Research Article

Identifying Infliximab- (IFX-) Responsive Blood Signatures for the Treatment of Rheumatoid Arthritis

ShiJian Ding 1, ZhanDong Li 2, Tao Zeng 3, Yu-Hang Zhang 4, Tao Huang 3,5 and Yu-Dong Cai 1

1 School of Life Sciences, Shanghai University, Shanghai, China
2 College of Food Engineering, Jilin Engineering Normal University, Changchun, China
3 Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China
4 Channing Division of Network Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA, USA
5 CAS Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China

Correspondence should be addressed to Tao Huang; tohuangtao@126.com and Yu-Dong Cai; cai_yud@126.com

Received 26 February 2021; Revised 18 March 2021; Accepted 24 March 2021; Published 7 April 2021

Academic Editor: Quan Zou

Copyright © 2021 ShiJian Ding et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Rheumatoid arthritis (RA) is a severe chronic pathogenic inflammatory abnormality that damages small joints [1, 2]. Some patients with severe and progressive RA may also suffer from pathogenic lesions in various body systems, including the skin, eyes, lungs, and blood vessels [1]. According to the statistics provided by the American College of Rheumatology in 2008, more than 1.3 million people in the USA suffer from pathogenic rheumatoid arthritis with typical symptoms [3, 4]; thus, RA is the leading cause of arthritis among multiple pathogeneses.

Comprehensive diagnosis and treatment procedures for RA have been established because of its severe symptoms and relatively high morbidity [1, 5–8]. The diagnosis of RA can be divided into two major procedures: symptom-dependent diagnosis [9] and clinical laboratory examinations [10]. A group of typical symptoms can be used to preliminary screen for RA. The typical symptoms of RA include the swelling of small joints (which may further extend to larger joints, such as the hips and shoulders), fatigue, fever, and anemia, which are quite easy to identify and recognize [11, 12]. However, the types and severity of these symptoms vary among patients. Thus, RA is difficult to diagnose based only on clinical symptoms. Apart from clinical symptoms, blood tests and imaging tests have also been applied for the accurate diagnosis of RA. Erythrocyte sedimentation rate [13] and C-reactive protein [13] are nonspecific biomarkers for the diagnosis and progression monitoring of RA that reflect the degree of inflammatory responses in the whole body. The
gold standard for RA diagnosis is clinical liquid biopsy, and rheumatoid factor [14] and anticyclic citrullinated peptide [15] antibodies are the specific biomarkers for the blood test screening of RA. Pathogenic lesions in the joints can be regarded as a diagnosis measurement for RA and can be easily identified by X-ray [16], ultrasound [17], and magnetic resonance imaging [18].

Medication and surgery are the two major therapeutic approaches for RA [19, 20]. Different from surgery-based approaches, which focus on the direct relief of joint damage, most medications focus on the prevention of abnormal inflammatory immune responses and therefore indirectly relieve systemic symptoms. Four subgroups of medications are clinically applied, namely, nonsteroidal anti-inflammatory drugs, steroids, disease-modifying antirheumatic drugs, and biological agents [21]. Among these drugs, infliximab (IFX) is a novel biological agent applied for the treatment of RA. As a chimeric monoclonal antibody, IFX has been approved by the Food and Drug Administration (FDA) for the treatment of multiple immune-associated diseases, including RA, early in 2007 [22]. IFX targets one of the pathogenic proinflammatory cytokines, tumor necrosis factor-alpha (TNF-α), and thus has been applied in classical TNF antagonist therapy against multiple chronic autoimmune inflammatory diseases, including RA [23]. The clinical application of IFX has been studied for nearly 30 years; IFX greatly improves physical functions and is beneficial for the achievement of clinical remission even under discontinuous medication [24, 25]. However, not all patients react to IFX, and its anti-inflammatory effects vary among patients with RA. IFX-sensitive and IFX-resistant patients are difficult to distinguish based on traditional clinical examination; thus, how to predict the therapeutic effects of IFX on patients with RA is one of the urgent translational medicine problems in the clinical treatment of RA.

