SUPPLEMENTARY FILE

Data analysis and multivariate mining methods

1. Basic statistics

In the first step, basic statistical methods (univariate and multivariate statistics) and visual network analysis were applied in parallel in order to understand the importance of expression of individual genes and their profiles in particular patient's subgroups. Statistical analysis was performed using the R statistical software package, a free software environment for statistical computing and graphics (http://www.r-project. org/) and using GraphPad Prism (San Diego, CA). Applied statistical methods were Mann-Whitney U-test, Benjamini-Hochberg correction, and Kruskal-Wallis test. Spearman correlation between gene expression and continuous DAS28 values were performed using Genex (MultiD Analyses AB, Sweden). Since our data did not meet the assumption of normality as assessed by the Shapiro-Wilk test, the non-parametric Mann-Whitney U-test was used for the comparison of data distribution between two groups. A p-value <0.05 was considered significant.

2. Patient similarity network (PSN) based on LRNet algorithm

We utilised the LRNet method of network construction based on the nearest neighbour analysis (1) for i) the visualisation of the individual gene expression profiles of patients and ii) investigation of the best combination of the two or more markers (= gene expression) able to discriminate between particular patient's subgroups (active and inactive RA).

The implementation of the LRNet algorithm has been prepared for paper (1) in which this method is described in detail. This implementation was also utilised to analyse data from this presented research. Briefly, the construction of a network starts from the calculation of the pairwise similarity between any two pairs of patient/gene expressions of a given data set and then, calculated similarities are analysed with regards to local connectivity of individual vertices (genes). In the resulting network, vertices represent different patients/genes, and the size of each vertex corresponds to its local importance (representativeness) based on analysis of its neighbourhood (*i.e.* other neighbouring patients/genes). Simply said, vertices which are nearest neighbours for most of their neighbours have higher representativeness (2).

Firstly, the LRNet method was applied to construct patient similarity network (3) from data. The internal structure of the resulting network represents the similarity among the patients, calculated by the Gaussian function (kernel) applied on normalised data. The vertices represent individual patient gene expression profiles and the edges (links) and their strength represent the similarity of connected vertices. Colours distinguish the particular patient's subgroups. The most informative markers for grouping (= clustering) are nominated by the quality parameters of constructed networks (modularity, silhouette), whereas those parameters distributed among all formed subgroups are taking as non-informative. Using network construction algorithms transforming vector data into networks, it is ensured that the edges of the vertices (patients) exist only in cases with a sufficiently strong similarity. The most important feature of networks is their ability to visualise data readily. In this visualisation, we can observe the densely con-

