Cross-Tissue Transcriptomic Analysis Leveraging Machine Learning Approaches Identifies New Biomarkers for Rheumatoid Arthritis

Dmitry Rychkov[1,2,3], Jessica Neely[3], Tomiko Oskotsky[1], Marina Sirota[1,3]

1. Bakar Computational Health Sciences Institute, UCSF, San Francisco, CA, USA; **2.** Department of Surgery, UCSF, San Francisco, CA, USA; **3.** Department of Pediatrics, UCSF, San Francisco, CA, USA

Abstract

Background/Purpose: There is an urgent need to identify effective biomarkers for early diagnosis of Rheumatoid Arthritis (RA) and accurate monitoring of disease activity. Here we define a RA meta-profile using publicly available cross-tissue gene expression data and apply machine learning to identify putative biomarkers, which we further validate on independent datasets.

Methods: We carried out a comprehensive search for publicly available microarray gene expression data in the NCBI Gene Expression Omnibus database for whole blood and synovial tissues from RA patients and healthy controls. The raw data from 13 synovium datasets with 284 samples and 14 blood datasets with 1,885 samples were downloaded and processed. The datasets for each tissue were merged and batch corrected and split into training and test sets. We then developed and applied a robust feature selection pipeline to identify genes dysregulated in both tissues and highly associated with RA. From the training data we identified a set of overlapping differentially expressed genes following the condition of co-directionality. The classification performance of each gene in the resulting set was evaluated on the testing sets using AUROC. Five independent datasets were used to validate and threshold the feature selected (FS) genes. Finally, we define the RAScore, composed of a geometric mean of the selected RAScore Panel genes and demonstrate its clinical utility.

Results: The result of the feature selection pipeline was a set of 25 upregulated and 28 downregulated genes. To assess the robustness of these feature selected genes, we trained a Random Forest machine learning model with this set of 53 genes and then with the set of 32 common differentially expressed genes and tested on the validation cohorts. The model with FS genes outperformed the model with common DE genes with AUC 0.89 ± 0.04 vs 0.86 ± 0.05. The FS genes were further thresholded on the 5 independent datasets resulting in 10 upregulated genes, *TNFAIP6, S100A8, TNFSF10, DRAM1, LY96, QPCT, KYNU, ENTPD1, CLIC1, ATP6V0E1*, that are involved in innate immune system pathways, including neutrophil degranulation and apoptosis and expressed in granulocytes, dendritic cells, and macrophages; and 3 downregulated genes, *HSP90AB1, NCL, CIRBP*, involved in metabolic processes and T-cell receptor regulation of apoptosis and expressed in lymphoblasts.

To investigate the clinical utility of the 13 validated genes, the RA Score was developed and found to be highly correlated with DAS28 ($r = 0.33 \pm 0.03$, $p = 7e-9$) and able to distinguish OA and RA samples (OR 0.57, 95% CI [0.34, 0.80], $p = 8e-10$). Moreover, the RA Scores were not significantly different for RF-positive and RF-negative RA sub-phenotypes ($p = 0.9$) suggesting the generalizability of this score in clinical applications. The RA Score was also able to monitor the treatment effect among RA patients (t-test of treated vs untreated, $p = 2e-4$) and distinguish polyJIA from healthy individuals in 10 independent pediatric cohorts (OR 1.15, 95% CI [1.01, 1.3], $p = 2e-4$).

Conclusion: The RAScore, consisting of 13 putative biomarkers, identified through a robust feature selection procedure on public data and validated using multiple independent data sets may be useful in the diagnosis and treatment monitoring of RA.

Introduction

Rheumatoid arthritis (RA) is a systemic inflammatory condition characterized by a symmetric and destructive distal polyarthritis. Undiagnosed and untreated, RA can progress to severe joint damage, involve other organ systems, and predispose individuals to cardiovascular disease (Battisha et al., 2020; Lai et al., 2020). While our understanding of disease pathogenesis has greatly improved, and the number of available, effective therapeutics has significantly increased, there remains significant barriers to caring for patients with RA, and they continue to suffer from the morbidity and mortality associated with the disease. There is an urgent need to develop objective biomarkers for the early diagnosis and prompt initiation of disease-modifying therapy during the so-called "window of opportunity" (Bergstra et al., 2020; Burgers et al., 2019; Coffey et al., 2019; Van Nies et al., 2015). Additionally, clinicians need tests to help accurately assess disease activity or treatment targets in order to adjust therapy appropriately. Identification of biomarkers would greatly add to clinicians' existing toolset used to evaluate patients with RA helping to improve outcomes and alleviate the suffering caused by this prevalent disease.

Over the past decade, advances in genomic sequencing technology have greatly contributed to our understanding of inflammatory diseases and informed development of effective therapeutics. Transcriptomics provides a lens into the specific genes over- or under-expressed in a disease providing insight into cellular responses. Given the numerous transcriptomic datasets that have been generated and made publicly available, there are now opportunities to combine these datasets in a meta-analytic fashion for unbiased computational biomarker discovery. Meta-analysis is a systematic approach to combine and integrate cohorts to study a disease condition which provides enhanced statistical power due to a higher number of samples when combined. Additionally, it provides an opportunity of leveraging all the disease heterogeneity combined from multiple smaller studies across diverse populations which allows creating a robust signature and better recognizing direct disease drivers as well as disease subtyping and patient stratification. Moreover, integrating datasets generated from the multiple target tissues within a given disease further strengthens the associations identified (Barbeira et al., 2019; Gomez-Cabrero et al., 2014; Haynes et al., 2020; Li et al., 2019; Lofgren et al., 2016; Pineda et al., 2020; L. Wang et al., 2015a). This approach has been successfully applied to the study of antineutrophil cytoplasmic antibody (ANCA)-associated vasculitis (Friedman et al., 2019), dermatomyositis (Neely et al., 2019) and systemic lupus erythematosus (Haynes et al., 2020). These large datasets also present an opportunity to apply advanced machine learning techniques that were not previously feasible computationally allowing for interrogation of the data with new and unbiased approaches.

Multiple studies attempted to identify an RA transcriptomics signature in blood (Batliwalla et al., 2005; Sumitomo et al., 2018; Teixeira et al., 2009a, Wang et al., 2015a) and in synovial tissue (Asif Amin et al., 2017; Dana E Orange et al., 2018; Dana E. Orange et al., 2018) separately or in a cross-tissue analysis (Afroz et al., 2017; Li et al., 2014). The tissue-specific studies have found very few overlapping signals. The integrative meta-analysis studies combined a few datasets from each tissue (Afroz et al., 2017; Li et al., 2014) to identify an overlap of dysregulated genes and to recognize similarities and differences in disease pathways in both tissues. While this type of approach allows better understanding of the disease, a corresponding set of biomarkers is often redundant and requires extensive prioritization analysis and validation. Thus, more rigorous approaches for biomarker search with a built-in prioritization procedure are needed.

In this study, we leveraged publicly available transcriptomic datasets generated from microarray and RNA sequencing (RNAseq) platforms from over 2,000 samples from whole blood and synovial tissue of patients with RA. After combining these datasets in using a well-described meta-analytic pipeline (Hughey and Butte, 2015) and describing the expression pathways and cell types present in RA tissues, we developed and applied a robust machine learning and feature selection approach to identify unique and independent biomarkers which were subsequently refined and validated on test data. We then evaluated the diagnostic utility of this set of biomarkers and the correlation with disease activity measures to inform future clinical studies. The development of an

effective blood test for the diagnosis and monitoring of RA can add valuable information to the physician's assessment and help inform decision-making to improve the morbidity and quality of life for patients with RA.

Methods

Discovery data collection and processing

We carried out a comprehensive search for publicly available microarray data in the NCBI Gene Expression Omnibus (Barrett et al., 2013) (GEO) database (http://www.ncbi.nlm.nih.gov/geo/) for whole blood and synovial tissues in Rheumatoid Arthritis and healthy controls using the keywords "rheumatoid arthritis", "synovium", "synovial", "biopsy", and "whole blood", among organisms "Homo Sapiens" and study type "Expression profiling by array" (**Figure 1a**). Datasets were excluded when samples were poorly annotated or run on platforms with few numbers of probes. This search yielded 13 synovial datasets, which included 257 biopsy samples from subjects with RA and 27 from healthy controls obtained during joint or trauma surgeries (**Supplementary Table 1**). We identified 14 whole blood datasets with 1,885 samples: 1,470 RA patients and 415 healthy controls (**Supplementary Table 1**).

Raw data was downloaded and processed using R language version 3.6.5 (R core team, 2018) and the Bioconductor (Huber et al., 2015) packages *SCAN.UPC* (Piccolo et al., 2012), *affy* (Gautier et al., 2004) and *limma* (Ritchie et al., 2015). Processing steps included background correction, log2-transformation, and intra-study quantile normalization (**Figure 1a**). Next, we performed probe-gene mapping, data merging and normalization across batches with Combat within the R package *sva* (Leek et al., 2012). The dimensionality reduction plots before and after normalization are shown in **Supplementary Figure 1**. After merging studies, the total number of common genes was 11,057 in synovium and 14,596 in whole blood.

Validation data collection and processing

Five additional datasets from GEO were identified and downloaded: synovium microarray and RNA-seq, PBMC microarray and RNA-seq, whole blood microarray datasets (**Supplementary Table 1**). Microarray data was processed as described above. RNA-seq data from GSE89408 were downloaded in a form of processed data of feature counts, which were normalized using the variance stabilizing transformation function *vst()* from the R package *DESeq2* (Love et al., 2014). RNA-seq data from GSE90081 were downloaded in a processed form of Fragments Per Kilobase Million (FPKM) counts, which were converted to Transcripts Per Kilobase Million (TPM) counts followed by log2 transformation with 0.1 offset.

Differential gene expression and pathway analysis

Differentially expressed genes were identified using a linear model from the R package *limma* (Ritchie et al., 2015). To account for factors related to gene expression, the imputed sex and treatment categories were used as covariates. Treatment types were categorized based on the drug class (**Supplementary Table 2)**. For 877 (40%) samples without sex annotations, sex was imputed using the average expression of Y chromosome genes. Significance for differential expression was defined using the cutoff of FDR p-value < 0.05 and abs(FC) > 1.2. Pathway analysis of differentially expressed genes was performed using the package *clusterProfiler* (Yu et al., 2012) with the Reactome database as well as the gene list enrichment analysis tool ToppGene (https://toppgene.cchmc.org/). To assess statistical significance of gene overlaps we computed p-values using the hypergeometric test.

Cell type enrichment analysis

In order to estimate the presence of certain cell types in a tissue, we leveraged the cell type enrichment analysis tool, xCell (Aran et al., 2017) which computes enrichment scores for 64 immune and stromal cells based on gene expression data. We limited our analysis to 53 types of stromal, hematopoietic, and immune cells we expected to be present in blood and synovium. The cell types with a detection p-value greater than 0.2 taken as a medium across all samples in a tissue were filtered. Non-parametric Wilcoxon-Mann test with multiple testing correction

with Benjamini-Hochberg approach (cut-off 0.05) was used to assess significantly enriched cell types in synovium and whole blood in RA compared to healthy control subjects. The effect size of each cell type was estimated by computing the ratio of the mean enrichment score in RA patients over mean score in healthy individuals.

Feature selection pipeline

The feature selection procedure was partially described by Perez-Riverol, et al. (Perez-Riverol et al., 2017) and is represented in **Figure 1b**. First, for each tissue, the data were split into training and testing sets in an 80:20 ratio with random sample selection and class distribution preservation using the function *createDataPartition()* from the R package *caret* (Kuhn, 2008). Within each training set, a set of significant genes was identified using *limma* FDR p-value < 0.05. Pearson correlation coefficient was computed with the case-control status for each significant gene and those with r < 0.25 were filtered out. For robustness and reducing gene redundancy, we computed gene pair-wise correlations and removed genes with correlation greater than 0.8. Next, we overlapped the gene sets from both tissues and filtered out any genes differentially expressed in opposite directions in synovium and blood. To monitor statistical significance of gene overlaps we computed p-values using the hypergeometric test. To evaluate each gene's performance in distinguishing RA from healthy samples, we trained a logistic regression model per gene on a training set for each tissue and tested on a testing set using area under receiver operating characteristic (AUROC) curve as a performance measure.

We repeated these steps 100 times to minimize bias of a random split into training and testing sets. From the resulting 100 gene sets, any gene that was found in each set was carried to the further analysis. The AUC performance of each gene was averaged, and its standard deviation was calculated. We then set the AUC threshold to 2/3 and applied this criterion to the testing results to identify the genes with the best performance, the feature selected genes.

Feature validation and RA Score

We used the five independent validation datasets to evaluate the feature selected genes. To evaluate and compare the value of the feature selected genes and the common DE genes in diagnostics, we proceeded with training machine learning models on the discovery blood data with these two gene sets and testing them on 5 validation sets. As some genes were not present in all validation sets, we reduced the gene sets to the genes that were found in all 5 validation sets. We used three machine learning algorithms, Logistic Regression, Elastic Net and Random Forest, for training classification models and AUROC for measuring their performance. We trained an ML model for each feature selected gene individually on the discovery blood data and tested on the validation sets. AUROC was used as a performance measure. The genes with average AUC greater than 0.8 were selected. The selected genes were used to create the RA Score, computed by subtracting the geometric mean expression of the down-regulated genes from the geometric mean expression of the up-regulated genes.

Next, to assess the clinical value of the selected genes and the RA Score, we identified datasets with samples that included values for DAS28, a measure of disease activity in RA. We computed the Pearson correlation coefficients of the RA Score and expression levels of the RA Score genes with DAS28. Eight datasets with both RA and Osteoarthritis (OA) samples (**Supplementary Table 1**) were used to evaluate the ability of the RA Score to distinguish RA from OA. GSE74143 was used to test the difference in RAScore between RA sub-phenotypes with positive and negative Rheumatoid Factors by applying Student's t-test. GSE45876 and GSE93272 were used to test the RAScore difference between treated and untreated RA patients via Student's t-test. Additionally, we leveraged 10 datasets to test the ability of the RA Score to recognize polyJIA by calculating the Odds Ratio (**Supplementary Table 1**).

