Supplementary Document

Shigeharu Ishida,[†] Hideaki Umeyama,[‡] Mitsuo Iwadate,[‡] and Y-H. Taguchi[*,†]

Department of Physics, Chuo University, Tokyo 112-8551, Japan, and Department of Biological Sciences, Chuo University, Tokyo 112-8551, Japan

E-mail: tag@granular.com

Thresholds and criteria for selection

Even if we develop a *score* that measures the importance of genes and we rank all genes based on this score, there are no strict criteria available that could be used to decide how many genes we should consider as significant for three autoimmune diseases. If we consider larger numbers of genes, although their significance decreases monotonically, this would allow us to analyze more factors that may have critical roles in three autoimmune diseases. However, using a limited number of genes increases significance but may be too stringent. Therefore, for practical use, a threshold value can be chosen to keep consistency between gene selections for each disease.

As described in the main text, we selected genes for each disease as follows:

- SLE: a set of genes G_{SLE} is defined as

$$G_{\mathrm{SLE}} \equiv \{g \mid \mathrm{PC2}_g < d\}$$

[*]To whom correspondence should be addressed
[†]Department of Physics, Chuo University, Tokyo 112-8551, Japan
[‡]Department of Biological Sciences, Chuo University, Tokyo 112-8551, Japan

- RA: a set of genes G_{RA} is defined as

$$G_{\mathrm{RA}} \equiv \{g \mid \mathrm{PC2}_g > d\}$$

- DM: a set of genes G_{DM} is defined as

$$G_{\mathrm{DM}} \equiv \{g \mid \mathrm{PC3}_g < d\}$$

where $\mathrm{PC}N_g$ is the Nth principal component score of gth gene and d is the threshold value. Then we computed the ratio r of the number of genes commonly selected for three diseases to the number of genes selected for a single disease,

$$r \equiv \frac{N(G_{\mathrm{SLE}} \bigcap G_{\mathrm{RA}} \bigcap G_{\mathrm{DM}})}{N(G_{\mathrm{SLE}} \bigcup G_{\mathrm{RA}} \bigcup G_{\mathrm{DM}})},$$

where $N(G)$ is the number of genes that belong to set G, as a function of d (Fig. S18). As d increases, r also increases, but as d further increases, r starts to decrease. Regarding values of r, $0.1 < d < 0.2$ values are equally significant. However, since a larger d implies greater significance in each selection, we chose a value of $d = 0.2$ as a maximum that does not reduce r.

Comparisons with previous feature selection methods

Although it was apparent that our proposed method successfully selected common genes that are critical for all three autoimmune diseases, it would be interesting to compare the standard or conventional feature selection methods to see if they were equal to or better than this method. Therefore, we compared the performance of methods employed by two previous studies[1, 2].

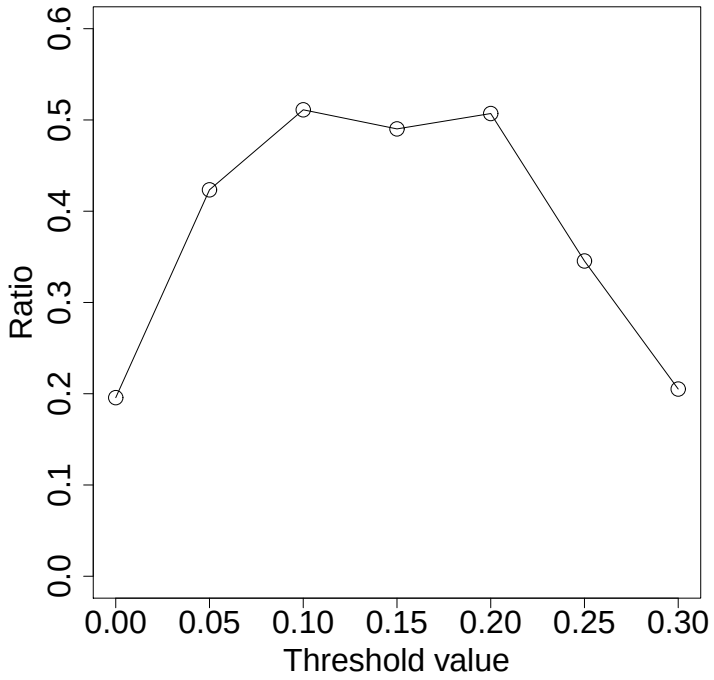


Figure S18: The dependence of the ratio r on threshold value d.