With the development of liquid biopsy and high-throughput sequencing technologies, a recent study [26] revealed that patients with different drug susceptibility against IFX have different pretherapeutic blood expression patterns; thus, liquid biopsy and the transcriptomic profiling of patients’ blood may help predict the therapeutic effects of IFX in RA and therefore assist in clinical medication. However, the application of whole transcriptome sequencing on every patient with RA is not feasible; therefore, the identification of accurate and efficient transcriptomic biomarkers and their respective pathogenic expression pattern would be the priority. In this study, we presented a novel computational method to identify the applicable and substantial blood gene biomarkers of IFX sensitivity by liquid biopsy. This study may assist in the establishment of a test standard for clinical drug sensitivity for RA treatment and contribute to the revelation of unique IFX-associated pharmacological mechanisms.

2. Materials and Methods

2.1. Data. The blood gene expression profiles of 140 patients with RA before IFX treatment were downloaded from the Gene Expression Omnibus under accession number GSE78068 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE78068) [26]. Among the patients, 42 showed remission and 98 showed nonremission. 19595 genes were involved to constitute the blood gene expression profiles. Predicting the response of patients with RA to IFX before therapy using their blood will help in deciding precise treatments, which is the ultimate goal of precision medicine. We would like to build such IFX response prediction models for patients with RA based on their blood gene expression profiles.

2.2. Monte Carlo Feature Selection (MCFS). In this work, the MCFS method, which is a decision tree- (DT-) based feature selection method [27–29], was used to select important gene candidates. MCFS can randomly select a few feature subsets from the original features. Each feature subset consists of a small number of features from the original ones. One bootstrap sample dataset can be induced from this feature subset, and then, multiple DTs can be learned and tested on this bootstrap sample datasets. The process is repeated several times to produce many feature subsets. Each feature subset can help learn the same number of DTs.

The contribution of each feature in these DTs can be evaluated by the relative importance (RI) score, which is calculated as follows:

\[ RI_f = \sum_{\tau=1}^{p!} (wAcc)^u \sum_{n_f(\tau)} IG(n_f(\tau)) \left( \frac{\text{no. in } n_f(\tau)}{\text{no. in } \tau} \right)^v, \]

where \( wAcc \) indicates the weighted accuracy and \( n_f(\tau) \) represents a node of feature \( f \) in DT \( \tau \). The information gain of \( n_f(\tau) \) in the DT is measured by \( IG(n_f(\tau)) \), and \( \text{no. in } n_f(\tau) \) is the number of training samples in node \( n_f(\tau) \). In addition, the weighting factors \( u \) and \( v \) have a value of 1 by default.

After each feature was assigned a RI value, we ranked all features in a list with the decreasing order of their RI values. In addition to the feature list, the MCFS method also outputs some most important features, called informative features, which are some top features in the list. These features are accessed by determining a threshold of RI values via a permutation test on class labels and one-sided Student’s \( t \)-test.

This study used the MCFS program retrieved from http://www.ipipan.eu/staff/m.draminski/mcfs.html. Its default parameters were used.

2.3. Incremental Feature Selection (IFS). IFS is widely used to determine the optimal number of features for constructing a classification model with an integrated supervised classifier [30]. On the basis of a ranked feature list obtained from MCFS, a series of feature subsets are produced with a step interval of 10. For example, the first feature subset includes the top 10 features and the second feature subset includes the top 20 features. For each feature subset, a classifier (e.g., support vector machine (SVM)) is trained on a training data induced from this feature subset. An optimal feature subset is selected when it has the highest performance among the candidate feature subsets, where performance is evaluated by the Matthews correlation coefficients (MCC) [31] under 10-fold cross-validation [32]. The classifier with such optimal feature subset can be built and was called the optimum classifier in this study.
2.4. Support Vector Machine (SVM). SVM is a classification algorithm suitable for linear and nonlinear data [33–39]. For a dataset with two classes, SVM tries to find out an optimum hyperplane, which can divide samples in two classes with a maximum margin. However, in many cases, such hyperplane is not easy or impossible to be discovered. SVM employs a kernel trick to convert the original sample in a low-dimensional space to a new sample in some high-dimensional space, in which the optimum hyperplane can be easily constructed. For a new sample, it is also mapped into the high-dimensional space and its class is determined according to the side of the hyperplane it lies. In this study, we used the tool “SMO” in Weka [40, 41] to quickly implement SVM. Such SVM is optimized by the sequential minimum optimization algorithm [42]. For convenience, default parameters were adopted, where the kernel was a polynomial function.