Results

Cross-Tissue Differential Expression and Pathway Analysis Reveals Significant Similarities on Gene and Pathway Levels

The differential gene expression analysis identified 1,370 genes with 771 up-and 599 down-regulated genes in the synovium (**Supplementary Figure 2ab** and **Supplementary Table 3**) and 155 genes with 110 up- regulated and 45 down-regulated genes in the blood (**Supplementary Figure 3ab** and **Supplementary Table 4**). The pathway analysis revealed that in both tissues up-regulated genes shared enrichments in neutrophil degranulation, interferon alpha/beta signaling, toll-like receptor (TLR) cascades, regulation of TLR by endogenous ligand, and caspase activation via extrinsic apoptotic signaling pathways (**Figure 2a, Supplementary Figures 2cd, 3cd, Supplementary Table 5**), while interferon gamma signaling, MHC class II antigen presentation, TCR signaling were specific for synovium and apoptosis, programmed cell death, antiviral mechanisms, DDX58/IFIH1-mediated induction of interferon-alpha/beta pathways were specific for blood (**Supplementary Table 6**). The down-regulated genes were commonly involved only in signaling by interleukins pathway (**Figure 2b, Supplementary Figures 2ef, 3ef**). However, the signaling by interleukins was also a common pathway with up-regulated genes in synovium coupled with enrichment in interleukin-5, interleukin-13 and GM-CSF signaling pathways. Many pathways were not shared suggesting different molecular mechanisms underlying in tissues. For example, the interleukin-4, interleukin-13 signaling, muscle contraction, FOXO-mediated transcription, and ESR-mediated signaling pathways were specific only for synovium (**Supplementary Tables 5,6**).

When evaluating the overlap between differentially expressed genes in synovium and blood, there were 28 genes commonly up-regulated: *TNFAIP6, S100A8, MMP9, S100A9, IFI27, EVI2A, NMI, BCL2A1, TNFSF10, LY96, SAMSN1, GPR65, DDX60, ISG15, MX1, OAS1, IFI44, ENTPD1, IFIT3, CSTA, CLIC1, IFIT1, DOCK4, NAT1, FAS, C1GALT1C1, CD58, COMMD8*; and 4 down-regulated genes: *S1PR1, TUBB2A, ABLIM1, MYC* (**Figure 2c, Supplementary Table 7,** hypergeometric test p = 9e-9). However, the overlap of down-regulated genes did not meet statistical significance (hypergeometric test p = 0.28, **Figure 2d**). The common differentially expressed (DE) genes formed more distinct clusters of RA and control samples for both synovium (**Figure 2ef**) and blood (**Figure 2gh**) than all DE genes for these tissues (**Supplementary Figure 2ab** and **Supplementary Figure 3ab**). The Gene Ontology biological processes of these common up-regulated genes included innate immune and defense response, neutrophil degranulation and type I interferon signaling pathways, whereas down-regulated genes are associated with PDGFR-beta signaling and Interleukin-4 and 13 signaling pathways. Interestingly, the genes involved in interferon pathways showed the negative correlation between tissues (r = -0.78, 95% CI [-0.97, -0.07], p = 0.04), whereas the genes involved in cell activation and neutrophil degranulation pathways correlated positively: r = 0.7 (p = 0.03) and rho = 0.8 (p = 0.1), respectively**.**

Cell-Type Deconvolution Analysis Identifies a Reverse Signal in Blood and Synovium

The cell type enrichment analysis with xCell in synovium revealed the enrichment of immune cell types, including, CD4+ and CD8+ T-cells, B-cells, macrophages and dendritic cells in RA samples (**Figure 3a**). However, opposite associations were seen in whole blood samples with enrichment of T- and B-cells in healthy controls (**Figure 3b**). Concordance in activation of innate immune cells and opposition in activation of lymphocytes in tissues from discovery cohorts (**Figure 3c**) were confirmed with validation datasets (**Figure 3d**). The significant cell types in synovium and blood showed high correlations in validation data: r = 0.71 (p = 1.3e-5) for synovium (**Figure 3e**) and r = 0.61 (p = 0.004) in blood (**Figure 3f**).

Machine Learning Feature Selection Strategy to Identify Robust Cross-Tissue Biomarkers of RA

Aiming to determine a more robust list of putative biomarkers that are strongly associated with RA in both synovium and whole blood tissues and have higher predictive power, we applied an iterative feature selection procedure leveraging the gene expression data from both tissues. In the pipeline, only 10,071 genes that were common between synovium and whole blood data were used. At each iteration, only genes found significantly dysregulated in both tissues following the condition of co-directionality were kept (p = 6.3e-10). As a result of these filtering steps, on average 65 up-regulated and 71 down-regulated were selected from each iteration (See

Methods). From 100 iterations, any gene significantly dysregulated in all the iterations was selected, resulting in a set of 53 genes: 25 upregulated and 28 down-regulated (**Supplementary Table 8**). A summary of the average AUC performance from the 100 iterations for each gene are shown in the **Figure 4a** and **Supplementary Table 8**. The AUC for selected genes in synovial tissue varied with mean 0.853 ± 0.005 for training and 0.866 ± 0.006 for testing sets, whereas for the whole blood the mean AUC was 0.744 ± 0.006 for both training and testing sets.

For validation purposes, we leveraged 5 publicly available independent datasets on synovium and blood (see *Methods*) (**Supplementary Table 1**). To compare classification performance of the set of 53 feature selected genes to the set of 32 common DE genes, we trained three ML models on the discovery blood data and tested them on the five validation datasets. However, since not all genes were measured across the validation studies, the sets were reduced to 25 of 32 common DE genes and 38 of 53 feature selected genes. We found the set of 38 feature selected genes has superior performance over the set of 25 common DE genes for all three ML methods (**Supplementary Figure 4**). The largest difference in performance was for the Random Forest model: the model with the common DE genes had an AUC of 0.856 ± 0.046 (95% CI [0.775, 0.937]) (**Figure 4b**), while the model with the feature selected genes performed with 0.889 ± 0.044 (95% CI [0.811, 0.966]) (**Figure 4c**).

Next, the set of 53 feature selected genes was thresholded to an average AUC of 0.8 using the validation sets resulting in the set of 10 up-regulated *TNFAIP6, S100A8, TNFSF10, DRAM1, LY96, QPCT, KYNU, ENTPD1, CLIC1, ATP6V0E1*, and 3 down-regulated *HSP90AB1, NCL, CIRBP* genes (**Figure 4a, Supplementary Figure 5**)**.** Five of the related proteins, *TSG-6, MRP-8/Calgranulin-A, TNFSF10/TRAIL, Ly-96, QC,* are known to be normally secreted into blood, while *Kynureninase*, *HSP 90-beta* and *CLIC1* are localized to cytosol (Uhlén et al., 2019).

Clinical Implications of Transcription Based Disease Score

In order to assess the clinical utility of the 13 validated genes, we introduced a scoring function, RA Score, which is derived by subtracting the geometric mean of expression values of down-regulated genes from the geometric mean of up-regulated genes. With this definition, the RAScore is 2-fold (95% CI [1.8, 2.2], p = 3e-15) larger for RA in comparison to healthy samples in synovium. In whole blood, the RA Score has a mean effect size of 1.37 (95% CI [1.34, 1.4], p = 1e-108). On the validation synovium data, the RAScore had a mean effect size 5.5 (95% CI [3.8, 8.2] , p = 1e-10) and 2.4 (95% CI [2.1, 2.8], p = 3e-23) on the validation blood data (i.e. whole blood and PBMC).

We identified 4 datasets with 411 samples with available disease activity score (DAS28) annotations. To determine if the RA Score Panel genes were associated with DAS28, and thus potentially useful as disease activity biomarkers, we assessed the correlation of the expression value of each gene with the DAS28 score. The RA score was overall positively correlated with DAS28 with the most correlated gene being *S100A8* with mean R = 0.28 (95% CI [0.19, 0.37]) and most anti-correlated gene *HSP90AB1* with mean r = -0.23 (95% CI [-0.32, -0.14]) (**Figure 5b**, **Supplementary Figure 6).** We also determined the correlation of the RA Score with DAS28 in these datasets and obtained Pearson correlation coefficient from 0.25 to 0.43 in blood and 0.31 in synovium (**Supplementary Figure 7**). The average correlation was 0.33 with 95% CI [0.24, 0.41] (**Figure 5a**).

To investigate the ability of the RAScore to differentiate RA from osteoarthritis (OA), we identified eight datasets that had both RA and OA samples available. The **Figure 5e** shows the distributions of RAScore for RA, OA and healthy samples in eight available datasets. In most datasets, the RA Score was able to significantly differentiate OA from RA (OR 0.57, 95% CI [0.34, 0.80], p = 8e-10) and healthy samples (OR 1.53, 95%CI [1.37, 1.69], p = 7.5e-4) implicating that this score may be useful diagnostically.

The RA Score performed similarly in both RF-positive and RF-negative rheumatoid arthritis samples in the whole blood dataset GSE74143 suggesting the applications of this score are generalizable to these RA subtypes (t-test, p = 0.9) (**Figure 5c**). Furthermore, we tested the utility of this score in datasets from polyarticular juvenile idiopathic arthritis (polyJIA) samples given that this subtype of JIA is most similar to RA, and also found comparable performance in the ability to differentiate polyJIA from healthy controls (OR 1.15, 95% CI [1.01, 1.3], p = 2e-4) (**Figure 5f**). Thus, this score may also be useful in the pediatric arthritis population.

Lastly, it appears the RA score also tracks with treatment response. In two datasets, RA patients had transcriptional measurements before and after treatment with DMARD. The RA score significantly (t-test, p = 2e-4) decreases between pre- and post-treatment measurements (**Figure 5d**).

Discussion

In this study, we leveraged publicly available microarray gene expression data from both synovium and peripheral blood tissues in search of putative biomarkers for RA. We first applied a conventional approach (Dudley et al., 2009; Hao et al., 2017; Shchetynsky et al., 2017; L. Wang et al., 2015b; Wu et al., 2010) of intersecting the differentially expressed genes from both tissues and obtained a list of 32 common genes. Some of our results showed that agreement with previous findings (Afroz et al., 2017; Teixeira et al., 2009b; L. Wang et al., 2015b) with the involvement into similar biological processes (e.g. cytokine signaling, immune and defense response, response to biotic stimulus), that were found and described before (Afroz et al., 2017; Teixeira et al., 2009b; L. Wang et al., 2015b). The common DE genes having a higher expression in both tissues formed denser and more distinct clusters of both RA and control samples in synovium (**Figure 2ef**) and blood (**Figure 2gh**) than all DE genes (**Supplementary Figure 2ab** and **Supplementary Figure 3ab**). While this list of overlapping genes provides valuable insight into disease biology, the predictive ability can be further improved by applying more advanced methods to prioritize the candidates and remove redundancy.

In order to identify a robust and non-redundant set of biomarkers, we developed a specific ML based feature selection pipeline that leveraged the concordantly expressed genes data for both tissues. This resulted in 53 protein coding genes that outperformed 32 common genes in outcome prediction tasks on independent data. In further validation steps, we identified and selected 10 up- and 3 down-regulated genes with the highest performance. The up-regulated genes are highly expressed in diseased synovial tissue, and their elevated protein levels in blood can be the direct markers for RA disease. We went further in combining the 13 validated genes into a transcriptional gene score, the RA Score, that potentially could serve as a clinical tool in a blood test for accurate disease diagnosis and monitoring disease progression and treatment. The RA Score was able to significantly discriminate RA from OA as well as polyJIA. Some genes/proteins, e.g. S100A8/MRP8, ENTPD1/CD39, KYNU, TNFAIP6, from the score were previously found to be associated with JIA (Bojko, 2017; Brachat et al., 2017; Griffin et al., 2009; Hinze et al., 2019; Holzinger et al., 2019; Jiang et al., 2013; Korte-Bouws et al., 2019; Moncrieffe et al., 2010). The RA Score did not differentiate between RF+ and RF - sub-types of RA suggesting the generalizability of this metric. The effect of the treatment was also captured with a significantly lower RA Score for DMARD treated patients in comparison to treatment-naïve ones. Although the genes were identified from not just blood but also synovium, we intentionally selected differentially expressed genes that followed the condition of co-directionality, i.e. upregulated in both synovium and blood or downregulated in both tissues thereby making a RA Score test based on blood only tenable.

The 13 genes identified using these machine learning methods in the feature selection pipeline represent candidate biomarkers in RA. Six of the 13 RAScore Panel genes (TNFAIP6, S100A8, TNFSF10, LY96, ENTPD1, and CLIC1) were also among the 32 Common DE Genes, whereas seven of the 13 RAScore Panel genes (DRAM1, QPCT, KYNU, ATP6V0E1, NCL, CIRBP and HSP90AB1) were not and found only in the RA Score set of genes.

Among the RA Score Panel genes that were also in the set of 32 Common DE Genes, a few have been associated with RA. ***TNFAIP6***, also known as TSG-6, encodes a secretory protein that is not a constituent of healthy adult tissues but produced in response to inflammatory mediators, with high levels detected in the synovial fluid of patients with rheumatoid arthritis and osteoarthritis (Wisniewski et al., 1993). TNFAIP6 is thought to affect the destruction of inflammatory tissue through its role in extracellular matrix remodeling (Milner and Day, 2003). ***S100A8*** along with S100A9 (also known as MRP8 and MRP14, respectively) encodes Calcium binding proteins that often exist in heterodimer form as Calprotectin (S100A8/A9) and are constituents of circulating neutrophils and

monocytes and in macrophages in RA synovial tissue (Nordal et al., 2016). S100A8, S100A9, and S100A8/A9 participate in modulating and restoring homeostasis during inflammation, although their overexpression and release may result in an imbalance of inflammatory processes (Wang et al., 2018). A study comparing markers of inflammation in patients with RA found that Calprotectin had the highest association with ultrasound-detected synovitis and predicted treatment response (Nordal et al., 2016). Of note, while the 32 Common DE Genes set includes S100A9, this gene is not one of the 53 FS genes or subsequent 13 RAScore Panel genes possibly since our process excludes genes that are highly correlated (pair-wise feature correlation > 0.8). ***TNFSF10***, also known as TNF-Related Apoptosis Inducing Ligand (TRAIL), encodes a protein that induces apoptosis of tumor cells. While TNSF10 has been considered for controlling the synovial hyperplasia and immune cell hyperactivity in RA, there is also evidence that TNSF10 may promote joint destruction and exacerbate RA (Audo et al., 2013). ***LY96***, also known as MD2, encodes a protein which often is a coreceptor with TLR4 forming the TLR4/MD2 complex (Nagai et al., 2002) and is a potential marker of rheumatic diseases such as rheumatoid arthritis (Teixeira et al., 2009b; L. Wang et al., 2015b). ***KYNU*** encodes the Kynureninase enzyme which is involved in a pathway of tryptophan metabolism related to immunomodulation and inflammation (Harden et al., 2016; Q. Wang et al., 2015). KYNU expression has been found to be increased in chondrocytes and synovial tissue of RA patients compared to healthy patients (Chen et al., 2018). Moreover, Increased tryptophan degradation has been observed in the blood of RA patients (Schroecksnadel et al., 2003). ***ENTPD1***: also known as CD39, a gene found to be an expression quantitative trait locus associated with levels of CD39+, CD4+ regulatory T-cells. In RA, high levels of CD39+ expressing Tregs were associated with methotrexate response suggesting this could be a biomarker to predict responders and non-responders. Overexpression of CD39 has also been explored as a possible therapeutic modality in RA (Spiliopoulou et al., 2019).

