Title

Compendium of synovial signatures identifies pathologic characteristics for predicting treatment response in rheumatoid arthritis patients.

Permalink
https://escholarship.org/uc/item/7t98602d

Authors
Kim, Ki-Jo
Kim, Minseung
Adamopoulos, Iannis E
et al.

Publication Date
2019-05-01

DOI
10.1016/j.clim.2019.03.002

Peer reviewed

Published in final edited form as:

Clin Immunol. 2019 May ; 202: 1–10. doi:10.1016/j.clim.2019.03.002.

Compendium of Synovial Signatures Identifies Pathologic Characteristics for Predicting Treatment Response in Rheumatoid Arthritis Patients

Ki-Jo Kim[a,b,c], **Minseung Kim**[b,c], **Iannis E Adamopoulos**[d], **Ilias Tagkopoulos**[b,c]

[a]Division of Rheumatology, St. Vincent Hospital, The Catholic University of Korea, Seoul, Republic of Korea

[b]Department of Computer Science, University of California at Davis, California, US

[c]Genome Center, University of California at Davis, California, US

[d]Division of Rheumatology, Allergy and Clinical Immunology, School of Medicine, University of California at Davis, US

Abstract

We describe a novel integration method for RA synovial transcriptional profiling to provide predictive insights on drug responses. A normalized compendium consisting of 256 RA synovial samples that cover an intersection of 11,769 genes from 11 datasets was compared with similar datasets derived from OA patients and healthy controls. RA-relevant pathway activation scores and four machine learning classification techniques led to a predictive model of patient treatment response. We identified 876 up-regulated DEGs including 24 known genetic risk factors and 8 drug targets. DEG-based subgrouping revealed 3 distinct RA patient clusters with distinct activity signatures for RA-relevant pathways. In the case of infliximab, we constructed a classifier of drug response that was highly accurate with an AUC/AUPR of 0.92/0.86. Our work argues that the construction and analysis of normalized synovial transcriptomic compendia can provide useful insights for understanding RA-related pathway involvement and drug responses for individual patients.

Keywords

rheumatoid arthritis; gene expression; machine learning; clustering; drug responsiveness

Corresponding author: Ilias Tagkopoulos, PhD, Genome Center, University of California, Davis, California, US, Address: 5313 Genome & Biomedical Science Facility, 451 Health Sciences Drive, University of California, Davis, CA 95616, itagkopoulos@ucdavis.edu.

Author Contributions

K-J Kim, I Tagkopoulos designed the study and K-J Kim acquired the data. M Kim, I Adamopoulos, I Tagkopoulos contributed to the analysis and interpretation of data. K-J Kim conducted all data analysis and drafted the initial manuscript. All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published.

Competing financial interests

The authors have no conflicts of interest to disclose.

1. INTRODUCTION

Rheumatoid arthritis (RA) is a complex autoimmune disease involving a multitude of environmental and genetic factors that exhibit nonlinear dynamic interactions [1]. The disease is characterized by chronic inflammation of the synovium, which results in irreversible damage to the bone tissue over time, leading to pain and joint function impairment. Severity and clinical course of the disease is highly variable across the different patients and hence difficult to predict [1]. Despite the success of tumor necrosis factor (TNF) inhibitors, over 30% of patients do not respond fully to therapy [2]. Moreover, a considerable subset of the patients who showed initial good response experience [2]. A personalized treatment that provides the best possible drug combination for a patient is likely to improve our ability to treat RA and avoid patient relapse. Despite the fact that RA pathophysiology is actively researched, we still have partial understanding regarding the mechanistic basis of disease progression, which is critical to administer personalized and precise care.

In RA, gene expression profiling has been used to gain insights regarding pathogenesis and drug response [3]. Unfortunately, these studies have been conducted in unrelated small sample size cohorts, that exhibit high heterogeneity (sex, age, and ethnicity), differences in technical protocols, microarray platform, and data analysis methods, thus hindering a comprehensive analysis across all available datasets. In addition, most studies have collected samples from whole blood or peripheral blood mononuclear cells, which are easier to acquire but have a limited capacity to adequately reflect local joint inflammation [4–6].

In this study, our aim is to elucidate the various transcriptional and signaling signatures of RA by performing a comprehensive meta-analysis of the publicly available datasets. We focus on the American College of Rheumatology (ACR) classification criteria and analyze exclusively synovial tissue samples to avoid the high false discovery rates coming from blood samples. We have applied several preprocessing and normalization steps to create a cohesive, homogenized compendium of genome-wide gene expression signatures for downstream analysis. We used this compendium to separate expression-driven subgroup, understand the key cellular components in each group and then use genes and pathways with high information value that we have identified to create predictive models for drug responsiveness.

2. METHODS

2.1. Systematic search and data collection

We used the keywords “Rheumatoid Arthritis (RA)”, “Synovium or synovial tissue”, “Transcriptomics or microarray”, “Dataset” in Google Scholar and PubMed to find relevant publications to the topic of synovial gene signatures of patients with rheumatoid arthritis (figure 1). We retrieved all publications that used the American College of Rheumatology (ACR) classification criteria for diagnosis of RA [7] and relevant criteria for OA [8] (20 studies in total). From the resulting set, we removed entries that had been duplicated and selected datasets measuring over 10,000 genes to secure the largest size of genes and samples. Since there was a trade-off between the number of studies to include and the number of genes that are within the intersection from all datasets, we optimized the product of the two by selecting the point where these two trends cross (Supplementary Fig. S1). The final RA sample count was 256, the osteoarthritis (OA) count 41, and 36 normal (NC) samples were included as controls. Clinical characteristics of the RA patients were summarized in Supplementary Table 1. Ultimately, the final RA compendium was constructed out of 11 studies with a total of 333 samples, one per patient, covering 22,721 genes total (common core of 11,769 genes).

2.2. Data normalization and removal of batch effects

For one-channel arrays, the image data was first imported and then the Robust Multi-array Average (RMA) method was applied for a set of replicates for background correction, normalization, probe-set summarization. For dual-channel arrays, the image data were imported and background correction was performed using normexp as it was shown to outperform other methods. Red and green channels were separated and quantile-normalized for each set of replicates. The vectors for the matrices were normalized using the quantile normalization method. Residual technical batch effects arising due to heterogeneous data integration were corrected using the ComBat function within the empirical Bayes package. Quality assurance and distribution bias was evaluated by Principal Component Analysis (Supplementary Fig. S2).

2.3. The RA compendium

After preprocessing, the gene expression profiles have a significant reduction of systematic, dataset-specific bias in comparison with the same dataset before normalization and batch correction (Supplementary Fig. S2). The resulting compendium has a gene size of 11,769 in 333 samples, including 256 RA patients, 41 OA patients, and 36 normal controls. In 105 of the RA samples, synovial tissue sampling was conducted before the start of certain drug: 11 for adalimumab, 62 for infliximab, 8 for methotrexate, 12 for rituximab, and 12 for tocilizumab. For these patients, assessment of disease activity and response was performed per the EULAR response criteria [9] 12–16 weeks after initiation of therapy: 32 were good, 47 were moderate, and 26 were poor responders

2.4. Filtering of differentially expressed genes

In order to identify the differentially expressed genes (DEGs), we employed three widely-used methods: (a) an empirical Bayesian method using the Benjamini and Hochberg procedure with a significance threshold at an adjusted p-value < 0.05; (b) the Significance Analysis of Microarray (SAM) method, with a significance threshold of false discovery rate (FDR) < 0.05; (c) the Rank Products (RP) method with a significance threshold set at percentage of false prediction pfp < 0.05. The resulting list of DEGs is the intersection of the three individual DEGs sets for each method to minimize the FDR statistic.

2.5. Functional enrichment analysis

We performed functional enrichment analysis focusing on the up-regulated DEGs using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) software [10].

Terms were regarded significant if the p-value (EASE score) is lower than 0.05, the enrichment score higher than 1.3, and the fold enrichment was larger than 1.5.

2.6. Gene set enrichment analysis

Gene set enrichment analysis (GSEA) analysis was carried out using the GSEA software from the Broad Institute to assess the overrepresentation of RA-related gene sets [11, 12]. The enrichment results were visualized with the Enrichment Map format, where nodes represent gene-sets and weighted links between the nodes represent an overlap score depending on the number of genes two gene-sets share (Jaccard coefficient) [13]. To intuitively identify redundancies between gene sets, the nodes were connected if their contents overlap by more than 25%. Clusters map to one or more functionally enriched groups, which were manually circled and assigned a label.

