Analysis of gene expression in rheumatoid arthritis and related conditions offers insights into sex-bias, gene biotypes and co-expression patterns

Alexander Platzer, Thomas Nussbaumer, Thomas Karonitsch, Josef S. Smolen, Daniel Aletaha

Angaben zur Veröffentlichung / Publication details:

Platzer, Alexander, Thomas Nussbaumer, Thomas Karonitsch, Josef S. Smolen, and Daniel Aletaha. 2019. “Analysis of gene expression in rheumatoid arthritis and related conditions offers insights into sex-bias, gene biotypes and co-expression patterns.” Edited by Pierre Bobé. *PLOS ONE* 14 (7): 219698. https://doi.org/10.1371/journal.pone.0219698.

RESEARCH ARTICLE

Analysis of gene expression in rheumatoid arthritis and related conditions offers insights into sex-bias, gene biotypes and co-expression patterns

Alexander Platzer[1]*, **Thomas Nussbaumer**[2,3], **Thomas Karonitsch**[1], **Josef S. Smolen**[1], **Daniel Aletaha**[1]

[1] Division of Rheumatology, Department of Medicine III, Medical University of Vienna, Vienna, Austria, [2] Chair and Institute of Environmental Medicine, UNIKA-T, Technical University and Helmholtz Zentrum München, Augsburg, Germany, [3] Institute of Network Biology (INET), Helmholtz Center Munich, Neuherberg, Germany

* alexander.platzer@meduniwien.ac.at

Abstract

The era of next-generation sequencing has mounted the foundation of many gene expression studies. In rheumatoid arthritis research, this has led to the discovery of important candidate genes which offered novel insights into mechanisms and their possible roles in the cure of the disease. In the last years, data generation has outstripped data analysis and while many studies focused on specific aspects of the disease, a global picture of the disease is not yet accomplished. Here, we analyzed and compared a collection of gene expression information from healthy individuals and from patients suffering under different arthritis conditions from published studies containing the following clinical conditions: early and established rheumatoid arthritis, osteoarthritis and arthralgia. We show comprehensive overviews of this data collection and give new insights specifically on gene expression in the early stage, into sex-dependent gene expression, and we describe general differences in expression of different biotypes of genes. Many genes that are related to cytoskeleton changes (actin filament related genes) are differently expressed in early rheumatoid arthritis in comparison to healthy subjects; interestingly, eight of these genes reverse their expression ratio significantly between men and women compared early rheumatoid arthritis and healthy subjects. There are some slighter changes between men and woman between the conditions early and established rheumatoid arthritis. Another aspect are miRNAs and other gene biotypes which are not only promising candidates for diagnoses but also change their expression grossly in average at rheumatoid arthritis and arthralgia compared to the healthy condition. With a selection of intersecting genes, we were able to generate simple classification models to distinguish between healthy and rheumatoid arthritis as well as between early rheumatoid arthritis to other arthritides based on gene expression.

Citation: Platzer A, Nussbaumer T, Karonitsch T, Smolen JS, Aletaha D (2019) Analysis of gene expression in rheumatoid arthritis and related conditions offers insights into sex-bias, gene biotypes and co-expression patterns. PLoS ONE 14 (7): e0219698. https://doi.org/10.1371/journal.pone.0219698

Editor: Pierre Bobé, Universite Paris-Sud, FRANCE

Received: January 24, 2019

Accepted: June 28, 2019

Published: July 25, 2019

Data Availability Statement: Data/supplementary archive is available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.315v28r.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Rheumatoid arthritis (RA) is a chronic, complex, systemic, multifactorial disease [1, 2] with a prevalence of 0.3–1% in the population worldwide [3], affecting women 2–3 times more often than men. The proven or at least strongly suspected etiopathogenetic factors include genomic variations [4], gene expression changes [5], autoimmunity [6] and environmental factors [7]. No factor is considered as single cause, except for the (currently unknown) cause of the first insult leading to the autoimmune inflammation characteristic of RA. Likely, there is not one single cause of RA, no single path to progression and no single curative approach, as ultimate success rates of single therapies are limited [8, 9]. This has led to the hypothesis that there might be RA subtypes with different RA disease manifestations that are dependent on sex, genotype, gene expression or on the composition of the microbiome, which would make RA an important showcase for personalized medicine.

A lot of effort has been undertaken in finding or assessing specific genes and pathways of importance for the progression of RA [10–13]. Less effort has been spent so far to obtain a more complete view of the complete genome and expression data. The analysis of whole genome sequencing data of RA is covered in broader GWAS approaches (e.g. [4, 14], where several associated SNPs were reported), while gene expression data from RNA-seq is more left open for the broader view and was generated for particular research questions [15, 16]. Some broader views of gene expression are published based on microarray data [17–19], as well as some narrower views based on quantitative polymerase chain reactions (PCRs) of several genes [20–22]. The prevalence of RA is lower in men than in women [23–26], but it is unclear whether this is also related to gene expression; reported relations with sex hormones such as estrogen and androgen would be supportive of this hypotheses [27]. If such major differences of gene expression between men and women exist specifically in RA compared to healthy subjects, these genes might be targets for further investigation as there might be sex-specific issues beside the prevalence.

Some miRNAs have been also reported as related to RA, with the main motivation to use them as diagnostic markers for RA [28, 29]. Despite the different alternatives for the initial starting points of RA and potential diagnostic markers, there is a consensus regarding the center of amplification and perpetuation of joint inflammation: the synovial tissue [30, 31]. Uncontrolled persistent inflammation of the synovial membrane leads to progressive joint damage and disability [2]. For this reason, we focus the present analyses on gene expression (RNA-seq) data from synovial tissue. We use published studies with large amounts of RNA-seq data in populations of early and established RA, as well as in patients with related diagnoses [15, 16].

Results

Clustering

Clustering of subjects and conditions. We applied different clustering and dimension reduction methods to obtain a comprehensive view of the transcriptome data from the 236 RNA-seq synovial biopsy samples. A PCA is shown in Fig 1, divided in panels, where only some conditions are shown (because of large overlaps; all conditions together are shown in S1 Fig). On a high-level view, gene expression of healthy subjects is quite different compared to the non-healthy conditions (Fig 1D; classification model accuracy of 95%, p-value of separation between healthy and non-healthy (excluding OA) is $4.4^{*}10^{-18}$ when the coordinates of the first two components of the PCA and the labels for the samples were taken and treated as a classification problem for the tree learner learner C4.5 [32]). Established RA is quite broad and

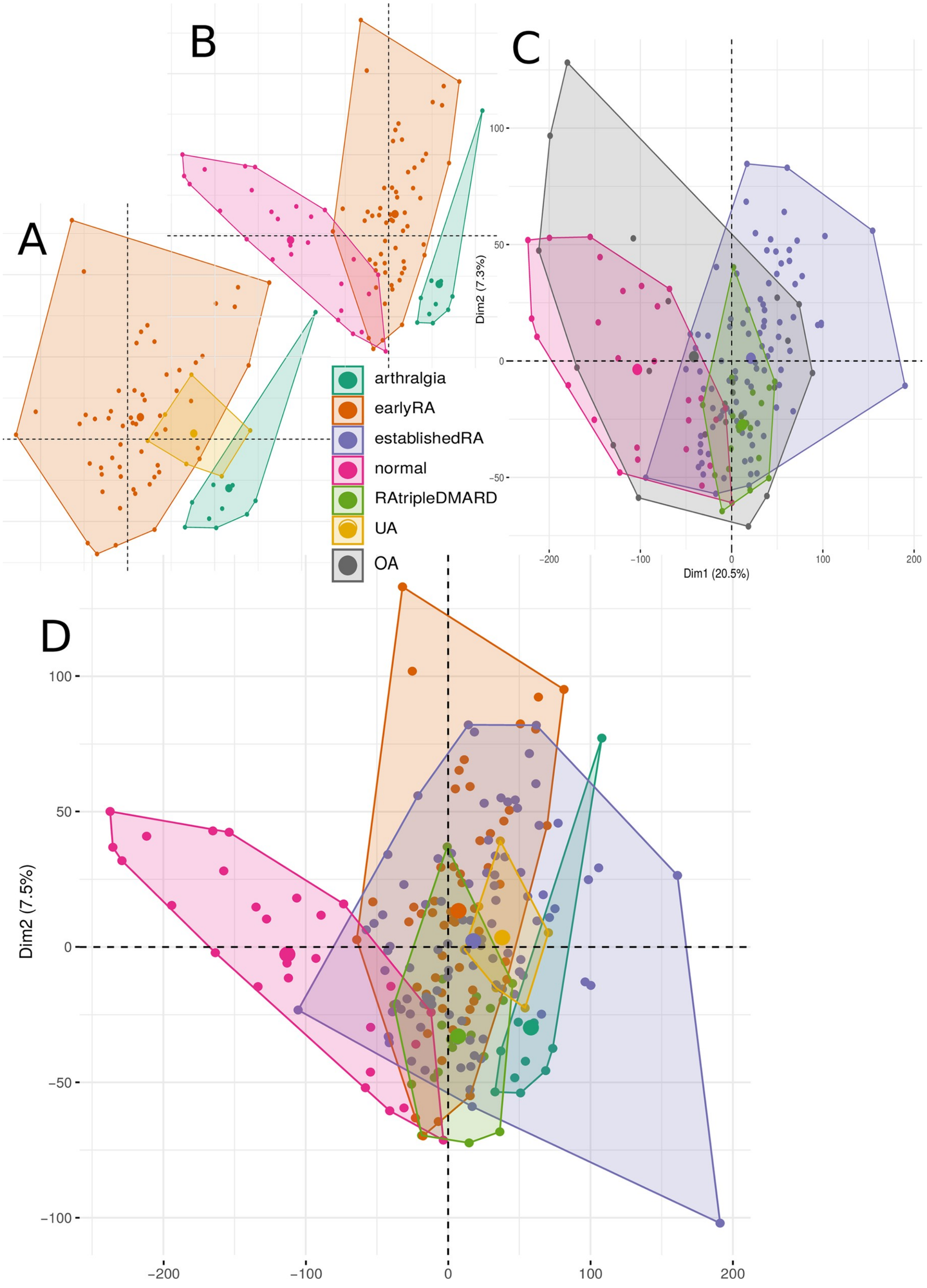


Fig 1. The first two principal components of the PCA based on the RPKMs of the coding genes. The areas are the convex hulls of the conditions. The largest point of one color depicts the center of a hull. A, B, and D are the same PCA analysis with the same coordinates,

where in D all conditions except OA are visible, in A and B only three of them for a better overview. C is a PCA with OA, where four conditions are shown to depict the variability of OA. Number of samples: 10 arthralgia, 57 earlyRA, 95 establishedRA, 27 normal, 22 OA, 19 RAtripleDMARD and 6 UA.

https://doi.org/10.1371/journal.pone.0219698.g001

overlaps with all other groups. Arthralgia is clearly different from early RA as the convex hull is not overlapping at all (Fig 1). Different clustering approaches do not entail more insights than PCA (Conformal Eigenmaps [33], multidimensional scaling [34] and Sammon mapping [35], see S2–S4 Figs). Fig 2 shows the same overview for the conditions when split by sex and by only using the three conditions where more than ten samples for men and women exist (conditions early RA, established RA and healthy; for an overview of the first 10 principal components see S5 Fig). There is not much visible difference between men and women, as all conditions are highly overlapping in the first two principal components whether samples originate from male or female individuals.

Looking into clustering within one diagnosis reveals potential subtypes (PCA and hierarchical clustering in Supplementary Archive FiguresClusteringWithinClasses/). For example, there is a kind of female cluster within early RA (Supplementary Fig 'earlyRA_hclust'). When filtering for genes, which are significantly differently expressed in at least one comparison to normal, the similarity of conditions based on the fold-change can be assessed (S6 Fig). OA is there located next to samples from healthy individuals, a strong treatment (RAtripleDMARD) constitute the next neighbor and early/established RA conditions group apart. When restricting to miRNAs, the picture is similar: only the closer conditions are a little bit rearranged (S7 Fig). Hierarchical clustering instead of neighbor joining and PCA are shown in S8 Fig for coding genes and in S9 Fig for miRNAs.