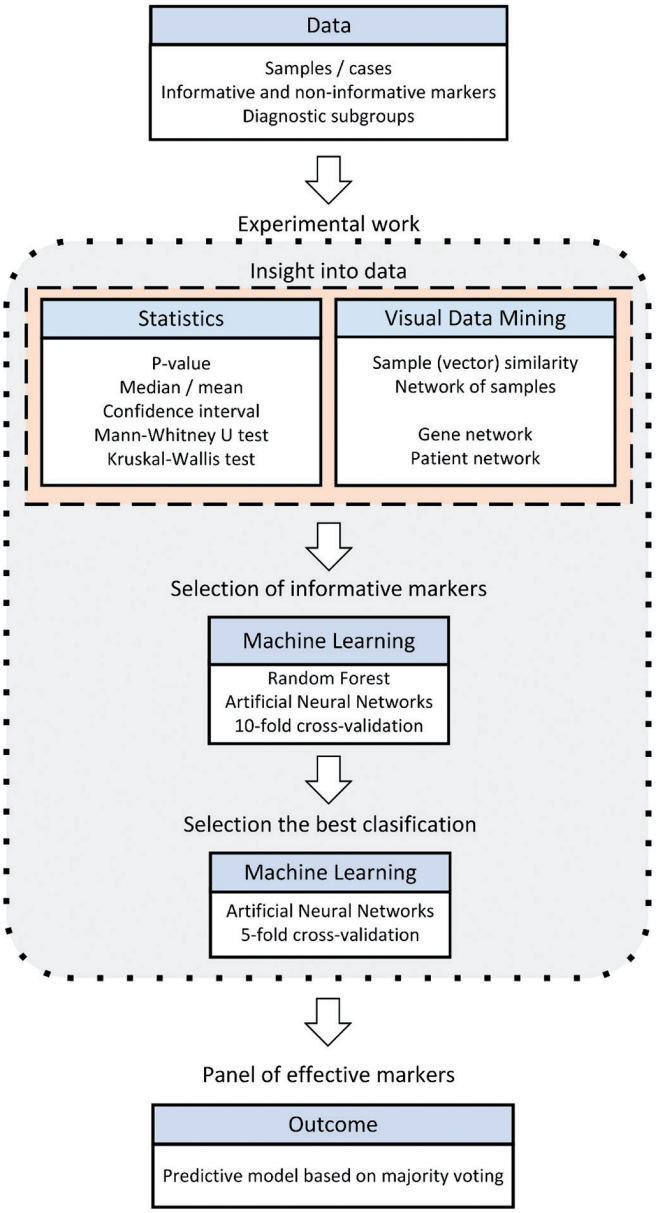

Fig. S1. Algorithm flow chart of statistics and advanced data-mining methods used in this study.

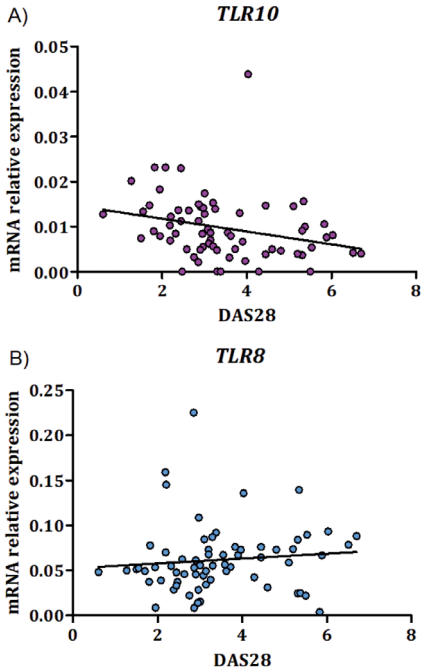

Fig. S2. Correlation analysis between DAS28 and relative mRNA expression of **A**) *TLR10* and **B**) *TLR8*.

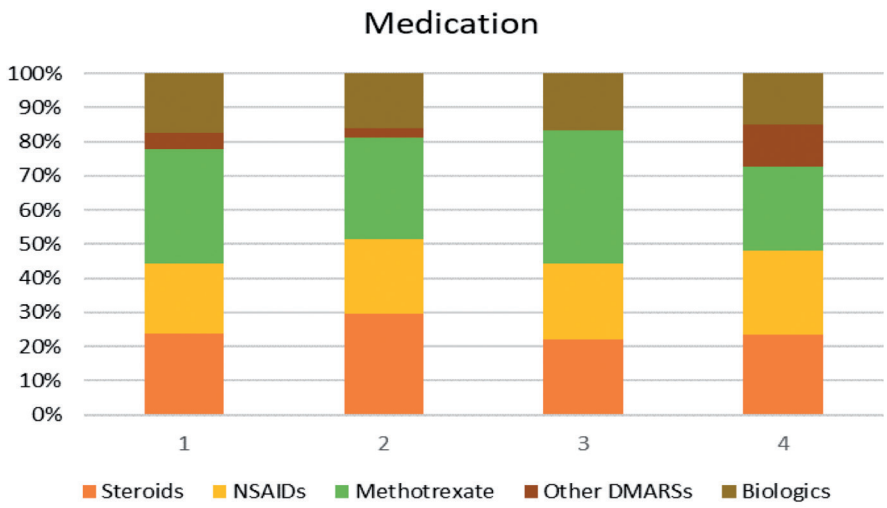

Fig. S3. The proportions of patients treated with different drugs in four subgroups (clusters) as identified from multivariate network analysis based on expression signature of *TLR8* and *IL1RN*.

nected groups of vertices, consistency and degree of connectedness of vertices within the groups, interconnections between groups and their distance, etc. Advantage of networks, in general, is the possibility to investigate network structures for individual combinations of features not only visually but also computationally.

