Supplementary Materials:

fig. S1

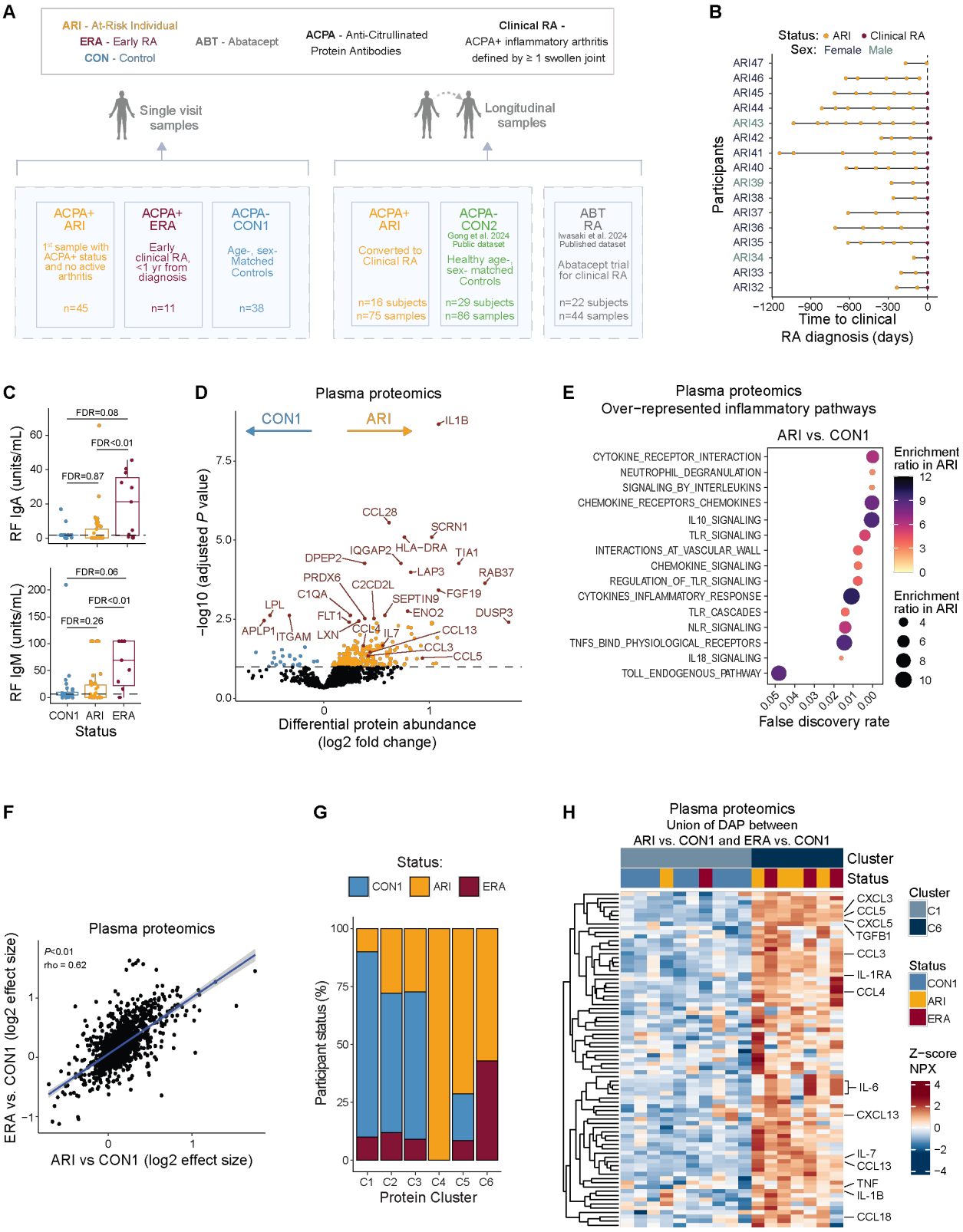

Fig S1; related to data figure 1.
(A) Consort diagram of groups and data sets used in this study. (B) Longitudinal blood sampling frequency of ARI who progressed to clinical RA. Onset of clinical RA is denoted by dashed vertical line (time to clinical RA diagnosis = 0). Biological sex is denoted by participant ID color. (C) Baseline (initial) RF-IgA and RF-IgM levels from CON1, ARI, ERA. (D) Volcano plot of baseline differential plasma protein abundance elevated in ARI (orange) or CON1 (blue). Each dot represents a single protein assayed. The 20 proteins with the smallest *P* values, in addition to select inflammatory proteins, are noted in red. (E) Over-represented inflammatory pathways (MSigDB Hallmark, KEGG, Reactome) in ARI. Enrichment ratios are shown by color and size. (F) Comparison of differential protein abundance between ARI vs. CON1 and ERA vs. CON1 (Spearman $\rho = 0.62$). (G) Number of ARI, ERA, CON1 participants comprising each protein cluster from Fig. 1C. (H) Z-scored NPX of differentially abundant inflammatory mediators between clusters Prot-C1 and Prot-C6. Columns indicate participants. Select proteins (rows) are labeled. Boxplots show median (centerline), first and third quartiles (lower and upper bound of the box) and whiskers show the 1.5x interquartile range of data. *P* values were calculated by Kruskal-Wallis followed by Dunn's post hoc testing (C), linear regression models (D), or hypergeometric tests (E). Nominal *P* value is indicated for (F). FDR values are indicated for remaining panels.

