067-4626). Libraries were sequenced on the Illumina NovaSeq platform with the following read lengths: 51nt read 1, 8nt i7 index, 16nt i5 index, 51nt read 2.

Plasma proteomics

Plasma samples were run on the Olink Explore 1536 platform. Analytes from the inflammation, oncology, cardiometabolic, and neurology panels were measured. Samples were randomized across plates to achieve a balanced distribution of age and sex. Resulting data were first normalized to an extension control that was included in each sample well. Plates were then standardized by normalizing to inter-plate controls run in triplicate on each plate. Data were then intensity normalized across all samples. Final normalized relative protein quantities were reported as log2 normalized protein expression (NPX) values by Olink. Three protein analytes were repeated across each of the four panels and treated as distinct measurements: TNF, IL-6, and CXCL8. Data, including QC flags, were reviewed for overall quality prior to analysis. Samples were measured across multiple batches.

To facilitate comparisons between batches, plasma from 12 donors was obtained commercially (BioIVT; Bloodworks Northwest) and randomly interspersed among the above study samples. Samples measured in later batches were bridge normalized to the earliest batch. Bridge offsets were determined for each batch and each analyte separately by taking the median of the per-sample NPX differences between the later batch result and the earliest (reference) batch result for the 12 commercial samples. Offsets were then subtracted from the analyte measurements of all samples in the later batch to obtain the normalized NPX values.

Dataset integration

Paired scRNA-seq and scATAC-seq datasets from each participant were obtained from 26 At-Risk individuals with elevated anti-citrullinated protein antibody (ACPA), 6 seropositive ERA patients and 35 controls (CON). The detailed clinical information is summarized in **Supplementary Table S1**.

10x scRNA-seq data. scRNA-seq data were aligned using 10x cellranger v3.1.0 and 10x transcriptome vGRCh38-3.0.0. Hashtag Oligo sequences were processed using CITE-Seq Count v1.4.3, and cells were assigned to sample-linked hashes, split by sample for each well, and merged across wells per sample using an AIFI pipeline. Cells were labeled using Seurat v4 labeling pipeline with default parameters. The reference was customized based on the recently described CITE-seq reference of 162,000 PBMC measured with 228 antibodies[40]. QC summary plots along with statistics can be found in **Supplementary Fig. S1A** and **Supplementary Table S2**.

10x scATAC-seq data. In the scATAC-seq pipeline, we implemented CellRanger alignment, followed by a rigorous quality control process. We retained cells with unique fragments between 1000 and 100,000, fragment size between 10 and 2000, >50% of fragments in Altius, >20% of fragments in transcription starting site (TSS), >4 TSS enrichment score. This ensures that cells from the scATAC-seq pipeline are high quality, reduces the number of doublets, and are available in a variety of formats for downstream analysis (.arrow, fragments.tsv.gz, and .h5-formatted count matrices). scATAC-seq data were aligned using 10x cellranger-atac v1.1.0, using reference vGRCh38-1.1.0. After alignment, data were processed through a custom QC and counting pipeline to generate a matrix of unique fragment counts in each peak. ArchR v1.0.2 was used to generate Arrow files, doublet filtering (filterRatio=0.5), dimensionality reduction with iterative latent semantic indexing (LSI) (iterations=4), and clustering (resolution=3). QC summary plots along with statistics can be found in **Supplementary Fig. S1C** and **Supplementary Table S2**.

Integration of scRNA-seq and scATAC-seq data. scATAC-seq data were integrated with the corresponding scRNA-seq using the “addGeneIntegrationMatrix” function in ArchR with default parameters. After alignment, each cell in the scATAC-seq space was assigned a gene expression signature from the cell in the scRNA-seq that is the most similar. Cells from both scRNA-seq and scATAC-seq were clustered in the same co-embedding space.

TF regulatory networks construction based on Taiji

Single cells within the same cluster were treated as one “pseudo-bulk” sample with the annotation as the cell type occurring most frequently in the cluster. The gene counts of scRNA-seq were added up and the fragments of scATAC-seq were combined to generate the RNA-seq input and ATAC-seq input for the pseudo-bulk samples respectively. Only pseudo-bulk samples with >2000 open chromatin peaks, >20 scATAC-seq cells and >20 scRNA-seq cells were kept on account of reliability of constructed regulatory networks. Additionally, to link promoters and enhancers, the promoter-enhancer contacts predicted by Epitensor v0.9 was used. Taiji v1.1.0 with default parameters was used for the integrative analysis of RNA-seq and ATAC-seq data. The motif file was downloaded directly from the CIS-BP database containing 1078 human motifs.

Taiji pipeline overview

To characterize TF activity in each pseudo-bulk cluster, we performed an integrated multi-omics analysis using the Taiji pipeline[13,16]. Taiji integrates gene expression and epigenetic modification data to build gene regulatory networks. The algorithm first predicts putative TF binding sites in each open chromatin region that mark active promoters and enhancers using motifs documented in the CIS-BP database[41]. These TFs are then linked to their target genes predicted by EpiTensor[42]. The regulatory interactions are assembled into a genetic network. Finally, the personalized PageRank algorithm is used to assess the global influences of the TFs. In the network, the node weights are determined by the z scores of gene expression levels, allocating higher ranks to the TFs that regulate more differentially expressed genes. Each edge weight is set to be proportional to the TF's expression level, its binding site's open chromatin peak intensity, and the motif binding affinity, thus representing the regulatory strength. Using this method, Taiji has more power than other methods that identify key regulators in individual transcriptome and chromatin accessibility and has been confirmed using simulated data, literature evidence and experimental validation in numerous studies of various biological problems[13–16]. For this dataset, the median number of nodes and edges of the networks were 17,046 and 3,002,662, respectively, including 1047 (6.14%) TF nodes. On average, each TF regulates 3417 genes, and each gene is regulated by 184 TFs.

TF regulatory networks weighting scheme

As described in the original Taiji paper[13], a personalized PageRank algorithm was applied to calculate the ranking scores for TFs. We first initialized the edge weights and node weights in the network. The node weight was calculated as e^{z_i}, where z_i is the gene's relative expression level in cell type i, which is computed by applying the z score transformation to its absolute expression levels. The edge weight was determined by $e_{ij} = \sqrt{g \sum_{k=1}^{n} p_k * m_k}$, where p is the peak intensity, calculated as $\frac{1}{1+e^{-(x-5)}}$, where x is $-log_{10}(p)$, represented by the p-value of the ATAC-seq peak at the predicted TF binding site, rescaled to [0, 1] by a sigmoid function; m is the motif binding affinity, represented by the p-value of the motif binding score, rescaled to [0, 1] by a sigmoid function; g is the TF expression value; n is the number of binding sites linked to gene j. Let s be the vector containing node weights and W be the edge weight matrix. The personalized PageRank score vector v was calculated by solving a system of linear equations $v = (1 - d)s + dWv$, where d is the damping factor (default to 0.85). The above equation can be solved in an iterative fashion, i.e., setting $v_{t+1} = (1 - d)s + dWv_t$.

If the TFs in the same protein family share the same motifs, their PageRank scores are distinguished by their own expression levels because their motifs and the target genes are the same. If a motif is weak, the PageRank score of the TF is decided by whether these motifs occur in the open chromatin regions (measured by the peak intensity of the ATAC-seq data), the TF expression and its target expression levels. The relative difference between the PageRank scores of TFs also helps to uncover important TFs with weak motifs.

Unsupervised clustering analysis

To identify the groups of samples showing similar TF activity profile, we clustered the samples based on the normalized PageRank across TFs. First of all, we performed the principal component analysis (PCA) for dimension reduction of the TF score matrix. We retained the first 500 principal components (PCs) for further clustering analysis based on “elbow” method, which explained 85% variance (**Supplementary Fig. S2B**). To find the optimal number of groups and similarity metric, we performed the Silhouette analysis to evaluate the clustering quality using five distance metrics: Euclidean distance, Manhattan distance, Kendall correlation, Pearson correlation, and Spearman correlation (**Supplementary Fig. S2C**). Pearson correlation was the most appropriate distance metric since the average Silhouette width was the highest among the five distance metrics. Based on these analyses, we identified 5 Kmeans groups showing distinct dynamic patterns of TF activity.

Identification of Kmeans group-specific TFs

To identify Kmeans group-specific TFs, we divided the clusters into two groups: target group and background group. Target group included the clusters in the Kmeans group of interest and the background group comprised the remaining clusters. We then performed the normality test using Shapiro-Wilk’s method to determine whether the two groups were normally distributed and we found that the PageRank scores of most clusters (95%) didn’t follow normal or log-normal distribution. Thus, Mann-Whitney U Test was used to calculate the P-value. Double cutoffs, i.e.

P-value $\leq$ 0.01 and log2 fold change $\geq$ 0.5, were used for calling specific TFs. Results were summarized in **Supplementary Table S5**.

TF regulatee analysis

Taiji generated the regulatory network file for each cluster showing the regulatory relationship between TF and regulatees with edge weight, which represents the regulatory strength. Regulatees in **Supplementary Fig. S2I, K** are top 500 regulatees ranked by mean edge weight across G2-specific TFs. Representative regulatees in **Supplementary Table S8** were selected as the top 10 genes regulated by the signature TFs involved in each pathway ranked by the mean edge weight.

Pathway enrichment analysis

The enriched functional terms in this study were analyzed by R package clusterProfiler_4.0.5. A cutoff of P-value $\leq$ 0.05 was used to select the significantly enriched Reactome pathways.

Cell-cell communication analysis

The R package CellChat_2.1.2[25] was used to analyze the intercellular interactions within each individual. First, input scRNA-seq data matrix was normalized by TPM (transcripts per million) method and log-transformed with pseudo count of 1. The assigned cell labels were the cell types identified from co-embedding. Ligand-receptor interaction database was CellChatDB v2 excluding non-protein signaling interactions, which finally includes ~2300 validated molecular interactions in the analysis. The default parameters were used following the standard CellChat pipeline. Finally, the intercellular communication networks were obtained for each individual and aggregated together for the downstream visualization.

Identification of candidate pathogenic genes related to signature group G2

We first curated a customized list of 186 genes including all the available cytokines, chemokines, growth factors, NOTCHs, MMPs, and ADMATS with gene expression in this study. The full gene list is shown in **Supplementary Table S9**. For each gene, the maximum gene expression across clusters was taken within each Kmeans group and each individual as input. Then, we identified the universal G2-important genes with mean gene expression across all patients ranked as top 50% and coefficients of variation (CV) less than 2. In total, 63 genes were identified as candidate predictors for the following classification model.

Classification model construction

To distinguish the controls from At-Risk/ERA patients, we developed a random forest classification model. The input data was gene expression of identified important genes across patients. For each At-Risk/ERA patient, the maximum gene expression across G2 clusters was taken. For each control, the maximum gene expression across G4 clusters was considered.

The samples were split into train and test subsets at a 7:3 ratio. The R package Caret_6.0.94[43] was used for feature importance evaluation based on recursive elimination algorithm implemented in “rfe” function. Only features with positive importance was kept. Random forest model was trained multiple times with an increasing number of predictors, from the most to least important, using 10-fold cross-validation and repeated 5 times. Each trained model was then evaluated on prediction accuracy on the unseen test set. The above process was repeated 20 times with different random seeds from 1 to 20. The mean and standard deviation of the training and testing accuracy was calculated for each number of predictors.

Comparison with AMP study

To confirm the expression patterns of newly identified predictors from classification model, we checked the gene expression levels in synovial tissues samples from established RA patients in AMP study[11]. To make it more compatible with cell types in PBMC samples, we only considered 22 clusters defined in original AMP paper that are also present in PBMC populations from 82 synovial tissue samples (**Fig. 5D**). We collapsed single-cell gene expression profiles into pseudo-bulk count matrices by summing the raw UMI counts for each gene across all cells from the same sample and cluster. For each gene, we normalized counts in each pseudo-bulk sample into counts per million. We averaged the normalized counts across samples, cell types, and genes and visualized the results as heatmaps (**Supplementary Table S11; Fig. 5D-F)**.

Abatacept treatment response analysis

We obtained gene expression matrices of bulk PBMC RNA-seq from 22 RA patients before and after abatacept treatment[28]. A deconvolution analysis using CIBERSORTx[44] was performed to estimate the expression of genes in monocytes. Top 30 mediators defined in our study were imputed and reference remained the same in the original work[28].

Longitudinal samples analysis

Longitudinal samples were processed and analyzed in the same way as cross-sectional samples. We first determined the optimal number of groups K and similarity metrics for unsupervised clustering. Pearson correlation was the most appropriate distance metric and K=4 was the optimal number since the average Silhouette width was the highest. We identified four Kmeans groups showing distinct dynamic patterns of TF activity. G4 is a multi-lineage group with cell type distribution similar to the overall PBMC distribution. Comparison of signature TFs identified from longitudinal and cross-sectional datasets showed a significant overlap (75%) as did the enriched Reactome pathways (86%). All of the signature pathways identified in cross-sectional dataset were also observed in the longitudinal dataset.