2.5. Random Forest (RF). RF [43] is a meta-classifier that is widely applied in biological and biomedical researches [44–52]. RF includes many DTs as members. To construct a DT, a dataset, in which samples are randomly selected, with replacement, from the original dataset, is constructed. Such dataset has the same size of the original dataset. The DT is grown at each node by determining an optimal splitting way on some features, which were randomly selected from all features. Given a new sample, each DT gives its prediction. RF integrates these predictions with majority voting. Similar to SVM, we also employed a tool “RandomForest” in Weka [40, 41], which implements RF. Likewise, the default parameters were used, where the number of DTs was set to ten.

2.6. Rule Learning. In addition to “black-box” classification algorithms, the interpretable rules for a classification model can also be extracted to explain the feature differences between groups of patients with particular response to drug treatment. To accelerate this procedure, we directly picked up the informative features extracted by the MCFS method. These features were further filtered by the Johnson Reducer algorithm [53]. The remaining features were fed into the repeated incremental pruning to produce error reduction (RIPPER) algorithm [54] to extract classification rules. Obtained rules were represented by IF-ELSE rules. The above rule learning procedures were also integrated in the MCFS program, which was directly adopted in this study.

2.7. Measurement. The MCC [31] within 10-fold cross-validation was applied in this work to evaluate the classification performance of different classification models. The MCC for the binary problem is calculated as follows:

\[
\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TN + FN) \times (TN + FP) \times (TP + FN) \times (TP + FP)}},
\]

where TP, TN, FP, and FN represent the number of true-positive, true-negative, false-positive, and false-negative samples, respectively. As a measurement for the classification model, the MCC value ranges from -1 to +1, where a value of +1 indicates that the classification model has the best performance. To date, it has wide application in bioinformatics for evaluating the performance of different classification models [55–61].

Besides, we also employed other five measurements to give a full evaluation on different classification models. They were sensitivity (SN), specificity (SP), accuracy (ACC), precision, and F1-measure, which can be calculated by the following equations:

\[
\begin{align*}
\text{SN} &= \frac{TP}{TP + FN}, \\
\text{SP} &= \frac{TN}{TN + FP}, \\
\text{ACC} &= \frac{TP + TN}{TP + FN + TN + FP}, \\
\text{Precision} &= \frac{TP}{TP + FP}, \\
F1\text{-measure} &= \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}.
\end{align*}
\]

3. Results

In this study, we gave a computational investigation on the blood gene expression profiles of patients with RA before IFX treatment. The entire procedures are illustrated in Figure 1. This section gave the results of all procedures.

3.1. Results of MCFS. The blood gene expression profiles were first analyzed by the MCFS method. As a result, each feature was assigned a RI value, which indicated its importance. The RI values of all features are listed in Table S1. Accordingly, all features were ranked in a list by the decreasing order of features’ RI values. Such list is also provided in Table S1.