2.7. Construction of protein-protein interaction network

To assess the interconnectivity of DEGs in the RA synovium samples, we constructed a protein-protein network based on the interaction data obtained from public databases including BIOGRID [14], HPRD [15], IntAct [16], Reactome [17], and STRING [18]. In the network, nodes and edges represent genes and functional or physical relationships between them, respectively. Graph theory concepts such as degree, closeness, and betweenness were employed to assess the topology of this network. Hub molecules were defined as the shared genes in top 10% with the highest rank in each arm of the three centrality parameters [19].

2.8. Non-negative matrix factorization and determination of the optimal number of clusters

To classify the RA patients into subgroups based on their molecular signatures, we used the non-negative matrix factorization (NMF) method. NMF clustering is a powerful unsupervised approach to identify the disease subtype or patient subgroup and discover biologically meaningful molecular pattern [20, 21]. We applied the consensus NMF clustering method and initialized 100 times for each rank k (range from 2 to 6), where k was a presumed number of subtypes in the dataset. For each k, 100 matrix factorizations were used to classify each sample 100 times. The consensus matrix was used to assess how consistently sample-pairs cluster together. We then computed the cophenetic coefficients and silhouette scores for each k, to quantitatively assess global clustering robustness across the consensus matrix. The maximum peak of the cophenetic coefficient and silhouette score plots determined the optimal number of clusters [20]. To confirm unsupervised clustering results, we used t-distributed stochastic neighborhood embedding (t-SNE) [22], a powerful dimensionality reduction method. The t-SNE method captures the variance in the data by attempting to preserve the distances between data points from high to low dimensions without any prior assumptions about the data distribution.

2.9. Scoring of pathway activation

To quantify certain biological pathway activity, we calculated the gene expression z-scores [21, 23]. Briefly, a Z-score is defined as the difference between the error-weighted mean of the expression values of the genes in each pathway and the error-weighted mean of all genes in a sample after normalization. BCR-, chemokine-, Jack-STAT-, MAPK-, NFκB-, p53-, PI3K-AKT-, RIG-I-like receptor-, Fc ε RI-, TCR-, TGFβ-, TLR-, TNF-, VEGF-, and Wnt signaling pathways and their gene sets were imported from Kyoto Encyclopedia of Genes and Genomes (KEGG) database [24] and IFN type I- and type II signaling pathways and their gene sets referred to Reactome database [17]. *Z*-scores were computed using each pathway in the signature collection for each of the samples, resulting in a matrix of pathway activation scores.

2.10. Supervised learning analyses for the prediction of drug responsiveness

We used Naïve Bayes (NB), Decision Trees (DT), *k*-Nearest-Neighbors (KNN), and Support Vector Machines (SVM) to create drug responsiveness predictors.[25, 26] Each binary SVM was built using Gaussian Radial Basis Function (RBF) kernel and the Sigma hyperparameter was determined from the estimation based upon the 0.1 and 0.9 quantiles of the samples. For soft margins, the C parameter that achieved the best performance was in the range of 2^{-4} to 2^{7}. For KNN, the *k* parameter was tuned in the range 2 to 20. All tuning hyperparameters were separately determined for each bootstrapped training dataset.

To determine the optimal feature set that enables distinguishing ‘good’ from ‘not good’ responders with the highest accuracy according to the EULAR response criteria [9], we employed the wrapper feature selection method [26]. The wrapper method uses the classifier as a black box to rank different subsets of the features according to their predictive power. In the wrapper method, a feature set is fed to the classifier and its performance is scored and the feature set with the highest rank is selected as the optimal feature set. The predictive power of each predictor was assessed through Receiver-Operator Characteristics (ROC) and Precision-Recall (PR) curve [27]. Data was separated into independent training and test sets in a three-to-one sample-size ratio in a way of stratified random sampling. To make up for small sample size and minimize the error, we constructed the pool of resampled dataset by applying bootstrapping with 1000 iterations and subsequently applying a stratified 10-fold cross-validation (CV) for each bootstrapped dataset [25, 26]. Tenfold CV measures the prediction performance in a self-consistent way by systematically leaving out part of the dataset during the training process and testing against those left-out subset of samples. Compared to the test on independent dataset, CV has less bias and better predictive and generalization power. The predictive ability of the models generated from all the approaches was tested by performing the CV test at all the ten locations under study. Given the unequal numbers of trials in each class, balanced accuracy formula was employed to calculate the accuracy [28]. The baseline is estimated by random expectation based on the pre-determined ratio of each condition. In case of infliximab, a probability of 0.29 (18/62) for a “good” and 0.71 (44/62) for a “not good” responder was applied.

2.11. Statistical analysis

For continuous distributed data, between-group comparisons were performed using the one-way ANOVA, unpaired *t*-test or Mann-Whitney *U* test. Categorical or dichotomous variables were compared using the chi-squared test or Fisher’s exact test. To investigate the difference of pathway activation score across the subgroups, we fitted the one-way ANOVA model using logistic regression. All analyses were conducted in *R* (The R Project for Statistical

Computing, www.r-project.org) and R packages used in the analysis and their references were summarized in the Supplementary Table S3.

3. RESULTS

3.1. The RA transcriptomics compendium

To get a list of RA-related DEGs, gene expression profiles of RA patients were compared with samples from the OA and NC groups. We identified 2,762 DEGs for RA versus OA, and 3,087 DEGs for RA versus NC (Fig. 1). Distribution of DEGs was assessed after the DEGs were divided into up- and down-regulated groups (Fig. 2A). The number of up-regulated DEGs was 1,486 for RA versus OA and 1,774 for RA versus NC. The intersection between two up-regulated DEG sets was 876, which we considered as RA-unique (Fig. 2A and **supplementary File S1**).

3.2. Enriched biological processes and protein-to-protein interaction network

We performed a gene-set enrichment analysis [11, 12] where 206 gene ontology processes were identified (Fig. 2B and Supplementary Fig. S3). As expected, immune-related biological processes including adaptive and innate immune response, T and B cell activation and response, and cytokine-related responses, were enriched. These occupied the main positions in the network and closely connected to each other. Among cytokine-related processes, interferon-β (IFN–β), interferon-γ (IFN-γ), interleukin (IL)-4, IL-10, IL-12, IL-17, toll-like receptor (TLR), and TNF-related processes stood out as being substantially more enriched.

Interestingly, several biological processes associated with viral invasion and defense response against viruses were newly identified (Supplementary Fig. S4). Metabolic processes such as calcium ion regulation and protein synthesis/transportation were enriched (all $P<0.01$), suggestive of active intracellular signaling and enhanced protein production and enzyme activity.

Identification of central attractors in the gene and protein network can provide targets for further experimentation and/or drug discovery. For this reason, we constructed the protein-to-protein interaction network of RA (Fig. 2C). We identified 3563 interactions among the 876 DEGs. Thirty-one of DEGs were overlapped with RA genetic susceptibility loci previously discovered [29] (Supplementary Fig. S5) and a total of 56 genes were ranked as hub molecules based on the centrality analysis. The *CD2*, *PTPRC* (protein tyrosine phosphatase, receptor type C, also known as *CD45*), and *PRKCQ* (protein kinase C theta) were RA-susceptible genes having hub position in the network and products of these genes are involved in signal transduction of T cells. Eight genes including primary targets (*JAK2*, *SYK*, *CTLA4*, *MS4A1*) and counterpart receptor molecules (*TNFRSF14*, *TNFRSF17*, *TNFRSF18*, and *IL21R*) of cytokines targeted by the drugs currently in use or under clinical trial or development are also differentially expressed [30, 31]. Interestingly, the targets of small molecule therapeutics, *JAK2* and *SYK* are central hub nodes, in contrast to the targets of biologic agents, such as *CTLA4*, *MS4A1* (also known as *CD20*), *TNFRSF14*, *TNFRSF17*, and *TNFRSF18*. We found 219 RA-associated genes from the DisGeNet database [32], which are genes and variants having an important role in RA pathophysiology. Forty-six of them were overlapped with the RA synovial DEG. To assess topological proximity between RA-associated genes and drug targets in PPI network of synovial DEGs, the shortest distance between nodes was calculated (Supplementary Fig. S6). Mean distance of *JAK2* and *SYK* was 2.11 ± 0.69 S.D. and 2.09 ± 0.68, respectively, and significantly shorter than those of other target molecules (range, 2.65 ~ 3.39) (in all cases $P<0.05$).