fig. S2

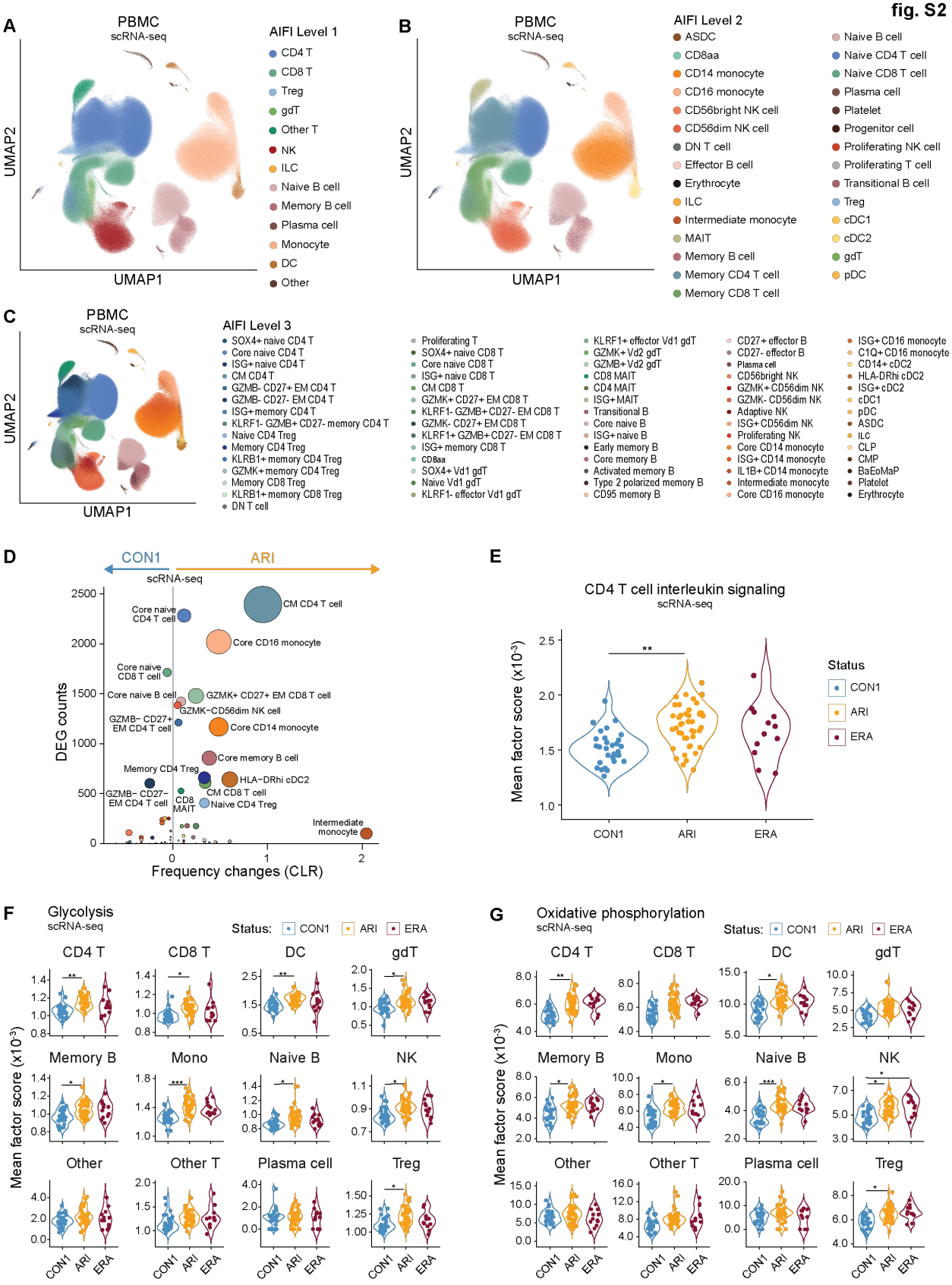

Fig. S2; related to data figure 1.
(A-C) UMAPs of cell subsets labeled using the Allen Institute for Immunology Immune Cell Atlas at levels 1 (A), 2 (B), and 3 (C). (D) Comparison DEG number with the change in frequency (centered-log ratio (CLR) transformation) between ARI and CON1 at level 3 cell types. Bubble size corresponds to the aggregate score calculated by [-log(padj CLR frequency changes) x total number of DEGs]. (E-G) Spectra factor scores for interleukin signaling in CD4 T cells (E), and oxidative phosphorylation (F) and glycolysis (G) from ARI, ERA, CON1 across all level 1 cell types. *P* values were calculated using linear regression modeling (D-G). For (E-G), all pairwise comparisons were tested and FDR values are indicated for those that are significant. *FDR < 0.05; **FDR < 0.01.