Weka

Weka[3] is a popular and conventional package that includes several feature selection methods. Wang et al. [1] used Weka to check the abilities of several feature selection methods when applied to gene selection based on gene expression. Here, we applied their trials to our data set.

Minimum description length (MDL) based discretization

First, we tried to discretize methylation values so we could apply ranking methods based on the discretized values similar to the study by Wang et al. However, discretization obtained by MDL based discretization methods attributed the same constant to most values of many features. This may be because we only used 20 samples, consisting of five patients and 14 or 15 healthy controls. The discretized values obtained were used for χ^2-statistic, InfoGain, ReliafF and Symmetrical uncertainty.

Gene selection

In Tables S2, S3, S4, and S5, we listed genes that were highly ranked based on several scores. Although we tried to rank the top 10 genes as in Wang et al., because of many tied values, this was not possible. Thus, we applied an alternative approach. As SLE had many tied top ranked genes, we first listed all of them. Then, we listed genes for RA whose number was not less than that of the genes listed in SLE but as small as possible. Because of tied values, we could not select the exact same number of highly ranked genes as in SLE. As DM always had a limited number of genes to which non-zero scores were attributed and the number was always less than those listed genes for both SLE and RA, we listed all genes to which non-zero scores were attributed in DM. In Tables S6 and S7, we listed genes selected by correlation-based feature selection (CFS) and several wrapper methods with best first criterion.

Genes overlaps

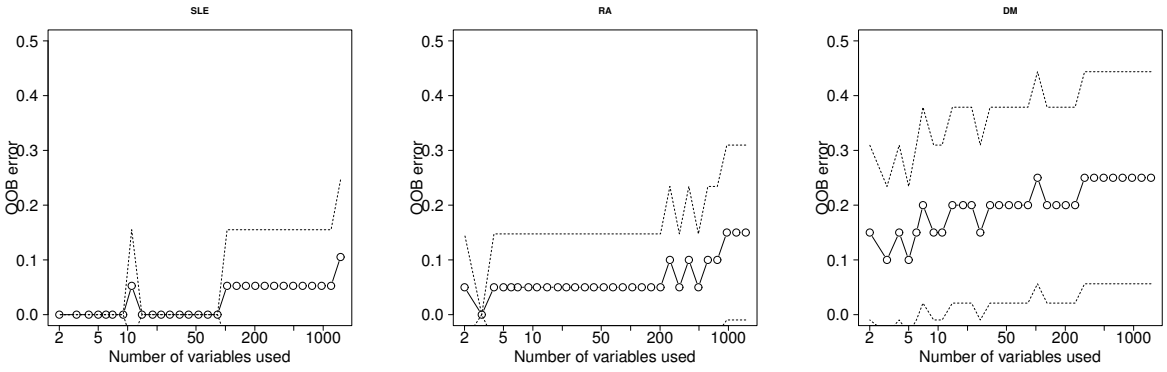
For tables S2, S3,S4,S5,S6, and S7, we identified gene overlaps as follows:

- **Bold** indicates overlap with genes selected by the present study (Table 2).
- *Italic* indicates genes selected by only one method for each disease.
- Underlined indicates genes selected for more than one disease for each method.

It is clear that the number of bold faced genes outperforms the number of both italic and underlined genes. This demonstrates the plausibility of our method. In contrast to this, there is only one underlined gene (TES in Table S6). This suggests that none of the methods listed here have the ability to list selected genes common for three diseases. Despite this, the dependence of gene selection by the method used is low. There are few Italic faced genes. Although there is a relatively high number of Italic faced genes in Table S6, this is simply because CFS for SLE lists more genes than for other methods. Thus, any methods that make use of classification or labeling information cannot list selected genes common for three diseases.