Interestingly, seven of the 13 RAScore Panel genes (DRAM1, QPCT, KYNU, ATP6V0E1, NCL, CIRBP and HSP90AB1) were not found within the 32 Common DE Genes set, and a few of these genes unique to the RA Score Panel have been associated with RA. The gene **QPCT** encodes Glutaminyl-Peptide Cyclotransferase, or QC for short (NCBI, 2020a), which plays a role in maintaining inflammation (Kehlen et al., 2017). A clinical study found that expression of QPCT was significantly increased in the blood and in the fluid from gingival crevices of RA patients compared to healthy controls (Bender et al., 2019). **CIRBP** encodes cold-inducible RNA-binding protein which can be induced under stress and have a cytoprotective role but has also been increasingly recognized for participating in proinflammatory response (Liao et al., 2017). CIRBP binds to the TLR4/MD2 complex of macrophages and monocytes in the circulation or tissues, thereby activating the NF-Kappa B pathway and resulting in the release of inflammatory mediators (Liao et al., 2017; Qiang et al., 2013). A clinical study measured CIRBP mRNA expression of CD14+ monocytes of five healthy and five RA patients and found the relative expression of CIRBP was higher in RA patients (Yoo et al., 2018). Our analysis found down-regulated CIRBP expression in whole blood from RA compared to healthy control, but no statistically significant difference in CIRBP expression in white blood cells ($p = 0.18$, GSE117769). Further study of CIRBP in RA patients, particularly single cell analysis, is warranted. **HSP90AB1** encodes the protein HSP90β (NCBI, 2020b). Post-translationally citrullinated isoforms of heat shock protein 90, including citHSP90β, have been identified as potential autoantigens in patients with RA-associated interstitial lung disease (Harlow et al., 2014; Travers et al., 2016). While there is evidence to support a role in RA for a number of these genes, others such as DRAM1, ATP6V0E1, and NCL have not been previously reported to our knowledge, though DRAM1 and NCL were shown to be associated with other autoimmune diseases, such as Systemic Lupus Erythematosus (Molineros et al., 2013; Yang et al., 2013) and Multiple Sclerosis (Liu et al., 2012).

Several limitations of this study should be recognized. The data was collected from the public repository NCBI GEO where often the case-control ratio was highly imbalanced up to a full absence of healthy controls especially in whole blood. We separately collected two datasets of healthy individuals to enrich the blood data with the control class. All sample annotations were kept from the original publications, though for 40% of samples the sex annotations were not available, and they were imputed based on the expression levels of Y chromosome genes. Another limitation to the study results were the limited availability of validation cohorts that would have a fair case-control balance. Out of three validation blood datasets, two were from PBMC in contrast to the whole blood discovery data. This could possibly lead to lower AUC performance for genes in the validation datasets. Additionally, most case samples were from RA patients who were taking various medications. Even though the

treatments were used in the differential expression analysis as covariates (including untreated patients) there still exists the possibility of their impact on the results.

In this study we presented a robust machine learning based pipeline for search for putative biomarkers: each gene went individually through a feature selection procedure with multiple iterations on the discovery data and was independently tested on the validation cohorts. The gene redundancy was decreased selecting the best performing genes in RA association prediction. We apply the pipeline to a set of over 2000 samples to identify the RA Score as a potential diagnostic panel and show its clinical utility. The strength of RAScore is in the independence of its composing genes. Further development of the RA Score as a clinical tool requires greater understanding and validation of its composing genes with experimental analysis of the protein levels in RA patients and healthy individuals. An independent longitudinal study would bring better understanding of the early diagnostic and disease monitoring capability of the proposed panel. Additional experiments leveraging single cell technologies will further enhance our understanding of the biology and cell type specific effects in RA allowing us to refine the proposed diagnostic strategies.

Acknowledgements

This work was in part supported by Pfizer ASPIRE Grant WI215080 and P30 AR070155. We would like to thank Dr. Boris Oskotsky for technical support and the members of the Sirota Lab for useful discussion. We would like to thank Cindy Lin for her initial work aggregating gene expression datasets and carrying out preliminary analysis of a subset of data and Dr. Judith Ashouri for useful discussions.

References

Afroz S, Giddaluru J, Vishwakarma S, Naz S, Khan AA, Khan N. 2017. A Comprehensive Gene Expression Meta-analysis Identifies Novel Immune Signatures in Rheumatoid Arthritis Patients. *Frontiers in Immunology* **8**:74–74. doi:10/gg277s

Aran D, Hu Z, Butte AJ. 2017. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biology* **18**:220. doi:10/gckmjs

Asif Amin M, Fox DA, Ruth JH. 2017. Synovial cellular and molecular markers in rheumatoid arthritis. *Seminars in Immunopathology* **39**:385–393. doi:10.1007/s00281-017-0631-3

Audo R, Combe B, Hahne M, Morel J. 2013. The two directions of TNF-related apoptosis-inducing ligand in rheumatoid arthritis. *Cytokine* **63**:81–90. doi:10/f44tmt

Barbeira AN, Pividori MD, Zheng J, Wheeler HE, Nicolae DL, Im HK. 2019. Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genetics* 1–20. doi:10.1371/journal.pgen.1007889

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. 2013. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* **41**:D991–D995. doi:10/f3mn62

Batliwalla FM, Baechler EC, Xiao X, Li W, Balasubramanian S, Khalili H, Damle A, Ortmann WA, Perrone A, Kantor AB, Gulko PS, Kern M, Furie R, Behrens TW, Gregersen PK. 2005. Peripheral blood gene expression profiling in rheumatoid arthritis. *Genes and Immunity* **6**:388–397. doi:10.1038/sj.gene.6364209

Battisha A, Sawalha K, Altibi AM, Madoukh B, Al-Akchar M, Patel B. 2020. Cardiogenic shock in autoimmune rheumatologic diseases: an insight on etiologies, management, and treatment outcomes. *Heart Fail Rev*. doi:10/gg3dh9

Bender P, Egger A, Westermann M, Taudte N, Sculean A, Potempa J, Möller B, Buchholz M, Eick S. 2019. Expression of human and Porphyromonas gingivalis glutaminyl cyclases in periodontitis and rheumatoid arthritis–A pilot study. *Archives of Oral Biology* **97**:223–230. doi:10/gg3dkd

Bergstra SA, Van Der Pol JA, Riyazi N, Goekoop-Ruiterman YPM, Kerstens PJSM, Lems W, Huizinga TWJ, Allaart CF. 2020. Earlier is better when treating rheumatoid arthritis: but can we detect a window of opportunity? *RMD Open* **6**. doi:10.1136/rmdopen-2020-001242

Bojko J. 2017. Measurement of blood calprotectin (MRP-8/MRP-14) levels in patients with juvenile idiopathic arthritis. *Reumatologia* **55**:10–14. doi:10/gg277m

Brachat AH, Grom AA, Wulffraat N, Brunner HI, Quartier P, Brik R, McCann L, Ozdogan H, Rutkowska-Sak L, Schneider R, Gerloni V, Harel L, Terreri M, Houghton K, Joos R, Kingsbury D, Lopez-Benitez JM, Bek S, Schumacher M, Valentin M-A, Gram H, Abrams K, Martini A, Lovell DJ, Nirmala NR, Ruperto N, for the Pediatric Rheumatology International Trials Organization (PRINTO) and the Pediatric Rheumatology Collaborative Study Group (PRCSG). 2017. Early changes in gene expression and inflammatory proteins in systemic juvenile idiopathic arthritis patients on canakinumab therapy. *Arthritis Research & Therapy* **19**:13. doi:10/gg4jfc

Burgers LE, Raza K, van der Helm-van Mil AH. 2019. Window of opportunity in rheumatoid arthritis - definitions and supporting evidence: from old to new perspectives. *RMD Open* **5**:e000870. doi:10/gg3djt

Chen Y-J, Chang W-A, Wu L-Y, Hsu Y-L, Chen C-H, Kuo P-L. 2018. Systematic Analysis of Differential Expression Profile in Rheumatoid Arthritis Chondrocytes Using Next-Generation Sequencing and Bioinformatics Approaches. *Int J Med Sci* **15**:1129–1142. doi:10/gd39zb

Coffey CM, Crowson CS, Myasoedova E, Matteson EL, Davis JM. 2019. Evidence of Diagnostic and Treatment Delay in Seronegative Rheumatoid Arthritis: Missing the Window of Opportunity. *Mayo Clin Proc* **94**:2241–2248. doi:10/gg3djs

Dudley JT, Tibshirani R, Deshpande T, Butte AJ. 2009. Disease signatures are robust across tissues and experiments. *Molecular Systems Biology* **5**:307–307. doi:10/cmf4mg

Friedman MA, Choi D, Planck SR, Rosenbaum JT, Sibley CH. 2019. Gene Expression Pathways across Multiple Tissues in Antineutrophil Cytoplasmic Antibody-associated Vasculitis Reveal Core Pathways of Disease Pathology. *The Journal of rheumatology* jrheum.180455. doi:10.3899/jrheum.180455

Gautier L, Cope L, Bolstad BM, Irizarry RA. 2004. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**:307–315. doi:10/bwk7sk

Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, Gisel A, Ballestar E, Bongcam-

Rudloff E, Conesa A, Tegnér J. 2014. Data integration in the era of omics: current and future challenges. *BMC Systems Biology* **8**:I1. doi:10.1186/1752-0509-8-S2-I1

Griffin TA, Barnes MG, Ilowite NT, Olson JC, Sherry DD, Gottlieb BS, Aronow BJ, Pavlidis P, Hinze CH, Thornton S, Thompson SD, Grom AA, Colbert RA, Glass DN. 2009. Gene expression signatures in polyarticular juvenile idiopathic arthritis demonstrate disease heterogeneity and offer a molecular classification of disease subsets. *Arthritis & Rheumatism* **60**:2113–2123. doi:10/c9ct9g

Hao R, Du H, Guo L, Tian F, An N, Yang T, Wang C, Wang B, Zhou Z. 2017. Identification of dysregulated genes in rheumatoid arthritis based on bioinformatics analysis. *PeerJ* **5**:e3078–e3078. doi:10/gbnkhj

Harden JL, Lewis SM, Lish SR, Suárez-Fariñas M, Gareau D, Lentini T, Johnson-Huang LM, Krueger JG, Lowes MA. 2016. The tryptophan metabolism enzyme L-kynureninase is a novel inflammatory factor in psoriasis and other inflammatory diseases. *Journal of Allergy and Clinical Immunology* **137**:1830–1840. doi:10/f8q2r7

Harlow L, Gochuico BR, Rosas IO, Doyle TJ, Osorio JC, Travers TS, Camacho CC, Oddis CV, Ascherman DP. 2014. Anti-citrullinated heat shock protein 90 antibodies identified in bronchoalveolar lavage fluid are a marker of lung-specific immune responses. *Clinical Immunology* **155**:60–70. doi:10/f7qrbs

Haynes WA, Haddon DJ, Diep V, Khatri A, Bongen E, Yiu G, Balboni I, Bolen CR, Mao R, Utz PJ, Khatri P. 2020. Integrated, multi-cohort analysis reveals unified signature of systemic lupus erythematosus. *JCI Insight*. doi:10.1172/jci.insight.122312

Hinze CH, Foell D, Johnson AL, Spalding SJ, Gottlieb BS, Morris PW, Kimura Y, Onel K, Li SC, Grom AA, Taylor J, Brunner HI, Huggins JL, Nocton JJ, Haines KA, Edelheit BS, Shishov M, Jung LK, Williams CB, Tesher MS, Costanzo DM, Zemel LS, Dare JA, Passo MH, Ede KC, Olson JC, Cassidy EA, Griffin TA, Wagner-Weiner L, Weiss JE, Vogler LB, Rouster-Stevens KA, Beukelman T, Cron RQ, Kietz D, Schikler K, Mehta J, Ting TV, Verbsky JW, Eberhard AB, Huang B, Giannini EH, Lovell DJ. 2019. Serum S100A8/A9 and S100A12 Levels in Children With Polyarticular Forms of Juvenile Idiopathic Arthritis: Relationship to Maintenance of Clinically Inactive Disease During Anti–Tumor Necrosis Factor Therapy and Occurrence of Disease Flare After Discontinu. *Arthritis and Rheumatology* **71**:451–459. doi:10/gg277p

Holzinger D, Tenbrock K, Roth J. 2019. Alarmins of the S100-family in juvenile autoimmune and auto-inflammatory diseases. *Frontiers in Immunology* **10**:1–13. doi:10/gg277r

Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R, Hahne F, Hansen KD, Irizarry RA, Lawrence M, Love MI, MacDonald J, Obenchain V, Oleś AK, Pagès H, Reyes A, Shannon P, Smyth GK, Tenenbaum D, Waldron L, Morgan M. 2015. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods* **12**:115–121. doi:10/bb35

Hughey JJ, Butte AJ. 2015. Robust meta-analysis of gene expression using the elastic net. *Nucleic acids research* **43**:e79–e79. doi:10/f7nnbm

Jiang K, Frank MB, Chen Y, Osban J, Jarvis JN. 2013. Genomic characterization of remission in juvenile idiopathic arthritis. *Arthritis Res Ther* **15**:R100. doi:10/gb9z3v

Johnson WE, Li C, Rabinovic A. 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**:118–127. doi:10/dsf386

Kehlen A, Haegele M, Böhme L, Cynis H, Hoffmann T, Demuth H-U. 2017. N-terminal pyroglutamate formation in CX3CL1 is essential for its full biologic activity. *Biosci Rep* **37**. doi:10/gbtbff

