



UNIVERSITÉ
DE NAMUR

Institutional Repository - Research Portal Dépôt Institutionnel - Portail de la Recherche

researchportal.unamur.be

RESEARCH OUTPUTS / RÉSULTATS DE RECHERCHE

Blood transcriptomics to facilitate diagnosis and stratification in pediatric rheumatic diseases – a proof of concept study

Ha, My Kieu; Bartholomeus, Esther; Van Os, Luc; Dandelooy, Julie; Leysen, Julie; Aerts, Olivier; Siozopoulou, Vasiliki; De Smet, Eline; Gielen, Jan; Guerti, Khadija; De Maeseneer, Michel; Herregods, Nele; Lechkar, Bouchra; Wittoek, Ruth; Geens, Elke; Claes, Laura; Zaqout, Mahmoud; Dewals, Wendy; Lemay, Annelies; Tuerlinckx, David; Weynants, David; Vanlede, Koen; van Berlaer, Gerlant; Raes, Marc; Verhelst, Helene; Boiy, Tine; Van Damme, Pierre; Jansen, Anna C.; Meuwissen, Marije; Sabato, Vito; Van Camp, Guy; Suls, Arvid; Werff ten Bosch, Jutte Van der; Dehoorne, Joke; Joos, Rik; Laukens, Kris; Meysman, Pieter; Ogunjimi, Benson

Published in:

Pediatric Rheumatology

DOI:

[10.1186/s12969-022-00747-x](https://doi.org/10.1186/s12969-022-00747-x)

Publication date:

2022

Document Version

Publisher's PDF, also known as Version of record

[Link to publication](#)

Citation for published version (HARVARD):

Ha, MK, Bartholomeus, E, Van Os, L, Dandelooy, J, Leysen, J, Aerts, O, Siozopoulou, V, De Smet, E, Gielen, J, Guerti, K, De Maeseneer, M, Herregods, N, Lechkar, B, Wittoek, R, Geens, E, Claes, L, Zaqout, M, Dewals, W, Lemay, A, Tuerlinckx, D, Weynants, D, Vanlede, K, van Berlaer, G, Raes, M, Verhelst, H, Boiy, T, Van Damme, P, Jansen, AC, Meuwissen, M, Sabato, V, Van Camp, G, Suls, A, Werff ten Bosch, JVD, Dehoorne, J, Joos, R, Laukens, K, Meysman, P & Ogunjimi, B 2022, 'Blood transcriptomics to facilitate diagnosis and stratification in pediatric rheumatic diseases – a proof of concept study', *Pediatric Rheumatology*, vol. 20, no. 1, 91. <https://doi.org/10.1186/s12969-022-00747-x>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.


- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

RESEARCH ARTICLE

Open Access



Blood transcriptomics to facilitate diagnosis and stratification in pediatric rheumatic diseases – a proof of concept study

My Kieu Ha^{1,2,3*} , Esther Bartholomeus^{2,3,4}, Luc Van Os⁵, Julie Dandelooy⁶, Julie Leysen^{6,7}, Olivier Aerts^{6,7}, Vasiliki Siozopoulou⁸, Eline De Smet⁹, Jan Gielen^{9,10}, Khadija Guerti¹¹, Michel De Maeseneer¹², Nele Herregods¹³, Bouchra Lechkar¹⁴, Ruth Wittoek^{15,16}, Elke Geens¹⁶, Laura Claes¹⁷, Mahmoud Zaqout^{18,19}, Wendy Dewals¹⁸, Annelies Lemay²⁰, David Tuerlinckx^{21,22}, David Weynants²³, Koen Vanlede²⁴, Gerlant van Berlaer²⁵, Marc Raes²⁶, Helene Verhelst²⁷, Tine Bojy²⁸, Pierre Van Damme^{2,29}, Anna C. Jansen¹⁷, Marije Meuwissen³, Vito Sabato^{14,30}, Guy Van Camp⁴, Arvid Suls^{2,2}, Jutte Van der Werff ten Bosch³¹, Joke Dehoorne³², Rik Joos^{16,28,30,32}, Kris Laukens^{2,33,34}, Pieter Meysman^{2,33,34} and Benson Ogunjimi^{1,2,16,28,30,35*}

Abstract

Background Transcriptome profiling of blood cells is an efficient tool to study the gene expression signatures of rheumatic diseases. This study aims to improve the early diagnosis of pediatric rheumatic diseases by investigating patients' blood gene expression and applying machine learning on the transcriptome data to develop predictive models.

Methods RNA sequencing was performed on whole blood collected from children with rheumatic diseases. Random Forest classification models were developed based on the transcriptome data of 48 rheumatic patients, 46 children with viral infection, and 35 controls to classify different disease groups. The performance of these classifiers was evaluated by leave-one-out cross-validation. Analyses of differentially expressed genes (DEG), gene ontology (GO), and interferon-stimulated gene (ISG) score were also conducted.

Results Our first classifier could differentiate pediatric rheumatic patients from controls and infection cases with high area-under-the-curve (AUC) values ($AUC = 0.8 \pm 0.1$ and 0.7 ± 0.1 , respectively). Three other classifiers could distinguish chronic recurrent multifocal osteomyelitis (CRMO), juvenile idiopathic arthritis (JIA), and interferonopathies (IFN) from control and infection cases with $AUC \geq 0.8$. DEG and GO analyses reveal that the pathophysiology of CRMO, IFN, and JIA involves innate immune responses including myeloid leukocyte and granulocyte activation, neutrophil activation and degranulation. IFN is specifically mediated by antibacterial and antifungal defense responses, CRMO by cellular

*Correspondence:

My Kieu Ha
my.ha@uantwerpen.be
Benson Ogunjimi
benson.ogunjimi@uantwerpen.be

Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

response to cytokine, and JIA by cellular response to chemical stimulus. IFN patients particularly had the highest mean ISG score among all disease groups.

Conclusion Our data show that blood transcriptomics combined with machine learning is a promising diagnostic tool for pediatric rheumatic diseases and may assist physicians in making data-driven and patient-specific decisions in clinical practice.