Data availability

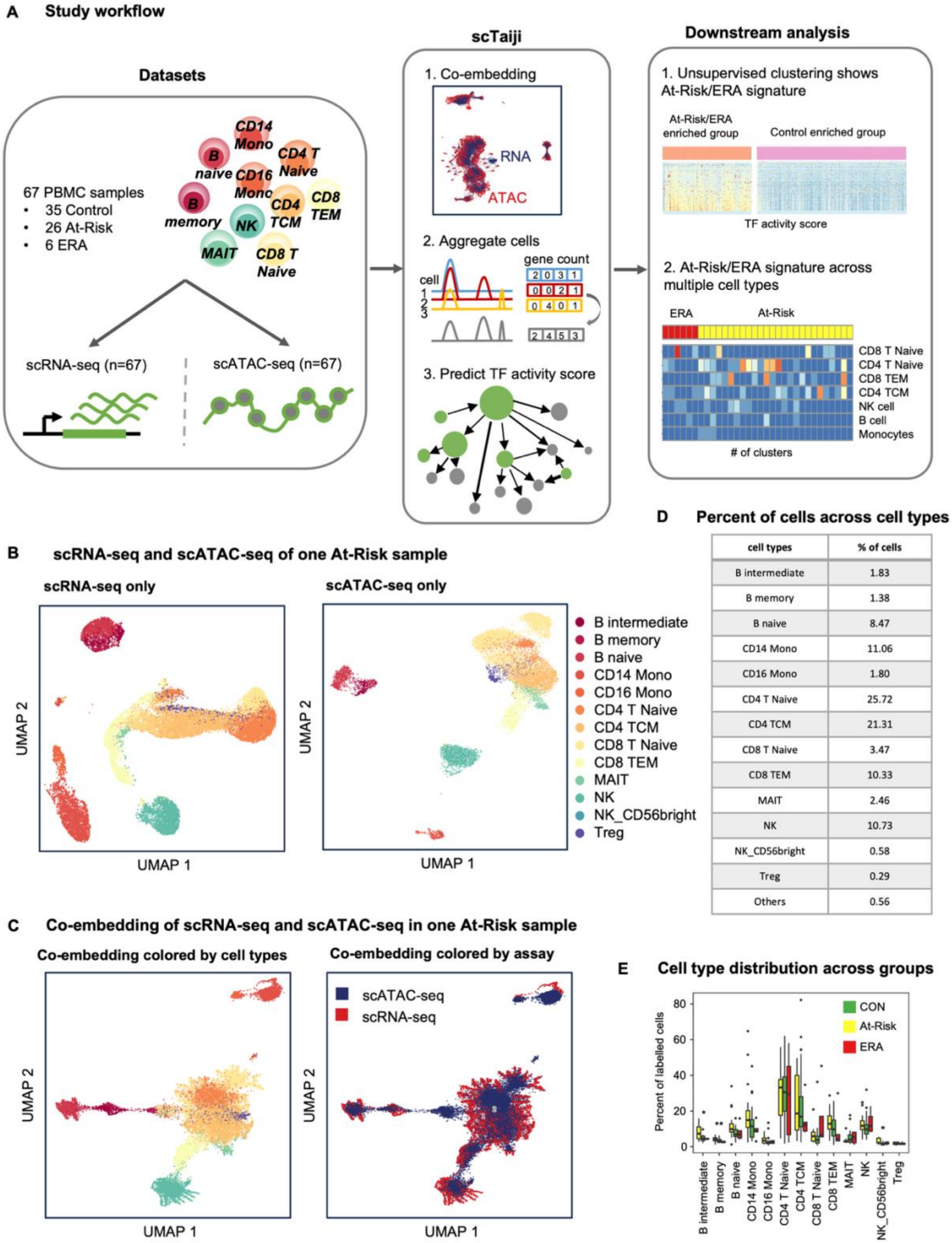
scRNA-seq and scATAC-seq data from this paper are deposited in the GEO database (GSE278746). The output of this study (TF activity heatmap, individual UMAP and cellular network plots) will be available at our Taiji-altra portal (https://wangweilab.shinyapps.io/Taiji_Altra/). All other raw data are available from the corresponding author upon request.

Code availability:

The code to reproduce the data analysis and related figures in this study can be found at https://github.com/Wang-lab-UCSD/Taiji_ALTRA

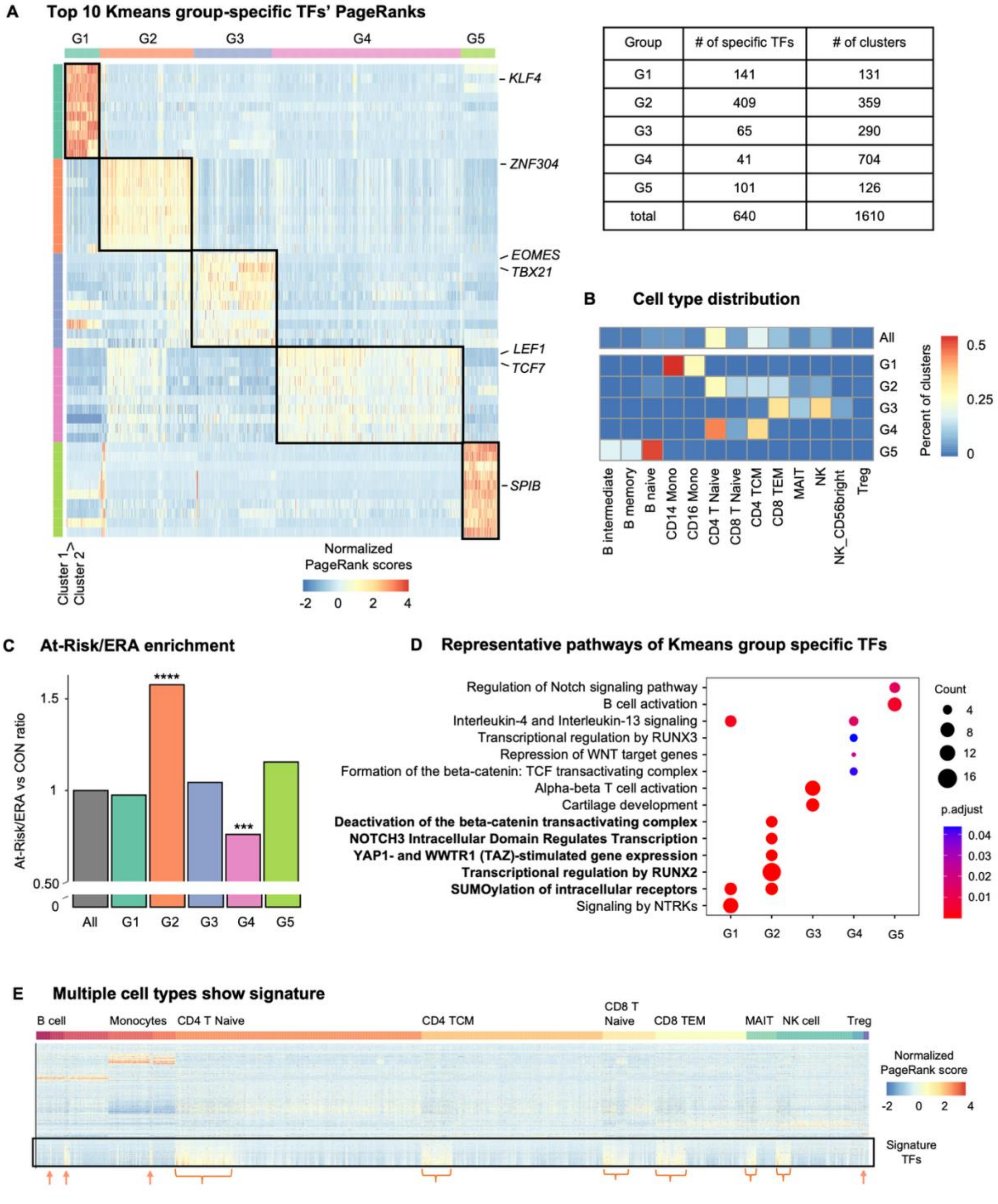
Figures

Fig.1 Study overview of multi-omics integrative analysis. (A) Study workflow. PBMC samples including 35 matched controls (CON), 26 ACPA positive (At-Risk) and 6 early RA (ERA) were utilized for scRNA-seq and scATAC-seq respectively. For each sample, matched data were co-embedded into clusters. Cells in each cluster were aggregated in terms of gene count and open chromatin regions. Then each cluster was used as input of scTaiji to construct a regulatory network and generate the PageRank scores as output. The following unsupervised clustering revealed At-Risk/ERA signatures that were shared across multiple participants and cell types. **(B)** UMAP colored by cell types in scRNA-seq cells (left) and scATAC-seq cells (right) respectively for one At-Risk sample. Clusters in both scRNA-seq and scATAC-seq were well separated by cell types. The selected sample represents the typical situation for all the 67 samples. Thirteen cell types include B memory cells, B intermediate cells, B naive cells, CD14 monocytes (CD14 Mono), CD16 monocytes (CD16 Mono), CD4 naive T cells (CD4 T Naive), central memory CD4 T cells (CD4 TCM), CD8 naive T cells (CD8 T Naive), effector memory CD8 T cells (CD8 TEM), mucosal-associated invariant T cells (MAIT cells), natural killer cells (NK), CD56 birght natural killer cells (NK_CD56bright), and regulatory T cells (Treg). **(C)** UMAP colored by cell types (left) and assays (right) in cells from both scRNA-seq and scATAC-seq for the same sample in **Fig. 1B**. The color palette of the left plot is the same as **Fig. 1B**. Blue and red represent scATAC-seq and scRNA-seq. Clusters in co-embedding space were still separated by cell types while scRNA-seq and scATAC-seq cells were well aligned. **(D)** Percent of total cells across cell types. CD4 Naive and CD4 TCM were the most abundant cell type while B memory cells, CD16 Mono, MAIT, and Treg cells were the relatively rare cell subsets. **(E)** Cell type distribution across 3 groups of PBMC samples. Yellow, red, green represent At-Risk, ERA, and CON. The color palette is maintained throughout all figures. Centered Log-Ratio (CLR) transformation before Kruskal-Wallis test, *p< 0.1, **p < 0.01. Most cell types showed similar distribution across groups except for B intermediate, B memory, and NK_CD56bright, which were modestly higher in At-Risk compared to other two groups.



cell types	% of cells
B intermediate	1.83
B memory	1.38
B naive	8.47
CD14 Mono	11.06
CD16 Mono	1.80
CD4 T Naive	25.72
CD4 TCM	21.31
CD8 T Naive	3.47
CD8 TEM	10.33
MAIT	2.46
NK	10.73
NK_CD56bright	0.58
Treg	0.29
Others	0.56

Fig.2 Unsupervised clustering identified signature TFs and pathways . (A) PageRank scores heatmap of 5 Kmeans group-specific TFs across 1613 clusters. Top 10 TFs from each Kmeans group are selected as rows and colored by their group specificity. Color palette for Kmeans groups is RColorBrewer palette Set2. The color palette is maintained throughout all figures. Clusters in columns are ordered by Kmeans group. Color of the cell indicates the normalized PageRank scores with red displaying high scores. Each Kmeans group displayed distinct dynamic patterns of TF activity. Side table is the number of the specific TFs for each Kmeans group. G2 has the largest number of specific TFs. **(B)** Cell type distribution across Kmeans groups. The separate top row represents the overall cell type distribution across all the clusters. The bottom five rows are distributions for five Kmeans groups. Color represents the percentage of clusters of each cell type with red displaying a high percentage. G2 is a multi-lineage group with distribution similar to the overall distribution. Other 4 groups had predominant cell types. **(C)** At-Risk/ERA vs CON ratio distribution across Kmeans groups. The first gray bar is the overall ratio adjusted to 1 while other bars represent 5 Kmeans groups. G2 is significantly enriched in At-Risk/ERA while G4 is enriched in CON. G1, G3, and G5 show no significant enrichment. **(D)** Representative Reactome pathways enriched in each Kmeans group-specific TFs. The horizontal axis represents Kmeans groups and the vertical axis represents pathways. Circle size represents the number of TFs in the pathway and color represents the adjusted p-values. Bold text represents signature pathways. G2 exhibits unique enrichment of several RA-related pathways e.g. SUMOylation of intracellular receptors (adjusted p-value < 1e-5), Transcriptional regulation by RUNX2 (adjusted p-value < 1e-5), etc. **(E)** Heatmap of PageRank scores of all TFs across all clusters with columns ordered by cell types and Kmeans groups. The signature TF group is marked by black box. Each cell type displayed high activity in signature TFs; Chi-squared test, ***p< 0.001, ****p < 0.0001.



Group	# of specific TFs	# of clusters
G1	141	131
G2	409	359
G3	65	290
G4	41	704
G5	101	126
total	640	1610

Fig.3 Patterns of cell types with At-Risk/ERA signature vary across individuals. (A) G2 clusters per cell type of the total clusters per cell type in CON and At-Risk/ERA respectively. CD4 T Naive, CD4 TCM, CD8 T Naive, and CD8 TEM are mostly enriched in At-Risk/ERA. MAIT cells with the signature TFs were only found in CON clusters. **(B)** Mean PageRank scores of top 50 G2-specific TFs across cell types in G2 and other groups respectively. Rows represent TFs while columns represent cell types in G2 and other groups. Gray represents the average across other 4 groups. **(C)** Representative enriched pathways of G2-specific TFs across cell types. Bold text represents signature pathways. All the cell types were enriched in signature pathways. **(D)** Heatmap of At-Risk/ERA participants in G2 across cell types. Row shows each cell type while column shows each participant. Top bar represents the disease states of participants. Color represents the number of clusters per cell type for each participant. Twenty-five out of 26 At-Risk and ERA participants had the signature in at least one cell type but the combination and distribution of cell types are highly variable. **(E)** Heatmap of established RA participants in signature clusters across cell types. All the RA participants had the signature in at least one cell type. The signatures were enriched in CD4 T Naive, CD14 Mono and NK cells. **(F, G)** Venn diagram showing the overlap of signature transcription factors **(F)** and signature pathways **(G)** between established RA and At-Risk/ERA groups; Chi-squared test, *p < 0.1, **p < 0.05, ***p<0.01.