3.2. Results of IFS. Based on the feature list obtained in the above section, the IFS method is followed. It first constructed several feature subsets. Then, on each feature subset, a classifier was built using SVM or RF as the classification algorithm. Each classifier was evaluated by 10-fold cross-validation. Six measurements (see equations (2) and (3)) were obtained for each classifier, which are listed in Tables S2 and S3. For an easy observation, a curve was plotted for each classification algorithm, as shown in Figure 2, in which MCC was set as the y-axis and the number of features was set as the x-axis. It can be observed that for SVM, the highest MCC was 0.760 when top 1260 features were used. Thus, these 1260 features constituted the optimum feature subset of SVM and the SVM classifier with these features was the optimum SVM classifier. Other measurements of such classifier are listed in Table 1. As for RF, the highest MCC was 0.611 when only top ten features were adopted. An optimum RF classifier was built based on these top ten features. The MCC of the optimum RF classifier was much lower than that of the optimum SVM classifier. Other five measurements of such RF classifier are provided in Table 1. Evidently, each measurement was inferior to that of the optimum SVM classifier.
Figure 1: Entire procedures to investigate the blood gene expression profiles of rheumatoid arthritis patients. Profiles are retrieved from Gene Expression Omnibus, which are analyzed by the Monte Carlo feature selection method. A feature list is obtained, which is fed into the incremental feature selection method to construct efficient classifiers and extract essential genes. On the other hand, informative features, which are some top features in the list, are used to construct classification rules via Johnson Reducer and RIPPER algorithms.

Figure 2: IFS curves with different classification algorithms on different numbers of features (genes). The support vector machine yields the highest MCC of 0.760 when top 1260 features are used, whereas the random forest generates the highest MCC of 0.611 when top 10 features are adopted.

Table 1: Performance of some key support vector machine (SVM) and random forest (RF) classifiers.

<table>
<thead>
<tr>
<th>Classification algorithm</th>
<th>Number of features</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>Precision</th>
<th>F1-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>1260</td>
<td>0.690</td>
<td>0.990</td>
<td>0.900</td>
<td>0.967</td>
<td>0.806</td>
</tr>
<tr>
<td>SVM</td>
<td>60</td>
<td>0.619</td>
<td>0.980</td>
<td>0.871</td>
<td>0.929</td>
<td>0.743</td>
</tr>
<tr>
<td>RF</td>
<td>10</td>
<td>0.643</td>
<td>0.929</td>
<td>0.843</td>
<td>0.794</td>
<td>0.711</td>
</tr>
</tbody>
</table>
classifier. Therefore, it can be concluded that the optimum SVM classifier is better than the optimum RF classifier.

As mentioned above, the optimum SVM classifier needed much more features than the optimum RF classifier. In fact, the SVM classifier can yield good performance when much less features were used. As shown in Figure 3, when top 60 features were used, the SVM classifier can generate the MCC of 0.686, which was still higher than that of the optimum RF classifier. The detailed performance of such SVM classifier is listed in Table 1. It can be seen that all measurements, except SN, of this SVM classifier were higher than those of the optimum RF classifier. Thus, we picked up such SVM classifier as the proposed classifier because it can provide good performance and was much more efficient than the optimum SVM classifier.

3.3. Rule Learning Results. The optimum SVM and RF classifiers gave good performance. However, they were black-box algorithms. Few medical insights can be captured from these classifiers. In view of this, we further employed a rule learning procedure. The informative features yielded by MCFS were processed by the Johnson Reducer and RIPPER algorithms one by one. As a result, three rules were constructed, as shown in Table 2, where two rules were for prediction of IFX-sensitive patients and the last one was for identification of IFX-resistant patients. We counted two measurements: support and accuracy for each rule, as listed in Table 2. Each rule covered some samples, and all accuracies were quite high, implying the utility of the rules.

Furthermore, to test the effectiveness of the above rule learning procedures, we did the 10-fold cross-validation three times. Six measurements calculated by equations (2) and (3) were counted. The MCC was 0.439, and other five measurements were 0.421 (SN), 0.939 (SP), 0.783 (ACC), 0.746 (precision), and 0.538 (F1-measure), respectively. Clearly, the rule classifier was inferior to the optimum SVM and RF classifiers. However, it can explain the detailed gene expression pattern for distinguishing patients with RA who respond to IFX treatment or not.