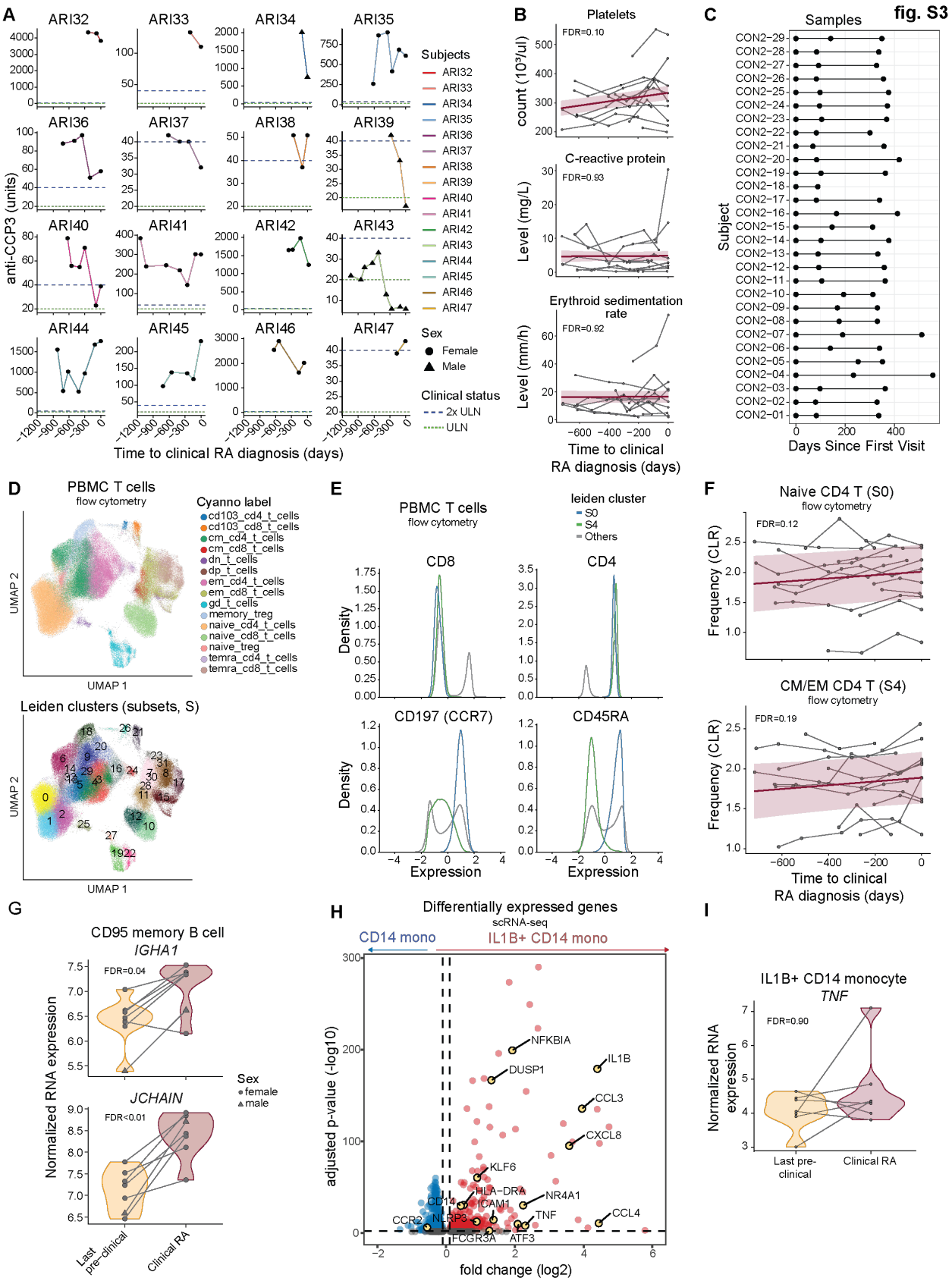

Fig. S3; related to data figure 2.
(A) Longitudinal anti-CCP3 serum levels in ARI who progressed to clinical RA. Each plot represents a different participant, and biological sex is indicated by point shape. Dashed horizontal lines indicate the upper limit of normal (ULN; 20 units; green) and 2x ULN (40 units; blue). (B) Clinical lab features in ARI who progressed to clinical RA. Each participant's longitudinal series is connected by a gray line, with a group trendline and 95% confidence interval in purple. (C) Longitudinal sample inclusion from CON2 for comparison of intra-donor coefficients of variation in Fig. 2B. (D-F) PBMC were analyzed by flow cytometry and clustered into subsets (S) for abundance comparisons. Significant subsets were manually reviewed and annotated. (D) UMAPs annotated for Cyanno cell type labels (top) and clustered subsets (bottom). (E) T cells markers in S0 and S4 compared to all other subsets. (F) Centered log-ratio (CLR)-transformed frequency changes of subsets annotated as naive CD4 and central memory/effector memory (CM/EM) CD4 T cells over time as ARI progress to clinical RA. (G) RNA expression of IGHA1 and JCHAIN in CD95 memory B cells. Paired donor samples from their last pre-symptomatic and diagnosis of clinical RA visits are connected by lines. (H) Volcano plot comparing genes with elevated expression in IL1B+ CD14 monocytes (red) vs. core CD14 monocytes (blue). (I) Normalized RNA expression of TNF in IL1B+ CD14 monocytes, as in (G). *P* values were calculated using linear mixed models (B-C), paired Wald test (G,I), or Wilcoxon rank-sum test (H). FDR values are indicated for all panels.

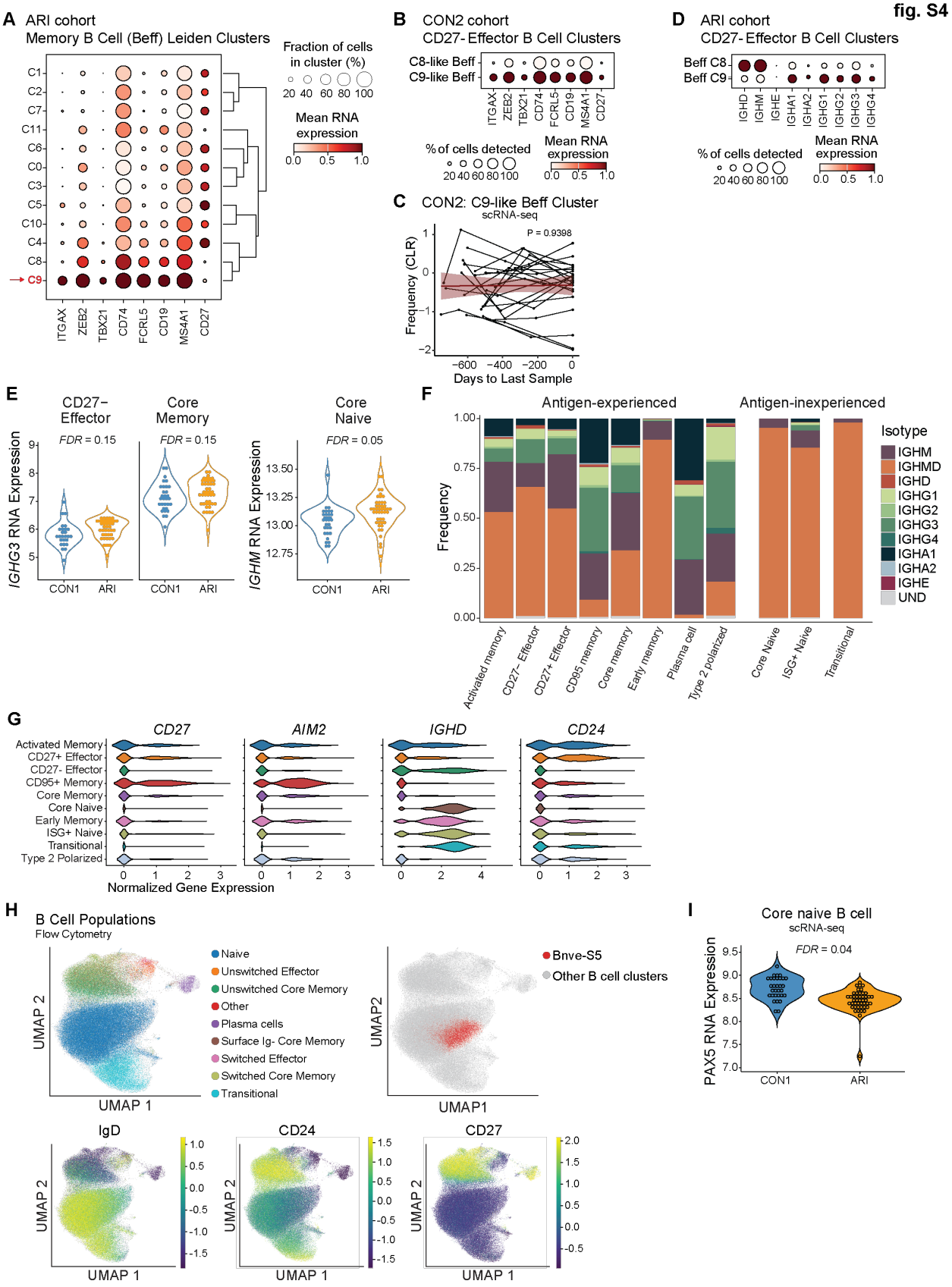

Fig. S4; related to data figure 3.
(A) Subset-defining expression among ARI memory B cell leiden clusters from scRNA-seq. (B) Subset-defining gene expression for CD27- effector B cell identity clusters. (C) Longitudinal centered log-ratio (CLR) transformed frequencies for Beff C9-like cluster in CON2 over a 2-year span. Each participant's longitudinal series is connected by a gray line, with a group trendline and 95% confidence interval in purple. (D) IGH gene expression levels for Beff-C8 and Beff-C9 from ARI. (E) *IGHG3* RNA expression by core memory (P=0.02; FDR=0.15) and CD27- effector (P=0.004; FDR=0.15) B cells and *IGHM* expression by naive B cells (P=0.02; FDR=0.05) of ARI and CON1. (F) B cell IgH isotype composition, as frequency within population, for all subsets. (G) CD27, AIM2, CD24 and IGHD gene expression by B cell subsets in ARI. (H) Flow cytometry UMAP plots for B cells showing population labels determined by Cyanno model-based approach (top left), Bnve-S5 cells (top right, red dots), and overlaid subset-defining marker expression (bottom). (I) PAX5 gene expression in naive B cells from ARI and CON1. *P* values were calculated using linear mixed models (C), Wald test in DESeq2 with Storey-Tibshirani procedure (E), and Wilcoxon rank-sum test (I). Nominal *P* value is indicated for (C). FDR values are indicated (E,I).