3.3. Identification and characterization of molecular subgroups

Next, we assessed whether RA patients can be categories in subgroups based on their expression profiles through consensus non-negative matrix factorization (NMF) clustering [20]. To identify the optimal number of clusters and to assess robustness of the clustering result, we computed the cophenetic coefficient and silhouette score for different numbers of clusters from 2 to 6, where we found that 3 clusters are the optimal representation of the data (Fig. 3A, Supplementary Fig. S7, and **Supplementary Methods**). Segregation of RA subgroups was also reproduced by *t*-distributed stochastic neighborhood embedding (*t*-SNE) and principal component analysis (PCA) (Fig. 3B **and** 3C). To understand the differences among the three clusters, we curated the 17 representative RA-relevant signaling pathways from the result of gene-set enrichment analysis (Fig. 2B) based on the literatures[31, 33–35] and analyzed the activation of individual pathways. As shown in the chord diagram, these pathways are strongly connected, with only TGFβ-, P53-, and Wnt signaling pathways more isolated than others (less shared DEGs). Especially TGFβ- and Wnt, have an opposite trend in their DEG expression (higher in cluster 1, mid in cluster 2 and low in cluster 3), which is the opposite of the trend we observe in most of the other pathways (Fig. 4 and Supplementary Fig. S8). P53 signaling pathways shared fewer genes with other pathways but strongly correlated with BCR-, chemokine-, TCR-, TLR-, and TNF signaling pathways.

While the activation scores of all pathways exhibited significant difference across the various clusters, all clusters exhibited one of the two trends in a statistically significant manner ($P<0.05$ in all cases) and in accordance with the observation through DEG-driven enrichment (all cases except TNF). Compared with RA cluster 2 and 3, RA cluster 1 had moderate activation scores for most of the proinflammatory signaling pathways but high for PI3K-AKT-, TGFβ- and Wnt signaling pathways, which are principally involved in synovial proliferation and tissue remodeling.[36] RA cluster 2 and 3 showed comparable activities for most of the proinflammatory pathways. More active in RA cluster 2 were the P53- and PI3K-AKT signaling pathways, which were reported to play a role in regulating survival of synoviocytes or macrophages [37, 38]. In RA cluster 3, TCR-, Jak-STAT-, and NFκB signaling pathways were more activated and it is noteworthy that IFN signaling pathways were most scored. Cellular processes affected by these pathways are in agreement with the DEG-driven enriched gene ontology (GO) terms in each cluster (Supplementary Fig. S9). This result indicates that there exist RA subgroups representing a distinct mode of inflammation deflected toward a certain combination of signaling pathways (Supplementary Table S4).

3.4. Clinical implication of the 3 molecular subgroups

Next, we examined the relationship between identified 3 subgroups and the pertinent clinical features based on the provided information. There was no difference in gender ratio, age distribution, and tissue sampling method across the subgroups (*P*>0.10 in all cases, see Supplementary Fig. S10). The frequency and distribution of 3 subgroups by seropositivity was estimated on basis of the information available in the 9 datasets (233 samples). Cluster 2 and 3 were predominant in the seropositive, while cluster 1 prevailed in the seronegative (*P*<0.001) (Fig. 5A). Because data on the disease duration and activity were not fully provided for each sample, we compared two distinctively opposing datasets from compendium: the first (GSE45867) includes naïve, untreated RA patients with disease duration of <1 year, moderate disease activity and with arthroscopic needle biopsy performed before methotrexate or tocilizumab therapy [39]. The second (GSE21537) is a cohort of the long-standing RA patients with high disease activity who had failed at least two DMARDs (including methotrexate) and did arthroscopic needle biopsy before infliximab therapy.[40] Disease duration and activity were significantly longer and higher in the latter dataset (all *P*<0.001) while there was no difference in age, gender, and RF positive between two datasets (all *P*>0.10). Distribution of 3 subgroups did not differ between two datasets (*P*=0.754) (Fig. 5B), indicating gene expression pattern by 3 subgroups would be an intrinsic characteristic irrespective of disease duration and activity.

3.5. Towards a predictor of drug response

For 105 RA samples that we had drug effectiveness data, we tested the hypothesis that there is an association between drug responsiveness and cluster membership. Out of the 5 drugs that we had data on (adalimumab, infliximab, methotrexate, rituximab, and tocilizumab) we were not able to identify any such association (Supplementary Fig. S11). Cluster 1 patients had an encouraging response to tocilizumab but at a low statistical significance level (*P*=0.082). In addition to the intricacy of the pertinent pathways, the small size of samples treated by the specific drug, and their potential heterogeneity make the association between drug responsiveness and RA clusters difficult.

Since the differential expression of genes and pathways is at a higher resolution than general clustering signatures, we tested whether drug response can be predicted by using such features. We focused on the patients that were treated with infliximab due to the larger sample size (n=62). To test this hypothesis, we applied outcome to a binary classification (labels of “good” and “not good” responder according to the EULAR response criteria [9]) and tried two approaches: pathway-driven and DEG-driven models. Note that PCA analysis does not reveal separating distributions between the “good” and “not good” responders both for pathway activation score and DEG values (Supplementary Fig. S12).

As features, we used the 17 pathways that are represented by continuous variables through their activation scores (refer to the pathway activation score for each pathway in the **Supplementary File S2**). To reduce the number of dimensions we performed feature selection through recursive elimination (Supplementary Table S5). Based on those results made a predictive model using 4 supervised machine learning methods (NB, DT, KNN, and SVM) for selected key pathway scores and calculated the performance. All models outperformed the baseline (all *P*<0.001) (Fig. 6A, left plot) and SVM, the best performing model, had an average performance AUC (area-under-curve of ROC / AUPR (AUC of PR)) of 0.87/0.78 (all *P*<0.001) (Fig. 6A, middle and right plots). The selected key predictors for SVM model were NFκB-, FcεRI-, TCR-, and TNF signaling pathways. Next, models based on expression values of DEG were fit in order to sort out the informative genes and compare their performance with pathway-driven models. DEG-driven models showed superior performance as compared with pathway-driven models (Fig. 6B, left plot). The AUC of the ROC curves exceeds 0.85 (Fig. 6B, middle and right plots). SVM showed the best performance AUC/AUPR of 0.92/0.86 and with the *HMMR*, *PRPF4B*, *EVI2A*, *RAB27A*, *MALT1*, *SNX6*, and *IFIH1* genes as features. The expression of these genes provide a distinct signature between two different outcomes (*P*<0.05 in all cases, see Supplementary Fig. S13).

4. DISCUSSION

Here, we built the largest RA compendium made by synovial transcriptomes. DEGs extracted from this compendium encompassed the susceptible genes and target molecules. Their topology in the network has opened new possibilities to elucidate biological roles and offer a cue for existing clinical questions. Unbiased cluster analysis of RA compendium resulted in meaningful categories of RA patients with distinct activity for relevant pathways. The pathway-based analysis allowed refinement in our understanding of RA subgroups and it was also feasible to construct pathway- or DEG-driven predictive model for intended treatment by machine learning methods.

Synovial tissues are considerably more difficult samples to obtain, as they are obtained during joint replacement surgery, synovectomy or by arthroscopy at 4–8 sites of the affected joint. However, they are more suitable to understand the mechanism and response to RA, since blood-derived samples are a distant and hence more noisy proxy to the disease, with known quality issues [4–6]. Moreover, to refine the RA-unique genes, we compared RA samples with two control sets (OA and NC groups) and adopted the DEGs shared by three independent methods. We found that 24 of the DEGs are the known RA-associated genetic loci and take a central position in the synovial network. Since functional implications of risk allele were often obscure, it would be helpful to elucidate the biological mechanisms in which risk alleles operate. *STAT1*, a transcription factor downstream of IFN signaling pathway, highlighted as a key molecule in the previous reports [41, 42], was found to be one of the hub genes. Other hub genes, such as *JAK2*, *SYK*, and *BTK* are small molecules that have increasingly drawn attention as novel therapeutic targets following the cytokine-targeting biologics [31]. In contrast, molecules such as TNF receptor molecules, *CTLA4*, *IL6R*, and *MS4A1* were located at the functional periphery of the network although drugs against these molecules are widely used in clinical practice. Moreover, these molecules were placed farther from RA-associated genes than *JAK2* and *SYK* in the network, inferring part of their less potent efficacy in active RA. This was in good harmony with a recent clinical trial that baricitinib, an inhibitor of the Janus kinases *JAK1* and *JAK2*, showed a stronger therapeutic effect as compared with ADLM, a TNF inhibitor [43].