Random forest

Diaz-Uriarte et al. employed random forest for gene selection based on gene expression [2]. In this subsection, we tried to apply their methodology to our data set. We used varSelRF[4] (ntree = 2000, ntreeIterat = 1000, vars.drop.frac = 0.2) following the procedure described by Diaz-Uriarte et al. We executed varSelRF a number of times, but it provided only a few genes (see typical examples in Fig. S19) and the selected genes heavily fluctuated between trials. Thus, to collect enough numbers of selected genes, we executed varSelRF 100 times (see results in Table S8). The meanings of Bold, Italic and underlined are the same as those in the previous subsection. The outcome was quite similar to the previous section. Again, there were high numbers of bold faced genes and TES was the only gene selected for more than one disease. One difference was increased numbers of Italic faced genes in DM, but this reflected that very few genes were selected for DM in the previous subsection, while random forest listed more genes. In conclusion, random forest is not useful for the selection of critical genes common for three autoimmune diseases.



(a) (b) (c)

Figure S19: Typical examples of error rates of random forest as a function of the number of selected genes. Solid curves: mean, upper/lower dotted curves: upper/lower bounds, respectively. (a) SLE, selected genes are RARA_P1076_R and S100A2_E36_R. (b) RA, selected genes are PLAGL1_P334_F, CDC25B_E83_F, and ABCA1_P45_F. (c) DM, selected genes are ESR1_P151_R and IL4_P262_R.

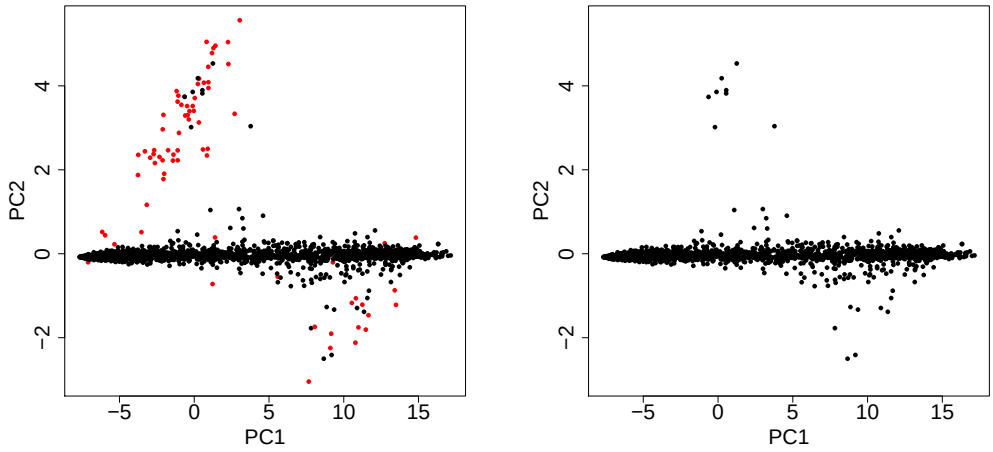
Comparisons with other data sets

Although we successfully selected important genes common for three autoimmune diseases, if we could reproduce the outcome using an independent dataset, our results would be more feasible. We could not find an identical data set, i.e., one that consisted of MZ twins of SLE, RA and DM patients with samples taken from white blood cells, so we used two similar data sets for RA[5] and SLE[6]. Here, we present the results for these two data sets.

RA

Liu et al. [5] recently reported methylation patterns of genomes from peripheral blood lymphocytes (PBL) of RA patients and healthy controls. Since they did not collect MZ twins, they collected a large number of samples (335 patients and 354 healthy controls). Although the study did not include SLE or DM, the number of samples is huge, and therefore is a good data set to check the feasibility of our method. First, we downloaded GSE42861_processed_methylation_matrix.txt.gz from GEO ID: GSE42861. Then we selected probes annotated as either TSS200 or TSS1500 to collect methylation patterns limited to promoter regions. Then, promoters of genes that exist on microarray plates used in our study were chosen. Finally, 3540 probes were chosen, more than twice as many probes as used in our study (1505). Fig. S20(a) shows two dimensional embeddings by PCA (the first and second PCs). Similar to Fig. S1(a), it shows a barb-like structure. This structure vanished after removing genes located on the X chromosome (red dots in Fig. S20(a) and see Fig. S20(b) for the embeddings after removing genes on the X chromosome). Next, we checked which PC represented the distinction between RA patients and normal controls. Fig. S21 shows the contributions of each sample to the first, second and third principal components. As expected, the first principal components have almost constant values for both RA patients and normal controls. However, the second principal components seem to represent a distinction between RA patients and normal controls. To confirm this, we applied Wilcoxon Rank sum test between the second principal components of patients and normal controls and found that $P < 1 \times 10^{-16}$. Thus, PC2