Keywords Pediatric rheumatic diseases, RNA sequencing, Blood transcriptomics, Classification model

Background

Pediatric rheumatic diseases encompass a spectrum of autoimmune and autoinflammatory diseases that can affect the joints, muscles, bones, and other organs in children under the age of 16–18 years. Although many pediatric rheumatic diseases typically present with joint manifestations, other organs, including the eyes, skin, muscles, and gastrointestinal tract, may also be affected. Children who present with rheumatic symptoms often pose several challenges to their physicians. First, typical symptoms such as fever, rash, redness, pain and/or swelling at joints are common in many rheumatic diseases: (i) a transient/self-limiting process such as (reactive) infectious arthritis, (ii) a relapsing process like autoinflammatory diseases (AID), (iii) a chronic condition like vasculitis, chronic recurrent multifocal osteomyelitis (CRMO), or juvenile idiopathic arthritis (JIA), (iv) an interferonopathy (IFN) such as dermatomyositis or systemic lupus erythematosus (SLE) characterized by dysregulation of type I interferon, or (v) diseases related to the human leukocyte antigen B51 serotype (HLA-B51). Second, once a disease has been confirmed, it is still difficult to make a rapid and definitive classification amongst the different subtypes within the disease due to its rarity and clinical presentation variability. This has challenged physicians' efforts in making specific diagnoses and assigning proper treatments.

Transcriptome profiling of blood cells has proven to be useful and efficacious in identifying gene expression signatures in rheumatic diseases [1, 2], enabling physicians to make data-driven and patient-specific decisions. Thanks to the use of transcriptomics, the participation of type I interferons participate in the pathophysiology of SLE [3] and dermatomyositis [4] was discovered and has become one of the paramount findings in rheumatology. Psoriatic arthritis is another exemplary disease demonstrating the usefulness of applying transcriptomics to identify important pathways related to interleukins IL-12, IL-17 and IL-23 [1]. Large datasets of gene expression generated from such studies are highly interesting, yet their systematic analysis and interpretation is quite challenging. Thus, there has been considerable interest in applying machine learning on blood cells' gene expression in order to obtain new insights into the pathophysiology of rheumatic diseases which in turn may have

important implications for their clinical management [5, 6].

By investigating the whole blood gene expression of children with rheumatic diseases in comparison with reactive/post-infection controls, we aim to develop computational classifiers based on the obtained transcriptome data that allow us to identify pediatric patients with rheumatic diseases and distinguish different rheumatic groups (e.g., CRMO, JIA, and IFN), and thus, ultimately, to improve the diagnosis of future patients.

Methods

Patients and controls

After obtaining written consent, 48 children (1 to 16 years old) with rheumatic diseases (i.e., AID, CRMO, IFN, JIA, vasculitis, and HLA-B51 related rheumatic diseases) were recruited, prior to any treatment except non-steroidal anti-inflammatory drugs (NSAIDs), from May 2016 until August 2020, at the Divisions of Pediatric Rheumatology of four hospitals in Belgium (Antwerp Hospital Network, Antwerp University Hospital, Brussels University Hospital, and Ghent University Hospital). Venous blood was collected into PAXgene® Blood RNA tubes (PreAnalytiX, Switzerland). Clinical details of the patients can be found in Table S1. As controls, 46 children with PCR-confirmed viral (mainly enterovirus) infections were also recruited and requested to provide blood twice: first while being actively infected, and second following remission [7]. Only 35 children agreed to a second venapuncture as it was not obligatory.

RNA extraction

PAXgene® Blood RNA tubes were kept at -80 °C within 72 h after blood collection until use. RNA extraction was performed via a column-based RNA extraction using the PAXgene® Blood RNA extraction kit (Qiagen, Germany). To optimize RNA concentrations, we used the RNA Clean & Concentrator™-5 kit (Zymo Research, USA). We verified the RNA quality using the RNA ScreenTape Analysis on the TapeStation (Agilent, USA).

3' mRNA library preparation and sequencing

All RNA samples were prepared with the QuantSeq3' mRNA-Seq Library Prep Kit for Illumina (Lexogen GmbH, Austria) following the standard supplier's

protocol for long fragments. During the RNA removal step, we also added globin blockers, so none of the abundant globin mRNA was copied to double stranded cDNA. The resulting amplified cDNA libraries were equimolarly pooled and sequenced on NextSeq 500 (high output v2,5 kit, 150 cycli, single read) (Illumina, USA) with up to 40 samples per batch. This gave us an optimum of 10 million reads for each sample. All samples were prepared and sequenced in 4 batches.

Raw data processing

Raw data from the NextSeq was demultiplexed and further processed per batch through an in-house pipeline. The quality of all reads was evaluated using FastQC (v0.11.9) before and after processing with Trimmomatic (v0.36). Trimmomatic removed the leading 20 bases from reads, ensure a minimum quality score of 15 over a sliding window of 4 bases and required a minimum read length of 30 bases. As usage of oligodT primers might cause poly-A stretches at the 3' end, the latter end was trimmed with our own in-house poly-A removal script. All sequences that remained after trimming were mapped against the human reference genome build 38 (polymorph variants excluded) with HISAT2 (v2.0.4). HTseq (v0.6.1) was used to count all reads for each gene and generate a readcount table.

Cluster identification

Clustering is an unsupervised learning algorithm that is used to discover patterns in high-dimensional data that would not be easily identified by conventional statistics. The readcount table obtained after raw data processing was normalized by median scaling. Then, the t-distributed stochastic neighbor embedding (t-SNE) method and hierarchical clustering algorithms were applied on the normalized readcount data to assign data into different clusters based on the (dis)similarity in the gene expression between samples.

Differential gene expression and gene ontology enrichment analyses

Preliminary filtering of the normalized readcount data was performed by removing genes with fewer than 10 readcounts over all samples. Differential gene expression analyses were performed using the DESeq2 Bioconductor package in the open-source statistical software R. To account for batch effects, the batch number was included in the DESeq2 design. Differentially expressed genes with \log_2 fold change ≥ 2 (either up- or down-regulated) and p -value < 0.01 (adjusted for multiple testing by the False Discovery Rate method) were passed on to gene ontology enrichment analysis using the topGO package in R. Fisher's exact test was performed to determine significantly

enriched/depleted gene ontology terms relating to biological processes.