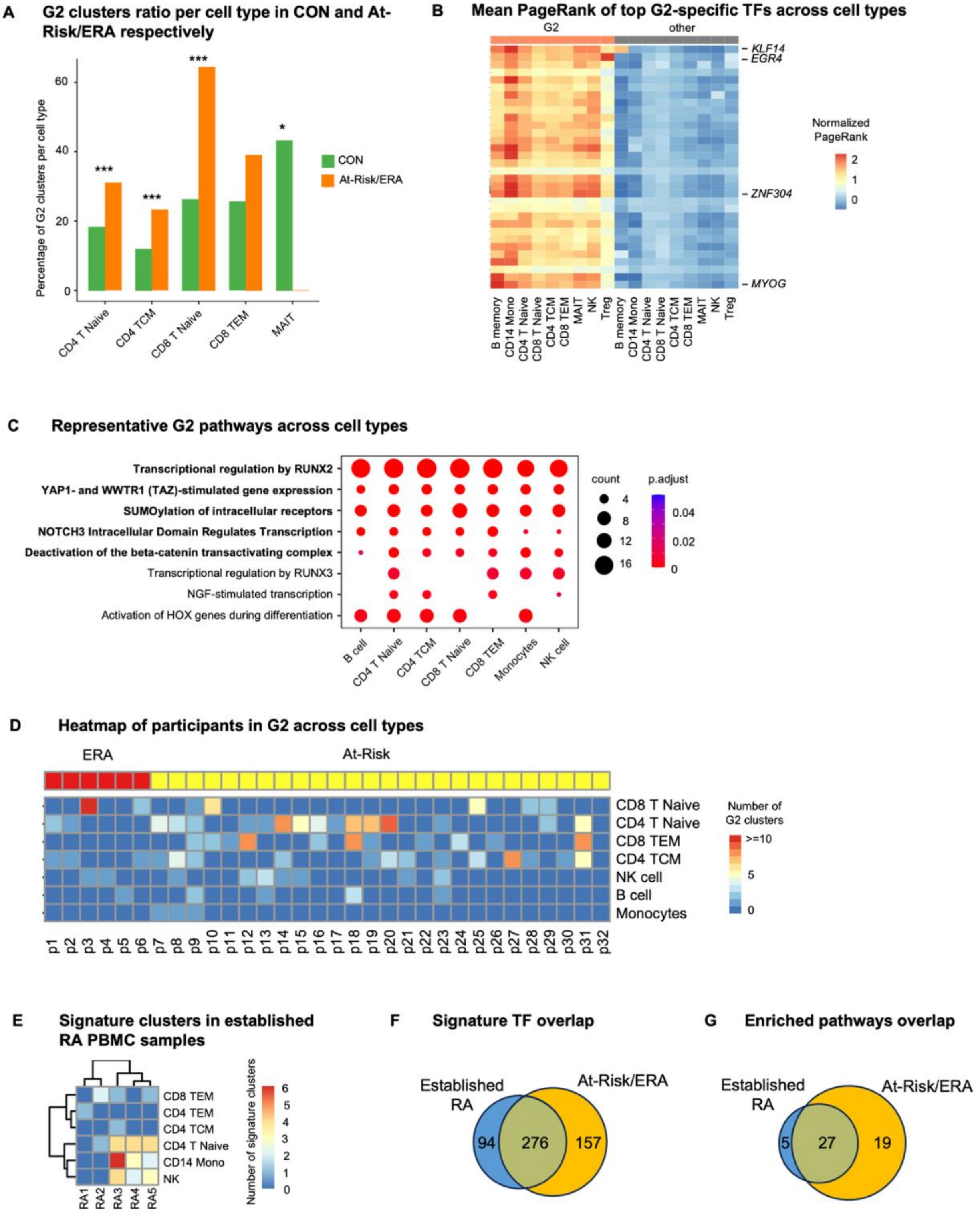
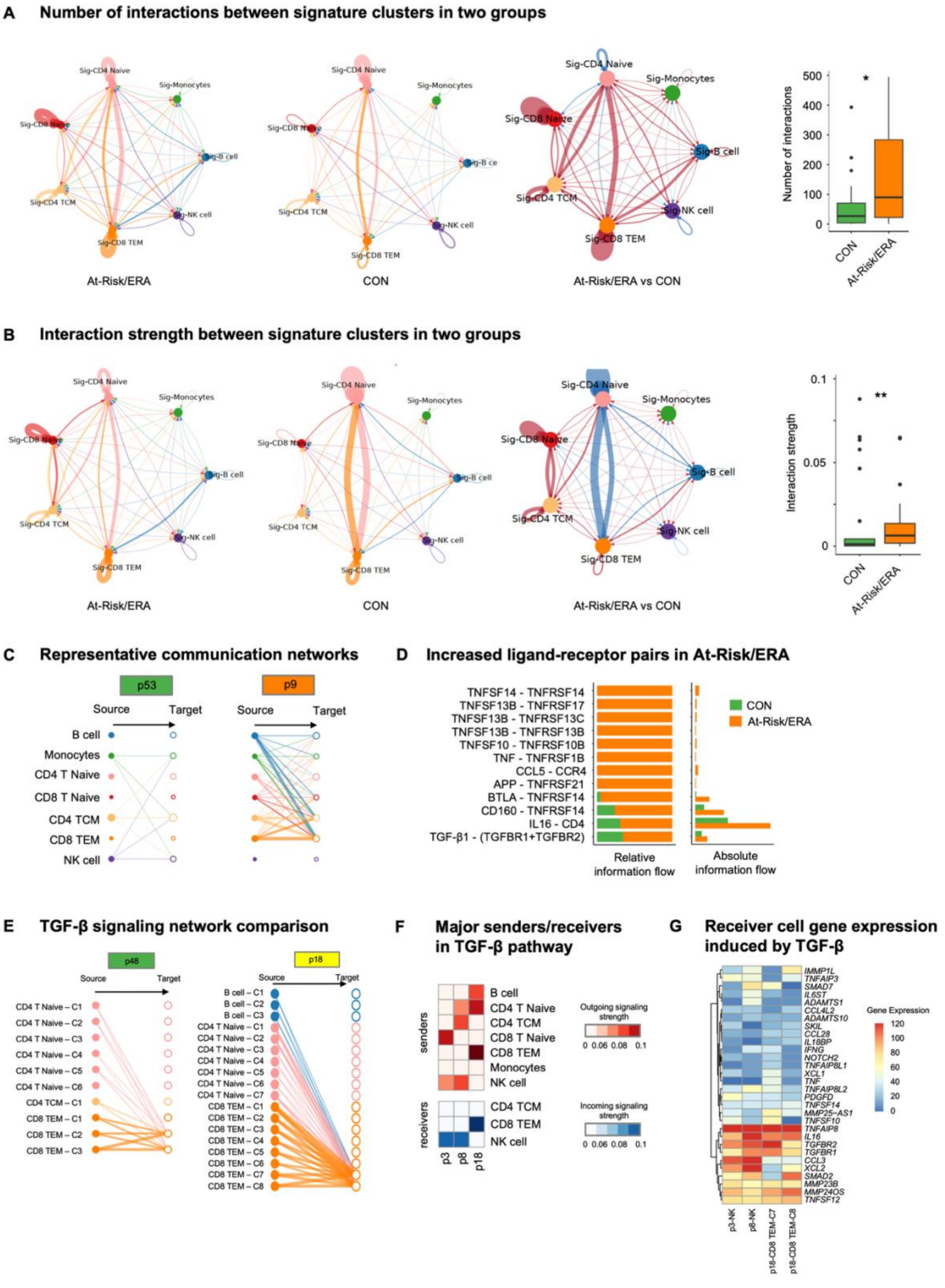

A
G2 clusters ratio per cell type in CON and At-Risk/ERA respectively
Percentage of G2 clusters per cell type
0
20
40
60
***
***
***
*
CON
At-Risk/ERA
CD4 T Naive
CD4 TCM
CD8 T Naive
CD8 TEM
MAIT
B
Mean PageRank of top G2-specific TFs across cell types
G2
other
KLF14
EGR4
ZNF304
MYOG
Normalized PageRank
2
1
0
B memory
CD14 Mono
CD4 T Naive
CD8 T Naive
CD4 TCM
CD8 TEM
MAIT
NK
Treg
B memory
CD14 Mono
CD4 T Naive
CD8 T Naive
CD4 TCM
CD8 TEM
MAIT
NK
Treg
C
Representative G2 pathways across cell types
Transcriptional regulation by RUNX2
YAP1- and WWTR1 (TAZ)-stimulated gene expression
SUMOylation of intracellular receptors
NOTCH3 Intracellular Domain Regulates Transcription
Deactivation of the beta-catenin transactivating complex
Transcriptional regulation by RUNX3
NGF-stimulated transcription
Activation of HOX genes during differentiation
B cell
CD4 T Naive
CD4 TCM
CD8 T Naive
CD8 TEM
Monocytes
NK cell
count
4
8
12
16
p.adjust
0.04
0.02
0
D
Heatmap of participants in G2 across cell types
ERA
At-Risk
CD8 T Naive
CD4 T Naive
CD8 TEM
CD4 TCM
NK cell
B cell
Monocytes
p1
p2
p3
p4
p5
p6
p7
p8
p9
p10
p11
p12
p13
p14
p15
p16
p17
p18
p19
p20
p21
p22
p23
p24
p25
p26
p27
p28
p29
p30
p31
p32
Number of G2 clusters
>=10
5
0
E
Signature clusters in established RA PBMC samples
CD8 TEM
CD4 TEM
CD4 TCM
CD4 T Naive
CD14 Mono
NK
RA1
RA2
RA3
RA4
RA5
Number of signature clusters
6
5
4
3
2
1
0
F
Signature TF overlap
Established RA
At-Risk/ERA
94
276
157
G
Enriched pathways overlap
Established RA
At-Risk/ERA
5
27
19

Fig.4 Enhanced cell-cell communication networks in At-Risk/ERA. (A) Number of cellular interactions within signature clusters in two groups. Edge thickness is proportional to the number of interactions. Thicker edge indicates more interactions. Left and middle circular plots represent networks in At-Risk/ERA and CON groups. Color represents the cell type. Right circular plot represents the differential network between At-Risk/ERA and CON. Red edge indicates more interactions in At-Risk/ERA and blue is vice versa. Rightmost panel shows the number of interactions in two groups. At-Risk/ERA group has significantly more interactions than CON. **(B)** Interaction strength within signature clusters in two groups. Edge thickness is proportional to the interaction strength. Thicker edge indicates stronger signals. Red edge indicates more intense interactions in At-Risk/ERA and blue is vice versa in the right circular plot. Rightmost panel shows the interaction strength in two groups. At-Risk/ERA group has significantly stronger interactions than CON. **(C)** Representative cellular communication networks within signature clusters in CON and At-Risk patients. Color represents the cell type and thickness of edge weight is proportional to the interaction strength. Thicker edge line indicates stronger signal. Solid and open circles represent source and target respectively. Circle size is proportional to the number of clusters. Both the edge thickness and circle size were normalized and comparable across different networks. At-Risk patient showed much denser and stronger interactions than control across almost all cell types. **(D)** Increased ligand-receptor pairs in At-Risk/ERA group. The rank is based on the difference in total information flow between At-Risk/ERA and control groups. The total information flow is calculated by summing the probability of all communications between the signature clusters. The left panel showed the relative information flow while the right panel showed the absolute information flow values. **(E)** Representative TGF-β signaling networks within signature clusters in CON and At-Risk patients. Each circle represents one Seurat cluster instance with cell type label. The At-Risk patient showed much denser and stronger interactions than CON. (**F**) Outgoing and incoming signaling strength of TGF-β pathway across cell types in At-Risk/ERA. The horizontal axis represents the cell types and vertical axis represents each individual, in which TGF-β signaling pathway is significant. Gradient red colors represent the total outgoing signaling strength with red displaying higher values. Gradient blue colors represent the total incoming signaling strength with blue displaying higher values **(G)** Expression levels of TGF-β-induced genes in each receptor cluster. *TGFBR1* and *TGFBR2* are highly expressed across all receivers**;** Wilcoxon rank-sum test, *p < 0.1, **p < 0.05.