Korte-Bouws GAH, Albers E, Voskamp M, Hendriksen H, De Leeuw LR, Güntürkün O, De Roock S, Vastert SJ, Korte SM. 2019. Juvenile Arthritis Patients Suffering from Chronic Inflammation Have Increased Activity of Both IDO and GTP-CH1 Pathways But Decreased BH4 Efficacy: Implications for Well-Being, Including Fatigue, Cognitive Impairment, Anxiety, and Depression. *Pharmaceuticals* **12**:9. doi:10/gg4jb3

Kuhn M. 2008. Building predictive models in R using the caret package. *Journal of Statistical Software* **28**:1–26. doi:10.18637/jss.v028.i05

Lai C-H, Hsieh C-Y, Barnado A, Huang L-C, Chen S-C, Tsai L-M, Shyr Y, Li C-Y. 2020. Outcomes of acute cardiovascular events in rheumatoid arthritis and systemic lupus erythematosus: a population-based study. *Rheumatology (Oxford)* **59**:1355–1363. doi:10.1093/rheumatology/kez456

Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. 2012. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**:882–883. doi:10/fx9kw2

Li Z, Gao N, Martini JWR, Simianer H. 2019. Integrating gene expression data into genomic prediction. *Frontiers in Genetics* **10**:1–11. doi:10.3389/fgene.2019.00126

Li Z-C, Xiao J, Peng J-L, Chen J-W, Ma T, Cheng G-Q, Dong Y-Q, Wang W, Liu Z-D. 2014. Functional Annotation of

Rheumatoid Arthritis and Osteoarthritis Associated Genes by Integrative Genome-Wide Gene Expression Profiling Analysis. *PLoS ONE* **9**:e85784–e85784. doi:10/gg2772

Liao Y, Tong L, Tang L, Wu S. 2017. The role of cold-inducible RNA binding protein in cell stress response. *International Journal of Cancer* **141**:2164–2173. doi:10/gg3qwt

Liu J, Yin L, Dong H, Xu E, Zhang L, Qiao Y, Liu Y, Li L, Jia J. 2012. Decreased serum levels of nucleolin protein fragment, as analyzed by bead-based proteomic technology, in multiple sclerosis patients compared to controls. *J Neuroimmunol* **250**:71–76. doi:10/f35q4t

Lofgren S, Hinchcliff M, Carns M, Wood T, Aren K, Arroyo E, Cheung P, Kuo A, Valenzuela A, Haemel A, Wolters PJ, Gordon J, Spiera R, Assassi S, Boin F, Chung L, Fiorentino D, Utz PJ, Whitfield ML, Khatri P. 2016. Integrated, multicohort analysis of systemic sclerosis identifies robust transcriptional signature of disease severity. *JCI Insight* **1**:1–12. doi:10.1172/jci.insight.89073

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**:1–21. doi:10.1186/s13059-014-0550-8

Milner CM, Day AJ. 2003. TSG-6: a multifunctional protein associated with inflammation. *J Cell Sci* **116**:1863–1873. doi:10/fhm8pw

Molineros JE, Maiti AK, Sun C, Looger LL, Han S, Kim-Howard X, Glenn S, Adler A, Kelly JA, Niewold TB, Gilkeson GS, Brown EE, Alarcón GS, Edberg JC, Petri M, Ramsey-Goldman R, Reveille JD, Vilá LM, Freedman BI, Tsao BP, Criswell LA, Jacob CO, Moore JH, Vyse TJ, Langefeld CL, Guthridge JM, Gaffney PM, Moser KL, Scofield RH, Alarcón-Riquelme ME, Network on behalf of the B, Williams SM, Merrill JT, James JA, Kaufman KM, Kimberly RP, Harley JB, Nath SK. 2013. Admixture Mapping in Lupus Identifies Multiple Functional Variants within IFIH1 Associated with Apoptosis, Inflammation, and Autoantibody Production. *PLOS Genetics* **9**:e1003222. doi:10/f3vcfv

Moncrieffe H, Nistala K, Kamhieh Y, Evans J, Eddaoudi A, Eaton S, Wedderburn LR. 2010. High Expression of the Ectonucleotidase CD39 on T Cells from the Inflamed Site Identifies Two Distinct Populations, One Regulatory and One Memory T Cell Population. *The Journal of Immunology* **185**:134–143. doi:10/c52pnw

Nagai Y, Akashi S, Nagafuku M, Ogata M, Iwakura Y, Akira S, Kitamura T, Kosugi A, Kimoto M, Miyake K. 2002. Essential role of MD-2 in LPS responsiveness and TLR4 distribution. *Nat Immunol* **3**:667–672. doi:10/cn9sgf

NCBI. 2020a. QPCT glutaminyl-peptide cyclotransferase [Homo sapiens (human)] - Gene - NCBI. https://www.ncbi.nlm.nih.gov/gene/25797

NCBI. 2020b. HSP90AB1 heat shock protein 90 alpha family class B member 1 [Homo sapiens (human)] - Gene - NCBI. https://www-ncbi-nlm-nih-gov.ucsf.idm.oclc.org/gene/3326

Neely J, Rychkov D, Paranjpe M, Waterfield M, Kim S, Sirota M. 2019. Gene Expression Meta-Analysis Reveals Concordance in Gene Activation, Pathway, and Cell-Type Enrichment in Dermatomyositis Target Tissues. *ACR Open Rheumatology* acr2.11081-acr2.11081. doi:10/gg2776

Nordal HH, Brun JG, Hordvik M, Eidsheim M, Jonsson R, Halse AK. 2016. Calprotectin (S100A8/A9) and S100A12 are associated with measures of disease activity in a longitudinal study of patients with rheumatoid arthritis treated with infliximab. *Scandinavian Journal of Rheumatology* **45**:274–281. doi:10/gg277f

Orange Dana E, Agius P, DiCarlo EF, Robine N, Geiger H, Szymonifka J, McNamara M, Cummings R, Andersen KM, Mirza S, Figgie M, Ivashkiv LB, Pernis AB, Jiang CS, Frank MO, Darnell RB, Lingampali N, Robinson WH, Gravallese E, Bykerk VP, Goodman SM, Donlin LT. 2018. Identification of Three Rheumatoid Arthritis Disease Subtypes by Machine Learning Integration of Synovial Histologic Features and RNA Sequencing Data. *Arthritis & Rheumatology* **70**:690–701. doi:10.1002/art.40428

Orange Dana E., Agius P, DiCarlo EF, Robine N, Geiger H, Szymonifka J, McNamara M, Cummings R, Andersen KM, Mirza S, Figgie M, Ivashkiv LB, Pernis AB, Jiang CS, Frank MO, Darnell RB, Lingampali N, Robinson WH, Gravallese E, Bykerk VP, Goodman SM, Donlin LT. 2018. Identification of Three Rheumatoid Arthritis Disease Subtypes by Machine Learning Integration of Synovial Histologic Features and RNA Sequencing Data. *Arthritis and Rheumatology* **70**:690–701. doi:10.1002/art.40428

Perez-Riverol Y, Kuhn M, Vizcaíno JA, Hitz M-P, Audain E. 2017. Accurate and fast feature selection workflow for high-dimensional omics data. *PLOS ONE* **12**:e0189875. doi:10.1371/journal.pone.0189875

Piccolo SR, Sun Y, Campbell JD, Lenburg ME, Bild AH, Johnson WE. 2012. A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics* **100**:337–344. doi:10/f4gwz9

Pineda S, Bunis DG, Kosti I, Sirota M. 2020. Data Integration for Immunology.

R core team. 2018. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**:e47–e47. doi:10.1093/nar/gkv007

Schroecksnadel K, Kaser S, Ledochowski M, Neurauter G, Mur E, Herold M, Fuchs D. 2003. Increased degradation of tryptophan in blood of patients with rheumatoid arthritis. *The Journal of Rheumatology* **30**:1935–1939.

Shchetynsky K, Diaz-Gallo LM, Folkersen L, Hensvold AH, Catrina AI, Berg L, Klareskog L, Padyukov L. 2017. Discovery of new candidate genes for rheumatoid arthritis through integration of genetic association data with expression pathway analysis. *Arthritis Research and Therapy* **19**:1–9. doi:10/gbvt7k

Spiliopoulou A, Colombo M, Plant D, Nair N, Cui J, Coenen MJ, Ikari K, Yamanaka H, Saevarsdottir S, Padyukov L, Jr SLB, Kimberly RP, Okada Y, Riel PLC van, Wolbink G, Horst-Bruinsma IE van der, Vries N de, Tak PP, Ohmura K, Canhão H, Guchelaar H-J, Huizinga TW, Criswell LA, Raychaudhuri S, Weinblatt ME, Wilson AG, Mariette X, Isaacs JD, Morgan AW, Pitzalis C, Barton A, McKeigue P. 2019. Association of response to TNF inhibitors in rheumatoid arthritis with quantitative trait loci for CD40 and CD39. *Annals of the Rheumatic Diseases* **78**:1055–1061. doi:10/gg4ffq

Sumitomo S, Nagafuchi Y, Tsuchida Y, Tsuchiya H, Ota M, Ishigaki K, Suzuki A, Kochi Y, Fujio K, Yamamoto K. 2018. Transcriptome analysis of peripheral blood from patients with rheumatoid arthritis: a systematic review. *Inflammation and Regeneration* **38**:1–5. doi:10.1186/s41232-018-0078-5

Teixeira VH, Olaso R, Martin-Magniette ML, Lasbleiz S, Jacq L, Oliveira CR, Hilliquin P, Gut I, Cornelis F, Petit-Teixeira E. 2009a. Transcriptome analysis describing new immunity and defense genes in peripheral blood mononuclear cells of rheumatoid arthritis patients. *PLoS ONE* **4**. doi:10.1371/journal.pone.0006803

Teixeira VH, Olaso R, Martin-Magniette ML, Lasbleiz S, Jacq L, Oliveira CR, Hilliquin P, Gut I, Cornelis F, Petit-Teixeira E. 2009b. Transcriptome analysis describing new immunity and defense genes in peripheral blood mononuclear cells of rheumatoid arthritis patients. *PLoS ONE* **4**. doi:10/d4xpcp

Travers TS, Harlow L, Rosas IO, Gochuico BR, Mikuls TR, Bhattacharya SK, Camacho CJ, Ascherman DP. 2016. Extensive Citrullination Promotes Immunogenicity of HSP90 through Protein Unfolding and Exposure of Cryptic Epitopes. *The Journal of Immunology* **197**:1926–1936. doi:10/f87dn8

Uhlén M, Karlsson MJ, Hober A, Svensson A-S, Scheffel J, Kotol D, Zhong W, Tebani A, Strandberg L, Edfors F, Sjöstedt E, Mulder J, Mardinoglu A, Berling A, Ekblad S, Dannemeyer M, Kanje S, Rockberg J, Lundqvist M, Malm M, Volk A-L, Nilsson P, Månberg A, Dodig-Crnkovic T, Pin E, Zwahlen M, Oksvold P, Feilitzen K von, Häussler RS, Hong M-G, Lindskog C, Ponten F, Katona B, Vuu J, Lindström E, Nielsen J, Robinson J, Ayoglu B, Mahdessian D, Sullivan D, Thul P, Danielsson F, Stadler C, Lundberg E, Bergström G, Gummesson A, Voldborg BG, Tegel H, Hober S, Forsström B, Schwenk JM, Fagerberg L, Sivertsson Å. 2019. The human secretome. *Sci Signal* **12**. doi:10/ggdq74

Van Nies JAB, Tsonaka R, Gaujoux-Viala C, Fautrel B, Van Der Helm-Van Mil AHM. 2015. Evaluating relationships between symptom duration and persistence of rheumatoid arthritis: Does a window of opportunity exist? Results on the Leiden Early Arthritis Clinic and ESPOIR cohorts. *Annals of the Rheumatic Diseases* **74**:806–812. doi:10/gg278b

Wang L, Wu LF, Lu X, Mo XB, Tang ZX, Lei SF, Deng FY. 2015a. Integrated analyses of gene expression profiles digs out common markers for rheumatic diseases. *PLoS ONE* **10**:1–11. doi:10.1371/journal.pone.0137522

Wang L, Wu LF, Lu X, Mo XB, Tang ZX, Lei SF, Deng FY. 2015b. Integrated analyses of gene expression profiles digs out common markers for rheumatic diseases. *PLoS ONE* **10**:1–11. doi:10/gg277n

Wang Q, Liu D, Song P, Zou M-H. 2015. Deregulated tryptophan-kynurenine pathway is linked to inflammation, oxidative stress, and immune activation pathway in cardiovascular diseases. *Front Biosci (Landmark Ed)* **20**:1116–1143.

Wang Siwen, Song R, Wang Z, Jing Z, Wang Shaoxiong, Ma J. 2018. S100A8/A9 in Inflammation. *Front Immunol* **9**. doi:10/gdrkmp

Wisniewski HG, Maier R, Lotz M, Lee S, Klampfer L, Lee TH, Vilcek J. 1993. TSG-6: a TNF-, IL-1-, and LPS-inducible secreted glycoprotein associated with arthritis. *The Journal of Immunology* **151**:6593–6601.