Classifier development

Due to the larger number of genes in the dataset, a feature selection step was performed using the Boruta package in R [8] with a Bonferroni correction to identify genes that had good predictive power for disease classification. The Boruta algorithm was applied on the normalized gene expression values obtained from DESeq2. With the Boruta selected genes, classification models were trained using the Random Forest algorithm [9], also in R. Validation of the trained classifiers was performed using a leave-one-out cross-validation strategy. In this strategy, a single sample would be removed from the dataset and a classification model was trained on the remaining samples. The model was then used to classify the left-out sample. This process would be repeated for all samples in the dataset. Receiver-operator-characteristic (ROC) curves and the area-under-the-curve (AUC) were employed using the package pROC [10] in R to evaluate the classifier performance. AUC values were displayed in figures as mean \pm confidence interval (95%).

Calculation of interferon-stimulated genes' scores

First, relative expression (RE) was calculated based on the readcounts of interferon-stimulated genes (ISG) and the housekeeping gene GAPDH: $RE = 2^{-(\text{count}_{ISG} - \text{count}_{GAPDH})}$. The ISG score was calculated by summing up the individual RE per gene after normalization to the control group as follows: $\sum (RE_{\text{subject}} - \text{Mean}_{\text{control}}) / \text{Standard Deviation}_{\text{control}}$. 28 ISG were selected for ISG score calculation (Table S4) [11]. Statistical significance was assessed using Mann-Whitney test. * $P < 0.05$; ** $P < 0.005$.

Results

Transcriptome profiles of rheumatic diseases, viral infection, and convalescent controls

We compared the transcriptome profiles of six rheumatic disease groups (i.e., JIA, AID, CRMO, HLA-B51, IFN, and vasculitis) with viral infection and convalescent controls. Clustering analyses using t-SNE and hierarchical algorithms were displayed in Fig. 1 and Figure S1, respectively. As shown in the t-SNE plot in Fig. 1, most controls were gathered in cluster 1 while infection cases were grouped into a separate cluster 2, which implies that the gene expression of actively infected cases and remission cases (i.e., controls) is substantially independent despite coming from the same participants. However, patients with different rheumatic diseases were not well distinguished and assigned mostly to cluster 3, while cluster 4 contained a mixture of different categories.

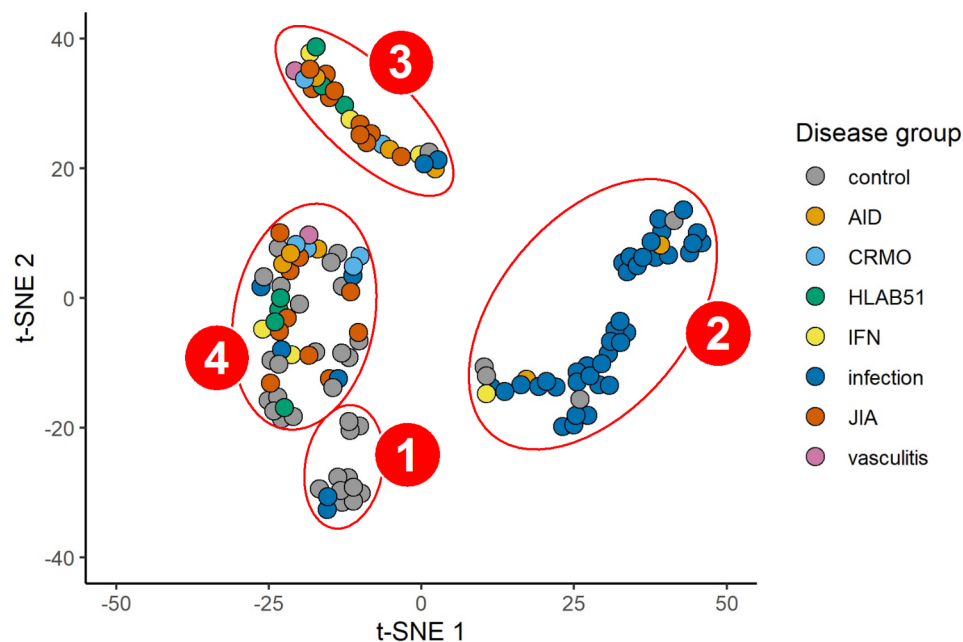


Fig. 1 t-SNE plot of 4 different clusters

Classifier development

The Random Forest algorithm was used for classifier development because it uses the ensemble learning technique that is robust to outliers, stable with new data, and can handle non-linear correlations. The first classifier was developed to distinguish between control, infection, and pediatric rheumatic cases based on normalized transcriptome data. Leave-one-out cross-validation results in Fig. 2a confirm that the classifier could differentiate pediatric rheumatic patients from negative controls ($AUC=0.8 \pm 0.1$) and from viral infection cases ($AUC=0.7 \pm 0.1$). The Boruta algorithm selected 349 genes out of 31,319 initial genes (Table S2) for the training of this classifier between control, infection, and pediatric rheumatic cases. Some of the notable selected genes were CD3G, CD96, and CD200R1 (CD200 receptor 1). The gene CD3G encodes the CD3 γ polypeptide, which forms a part of the CD3-TCR (T-cell receptor) complex. This complex plays an important role in antigen recognition and several intracellular signal-transduction pathways. This finding indicates that some of the rheumatic diseases are specifically connected to the alteration and malfunction of γ T-cells. Previous studies have also reported the association of γ and δ T-cells with (immunodeficiency and) autoimmune diseases [12]. CD96 is expressed on T-cells and natural killer cells. It belongs to a family of molecules that provide costimulatory and coinhibitory signals during T-cell activation. It was shown to inhibit the expansion and IL-9 production of Th17 cells and thus, reduce inflammation and pathogenicity [13]. CD200R1 is also expressed on T-cells, as well as myeloid cells. It was reported to alter the balance

between Th17 cells and regulatory T-cells in SLE patients [14] and has also been confirmed as one of the genetic factors susceptible to JIA, especially oligoarticular JIA [15]. Aberrant expression of CD200R1 was shown to contribute to abnormal Th17 cell differentiation and chemotaxis in patients with rheumatoid arthritis [15].