Biological processes and pathways identified from RA compendium show what is happening in the inflamed synovium of RA and are in good line with the previous studies [5, 41, 44]. It is worthy of note that processes concerning viral cycle and anti-viral response were found to be enriched. This could be the internal process analogous to or the vestige of viral infection such as *Chikungunya* virus [45, 46]. A series of studies pointed out activation of IFN-related gene signatures in a subset of RA patients and its substantial similarity to viral infection [5, 41, 44, 46–48] and one reported that the type I IFN signature negatively predicts the clinical response to rituximab treatment in patients with RA [47]. Here, our results suggest that such a probable link between the IFN signature and the anti-viral response may exist [46].

Interestingly, we were able to identify three distinct subgroups through NMF analysis of the RA compendium and they differed in activation level of RA-relevant signaling pathways [20, 21]. Various combinations of molecular perturbations might converge to dysregulation of common pathways and lead to the similar phenotype [49]. Since combinations of genomic perturbations are variable across the patients, pathway- or module-based approaches are desirable for a better understanding of complex inflammatory disease like RA. We looked at the enriched pathways derived from DEGs, which were commensurate with the pathway activation scores calculated from the whole gene list in the compendium. The RA cluster 1 was weighted toward signals regarding synovial proliferation and tissue remodeling (PI3K-AKT-, TGFβ- and Wnt signaling pathways) [36]. RA clusters 2 and 3 showed a strong disposition for proinflammatory signaling pathways (Chemokine-, TNF-, TLR- and VEGF signaling pathways). Apoptosis-related pathways (P53- and PI3K-AKT signaling pathway) were much prominent in RA cluster 2 [37, 38], while BCR-, Jak-STAT-, NFκB-, and TCR signaling pathways were stronger in RA cluster 3. It is known that synoviocytes are the main culprit of invasive synovium and quantitative and qualitative activities of synovial macrophage reflect therapeutic efficacy [50, 51]. They add to the cellular resistance to apoptosis and increase of the potential for proliferation, hence they contribute to the progression and perpetuation of destructive joint inflammation. Therefore, we speculate that an aggressive suppression of pro-inflammatory signals would be better pertinent to RA cluster 3, while therapeutic strategies to control propagation and survival of synoviocytes and macrophages together with anti-inflammatory treatment should be considered in RA cluster 1 and 2 (Supplementary Table S4) [52]. This insight, together with the candidate gene targets for drug development that we have identified in each cluster, may provide good starting points for delivering precision and personalized treatment.

Machine learning has become ubiquitous and indispensable for solving complex problems in most sciences [53]. Since the problem of unresolved heterogeneity is prevalent to medicine, the same methods are expected to open up vast new possibilities in medicine and actively employed in a variety of clinical research [53]. We tried to make a predictive model for 62 samples that were obtained from the synovial tissue of RA patients before administration of IFXM. Because key features are informative for predicting the outcome rather than being directly implicated in the major pathways or usual suspects related to the RA synovium, they could be different depending on drugs and models. The fact that we achieved high performance scores in RA response prediction from mining the RA compendium, despite this was not attainable through individual statistical techniques in the past [40], argues that similar techniques can guide us to narrow choices for more effective drugs. Interestingly, DEG-driven models outperformed models that were relying on pathways as features. Among 7 featured genes in SVM model, HMMR (Hyaluronan-mediated motility receptor, also known as RHAMM) exacerbated collagen-induced arthritis by supporting cell migration and up-regulating genes involved with inflammation[54] and MATL1 (Mucosa associated lymphoid tissue lymphoma translocation gene 1) was recently identified to play a crucial role in the pathogenesis of RA as MATL1-deficient mice were completely resistant to collage-induced arthritis [55]. Direct connection to RA was not revealed for the rest of the identified informative genes so far and it remains to be investigated how and why these features are indicative of drug response.

There are some limitations to be addressed in this study. First, removal of batch effects is not ideal which adds to the noise in the compendium. Second, we did not fully address the association of RA subgroup with clinical factors including age, sex, disease duration, and antibodies against anti-cyclic citrullinated protein due to lack of complete annotation for each RA sample. Third, a limited number of samples were treated with other drugs except for infliximab precluded us from making a predictive model. In general, more meta-data would be desired, although this is to be expected as these studies were performed in different clinical environments, with different procedures and goals, which did not include their aggregation to a single compendium and application of advanced machine learning techniques. In the future, we anticipate that the construction of datasets with sufficient metadata for machine learning analysis would enable critical insights and may lead to novel drug targets for RA treatment.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Abbreviations

BCR	B cell receptor
CV	cross-validation
DAVID	Database for Annotation, Visualization, and Integrated Discovery
DEG	differentially expressed genes
DT	Decision Trees
FDR	false discovery rate
GO	Gene ontology
GSEA	Gene set enrichment analysis
IFN	interferon
IL	interleukin
KEGG	Kyoto Encyclopedia of Genes and Genomes

KNN	k-Nearest-Neighbors
NB	Naïve Bayes
NC	normal control
NMF	non-negative matrix factorization
OA	osteoarthritis
PR	Precision-Recall
RA	Rheumatoid arthritis
ROC	Receiver-Operator Characteristics
SVM	Support Vector Machines
t-SNE	t-distributed stochastic neighborhood embedding
TCR	T cell receptor
TGF	transforming growth factor
TLR	toll-like receptor
TNF	tumor necrosis factor
VEGF	vascular endothelial growth factor

REFERENCES

[1]. Lee DM, Weinblatt ME. Rheumatoid arthritis. Lancet, 2001;358:903–11. [PubMed: 11567728]

[2]. Smolen JS, Aletaha D. Rheumatoid arthritis therapy reappraisal: strategies, opportunities and challenges. Nat Rev Rheumatol, 2015;11:276–89. [PubMed: 25687177]

[3]. Burska AN, Roget K, Blits M, Soto Gomez L, van de Loo F, Hazelwood LD et al. Gene expression analysis in RA: towards personalized medicine. Pharmacogenomics J, 2014;14:93–106. [PubMed: 24589910]

[4]. You S, Cho CS, Lee I, Hood L, Hwang D, Kim WU. A systems approach to rheumatoid arthritis. PLoS One, 2012;7:e51508. [PubMed: 23240033]

[5]. van Baarsen LG, Wijbrandts CA, Timmer TC, van der Pouw Kraan TC, Tak PP, Verweij CL. Synovial tissue heterogeneity in rheumatoid arthritis in relation to disease activity and biomarkers in peripheral blood. Arthritis Rheum, 2010;62:1602–7. [PubMed: 20178127]

[6]. Haupl T, Stuhlmuller B, Grutzkau A, Radbruch A, Burmester GR. Does gene expression analysis inform us in rheumatoid arthritis? Ann Rheum Dis, 2010;69 Suppl 1:i37–42. [PubMed: 19995742]

[7]. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum, 1988;31:315–24. [PubMed: 3358796]

[8]. Altman R, Asch E, Bloch D, Bole G, Borenstein D, Brandt K et al. Development of criteria for the classification and reporting of osteoarthritis. Classification of osteoarthritis of the knee. Diagnostic and Therapeutic Criteria Committee of the American Rheumatism Association. Arthritis Rheum, 1986;29:1039–49. [PubMed: 3741515]

[9]. van Gestel AM, Prevoo ML, van ‘t Hof MA, van Rijswijk MH, van de Putte LB, van Riel PL. Development and validation of the European League Against Rheumatism response criteria for

rheumatoid arthritis. Comparison with the preliminary American College of Rheumatology and the World Health Organization/International League Against Rheumatism Criteria. Arthritis Rheum, 1996;39:34–40. [PubMed: 8546736]

[10]. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc, 2009;4:44–57. [PubMed: 19131956]

[11]. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005;102:15545–50. [PubMed: 16199517]

[12]. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet, 2003;34:267–73. [PubMed: 12808457]

[13]. Merico D, Isserlin R, Stueker O, Emili A, Bader GD. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PLoS One, 2010;5:e13984. [PubMed: 21085593]

[14]. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res, 2006;34:D535–9. [PubMed: 16381927]

[15]. Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B et al. Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res, 2004;32:D497–501. [PubMed: 14681466]

[16]. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F et al. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res, 2014;42:D358–63. [PubMed: 24234451]

[17]. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R et al. The Reactome pathway Knowledgebase. Nucleic Acids Res, 2016;44:D481–7. [PubMed: 26656494]

[18]. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res, 2017;45:D362–d8. [PubMed: 27924014]

[19]. Koschutzki D, Schreiber F. Centrality analysis methods for biological networks and their application to gene regulatory networks. Gene Regul Syst Bio, 2008;2:193–201.