likely represents a distinction between RA patients and normal controls. Furthermore, the third PC may represent an odd structure that is beyond the scope of this paper.



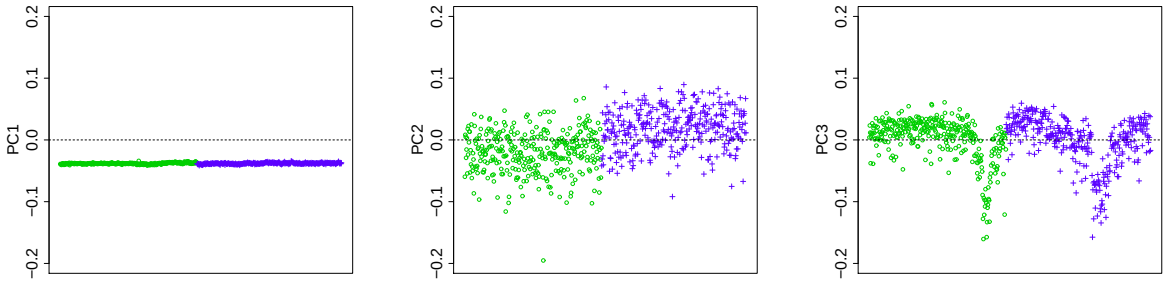
(a) (b)

Figure S20: Two dimensional PCA embeddings of 3540 probes extracted from GEO ID: GSE41861. (a) PC1 vs PC2 before removing genes located on the X chromosome (red dots). (b) PC1 vs PC2 after removing genes located on the X chromosome.

Next, we attempted to select genes as we did in the present study (red dots in Fig. S22 and Table S9). The number of genes was not uniquely decided but we selected 40 genes. However, since more than one probe is sometimes attributed to one gene only, 34 genes were selected. In this study, we selected 36 genes among 813 genes (Table 2). Thus, the expected number of bold faced genes in Table S9 was $34 \times \frac{36}{813} \simeq 1.5$, which was markedly lower than the number of bold faced genes, 11, in Table S9. This strong coincidence demonstrates the feasibility of our method, if we consider the fact that samples were not taken from white blood cells (as in our study) but were extracted from peripheral blood lymphocytes (PBLs).

SLE

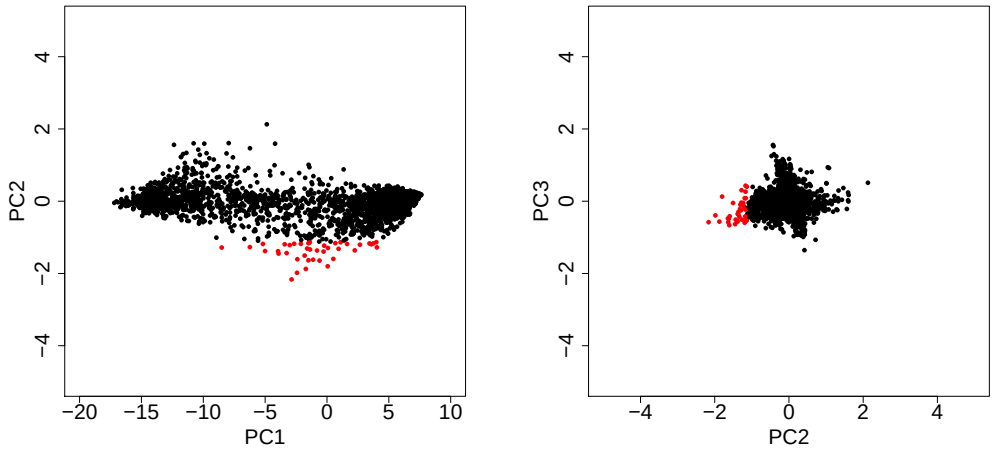
Jeffries et al. [6] collected genomes from T-cells of 11 female SLE patients and 12 normal female controls. Although the number of samples was not much larger than the number of SLE sam-