Wu G, Zhu L, Dent JE, Nardini C. 2010. A comprehensive molecular interaction map for rheumatoid arthritis. *PLoS ONE* **5**. doi:10/dbvvr8

Yang W, Tang H, Zhang Y, Tang X, Zhang J, Sun L, Yang J, Cui Y, Zhang Lu, Hirankarn N, Cheng H, Pan H-F, Gao J, Lee TL, Sheng Y, Lau CS, Li Y, Chan TM, Yin X, Ying D, Lu Q, Leung AMH, Zuo X, Chen X, Tong KL, Zhou F, Diao Q,

Tse NKC, Xie H, Mok CC, Hao F, Wong SN, Shi B, Lee KW, Hui Y, Ho MHK, Liang B, Lee PPW, Cui H, Guo Q, Chung BH-Y, Pu X, Liu Q, Zhang Xiaoguang, Zhang C, Chong CY, Fang H, Wong RWS, Sun Y, Mok MY, Li X-P, Avihingsanon Y, Zhai Z, Rianthavorn P, Deekajorndej T, Suphapeetiporn K, Gao F, Shotelersuk V, Kang X, Ying SKY, Zhang Lijuan, Wong WHS, Zhu D, Fung SKS, Zeng F, Lai WM, Wong C-M, Ng IOL, Garcia-Barceló M-M, Cherny SS, Shen N, Tam PK-H, Sham PC, Ye D-Q, Yang S, Zhang Xuejun, Lau YL. 2013. Meta-analysis Followed by Replication Identifies Loci in or near CDKN1B, TET3, CD80, DRAM1, and ARID5B as Associated with Systemic Lupus Erythematosus in Asians. *The American Journal of Human Genetics* **92**:41–51. doi:10/f4k244

Yoo IS, Lee SY, Park CK, Lee JC, Kim Y, Yoo SJ, Shim SC, Choi YS, Lee Y, Kang SW. 2018. Serum and synovial fluid concentrations of cold-inducible RNA-binding protein in patients with rheumatoid arthritis. *International Journal of Rheumatic Diseases* **21**:148–154. doi:10/gg3qws

Yu G, Wang LG, Han Y, He QY. 2012. ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS A Journal of Integrative Biology* **16**:284–287. doi:10.1089/omi.2011.0118

Tables

Gene	Gene name	Regulation	Discovery Synovium			Discovery Blood			Validation	Protein Secretion
			FC (FDR adj. p-value)	⍰ (BH adj. p-value)	AUC	FC (FDR adj. p-value)	⍰ (BH adj. p-value)	AUC	AUC	
TNFAIP6	TNF Alpha Induced Protein 6	up	2.46 (4E-06)	0.39 (7E-11)	0.81	1.36 (8E-16)	0.39 (3E-67)	0.77	0.88	Secreted in blood
S100A8	S100 Calcium Binding Protein A8	up	2.28 (7E-05)	0.34 (1E-08)	0.81	1.46 (7E-32)	0.48 (9E-108)	0.81	0.94	Secreted in blood
DRAM1	DNA Damage Regulated Autophagy Modulator 1	up	1.55 (6E-07)	0.46 (3E-15)	0.93	1.18 (8E-15)	0.41 (6E-76)	0.79	0.81	
TNFSF10	TNF Superfamily Member 10	up	1.55 (1E-09)	0.52 (3E-19)	0.9	1.27 (1E-23)	0.44 (4E-88)	0.8	0.84	Secreted in blood
LY96	Lymphocyte Antigen 96	up	1.54 (1E-09)	0.51 (2E-18)	0.94	1.22 (7E-11)	0.28 (2E-35)	0.69	0.87	Secreted in blood
QPCT	Glutaminyl-Peptide Cyclotransferase	up	1.46 (4E-05)	0.39 (7E-11)	0.92	1.19 (4E-10)	0.29 (1E-37)	0.71	0.82	Secreted in blood
KYNU	Kynureninase	up	1.41 (5E-05)	0.36 (1E-09)	0.84	1.17 (2E-11)	0.28 (3E-34)	0.69	0.82	Intracellular or membrane-bound
ENTPD1	Ectonucleoside Triphosphate Diphosphohydrolase 1	up	1.33 (1E-08)	0.52 (2E-19)	0.94	1.21 (2E-16)	0.4 (5E-71)	0.78	0.86	
CLIC1	Chloride Intracellular Channel 1	up	1.32 (5E-08)	0.47 (7E-16)	0.91	1.2 (5E-27)	0.47 (4E-103)	0.84	0.8	Intracellular or membrane-bound
ATP6V0E1	ATPase H+ Transporting V0 Subunit E1	up	1.23 (3E-04)	0.37 (8E-10)	0.84	1.08 (4E-10)	0.28 (3E-35)	0.7	0.82	
NCL	Nucleolin	down	0.83 (2E-05)	-0.39 (4E-11)	0.82	0.88 (4E-09)	-0.32 (2E-44)	0.72	0.82	
CIRBP	Cold Inducible RNA Binding Protein	down	0.8 (3E-05)	-0.41 (4E-12)	0.83	0.91 (2E-10)	-0.33 (2E-47)	0.74	0.89	
HSP90AB1	Heat Shock Protein 90 Alpha Family Class B Member 1	down	0.79 (2E-04)	-0.37 (3E-10)	0.82	0.84 (4E-12)	-0.36 (7E-56)	0.73	0.8	Intracellular or membrane-bound

Table 1. Summary of 13 validated RAScore Panel genes.

Figures

Figure 1. Study overview. a) Public data collection, processing and DGE analysis. b) Feature selection pipeline. c) Gene list validation on the independent datasets. Introducing the RAScore as a geometric mean of validated genes and its association with clinical outcomes.

Figure 2. DE genes overlapped between synovium and whole blood tissues. Top Reactome common and different pathways for a) Up-regulated and b) down-regulated genes. Venn diagram of up- and down-regulated genes in synovium and blood: 28 common up-regulated genes (p =9e-09) and 4 common down-regulated genes (p = 0.28). d) Comparison scatter plot of fold changes between common genes in synovium and blood. Heatmap and PCA plots of common genes in e) and f synovium and f) blood. Vertical bars in the heatmap plots represent the color-coded coefficients of variation, Pearson correlations and log2 fold changes.

Figure 3. Cell type enrichment analysis for synovium and whole blood tissues. BH adj p-values <0.05. 30 significant cell types in synovium, 20 significant cell types in WB, 11 common significant cell types.

Figure 4. Feature selected genes. a) Mean AUC performance with standard error for each feature selected gene on testing synovium and blood datasets (green) and on five independent validation datasets (black). 13 genes with AUC greater than 0.8 on validation datasets were chosen as the best performing genes. Mean AUC performance with standard errors of a RF model trained on discovery blood data with b) common DE genes and c) feature selected genes on five independent validation datasets.

Figure 5. Clinical Interpretation of the RAScore. a) A forest plot of correlation RAScore with DAS28. b) Forest plots of correlations of some RAScore Panel genes with DAS28. c) RAScore tracks the treatment effect in both synovium and blood but shows no difference between RF+ and RF- phenotypes. d) RAScore tracks the treatment effect in both synovium and blood. RA_utr: RA untreated. RA_tr: RA treated. e) RAScore distinguishes Healthy, OA and RA samples in synovium. f) RAScore distinguishes Healthy and polyJIA samples. The p-values were obtained using Student's t-test.

Supplementary Tables

Supplementary Table 1. Overview of the Discovery and Validation Studies

Supplementary Table 2. Treatment Categories

Supplementary Table 3. Significantly Differentially Expressed Genes - Synovium

Supplementary Table 4. Significantly Differentially Expressed Genes - Blood

Supplementary Table5. Significant Pathways - Synovium

Supplementary Table6. Significant Pathways - Blood

Supplementary Table7. 32 Common DE Genes

Supplementary Table 8. 53 Feature Selected Genes

Supplementary Figures

Supplementary Figure 1. Discovery Data Batch Correction, QC and PCA plots

For synovium and whole blood: a) and b) before batch correction; c) and d) after normalization colored by batch; e) and f) after normalization colored by treatment type; g) and h) after normalization colored by phenotype.

Supplementary Figure 2. Blood Differential Expression and Pathway Analysis

DGE analysis in synovial tissue. a) Heatmap and b) PCA plot with DE genes. Reactome pathways for c) , d) up- and e), f) down-regulated genes.

Supplementary Figure 3. Synovium Differential Expression and Pathway Analysis

DGE analysis in whole blood. a) Heatmap and b) PCA plot with DE genes. Reactome pathways for c) , d) up- and e), f) down-regulated genes.

Supplementary Figure 4. Machine Learning Model Comparison

AUROC plots for common and feature selected genes. Three models, a Logistic Regression, Elastic Net and Random Forest, were trained on the discovery whole blood data using either common genes or feature selected genes and validated on 5 validation datasets. The summary curves are the averaged curves with bars of standard errors and colored by red. The dashed and solid lines represent synovium and blood data, respectively.

Supplementary Figure 5. Validation Data Batch Correction, QC and PCA Plots

Heatmap and PCA plots of 13 best performing genes on the independent validation a) synovium RNA-seq GSE89408, b) synovium microarray GSE1919, c) whole blood microarray GSE90081, d) PBMC RNA-seq GSE17755 and e) PBMC microarray GSE15573 datasets.

Supplementary Figure 6. Correlation forest plots with DAS28 for all 13 feature selected genes.

Supplementary Figure 7. Correlation of DAS score with RA Score for synovium GSE45867 and blood GSE15258, GSE58795, GSE93272 datasets.

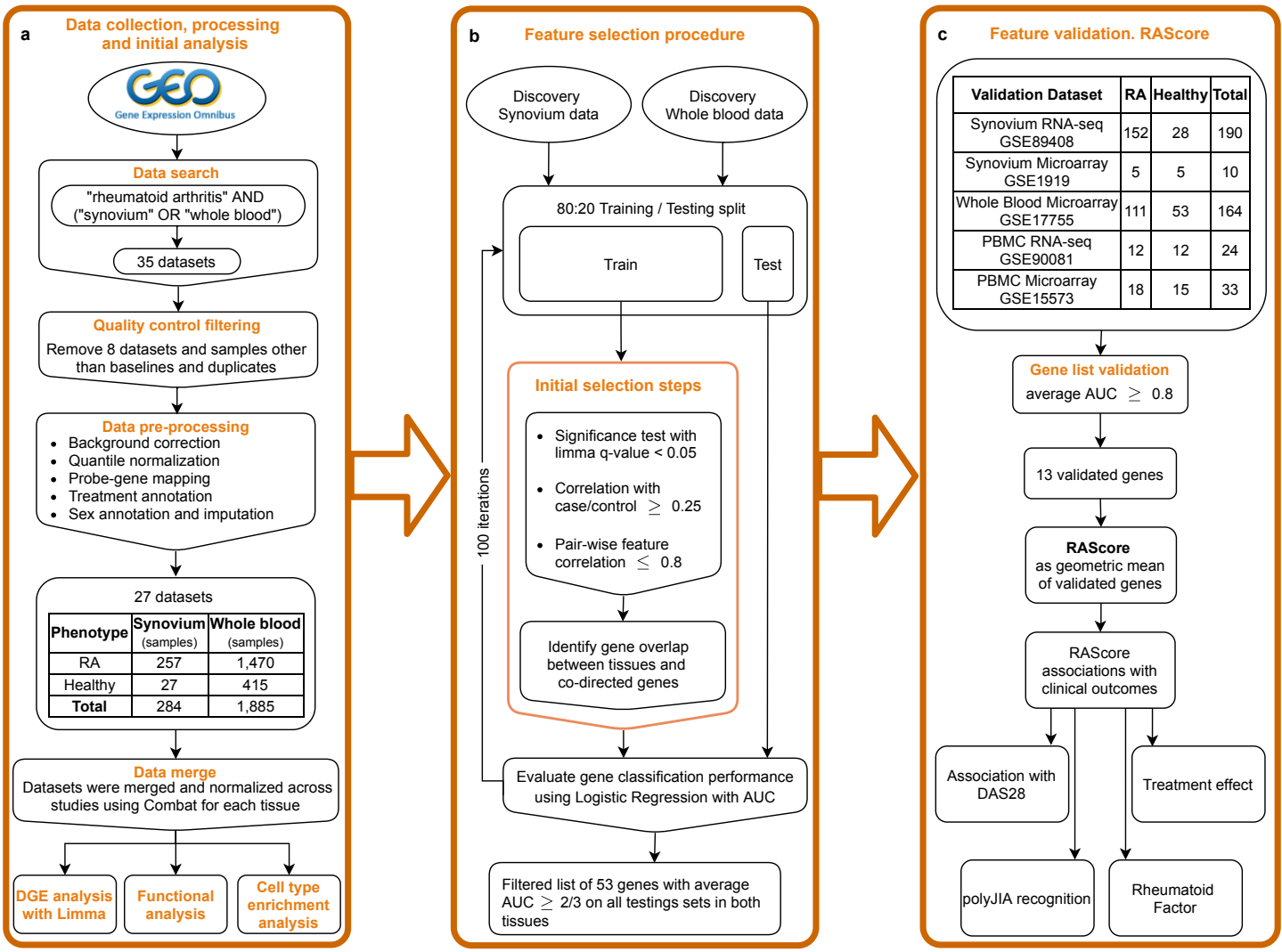


Figure 1. Study overview. a) Public data collection, processing and DGE analysis. b) Feature selection pipeline. c) Gene list validation on the independent datasets. Introducing the RAScore as a geometric mean of validated genes and its association with clinical outcomes.