Gene enrichment analysis

In order to better understand the differences in gene functions between the different conditions, we performed a gene enrichment analysis. We looked for GO [36, 37], KEGG [38] and REACTOME [39] enriched terms in the significantly differentially expressed genes between the clinical conditions and various derived gene lists. This includes all gene lists used and generated in this article (see in the Supplementary Archive tables/ and geneSets/ for all enrichments and gene lists). An overview of GO (BP) term enrichments of the comparisons of normal to earlyRA, arthralgia, OA and undifferentiated arthritis is shown in Fig 3. More details about tools and strict filtering settings needed for a diagram fitting onto a single page are in the method section. The GO terms in a larger font therein were selected for their specifity for earlyRA and meaningfulness. For example, in the upper right cluster, the term 'vesicle-mediated transport' might be interesting, but is enriched in the up-regulated genes of all four conditions. The term 'cell activation' is specifically enriched in up-regulated genes in earlyRA, but the term itself is rather nonspecific. Taken the GO terms of DEGs in earlyRA together, there is specifically more expression for chromatin (lower right in Fig 3), coagulation factors (as also reported in several articles [40–42]), less activity of polymerase II (as can also be seen in section 'Different gene expression at different gene biotypes'), less muscle cell activity (see section 'Different RA gene expression in men and women' for a more detailed different view on that) and more antigen presentation (left side in Fig 3). Other patterns in this view are also interesting, like enrichment specifically for earlyRA and undifferentiated arthritis as these conditions are clinically quite close. For example, the Gene Ontology terms 'biological adhesion', 'regulation of cell-cell adhesion' and 'immunoglobulin production' are enriched in earlyRA and UA, but not in OA and arthralgia (left in Fig 3). Unfortunately, there are only few samples for undifferentiated arthritis, which weakens the hints from these patterns. The complete enrichment lists

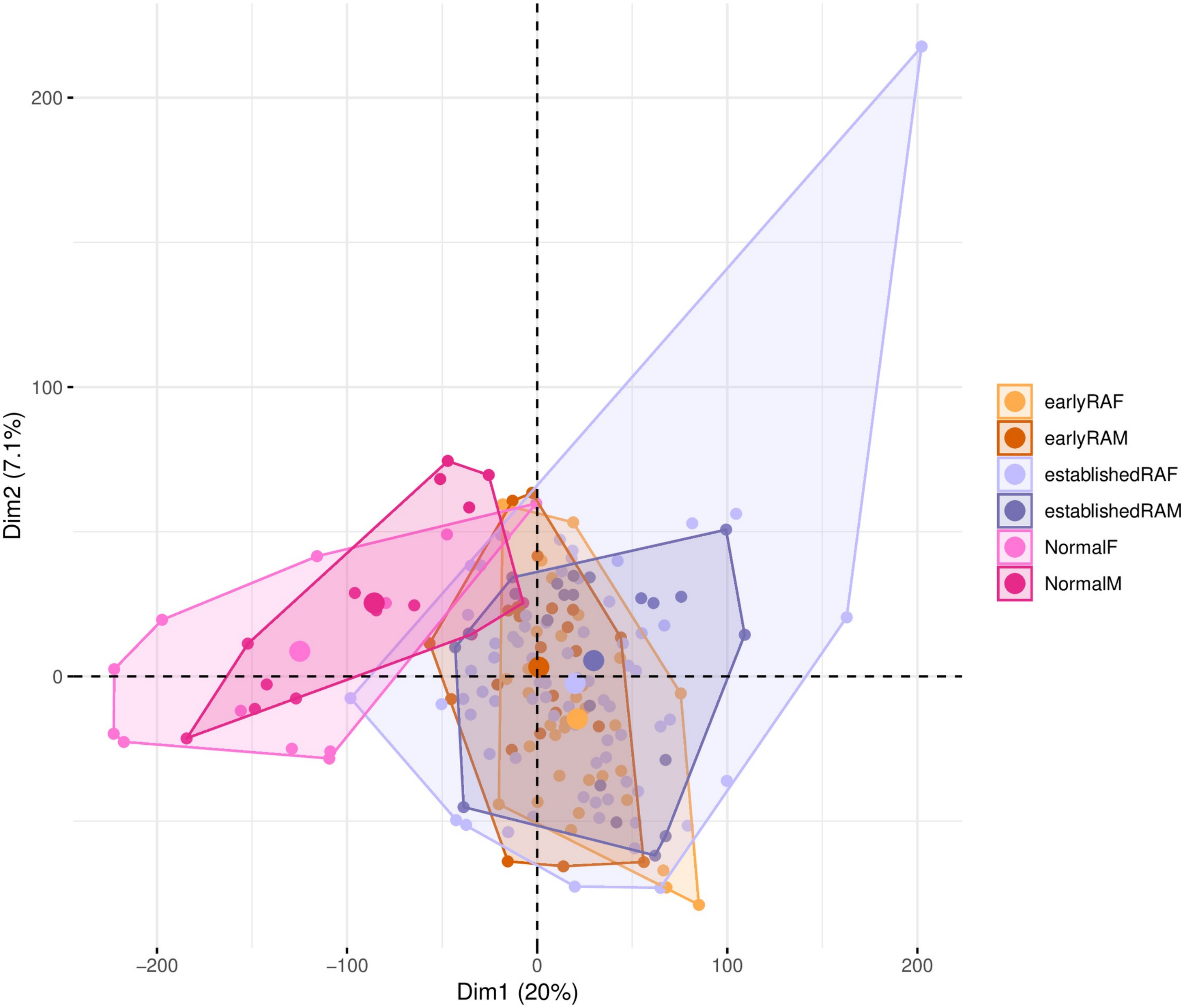


Fig 2. The first two principal components of the PCA considering RPKMs of the coding genes. The areas are the convex hull of the condition. The largest point of one color depicts the centers of the hull. Only those conditions are shown where more than ten samples were available for male and female individuals. Number of samples: 33 earlyRAF, 24 earlyRAM, 73 establishedRAF, 22 establishedRAM, 13 NormalF, 14 NormalM.

https://doi.org/10.1371/journal.pone.0219698.g002

in the Supplement give a more detailed view, in the main text and in the next sections we focus on single effects on the top-level.

Different RA gene expression in men and women

Within each clinical condition there are 85 to 101 genes differentially expressed when comparing men and women within the 236 RNA-seq synovial biopsy samples. Some of these genes are also differentially expressed in early RA compared to normal condition and some of these genes reverse their expression sex-ratio between normal and early RA. This is shown in Fig

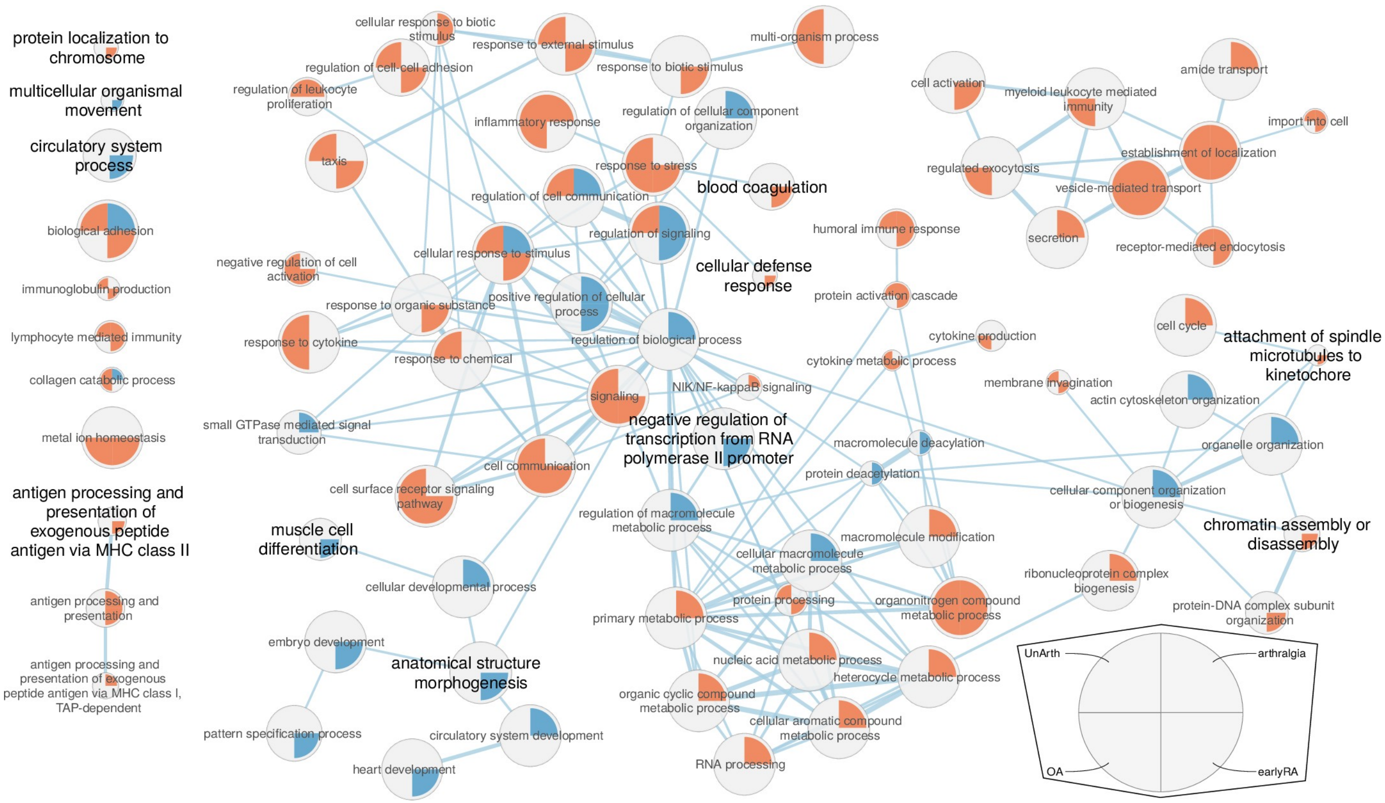

protein localization to chromosome
multicellular organismal movement
circulatory system process
biological adhesion
immunoglobulin production
lymphocyte mediated immunity
collagen catabolic process
metal ion homeostasis
antigen processing and presentation of exogenous peptide antigen via MHC class II
antigen processing and presentation
antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent
cellular response to biotic stimulus
response to external stimulus
regulation of cell-cell adhesion
regulation of leukocyte proliferation
response to biotic stimulus
multi-organism process
inflammatory response
regulation of cellular component organization
taxis
response to stress
blood coagulation
regulation of cell communication
regulation of signaling
cellular response to stimulus
negative regulation of cell activation
cellular defense response
humoral immune response
protein activation cascade
positive regulation of cellular process
response to organic substance
response to cytokine
regulation of biological process
response to chemical
signaling
NIK/NF-kappaB signaling
small GTPase mediated signal transduction
negative regulation of transcription from RNA polymerase II promoter
cell communication
cell surface receptor signaling pathway
regulation of macromolecule metabolic process
muscle cell differentiation
cellular developmental process
embryo development
anatomical structure morphogenesis
pattern specification process
circulatory system development
heart development
cytokine production
cytokine metabolic process
macromolecule deacylation
protein deacetylation
macromolecule modification
cellular macromolecule metabolic process
protein processing
organonitrogen compound metabolic process
primary metabolic process
nucleic acid metabolic process
heterocycle metabolic process
organic cyclic compound metabolic process
cellular aromatic compound metabolic process
RNA processing
cell activation
myeloid leukocyte mediated immunity
amide transport
import into cell
establishment of localization
regulated exocytosis
vesicle-mediated transport
secretion
receptor-mediated endocytosis
cell cycle
attachment of spindle microtubules to kinetochore
membrane invagination
actin cytoskeleton organization
organelle organization
cellular component organization or biogenesis
chromatin assembly or disassembly
ribonucleoprotein complex biogenesis
protein-DNA complex subunit organization
UnArth
arthralgia
OA
earlyRA