More specific classifiers were then developed per disease group. As the number of rheumatic patients in our dataset was limited, these classifiers focused only on CRMO, IFN, and JIA groups, which had more subjects for model training and validation than the other disease groups. Three classifiers were developed to distinguish patients with CRMO ($n=6$), IFN ($n=6$), and JIA ($n=20$) from control ($n=35$) and infection ($n=46$) cases. They worked quite well as their AUC values are above or equal to 0.8 (Fig. 2b). Since CRMO, IFN, and JIA were differentiated well from control and infection cases, it was subsequently important to examine how they could be distinguished from one another. ROC curves and AUC values of a classifier between CRMO, IFN, and JIA (Figure S2) indicated that IFN could be distinguished relatively well from CRMO and JIA ($AUC=0.7 \pm 0.2/0.3$), however CRMO is not easily differentiated from JIA ($AUC=0.5 \pm 0.3$), likely explained by the limited sample size. The Boruta-identified genes for these classifiers are also presented in Table S2. There were 349 selected genes from the CRMO-control-infection classifier, 247 genes from the IFN-control-infection classifier, and 286 genes from the JIA-control-infection classifier. As expected, more interferon-related genes were selected for the IFN classifier compared to those of CRMO and JIA.

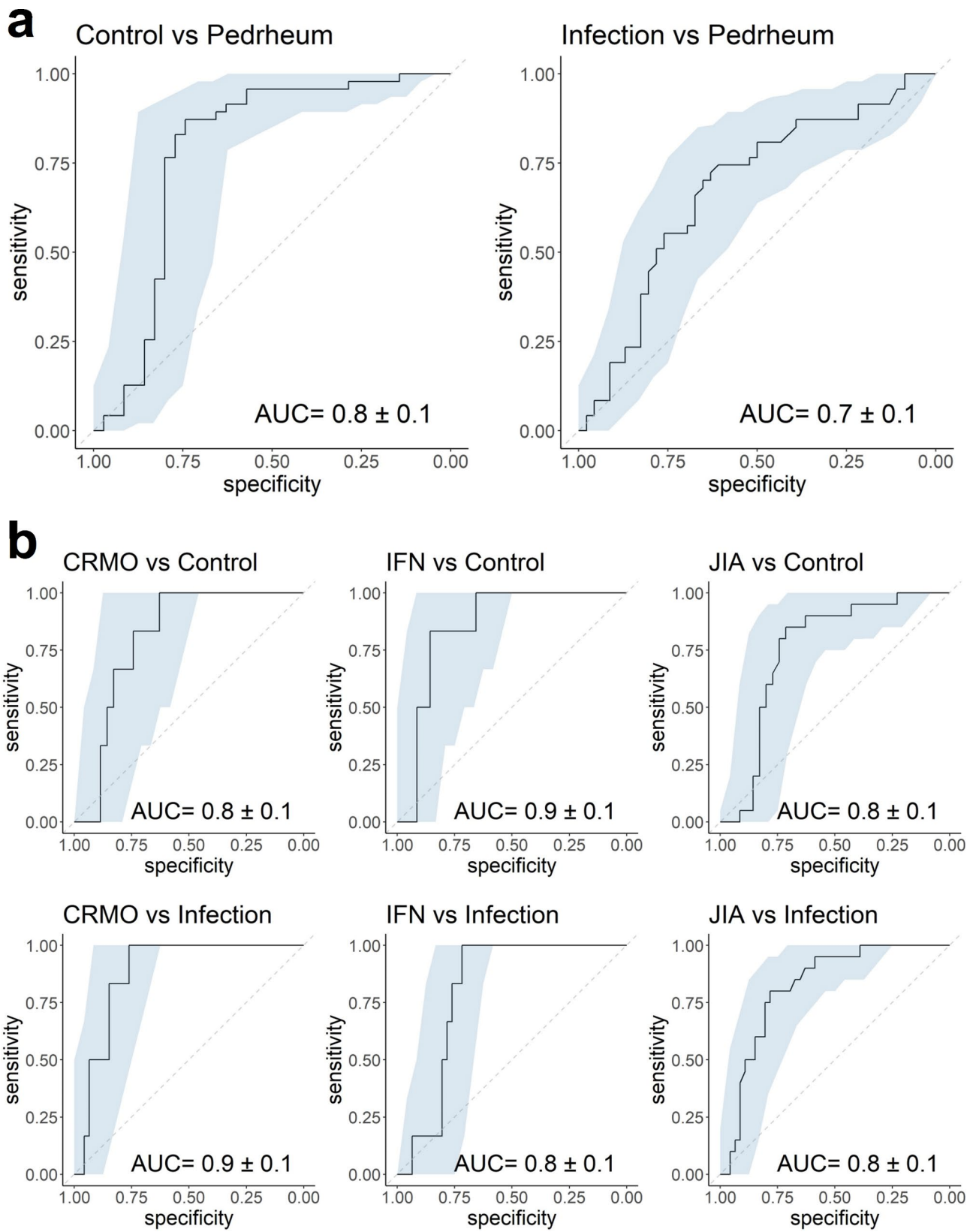


Fig. 2 ROC curves and AUC values from leave-one-out cross-validation of classifier between (a) negative controls (i.e., control), viral infected subjects (i.e., infection) and subjects with rheumatic diseases (i.e., Pedrheum); and more specifically between (b) CRMO, IFN, JIA and control/infection cases