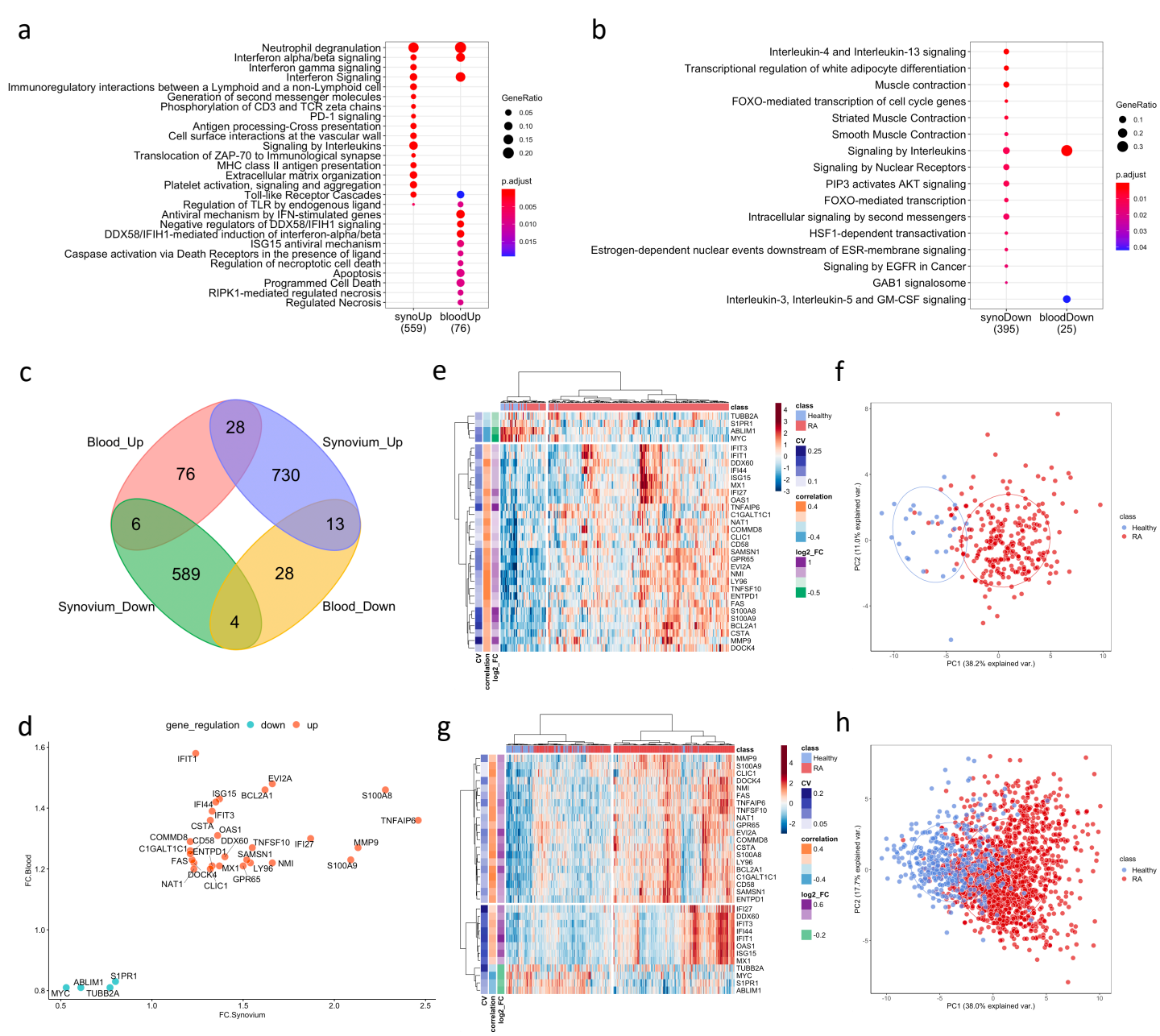


Figure 2. <u>DE genes overlapped between synovium and whole blood tissues.</u> Top Reactome common and different pathways for a) Up-regulated and b) down-regulated genes. Venn diagram of up- and down-regulated genes in synovium and blood: 28 common up-regulated genes (p =9e-09) and 4 common down-regulated genes (p = 0.28). d) Comparison scatter plot of fold changes between common genes in synovium and blood. Heatmap and PCA plots of common genes in e) and f synovium and f) blood. Vertical bars in the heatmap plots represent the color-coded coefficients of variation, Pearson correlations and log2 fold changes.

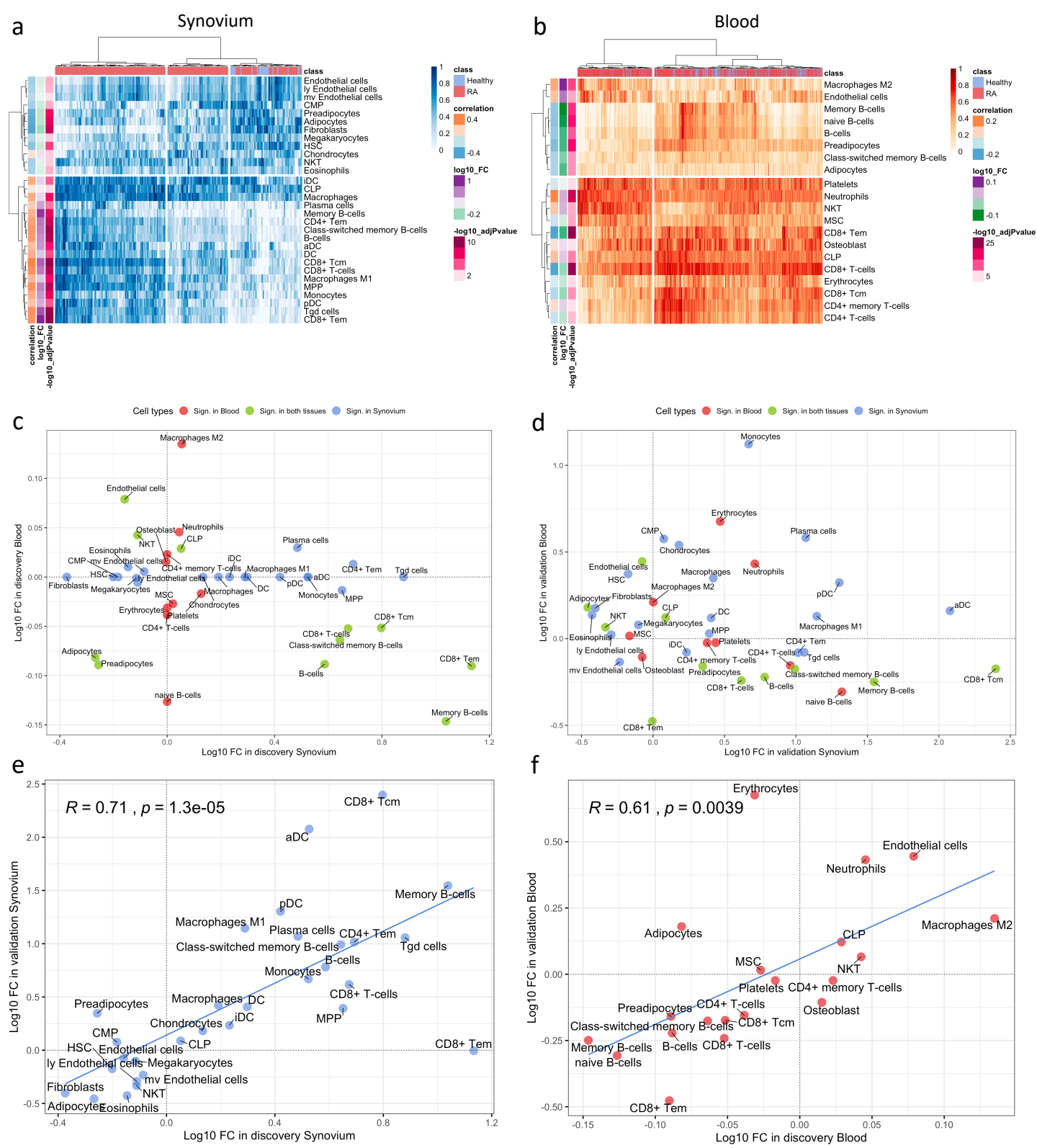


Figure 3. <u>Cell type enrichment analysis for synovium and and whole blood tissues.</u> BH adj p-values <0.05. 30 significant cell types in synovium, 20 significant cell types in WB, 11 common significant cell types.

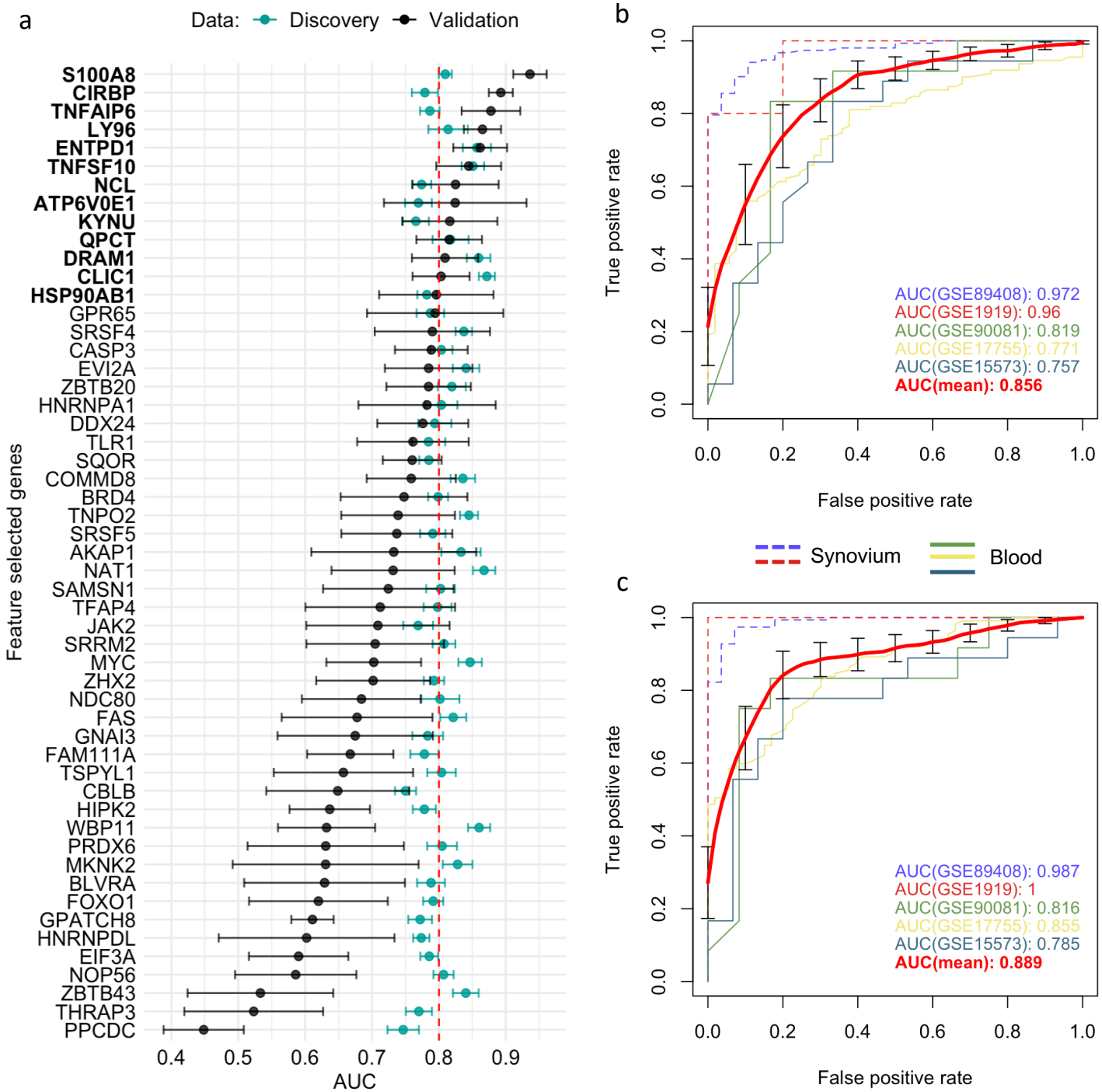


Figure 4. Feature selected genes. a) Mean AUC performance with standard error for each feature selected gene on testing synovium and blood datasets (green) and on five independent validation datasets (black). 13 genes with AUC greater than 0.8 on validation datasets were chosen as the best performing genes. Mean AUC performance with standard errors of a RF model trained on discovery blood data with b) common DE genes and c) feature selected genes on five independent validation datasets.