Fig 3. GO (BP) term enrichments of DEGs in earlyRA, arthralgia, OA and undifferentiated arthritis. Base state is normal, each term has the DEG enrichment of the four conditions in the circle's quadrants according to the legend bottom right. Red indicates there is an enrichment in the upregulated DEGs, blue indicates an enrichment in the down-regulated DEGs and gray indicates no enrichment. The node size represents the number of genes in the annotation for that term. The edge thickness represents the degree of overlap between the gene-sets of two terms. The terms in a larger font are a selection for meaningful terms specifically for earlyRA (somewhat arbitrary). See methods and main text for filtering and discussion. Number of samples: 10 arthralgia, 57 earlyRA, 22 OA and 6 UA.

https://doi.org/10.1371/journal.pone.0219698.g003

4A, where all genes exhibiting a threefold change can be found in the lower right quadrant. This means that there exist not only genes which are significantly differentially expressed when comparing the normal condition with early RA, but of which some are also significantly lower in healthy males and significantly higher in males with early RA. These genes are ATP2A1, LMOD2, ACTN2, DES, CKM, NRAP, MYH2, XIRP2 and RP11−766F14.2 where all but the last are related to muscles according to the GeneCards database [43] (muscle filament, sarcomere and actin filament). A Gene Ontology term enrichment shows the same, all terms assigned to more than one of the nine genes are related to muscles and actin organization (see Supplementary Archive tables/ for enrichment, where this set is named 'earlyRAdown_NormalFM_down_earlyRAFMup_ageFilter'). As no muscle cells should be present in synovial biopsies, the substantial GO-term in this context refers to cytoskeleton changes (change of expression of actin filament related genes). For RP11-766F14.2 only little is known—maybe because of its just recent aliases [44], where its role in obliterative portal venopathy is described.

The same comparison between early and established RA (Fig 4B) has no threefold significantly differentially expressed genes, but shows two things: the aforementioned genes for cytoskeleton changes are different between earlyRA and establishedRA, this special difference in men and women is only present at earlyRA (see also S10 Fig for the same comparison between establishedRA and normal condition), and secondly, that some cytokines and the two genes

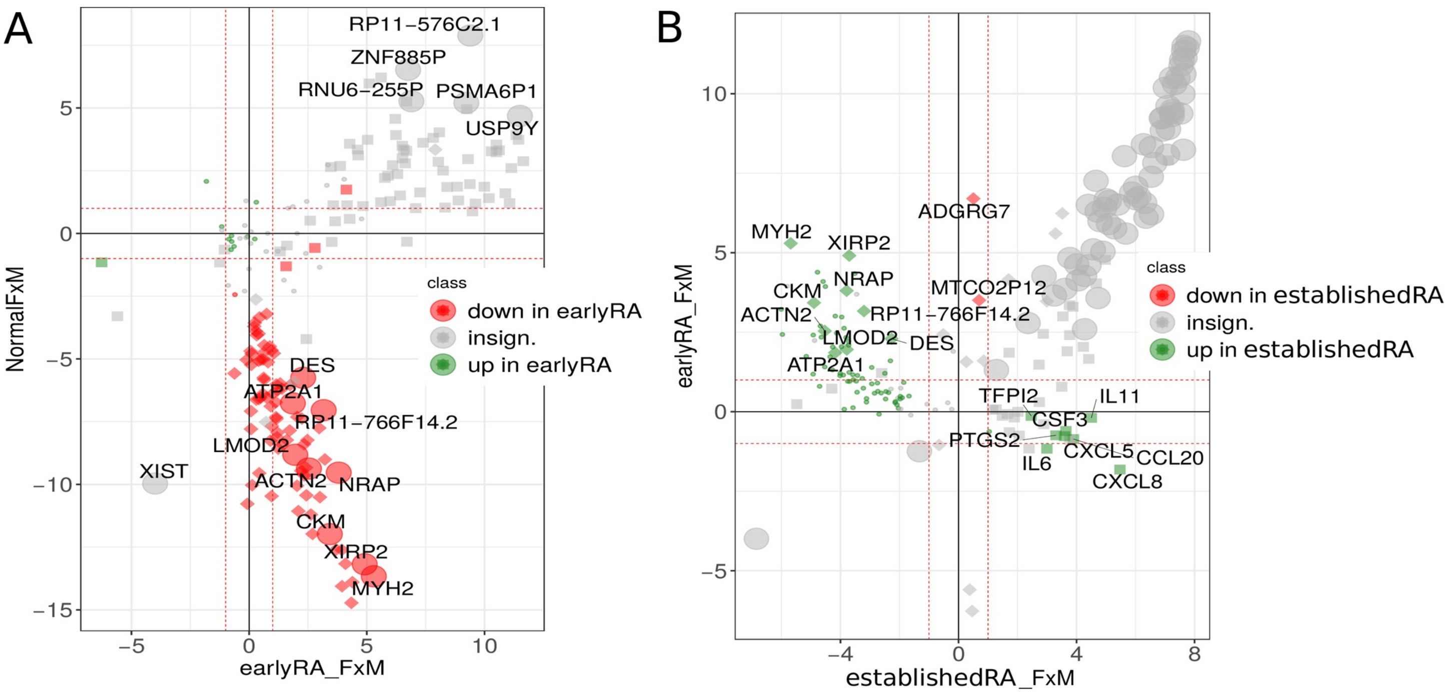


Fig 4. Log$_2$ fold-changes of gene expression between men and women in early RA, established RA and normal condition. Only genes are shown which are significantly differentially expressed in men and women. (A) The sex-ratio of gene expression in early RA and normal condition. The size and shape shows the significance in differences of men and women: the large circles are genes significantly differentially expressed between the sexes in early RA and normal condition, these genes are also labelled. Small squares mean a significant difference between men and woman only in early RA, small diamonds mean a significant difference only between healthy men and woman. The color represents the significance of the difference in expression between normal and early RA. (B) The sex-ratio of gene expression in established RA and early RA. There are no gene significantly differentially expressed in men and women in established RA and early RA and differentially expressed between established RA and early RA. Genes are named if one of the gene expression sex-ratios is significant: squares mean a significant difference between men and woman only in established RA, small diamonds mean a significant difference only between men and woman in early RA.

https://doi.org/10.1371/journal.pone.0219698.g004

PTGS2 and TFPI2 are stronger expressed in men with established RA. Beside the cytokines, also the latter two genes have been also investigated for RA (PTGS2 is more often referred to as COX-2) [45–47].

The comparison between OA and normal condition is unremarkable (S11 Fig).

Different gene expression at different gene biotypes

We then assessed the average expression change of different gene biotypes (as defined as biotypes by Ensembl [48]) in the 236 RNA-seq synovial biopsy samples (Fig 5 and S2 Table). The base state is defined there as the normal condition. The positive and negative average fold-changes in Fig 5B are roughly corresponding to the count of significantly differently expressed genes (corresponds to the difference in counts; Fig 5A). rRNAs of mitochondria are less expressed in RA conditions and arthralgia, while the normal rRNA is much higher expressed in arthralgia and miRNAs are less expressed in RA conditions and arthralgia. Generally, there is a pattern of lower gene expression in RA and arthralgia. It seems unexpected that in this sense undifferentiated arthritis is not similar to the RA conditions, as undifferentiated arthritis has clinical signs of synovitis, but 'just' failing to meet the 2010 American College of Rheumatology criteria [49] for RA. When taking a closer look, it does not look contradictory; in both conditions many genes related to the immune system are highly up-regulated, but in RA even more genes are down-regulated (more than up-regulated and many more than down-regulated in undifferentiated arthritis). The down-regulated genes in early RA seem to have very

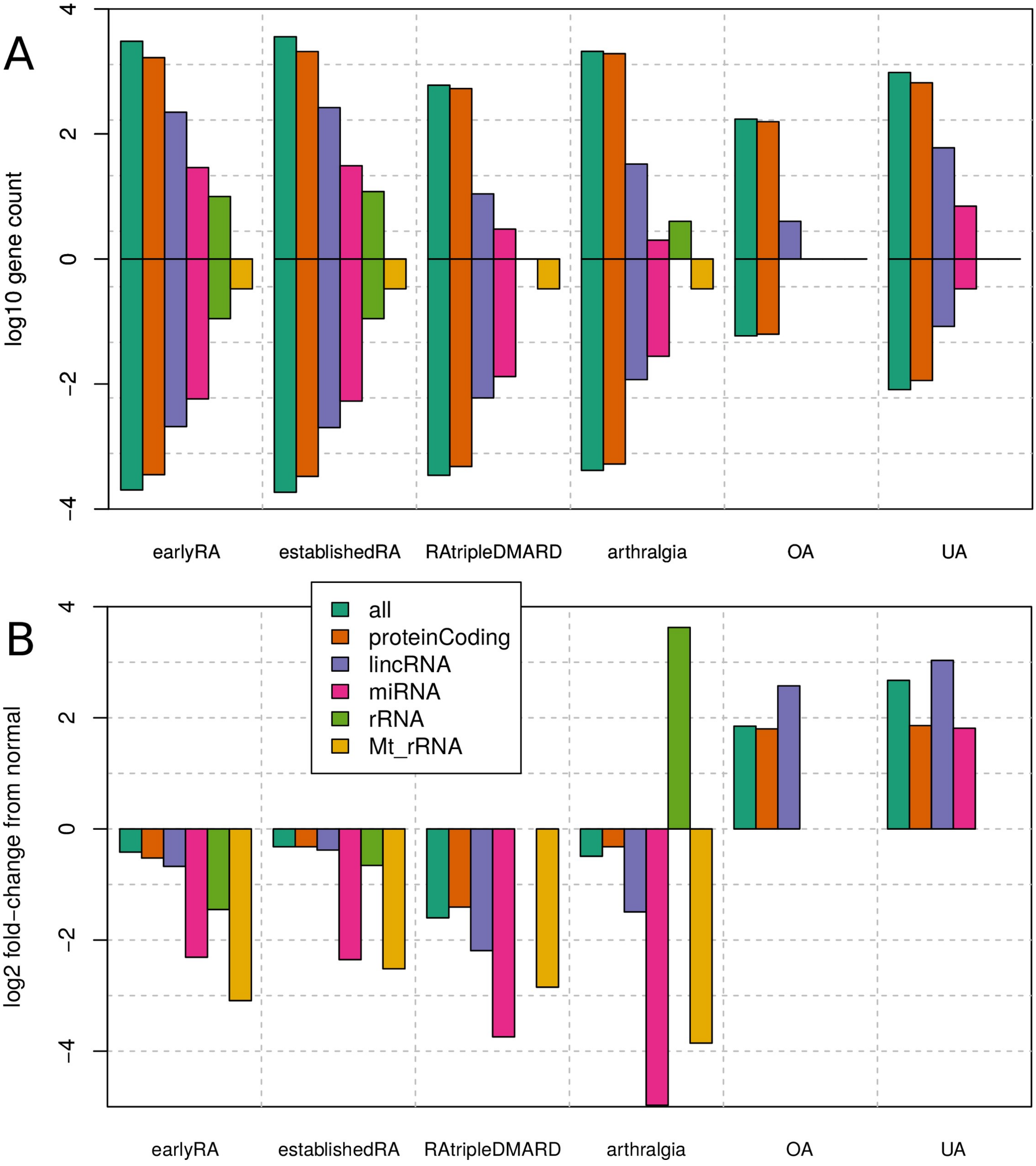


Fig 5. Average fold-changes and counts of different biotypes of genes. The labels on the x-axis mean the change of this condition relative to normal. Only significant changes are regarded. The labels for the biotypes of genes are defined by Ensembl (the biotypes of genes). Missing bars mean that there was no significant change in any

gene of this biotype. (A) The $\log_{10}$ count of the significantly differentially expressed gene by biotype. Each column consists of two bars: from 0 to the positive side are the ($\log_{10}$) numbers of significantly higher expressed genes, from 0 to the negative side are the numbers of significantly lower expressed genes. (B) The average $\log_2$ fold-change of the significantly differentially expressed genes. It is to see that the average $\log_2$ fold-change roughly corresponds in the difference of the counts of the significantly higher and lower expressed genes. Number of samples: 10 arthralgia, 57 earlyRA, 95 establishedRA, 27 normal, 22 OA, 19 RAtripleDMARD and 6 UA.

https://doi.org/10.1371/journal.pone.0219698.g005

different functions compared with the up-regulated genes, except genes related to immune system activation; the most prominently enriched GO-terms are related to cytoskeleton changes (see Supplementary Archive tables/ for the various term enrichments of these genes).