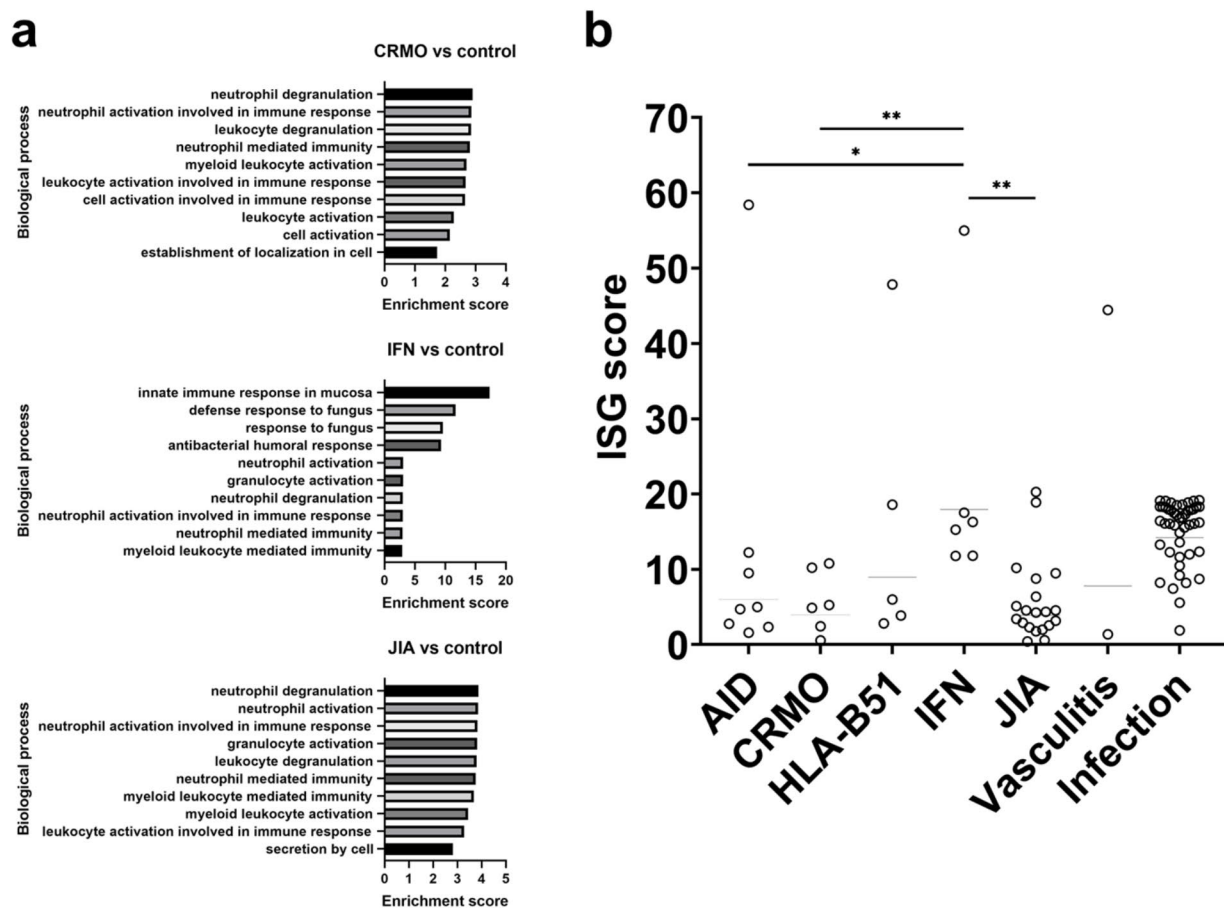


Fig. 3 (a) Gene ontology enrichment analysis of CRMO-, IFN-, and JIA-associated genes. Bar charts the top 10 GO terms for biological process. (b) ISG score by disease group; horizontal lines represent median values of each group; Mann-Whitney test for statistical significance. * $P < 0.05$; ** $P < 0.005$

Differential expression and gene ontology enrichment analyses

We analyzed the differentially expressed genes (DEGs) of CRMO, IFN, and JIA versus controls. The resulting DEGs were translated to corresponding Gene Ontology (GO) categories to understand which pathways were involved in the disease pathophysiology. Many of the top 10 GO categories of CRMO, IFN, and JIA groups are related to innate immunity including myeloid leukocyte and granulocyte activation, neutrophil activation and degranulation (Fig. 3a and Table S3). In IFN particularly, the immunity is largely mediated by antibacterial and antifungal defense responses. Results from GO analyses of CRMO, IFN, and JIA against the other Pedrheum groups are displayed in Figure S3 and Table S3. Although the classifiers could not adequately differentiate between CRMO and JIA, we noted that 1,106 DEGs could be found between CRMO and all other Pedrheum groups, 1,730 DEGs in the case of IFN, and 1,216 DEGs for JIA (Table S5). Additionally, more than 170 DEGs were found between

CRMO and IFN, CRMO and JIA, as well as between IFN and JIA (Table S5).

ISG scores

Using the whole blood gene expression obtained from 3' mRNA sequencing, we calculated the ISG scores of IFN patients and compared them with those from other disease groups. As displayed in Fig. 3b, IFN patients had the highest mean score of 18. Other disease groups, although displaying lower mean scores than IFN (6.0 for AID, 4.0 for CRMO, 9.0 for HLA-B51, 3.9 for JIA, 7.8 for vasculitis, and 14 for infection cases), did include some patients with particularly high scores: one AID patient had a score of 58, one HLA-B51 patient had score 48, and one vasculitis patient had score 44. Interestingly, longitudinal tracking of ISG scores was proven feasible using 3' mRNA sequencing. Indeed, we showed that one patient with Aicardi-Goutières syndrome had significantly high ISG scores at early presentations that decreased following initiation of JAK-inhibition via tofacitinib (see Figure S4).