3. Neural network-based feature selection algorithm

In the next step, the machine learning method was applied to design a predictive model (diagnostic classifier) based on the effective panel of the most informative genes. For the pre-selection of informative genes, the Random Forest (RF) machine learning classifier was applied. Briefly described, RF is based on the average decision outcome on a large number of decision trees that were tested efficiently on a subset of gene expression data from randomly selected patients (training sets) and the evaluation of misclassification errors in a fraction of randomly selected patients (test sets). RF features, along with a 10-fold cross-validation (performed 10-times), make RF robust against

Table S1. Investigated genes and primers used for qRT-PCR.

Gene symbol	Gene product	RefSeq accession no.	Forward/reverse primer sequence (5' to 3')	Amplicon length (bp)
PGK1	Phosphoglycerate kinase 1	NM_000291	CTCAACAACATGGAGATTGG/CTTTGGACATTAGGTCTTTGAC	76
TLR1	Toll-like receptor 1	NM_003263	CCCTACAAAAGGAATCTGTATC/TGCTAGTCATTTTGGAACAC	89
TLR2	Toll-like receptor 2	NM_003264	CTTTCAACTGGTAGTTGTGG/GGAATGGAGTTTAAAGATCCTG	176
TLR3	Toll-like receptor 3	NM_003265	AGATTCAAGGTACATCATGC/CAATTTATGACGAAAGGCAC	195
TLR4	Toll-like receptor 4	NM_138557	GATTTATCCAGGTGTGAAATCC/TATTAAGGTAGAGAGGTGGC	75
TLR5	Toll-like receptor 5	NM_003268	ATCTTTCACATGGGTTTGTC/TTCCCCCAGAAGGTTATATG	170
TLR6	Toll-like receptor 6	NM_006068	CTGCCCAAGATTCAGGAGTG/CCATTGCCTTACAACAAAGTTCT	63
TLR7	Toll-like receptor 7	NM_016562	AGATATAGGATCACTCCATGC/CTTCCAAAATGGAATGTAGAGG	120
TLR8	Toll-like receptor 8	NM_138636	TGGAAAACATGTTCCTTCAG/TGCTTTTTCTCATCACAAGG	121
TLR9	Toll-like receptor 9	NM_017442	AAATCCCTCATATCCCTGTC/TTGTAATAACAGTTGCCGTC	116
TLR10	Toll-like receptor 10	NM_030956	AGATTGCTTTTGCCACCAAC/TCTCACATCTCCTTTTGATAGCC	114
IL1B	Interleukin 1 beta	NM_000576	CTAAACAGATGAAGTGCTCC/GGTCATTCTCCTGGAAGG	183
IL1RN	Interleukin 1 receptor antagonist	NM_173842	ATACTTGCAAGGACCAAATG/TGTTAACTGCCTCCAGC	155
IL18	Interleukin 18	NM_001562	CCTTTAAGGAAATGAATCCTCC/CATCTTATTATCATGTCCTGGG	95
IL1R1	Interleukin 1 receptor type 1	NM_000877	TGTTCATTTATGGAAGGGATGA/TTCTGCTTTTCTTTACGTTTTCATT	78
IL1RAP	Interleukin 1 receptor accessory protein	NM_002182	AACTTGAGTTTCCTCATTGC/AGCCTACTACCTTTACAGTC	121
IL18R1	Interleukin 18 receptor 1	NM_003855	GTGAGAAAAGCAGACATGG/AAATGACACACACAGTCAC	112
SIGIRR/IL1R8	Single Ig and TIR domain containing	NM_001135054	ACCCATCTTCATCACCTTC/AAAATCGGAGGAAGGAGTC	133
IL8/CXCL8	C-X-C motif chemokine ligand 8	NM_000584	GGCACAAACTTTCAGAGACAG/ACACAGAGCTGCAGAAATCAG	153

Table S2. Relative mRNA expression levels of genes differentially expressed between A) RA *vs*. healthy controls, B) active *vs*. inactive RA.