4. Discussion

IFX is one of the major clinically applied drugs for RA. However, the sensitivity and effectiveness of this drug vary among patients. Recent publications confirmed that the sensitivity of this drug against RA can be predicted by obtaining the expression profiling pattern of patients’ pretherapeutic blood. However, the core signatures/biomarkers for the prediction and understanding of IFX sensitivity are difficult to identify. We identified gene signatures for drug therapeutic effect evaluation and established a series of quantitative rules that explain the detailed accurate recognition of patients with different IFX sensitivity using a novel computational approach on the expression profiling of pretherapeutic blood. All the identified signatures have been confirmed by recent publications, and the detailed analysis of the representative genes and rules is discussed below.

4.1. Gene Signatures Associated with IFX Response. In this study, with some computational methods, several genes associated with IFX response were identified. Here, we selected some of them for detailed analysis, which are listed in Table 3.

DISC1, which encodes a scaffold protein, participates in the synthesis of hemoglobin in peripheral blood [62, 63]. Although no direct evidence confirmed that the blood expression of DISC1 may directly contribute to the pathogenesis of RA, recent publications validated that DISC1 may
participate in TNF-α-associated biological processes [64, 65]. A further study on the biological processes of TNF-α receptors reported that our predicted gene, DISC1, may be functionally related to MIPT3, a microtubule-interacting protein associated with TNF receptors; thus, DISC1 has potential regulatory effects on TNF-associated biological processes [66]. The therapeutic effects of IFX rely on the pharmacological regulation on TNF-α-associated biological processes [67, 68]; therefore, as a regulator for the TNF receptor, the different expression patterns of DISC1 can indicate different TNF-related biological processes and affect the pharmacological effects of IFX.

SAMD11, as a transcription coactivator, contributes to the development of a photoreceptor [69]. In 2009, SAMD11 was identified as a susceptibility gene for RA by high-throughput genotyping techniques [70]. Multiple publications have also confirmed its specific role in inflammatory regulation [71, 72] and even validated its functional relationship with IL17 [73, 74]. The pharmacological effects of IFX are mediated by TNF-α inhibition; thus, IFX is functionally related to the regulation of immune responses via the interactions between IL17 and TNF-α [75, 76]. As a specific regulator of IL17, our candidate gene, SAMD11, may also participate in the regulation of IFX-mediated pharmacological responses in patients with RA. Therefore, the expression level of SAMD11 in the blood may indirectly reflect the reactivity of IFX in patients with RA.

EID2B, as an RA-associated gene, participates in MyoD-dependent transcription, glucocorticoid receptor-dependent transcription, and muscle differentiation as a functional repressor [77, 78]. MyoD-associated biological functions induce muscle lesions by interacting with abnormal regulation on TNF-α pathways, targeted by IFX during the initiation and progression of RA [79]. Further studies on the pharmacological effects of IFX confirmed that MyoD may directly participate in IFX-associated therapeutic metabolism in Crohn’s disease, which is another autoimmune disease [80]. Our predicted gene EID2B may also participate in the regulation of the therapeutic effects of IFX by interacting with the core regulator, MyoD. Therefore, the expression level of EID2B in the peripheral blood may reflect EID2B expression in mesenchymal cells and immune cells in the blood and may also be a novel parameter for the prediction of the prognosis of IFX-dependent therapeutics.

NTS, which encodes a common precursor for peptides neuromedin N and neurotensin, participates in the regulation of fat metabolism in muscles [81, 82]. A recent research confirmed its specific pathogenic role in the pathophysiology of RA [83]. As for its distinctive role for the sensitivity of IFX, NTS participates in TNF-α-associated biological processes [84, 85]. Considering that IFX acts as the inhibitor of TNF-α-associated biological processes [23], the therapeutic effects of IFX are functionally associated with the abnormal activation status of TNF-α-associated biological processes. Therefore, as an effective participator of TNF-α-associated biological processes, the expression pattern of NTS may reflect the activation status of TNF-α-associated biological processes and therefore indicates the therapeutic effects of IFX.