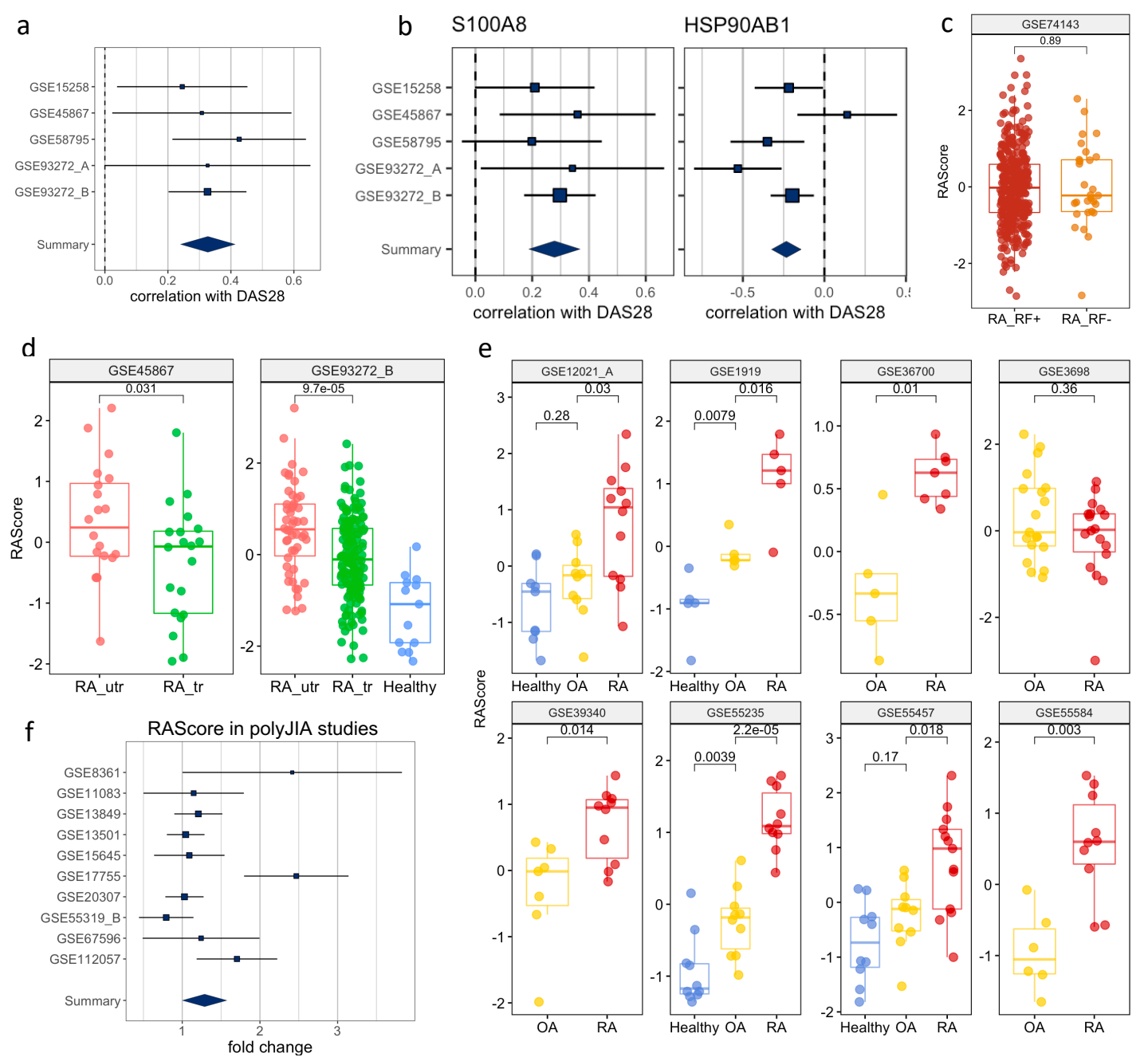
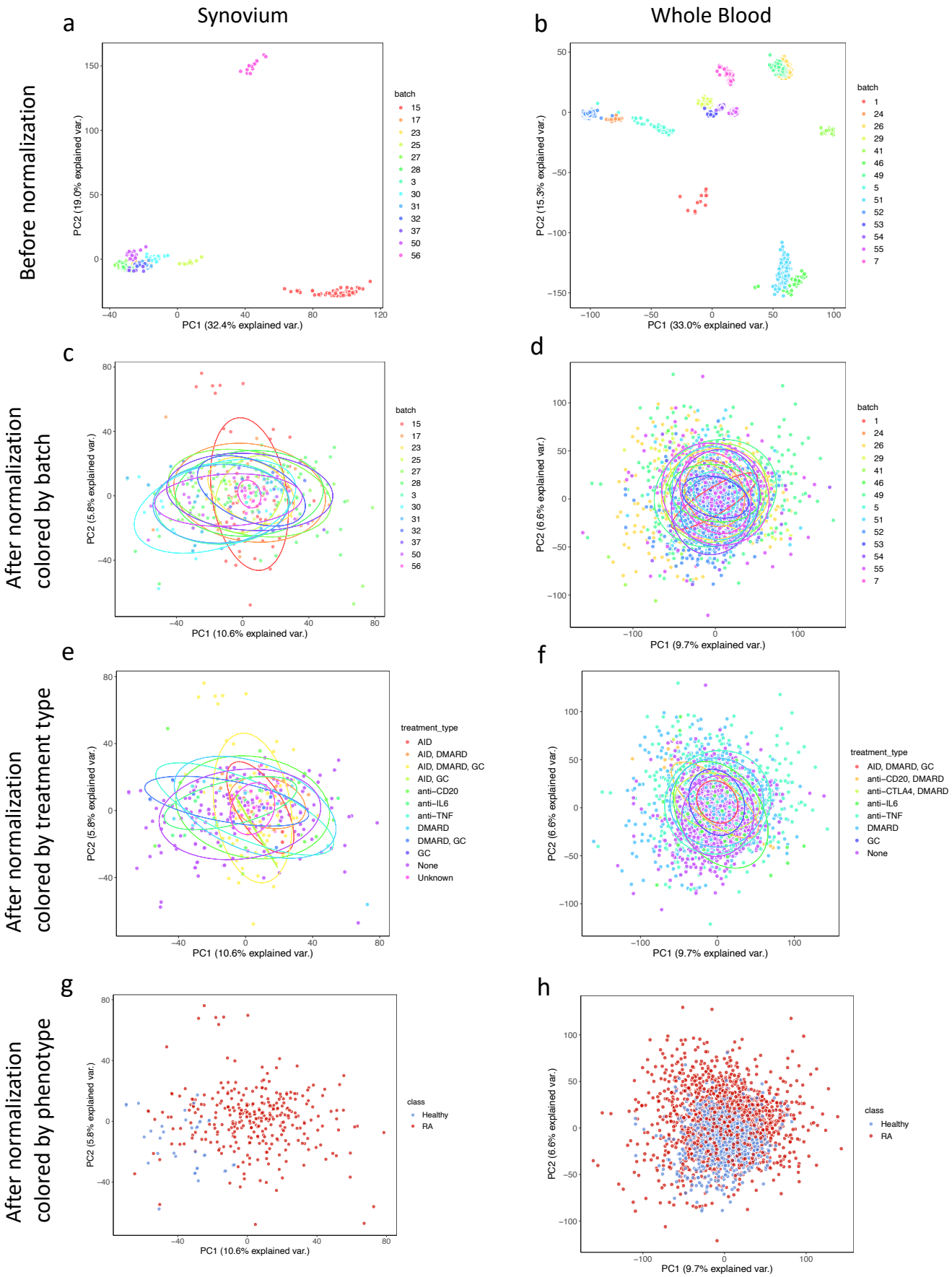
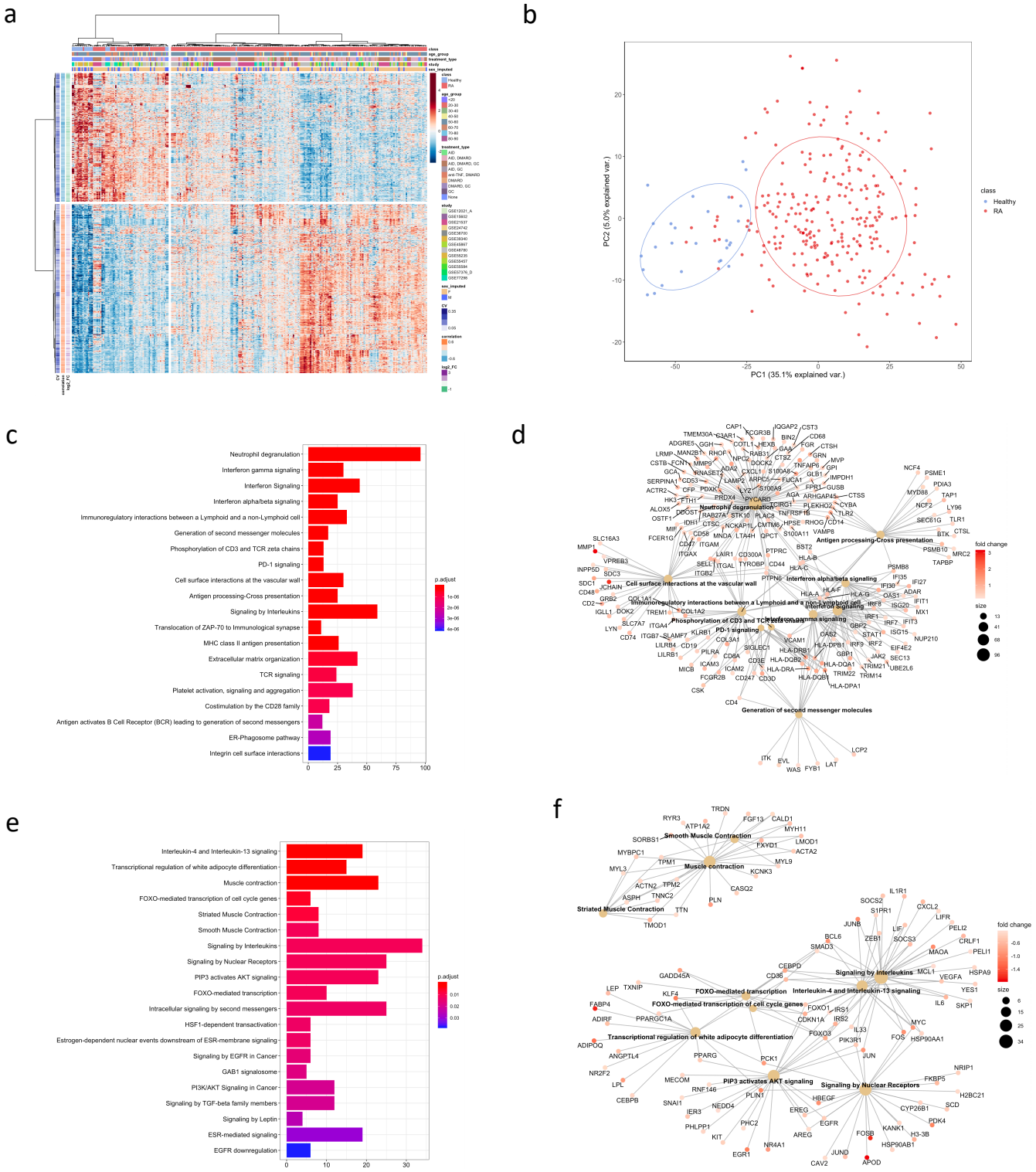


Figure 5. Clinical Interpretation of the RAScore. a) A forest plot of correlation RAScore with DAS28. b) Forest plots of correlations of some RAScore Panel genes with DAS28. c) RAScore tracks the treatment effect in both synovium and blood but shows no difference between RF+ and RF- phenotypes. d) RAScore tracks the treatment effect in both synovium and blood. RA_utr: RA untreated. RA_tr: RA treated. e) RAScore distinguishes Healthy, OA and RA samples in synovium. f) RAScore distinguishes Healthy and polyJIA samples. The p-values were obtained using Student's t-test.



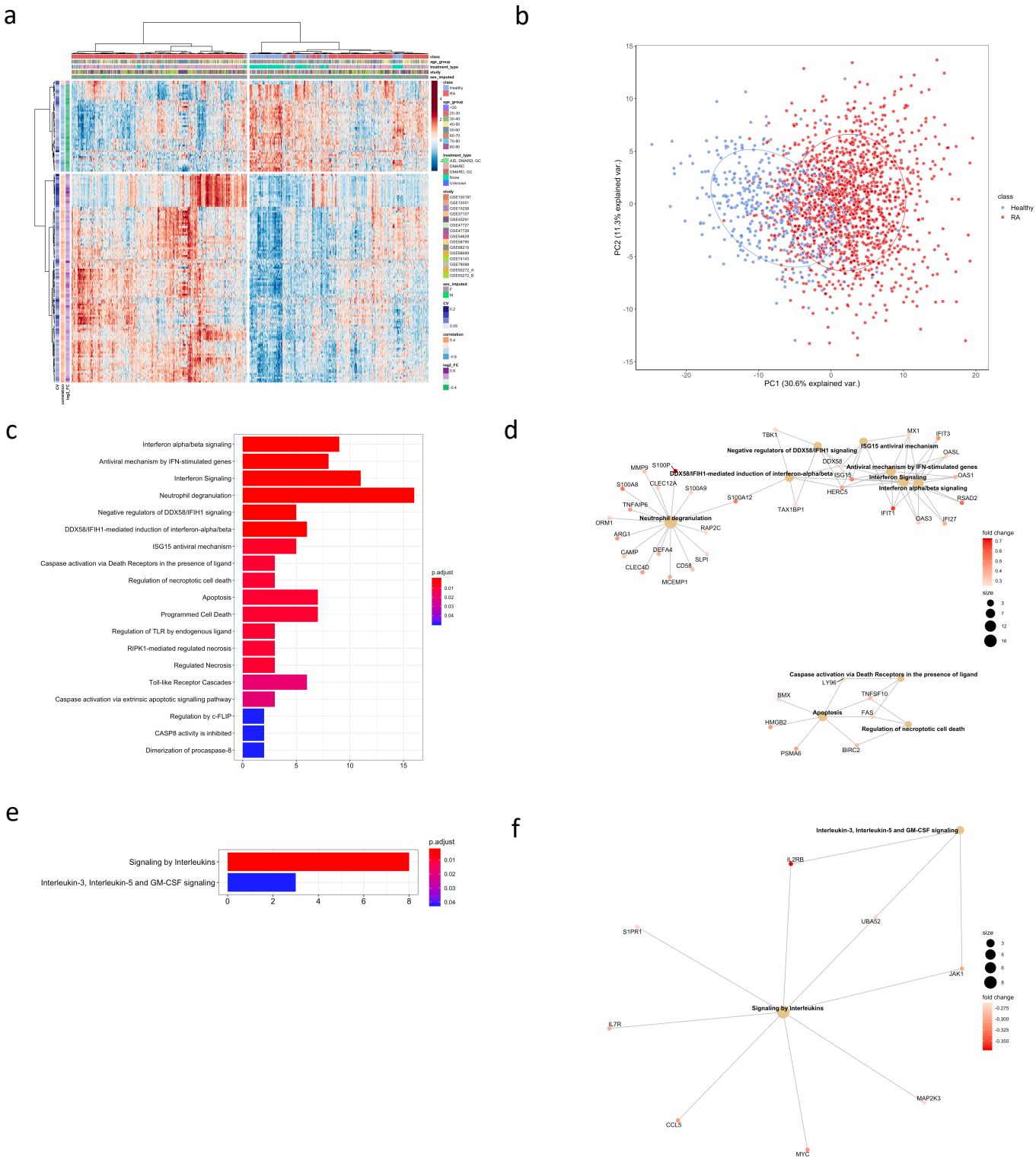
Supplementary Figure 1. Discovery Data Batch Correction, QC and PCA plots

For synovium and whole blood: a) and b) before batch correction; c) and d) after normalization colored by batch; e) and f) after normalization colored by treatment type; g) and h) after normalization colored by phenotype.



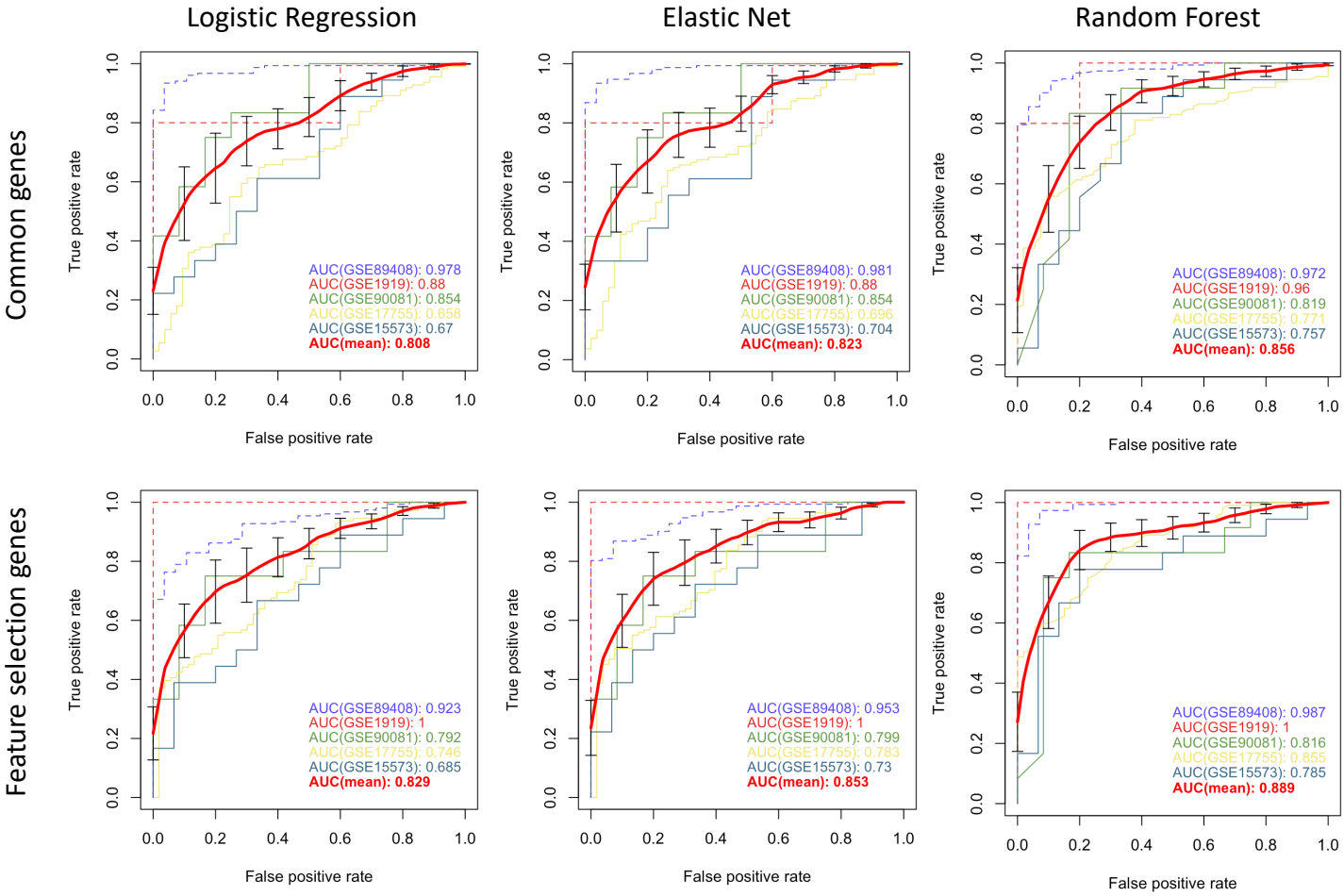
Supplementary Figure 2. Blood Differential Expression and Pathway Analysis

DGE analysis in synovial tissue. a) Heatmap and b) PCA plot with DE genes. Reactome pathways for c) , d) up- and e), f) down-regulated genes.



