Supplementary file 1.0: Additional methods

RNA extraction methods

As described in [1] RA-MAP blood was collected in Tempus Blood RNA Tubes (Applied Biosystems) and the RNA extracted using MagMAX RNA isolation kits (Ambion). This included removal of globin mRNA using GLOBINclear human 20 reaction kits (Ambion) according to the manufacturer's protocol. Additional blood for isolation of PBMCs was collected into EDTA Vacutainer collection tubes (BD) and separated using Leucosep separation tubes (Greiner). For subsequent microarray analysis, amplified RNA was hybridized to beadchips and scanned on an Illumina Beadstation 500. Illumina's GenomeStudio version 2011.1 with the Gene Expression Module v1.9.0 was used to generate signal intensity values. Non-normalised control and sample probe data were exported from GenomeStudio, background subtracted, and quantile normalised using Limma's neqc function [2]. The data were then filtered for probe signal intensity where probes were retained if they had a p value < 0.05 in at least 10% of samples. Probes were aggregated to genes using Limma's avereps function.

Cell specific RNA from the NEAC cohorts was extracted as described previously [3, 4]. In brief RNA was immediately extracted from total CD4+ T cells or B lymphocytes using an RNeasy Mini kit or AllPrep DNA/RNA Mini kit (both from Qiagen), and then subject to quality control using an Agilent 2100 Bioanalyzer (Agilent). The median RNA integrity number in the samples analyzed was 9.4. Complementary RNA generated from 250 ng total RNA (Illumina TotalPrep RNA Amplification kit) was hybridized to either an Illumina Whole Genome 6 version 3 (using CD4+ lymphocyte samples obtained prior to 2012) or a 12HT BeadChip (using CD4+ T cell samples obtained in or after 2012, and all B cell samples) (both from Illumina). For the subsequent methylation analyses four hundred nanogram of DNA was bisulphite-converted and DNAm quantified using the Infinium MethylationEPIC BeadChip (Illumina). After independent preprocessing and functional normalization of $CD4^+$ T- and B-cell data, probe filtering was performed and surrogate variable analysis used to estimate confounding variables (surrogate variable analysis package) as described in [4].

Expression and methylation analysis

Analysis of differential expression (DEG) and methylation (DMS) between IGS high and low eRA groups was performed using linear models using the limma R package. Differentially expressed genes (DEGs) were defined as genes with unadjusted p-value <0.05 and an absolute fold change >1.5. For methylation analysis, CpG sites exhibiting an absolute delta beta value of >0.1 and unadjusted p-value <0.05 were considered differentially methylated (DMSs). For both no multiple test correction was done for this as the basis for exploratory pathway analysis.

To assess the association between DNA methylation and gene expression, all CpGs remaining after pre-processing were first mapped to annotated genes based on the Illumina EPIC Human Methylation array annotation file, i.e. CpGs that were within a gene or its promoter region within 2kb. Only CpG-gene pairs containing a DMS or a DEG were analysed. Then, for each DEG or DMS, the paired gene expression and methylation profiles were tested for significant Pearson correlation coefficients (Benjamini-Hochberg, BH-adjusted p-value<0.05).

DMSs positions were overlapped with defined chromatin states corresponding to primary peripheral blood T helper (E043) and B cell (E032) epigenomes obtained from NIH Roadmap Epigenomics [5] using the LOLA R package [6]. The E043 and E032 15-state HMM models were downloaded from http://egg2.wustl.edu/roadmap and collapsed into five intuitive chromatin state annotations (Transcription start site, TSS flank, Enhancer, Transcribed and Repressed) as described in [4]. Transcription factor binding site (TFBS) enrichment analyses were performed using ENCODE phase 2 data generated using ChIPseq [7] and predicted TFBSs from the JASPAR database [8] (datasets downloaded from http://databio.org/regiondb). The 20bp region encompassing the DMSs were overlapped with TFBSs using the runLOLA function which applies Fisher's exact tests to calculate their significance (p-value < 0.05).

1. *The RA-MAP Consortium. RA-MAP, molecular immunological landscapes in early rheumatoid arthritis and healthy vaccine recipients* Sci Data, 2022. **In press**.
2. Smyth, G., *Limma: linear models for microarray data*, in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, R. Gentleman, Carey,V., Dudoit,S., Irizarry,R. and Huber,W. (eds), Editor. 2005, Springer: New York. p. 397–420.
3. Thalayasingam, N., et al., *CD4+ and B Lymphocyte Expression Quantitative Traits at Rheumatoid Arthritis Risk Loci in Patients With Untreated Early Arthritis: Implications for Causal Gene Identification.* Arthritis Rheumatol, 2018. **70**(3): p. 361-370.

4. Clark, A.D., et al., *Lymphocyte DNA methylation mediates genetic risk at shared immune-mediated disease loci.* J Allergy Clin Immunol, 2020. **145**(5): p. 1438-1451.
5. Roadmap Epigenomics, C., et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317-30.
6. Sheffield, N.C. and C. Bock, *LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor.* Bioinformatics, 2016. **32**(4): p. 587-9.
7. *A User's Guide to the Encyclopedia of DNA Elements (ENCODE)* in *https://doi.org/10.1371/journal.pbio.1001046*.
8. Fornes, O., et al., *JASPAR 2020: update of the open-access database of transcription factor binding profiles.* Nucleic Acids Res, 2020. **48**(D1): p. D87-D92.