STAT2, as an effective member of the STAT protein family, participates in the transcriptional regulation mediated by type I interferons (IFNs) [86] and the Jak kinase signaling cascade [87, 88]. The treatment effects of IFX are functionally connected to the IFN signaling cascade [89–91]; hence, IFN-associated biological processes may be crucial for the pharmacological effects of IFX. Therefore, as a transcriptional regulator at the downstream of IFN-associated biological processes, the expression level of our predicted gene STAT2 may also be an alternative following the expression alteration of genes in the type I IFN-associated pathways. This finding indicates the potential identification ability of STAT2 on the therapeutic effects of IFX.

The HELZ gene can encode a member of the RNA helicase superfamily I class. HELZ participates in RNA hydrolysis in multiple tissues [92]. In fact, TNF-α and its related immune regulatory functions are regulated by multiple RNA helicases, including helicase with zinc finger (HELZ) [93–95]. Therefore, the expression pattern of HELZ may also affect the therapeutic effects of IFX by interfering with TNF-α-associated biological processes. Similarly, SUMO2 also affects IFX sensitivity by interfering with TNF-α-associated biological processes [96, 97].

Apart from unannotated RNA transcripts with no validated protein products, all the predicted genes have been confirmed to be functionally related to TNF-α-associated
biological processes in either physical or pathological conditions. Therefore, these genes may further affect the therapeutic effects and sensitivity of IFX in patients with RA. This finding validates the efficacy and accuracy of our prediction method and analysis.

4.2. Signature Rules Associated with IFX Response. Apart from the qualitative analysis of each top-ranked gene signatures in our prediction list, we also set up a series of quantitative recognition rules for the detailed and accurate recognition of IFX-sensitive and IFX-resistant patients. The first rule involves only the DISC1 gene. A low DISC1 expression (<5 FPKM) indicates that the patient may be sensitive to IFX. Based on the analysis, a low DISC1 expression may indicate abnormal TNF-α-associated immune activation status in RA [66]. Therefore, patients with a low DISC1 expression pattern may have an activated TNF-associated signaling pathway and are definitely sensitive to therapeutic effects. The expression level of DISC1 is quite high in normal conditions (>10 FPKM). Therefore, our threshold may indicate low DISC1 expression level and corresponds to our analysis above.

The second rule involves the SAMD11 gene. Different from DISC1, a high SAMD11 expression level may indicate an activated inflammatory status in the whole body and a high TNF-α expression level [75, 76]. Therefore, the therapeutic effects of IFX may also be more effective in conditions with more potential pharmacological targets. This finding corresponds to our prediction rules. According to recent publications, the expression level of SAMD11 in normal whole blood is <1 FPKM. Therefore, under the predicted conditions, the expression level of SAMD11 may be upregulated and result in the stronger activation of the TNF-α signaling pathway. Patients with specific SAMD11 expression levels that follow our rules would be sensitive to IFX. By contrast, patients with expression profiling that does not follow these quantitative rules may be resistant to IFX-mediated RA therapeutics.

5. Conclusions

The identified blood gene signatures participate in IFX-sensitive pharmacological processes in patients with RA. Thus, these genes may be potential biomarkers for the distinction of IFX-sensitive and IFX-resistant patients at the transcriptomic level. Several quantitative signature rules for the distinction of patients have also been verified by other recent publications. Therefore, our newly presented method provides comprehensive qualitative and quantitative prediction standards for prognosis guidance on the clinical application of IFX on patients with RA.

Data Availability

The data used to support the findings of this study have been deposited in the Gene Expression Omnibus repository (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE78068).

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Authors’ Contributions

Shijian Ding and ZhanDong Li contributed equally to this work.

Acknowledgments

This research was funded by the Strategic Priority Research Program of Chinese Academy of Sciences (XDB38050200), the National Key R&L Program of China (2017YFC1201200, 2018YFC0910403), the Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), the National Natural Science Foundation of China (31701151), the Shanghai Sailing Program (16YF1413800), the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (2016245), and the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences (202002).

Supplementary Materials

Table S1: list of genes ranked based on the RI score from MCFS. Table S2: performance of IFS with SVM. Table S3: performance of IFS with RF. (Supplementary materials)

References