A: RA *vs*. healthy controls

Gene	Mean (95 % CI)		FC	p	p_{corr}
	Healthy controls	RA			
SIGIRR	0.196 (0.167-0.225)	0.367 (0.329-0.405)	1.87	3.9×10^{-10}	7.1×10^{-9}
IL18	0.036 (0.031-0.042)	0.060 (0.054-0.067)	1.56	4.1×10^{-8}	3.7×10^{-7}
IL1RN	0.018 (0.013-0.024)	0.039 (0.034-0.044)	2.75	1.4×10^{-7}	8.6×10^{-7}
TLR5	0.029 (0.020-0.037)	0.060 (0.052-0.067)	3.20	4.4×10^{-7}	2.0×10^{-6}
IL18R1	0.006 (0.004-0.007)	0.011 (0.009-0.012)	1.99	3.4×10^{-6}	1.2×10^{-5}
TLR3	0.003 (0.002-0.004)	0.006 (0.005-0.007)	6.59	1.8×10^{-5}	5.4×10^{-5}
IL1RAP	0.008 (0.006-0.010)	0.014 (0.012-0.017)	2.08	4.2×10^{-5}	1.1×10^{-4}
TLR8	0.040 (0.032-0.049)	0.062 (0.053-0.071)	1.59	4.2×10^{-4}	9.5×10^{-4}
IL1B	0.035 (0.002-0.067)	0.062 (0.033-0.091)	1.79	8.2×10^{-4}	1.6×10^{-3}
TLR2	0.049 (0.035-0.062)	0.067 (0.057-0.077)	1.91	1.3×10^{-3}	2.3×10^{-3}
CXCL8	0.108 (0.025-0.191)	0.145 (0.096-0.195)	2.48	2.2×10^{-3}	3.7×10^{-3}
TLR10	0.007 (0.006-0.008)	0.010 (0.008-0.011)	1.41	2.1×10^{-2}	3.2×10^{-2}
TLR4	0.050 (0.043-0.057)	0.041 (0.036-0.046)	0.86	3.2×10^{-2}	4.5×10^{-2}
IL1R1	0.003 (0.002-0.004)	0.004 (0.003-0.005)	1.35	1.7×10^{-1}	2.2×10^{-1}
TLR7	0.015 (0.013-0.017)	0.018 (0.015-0.020)	0.98	2.4×10^{-1}	2.9×10^{-1}
TLR1	0.056 (0.047-0.065)	0.058 (0.051-0.064)	1.04	5.8×10^{-1}	6.2×10^{-1}
TLR6	0.033 (0.021-0.044)	0.026 (0.022-0.030)	1.03	6.0×10^{-1}	6.2×10^{-1}
TLR9	0.009 (0.007-0.011)	0.010 (0.009-0.011)	0.93	7.4×10^{-1}	7.4×10^{-1}

B: Active *vs*. inactive RA

Gene	Mean (95 % CI)		FC	p	p_{corr}
	Inactive RA	Active RA			
TLR10	0.011 (0.009-0.013)	0.008 (0.005-0.011)	0.49	6.5×10^{-3}	1.2×10^{-1}
TLR8	0.057 (0.042-0.072)	0.067 (0.056-0.077)	1.37	1.4×10^{-2}	1.2×10^{-1}
TLR6	0.023 (0.017-0.028)	0.030 (0.024-0.036)	1.57	2.1×10^{-2}	1.3×10^{-1}
TLR2	0.057 (0.046-0.068)	0.078 (0.061-0.095)	1.40	3.3×10^{-2}	1.5×10^{-1}
TLR4	0.039 (0.031-0.048)	0.043 (0.037-0.049)	1.34	4.1×10^{-2}	1.5×10^{-1}
TLR1	0.051 (0.043-0.060)	0.064 (0.055-0.074)	1.19	6.0×10^{-2}	1.6×10^{-1}
SIGIRR	0.405 (0.350-0.461)	0.325 (0.274-0.377)	0.87	6.0×10^{-2}	1.6×10^{-1}
IL1R1	0.003 (0.002-0.005)	0.004 (0.003-0.006)	1.24	1.9×10^{-1}	4.1×10^{-1}
IL18R1	0.011 (0.009-0.013)	0.010 (0.008-0.013)	0.69	2.0×10^{-1}	4.1×10^{-1}
IL18	0.062 (0.054-0.071)	0.058 (0.049-0.068)	0.90	2.4×10^{-1}	4.2×10^{-1}
TLR3	0.007 (0.005-0.008)	0.006 (0.004-0.007)	0.80	3.7×10^{-1}	5.7×10^{-1}
CXCL8	0.167 (0.090-0.245)	0.121 (0.058-0.185)	1.15	3.9×10^{-1}	5.7×10^{-1}
TLR5	0.056 (0.047-0.065)	0.063 (0.051-0.076)	1.05	4.1×10^{-1}	5.7×10^{-1}
IL1RAP	0.012 (0.010-0.015)	0.017 (0.011-0.022)	1.04	4.7×10^{-1}	5.8×10^{-1}
IL1RN	0.037 (0.032-0.041)	0.041 (0.032-0.050)	1.15	4.8×10^{-1}	5.8×10^{-1}
IL1B	0.074 (0.020-0.129)	0.049 (0.032-0.067)	2.40	5.3×10^{-1}	6.0×10^{-1}
TLR9	0.010 (0.008-0.011)	0.010 (0.007-0.012)	1.03	7.2×10^{-1}	7.6×10^{-1}
TLR7	0.018 (0.014-0.022)	0.017 (0.014-0.019)	1.16	1.0	1.0