fig. S5

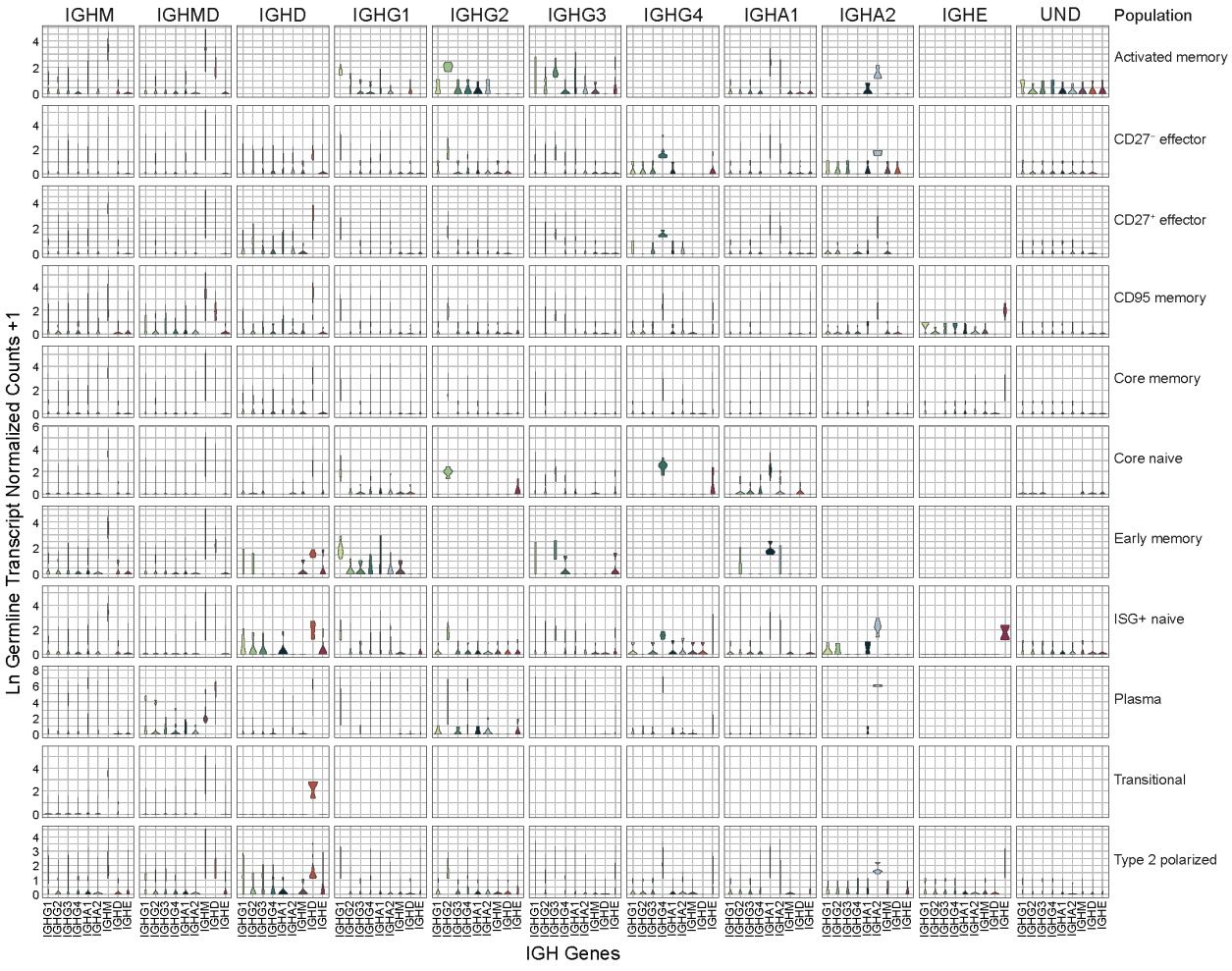

Fig. S5; related to data figure 3 and methods.
Log1p-transformed IGH gene germline transcription (GLT) normalized counts for each B cell isotype and subset in scRNA-seq data.