[20]. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A, 2004;101:4164–9. [PubMed: 15016911]

[21]. You S, Knudsen BS, Erho N, Alshalalfa M, Takhar M, Al-Deen Ashab H et al. Integrated Classification of Prostate Cancer Reveals a Novel Luminal Subtype with Poor Outcome. Cancer Res, 2016;76:4948–58. [PubMed: 27302169]

[22]. Maaten LVD, Hinton GE. Visualizing Data using t-SNE. J Machine Learning Res, 2008;9:2579–605.

[23]. Levine DM, Haynor DR, Castle JC, Stepaniants SB, Pellegrini M, Mao M et al. Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways. Genome Biol, 2006;7:R93. [PubMed: 17044931]

[24]. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res, 2017;45:D353–d61. [PubMed: 27899662]

[25]. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Springer; 2013.

[26]. Kuhn M, Johnson K. Applied predictive modeling. Springer; 2013.

[27]. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One, 2015;10:e0118432. [PubMed: 25738806]

[28]. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The Balanced Accuracy and Its Posterior Distribution. 2010 20th International Conference on Pattern Recognition; 2010, p. 3121–4.

[29]. Yamamoto K, Okada Y, Suzuki A, Kochi Y. Genetics of rheumatoid arthritis in Asia--present and future. Nat Rev Rheumatol, 2015;11:375–9. [PubMed: 25668139]

[30]. Koenders MI, van den Berg WB. Novel therapeutic targets in rheumatoid arthritis. Trends Pharmacol Sci, 2015;36:189–95. [PubMed: 25732812]

[31]. Kelly V, Genovese M. Novel small molecule therapeutics in rheumatoid arthritis. Rheumatology (Oxford), 2013;52:1155–62. [PubMed: 23297340]

[32]. Pinero J, Bravo A, Queralt-Rosinach N, Gutierrez-Sacristan A, Deu-Pons J, Centeno E et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res, 2017;45:D833–d9. [PubMed: 27924018]

[33]. Choy E. Understanding the dynamics: pathways involved in the pathogenesis of rheumatoid arthritis. Rheumatology (Oxford), 2012;51 Suppl 5:v3–11. [PubMed: 22718924]

[34]. Rabelo Fde S, da Mota LM, Lima RA, Lima FA, Barra GB, de Carvalho JF et al. The Wnt signaling pathway and rheumatoid arthritis. Autoimmun Rev, 2010;9:207–10. [PubMed: 19683077]

[35]. Smolen JS, Aletaha D, Barton A, Burmester GR, Emery P, Firestein GS et al. Rheumatoid arthritis. Nat Rev Dis Primers, 2018;4:18001. [PubMed: 29417936]

[36]. Miao CG, Yang YY, He X, Li XF, Huang C, Huang Y et al. Wnt signaling pathway in rheumatoid arthritis, with special emphasis on the different roles in synovial inflammation and bone remodeling. Cell Signal, 2013;25:2069–78. [PubMed: 23602936]

[37]. Pope RM. Apoptosis as a therapeutic tool in rheumatoid arthritis. Nat Rev Immunol, 2002;2:527–35. [PubMed: 12094227]

[38]. Smith MD, Walker JG. Apoptosis a relevant therapeutic target in rheumatoid arthritis? Rheumatology (Oxford), 2004;43:405–7. [PubMed: 14679296]

[39]. Ducreux J, Durez P, Galant C, Nzeusseu Toukap A, Van den Eynde B, Houssiau FA et al. Global molecular effects of tocilizumab therapy in rheumatoid arthritis synovium. Arthritis Rheumatol, 2014;66:15–23. [PubMed: 24449571]

[40]. Lindberg J, Wijbrandts CA, van Baarsen LG, Nader G, Klareskog L, Catrina A et al. The gene expression profile in the synovium as a predictor of the clinical response to infliximab treatment in rheumatoid arthritis. PLoS One, 2010;5:e11310. [PubMed: 20593016]

[41]. van der Pouw Kraan TC, van Gaalen FA, Kasperkovitz PV, Verbeet NL, Smeets TJ, Kraan MC et al. Rheumatoid arthritis is a heterogeneous disease: evidence for differences in the activation of the STAT-1 pathway between rheumatoid tissues. Arthritis Rheum, 2003;48:2132–45. [PubMed: 12905466]

[42]. Yoshida S, Arakawa F, Higuchi F, Ishibashi Y, Goto M, Sugita Y et al. Gene expression analysis of rheumatoid arthritis synovial lining regions by cDNA microarray combined with laser microdissection: up-regulation of inflammation-associated STAT1, IRF1, CXCL9, CXCL10, and CCL5. Scand J Rheumatol, 2012;41:170–9. [PubMed: 22401175]

[43]. Taylor PC, Keystone EC, van der Heijde D, Weinblatt ME, Del Carmen Morales L, Reyes Gonzaga J et al. Baricitinib versus Placebo or Adalimumab in Rheumatoid Arthritis. N Engl J Med, 2017;376:652–62. [PubMed: 28199814]

[44]. Woetzel D, Huber R, Kupfer P, Pohlers D, Pfaff M, Driesch D et al. Identification of rheumatoid arthritis and osteoarthritis patients by transcriptome-based rule set generation. Arthritis Res Ther, 2014;16:R84. [PubMed: 24690414]

[45]. Miner JJ, Aw Yeang HX, Fox JM, Taffner S, Malkova ON, Oh ST et al. Chikungunya viral arthritis in the United States: a mimic of seronegative rheumatoid arthritis. Arthritis Rheumatol, 2015;67:1214–20. [PubMed: 25605621]

[46]. Nakaya HI, Gardner J, Poo YS, Major L, Pulendran B, Suhrbier A. Gene profiling of Chikungunya virus arthritis in a mouse model reveals significant overlap with rheumatoid arthritis. Arthritis Rheum, 2012;64:3553–63. [PubMed: 22833339]

[47]. Thurlings RM, Boumans M, Tekstra J, van Roon JA, Vos K, van Westing DM et al. Relationship between the type I interferon signature and the response to rituximab in rheumatoid arthritis patients. Arthritis Rheum, 2010;62:3607–14. [PubMed: 20722020]

[48]. van der Pouw Kraan TC, Wijbrandts CA, van Baarsen LG, Voskuyl AE, Rustenburg F, Baggen JM et al. Rheumatoid arthritis subtypes identified by genomic profiling of peripheral blood cells: assignment of a type I interferon signature in a subpopulation of patients. Ann Rheum Dis, 2007;66:1008–14. [PubMed: 17223656]

[49]. Kim YA, Wuchty S, Przytycka TM. Identifying causal genes and dysregulated pathways in complex diseases. PLoS Comput Biol, 2011;7:e1001095. [PubMed: 21390271]

[50]. Bottini N, Firestein GS. Duality of fibroblast-like synoviocytes in RA: passive responders and imprinted aggressors. Nat Rev Rheumatol, 2013;9:24–33. [PubMed: 23147896]

[51]. Haringman JJ, Gerlag DM, Zwinderman AH, Smeets TJ, Kraan MC, Baeten D et al. Synovial tissue macrophages: a sensitive biomarker for response to treatment in patients with rheumatoid arthritis. Ann Rheum Dis, 2005;64:834–8. [PubMed: 15576415]

[52]. Martinez-Lostao L, Garcia-Alvarez F, Basanez G, Alegre-Aguaron E, Desportes P, Larrad L et al. Liposome-bound APO2L/TRAIL is an effective treatment in a rabbit model of rheumatoid arthritis. Arthritis Rheum, 2010;62:2272–82. [PubMed: 20506326]

[53]. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med, 2016;375:1216–9. [PubMed: 27682033]

[54]. Nedvetzki S, Gonen E, Assayag N, Reich R, Williams RO, Thurmond RL et al. RHAMM, a receptor for hyaluronan-mediated motility, compensates for CD44 in inflamed CD44-knockout mice: a different interpretation of redundancy. Proc Natl Acad Sci U S A, 2004;101:18081–6. [PubMed: 15596723]

[55]. Gilis E, Staalj J, Beyaert R, Elewaut D. The Paracaspase MALT1 Plays a Central Role in the Pathogenesis of Rheumatoid Arthritis [Abstract]. Arthritis Rheumatology, 2017;69.