A Number of interactions between signature clusters in two groups

B Interaction strength between signature clusters in two groups

C Representative communication networks

D Increased ligand-receptor pairs in At-Risk/ERA

E TGF-β signaling network comparison

F Major senders/receivers in TGF-β pathway

G Receiver cell gene expression induced by TGF-β

Fig.5 Biologic validation of identified key mediators in At-Risk/ERA. (A) Normalized gene expression of top 30 predictors for At-Risk/ERA and CON respectively. For each gene, the maximum gene expression across clusters was taken within each Kmeans group and each individual. Rows represent mediators while columns represent patients. Red cell represents a higher expression level of the cytokine in the patient. Top 30 predictor cytokines are uniformly more active in At-Risk/ERAs compared to controls. Example genes include *MMP23B*, *CCL4*, *IL12A*, *TNFSF14*, *IL15*, *NOTCH1*, *CCL5*, and *TGFB1.* **(B)** Protein expression level of six key mediators in each individual. At-Risk/ERA has significantly higher protein expression levels. **(C)** Normalized *TGFB1* gene expression levels in diverse cell types across At-Risk/ERAs. p32 is not shown due to lack of signature cells. The color scale is the same as Fig. 5A. **(D)** Normalized gene expression of top 30 mediators across pseudo-bulk clusters from AMP synovial tissues. Rows represent genes while columns represent pseudo-bulk clusters. Both rows and columns are hierarchically clustered. Color represents the average normalized expression across cells in the cluster, scaled for each gene across clusters. Column annotation legend represents cell types. Top mediators displayed gene expression across multiple cell types. **(E)** Normalized gene expression of top 30 mediators across synovial tissue samples. Rows represent genes while columns represent samples. Both rows and columns are hierarchically clustered. Color represents the average normalized expression across cells in the sample, scaled for each gene across samples. Each sample has its own group of highly expressed genes. **(F)** Mean gene expression across 30 mediators in various cell types across samples. Rows represent cell types while columns represent samples. Both rows and columns are hierarchically clustered. Color represents the average normalized expression across 30 mediators, scaled for each cell type across samples. Each sample has its own combinations of dominant cell types expressing the top mediators. **(G)** Estimated gene expression in monocytes across patients with RA which responded to abatacept treatment (responders) or not (non-responders) Rows represent mediators while columns represent patients. Color represents the deconvoluted expression level from bulk sample, scaled for each gene across patients. Non-responders were characterized by signature gene expression in monocytes rather than T cells as predicted; Wilcoxon rank-sum test, $^{**}p < 0.05$, $^{***}p < 0.01$, $^{****}p < 0.001$.

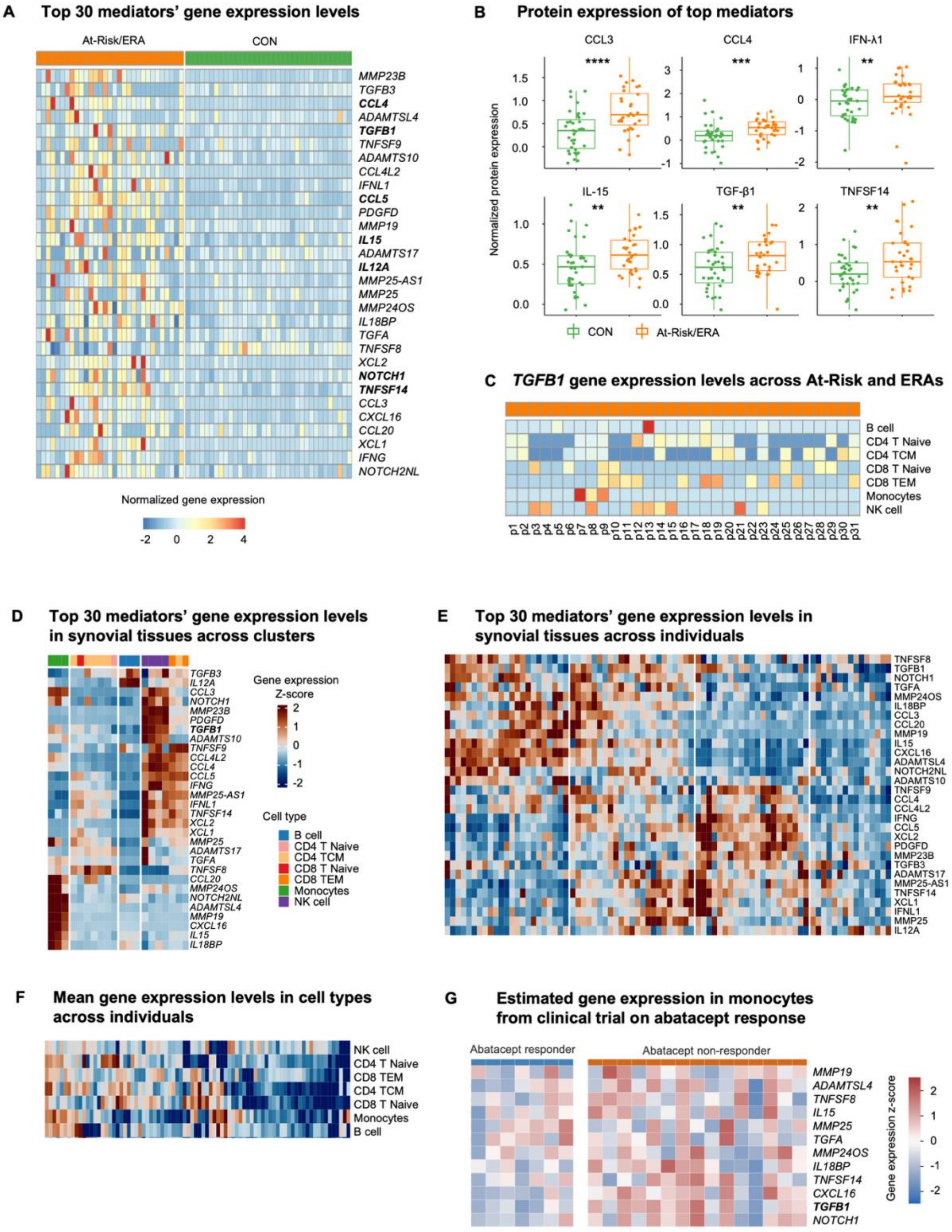
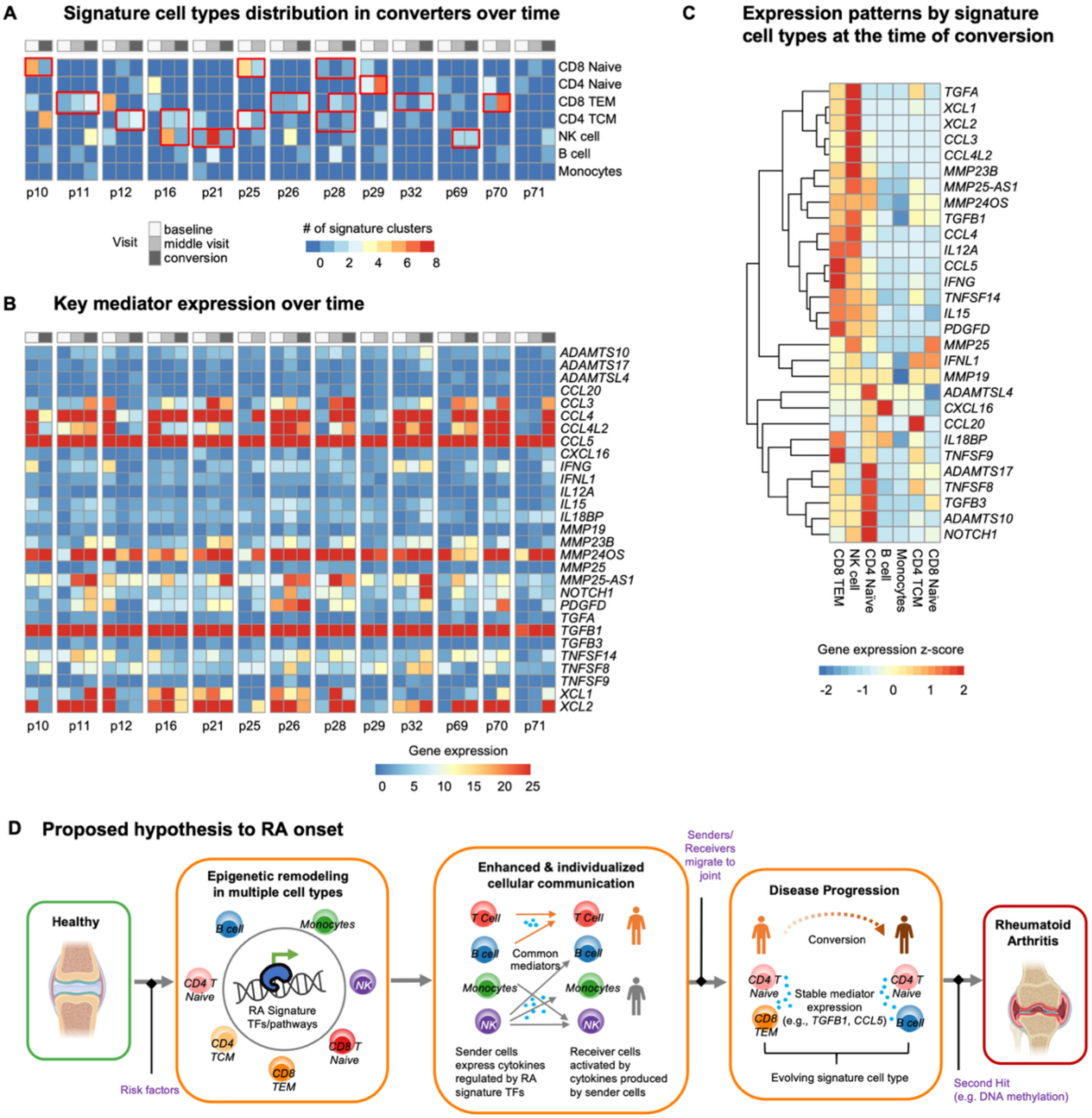

A
Top 30 mediators' gene expression levels
At-Risk/ERA
CON
MMP23B
TGFB3
CCL4
ADAMTSL4
TGFB1
TNFSF9
ADAMTS10
CCL4L2
IFNL1
CCL5
PDGFD
MMP19
IL15
ADAMTS17
IL12A
MMP25-AS1
MMP25
MMP24OS
IL18BP
TGFA
TNFSF8
XCL2
NOTCH1
TNFSF14
CCL3
CXCL16
CCL20
XCL1
IFNG
NOTCH2NL
Normalized gene expression
-2
0
2
4
B
Protein expression of top mediators
CCL3
CCL4
IFN-λ1
****
***
**
Normalized protein expression
1.5
1.0
0.5
0.0
3
2
1
0
-1
1
0
-1
-2
IL-15
TGF-β1
TNFSF14
**
**
**
1.0
0.5
0.0
1.5
1.0
0.5
0.0
2
1
0
CON
At-Risk/ERA
C
TGFB1 gene expression levels across At-Risk and ERAs
B cell
CD4 T Naive
CD4 TCM
CD8 T Naive
CD8 TEM
Monocytes
NK cell
p1
p2
p3
p4
p5
p6
p7
p8
p9
p10
p11
p12
p13
p14
p15
p16
p17
p18
p19
p20
p21
p22
p23
p24
p25
p26
p27
p28
p29
p30
p31
D
Top 30 mediators' gene expression levels in synovial tissues across clusters
TGFB3
IL12A
CCL3
NOTCH1
MMP23B
PDGFD
TGFB1
ADAMTS10
TNFSF9
CCL4L2
CCL4
CCL5
IFNG
MMP25-AS1
IFNL1
TNFSF14
XCL2
XCL1
MMP25
ADAMTS17
TGFA
TNFSF8
CCL20
MMP24OS
NOTCH2NL
ADAMTSL4
MMP19
CXCL16
IL15
IL18BP
Gene expression Z-score
2
1
0
-1
-2
Cell type
B cell
CD4 T Naive
CD4 TCM
CD8 T Naive
CD8 TEM
Monocytes
NK cell
E
Top 30 mediators' gene expression levels in synovial tissues across individuals
TNFSF8
TGFB1
NOTCH1
TGFA
MMP24OS
IL18BP
CCL3
CCL20
MMP19
IL15
CXCL16
ADAMTSL4
NOTCH2NL
ADAMTS10
TNFSF9
CCL4
CCL4L2
IFNG
CCL5
XCL2
PDGFD
MMP23B
TGFB3
ADAMTS17
MMP25-AS1
TNFSF14
XCL1
IFNL1
MMP25
IL12A
F
Mean gene expression levels in cell types across individuals
NK cell
CD4 T Naive
CD8 TEM
CD4 TCM
CD8 T Naive
Monocytes
B cell
G
Estimated gene expression in monocytes from clinical trial on abatacept response
Abatacept responder
Abatacept non-responder
MMP19
ADAMTSL4
TNFSF8
IL15
MMP25
TGFA
MMP24OS
IL18BP
TNFSF14
CXCL16
TGFB1
NOTCH1
Gene expression z-score
2
1
0
-1
-2

Fig.6 Longitudinal analysis revealed plasticity in signature cell types and stable mediator expression. (A) Signature cell type distribution in converters over time. Row shows each cell type while column shows individual participant in 2 or 3 timepoints. Top bar represents the timepoint of each converter. Color represents the number of signature clusters with red displaying more clusters. Signature cell types in an individual are not fixed but can evolve over time. **(B)** Gene expression of top 30 predictors for At-Risk/ERA and CON respectively. For each gene, the maximum gene expression across clusters was taken within each Kmeans group and each individual. Rows represent mediators while columns represent patients. Red cell represents a higher expression level of the cytokine in the patient. Top 30 predictor cytokines are uniformly more active in At-Risk/ERAs compared to controls. Example genes include *MMP23B*, *CCL4*, *IL12A*, *TNFSF14*, *IL15*, *NOTCH1*, *CCL5*, and *TGFB1.* **(D)** Proposed hypothesis to RA onset. Under the influence of risk factors such as genetics and environmental exposures, epigenetic remodeling took place in multiple cell types involving signature pathways like SUMOylation, RUNX2, YAP1, NOTCH3, and β-Catenin Pathways. The signature TFs drive a characteristic set of pro-inflammatory genes in receiver cells that can, in turn, contribute to the onset and perpetuation of RA. Diverse cell types and pathogenic mechanisms can drive a common clinical phenotype known as RA and could explain the wide variation in clinical response to agents that target individual cytokines or cell types.