Classification models

Based on the RNA-seq samples and certain gene sets, we were able to generate classification models with significance in an internal cross-fold validation to distinguish early RA from the normal condition and early RA from other diagnoses (Fig 6 and Fig 7). For the comparison of early RA with normal condition, we selected the intersection of genes differentially expressed in the single-variable comparisons normal vs early RA, normal vs OA, normal vs. undifferentiated arthritis and normal vs. arthralgia. This resulted in 45 differentially expressed genes. RPKMs of these genes are the input for generating models. One of the best and also quite the simplest model is a PART [50] model with only one rule, which is visualized in Fig 6. For the comparison of early RA to other arthritides (undifferentiated arthritis, OA, arthralgia) we selected the intersection of genes differentially expressed in the single-variable comparisons of early RA vs normal, early RA vs arthralgia and early RA vs OA (it would be no gene left when including early

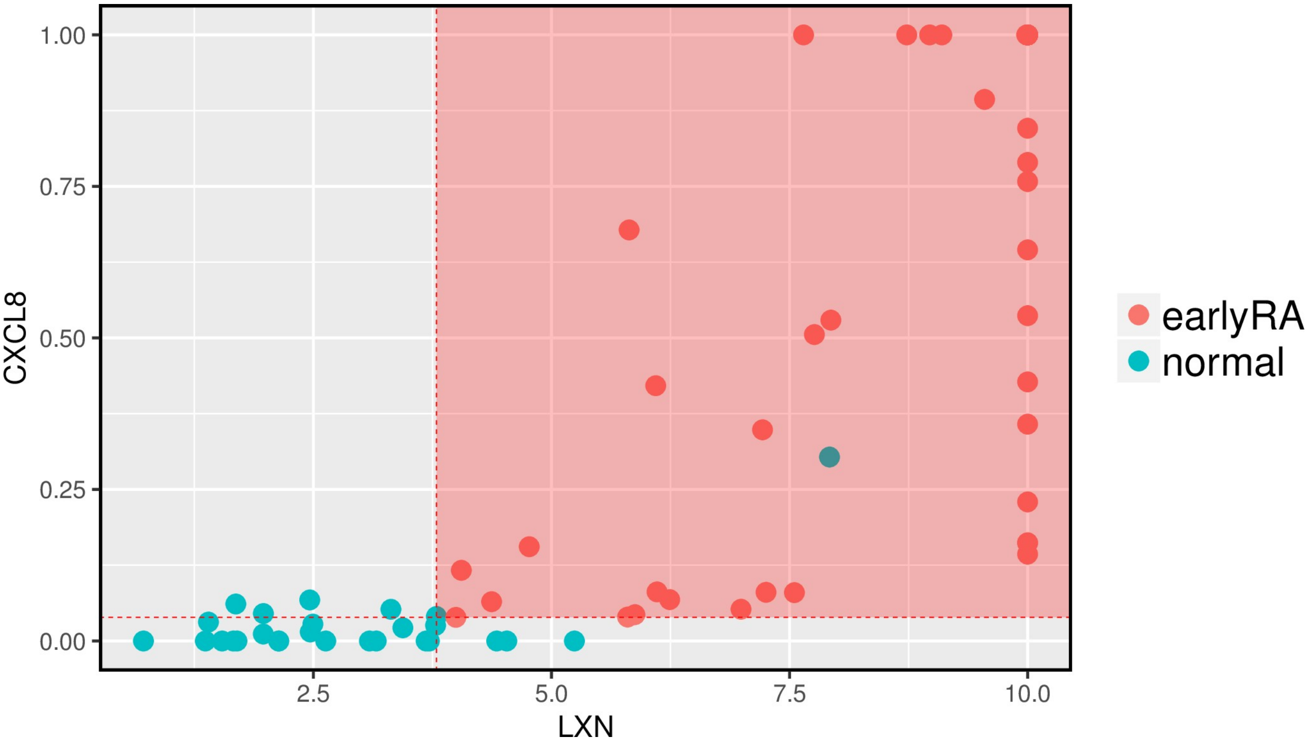
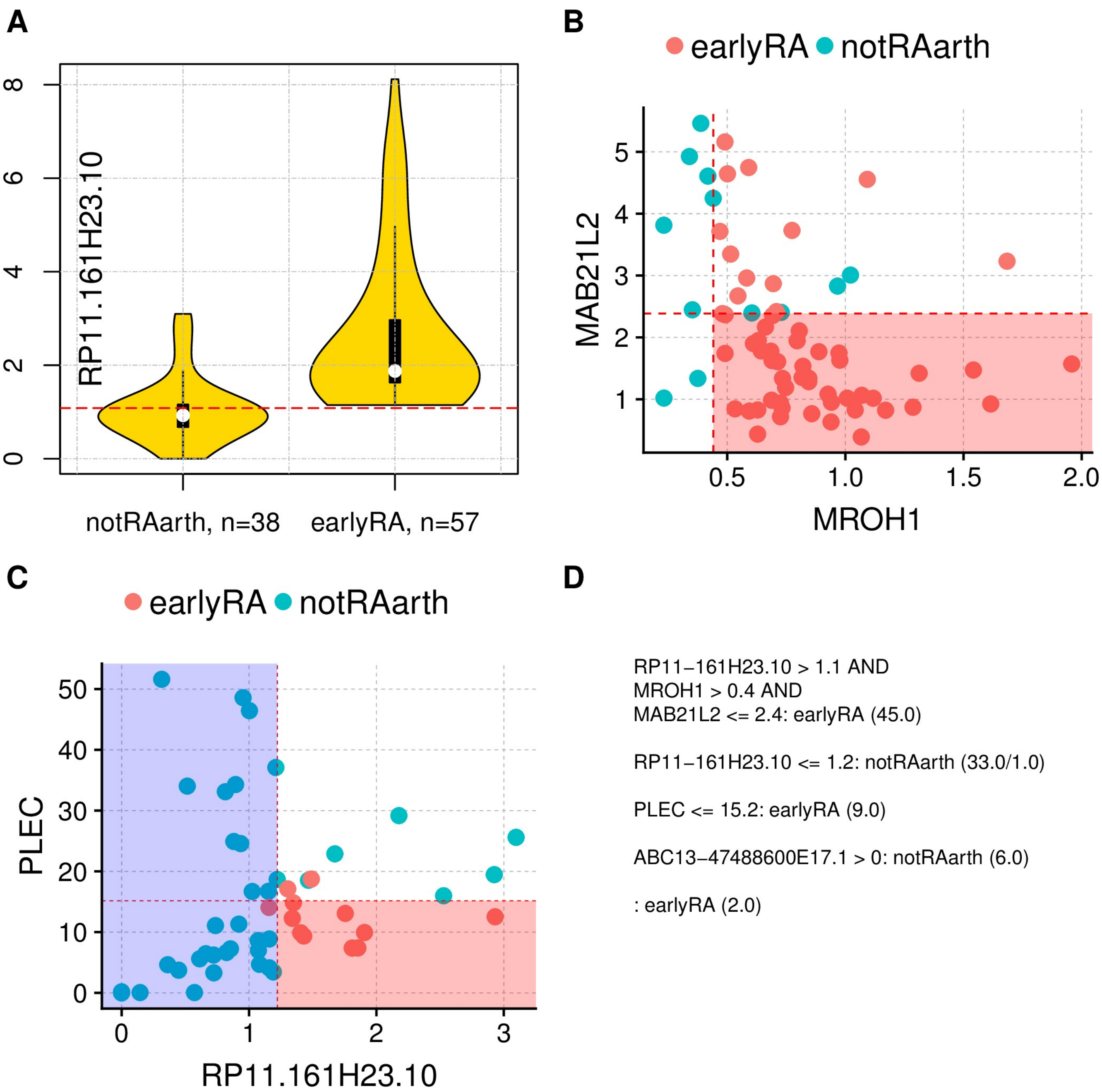


Fig 6. Classification model for distinguish between early RA and normal based on RPKMs. The model consists only of the single rule LXN > 3.8 AND CXCL8 > 0.04 -> early RA. It corresponds to an accuracy of 92% at the 10-fold cross-validation (p-value $2.02^{*}10^{-13}$). Variables for model-generation were pre-selected upon the intersection of single-variable comparisons. This pre-selection weakens the cross-validation as it is no part of it. This model is only intended for a distinction between early RA and normal as simple as possible based on gene expression. In total there are 84 samples. The RPKM values are cut at 10 and at 1, respectively.

https://doi.org/10.1371/journal.pone.0219698.g006



RP11–161H23.10 > 1.1 AND
MROH1 > 0.4 AND
MAB21L2 <= 2.4: earlyRA (45.0)

RP11–161H23.10 <= 1.2: notRAarth (33.0/1.0)

PLEC <= 15.2: earlyRA (9.0)

ABC13–47488600E17.1 > 0: notRAarth (6.0)

: earlyRA (2.0)

Fig 7. Classification model for distinguish between early RA and not healthy but also not RA based on RPKMs. 'not RA' as present in the data, which is OA, arthralgia and undifferentiated arthritis, shortcut 'notRAarth'. The model consists of four rules, where three of them are shown here graphically. Panels A and B depict the first rule, where in B are only cases left which are higher than the threshold in A. The red shaded area in B shows the cases hit by the first rule, which are all earlyRA. Panel C shows the second and third rule, where the model output for the red shaded area is earlyRA and for the blue shaded area is notRAarth. The complete model as text is shown in D. The threshold values are rounded to one decimal place. The model corresponds to an accuracy of 86% at the 10-fold cross-validation (p-value $1.7^{*}10^{-12}$). Variables for model-generation were pre-selected upon the intersection of single-variable comparisons. This pre-selection weakens the cross-validation as it is no part of it. This model is only intended for a distinction between early RA and other conditions as simple as possible based on gene expression. In total there were 95 samples as input data.

https://doi.org/10.1371/journal.pone.0219698.g007

RA vs undifferentiated arthritis in the intersection). This resulted in 94 differentially expressed genes. Also for this classification, a PART model is one of the best. It consists of four rules, which are shown in Fig 7. An overview of other model performances is shown in S3 Table. For an overview of single variable importance measured with information gain [51] and reliefF [52] see S4 Table. The selection of genes based on single-variable comparisons as input for generation classification models weakens the validity of the internal cross-fold validation, but is nevertheless more solid than the direct group wise comparison (e.g. just normal vs early RA) without any validation. The genes which were selected for the models are not unknown: LXN is known to be upregulated in early OA [53] and not only as part of the inflammatory response but also influencing the perception of pain [54]; several CXCL genes (chemokines) have been shown to be upregulated at RA, also CXCL8 [11, 12]; MAB21L2 is upregulated especially in OA [55]; about the RP11-* genes less is known as these labels are still their original clone ID [56].