(a) (b) (c)

Figure S21: The first (a), second (b) and third (c) principal components. Green circles and blue crosses are RA patients and normal controls, respectively.

ples in the present study, 19 (Table 1), and did not include MZ twins, application of our method to this data set may also validate our method. Promoter methylation pattern, GSE27895_non-normalized_ratios.txt was downloaded from GEO ID: GSE27895. Then, probes annotated as genes that existed on our microarray plate and used in the present study were collected. The number of probes collected was 2138. Since the sample consisted only of females, there was no need to exclude genes located on the X chromosome. Figs. S23 and S24 represent two dimensional embeddings and principal components, respectively. Initially, the principal component that represents distinction between SLE patients and normal controls did not appear to exist. However, if one compares Fig. 1 from Jeffries et al. with Fig. S24(b), PC2 corresponds to Fig. 1 from Jeffries et al., which represents the expression of genes selected to represent distinctions between SLE patients and normal controls. In both Fig. 1 from Jeffries et al. and Fig. S24(b), Control 9, 10 and 12 have a distinct expression from other control samples. PC2 simply reflects this feature correctly. It should be noted that this kind of anomaly, i.e., diversity not between but within normal controls or patient samples, was also observed in the present study. We also noted that promoter methylation differed within normal or patient samples dependent upon gender (e.g., Figs. S2(b), S11(a) and S11(b)). Although the anomaly in Fig. S24 was not due to gender because all samples were female, it is reasonable to treat Control 9, 10 and 12 as distinct samples from other controls



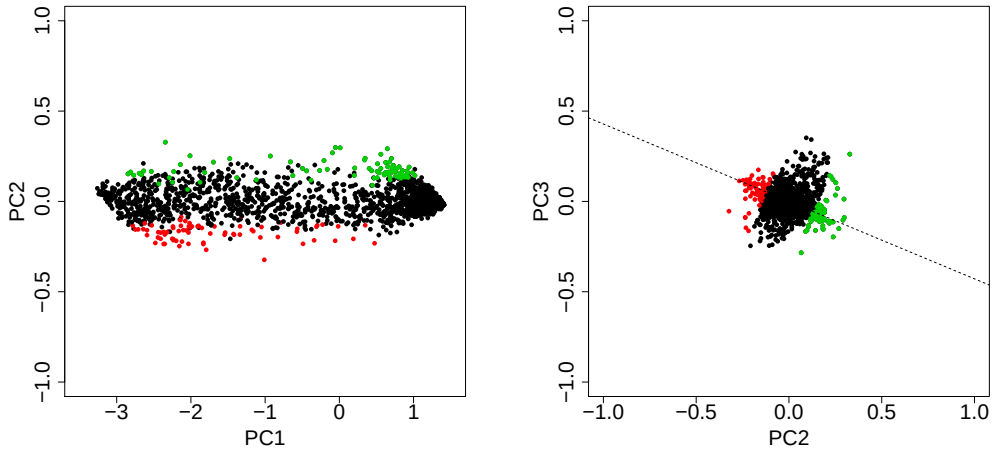
(a) (b)

Figure S22: Two dimensional PCA re-embeddings of 3540 probes extracted from GEO ID: GSE41861 after removing genes on the X chromosome. (a) PC1 vs PC2 (b) PC2 vs PC3. Red dots represent genes selected as critical for RA.

because of their unknown biologically distinct features.