A
Signature cell types distribution in converters over time
CD8 Naive
CD4 Naive
CD8 TEM
CD4 TCM
NK cell
B cell
Monocytes
p10
p11
p12
p16
p21
p25
p26
p28
p29
p32
p69
p70
p71
Visit
baseline
middle visit
conversion
# of signature clusters
0
2
4
6
8
B
Key mediator expression over time
ADAMTS10
ADAMTS17
ADAMTSL4
CCL20
CCL3
CCL4
CCL4L2
CCL5
CXCL16
IFNG
IFNL1
IL12A
IL15
IL18BP
MMP19
MMP23B
MMP24OS
MMP25
MMP25-AS1
NOTCH1
PDGFD
TGFA
TGFB1
TGFB3
TNFSF14
TNFSF8
TNFSF9
XCL1
XCL2
p10
p11
p12
p16
p21
p25
p26
p28
p29
p32
p69
p70
p71
Gene expression
0
5
10
15
20
25
C
Expression patterns by signature cell types at the time of conversion
TGFA
XCL1
XCL2
CCL3
CCL4L2
MMP23B
MMP25-AS1
MMP24OS
TGFB1
CCL4
IL12A
CCL5
IFNG
TNFSF14
IL15
PDGFD
MMP25
IFNL1
MMP19
ADAMTSL4
CXCL16
CCL20
IL18BP
TNFSF9
ADAMTS17
TNFSF8
TGFB3
ADAMTS10
NOTCH1
CD8 TEM
NK cell
CD4 Naive
B cell
Monocytes
CD4 TCM
CD8 Naive
Gene expression z-score
-2
-1
0
1
2
D
Proposed hypothesis to RA onset
Healthy
Risk factors
Epigenetic remodeling in multiple cell types
B cell
Monocytes
CD4 T Naive
NK
RA Signature TFs/pathways
CD4 TCM
CD8 T Naive
CD8 TEM
Enhanced & individualized cellular communication
T Cell
T Cell
B cell
Common mediators
B cell
Monocytes
Monocytes
NK
NK
Sender cells express cytokines regulated by RA signature TFs
Receiver cells activated by cytokines produced by sender cells
Senders/ Receivers migrate to joint
Disease Progression
Conversion
CD4 T Naive
CD8 TEM
Stable mediator expression (e.g., TGFB1, CCL5)
CD4 T Naive
B cell
Evolving signature cell type
Second Hit (e.g. DNA methylation)
Rheumatoid Arthritis

References:

1. Gravallese, E. M. & Firestein, G. S. Rheumatoid Arthritis - Common Origins, Divergent Mechanisms. *N. Engl. J. Med.* **388**, (2023).
2. Holers, V. M. *et al.* Mechanism-driven strategies for prevention of rheumatoid arthritis. *Rheumatology & autoimmunity* **2**, 109–119 (2022).
3. Holers, V. M. *et al.* Rheumatoid arthritis and the mucosal origins hypothesis: protection turns to destruction. *Nat. Rev. Rheumatol.* **14**, 542–557 (2018).
4. van Boheemen, L. *et al.* Atorvastatin is unlikely to prevent rheumatoid arthritis in high risk individuals: results from the prematurely stopped STAtins to Prevent Rheumatoid Arthritis (STAPRA) trial. *RMD open* **7**, e001591 (2021).
5. Gerlag, D. M. *et al.* Effects of B-cell directed therapy on the preclinical stage of rheumatoid arthritis: the PRAIRI study. *Ann. Rheum. Dis.* **78**, 179–185 (2019).
6. Krijbolder, D. I. *et al.* Intervention with methotrexate in patients with arthralgia at risk of rheumatoid arthritis to reduce the development of persistent arthritis and its disease burden (TREAT EARLIER): a randomised, double-blind, placebo-controlled, proof-of-concept trial. *Lancet* **400**, 283–294 (2022).
7. Deane K, Striebich C, Feser M, Demoruelle K, Moss L, Bemis E, Frazer-Abel A, Fleischer C, Sparks J, Solow E, James J, Guthridge J, Davis J, Graf J, Kay J, Danila M, Bridges, Jr. S, Forbess L, O'Dell J, McMahon M, Grossman J, Horowitz D, Tiliakos A, Schiopu E, Fox D, Carlin J, Arriens C, Bykerk V, Jan R, Pioro M, Husni M, Fernandez-Pokorny A, Walker S, Booher S, Greenleaf M, Byron M, Keyes-Elstein L, Goldmuntz E, Holers V. Hydroxychloroquine Does Not Prevent the Future Development of Rheumatoid Arthritis in a Population with Baseline High Levels of Antibodies to Citrullinated Protein Antigens and Absence of Inflammatory Arthritis: Interim Analysis of the StopRA Trial. *ARTHRITIS & RHEUMATOLOGY.* **74**, 3180–3182 (2022).

8. Rech, J. *et al.* Abatacept inhibits inflammation and onset of rheumatoid arthritis in individuals at high risk (ARIAA): a randomised, international, multicentre, double-blind, placebo-controlled trial. *Lancet* **403**, 850–859 (2024).

9. Weinand, K. *et al.* The chromatin landscape of pathogenic transcriptional cell states in rheumatoid arthritis. *Nature Communications* **15**, 4650 (2024).

10. Zhang, F. *et al.* Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. *Nat Immunol* **20**, 928–942 (2019).

11. Zhang, F. *et al.* Deconstruction of rheumatoid arthritis synovium defines inflammatory subtypes. *Nature* **623**, 616–624 (2023).

12. He, Z. *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* **17**, eadt7214 (2025).

13. Zhang, K., Wang, M., Zhao, Y. & Wang, W. Taiji: System-level identification of key transcription factors reveals transcriptional waves in mouse embryonic development. *Sci Adv* **5**, eaav3262 (2019).

14. Liu, C. *et al.* Systems-level identification of key transcription factors in immune cell specification. *PLoS Comput. Biol.* **18**, e1010116 (2022).

15. Chung, H. K. *et al.* Multiomics atlas-assisted discovery of transcription factors enables specific cell state programming. *bioRxiv* (2023).

16. Yu, B. *et al.* Epigenetic landscapes reveal transcription factors that regulate CD8 T cell differentiation. *Nature Immunology* **18**, 573–582 (2017).

17. Feinberg, M. W. *et al.* The Kruppel-like factor KLF4 is a critical regulator of monocyte differentiation. *EMBO J.* **26**, 4138–4148 (2007).

18. Intlekofer, A. M. *et al.* Effector and memory CD8+ T cell fate coupled by T-bet and eomesodermin. *Nat. Immunol.* **6**, 1236–1244 (2005).

19. Dehnavi, S. *et al.* The role of protein SUMOylation in rheumatoid arthritis. *J. Autoimmun.* **102**, 1–7 (2019).

20. Di Chen, Dongyeon J Kim, Jie Shen, Zhen Zou, Regis J O'Keefe. Runx2 plays a central role in Osteoarthritis development. *Journal of Orthopaedic Translation* **23**, 132–139 (2020).

21. Caire, R. *et al.* YAP/TAZ: Key Players for Rheumatoid Arthritis Severity by Driving Fibroblast Like Synoviocytes Phenotype and Fibro-Inflammatory Response. *Front. Immunol.* **12**, 791907 (2021).

22. Zhuang, Y. *et al.* A narrative review of the role of the Notch signaling pathway in rheumatoid arthritis. *Annals of Translational Medicine* **10**, 371–371 (2022).

23. Chen, S. *et al.* Wnt/β-catenin signaling pathway promotes abnormal activation of fibroblast-like synoviocytes and angiogenesis in rheumatoid arthritis and the intervention of Er Miao San. *Phytomedicine* **120**, 155064 (2023).

24. Vecellio, M., Cohen, C. J., Roberts, A. R., Wordsworth, P. B. & Kenna, T. J. RUNX3 and T-Bet in Immunopathogenesis of Ankylosing Spondylitis—Novel Targets for Therapy? *Front. Immunol.* **9**, 424898 (2018).

25. Jin, S. *et al.* Inference and analysis of cell-cell communication using CellChat. *Nat. Commun.* **12**, 1–20 (2021).

26. Galea, C. A., Nguyen, H. M., George Chandy, K., Smith, B. J. & Norton, R. S. Domain structure and function of matrix metalloprotease 23 (MMP23): role in potassium channel trafficking. *Cell. Mol. Life Sci.* **71**, 1191–1210 (2013).

27. Serum proteomic analysis identifies interleukin 16 as a biomarker for clinical response during early treatment of rheumatoid arthritis. *Cytokine* **78**, 87–93 (2016).

28. Iwasaki, T. *et al.* Monocyte-derived transcriptomes explain the ineffectiveness of abatacept in rheumatoid arthritis. *Arthritis Res Ther* **26**, 1 (2024).

29. Binvignat, M. *et al.* Single-cell RNA-Seq analysis reveals cell subsets and gene signatures associated with rheumatoid arthritis disease activity. *JCI Insight* **9**, e178499 (2024).

30. Inamo, J. *et al.* Deep immunophenotyping reveals circulating activated lymphocytes in individuals at risk for rheumatoid arthritis. *bioRxiv* 2023.07.03.547507 (2023) doi:10.1101/2023.07.03.547507.

31. Choi, E. *et al.* Joint-specific rheumatoid arthritis fibroblast-like synoviocyte regulation identified by integration of chromatin access and transcriptional activity. *JCI Insight* **9**, e179392 (2024).

32. Moreland, L. W. *et al.* Double-blind, placebo-controlled multicenter trial using chimeric monoclonal anti-CD4 antibody, cM-T412, in rheumatoid arthritis patients receiving concomitant methotrexate. *Arthritis Rheum* **38**, 1581–1588 (1995).

33. Prideaux, E. B. *et al.* Epigenetic trajectory predicts development of clinical rheumatoid arthritis in ACPA+ individuals: Targeting Immune Responses for Prevention of Rheumatoid Arthritis (TIP-RA). *bioRxiv* 2024.10.15.618490 (2025) doi:10.1101/2024.10.15.618490.

34. Joehanes, R. *et al.* Epigenetic Signatures of Cigarette Smoking. *Circ. Cardiovasc. Genet.* **9**, 436–447 (2016).

35. James, E. A. *et al.* Multifaceted immune dysregulation characterizes individuals at-risk for rheumatoid arthritis. *Nat. Commun.* **14**, 7637 (2023).

36. Warren, R. B. *et al.* Long-Term Efficacy and Safety of Bimekizumab and Other Biologics in Moderate to Severe Plaque Psoriasis: Updated Systematic Literature Review and Network Meta-analysis. *Dermatol Ther (Heidelb)* **14**, 3133–3147 (2024).

37. Aletaha, D. *et al.* 2010 Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. *Arthritis Rheum.* **62**, 2569–2581 (2010).

38. Swanson, E. *et al.* Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *Elife* **10**, e63632 (2021).

39. Swanson, E., Reading, J., Graybuck, L. T. & Skene, P. J. BarWare: efficient software tools for barcoded single-cell genomics. *BMC Bioinformatics* **23**, 106 (2022).

40. Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).

41. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, (2014).

42. Zhu, Y. *et al.* Constructing 3D interaction maps from 1D epigenomes. *Nat. Commun.* **7**, 10812 (2016).

43. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **28**, 1–26 (2008).

44. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* **37**, 773–782 (2019).

45. Ainsworth, R. I. *et al.* Systems-biology analysis of rheumatoid arthritis fibroblast-like synoviocytes implicates cell line-specific transcription factor function. *Nat. Commun.* **13**, 1–11 (2022).

46. Hilton, M. J. *et al.* Notch signaling maintains bone marrow mesenchymal progenitors by suppressing osteoblast differentiation. *Nat. Med.* **14**, 306–314 (2008).

47. Wei, K. *et al.* Notch signaling drives synovial fibroblast identity and arthritis pathology. *Nature* **582**, 259–264 (2020).

48. Bottini, A. *et al.* PTPN14 phosphatase and YAP promote TGFβ signalling in rheumatoid synoviocytes. *Ann. Rheum. Dis.* **78**, 600–609 (2019).

49. Ma, B. & Hottiger, M. O. Crosstalk between Wnt/β-Catenin and NF-κB Signaling Pathway during Inflammation. *Front. Immunol.* **7**, 221254 (2016).

50. Nagata, K. *et al.* Runx2 and Runx3 differentially regulate articular chondrocytes during surgically induced osteoarthritis development. *Nat. Commun.* **13**, 6187 (2022).

Acknowledgments:

This project was supported by grants from the Allen Institute for Immunology (GSF) and NIH (R01AR065466 to WW and GSF). We thank the study participants for their valuable time and contributions to this study. We thank the clinical research team at University of California, San Diego, University of Colorado, Benaroya Research Institute for recruitment and sample preparation. We thank the Allen Institute founder, P.G. Allen, for his vision, encouragement and support. We thank Adam Savage for support and critical review of the manuscript, the Allen Institute for Immunology operations team for maintaining the productive research environment, and the Human Immune System Explorer (HISE) software development team for their support and dedication. This paper and the research behind it would not have been possible without HISE, a collaborative computational data analysis environment for life sciences research.