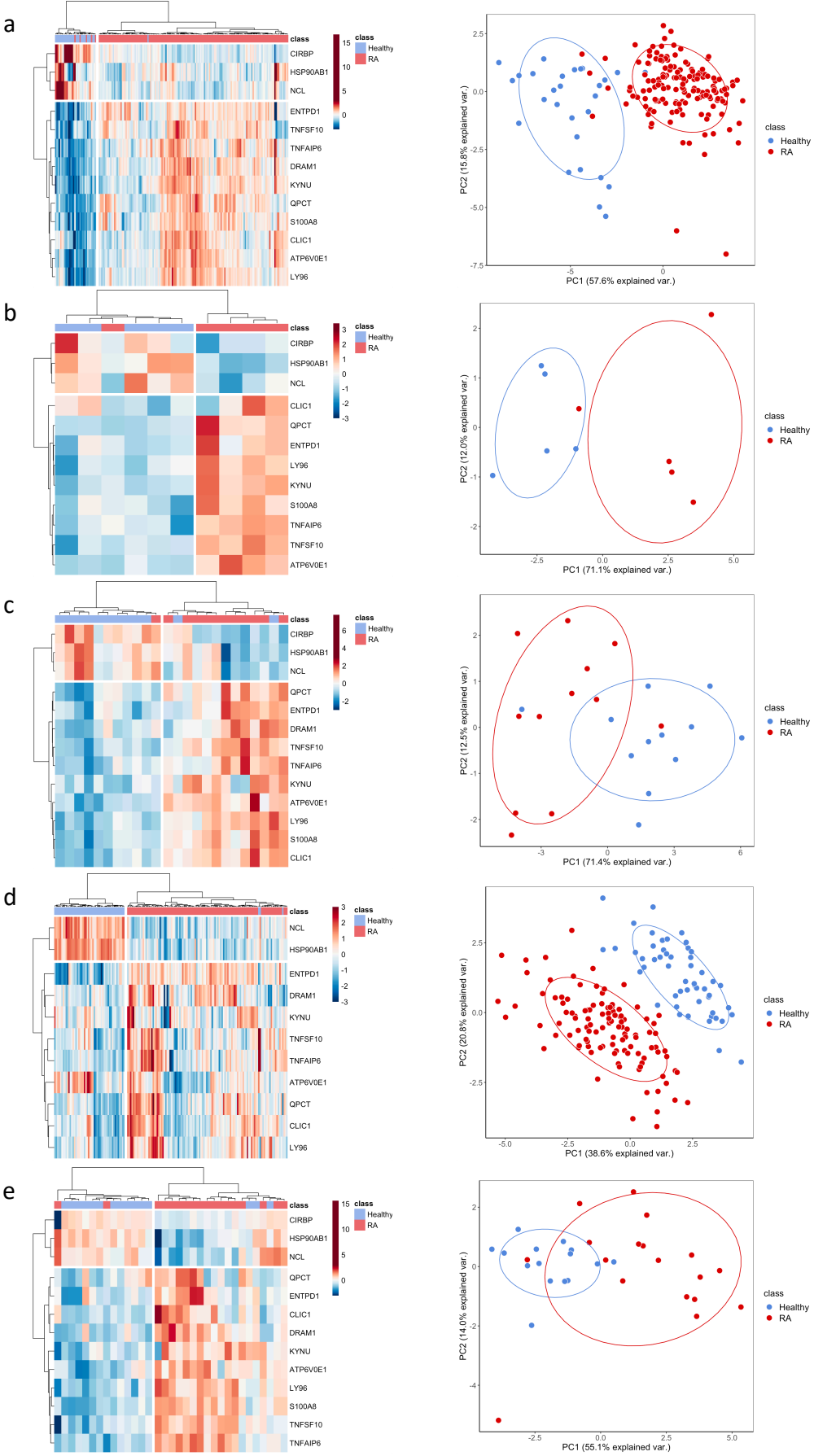
Supplementary Figure 3. Synovium Differential Expression and Pathway Analysis

DGE analysis in whole blood. a) Heatmap and b) PCA plot with DE genes. Reactome pathways for c) , d) up- and e), f) down-regulated genes.



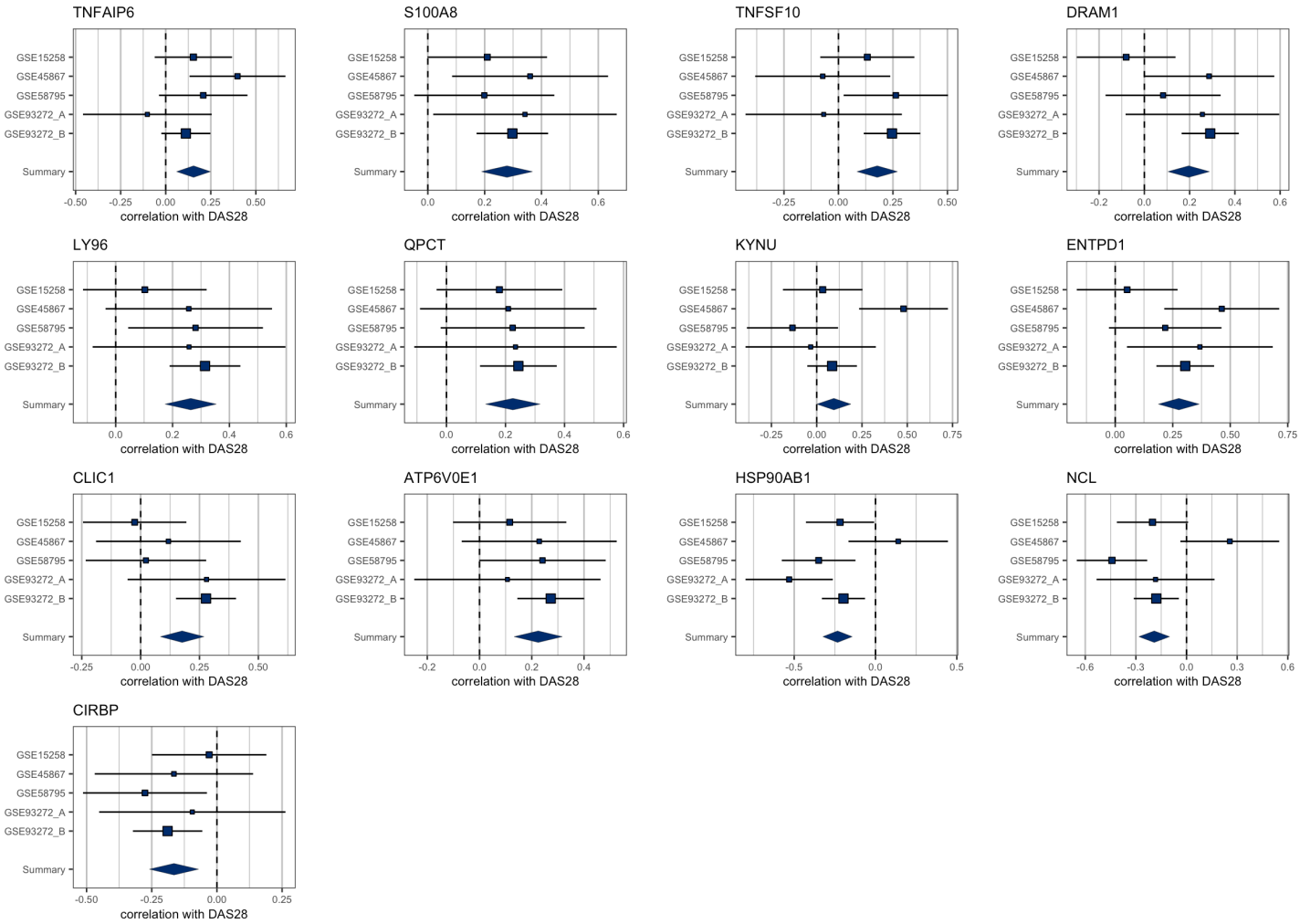
Supplementary Figure 4. Machine Learning Model Comparison

AUROC plots for common and feature selected genes. Three models, a Logistic Regression, Elastic Net and Random Forest, were trained on the discovery whole blood data using either common genes or feature selected genes and validated on 5 validation datasets. The summary curves are the averaged curves with bars of standard errors and colored by red. The dashed and solid lines represent synovium and blood data, respectively.

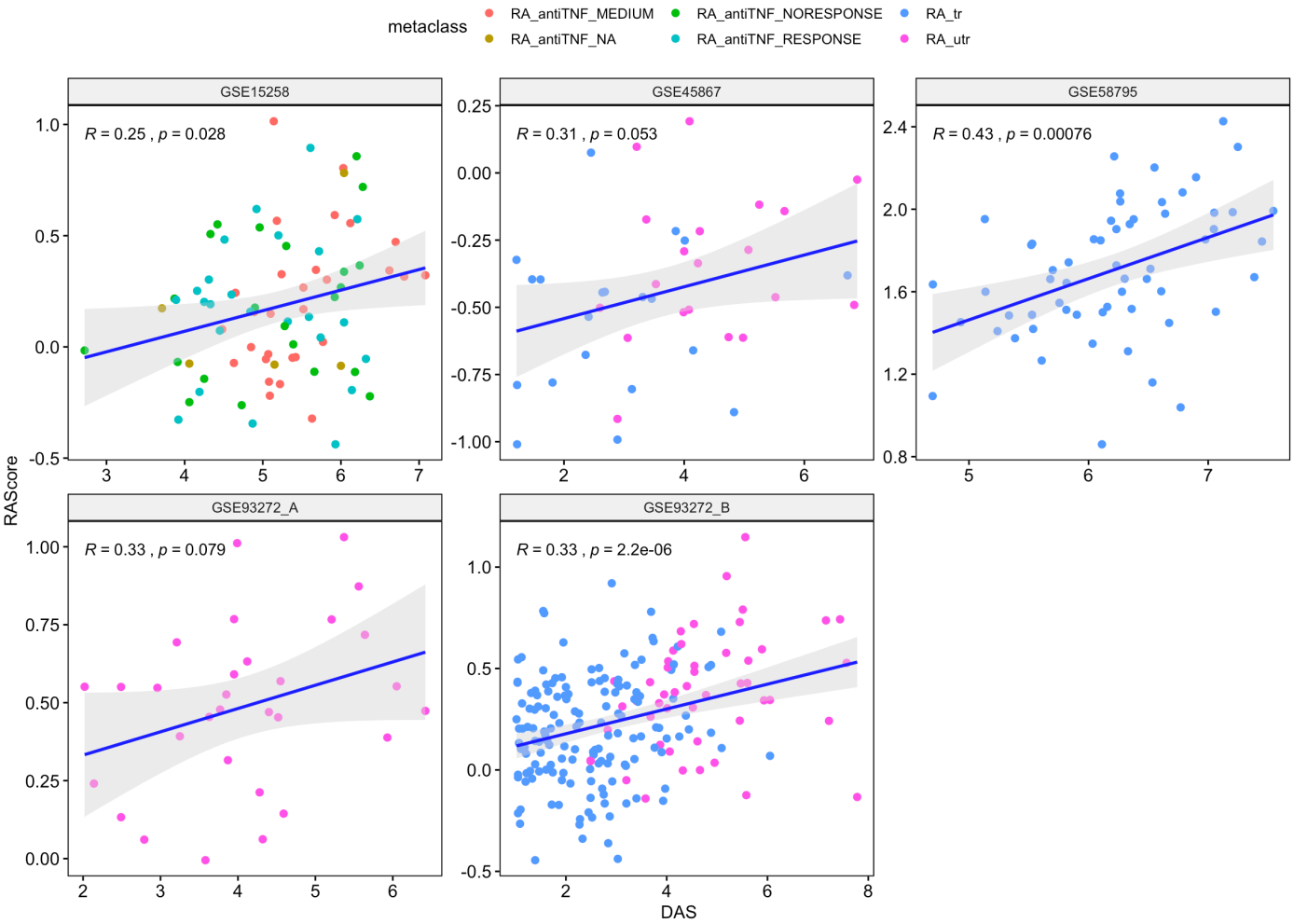


Supplementary Figure 5. Validation Data Batch Correction, QC and PCA Plots

Heatmap and PCA plots of 13 best performing genes on the independent validation a) synovium RNA-seq GSE89408, b) synovium microarray GSE1919, c) whole blood microarray GSE90081, d) PBMC RNA-seq GSE17755 and e) PBMC microarray GSE15573 datasets.



Supplementary Figure 6. Correlation forest plots with DAS28 for all 13 feature selected genes.



Supplementary Figure 7. Correlation of DAS score with RA Score for synovium GSE45867 and blood GSE15258, GSE58795, GSE93272 datasets.