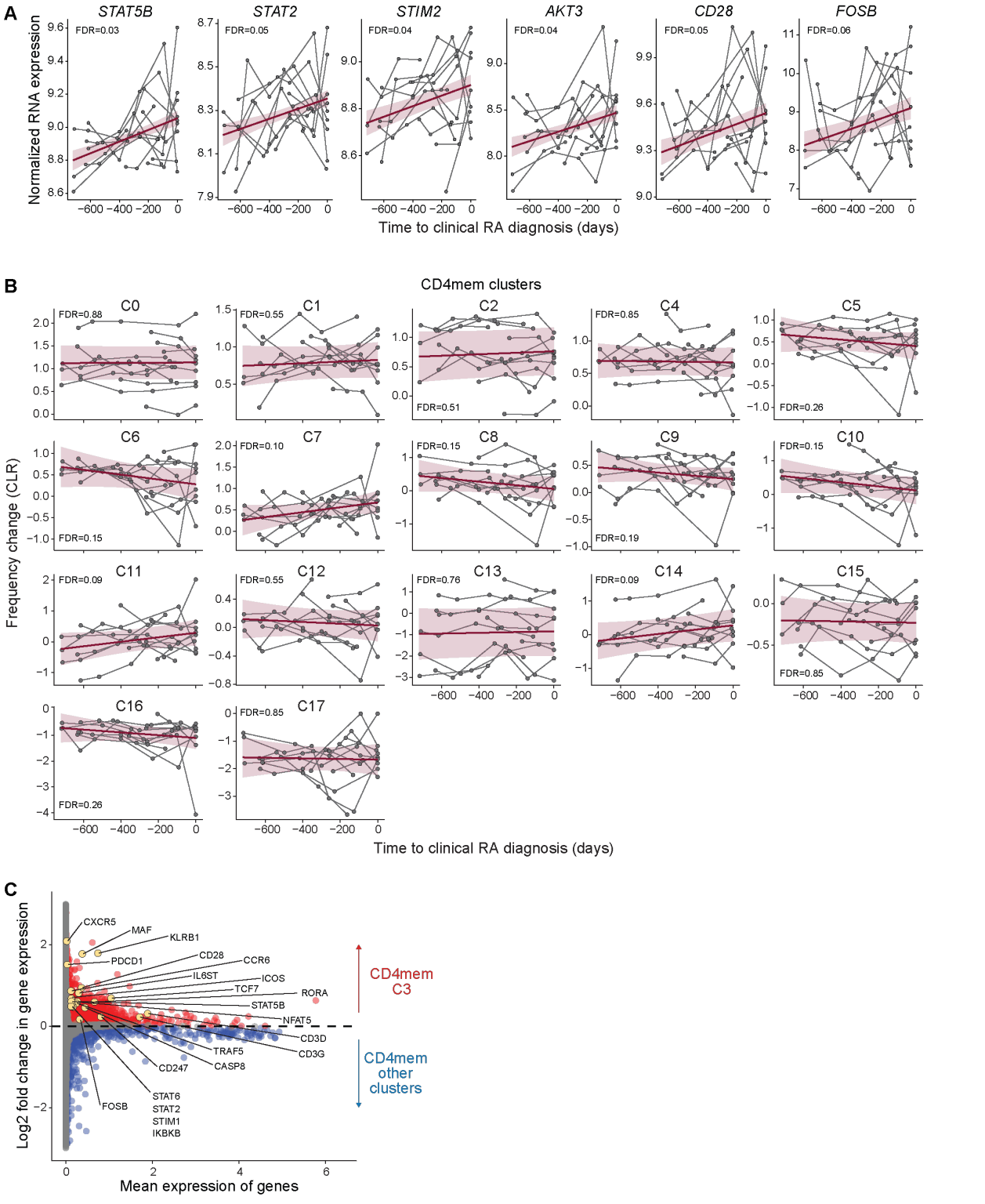

Fig. S6. related to data figure 4.

(A) RNA expression of select genes associated with activation in central memory (CM) CD4 T cells as ARI progress to clinical RA. Genes were selected based on Fig. 4A. Each participant's

longitudinal series is connected by a gray line, with a group trendline and 95% confidence interval in purple. (B-C) CD4mem Leiden clusters were derived from non-negative matrix factorization (NMF)-projected CD4 reference gene weights from Yasumizu *et al*. onto CD4mem T cells (see Fig. 4C). (B) Centered log-ratio (CLR)-transformed frequency for each cluster is shown over time as ARI progress to clinical RA. Group trendlines were determined as in (A). (C) Comparison of RNA expression in cluster CD4mem-C3 (red) vs all remaining clusters (blue) over time as ARI progress to clinical RA, with mean expression of each gene. *P* values were calculated using linear mixed models (A-B). FDR values are indicated for all panels.

fig. S7

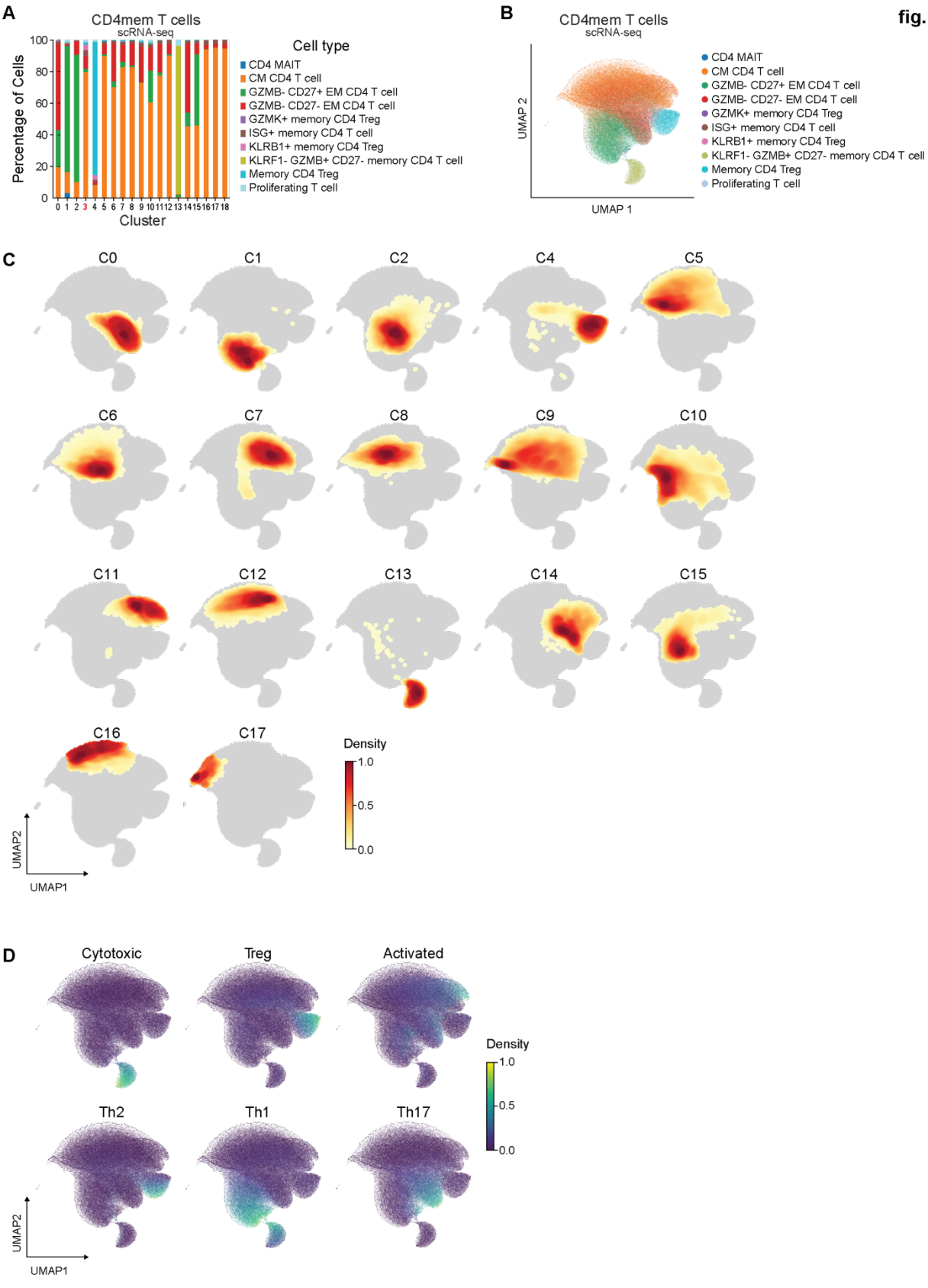

Fig S7. related to data figure 4.
(A-D) CD4mem T clusters were derived as in Fig. S6B. Quantitation (A) and UMAP (B) by Allen Institute for Immunology Immune Cell atlas level 3 labels. (C) UMAP density plots for each cluster. CD4mem-C3 is shown in Fig. 4H. (D) CD4mem T cells expressing polarized gene programs are distinguished based on the NMF projection using a pre-computed weight matrix of CD4 T cell population from Yasumizu *et al*. The Tfh density is shown in Fig. 4H.