Author contributions:

GSF, WW, KDD, VMH, JHB and TFB conceived and designed the project.
KN, VT, LL, AO, AW, MF, CS, JHB, CS identified and worked with the research subjects who participated and managed the project, with assistance from MLF, MKD, KAK, FZ, LKM, MC, BH, MS.
DB developed methodology and DB supervised the sample collection and processing.
PG, MW, VH, JR performed studies that generated data for the project. LO developed methodology and performed analysis. MAG, PS supervised data acquisition. LB is in charge of project management and TFB for cohort conceptualization.
CL and WW performed bioinformatics analysis with assistance from EBP and PW.
CL, WW, and GSF interpreted analytical results.
CL, WW and GSF drafted the initial manuscript.
All authors reviewed and edited the manuscript. All authors approved the final manuscript.

Competing interests:

J.H.B. is a Scientific Co-Founder and Scientific Advisory Board member of GentiBio, a consultant for Bristol Myers Squibb and Moderna and has past and current research projects sponsored by Amgen, Bristol Myers Squibb, Janssen, Novo Nordisk, and Pfizer. J.H.B also has a patent for tenascin-C autoantigenic epitopes in rheumatoid arthritis. The other authors declare they have no competing interests. A patent application based on these findings has been filed.

Supplementary Figures

Fig. S1 Quality control summary for scRNA-seq and scATAC-seq. (A) Violin plots showing distributions of QC metrics for scRNA-seq. Median (points) and 25th and 75th quantiles (whiskers and narrow bars) are overlaid on violin plots. Median values are also in **Supplementary Table S2**. From left to right are QC plots for percent of mitochondrial gene reads, percent of ribosomal gene reads, number of transcripts per cell, number of genes per cell, and complexity (number of genes detected per UMI). The QC metrics indicate the high quality of scRNA-seq data. **(B)** UMAP colored by samples (left) and cell types (right) in the scRNA-seq cells from all the samples. scRNA-seq cells are diffused evenly across the sample space, demonstrating a good integration across samples without batch effect. **(C)** Violin plots showing distributions of QC metrics for scATAC-seq. Median (points) and 25th and 75th quantiles (whiskers and narrow bars) are overlaid on violin plots. Median values are also in **Supplementary Table S2**. From left to right are QC plots for percent of mitochondrial gene reads, fraction of reads in TSS, fraction of reads in peaks, number of unique fragments per cell, and TSS enrichment. The QC metrics indicate the high quality of scATAC-seq data. **(D)** UMAP colored by samples (left) and cell types (right) in the scATAC-seq cells from all the samples. Color palette is the same as **Fig. S1B**. scATAC-seq cells are diffused evenly across the sample space, demonstrating a good integration across samples without batch effect. **(E)** Cluster purity of each cluster across cell types. Cluster purity is the percentage of the cells of most abundant cell type. B naive, CD14 Mono, CD16 Mono, MAIT, and NK displayed the highest purity while purity scores for T cell subsets were more diverse across clusters and relatively lower.

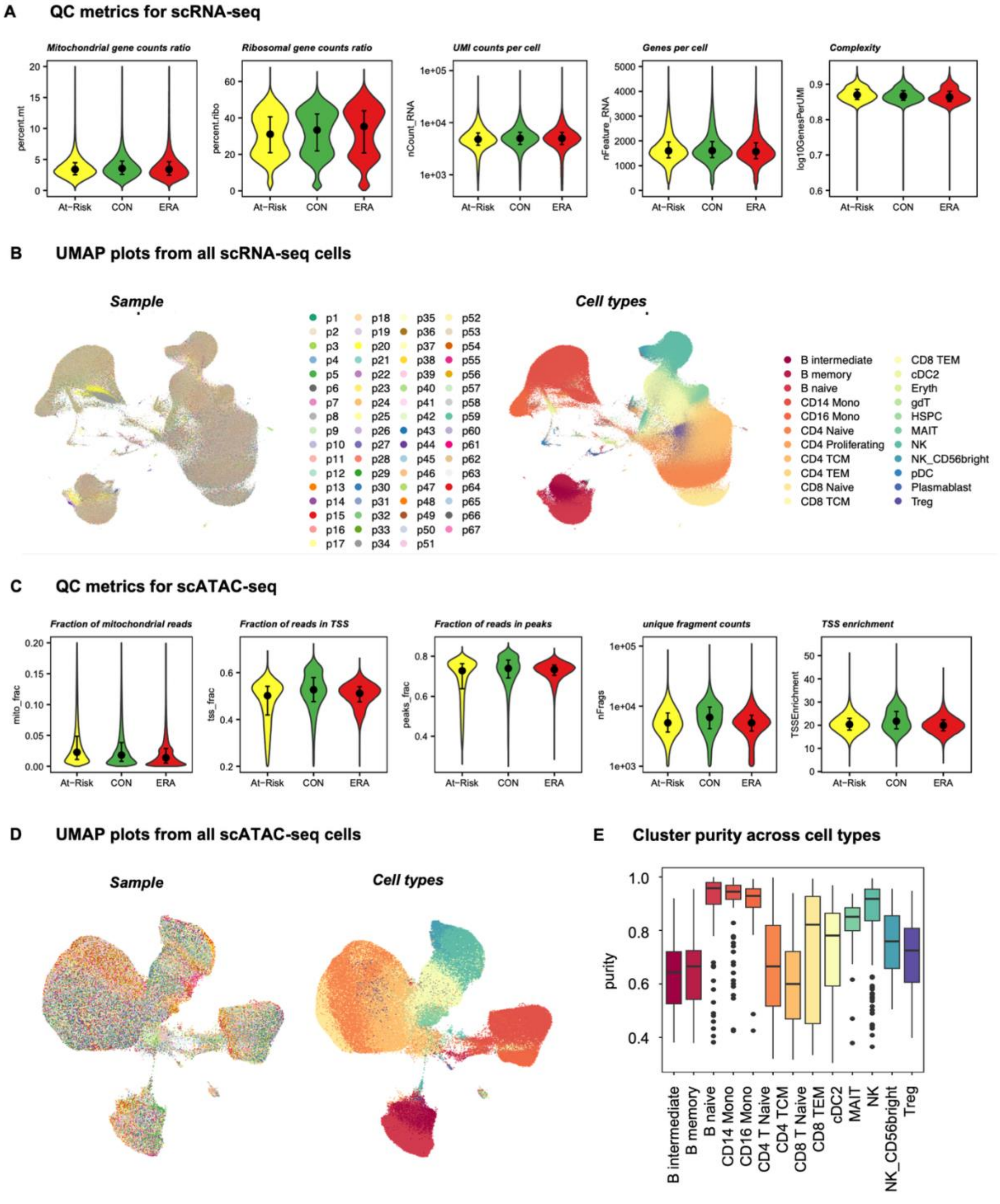

A QC metrics for scRNA-seq
Mitochondrial gene counts ratio
percent.mt
Ribosomal gene counts ratio
percent.ribo
UMI counts per cell
nCount_RNA
Genes per cell
nFeature_RNA
Complexity
log10GenesPerUMI
At−Risk
CON
ERA
B UMAP plots from all scRNA-seq cells
Sample
p1 p2 p3 p4 p5 p6 p7 p8 p9 p10 p11 p12 p13 p14 p15 p16 p17
p18 p19 p20 p21 p22 p23 p24 p25 p26 p27 p28 p29 p30 p31 p32 p33 p34
p35 p36 p37 p38 p39 p40 p41 p42 p43 p44 p45 p46 p47 p48 p49 p50 p51
p52 p53 p54 p55 p56 p57 p58 p59 p60 p61 p62 p63 p64 p65 p66 p67
Cell types
B intermediate
B memory
B naive
CD14 Mono
CD16 Mono
CD4 Naive
CD4 Proliferating
CD4 TCM
CD4 TEM
CD8 Naive
CD8 TCM
CD8 TEM
cDC2
Eryth
gdT
HSPC
MAIT
NK
NK_CD56bright
pDC
Plasmablast
Treg
C QC metrics for scATAC-seq
Fraction of mitochondrial reads
mito_frac
Fraction of reads in TSS
tss_frac
Fraction of reads in peaks
peaks_frac
unique fragment counts
nFrags
TSS enrichment
TSSEnrichment
At−Risk
CON
ERA
D UMAP plots from all scATAC-seq cells
Sample
Cell types
E Cluster purity across cell types
purity
B intermediate
B memory
B naive
CD14 Mono
CD16 Mono
CD4 T Naive
CD4 TCM
CD8 T Naive
CD8 TEM
cDC2
MAIT
NK
NK_CD56bright
Treg

Fig. S2 Unsupervised clustering shows distinct TF regulatory patterns. (A) PCA for dimension reduction of the TF score matrix. The cumulative proportion of variance explained increased with a larger number of principal components (PCs). The first 300 PCs were kept according to the “elbow” method, which explained 85% variance. **(B)** First and second PCs of all clusters with color coded by cell types and shape coded by disease state. Circle represents At-Risk; triangle represents CON and square represents ERA. Each point is one cluster. Clusters are mostly separated by cell types instead of disease states. **(C)** Selecting the best distance metric and number of groups K according to the silhouette metric. The Pearson Correlation was chosen and K=5 was the ideal number, marked as the red point in the figure. **(D)** PageRank scores heatmap of all TFs across all clusters. TFs in rows (z-normalized), clusters in columns ordered by Kmeans group, and color of the cell in the matrix indicates the normalized PageRank scores with red displaying high scores. A group of TFs are significantly active in G2. **(E)** Intersection of Kmeans group-specific TFs. Each group has its own unique set of active TFs. **(F)** Overlap of G2-specific TFs between At-Risk and ERA group. **(G)** Overlap of top 20 enriched pathways between At-Risk and ERA group. **(H)** G2 clusters out of the total clusters per cell type in CON and At-Risk/ERA respectively. Labels within the bar follow the following format: number of G2 clusters (percentage of the total clusters). **(I)** Reactome pathways enriched in the top 500 downstream genes of SUMOylation-related G2-specific TFs. Circle size represents the number of regulatees in the pathway and color represents the adjusted p-values. p-value scale bar is the same as Fig. 2D. **(J)** Intersection of TFs enriched in 5 representative signature pathways. The side horizontal bars are the original size of each pathway. RUNX2 pathway shared 3 TFs with NOTCH3 pathway, 2 TFs with SUMO pathway, and 1 TF with Wnt pathway. YAP1 has its own distinct set of TFs and have no overlap with other signature pathways. **(K)** G2-specific TFs whose regulatees are enriched in At-Risk/ERA signature pathways. p-value scale bar is the same as Fig. 2D.