Next, we attempted to select genes based on the proposed method. Close examination of Fig. S23(b) shows the hump is directly along the second PC but is slightly tilted (dotted line in Fig. S23(b)). To determine whether this direction represented a distinction between normal controls and SLE patients, we generated combined PC, $0.7\text{PC2} - 0.3\text{PC3}$ (Fig. S24(d)). Although it was not apparent, the new combined PC, $0.7\text{PC2} - 0.3\text{PC3}$, does represent a distinction between SLE patients and normal controls if we compare Control 9, 10 and 12 or other controls with SLE patients. *P*-values computed by Wilcoxon Rank sum test of combined PC,$0.7\text{PC2} - 0.3\text{PC3}$, between Control 9, 10 and 12 or other controls and SLE patients were 0.005 and 0.03, respectively. Thus, the difference is statistically significant. Selected genes based on the combined PC (red and green dots in Fig. S23) are listed in Table S9. *P*-values for the number of bold faced genes (i.e., overlaps between Table 2 and S9) were computed by binomial distribution. $P = 0.02$ and 0.28 for red and green dots, respectively. Thus, red dots significantly overlapped with genes selected by the present study (Table 2). Although this coincidence is not very strong, it is remarkable if we consider that samples were taken not from white blood cells as in the present study, but were

extracted from T-cells. Thus, this demonstrates that the PCA based feature extraction method can work for other data sets.



(a) (b)

Figure S23: The two dimensional PCA embeddings of 2138 probes extracted from GEO ID: GSE27895. (a) PC1 vs PC2 (b) PC2 vs PC3. Red and green dots represent genes selected as critical for SLE. Dotted line, $0.7\text{PC2} - 0.3\text{PC3} = 0$, represents the distinction between normal controls and SLE patients (see text for more details).

References

[1] Wang, Y.; Tetko, I. V.; Hall, M. A.; Frank, E.; Facius, A.; Mayer, K. F.; Mewes, H. W. Gene selection from microarray data for cancer classification–a machine learning approach. *Comput Biol Chem* **2005**, *29*, 37–46.

[2] Diaz-Uriarte, R.; Alvarez de Andres, S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **2006**, *7*, 3.

[3] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* **2009**, *11*.

[4] Diaz-Uriarte, R.; Alvarez de Andres, S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **2006**, *7*, 3.

[5] Liu, Y. et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat. Biotechnol.* **2013**,

[6] Jeffries, M. A.; Dozmorov, M.; Tang, Y.; Merrill, J. T.; Wren, J. D.; Sawalha, A. H. Genome-wide DNA methylation patterns in CD4+ T cells from patients with systemic lupus erythematosus. *Epigenetics* **2011**, *6*, 593–601.

Table S2: χ^2-statistic based feature extraction.

SLE		RA		DM	
Score	gene	Score	gene	Score	gene
19	CD34_P780_R	20	ABCA1_P45_F	10.588	**CARD15_P302_R**
19	CD9_P585_R	15.556	NQO1_E74_R	10.588	ESR1_P151_R
19	**CSF3R_P8_F**	15.556	PLAGL1_P334_F	10.588	IL4_P262_R
19	**DHCR24_P652_R**	15.556	PTGS1_E80_F	10.588	MCAM_P265_R
19	EMR3_E61_F	15	CDC25B_E83_F	10.588	*PRSS1_E45_R*
19	EPHA2_P203_F	15	CDH3_P87_R	0	
19	EPHA5_P66_F	15	CDKN1C_P6_R	0	
19	**HOXB2_P99_F**	15	DLL1_P386_F	0	
19	**IFNGR2_P377_R**	15	GNMT_E126_F	0	
19	IL10_P348_F	15	IGSF4C_P533_R	0	
19	IL10_P85_F	15	IRAK3_P13_F	0	
19	**LMO2_E148_F**	15	ONECUT2_P315_R	0	
19	**LMO2_P794_R**	15	PLAU_P176_R	0	
19	MAP3K8_P1036_F	15	TERT_E20_F	0	
19	**MMP14_P13_F**	15	TK1_E47_F	0	
19	**PECAM1_P135_F**	12.381	APC_E117_R	0	
19	PMP22_P1254_F	12.381	ARHGAP9_P260_F	0	
19	**RARA_P1076_R**	12.381	E2F3_P840_R	0	
19	**S100A2_E36_R**	12.381	EDN1_E50_R	0	
19	**SLC22A18_P216_R**	12.381	FVT1_P225_F	0	
19	**SPI1_E205_F**	12.381	IL12B_P1453_F	0	
19	**SPI1_P48_F**	12.381	ITGB1_P451_F	0	
19	**SPI1_P48_F**	12.381	JAK3_P156_R	0	
		12.381	PLAU_P11_F	0	
		12.381	TES_E172_F	0	
		12.381	TNC_P198_F	0	

Table S3: InfoGain based feature extraction.