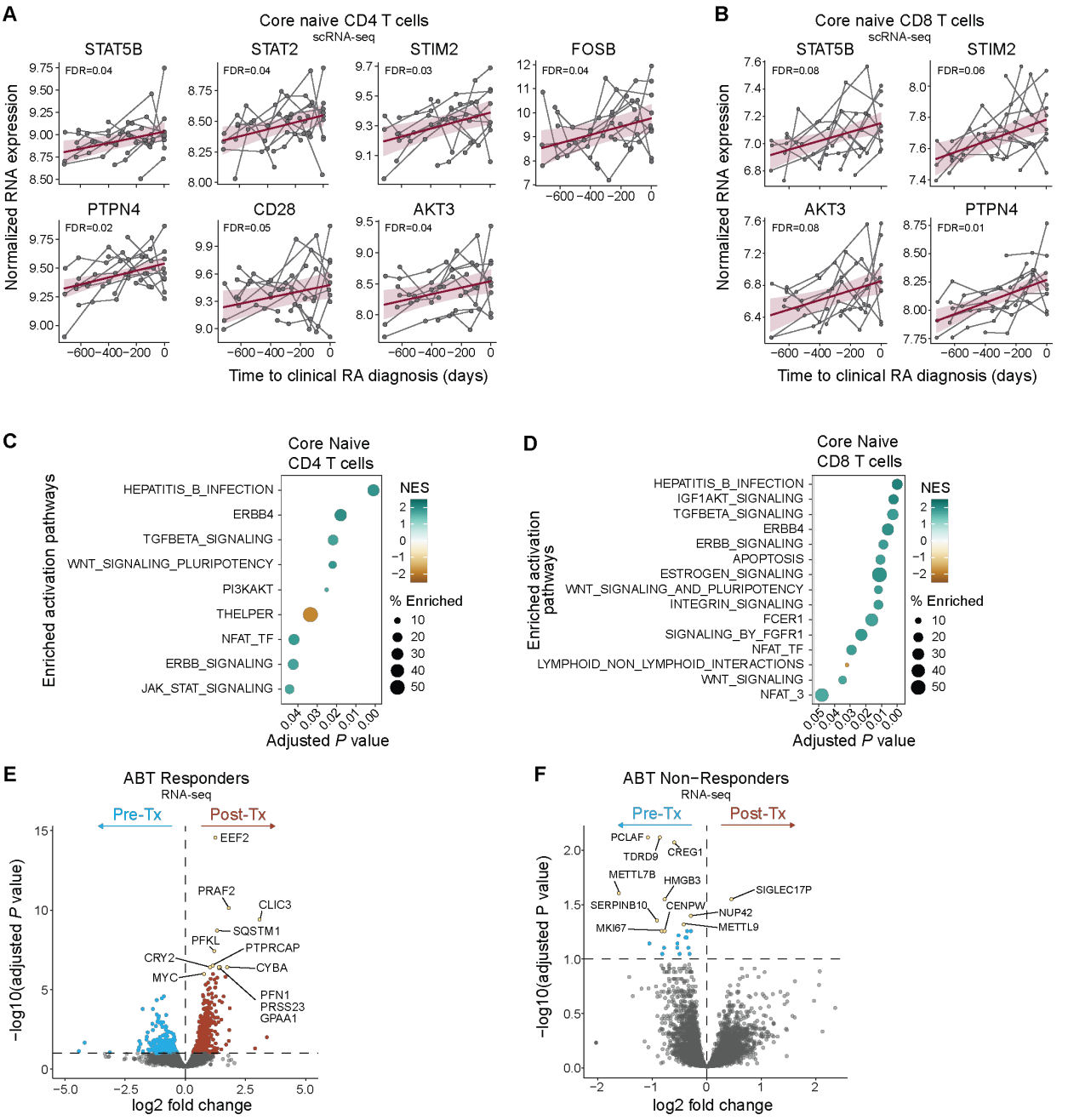

Fig. S8. related to data figure 5.

(A,B) Changes in RNA expression of select genes associated with T cell activation over time as ARI progress to clinical RA in core naive CD4 (A) and CD8 (B) T cells. Genes were selected based on Fig. 5A and Fig. 5C. Group trendlines, derived from linear mixed models, are shown with a purple line with 95% confidence interval in the shaded area. (C,D) Enriched pathways associated with T cell activation in core naive CD4 (C) and CD8 (D) T cells over time as ARI progress to clinical RA. Normalized enrichment scores (NES), by GSEA, are shown. (E,F) Volcano plots, derived from reanalysis of Iwasaki et al. (Fig. 5E), showing the differential expressed genes in responders (D) and non-responders (E) before and after abatacept (ABT)

treatment. *P* values were calculated using linear mixed models (A-B) or Wald test (E-F). FDR values are indicated for all panels.