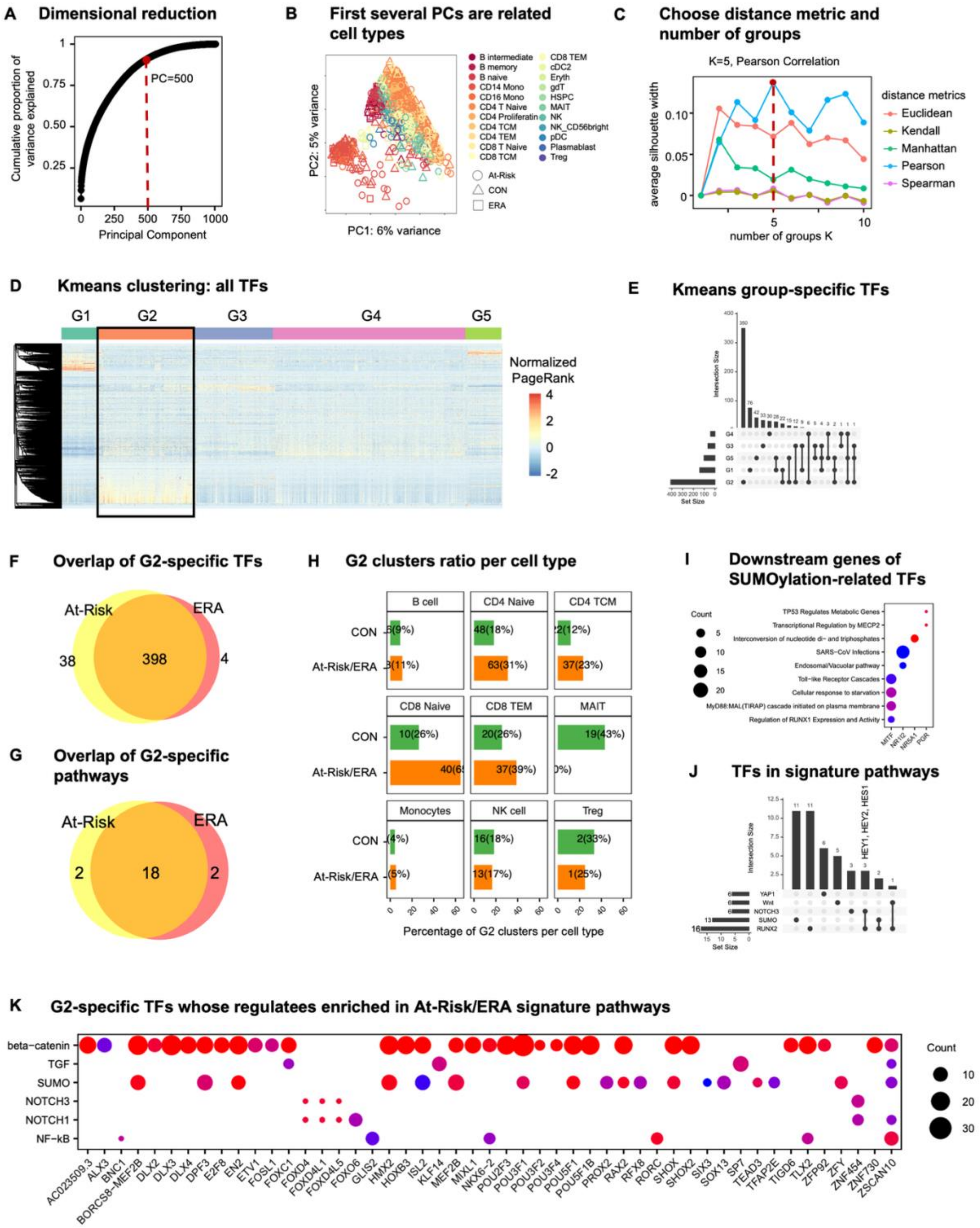
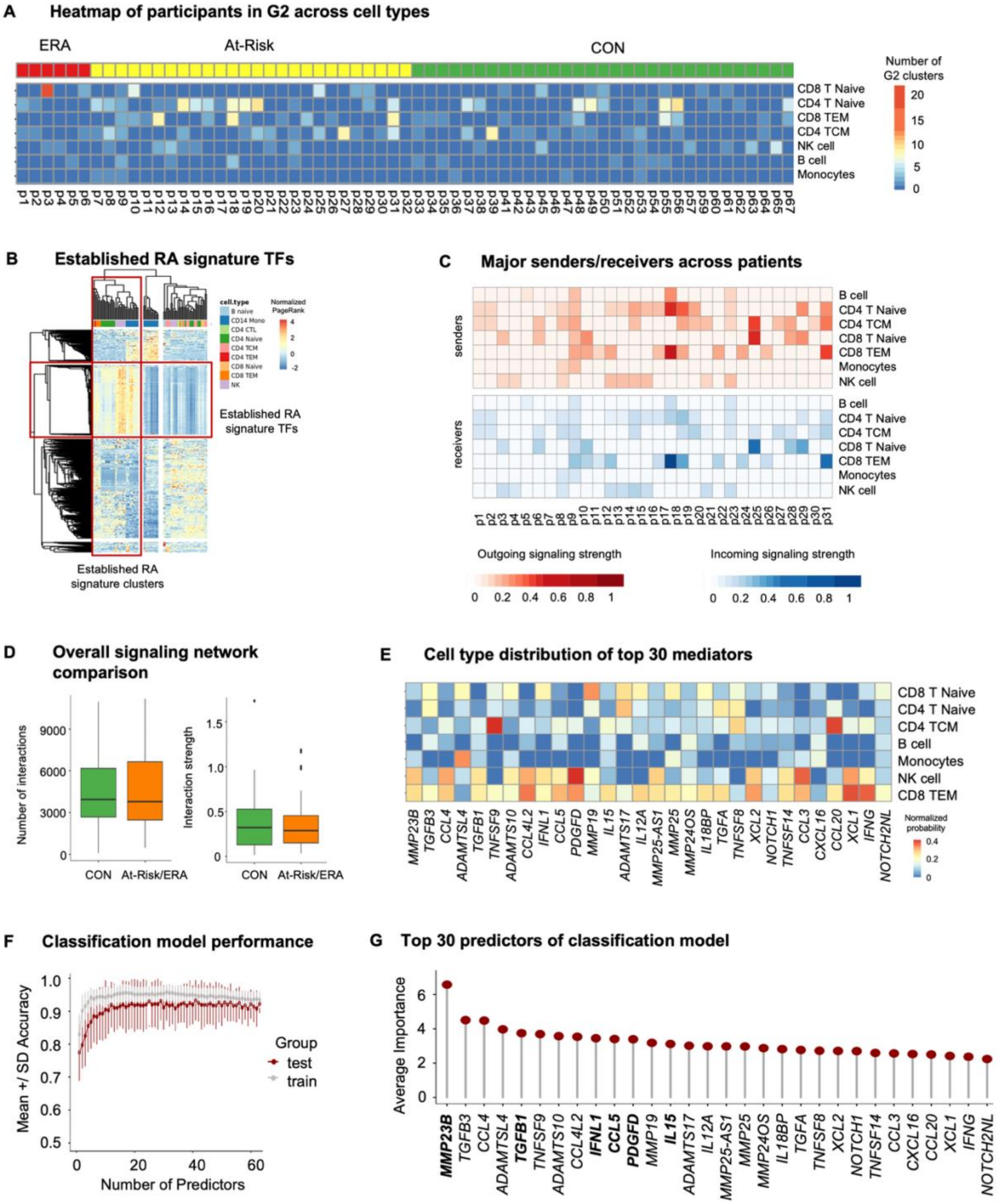

A
Dimensional reduction
Cumulative proportion of variance explained
PC=500
Principal Component
B
First several PCs are related cell types
PC1: 6% variance
PC2: 5% variance
B intermediate
B memory
B naive
CD14 Mono
CD16 Mono
CD4 T Naive
CD4 Proliferatin
CD4 TCM
CD4 TEM
CD8 T Naive
CD8 TCM
CD8 TEM
cDC2
Eryth
gdT
HSPC
MAIT
NK
NK_CD56bright
pDC
Plasmablast
Treg
At-Risk
CON
ERA
C
Choose distance metric and number of groups
K=5, Pearson Correlation
average silhouette width
number of groups K
distance metrics
Euclidean
Kendall
Manhattan
Pearson
Spearman
D
Kmeans clustering: all TFs
G1
G2
G3
G4
G5
Normalized PageRank
E
Kmeans group-specific TFs
Intersection Size
Set Size
G4
G3
G5
G1
G2
F
Overlap of G2-specific TFs
At-Risk
ERA
38
398
4
G
Overlap of G2-specific pathways
At-Risk
ERA
2
18
2
H
G2 clusters ratio per cell type
B cell
CD4 Naive
CD4 TCM
CD8 Naive
CD8 TEM
MAIT
Monocytes
NK cell
Treg
CON
At-Risk/ERA
5(9%)
48(18%)
22(12%)
3(11%)
63(31%)
37(23%)
10(26%)
20(26%)
19(43%)
40(6
37(39%)
0%)
(4%)
16(18%)
2(33%)
(5%)
13(17%)
1(25%)
Percentage of G2 clusters per cell type
I
Downstream genes of SUMOylation-related TFs
Count
TP53 Regulates Metabolic Genes
Transcriptional Regulation by MECP2
Interconversion of nucleotide di- and triphosphates
SARS-CoV Infections
Endosomal/Vacuolar pathway
Toll-like Receptor Cascades
Cellular response to starvation
MyD88:MAL(TIRAP) cascade initiated on plasma membrane
Regulation of RUNX1 Expression and Activity
MITF
NR1I2
NR5A1
PGR
J
TFs in signature pathways
Intersection Size
HEY1, HEY2, HES1
Set Size
YAP1
Wnt
NOTCH3
SUMO
RUNX2
K
G2-specific TFs whose regulatees enriched in At-Risk/ERA signature pathways
beta-catenin
TGF
SUMO
NOTCH3
NOTCH1
NF-kB
Count
AC023509.3
ALX3
BNC1
BORCS8-MEF2B
DLX2
DLX3
DLX4
DPF3
E2F8
EN2
ETV1
FOSL1
FOXC1
FOXD4
FOXD4L1
FOXD4L5
FOXO6
GLIS2
HMX2
HOXB3
ISL2
KLF14
MEF2B
MIXL1
NKX6-2
POU2F3
POU3F1
POU3F2
POU3F4
POU5F1
POU5F1B
PROX2
RAX2
RFX8
RORC
SHOX
SHOX2
SIX3
SOX13
SP7
TEAD3
TFAP2E
TIGD6
TLX2
ZFP92
ZFY
ZNF454
ZNF730
ZSCAN10

Fig. S3 At-Risk/ERA signature is shared across multiple cell types. (A) Heatmap of all participants in G2 across cell types. The horizontal axis shows the individual participants and the vertical axis shows each cell type. Top bar represents the disease states of participants. Color represents the percent of clusters per total clusters per cell type for each participant. All the At-Risk and ERA participants had the signature in at least one cell type but not all CON participants had the signature. T cells showed higher enrichment in At-Risk/ERA. **(B)** Hierarchical clustering of pseudo-bulk clusters from five established RA samples. The heatmap shows the normalized PageRank scores of all TFs across all clusters. A group of TFs are significantly active in a group of clusters, referred to as “Established RA signature TFs” and “Established RA signature clusters”, marked by red boxes. **(C)** Sum of outgoing and incoming signaling strength across all signaling pathways in At-Risk/ERA groups. **(D)** Comparison of overall intercellular interactions between control and At-Risk/ERA groups, which didn’t show any significant difference on account of number and intensity of interactions. **(E)** Row-wise normalized probability of cell types having the maximum expression of each gene in At-Risk/ERAs. Rows represent genes and columns represent cell types. Red cell represents a higher probability of the high gene expression in the specific cell type. Each cell type has its own set of highly expressed genes. For instance, *ADAMTSL4* and *CXCL16* in monocytes, *TNFSF9* in CD4 TCM cells. **(F)** Performance of random forest classification model to distinguish CON and At-Risk/ERAs. x-axis represents the number of predictors and y-axis represents the accuracy with dot as mean and the error bar as standard deviation. The test performance was monotonically increasing with more predictors. **(G)** Top 30 predictors of classification model ranked by the average importance across 20 experiments. Example top predictors include *MMP23B*, *TGFB1*, *IFNL1*, *IL15*, and *CCL5*.



A Heatmap of participants in G2 across cell types

B Established RA signature TFs

C Major senders/receivers across patients

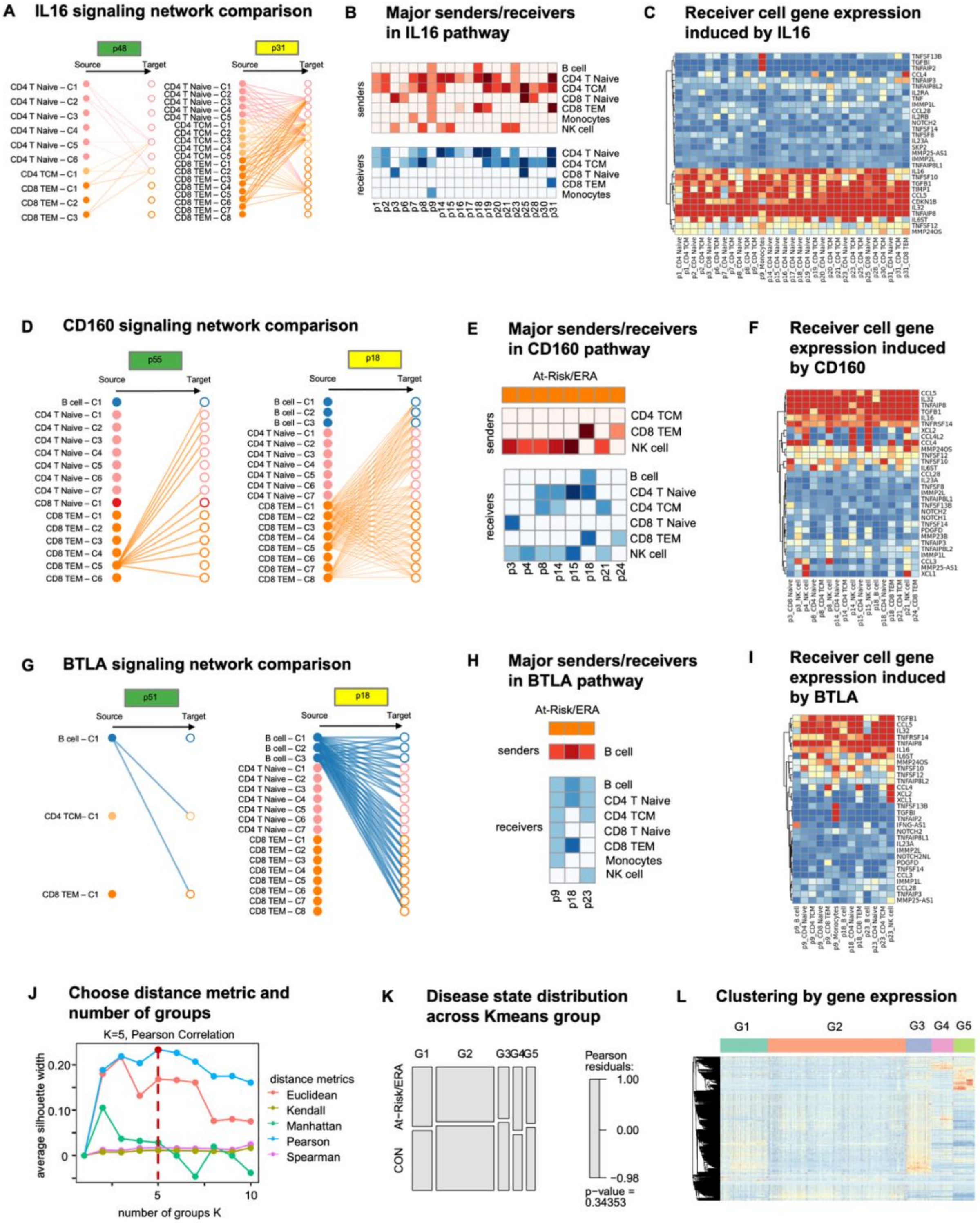
D Overall signaling network comparison

E Cell type distribution of top 30 mediators

F Classification model performance

G Top 30 predictors of classification model

Fig. S4 Cellular network comparison between two groups. **(A, D, G)** Representative signaling networks of IL16 **(A)**, CD160 **(D)**, and BTLA **(G)** within signature clusters in CON and At-Risk patients. At-Risk patient showed much denser and stronger interactions than controls. **(B, E, H)** Outgoing and incoming signaling strength of IL16 **(B)**, CD160 **(E)**, and BTLA **(H)** pathway across cell types in At-Risk/ERA. Color scale is the same as Fig. 4F. **(C, F, I)** Expression levels of IL16- **(C)**, CD160- **(F)**, and BTLA- **(I)** induced genes in receptor cells. Color scale is the same as Fig. 4G. **(J)** Selecting the best distance metric and number of Kmeans group according to the Silhouette width. The Pearson correlation was chosen and K=5 was the ideal number, marked as the red point in the figure. **(K)** Mosaic plot showing the association between disease state and Kmeans groups. G1 has slightly higher enrichment in At-Risk/ERA but not statically significant. The disease state and Kmeans groups didn't have clear association. **(L)** Normalized gene expression values heatmap of all TFs across all clusters. TFs in rows (z-normalized), clusters in columns. Both columns and rows are hierarchically clustered, and color of the cell in the matrix indicates the normalized gene expression with red displaying high expression values. Color scale is the same as Fig. 2A. Different groups displayed different gene activity patterns.