Discussion

In this analysis, we have added several new views, hints and insights about gene expression in RA. Some of these approaches had been tried on former, limited microarray data in the one or other similar way [18, 19, 57–60]. Compared with these results, it makes sense to to repeat the same or similar types of analyses with recent RNA-seq data. With this recent data, we got several new conclusions. First, the clustering and dimension reduction gives a highly informative and high-level view of RA and related diagnoses; it shows the major difference of the healthy and the unhealthy conditions. Such high-level views are presented in several papers (as in the microarray papers mentioned before); here it is shown with more sensitive RNA-seq data.

Clustering

The clustering and dimension reduction shows in brief certain high-level differences between the conditions (Fig 1), no differences between men and woman on the highest level (Fig 2), no differences between men and woman at gene cluster level (S12 Fig) and expected vicinity of the conditions by their fold-changes (S6 Fig). OA and establishedRA covers the largest area in the PCA, which might be related to their loose definition, both are samples any time after their first diagnosis (establishedRA at least 12 months after treatment start).

Overlaps in gene lists

S13 and S14 Figs give a summary of all overlaps between three and four conditions. The amount of linking arcs might be overwhelming at the beginning, but every link can be located with one close look. Venn diagrams are the straightforward choice to visualize overlaps, but they are unsightly with more than 4 or 5 sets, except maybe the famous six-way banana Venn diagram [61]. In the S13 and S14 Figs we have 9 and 12 sets.

Men and women

We showed with gene expression data that significant differences between men and women exist at least on a detailed view on RA (Fig 4; Fig 2 for the high-level view). The set of genes being three-fold significant (significantly differentially expressed in normal vs. early RA, in men vs women in normal and early RA, but with reversed sign, see Fig 4A) looks a bit out of place for synovial tissue with their term enrichment for muscles (muscle filament and sarcomere), but these are also enriched for actin filament reorganization. It is known that rearrangements in the cytoskeleton are associated with RA [62, 63]. This seems in this dataset specific for earlyRA. At establishedRA several genes related to the immune system are higher expressed. Some of these genes were already reported as sex-biased genes [64], here we show a particular instance. These 'RA-cytoskeleton-genes' and these sex-biased immune system genes are solid points for further investigation, as they might be targets for therapy, responsible for some side effects different in men and women, the effect of different behavior after progression of RA or just be an artifact unknown to now.

Gene biotypes

In the high-level view of gene biotypes, we see a clear pattern of generally lower gene expression and more genes significantly down-regulated in RA and arthralgia (Fig 5). This give rise to the hypothesis that the miRNA diagnostic marker(s) for RA—what many are looking for [29, 65, 66]—might be negative ones (that means the lack of certain miRNAs would point to RA), although it might be confused with arthralgia. The rRNA is worth extra attention as it is the major difference between arthralgia and the rest. rRNAs of mitochondria are less expressed in RA conditions and arthralgia, likely because of the hypoxic microenvironment [67, 68], but also possible associated with general exhaustion [69, 70].

Classification models

For the simplest distinction between RA and other conditions, we provide two classification models. The genes used in the models might not be causative or functionally most related, but are a minimal set of genes to classify the data, which classification in that way is also significant at 10-fold cross-validation. The pre-selection of genes based on intersecting single-variable comparisons is needed for escaping the curse of dimensionality for multivariable classification methods. This pre-selection has some limitations: it has itself no internal validation and the gene sets from the single-variable comparisons are differently solid, as there are conditions with different sample sizes (smallest: 6 samples of UA and 10 of arthralgia). This increases the chance to lose the 'best' (= most likely causative) predictors and getting instead the most correlated (to the 'best' predictors) variables in the model. This pre-selection weakens the validity of the internal cross-fold validation. The used classification method (PART [50], a tree learner based rule generator) is likely over-simplifying RA. For final assessments of the particular—potential causal—functions of the selected predictors, dedicated wet lab experiments are needed. The presented classification models are only intended for a distinction between RA and other conditions as simple as possible based on gene expression.

Previous microarray studies

Many RNA-seq data is already published for RA, as used in this article; more sample data is still available from microarrays. Single cell sequencing RNA-seq samples are catching up in number and of course depth, but previous microarray data are still a large source to compare with. The comparison of the collected RNAseq data with suitable previous microarray studies shows expected results (S1 Text, especially S15 Fig). The fact that the relative overlap of the up-regulated genes is always higher than of the down-regulated genes could point to a bias of seen importance (= more solid annotation, as this have changed over time) of these up-regulated genes or to a higher biological importance of these genes (as they are more overlapping in independent studies). Overall, it seems important to use RNA-seq data instead of microarray data for the transcriptome, to use the very same gene annotation and to process the data in the very same way. In such collections are likely still plenty of hidden insights.

Material and methods

Data collection

We have combined and compared 236 RNA-seq synovial biopsy samples from the papers of Walsh et al. and Guo et al. [15, 16] and microarray data from the papers of Liu et al., Teixeira et al., Niu et al. and Yoshida et al. [18, 58–60] in this study. This RNA-seq sample collection was chosen, because it is consistent, large and there are open questions for which insights or at least hints are in this data. The main data are the 236 RNA-seq samples, the microarray data are only used for showing the overlaps with the main data.

The samples of Teixeira et al. and Niu et al. are from peripheral blood (Peripheral Blood Mononuclear Cells—PBMCs), whereas all other samples are from synovial tissue. RNA-seq data were downloaded from the Sequence Read Archive (SRA) [71] (see S1 Table for accession identifiers and number of samples per diagnosis), microarray and clinical data were taken as presented in the original papers. The raw RNA-seq data were processed as described in the next section, the clinical information was taken as provided in the source papers [15, 16], where 'normal' refers to healthy patients, 'arthralgia' refers to a population based on this symptom, rather than a specific diagnosis, 'earlyRA' means treatment naïve RA within 12 months of first diagnosis, 'establishedRA' means treatment experienced RA of >12 months disease duration, 'RAtripleDMARD' means RA about 6 months after treatment initiation with methotrexate, sulfasalazine and hydroxychloroquine, finally 'OA' means osteoarthritis and 'UA' means undifferentiated arthritis. 'establishedRA' and 'RAtripleDMARD' are both patients under ongoing treatment, where the latter was a specific treatment and the sample after a certain time (6 months after treatment start) and the former are patients at/after any treatment after a longer period of time (>12 months after treatment start) [16]. Undifferentiated arthritis is defined as clinical signs of synovitis, but failing to meet the 2010 American College of Rheumatology criteria for RA [16]. An 'F' or an 'M' appended to a label means the subset of female and male patients.

RNA-seq—primary data processing

Reads were mapped onto the human reference genome release hg38 (GRCh38) [72] with Ensembl transcript annotation version 87 [48] using Tophat version 2.1.1 [73] with Bowtie version 2.2.9 [74]. Reads were counted with featureCounts [75] and gene expression values (reads per kilobase exon per million mapped reads (RPKM)) were calculated with Cufflinks version 2.2 [76]. The differential expression between two sample groups was calculated with edgeR [77]. The filtering for differentially expressed genes is for p-value of 0.05 (FWER corrected) and minimal fold-change of 2. In the more specific analyses for single genes, for the differences in men and woman and for classification models, (healthy) age-related genes are removed. This is because the sampled healthy subjects are in average quite younger than the subjects with different arthritis conditions and at comparisons between them age-related genes are expected to be significantly different. Age-related genes are taken from Yang et. al. [78]. We performed also comparisons of gene expression between groups adjusted for age. At the most changing adjustment in the comparison between healthy subjects and early RA (an average age of 35.2 vs. 55.9), we realized that many genes well known for RA are filtered (as CCL19 [79], CCL22 [80], CCR6 [81], CD6 [82], CDH11 [83], IFIT1B (as a paralog to IFIT1 [84]), IL26 [85], IL2RB [86], MMP10 [87], MMP12 [88], MMP8 [89] and MMP9 [90]). Similar worrying are the overlaps between unique DEGs in the comparison unadjusted and adjusted by age with the external age-related genes (as used for filtering from Yang et al. [78]), we see even a higher overlap between age-adjusted DEGs (healthy vs early RA) and the external age-related genes. Given that, we used the comparison without adjustment for further analyses. Age-adjusted comparisons are available in the Supplementary Archive.

Clustering and dimension reduction

The RPKM values per gene were the input for clustering. The standard R [91] functions were used for PCA and hierarchical clustering, as well as the interfaces of the visualization libraries (described in section visualization). For other dimension reduction methods the Matlab Toolbox for Dimensionality Reduction was used [92].

Co-expression analysis

For clustering genes into modules based on their expression profiles over all conditions, we used the expression of the transcript with the highest expression per gene. We grouped samples into conditions by choosing the median expression per gene and used this information as input for the Weighted Gene Co-expression Network Analysis (WGCNA) method [93]. Genes were kept only if the total cumulative RPKM over all samples was more than 10 and when exceeding the standard deviation of 0.5 along all conditions.

Gene enrichment analyses

For Gene Ontology (GO) [37] enrichment analysis of a gene set, GOstats version 2.46.0 [94] was used with default parameters, except the parameter ’conditional’, which was set to TRUE (which removes genes from significant terms deeper in the hierarchy). For the detection of enriched KEGG [38] and REACTOME [39] terms geneSCF version 1.1 [95] was used. All complete lists are available in the Supplementary Archive (tables/). For having a diagram of the GO (BP) terms fitting onto a single page (in Fig 3), we used REVIGO [96] for reducing overlapping terms with an allowed similarity threshold of 0.4 and the Cytoscape [97] plugin EnrichmentMap [98] for visualization with a threshold of 10^{-6} for the raw p-values and an edge similarity threshold of 0.5.

Visualization

For the visualization of clusters, distributions, overlaps, correlations and ratios we used the following R packages: ape [99], vioplot [100], dplyr [101], ggplot2 [102], ggrepel [103], FactoMineR [104], factoextra [105] and WGCNA [93]. Additionally, we used the tools Cytoscape [97] and Circos [106].

Machine learning for classification models

For the following classification methods, the reference implementation in WEKA [107] was used: C4.5 [32] (implemented as J48), PART [50], Alternating Decision Trees [108], naive Bayes [109], SMO [110]. The importance of variables was measured with their information gain [51] and reliefF [52] as implemented in WEKA.