Discussion

Application of transcriptomics techniques such as microarray or sequencing on blood or synovial fluid of rheumatic patients has been a key transforming factor in rheumatology [1]. In the current study, we showed that 3' mRNA from whole blood can provide adequate information for the differentiation between controls, viral infections, and pediatric rheumatic diseases. We demonstrated that the Random Forest algorithm can be applied on blood transcriptome data to determine whether a pediatric patient has reactive/post-infection phenomena or an autoimmune/autoinflammatory disease. Furthermore, studies on autoimmune and autoinflammatory diseases have largely focused on adaptive immunity, that is, regulatory and autoreactive T-cells [16]. Our study hereby provides evidence that innate immunity may also have an important role in the pathophysiology of pediatric rheumatic autoimmune and autoinflammatory diseases, including CRMO, IFN, and JIA. Moreover, we showed that the activation and immune response of myeloid cells form participate in the biological pathways underlying JIA, CRMO and IFN – three of the most common rheumatic diseases in children. Via the DEG and GO analyses, we found that the immunological activities of innate cells, such as neutrophils and granulocytes, were highly associated with CRMO, IFN, and JIA compared to the viral convalescent controls. Interestingly, in the IFN group, the immunity seems to be largely mediated by antibacterial and antifungal defense responses. This could be a sign of molecular mimicry where self-derived peptides resemble foreign antigens and thus stimulate the antigen-specific autoreactive T-cells or B-cells, which in turn results in the production of pro-inflammatory cytokines [17, 18]. One of them is type I interferon, which plays an important role in the pathophysiology of IFN disease and which also has been reported to participate in the immune response against viral, bacterial, fungal pathogens, and parasites [19]. GO analyses comparing CRMO, IFN, and JIA with the other Pedrheum groups (Figure S3 and Table S3) reveal that CRMO is specifically driven by cellular response to cytokine, JIA by cellular response to chemical stimulus, and IFN by the activation of myeloid leukocytes, neutrophils, and granulocytes. The dysregulated cytokine expression from innate immune cells has been concluded to have central contribution to the inflammatory phenotype of CRMO by Hofmann et al. [20]. JIA on the other hand has been found to be associated with antibiotics exposure in a dose- and time- dependent fashion in a large pediatric population by Horton et al. [21].

The over-expression of ISG is a useful biomarker of IFN diseases, including SLE. Although the interferon signature was first defined in SLE patients, different ISG are investigated to classify pathological conditions of other interferonopathies and lupus-like disorders (e.g.,

dermatomyositis, Aicardi-Goutières syndrome) to guide molecular diagnostics and to formulate targeted therapy approaches [22]. Quantitative polymerase chain reaction (qPCR) has been the method of choice to measure the expression of ISG and estimate the ISG score, but this approach has low throughput as qPCR can only analyze pre-determined genes and the number of genes to be analyzed simultaneously is limited [23]. Using whole blood gene expression obtained from 3' mRNA sequencing, we were able to calculate the ISG scores without gene pre-selection or gene number limitation. Our data show that the average score of IFN group was the highest among all disease categories, thereby confirming that the highly expressed ISG are signatures of interferonopathies. More importantly, we showed that robust longitudinal tracking of ISG scores is possible with the use of 3' mRNA sequencing. Beside the IFN group, viral infection cases also displayed high average ISG scores due to the participation of type I interferon in the immune responses against viral infection [19].

Despite demonstrating that blood RNA sequencing can distinguish autoimmune/autoinflammatory diseases from viral infection/post-infection cases, as well as reveal their key genes and pathways, our study is limited by the sample size of each rheumatic disease, due to which cross-validation was done to validate the classifiers' performance instead of training and testing on independent datasets. The limited number of rheumatic cases may also be responsible for the modest cross-validation performance of the CRMO-IFN-JIA classifier although rheumatic diseases in general could be identified well from controls and infections. Another limitation that challenges the predictive performance of our classifiers is disease heterogeneity. IFN and JIA are examples of heterogeneous groups that pose great difficulties for the classifiers to differentiate from other groups because they contain several disease subgroups (e.g., dermatomyositis and systemic lupus erythematosus were grouped together as IFN, and JIA included polyarticular, oligoarticular types, as well as spondyloarthropathies). The heterogeneity of diseases often obstructs explicit modelling of underlying distributions of individual features, which can be even more problematic when the sample population is small [24]. Finally, although efforts were made to minimize batch effects, these cannot always be completely avoided.

To incorporate the expression-based assessment into clinical use, it is necessary to demonstrate that the reliability and accuracy of this approach is comparable or superior to the currently used methods. The development of diagnostic criteria is challenging in rheumatology due to the heterogeneity of many rheumatic diseases, variable clinical presentations, and complex pathophysiology. Given the lack of optimal diagnostic criteria, physicians

must rely on a complicated decision-making process based on a combination of symptoms, physical examination, exclusion of competing diagnoses, and geographic prevalence [25]. Moreover, there is an ongoing concern that physical examination is insensitive in detecting subtle, smoldering synovitis [26]. Since the completion of the Human Genome Project in 2003, next-generation sequencing has seen great improvements in technique and decline in cost. In addition, the whole protocol from RNA extraction, library preparation, and sequencing until data pre-processing and disease classification using machine learning only takes 4–5 days in our experience, given that a classifier is already developed and validated. This is a similar average amount of time it would take to complete a blood test and other assessments for the diagnosis of rheumatic diseases. Thus, the approach proposed in this study, where machine learning is developed based on blood transcriptome data, should be highly affordable and applicable in clinical practice. Since gene expression variations in blood cells of rheumatic patients can predate the clinical manifestations [27], blood gene expression profiling is useful in identifying new biomarkers of pediatric rheumatic diseases and, together with the machine learning classifiers presented in this study – after further development and validation – will be an efficient tool for early diagnosis and heterogeneity exploration of pediatric rheumatic diseases. We believe that adding more cases of pediatric rheumatic diseases to the database will provide more data for the classifier to be trained on, allowing it to capture more distinct transcriptome features and variances of each disease, as well as get validated on an independent test set. It would also be useful to collect blood from the same patients over fixed time periods or before and after therapy to obtain longitudinal transcriptome data so that we can update and improve the model to foresee patients' clinical course or treatment response.

Conclusion

Overall, our study indicates that blood transcriptomics is a promising tool to improve the diagnosis of pediatric rheumatic diseases. The ease of sample collection as well as the continuous enhancement and affordability of sequencing techniques can overcome the challenges of patient heterogeneity and allow for further fruitful research.

Abbreviations

AID	Auto-Inflammatory Diseases
AUC	Area Under the Curve
CRMO	Chronic Recurrent Multifocal Osteomyelitis
DEG	Differentially Expressed Gene
GO	Gene Ontology
HLA-B51	Diseases related to the Human Leukocyte Antigen B51 serotype
IFN	Interferonopathy
IL	Interleukin
ISG	Interferon-Stimulated Gene
JIA	Juvenile Idiopathic Arthritis

RE	Relative Expression
RNA	Ribonucleic Acid
ROC	Receiver-Operator-Characteristic curve
SLE	Systemic Lupus Erythematosus
t-SNE	t-distributed Stochastic Neighbor Embedding method

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12969-022-00747-x>.