SLE		RA		DM	
Score	gene	Score	gene	Score	gene
0.831	CD34_P780_R	0.811	ABCA1_P45_F	0.367	**CARD15_P302_R**
0.831	CD9_P585_R	0.616	NQO1_E74_R	0.367	ESR1_P151_R
0.831	**CSF3R_P8_F**	0.616	PLAGL1_P334_F	0.367	IL4_P262_R
0.831	**DHCR24_P652_R**	0.616	PTGS1_E80_F	0.367	MCAM_P265_R
0.831	EMR3_E61_F	0.541	CDC25B_E83_F	0.367	PRSS1_E45_R
0.831	EPHA2_P203_F	0.541	CDH3_P87_R	0	
0.831	EPHA5_P66_F	0.541	CDKN1C_P6_R	0	
0.831	**HOXB2_P99_F**	0.541	DLL1_P386_F	0	
0.831	**IFNGR2_P377_R**	0.541	GNMT_E126_F	0	
0.831	IL10_P348_F	0.541	IGSF4C_P533_R	0	
0.831	IL10_P85_F	0.541	IRAK3_P13_F	0	
0.831	**LMO2_P794_R**	0.541	ONECUT2_P315_R	0	
0.831	**LMO2_E148_F**	0.541	PLAU_P176_R	0	
0.831	MAP3K8_P1036_F	0.541	TERT_E20_F	0	
0.831	**MMP14_P13_F**	0.541	TK1_E47_F	0	
0.831	**PECAM1_P135_F**	0.509	APC_E117_R	0	
0.831	PMP22_P1254_F	0.509	ARHGAP9_P260_F	0	
0.831	**RARA_P1076_R**	0.509	E2F3_P840_R	0	
0.831	**SLC22A18_P216_R**	0.509	EDN1_E50_R	0	
0.831	**SPI1_E205_F**	0.509	FVT1_P225_F	0	
0.831	**SPI1_P48_F**	0.509	IL12B_P1453_F	0	
0.831	**S100A2_E36_R**	0.509	ITGB1_P451_F	0	
0.831	**SPI1_P48_F**	0.509	JAK3_P156_R	0	
		0.509	PLAU_P11_F	0	
		0.509	TES_E172_F	0	
		0.509	TNC_P198_F	0	

Table S4: Relief based feature extraction.

SLE		RA		DM	
Score	gene	Score	gene	Score	gene
1	CD34_P780_R	1	ABCA1_P45_F	0.45	**CARD15_P302_R**
1	CD9_P585_R	0.87	NQO1_E74_R	0.45	ESR1_P151_R
1	**CSF3R_P8_F**	0.87	PTGS1_E80_F	0.45	IL4_P262_R
1	**DHCR24_P652_R**	0.825	PLAGL1_P334_F	0.45	MCAM_P265_R
1	EMR3_E61_F	0.76	E2F3_P840_R	0.45	PRSS1_E45_R
1	EPHA2_P203_F	0.76	ITGB1_P451_F	0	
1	EPHA5_P66_F	0.76	JAK3_P156_R	0	
1	**HOXB2_P99_F**	0.76	PLAU_P11_F	0	
1	**IFNGR2_P377_R**	0.75	APC_E117_R	0	
1	IL10_P348_F	0.75	FVT1_P225_F	0	
1	IL10_P85_F	0.75	IL12B_P1453_F	0	
1	**LMO2_E148_F**	0.74	ARHGAP9_P260_F	0	
1	**LMO2_P794_R**	0.705	EDN1_E50_R	0	
1	MAP3K8_P1036_F	0.7	CDC25B_E83_F	0	
1	**MMP14_P13_F**	0.7	CDH3_P87_R	0	
1	**PECAM1_P135_F**	0.7	CDKN1C_P6_R	0	
1	PMP22_P1254_F	0.7	DLL1_P386_F	0	
1	**RARA_P1076_R**	0.7	GNMT_E126_F	0	
1	**S100A2_E36_R**	0.7	IGSF4C_P533_R	0	
1	**SLC22A18_P216_R**	0.7	IRAK3_P13_F	0	
1	**SPI1_E205_F**	0.7	ONECUT2_P315_R	0	
1	**SPI1_P48_F**	0.7	PLAU_P176_R	0	
1	**SPI1_P48_F**	0.7	TERT_E20_F	0	
		0.7	TK1_E47_F	0	

Table S5: Symmetrical Uncertainty based feature extraction.