Supporting information

S1 Fig. The first two principal components of the PCA based on the RPKMs of the coding genes. The areas are the convex hulls of the conditions. The largest point of one color depicts the center of a hull. Number of samples: 22 OA, 10 arthralgia, 57 earlyRA, 95 longRA, 27 normal, 19 RApost and 6 UnArth.
(TIF)

S2 Fig. Conformal Eigenmaps (CCA) based on the RPKMs of the coding genes. Number of samples: 22 OA, 10 arthralgia, 57 earlyRA, 95 longRA, 27 normal, 19 RApost and 6 UnArth.
(TIF)

S3 Fig. Sammon mapping based on the RPKMs of the coding genes. Number of samples: 22 OA, 10 arthralgia, 57 earlyRA, 95 longRA, 27 normal, 19 RApost and 6 UnArth.
(TIF)

S4 Fig. Multidimensional scaling (MDS) based on the RPKMs of the coding genes. Number of samples: 22 OA, 10 arthralgia, 57 earlyRA, 95 longRA, 27 normal, 19 RApost and 6 UnArth.
(TIF)

S5 Fig. The first ten principal components of the PCA considering RPKMs of the coding genes. The areas are the convex hull of the condition. The largest point of one color depicts the centers of the hull. Only those conditions are shown where more than ten samples were available for male and female individuals. Number of samples: 33 earlyRAF, 24 earlyRAM, 73 establishedRAF, 22 establishedRAM, 13 NormalF, 14 NormalM.
(PDF)

S6 Fig. Neighbor joining tree based on the log fold-changes of significantly different genes (significant in any comparison). Origin for the fold-changes is normal/healthy (normal/healthy is the 0-vector). The x-axis is based on the Manhattan distance of significant fold-changes. The distance might be meaningless as an absolute value, but informative as relative distance.
(TIF)

S7 Fig. Neighbor joining tree based on the log fold-changes of significantly different miRNA genes (significant in any comparison). Origin for the fold-changes is normal/healthy (normal/healthy is the 0-vector). The x-axis is the distance. The distance might be meaningless as an absolute value, but informative as relative distance.
(TIF)

S8 Fig. Hierarchical clustering based on the log fold-changes of significantly different genes (significant in any comparison). Origin for the fold-changes is normal/healthy (normal/healthy is the 0-vector). The x-axis is the distance. The distance might be meaningless as an absolute value, but informative as relative distance.
(TIF)

S9 Fig. Hierarchical clustering based on the log fold-changes of significantly different miRNA genes (significant in any comparison). Origin for the fold-changes is normal/healthy (normal/healthy is the 0-vector). The x-axis is the distance. The distance might be meaningless as an absolute value, but informative as relative distance.
(TIF)

S10 Fig. Log2 fold-changes of gene expression between men and women in established/long RA and normal condition. Only genes are shown which are significantly differentially expressed in men and women. The size and shape shows the significance in differences of men and women: the large circles are genes significantly differentially expressed between the sexes in established/long RA and normal condition, these genes are also labelled. Small squares mean a significant difference between men and woman only in established/long RA, small diamonds mean a significant difference only between healthy men and woman. The color represents the significance of the difference in expression between normal and established/long RA.
(TIF)

S11 Fig. Log2 fold-changes of gene expression between men and women in OA and normal condition. Only genes are shown which are significantly differentially expressed in men and women. The size and shape shows the significance in differences of men and women: the large circles are genes significantly differentially expressed between the sexes in OA and normal condition. Small squares mean a significant difference between men and woman only in OA, small diamonds mean a significant difference only between healthy men and woman. The

color represents the significance of the difference in expression between normal and OA.
(TIF)

S12 Fig. Gene clusters. Genes were clustered into modules of co-expression and their module eigengenes (ME) normalized expressions are shown here for all conditions. In panel B the patients are split in male/female and diagnosis, in panel A the average value of male and female is shown. The color and the size of the dots depict the first principal component per module using the scaled expression of the respective genes of a module over all conditions. Parts of panel B should be seen with caution as some groups have a very small sample size (arthralgiaM and UAM). Pairwise significant differences of conditions within modules are in S5 Table. These are summarized in panel B as color-code: purple rows and columns are not significantly different to any condition in any module, gray cells are not significantly different to any condition within the particular module. Number of samples: 8 arthralgiaF, 2 arthralgiaM, 13 normalF, 14 normalM, 13 OAF, 9 OAM, 19 RAtripleDMARD, 5 UAF, 1 UAM, 33 earlyRAF, 24 earlyRAM, 73 establishedRAF and 22 establishedRAM.
(TIF)

S13 Fig. The amount and the overlaps of up- and down-regulated genes for different RA conditions. The base condition is normal, so the label 'earlyRA' means normal compared with early RA. Up-regulated fractions are shown in green, down-regulated fractions are shown in red; gray are fractions of genes which are not significantly differentially expressed. The full set is the union of significantly differentially expression genes in all comparisons. The colors of the arc connections are dependent on what they are connecting. Number of samples: 57 earlyRA, 95 establishedRA, 27 normal and 19 RAtripleDMARD.
(TIF)

S14 Fig. The amount and the overlaps of up- and down-regulated genes for early RA, OA, arthralgia and undifferentiated arthritis. The base condition is normal, so the label 'earlyRA' means normal compared with early RA. Up-regulated fractions are shown in green, down-regulated fractions in red; in gray are fractions of genes which are not significantly differentially expressed. The full set is the union of significantly differentially expression genes in all comparisons in this Fig. The colors of the arc connections are dependent on what they are connecting. Number of samples: 10 arthralgia, 57 earlyRA, 27 normal, 22 OA and 6 UA.
(TIF)

S15 Fig. Comparison of significantly differentially expressed genes. The base state of genes is the normal condition, except for 'OA->RA' where it is OA (base state is the first one in 'condition <one> compared to condition <two>'). The origins of the different sources or sets (the 'Set1' to 'Set5') are listed in Table B in S1 Text.
(TIF)

S1 Table. Accession identifiers (SRA database) and number of samples per diagnosis.
(XLS)

S2 Table. The average fold-changes and counts of genes in certain comparisons when genes are grouped according to their biotype.
(XLS)

S3 Table. Performance of different ML-methods for the classification Normal vs. earlyRA; weka is used.
(XLS)

S4 Table. The importance of variables for the classifications Normal vs. earlyRA and earlyRA vs. notRAarth; information gain and reliefF is used from weka.
(XLS)

S5 Table. Pairwise significant differences of conditions within modules of the Weighted gene correlation network analysis (WGCNA). The differences were tested with a Wilcoxon test comparing the eigengene values of the samples of two conditions. The 'significant?' column is on the Bonferroni-corrected p-value. Each combination is twice in the table (e.g. NormalF vs. earlyRAF and earlyRAF vs. NormalF are the same).
(XLS)

S1 Text. Relocated supporting text. Contains the sections 'Clustering of genes', 'Overlaps in differentially expressed genes between clinical conditions' and 'Comparisons with other studies'.
(DOCX)

Author Contributions

Conceptualization: Alexander Platzer, Thomas Nussbaumer.

Data curation: Alexander Platzer.

Investigation: Alexander Platzer, Thomas Nussbaumer.

Methodology: Alexander Platzer, Thomas Nussbaumer.

Resources: Thomas Karonitsch.

Validation: Alexander Platzer, Thomas Nussbaumer.

Visualization: Alexander Platzer, Thomas Nussbaumer.

Writing – original draft: Alexander Platzer, Thomas Nussbaumer.

Writing – review & editing: Alexander Platzer, Thomas Nussbaumer, Thomas Karonitsch, Josef S. Smolen, Daniel Aletaha.

References

1. Smolen JS, Aletaha D, Barton A, Burmester GR, Emery P, Firestein GS, et al. Rheumatoid arthritis. Nat Rev Dis Primers. 2018; 4:18001. https://doi.org/10.1038/nrdp.2018.1 PMID: 29417936.
2. Smolen JS, Aletaha D, McInnes IB. Rheumatoid arthritis. Lancet. 2016; 388(10055):2023–38. Epub 2016/10/30. https://doi.org/10.1016/S0140-6736(16)30173-8 PMID: 27156434.
3. WHO. Chronic rheumatic conditions [cited 2018 1th Oct]. Available from: http://www.who.int/chp/topics/rheumatic/en/.
4. Wellcome Trust Case Control C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447(7145):661–78. Epub 2007/06/08. https://doi.org/10.1038/nature05911 PMID: 17554300; PubMed Central PMCID: PMC2719288.
5. Tasaki S, Suzuki K, Kassai Y, Takeshita M, Murota A, Kondo Y, et al. Multi-omics monitoring of drug response in rheumatoid arthritis in pursuit of molecular remission. Nat Commun. 2018; 9(1):2755. Epub 2018/07/18. https://doi.org/10.1038/s41467-018-05044-4 PMID: 30013029; PubMed Central PMCID: PMC6048065.
6. Catrina AI, Joshua V, Klareskog L, Malmstrom V. Mechanisms involved in triggering rheumatoid arthritis. Immunol Rev. 2016; 269(1):162–74. Epub 2015/12/20. https://doi.org/10.1111/imr.12379 PMID: 26683152.
7. Angelotti F, Parma A, Cafaro G, Capecchi R, Alunno A, Puxeddu I. One year in review 2017: pathogenesis of rheumatoid arthritis. Clin Exp Rheumatol. 2017; 35(3):368–78. Epub 2017/06/21. PMID: 28631608.

8. O'Dell JR, Mikuls TR, Taylor TH, Ahluwalia V, Brophy M, Warren SR, et al. Therapies for active rheumatoid arthritis after methotrexate failure. N Engl J Med. 2013; 369(4):307–18. Epub 2013/06/13. https://doi.org/10.1056/NEJMoa1303006 PMID: 23755969.

9. Goekoop-Ruiterman YP, de Vries-Bouwstra JK, Allaart CF, van Zeben D, Kerstens PJ, Hazes JM, et al. Clinical and radiographic outcomes of four different treatment strategies in patients with early rheumatoid arthritis (the BeSt study): a randomized, controlled trial. Arthritis Rheum. 2005; 52 (11):3381–90. Epub 2005/11/01. https://doi.org/10.1002/art.21405 PMID: 16258899.

10. Yoshizawa T, Hammaker D, Sweeney SE, Boyle DL, Firestein GS. Synoviocyte innate immune responses: I. Differential regulation of interferon responses and the JNK pathway by MAPK kinases. J Immunol. 2008; 181(5):3252–8. Epub 2008/08/21. https://doi.org/10.4049/jimmunol.181.5.3252 PMID: 18713996; PubMed Central PMCID: PMC2725405.

11. Proost P, Struyf S, Loos T, Gouwy M, Schutyser E, Conings R, et al. Coexpression and interaction of CXCL10 and CD26 in mesenchymal cells by synergising inflammatory cytokines: CXCL8 and CXCL10 are discriminative markers for autoimmune arthropathies. Arthritis Res Ther. 2006; 8(4): R107. Epub 2006/07/19. https://doi.org/10.1186/ar1997 PMID: 16846531; PubMed Central PMCID: PMC1779382.

12. Kraan MC, Patel DD, Haringman JJ, Smith MD, Weedon H, Ahern MJ, et al. The development of clinical signs of rheumatoid synovial inflammation is associated with increased synthesis of the chemokine CXCL8 (interleukin-8). Arthritis Res. 2001; 3(1):65–71. Epub 2001/02/15. https://doi.org/10.1186/ar141 PMID: 11178128; PubMed Central PMCID: PMC17826.

13. Kirkham BW, Kavanaugh A, Reich K. Interleukin-17A: a unique pathway in immune-mediated diseases: psoriasis, psoriatic arthritis and rheumatoid arthritis. Immunology. 2014; 141(2):133–42. Epub 2013/07/04. https://doi.org/10.1111/imm.12142 PMID: 23819583; PubMed Central PMCID: PMC3904234.

14. Walsh AM, Whitaker JW, Huang CC, Cherkas Y, Lamberth SL, Brodmerkel C, et al. Integrative genomic deconvolution of rheumatoid arthritis GWAS loci into gene and cell type associations. Genome Biol. 2016; 17:79. Epub 2016/05/04. https://doi.org/10.1186/s13059-016-0948-6 PMID: 27140173; PubMed Central PMCID: PMC4853861.

15. Walsh AM, Wechalekar MD, Guo Y, Yin X, Weedon H, Proudman SM, et al. Triple DMARD treatment in early rheumatoid arthritis modulates synovial T cell activation and plasmablast/plasma cell differentiation pathways. PLoS One. 2017; 12(9):e0183928. Epub 2017/09/02. https://doi.org/10.1371/journal.pone.0183928 PMID: 28863153; PubMed Central PMCID: PMC5580991.

16. Guo Y, Walsh AM, Fearon U, Smith MD, Wechalekar MD, Yin X, et al. CD40L-Dependent Pathway Is Active at Various Stages of Rheumatoid Arthritis Disease Progression. J Immunol. 2017; 198 (11):4490–501. Epub 2017/04/30. https://doi.org/10.4049/jimmunol.1601988 PMID: 28455435.

17. Kasperkovitz PV, Timmer TC, Smeets TJ, Verbeet NL, Tak PP, van Baarsen LG, et al. Fibroblast-like synoviocytes derived from patients with rheumatoid arthritis show the imprint of synovial tissue heterogeneity: evidence of a link between an increased myofibroblast-like phenotype and high-inflammation synovitis. Arthritis Rheum. 2005; 52(2):430–41. Epub 2005/02/05. https://doi.org/10.1002/art.20811 PMID: 15692990.