A IL16 signaling network comparison
p48
Source
Target
CD4 T Naive – C1
CD4 T Naive – C2
CD4 T Naive – C3
CD4 T Naive – C4
CD4 T Naive – C5
CD4 T Naive – C6
CD4 TCM – C1
CD8 TEM – C1
CD8 TEM – C2
CD8 TEM – C3
p31
CD4 TCM – C1
CD4 TCM – C2
CD4 TCM – C3
CD4 TCM – C4
CD4 TCM – C5
CD8 TEM – C4
CD8 TEM – C5
CD8 TEM – C6
CD8 TEM – C7
CD8 TEM – C8
B Major senders/receivers in IL16 pathway
senders
receivers
B cell
CD4 T Naive
CD4 TCM
CD8 T Naive
CD8 TEM
Monocytes
NK cell
p1 p2 p3 p6 p7 p8 p9 p14 p15 p16 p17 p18 p19 p20 p21 p23 p25 p28 p30 p31
C Receiver cell gene expression induced by IL16
TNFSF13B
TGFBI
TNFAIP2
CCL4
TNFAIP3
TNFAIP8L2
IL2RA
TNF
IMMP1L
CCL28
IL2RB
NOTCH2
TNFSF14
TNFSF8
IL23A
SKP2
MMP25-AS1
IMMP2L
TNFAIP8L1
IL16
TNFSF10
TGFB1
TIMP1
CCL5
CDKN1B
IL32
TNFAIP8
IL6ST
TNFSF12
MMP24OS
p1_CD4 Naive
p1_CD4 TCM
p2_CD4 Naive
p2_CD4 TCM
p3_CD8 Naive
p6_CD4 TCM
p7_CD4 Naive
p7_CD4 TCM
p8_CD4 Naive
p8_CD4 TCM
p9_CD4 TCM
p9_Monocytes
p14_CD4 Naive
p15_CD4 Naive
p16_CD4 Naive
p17_CD4 Naive
p18_CD4 Naive
p19_CD4 Naive
p19_CD4 TCM
p20_CD4 Naive
p20_CD4 TCM
p21_CD4 TCM
p23_CD4 Naive
p23_CD4 TCM
p25_CD4 TCM
p25_CD8 Naive
p28_CD4 TCM
p30_CD4 TCM
p31_CD4 Naive
p31_CD4 TCM
p31_CD8 TEM
D CD160 signaling network comparison
p55
Source
Target
B cell – C1
CD4 T Naive – C1
CD4 T Naive – C2
CD4 T Naive – C3
CD4 T Naive – C4
CD4 T Naive – C5
CD4 T Naive – C6
CD4 T Naive – C7
CD8 T Naive – C1
CD8 TEM – C1
CD8 TEM – C2
CD8 TEM – C3
CD8 TEM – C4
CD8 TEM – C5
CD8 TEM – C6
p18
B cell – C2
B cell – C3
CD8 TEM – C7
CD8 TEM – C8
E Major senders/receivers in CD160 pathway
At-Risk/ERA
senders
CD4 TCM
CD8 TEM
NK cell
receivers
B cell
CD4 T Naive
CD4 TCM
CD8 T Naive
CD8 TEM
NK cell
p3 p4 p8 p14 p15 p18 p21 p24
F Receiver cell gene expression induced by CD160
CCL5
IL32
TNFAIP8
TGFB1
IL16
TNFRSF14
XCL2
CCL4L2
CCL4
MMP24OS
TNFSF12
TNFSF10
IL6ST
CCL28
IL23A
TNFSF8
IMMP2L
TNFAIP8L1
TNFSF13B
NOTCH2
NOTCH1
TNFSF14
PDGFD
MMP23B
TNFAIP3
TNFAIP8L2
IMMP1L
CCL3
MMP25-AS1
XCL1
p3_CD8 Naive
p3_NK cell
p4_NK cell
p8_CD4 Naive
p8_CD4 TCM
p8_NK cell
p14_CD4 Naive
p14_CD4 TCM
p14_NK cell
p15_CD4 Naive
p15_NK cell
p18_B cell
p18_CD4 Naive
p18_CD8 TEM
p21_CD4 TCM
p21_NK cell
p24_CD8 TEM
G BTLA signaling network comparison
p51
Source
Target
B cell – C1
CD4 TCM– C1
CD8 TEM – C1
p18
B cell – C1
B cell – C2
B cell – C3
CD4 T Naive – C1
CD4 T Naive – C2
CD4 T Naive – C3
CD4 T Naive – C4
CD4 T Naive – C5
CD4 T Naive – C6
CD4 T Naive – C7
CD8 TEM – C1
CD8 TEM – C2
CD8 TEM – C3
CD8 TEM – C4
CD8 TEM – C5
CD8 TEM – C6
CD8 TEM – C7
CD8 TEM – C8
H Major senders/receivers in BTLA pathway
At-Risk/ERA
senders
B cell
receivers
B cell
CD4 T Naive
CD4 TCM
CD8 T Naive
CD8 TEM
Monocytes
NK cell
p9 p18 p23
I Receiver cell gene expression induced by BTLA
TGFB1
CCL5
IL32
TNFRSF14
TNFAIP8
IL16
IL6ST
MMP24OS
TNFSF10
TNFSF12
TNFAIP8L2
CCL4
XCL2
XCL1
TNFSF13B
TGFBI
TNFAIP2
IFNG-AS1
NOTCH2
TNFAIP8L1
IL23A
IMMP2L
NOTCH2NL
PDGFD
TNFSF14
CCL3
IMMP1L
CCL28
TNFAIP3
MMP25-AS1
p9_B cell
p9_CD4 Naive
p9_CD4 TCM
p9_CD8 Naive
p9_CD8 TEM
p9_Monocytes
p18_B cell
p18_CD4 Naive
p18_CD8 TEM
p23_B cell
p23_CD4 Naive
p23_CD4 TCM
p23_NK cell
J Choose distance metric and number of groups
K=5, Pearson Correlation
average silhouette width
0.20
0.10
0
5
10
number of groups K
distance metrics
Euclidean
Kendall
Manhattan
Pearson
Spearman
K Disease state distribution across Kmeans group
G1 G2 G3 G4 G5
At−Risk/ERA
CON
Pearson residuals:
1.00
0.00
−0.98
p−value = 0.34353
L Clustering by gene expression
G1 G2 G3 G4 G5

Supplementary Tables

Supplementary Table S1. Summary of cohort information

Supplementary Table S2. QC metrics summary for each sample

Supplementary Table S3. Cell counts in samples across different cell types

Supplementary Table S4. Co-embedded cluster distribution in Kmeans groups

Supplementary Table S5. Identified Kmeans group-specific TFs

Supplementary Table S6. Co-embedded cluster counts in Kmeans groups across disease states

Supplementary Table S7. Co-embedded cluster counts in Kmeans groups across cell types

Supplementary Table S8. Identified signature pathways with signature TFs and its representative downstream genes.

Supplementary Table S9. Curated pathogenic gene list

Supplementary Table S10. Pathway enrichment statistics for the identified Kmeans group-specific TFs (p.adjust<0.05)

Supplementary Table S11. Heatmap statistics for the comparison with AMP study in figure 6.

Supplementary Notes

Advantages of Taiji framework

We measured single cell chromatin accessibility and transcriptomic profiles of PBMCs of At-Risk individuals, ERA patients and controls. Rather than relying on individual technologies and datasets to understand important pathways, we used a novel integrative analysis (Taiji)[13] to identify potential pathogenic pathways and cell types. This method uses chromatin accessibility for TF motifs and transcription of the putatively regulated genes to prioritize TFs based on their functional relevance in a particular cell.

Taiji was previously used by our group to identify critical TFs in primary fibroblast-like synoviocytes isolated from RA synovium[45]. We were able to stratify RA patients into two groups based on divergent functions of multiple TFs and pathways. Extensive biologic validation confirmed the Taiji computational predictions related to TFs like RARA and showed that individual TFs could have diametrically opposed functions in individual patients.

Using this integrative analysis, we now report that distinctive TF profiles are significantly enriched in At-Risk and ERA individuals compared to controls in multiple cell types, especially in CD8 and CD4 T cells. This is the first time that Taiji framework has been applied to integrative single cell analysis.

Unsupervised clustering of our data based on Taiji prediction delineated five groups of individuals with distinctive TF activity profiles. At-Risk/ERA-enriched signature pathways were found in G2. Of interest, CON-enriched pathways were found in G4 and included pathways like RUNX3 that can be protective. It's worth noting that unsupervised clustering using solely gene expression profiles failed to reveal significant cohort enrichment and demonstrates the importance of integrating multiple omics data types (**Supplementary Fig. S4J-L**). These results support the importance of PageRank over individual gene expression or open chromatin analyses to delineate TF activities and differences across cohorts.

G2 At-Risk/ERA TF signature is enriched for pathways implicated in RA pathogenesis

A complete review of the biology of each pathway enriched in the TF signature is beyond the scope, but it is useful to point out some of the key features and genes and their potential relationship to RA.

Sumoylation pathway. Sumoylation plays important regulatory roles in synovial fibroblast biology including cell survival, inflammatory responses, and matrix metabolism[19]. Several SUMO related pathways including SUMOylation, SUMO E3 ligases SUMOylate target proteins, and SUMOylation of intracellular receptors were enriched in G2 (**Supplementary Table S10**). For example, several NR family members (NR1I2, NR5A1, PGR) and MITF were found in G2 (**Supplementary Table S10**). These TFs also regulate genes significantly enriched in RA-related pathways including RUNX1, Toll-like receptors, MECP2, and TP53 pathways (**Supplementary Fig. S2I**). NR1I2 regulates HLA-G, which plays an important role in RA susceptibility and regulation.

RUNX2 and NOTCH3 pathways. While most signature pathways possessed distinctive sets of active TFs, the RUNX2 pathway shared 37.5% TFs with other signature pathways, particularly with the NOTCH3 pathway (**Supplementary Fig. S2J**), which suggests the interdependence between RUNX2 and other signature pathways. For instance, three TFs (HEY1, HEY2, and HES1) were identified in both NOTCH3 and RUNX2 pathways and regulate osteoblast function[46]. NOTCH genes and signaling also play a critical role in the differentiation of synovial fibroblasts into pathogenic cells[47].

YAP1 pathway. Recent evidence suggests a critical regulatory role of the Hippo pathway in the RA pathogenesis. As one of the key components, YAP promotes the localization of SMAD3 in RA fibroblast-like synoviocytes and enhances aggressiveness[48]. Unlike RUNX2 and NOTCH3 pathways, YAP1 pathway did not share any TFs with other signature pathways (**Supplementary Fig. S2J**). Multiple TEAD family members were also identified as important in the YAP1/TAZ pathway by binding and promoting gene expression. Previous studies indicate that inhibiting YAP-TEAD interaction reduces RA synoviocytes invasion[21].

β-catenin pathway. Although only a few TFs were enriched in β-catenin pathways, many of the TFs including DLX3, MEF2B, POU3F1, RAX2, TLX2 from different families had regulatees involved in β-catenin-related pathways (**Supplementary Fig. S2K**). For example, PAX4 regulates many functional genes including BCL9L, CARD11, AURKB, NR3C1, and BIRC2 that are significantly enriched in the signature pathways.

Other pathways. Besides of the signature pathways which are well known to be involved in RA pathogenesis, we also found some novel pathways of potential interest in RA (**Supplementary**

Table S10). For example, regulation of beta-cell development was ranked as one of the top G2-specific pathways, in which many TFs from HNF and NKX families are involved. Other examples are transcriptional regulation of pluripotent stem cells pathway and regulation of gene expression in late-stage pancreatic bud precursor cells pathway.

G4 is significantly enriched in the CON cohort (32% higher in CON, p-value < 0.0001; Chi-squared test) (**Fig. 2C**). β-catenin, Wnt, and RUNX related pathways were enriched in both G2 and G4 (**Fig. 3C**), suggesting potential multifaceted roles for specific signature pathways across different participant groups. However, the genes implicated in G4 were distinct from G2. For example, the former was enriched in the formation of the β-catenin:TCF transactivating complex while latter was associated with the deactivation of the β-catenin transactivating complex. This is consistent with the observations that Wnt/β-catenin signaling can exert either anti-inflammatory and proinflammatory functions depending on the context[49]. Similarly, RUNX3, which is also a pathway associated with CON, is chondroprotective in preclinical models of arthritis[50].