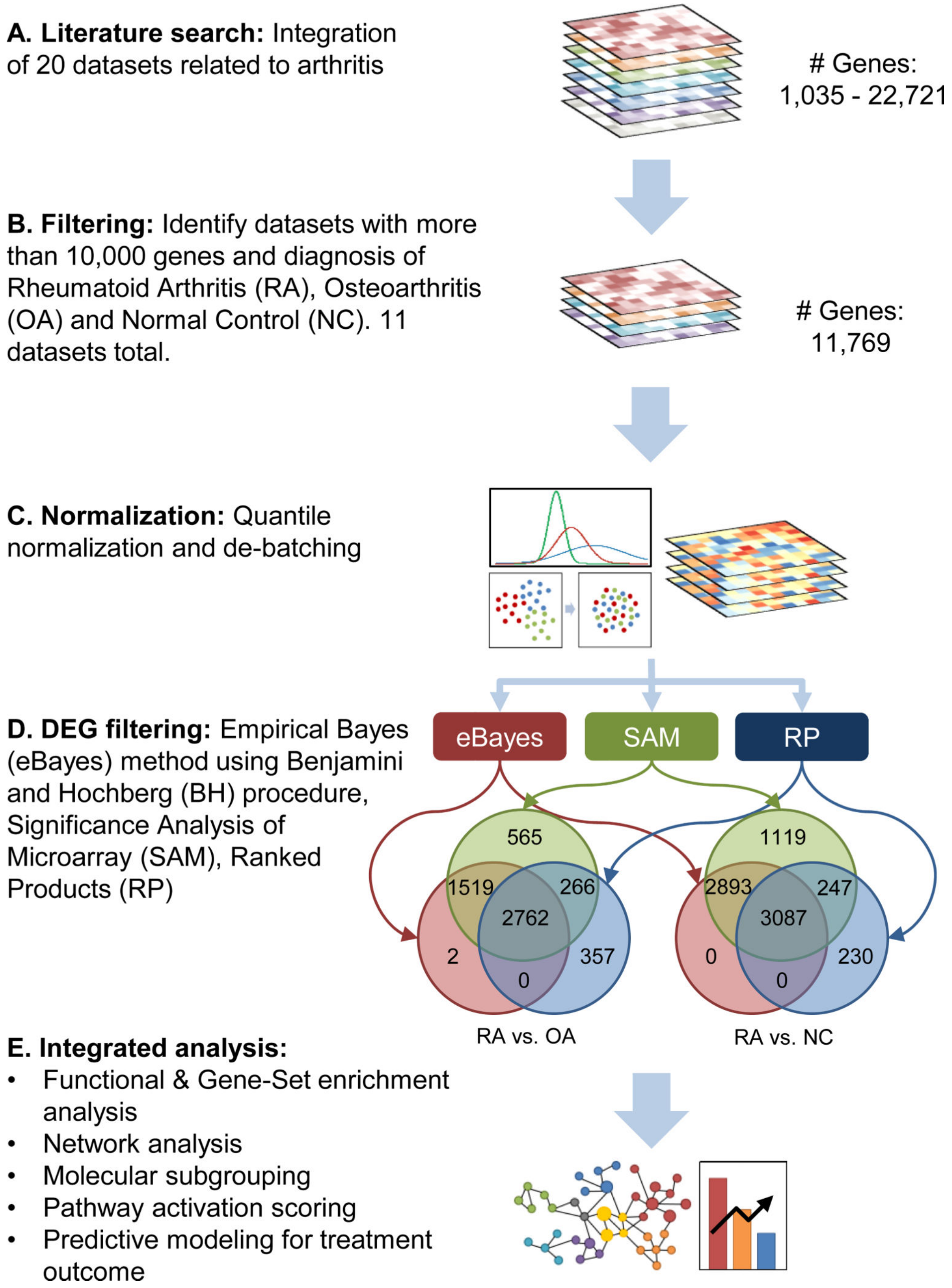

A. Literature search: Integration of 20 datasets related to arthritis
# Genes: 1,035 - 22,721
B. Filtering: Identify datasets with more than 10,000 genes and diagnosis of Rheumatoid Arthritis (RA), Osteoarthritis (OA) and Normal Control (NC). 11 datasets total.
# Genes: 11,769
C. Normalization: Quantile normalization and de-batching
D. DEG filtering: Empirical Bayes (eBayes) method using Benjamini and Hochberg (BH) procedure, Significance Analysis of Microarray (SAM), Ranked Products (RP)
eBayes
SAM
RP
565
1519
266
2762
2
357
0
RA vs. OA
1119
2893
247
3087
0
230
0
RA vs. NC
E. Integrated analysis:
• Functional & Gene-Set enrichment analysis
• Network analysis
• Molecular subgrouping
• Pathway activation scoring
• Predictive modeling for treatment outcome

Fig. 1. Overview of the data processing steps.

(A) Twenty studies maximally covering 20,511 genes were retrieved from the literature. **(B)** Selected were 11 datasets adequate to integrated analysis, which included 256 RA, 41 OA, and 36 NC samples covering 11,769 gene. **(C)** The merged dataset was normalized using quantile method and its batch effect was corrected. **(D)** DEG of RA compared to OA or NC were obtained using three methods, eBayes, SAM, and RP. Intersection of three DEG sets was chosen as significant DEG. The number of DEG was 2762 in RA versus OA and 3087 in RA versus NC. (E) A list of strategies for integrated analysis. (Abbreviation: RA, rheumatoid arthritis; OA, osteoarthritis; NC, normal controls; DEG, differentially expressed genes; eBayes, empirical Bayes; SAM, significance analysis of microarray; RP, rank products).

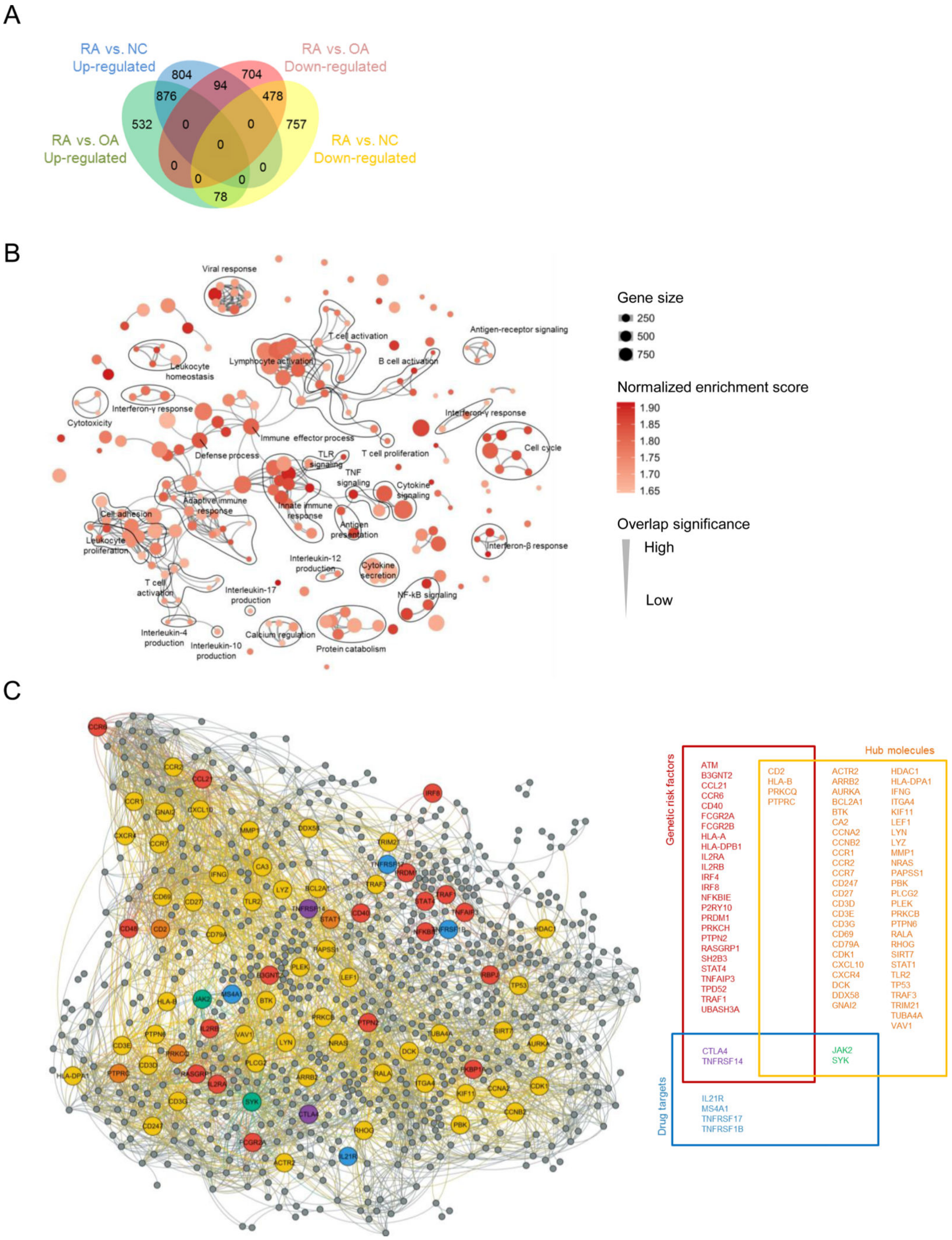


Fig. 2. Differentially expressed genes and their functional network.