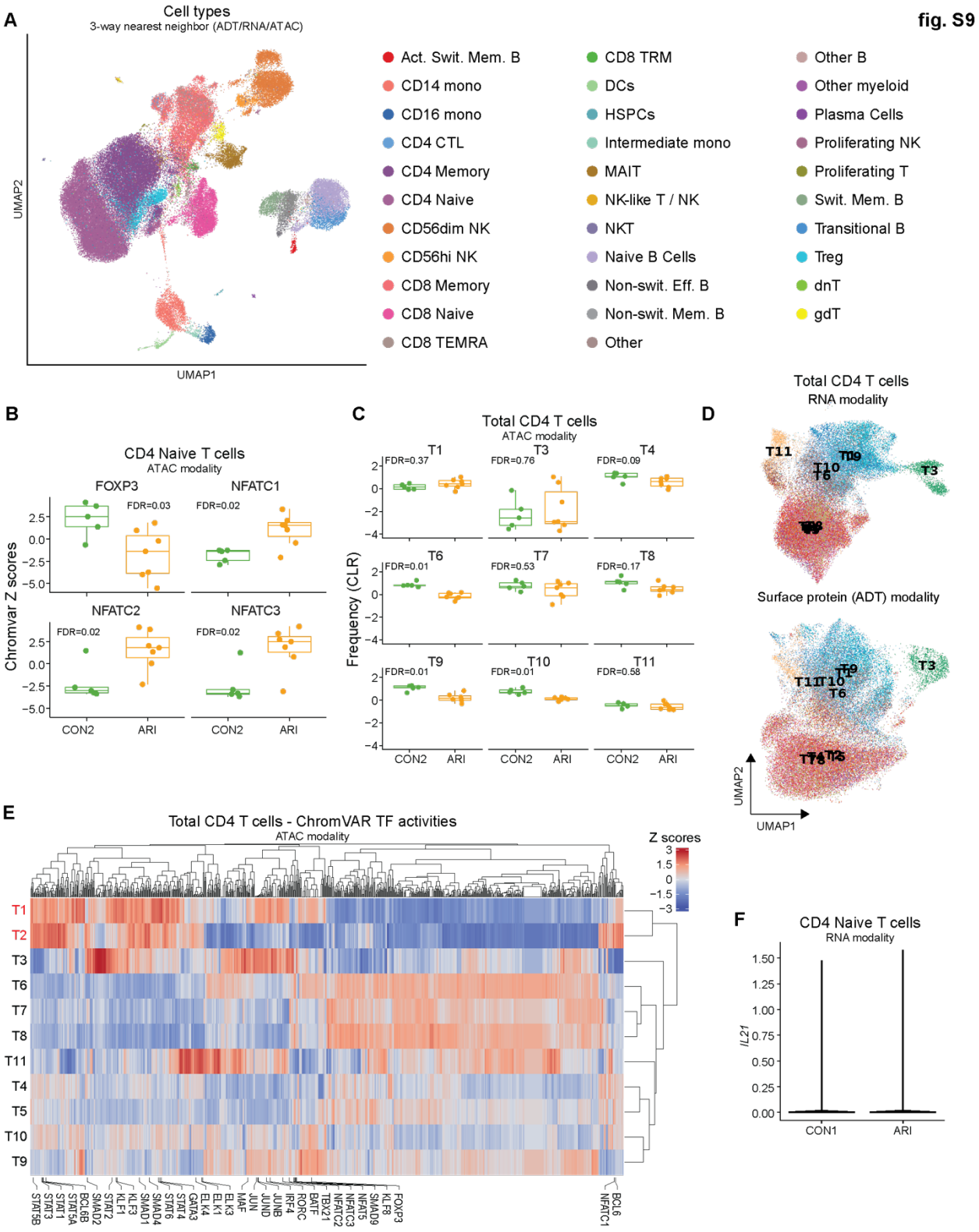

Fig. S9. related to data figure 6.

(A) Three-way weighted nearest neighbor UMAP of TEA-seq data incorporating surface protein, transcript, and chromatin accessibility, colored by cell type label. (B) ChromVAR Z scores of

selected transcription factor activities in naive CD4 T cells. (C) CLR frequency of CD4 T cells ATAC clusters. (D) UMAP of CD4 T cells in RNA modality (top) and surface protein (ADT) modality (bottom). (E) Heatmap of ChromVAR TF activity scores of 870 TFs among clusters, scaled by column. Selected TFs related to T cell activation and differentiation are labeled. (F) Normalized RNA expression of IL21 in ARI and CON1. Boxplots show median (centerline), first and third quartiles (lower and upper bound of the box) and whiskers show the 1.5x interquartile range of data. *P* values were calculated using the Wilcoxon rank-sum test (B-C). FDR values are indicated (B-C).

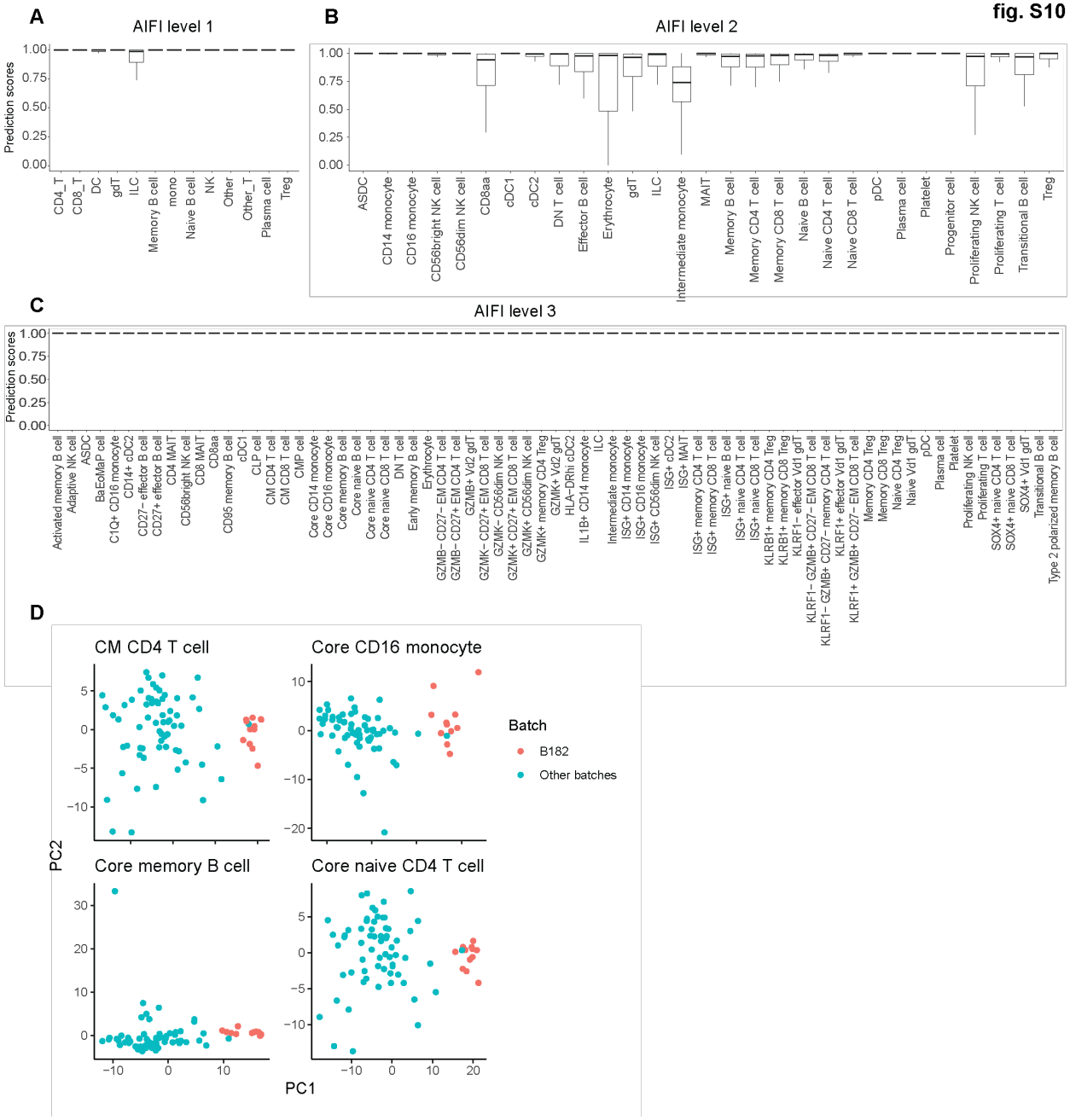

Fig. S10: related to methods.
Prediction scores of the CellTypist model generated from Allen Institute for Immunology (AIFI) Immune Cell Atlas cell types for level 1 (A), level 2 (B), and level 3 (C). Boxplots show median (centerline), first and third quartiles (lower and upper bound of the box) and whiskers show the 1.5x interquartile range of data. (D) Principal component analysis of pseudobulk scRNA data for CM CD4 T cells, core memory B cells, and core CD16 monocytes. Each dot represents a sample, colored by whether the sample was processed in batch 182 or other batches in the dataset.