p_{corr} value corrected for multiple comparisons (Benjamini-Hochberg correction)
FC (Fold change) between group medians of relative mRNA expression levels.

noisy data, irrelevant attributes (markers), unbalanced class distribution and small sample sets (4). Resulting nominated markers were used for artificial neural networks (ANN) - R package neuralnet (5), together with 10-fold cross-validation. Using ANN, we generated all possible combination of up to 10 different markers to detect the best combination of genes classifying active RA patients. We did experiments with different setting and structure of ANN, and finally, we selected the best combination of markers and ANN structure based on RMSE (root-mean-square deviation) and classification error. For the next step, we picked up top 30 combinations of markers with classification error lower than 25% and the smallest (RMSE). Next we again selected the best combination of markers and neural networks with the best classification of the training data by the 5-fold cross validation overall data and for each marker combination and fold which we repeated 500 times for different random initial weights setting. Based on the smallest average RMSE and the average classification error we selected the top two different markers sets. The resulting classifier contained ten neural networks, five for each marker set and markers used as the input to neural networks are TLR1, TLR2, TLR3, TLR7, TLR8, IL8, IL1RN, and IL18R1. Classification of unknown markers is done by the majority voting algorithm. For determining classification error, we used our developed classifier on ten blinded patients; the classification error was 20% (two patients from ten were misclassified).

4. Gene expression similarity network

Next, to assess the relationship between particular genes in active and inactive RA, we performed an analysis of networks constructed by the LRNet network construction method using nearest neighbour- and representativeness analysis (1). In the resulting networks, vertices represent particular genes, and the size of each vertex corresponds to its local importance (representativeness). The links (edges) between genes and their strength represent the similarities between pairs of genes. Again, as the similarity function, the Gaussian kernel was applied to normalised data. Schematic overview of statistics and advanced data-mining methods used in this study is shown in algorithm flow chart (Fig. S1).

References

1. OCHODKOVA E, ZEHNALOVA S, KUDELKA M: Graph Construction Based on Local Representativeness. In: Cao Y, Chen J (Eds) *Computing and Combinatorics*. COCOON 2017. Lecture Notes in Computer Science, vol. 10392. Springer, Cham.
2. TURCSANYI P, KRIEGOVA E, KUDELKA M *et al*.: Improving risk-stratification of patients with chronic lymphocytic leukemia using multivariate patient similarity networks. *Leuk Res* 2019; 79: 60-68.
3. PAI S, BADER GD: Patient Similarity Networks for Precision Medicine. *J Mol Biol* 2018; 430: 2924-38.
4. LIAW A, WIENER M: Classification and regression by randomForest. *R News* 2002; 2: 18-22.
5. FRITSCH S, GUENTHER F, SULING M, HUBERT M, MUELLER SM: 2016. R language Neuralnet package version 1.13.