(A) Venn diagram showing the overlap of up- and down-regulated DEG between RA versus OA and RA versus NC. **(B)** Gene-Set enrichment map for up-regulated DEG. Nodes represent GO-termed gene-sets. Their color intensity and size is proportional to the enrichment significance and the gene size, respectively. Edge thickness represents the degree of overlap between gene sets and only edges with a Jaccard coefficient larger than 0.25 were visualized. Clusters of functionally related gene-sets were manually curated based on the GO parent-child hierarchy and assigned a label. **(C)** Protein-Protein interaction network of up-regulated DEG. Red and blue nodes indicate the known RA-susceptible genes and drug target molecules, respectively. Drug targets were defined subject to the targets of drugs currently in use or under clinical trial and development. Yellow nodes correspond to the hub molecules, which are determined as the shared genes in top 10% with the highest rank in each arm of three centrality parameters; degree, closeness, and betweenness. Orange, green, and purple colored-nodes are the overlapped between red and yellow, yellow and blue, and red and blue ones, respectively. Right-side inset box is the schematic diagram of the interesting genes.

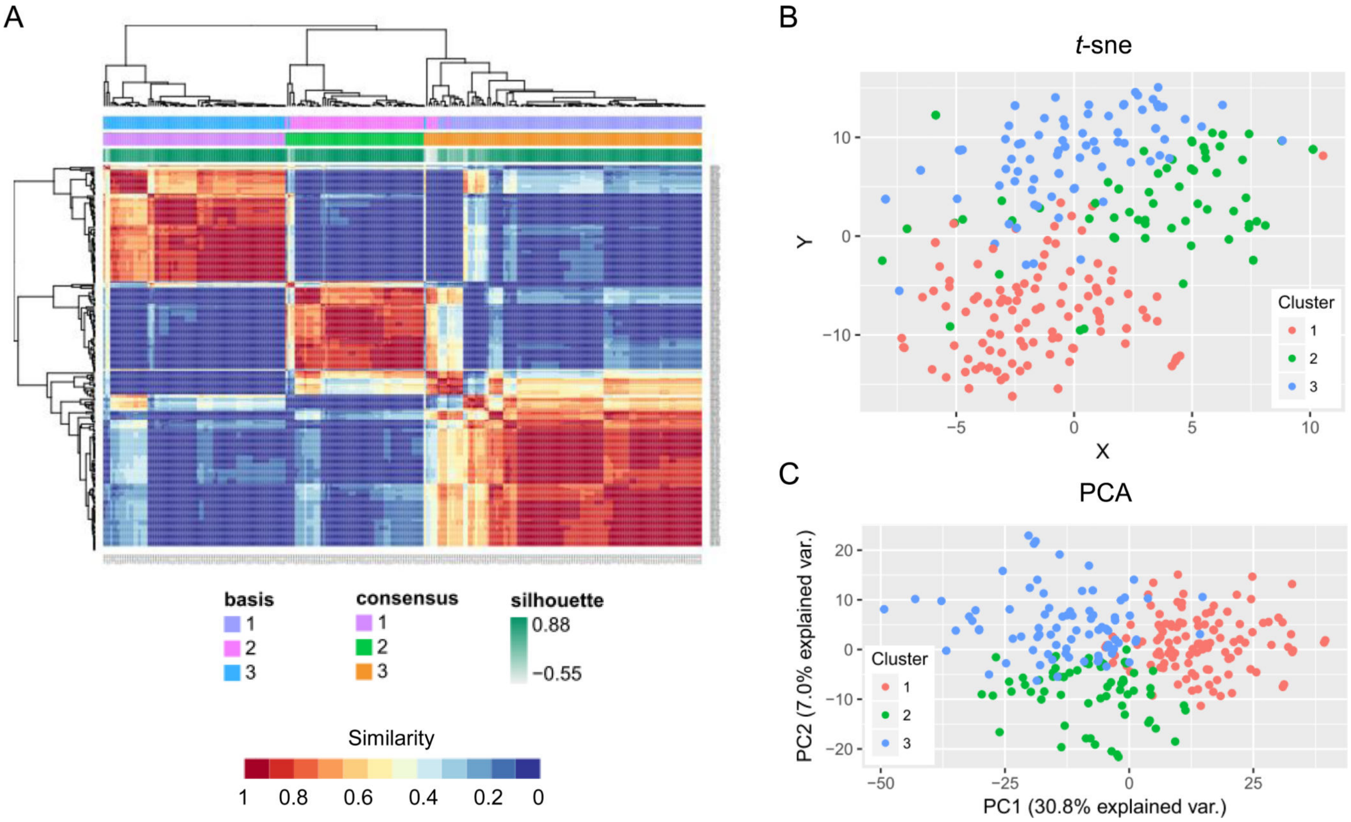


Fig. 3. Identification of novel RA subgroups according to synovial signatures.

(A) Reordered consensus matrices on RA compendium. The samples were clustered using average linkage and 1-correlation distances. Deep-red color indicates perfect agreement of the solution, whilst blue color indicates no agreement (Right-side color bar). Basis and consensus represent clusters based on the basis and consensus matrices, respectively. The silhouette score is a similarity measure within its own cluster compared to other clusters. **(B)** *t*-SNE and **(C)** PCA reduces the dimensions of a multivariate dataset. Each data point is assigned a location in a two-dimensional map to illustrate potential clusters of neighboring samples, which contain similar gene expression patterns.