Supplementary Material 1
 Supplementary Material 2
 Supplementary Material 3
 Supplementary Material 4
 Supplementary Material 5
 Supplementary Material 6

Acknowledgements

We appreciate the participation of all patients and their families in this study. We are grateful to all unmentioned clinicians, nurses, and lab colleagues. We thank Yanick Crow for helping us with certain analyses and Yves Vandeputte who organised the charity support from "Cycling for the smile of a child".

Authors' contributions

MH analyzed the data, prepared the first draft and final manuscript. LVO, JD, JL, OA, V. Siozopoulou, EDS, JG, KG, MDM, NH, BL, RW, EG, LG, MZ, WD, AL, DT, DW, KV, GVB, MR, HV, TB, PVD, ACJ, MM, V. Sabato, GVC, AA, JVDWTB, JD, and RJ contributed to obtaining IRB approval and recruiting participants. PM pre-processed and analyzed the data. BO conceptualised the study, obtained IRB approval and funding. BO, PVD, and CL supervised the study.

Funding

We received financial support from the Research Foundation Flanders (1861219N to BO), Al Thrasher Early Career Grant, academic investigator-initiated grants from AbbVie and charity support from "Cycling for the smile of a child". The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Data Availability

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

This non-commercial academic study was approved by the IRBs from the Antwerp University Hospital/University of Antwerp (IRB numbers 15/43/448, 18/16/215 and 19/30/358) and all local committees from each participating hospital. This study complies with the Declaration of Helsinki. Written informed consent has been obtained from the subjects (or their legally authorized representative).

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Center for Health Economics Research and Modelling Infectious Diseases (CHERMID), Vaccine and Infectious Disease Institute, University of Antwerp, Wilrijk, Belgium

²Antwerp Unit for Data Analysis and Computation in Immunology and Sequencing (AUDACIS), University of Antwerp, Antwerp, Belgium

³Antwerp Center for Translational Immunology and Virology (ACTIV), Vaccine and Infectious Disease Institute, University of Antwerp, Wilrijk, Belgium

⁴Center of Medical Genetics, University of Antwerp, Antwerp University Hospital, Edegem, Belgium

⁵Ophthalmology Department, Antwerp University Hospital, Edegem, Belgium

⁶Dermatology Department, Antwerp University Hospital, Edegem, Belgium

⁷Department of Translational Research in Immunology and Inflammation, University of Antwerp, Wilrijk, Belgium

⁸Pathology Department, Antwerp University Hospital, Edegem, Belgium

⁹Radiology Department, Antwerp University Hospital, Edegem, Belgium

¹⁰Department of Molecular – Morphology – Microscopy, University of Antwerp, Wilrijk, Belgium

¹¹Clinical Biology Department, Antwerp University Hospital, Edegem, Belgium

¹²Radiology Department, Brussels University Hospital, Jette, Belgium

¹³Radiology Department, Ghent University Hospital, Ghent, Belgium

¹⁴Department of Immunology, Allergology, and Rheumatology, Antwerp University Hospital, Edegem, Belgium

¹⁵Rheumatology Department, Ghent University Hospital, Ghent, Belgium

¹⁶Rheumatology Department, Antwerp Hospital Network, Antwerp, Belgium

¹⁷Pediatric Neurology Unit, Antwerp University Hospital, Edegem, Belgium

¹⁸Pediatric Cardiology Department, Antwerp University Hospital, Edegem, Belgium

¹⁹Pediatric Cardiology Department, Antwerp Hospital Network, Antwerp, Belgium

²⁰Department of Pediatrics, Turnhout General Hospital, Turnhout, Belgium

²¹Department of Pediatrics, Catholic University of Louvain, Louvain-la-Neuve, Belgium

²²Department of Pediatrics, Namur University Hospital Center, Site Dinant, Dinant, Belgium

²³Department of Pediatrics, Namur University Hospital Center, Site Sainte-Elisabeth, Namur, Belgium

²⁴Department of Pediatrics, Nikolaas General Hospital, Sint-Niklaas, Belgium

²⁵Department of Emergency Medicine/Pediatric Care, Brussels University Hospital, Jette, Belgium

²⁶Department of Pediatrics, Jessa Hospital, Hasselt, Belgium

²⁷Department of Pediatric Neurology, Ghent University Hospital, Ghent, Belgium

²⁸Department of Pediatric Rheumatology, Antwerp University Hospital, Edegem, Belgium

²⁹Center for the Evaluation of Vaccine, Vaccine and Infectious Disease Institute, University of Antwerp, Wilrijk, Belgium

³⁰Antwerp Center for Pediatric Rheumatology and Autoinflammatory Diseases, Antwerp, Belgium

³¹Pediatric Immunology Department, Brussels University Hospital, Jette, Belgium

³²Department of Pediatric Rheumatology, Ghent University Hospital, Ghent, Belgium

³³ADREM Data Lab, Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium

³⁴Biomedical Informatics Research Network Antwerp, University of Antwerp, Antwerp, Belgium

³⁵Department of Pediatric Rheumatology, Brussels University Hospital, Jette, Belgium