18. Liu T, Lin X, Yu H. Identifying genes related with rheumatoid arthritis via system biology analysis. Gene. 2015; 571(1):97–106. Epub 2015/06/29. https://doi.org/10.1016/j.gene.2015.06.058 PMID: 26117171.

19. 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(8):1008–14. Epub 2007/01/16. https://doi.org/10.1136/ard.2006.063412 PMID: 17223656; PubMed Central PMCID: PMC1954704.

20. Ospelt C, Brentano F, Rengel Y, Stanczyk J, Kolling C, Tak PP, et al. Overexpression of toll-like receptors 3 and 4 in synovial tissue from patients with early rheumatoid arthritis: toll-like receptor expression in early and longstanding arthritis. Arthritis Rheum. 2008; 58(12):3684–92. Epub 2008/11/28. https://doi.org/10.1002/art.24140 PMID: 19035519.

21. Ikeuchi H, Kuroiwa T, Hiramatsu N, Kaneko Y, Hiromura K, Ueki K, et al. Expression of interleukin-22 in rheumatoid arthritis: potential role as a proinflammatory cytokine. Arthritis Rheum. 2005; 52 (4):1037–46. Epub 2005/04/09. https://doi.org/10.1002/art.20965 PMID: 15818686.

22. Engler A, Aeschlimann A, Simmen BR, Michel BA, Gay RE, Gay S, et al. Expression of transient receptor potential vanilloid 1 (TRPV1) in synovial fibroblasts from patients with osteoarthritis and rheumatoid arthritis. Biochem Biophys Res Commun. 2007; 359(4):884–8. Epub 2007/06/15. https://doi.org/10.1016/j.bbrc.2007.05.178 PMID: 17560936.

23. Pennell LM, Galligan CL, Fish EN. Sex affects immunity. J Autoimmun. 2012; 38(2–3):J282–91. Epub 2012/01/10. https://doi.org/10.1016/j.jaut.2011.11.013 PMID: 22225601.

24. Lleo A, Battezzati PM, Selmi C, Gershwin ME, Podda M. Is autoimmunity a matter of sex? Autoimmun Rev. 2008; 7(8):626–30. Epub 2008/07/08. https://doi.org/10.1016/j.autrev.2008.06.009 PMID: 18603021.

25. Forslind K, Hafstrom I, Ahlmen M, Svensson B, Group BS. Sex: a major predictor of remission in early rheumatoid arthritis? Ann Rheum Dis. 2007; 66(1):46–52. Epub 2006/12/13. https://doi.org/10.1136/ard.2006.056937 PMID: 17158139; PubMed Central PMCID: PMC1798403.

26. Weyand CM, Schmidt D, Wagner U, Goronzy JJ. The influence of sex on the phenotype of rheumatoid arthritis. Arthritis Rheum. 1998; 41(5):817–22. Epub 1998/05/20. https://doi.org/10.1002/1529-0131(199805)41:5<817::AID-ART7>3.0.CO;2-S PMID: 9588732.

27. Cutolo M, Villaggio B, Craviotto C, Pizzorni C, Seriolo B, Sulli A. Sex hormones and rheumatoid arthritis. Autoimmun Rev. 2002; 1(5):284–9. Epub 2003/07/10. PMID: 12848982.

28. Tavasolian F, Abdollahi E, Rezaei R, Momtazi-Borojeni AA, Henrotin Y, Sahebkar A. Altered Expression of MicroRNAs in Rheumatoid Arthritis. J Cell Biochem. 2018; 119(1):478–87. Epub 2017/06/10. https://doi.org/10.1002/jcb.26205 PMID: 28598026.

29. Nakasa T, Miyaki S, Okubo A, Hashimoto M, Nishida K, Ochi M, et al. Expression of microRNA-146 in rheumatoid arthritis synovial tissue. Arthritis Rheum. 2008; 58(5):1284–92. Epub 2008/04/29. https://doi.org/10.1002/art.23429 PMID: 18438844; PubMed Central PMCID: PMC2749927.

30. McInnes IB, Schett G. The pathogenesis of rheumatoid arthritis. N Engl J Med. 2011; 365(23):2205–19. Epub 2011/12/14. https://doi.org/10.1056/NEJMra1004965 PMID: 22150039.

31. Gravallese EM, Manning C, Tsay A, Naito A, Pan C, Amento E, et al. Synovial tissue in rheumatoid arthritis is a source of osteoclast differentiation factor. Arthritis Rheum. 2000; 43(2):250–8. Epub 2000/02/29. https://doi.org/10.1002/1529-0131(200002)43:2<250::AID-ANR3>3.0.CO;2-P PMID: 10693863.

32. Quinlan JR. C4. 5: programs for machine learning: Elsevier; 2014.

33. Sha F, Saul LK, editors. Analysis and extension of spectral methods for nonlinear dimensionality reduction. Proceedings of the 22nd international conference on Machine learning; 2005: ACM.

34. Borg I, Groenen P. Modern multidimensional scaling: theory and applications. Journal of Educational Measurement. 2003; 40(3):277–80.

35. Sammon JW. A nonlinear mapping for data structure analysis. IEEE Transactions on computers. 1969; 100(5):401–9.

36. The Gene Ontology C. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res. 2017; 45(D1):D331–D8. Epub 2016/12/03. https://doi.org/10.1093/nar/gkw1108 PMID: 27899567; PubMed Central PMCID: PMC5210579.

37. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25(1):25–9. Epub 2000/05/10. https://doi.org/10.1038/75556 PMID: 10802651; PubMed Central PMCID: PMC3037419.

38. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000; 28(1):27–30. https://doi.org/10.1093/nar/28.1.27 PMID: 10592173

39. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018; 46(D1):D649–D55. Epub 2017/11/18. https://doi.org/10.1093/nar/gkx1132 PMID: 29145629; PubMed Central PMCID: PMC5753187.

40. Zacharski LR, Brown FE, Memoli VA, Kisiel W, Kudryk BJ, Hunt JA, et al. Pathways of coagulation activation in situ in rheumatoid synovial tissue. Clinical immunology and immunopathology. 1992; 63(2):155–62. PMID: 1611717

41. Onuora S. Blood coagulation factor drives arthritis pathogenesis. blood. 2014; 8:594754.

42. Knijff-Dutmer E, Koerts J, Nieuwland R, Kalsbeek-Batenburg E, Van De Laar M. Elevated levels of platelet microparticles are associated with disease activity in rheumatoid arthritis. Arthritis & Rheumatism: Official Journal of the American College of Rheumatology. 2002; 46(6):1498–503.

43. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, et al. GeneCards Version 3: the human gene integrator. Database (Oxford). 2010; 2010:baq020. Epub 2010/08/07. https://doi.org/10.1093/database/baq020 PMID: 20689021; PubMed Central PMCID: PMC2938269.

44. Besmond C, Valla D, Hubert L, Poirier K, Grosse B, Guettier C, et al. Mutations in the novel gene FOPV are associated with familial autosomal dominant and non-familial obliterative portal venopathy. Liver Int. 2018; 38(2):358–64. Epub 2017/08/10. https://doi.org/10.1111/liv.13547 PMID: 28792652.

45. Sugiyama T, Ishii S, Yamamoto J-i, Irie R, Saito K, Otuki T, et al. cDNA macroarray analysis of gene expression in synoviocytes stimulated with TNFα. FEBS letters. 2002; 517(1–3):121–8. https://doi.org/10.1016/s0014-5793(02)02588-7 PMID: 12062421

46. Malhotra S, Shafiq N, Pandhi P. COX-2 inhibitors: a CLASS act or Just VIGORously promoted. MedGenMed. 2004; 6(1):6. Epub 2004/06/23. PMID: 15208519; PubMed Central PMCID: PMC1140734.

47. Crofford LJ. COX-1 and COX-2 tissue expression: implications and predictions. J Rheumatol Suppl. 1997; 49:15–9. Epub 1997/07/01. PMID: 9249646.

48. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Res. 2018; 46(D1):D754–D61. Epub 2017/11/21. https://doi.org/10.1093/nar/gkx1098 PMID: 29155950; PubMed Central PMCID: PMC5753206.

49. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO III, et al. 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis & Rheumatism. 2010; 62(9):2569–81.

50. Frank E, Witten IH. Generating accurate rule sets without global optimization. 1998.

51. Quinlan JR. Induction of decision trees. Machine learning. 1986; 1(1):81–106.

52. Kira K, Rendell LA. A practical approach to feature selection. Machine Learning Proceedings 1992: Elsevier; 1992. p. 249–56.

53. Parra-Torres NM, Cázares-Raga FE, Kouri JB. Proteomic analysis of rat cartilage: the identification of differentially expressed proteins in the early stages of osteoarthritis. Proteome science. 2014; 12 (1):55. https://doi.org/10.1186/s12953-014-0055-0 PMID: 25435813

54. Jin M, Ishida M, Katoh-Fukui Y, Tsuchiya R, Higashinakagawa T, Ikegami S, et al. Reduced pain sensitivity in mice lacking latexin, an inhibitor of metallocarboxypeptidases. Brain Res. 2006; 1075 (1):117–21. Epub 2006/02/14. https://doi.org/10.1016/j.brainres.2005.12.099 PMID: 16469302.

55. Galligan C, Baig E, Bykerk V, Keystone E, Fish E. Distinctive gene expression signatures in rheumatoid arthritis synovial tissue fibroblast cells: correlates with disease activity. Genes and immunity. 2007; 8(6):480. https://doi.org/10.1038/sj.gene.6364400 PMID: 17568789

56. Ashurst J, Chen C-K, Gilbert JG, Jekosch K, Keenan S, Meidl P, et al. The vertebrate genome annotation (Vega) database. Nucleic acids research. 2005;33(suppl_1):D459-D65.

57. Batliwalla F, Baechler E, Xiao X, Li W, Balasubramanian S, Khalili H, et al. Peripheral blood gene expression profiling in rheumatoid arthritis. Genes and immunity. 2005; 6(5):388. https://doi.org/10.1038/sj.gene.6364209 PMID: 15973463

58. Niu X, Lu C, Xiao C, Zhang Z, Jiang M, He D, et al. The shared crosstalk of multiple pathways involved in the inflammation between rheumatoid arthritis and coronary artery disease based on a digital gene expression profile. PLoS One. 2014; 9(12):e113659. Epub 2014/12/17. https://doi.org/10.1371/journal.pone.0113659 PMID: 25514790; PubMed Central PMCID: PMC4267808.

59. Teixeira VH, Olaso R, Martin-Magniette ML, Lasbleiz S, Jacq L, Oliveira CR, et al. Transcriptome analysis describing new immunity and defense genes in peripheral blood mononuclear cells of rheumatoid arthritis patients. PLoS One. 2009; 4(8):e6803. Epub 2009/08/28. https://doi.org/10.1371/journal.pone.0006803 PMID: 19710928; PubMed Central PMCID: PMC2729373.

60. 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(3):170–9. Epub 2012/03/10. https://doi.org/10.3109/03009742.2011.623137 PMID: 22401175; PubMed Central PMCID: PMC3400100.

61. D'hont A, Denoeud F, Aury J-M, Baurens F-C, Carreel F, Garsmeur O, et al. The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature. 2012; 488(7410):213. https://doi.org/10.1038/nature11241 PMID: 22801500

62. Aidinis V, Carninci P, Armaka M, Witke W, Harokopos V, Pavelka N, et al. Cytoskeletal rearrangements in synovial fibroblasts as a novel pathophysiological determinant of modeled rheumatoid arthritis. PLoS genetics. 2005; 1(4):e48. https://doi.org/10.1371/journal.pgen.0010048 PMID: 16254600

63. Vasilopoulos Y, Gkretsi V, Armaka M, Aidinis V, Kollias G. Actin cytoskeleton dynamics linked to synovial fibroblast activation as a novel pathogenic principle in TNF-driven arthritis. Annals of the rheumatic diseases. 2007; 66(suppl 3):iii23–iii8.