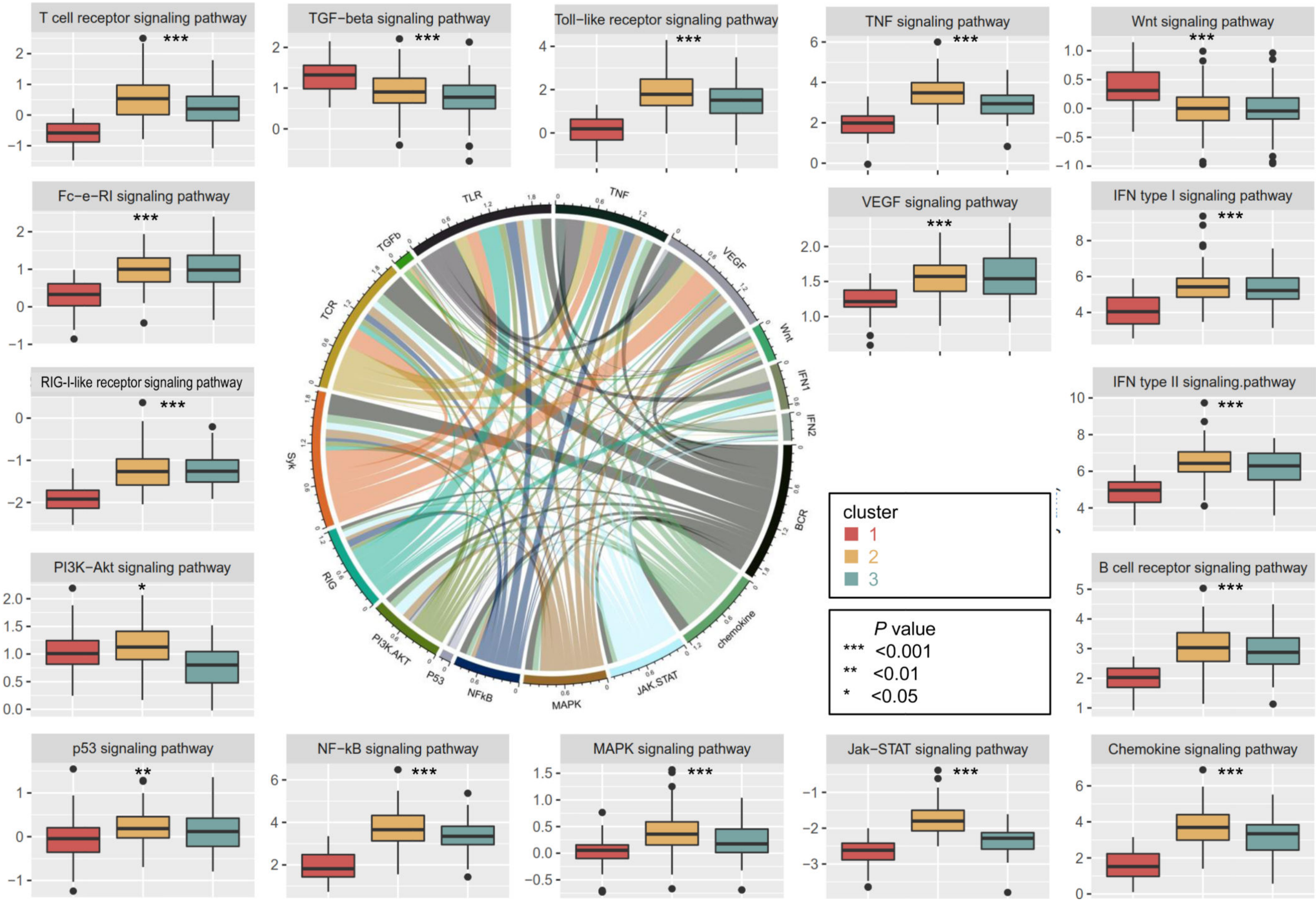
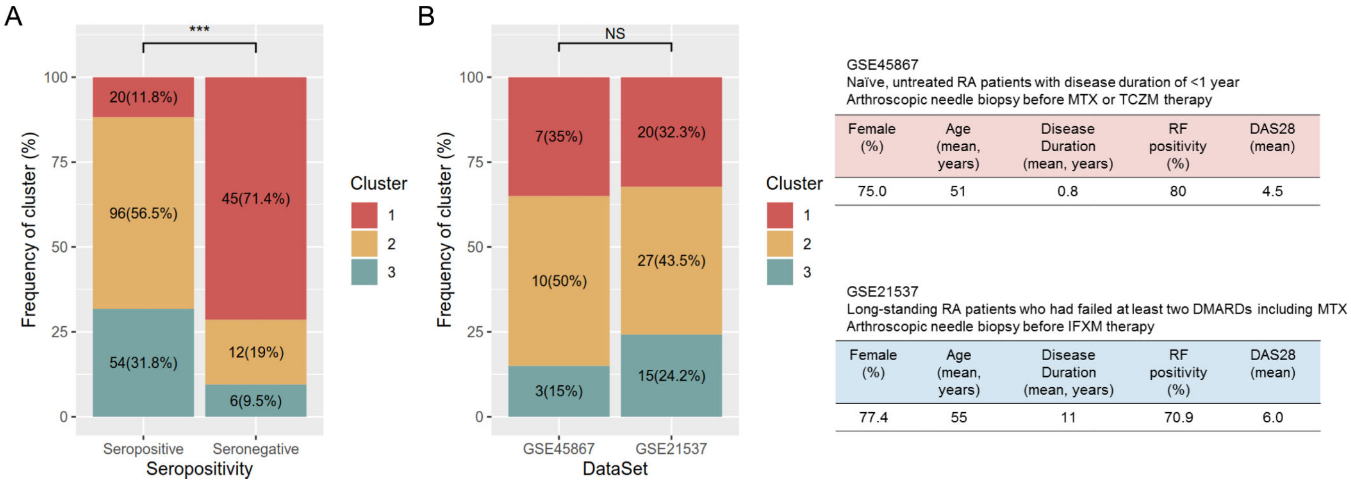


Fig. 4. Pathway activation scores according to RA subgroups.

Chord diagram shows interrelationship among pathways and link thickness is proportional to the overlap between two pathways, calculated using the Jaccard coefficients. Turkey boxplots reveals pathway activation scores across the RA subgroups and ANOVA test was used to analyze the differences among groups. *, *P*<0.05; **, *P*<0.01; ***, *P*<0.001.



GSE45867
Naïve, untreated RA patients with disease duration of <1 year
Arthroscopic needle biopsy before MTX or TCZM therapy

Female (%)	Age (mean, years)	Disease Duration (mean, years)	RF positivity (%)	DAS28 (mean)
75.0	51	0.8	80	4.5

GSE21537
Long-standing RA patients who had failed at least two DMARDs including MTX
Arthroscopic needle biopsy before IFXM therapy

Female (%)	Age (mean, years)	Disease Duration (mean, years)	RF positivity (%)	DAS28 (mean)
77.4	55	11	70.9	6.0

Fig. 5.

(A) Frequency and distribution of 3 subgroups by seropositivity. Estimation was on basis of the information available in the 9 datasets (233 samples). **(B) Frequency and distribution of 3 subgroups by the two-opposing datasets.** To examine the association of disease duration and activity, two distinctively opposing datasets were selected from the compendium for comparison. GSE45867 is a group of samples with shorter duration and moderate disease activity and GSE21537 is with longer duration and high disease activity. The number of samples assigned by subgroups and characteristics of the dataset was summarized. Distribution of 3 subgroups did not differ between two datasets (*P*=0.754).

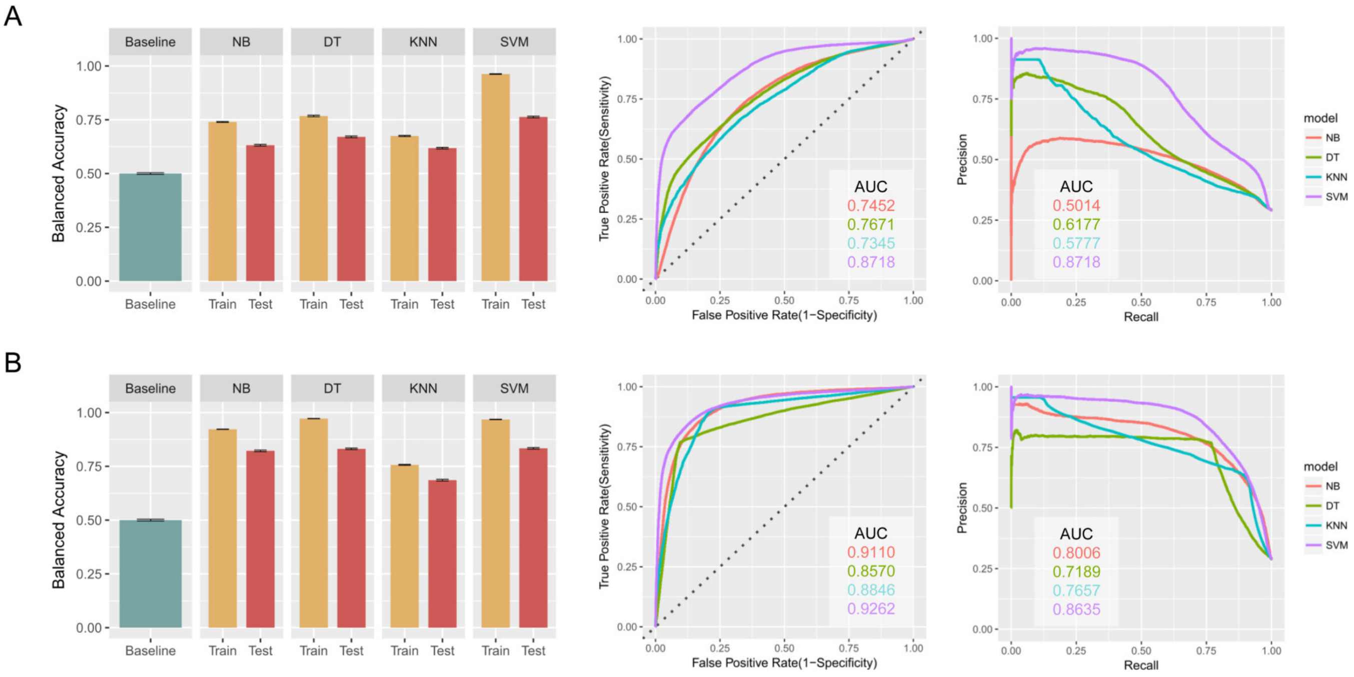


Fig. 6. Predictive models and their performance.

(**a**) **Pathway-driven models**. (**b**) **DEG-driven models**. (Left plot) The training and testing balanced accuracy for each classifier as compared with the baseline. All models outperformed the baseline (all *P*<0.001) and the performance of the trained models was significantly compromised in testing sets (all *P*<0.001). (Middle and right plots) Averaged ROC and PR curves showing the performance of each classifier.