Received: 19 July 2022 / Accepted: 24 September 2022

Published online: 17 October 2022

References

- Banchereau R, Cepika AM, Banchereau J, Pascual V. Understanding human autoimmunity and autoinflammation through transcriptomics. *Annu Rev Immunol*. 2017 Apr 26;35:337–70.
- Barturen G, Babaei S, Català-Moll F, Martínez-Bueno M, Makowska Z, Martorell-Marugán J, et al. Integrative Analysis Reveals a Molecular Stratification of Systemic Autoimmune Diseases. *Arthritis Rheumatol*. 2021 Jun 1;73(6):1073–85.
- Bennett L, Palucka AK, Arce E, Cantrell V, Borvak J, Banchereau J, et al. Interferon and Granulopoiesis Signatures in Systemic Lupus Erythematosus Blood. *J Exp Med*. 2003 Mar;17(6):711–23. 197.
- Walsh RJ, Kong SW, Yao Y, Jallal B, Kiener PA, Pinkus JL, et al. Type I interferon-inducible gene expression in blood is present and reflects disease activity in dermatomyositis and polymyositis. *Arthritis Rheum*. 2007 Nov 1;56(11):3784–92.
- Orange DE, Agius P, DiCarlo EF, Robine N, Geiger H, Szymonifka J, et al. Identification of Three Rheumatoid Arthritis Disease Subtypes by Machine Learning Integration of Synovial Histologic Features and RNA Sequencing Data. *Arthritis Rheumatol*. 2018 May 1;70(5):690–701.
- Figgett WA, Monaghan K, Ng M, Alhamdoosh M, Maraskovsky E, Wilson NJ, et al. Machine learning applied to whole-blood RNA-sequencing data uncovers distinct subsets of patients with systemic lupus erythematosus. *Clin Transl Immunol*. 2019 Jan 1;8(12):e01093.
- Bartholomeus E, De Neuter N, Lemay A, Pattyn L, Tuerlinckx D, Weynants D, et al. Diagnosing enterovirus meningitis via blood transcriptomics: An alternative for lumbar puncture? *J Transl Med*. 2019;17(1):1–9.
- Kursa MB, Rudnicki WR. Feature selection with the boruta package. *J Stat Softw*. 2010 Sep 16;36(11):1–13.
- Breiman L. Random forests. *Mach Learn*. 2001 Oct;45(1):5–32.
- Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011 Mar 17;12(1):1–8.
- Kim H, de Jesus AA, Brooks SR, Liu Y, Huang Y, VanTries R, et al. Development of a Validated Interferon Score Using NanoString Technology. *J Interf Cytokine Res*. 2018 Apr 1;38(4):171–85.
- Bank I. The Role of Gamma Delta T Cells in Autoimmune Rheumatic Diseases. *Cells*. 2020 Feb 18;9(2):462.
- Stanko K, Iwert C, Appelt C, Vogt K, Schumann J, Strunk FJ, et al. CD96 expression determines the inflammatory potential of IL-9-producing Th9 cells. *Proc Natl Acad Sci U S A*. 2018 Mar 27;115(13):E2940–9.
- Li Y, Zhao L, dan, Tong L, sha, Qian S ning, Ren Y, Zhang L, et al. Aberrant CD200/CD200R1 expression and function in systemic lupus erythematosus contributes to abnormal T-cell responsiveness and dendritic cell activity. *Arthritis Res Ther*. 2012 May 23;14(3):1–17.
- Nikopensius T, Niibo P, Haller T, Jagomägi T, Voog-Oras Ü, Tõnisson N, et al. Association analysis of juvenile idiopathic arthritis genetic susceptibility factors in Estonian patients. *Clin Rheumatol*. 2021 Jun 8;3:1–9.
- Schulert GS, Pickering AV, Do T, Dhakal S, Fall N, Schnell D, et al. Monocyte and bone marrow macrophage transcriptional phenotypes in systemic juvenile idiopathic arthritis reveal TRIM8 as a mediator of IFN-γ hyper-responsiveness and risk for macrophage activation syndrome. *Ann Rheum Dis*. 2021 May 1;80(5):617 LP – 625.
- Cusick MF, Libbey JE, Fujinami RS. Molecular mimicry as a mechanism of autoimmune disease. *Clin Rev Allergy Immunol*. 2012 Nov 19;42(1):102–11.
- Rojas M, Restrepo-Jiménez P, Monsalve DM, Pacheco Y, Acosta-Ampudia Y, Ramirez-Santana C, et al. Molecular mimicry and autoimmunity. Vol. 95, *Journal of Autoimmunity*. Academic Press; 2018. p. 100–23.
- Kovarik P, Castiglia V, Ivin M, Ebner F. Type I Interferons in Bacterial Infections: A Balancing Act. Vol. 7, *Frontiers in Immunology*. 2016. p. 652.
- Hofmann SR, Kapplusch F, Girschick HJ, Morbach H, Pablik J, Ferguson PJ, et al. Chronic Recurrent Multifocal Osteomyelitis (CRMO): Presentation, Pathogenesis, and Treatment. *Curr Osteoporos Rep*. 2017 Dec 1;15(6):542–54.
- Horton DB, Scott FI, Haynes K, Putt ME, Rose CD, Lewis JD, et al. Antibiotic exposure and juvenile idiopathic arthritis: A case-control study. *Pediatrics*. 2015 Aug 1;136(2):e333–43.
- Tesser A, De Carvalho LM, Sandrin-Garcia P, Pin A, Pastore S, Taddio A, et al. Higher interferon score and normal complement levels may identify a distinct clinical subset in children with systemic lupus erythematosus. *Arthritis Res Ther*. 2020 Apr;25(1):1–12. 22.
- Kim H, de Jesus AA, Brooks SR, Liu Y, Huang Y, VanTries R, et al. Development of a Validated Interferon Score Using NanoString Technology. *J Interf Cytokine Res*. 2018 Apr;38(4):171–85.
- Guan Y, Zhang H, Quang D, Wang Z, Parker SCJ, Pappas DA, et al. Machine Learning to Predict Anti-Tumor Necrosis Factor Drug Responses of Rheumatoid Arthritis Patients by Integrating Clinical and Genetic Markers. *Arthritis Rheumatol*. 2019 Dec 1;71(12):1987–96.
- June RR, Aggarwal R. The use and abuse of diagnostic/classification criteria. *Best Pract Res Clin Rheumatol*. 2014 Dec 1;28(6):921–34.

26. Griffin TA, Barnes MG, Ilowite NT, Olson JC, Sherry DD, Gottlieb BS, et al. Gene expression signatures in polyarticular juvenile idiopathic arthritis demonstrate disease heterogeneity and offer a molecular classification of disease subsets. *Arthritis Rheum.* 2009 Jul 1;60(7):2113–23.
27. Ma Y, Chen J, Wang T, Zhang L, Xu X, Qiu Y, et al. Accurate Machine Learning Model to Diagnose Chronic Autoimmune Diseases Utilizing Information From B Cells and Monocytes. *Front Immunol.* 2022;13.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.