Supplementary Table S2: Summary of baseline cohort information

Variable	ACPA-CON1 (n = 38)	ACPA+ ARI (n = 45)	ACPA+ ERA (n = 11)	*P* value[1]
Age at sample collection, mean years (SD)	56 (16)	57 (16)	49 (9)	0.23
Sex: Female, n (%)	34 (89%)	36 (80%)	10 (91%)	0.52
Sex: Male, n (%)	4 (11%)	9 (20%)	1 (9%)	-
Ethnicity: Non-Hispanic origin, n (%)	29 (76%)	43 (96%)	8 (73 %)	0.02
Race: White, n (%)	32 (84%)	37 (82%)	7 (64%)	0.27
BMI, mean kg/m2 (SD)	28 (6)	27 (5)	30 (8)	0.66
Met ACR/EULAR 2010 criteria, n (%)	-	-	11 (100%)	-
C-reactive protein, median mg/L (IQR)	1.6 (0.6-3.4)	1.5 (0.9-4.7)	3.3 (1.0-10)	0.39
Erythrocyte sedimentation rate, median mm/h (IQR)	9 (4-13)	11 (7-20)	20 (11-33)	0.07
Shared epitope present, n (%)[2]	13 (38%)	18 (40%)	6 (55%)	0.69
Serum ACPA, median (IQR)	6 (6-6)	62 (44-195)	560 (230-1101)	<0.01 (<0.01 Con vs. ARI; <0.01 Con vs. ERA)
Serum RF IgM, median (IQR)	5 (5-9.7)	0.9 (0.3-23)	69 (23-105)	0.01 (<0.01 ARI vs. ERA; 0.06 Con vs. ERA)
Serum RF IgA, median (IQR)	1.7 (1.7-1.7)	0.3 (0.3-5.2)	21 (1.4-35)	<0.01 (<0.01 ARI vs. ERA; 0.08 Con vs. ERA)

[1]*P* values were calculated using the Kruskal-Wallis test for continuous variables. For $P < 0.05$, a Dunn's post-hoc test with Bonferonni correction was performed and significant results are indicated in parentheses. *P* values were calculated using the Chi-squared test for discrete variables.

[2]Shared epitope present at any *HLA-DRB1 *01:01, *01:02, *04:01, *04:04, *04:05, *04:08, *04:09, *04:10, *04:13, *10:00* alleles. We were unable to measure 4 CON1 participants, and these were excluded from the percent calculation.

Supplementary Table S3: Longitudinal summary of ACPA+ ARI who progressed to clinical RA

	All participants (n = 16)			Female only (n = 13)		
Variable	**Baseline Visit**	**IA Onset**	***P* value[1]**	**Baseline Visit**	**IA Onset**	***P* value[1]**
Age at sample collection, mean years (SD)	51 (16)	52 (16)	<0.01	46 (14)	47 (14)	<0.01
Sex: Female, n (%)	13 (81%)	-	-	13 (100%)	-	-
Ethnicity: Non-Hispanic origin, n (%)	15 (94%)	-	-	12 (92%)	-	-
Race: White, n (%)	11 (69%)	-	-	9 (69%)	-	-
BMI, mean kg/m2 (SD)	28 (5)	28 (5)	0.49	27 (6)	28 (6)	0.07
Met ACR/EULAR 2010 criteria, n (%)	-	12 (75%)	-	-	10 (77%)	-
C-reactive protein, median mg/L (IQR)	2.8 (0.8-6.2)	2.3 (1.2-3.3)	0.98	3.3 (0.8-8.4)	2.9 (1.2-4.1)	1.00
Erythrocyte sedimentation rate, median mm/h (IQR)	12 (9-25)	14 (8-23)	0.71	12 (9-23)	13 (9-21)	0.67
Shared epitope present, n (%)[2]	6 (38%)	-	-	5 (38%)	-	-
Serum ACPA, median (IQR)	115 (49-1581)	171 (42-875)	0.11	133 (79-1557)	231(51-1242)	0.29
Serum RF IgM, median (IQR)	15 (0.3-34)	15 (3.2-35)	0.22	16 (0.3-30)	16 (3.8-50)	0.11
Serum RF IgA, median (IQR)	0.3 (0.3-5.6)	1.7 (0.3-2.7)	0.92	0.3 (0.3-5.2)	1.7 (0.3-4.1)	0.94
Average duration in the study, days (min, max)	-	467.9 (106-717)	-	-	497.8 (163-717)	-

[1]*P* values were calculated using the paired Wilcoxon rank-sum test for continuous variables. *P* values were calculated using the Fisher exact test for discrete variables.
[2]Shared epitope present at any *HLA-DRB1 *01:01, *01:02, *04:01, *04:04, *04:05, *04:08, *04:09, *04:10, *04:13, *10:00* alleles.

Supplementary Table S13: Summary of healthy controls (CON2) with longitudinal sampling

Variable	CON2 baseline visit (n=29)	CON2 last visit (n=29)	*P* value[1]
Age at sample collection, mean years (SD)	47.2 (14.2)	48.0 (14.2)	<0.001
Sex: Female, n (%)	22 (76%)	-	-
Ethnicity: Non-Hispanic origin, n (%)	28 (97%)	-	-
Race: White, n (%)	23 (79%)	-	-
BMI, mean kg/m2 (SD)	27.4 (5.0)	28.2 (4.4)	0.76
Met ACR/EULAR 2010 criteria, n (%)	0	0	-
C-reactive protein, median mg/L (IQR)	1.9 (1.1-3.0)	1.5 (0.7-2.0)	0.18
Erythrocyte sedimentation rate, median mm/h (IQR)	2 (2-9)	2 (2-9)	0.25
Serum ACPA, median (IQR)	0 (0 - 2)	-	-
Serum RF IgM, median (IQR)	0 (0 - 3.24)	-	-
Serum RF IgA, median (IQR)	0 (0 - 0)	-	-
Average duration in the study, Days (Min, Max)	-	354.7 (89 - 557)	-

[1]*P* values were calculated using the paired Wilcoxon rank-sum test for continuous variables. *P* values were calculated using the Fisher exact test for discrete variables.