64. Fish EN. The X-files in immunity: sex-based differences predispose immune responses. Nature Reviews Immunology. 2008; 8(9):737. https://doi.org/10.1038/nri2394 PMID: 18728636

65. Furer V, Greenberg JD, Attur M, Abramson SB, Pillinger MH. The role of microRNA in rheumatoid arthritis and other autoimmune diseases. Clin Immunol. 2010; 136(1):1–15. Epub 2010/03/13. https://doi.org/10.1016/j.clim.2010.02.005 PMID: 20223711.

66. Nakamachi Y, Kawano S, Takenokuchi M, Nishimura K, Sakai Y, Chin T, et al. MicroRNA-124a is a key regulator of proliferation and monocyte chemoattractant protein 1 secretion in fibroblast-like synoviocytes from patients with rheumatoid arthritis. Arthritis Rheum. 2009; 60(5):1294–304. Epub 2009/05/01. https://doi.org/10.1002/art.24475 PMID: 19404929.

67. Li G, Zhang Y, Qian Y, Zhang H, Guo S, Sunagawa M, et al. Interleukin-17A promotes rheumatoid arthritis synoviocytes migration and invasion under hypoxia by increasing MMP2 and MMP9 expression through NF-kappaB/HIF-1alpha pathway. Mol Immunol. 2013; 53(3):227–36. Epub 2012/09/11. https://doi.org/10.1016/j.molimm.2012.08.018 PMID: 22960198.

68. Konisti S, Kiriakidis S, Paleolog EM. Hypoxia—a key regulator of angiogenesis and inflammation in rheumatoid arthritis. Nat Rev Rheumatol. 2012; 8(3):153–62. Epub 2012/02/02. https://doi.org/10.1038/nrrheum.2011.205 PMID: 22293762.

69. Pollard L, Choy E, Gonzalez J, Khoshaba B, Scott D. Fatigue in rheumatoid arthritis reflects pain, not disease activity. Rheumatology. 2006; 45(7):885–9. https://doi.org/10.1093/rheumatology/kel021 PMID: 16449363

70. Nicklin J, Cramp F, Kirwan J, Greenwood R, Urban M, Hewlett S. Measuring fatigue in rheumatoid arthritis: A cross-sectional study to evaluate the bristol rheumatoid arthritis fatigue multi-dimensional questionnaire, visual analog scales, and numerical rating scales. Arthritis care & research. 2010; 62 (11):1559–68.

71. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database C. The sequence read archive. Nucleic Acids Res. 2011;39(Database issue):D19-21. Epub 2010/11/11. https://doi.org/10.1093/nar/gkq1019 PubMed PMID: 21062823; PubMed Central PMCID: PMC3013647.

72. Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P, et al. The UCSC Genome Browser database: 2016 update. Nucleic Acids Res. 2016; 44(D1):D717–25. Epub 2015/11/22. https://doi.org/10.1093/nar/gkv1275 PMID: 26590259; PubMed Central PMCID: PMC4702902.

73. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25(9):1105–11. Epub 2009/03/18. https://doi.org/10.1093/bioinformatics/btp120 PMID: 19289445; PubMed Central PMCID: PMC2672628.

74. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9(4):357–9. Epub 2012/03/06. https://doi.org/10.1038/nmeth.1923 PMID: 22388286; PubMed Central PMCID: PMC3322381.

75. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014; 30(7):923–30. Epub 2013/11/15. https://doi.org/10.1093/bioinformatics/btt656 PMID: 24227677.

76. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010; 28(5):511–5. Epub 2010/05/04. https://doi.org/10.1038/nbt.1621 PMID: 20436464; PubMed Central PMCID: PMC3146043.

77. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26(1):139–40. Epub 2009/11/17. https://doi.org/10.1093/bioinformatics/btp616 PMID: 19910308; PubMed Central PMCID: PMC2796818.

78. Yang J, Huang T, Petralia F, Long Q, Zhang B, Argmann C, et al. Synchronized age-related gene expression changes across multiple tissues in human and the link to complex diseases. Sci Rep. 2015; 5:15145. Epub 2015/10/20. https://doi.org/10.1038/srep15145 PMID: 26477495; PubMed Central PMCID: PMC4609956.

79. Radstake TR, van der Voort R, ten Brummelhuis M, de Waal Malefijt M, Looman M, Figdor CG, et al. Increased expression of CCL18, CCL19, and CCL17 by dendritic cells from patients with rheumatoid arthritis, and regulation by Fc gamma receptors. Ann Rheum Dis. 2005; 64(3):359–67. Epub 2004/08/28. https://doi.org/10.1136/ard.2003.017566 PMID: 15331393; PubMed Central PMCID: PMC1755402.

80. Flytlie HA, Hvid M, Lindgreen E, Kofod-Olsen E, Petersen EL, Jorgensen A, et al. Expression of MDC/CCL22 and its receptor CCR4 in rheumatoid arthritis, psoriatic arthritis and osteoarthritis. Cytokine. 2010; 49(1):24–9. Epub 2009/11/28. https://doi.org/10.1016/j.cyto.2009.10.005 PMID: 19942450.

81. Lee AY, Korner H. CCR6 and CCL20: emerging players in the pathogenesis of rheumatoid arthritis. Immunol Cell Biol. 2014; 92(4):354–8. Epub 2014/01/08. https://doi.org/10.1038/icb.2013.97 PMID: 24394994.

82. Rodriguez PC, Torres-Moya R, Reyes G, Molinero C, Prada D, Lopez AM, et al. A clinical exploratory study with itolizumab, an anti-CD6 monoclonal antibody, in patients with rheumatoid arthritis. Results Immunol. 2012; 2:204–11. Epub 2012/01/01. https://doi.org/10.1016/j.rinim.2012.11.001 PMID: 24371585; PubMed Central PMCID: PMC3862386.

83. Valencia X, Higgins JM, Kiener HP, Lee DM, Podrebarac TA, Dascher CC, et al. Cadherin-11 provides specific cellular adhesion between fibroblast-like synoviocytes. J Exp Med. 2004; 200(12):1673–9.

Epub 2004/12/22. https://doi.org/10.1084/jem.20041545 PMID: 15611293; PubMed Central PMCID: PMC2211995.

84. Castaneda-Delgado JE, Bastian-Hernandez Y, Macias-Segura N, Santiago-Algarra D, Castillo-Ortiz JD, Aleman-Navarro AL, et al. Type I Interferon Gene Response Is Increased in Early and Established Rheumatoid Arthritis and Correlates with Autoantibody Production. Front Immunol. 2017; 8:285. Epub 2017/04/05. https://doi.org/10.3389/fimmu.2017.00285 PMID: 28373872; PubMed Central PMCID: PMC5357778.

85. Corvaisier M, Delneste Y, Jeanvoine H, Preisser L, Blanchard S, Garo E, et al. IL-26 is overexpressed in rheumatoid arthritis and induces proinflammatory cytokine production and Th17 cell generation. PLoS Biol. 2012; 10(9):e1001395. Epub 2012/10/12. https://doi.org/10.1371/journal.pbio.1001395 PMID: 23055831; PubMed Central PMCID: PMC3463509.

86. Ruyssen-Witrand A, Lukas C, Nigon D, Dawidowicz K, Morel J, Sibilia J, et al. Association of IL-2RA and IL-2RB genes with erosive status in early rheumatoid arthritis patients (ESPOIR and RMP cohorts). Joint Bone Spine. 2014; 81(3):228–34. Epub 2013/11/10. https://doi.org/10.1016/j.jbspin.2013.10.002 PMID: 24200909.

87. Tolboom TC, Pieterman E, van der Laan WH, Toes RE, Huidekoper AL, Nelissen RG, et al. Invasive properties of fibroblast-like synoviocytes: correlation with growth characteristics and expression of MMP-1, MMP-3, and MMP-10. Ann Rheum Dis. 2002; 61(11):975–80. Epub 2002/10/16. https://doi.org/10.1136/ard.61.11.975 PMID: 12379519; PubMed Central PMCID: PMC1753950.

88. Liu M, Sun H, Wang X, Koike T, Mishima H, Ikeda K, et al. Association of increased expression of macrophage elastase (matrix metalloproteinase 12) with rheumatoid arthritis. Arthritis Rheum. 2004; 50 (10):3112–7. Epub 2004/10/12. https://doi.org/10.1002/art.20567 PMID: 15476203.

89. Vincenti MP, Brinckerhoff CE. Transcriptional regulation of collagenase (MMP-1, MMP-13) genes in arthritis: integration of complex signaling pathways for the recruitment of gene-specific transcription factors. Arthritis Res. 2002; 4(3):157–64. Epub 2002/05/16. https://doi.org/10.1186/ar401 PMID: 12010565; PubMed Central PMCID: PMC128926.

90. Burrage PS, Mix KS, Brinckerhoff CE. Matrix metalloproteinases: role in arthritis. Front Biosci. 2006; 11(1):529–43.

91. Team RC. R: A language and environment for statistical computing. 2015.

92. Van Der Maaten L, Postma E, Van den Herik J. Dimensionality reduction: a comparative. J Mach Learn Res. 2009; 10:66–71.

93. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008; 9:559. Epub 2008/12/31. https://doi.org/10.1186/1471-2105-9-559 PMID: 19114008; PubMed Central PMCID: PMC2631488.

94. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics. 2007; 23(2):257–8. Epub 2006/11/14. https://doi.org/10.1093/bioinformatics/btl567 PMID: 17098774.

95. Subhash S, Kanduri C. GeneSCF: a real-time based functional enrichment tool with support for multiple organisms. BMC Bioinformatics. 2016; 17(1):365. Epub 2016/09/14. https://doi.org/10.1186/s12859-016-1250-z PMID: 27618934; PubMed Central PMCID: PMC5020511.

96. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PloS one. 2011; 6(7):e21800. https://doi.org/10.1371/journal.pone.0021800 PMID: 21789182

97. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13(11):2498–504. Epub 2003/11/05. https://doi.org/10.1101/gr.1239303 PMID: 14597658; PubMed Central PMCID: PMC403769.

98. 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(11):e13984. https://doi.org/10.1371/journal.pone.0013984 PMID: 21085593

99. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2018. Epub 2018/07/18. https://doi.org/10.1093/bioinformatics/bty633 PMID: 30016406.

100. Adler D. vioplot: Violin plot. R package version 02, URL http://CRANR-projectorg/package=vioplot. 2005.

101. Wickham H, Francois R, Henry L, Müller K. dplyr: A grammar of data manipulation. R package version 04. 2015;3.

102. Wickham H. ggplot2: elegant graphics for data analysis: Springer; 2016.

103. Slowikowski K. ggrepel: Repulsive text and label geoms for 'ggplot2'. R package version 06. 2016; 5 (11).

104. Lê S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. Journal of statistical software. 2008; 25(1):1–18.

105. Kassambara A, Mundt F. Factoextra: extract and visualize the results of multivariate data analyses. R package version. 2016; 1(3).

106. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009; 19(9):1639–45. Epub 2009/06/23. https://doi.org/10.1101/gr.092759.109 PMID: 19541911; PubMed Central PMCID: PMC2752132.

107. Frank E, Hall M, Witten I. The WEKA workbench. Data mining: Practical machine learning tools and techniques. 2016; 4.

108. Freund Y, Mason L, editors. The alternating decision tree learning algorithm. icml; 1999.

109. John GH, Langley P, editors. Estimating continuous distributions in Bayesian classifiers. Proceedings of the Eleventh conference on Uncertainty in artificial intelligence; 1995: Morgan Kaufmann Publishers Inc.

110. Platt J. Fast Training of Support Vector Machines using Sequential Minimal Optimization. Smola BSaCBaA, editor: MIT Press; 1998.