SLE		RA		DM	
Score	gene	Score	gene	Score	gene
1	CD34_P780_R	1	ABCA1_P45_F	0.517	**CARD15_P302_R**
1	CD9_P585_R	0.728	NQO1_E74_R	0.517	ESR1_P151_R
1	**CSF3R_P8_F**	0.728	PLAGL1_P334_F	0.517	IL4_P262_R
1	**DHCR24_P652_R**	0.728	PTGS1_E80_F	0.517	MCAM_P265_R
1	EMR3_E61_F	0.706	CDC25B_E83_F	0.517	PRSS1_E45_R
1	EPHA2_P203_F	0.706	CDH3_P87_R		
1	EPHA5_P66_F	0.706	CDKN1C_P6_R		
1	**HOXB2_P99_F**	0.706	DLL1_P386_F		
1	**IFNGR2_P377_R**	0.706	GNMT_E126_F		
1	IL10_P348_F	0.706	IGSF4C_P533_R		
1	IL10_P85_F	0.706	IRAK3_P13_F		
1	**LMO2_E148_F**	0.706	ONECUT2_P315_R		
1	**LMO2_P794_R**	0.706	PLAU_P176_R		
1	MAP3K8_P1036_F	0.706	TERT_E20_F		
1	**MMP14_P13_F**	0.706	TK1_E47_F		
1	**PECAM1_P135_F**	0.583	APC_E117_R		
1	PMP22_P1254_F	0.583	ARHGAP9_P260_F		
1	**RARA_P1076_R**	0.583	E2F3_P840_R		
1	**S100A2_E36_R**	0.583	EDN1_E50_R		
1	**SLC22A18_P216_R**	0.583	FVT1_P225_F		
1	**SPI1_E205_F**	0.583	IL12B_P1453_F		
1	**SPI1_P48_F**	0.583	ITGB1_P451_F		
1	**SPI1_P48_F**	0.583	JAK3_P156_R		
		0.583	PLAU_P11_F		
		0.583	TES_E172_F		
		0.583	TNC_P198_F		

Table S6: CFS based feature extraction

SLE	RA	DM
ABCC2_P88_F	ABCA1_P45_F	*BCAP31_P1072_F*
ACTG2_E98_R	ARHGAP9_P518_R	**CARD15_P302_R**
AIM2_P624_F	CDC25B_E83_F	*E2F3_P840_R*
ATP10A_P147_F	*GABRA5_E44_R*	ESR1_P151_R
BIRC4_P500_F	IGSF4C_P533_R	
BTK_P105_F	*KLK10_P268_R*	
CD34_P780_R	NQO1_E74_R	
CD82_P557_R	PLAGL1_P334_F	
CD9_P585_R	PTGS1_E80_F	
CSF3R_P8_F	TES_E172_F	
DHCR24_P652_R	TK1_E47_F	
EMR3_E61_F		
EPHA2_P203_F		
EPHA5_P66_F		
EPHB2_P165_R		
HDAC9_P137_R		
HOXB2_P99_F		
HPN_P374_R		
HTR1B_E232_R		
IFNGR2_P377_R		
IGFBP6_P328_R		
IL10_P348_F		
IL10_P85_F		
LMO2_E148_F		
LMO2_P794_R		
MAP3K8_P1036_F		
MMP14_P13_F		
PECAM1_P135_F		
PMP22_P1254_F		
RARA_P1076_R		
S100A2_E36_R		
SLC22A18_P216_R		
SPI1_E205_F		
SPI1_P48_F		
SPI1_P48_F		
TES_E172_F		

Table S7: Wrapper based feature extraction.

	SLE	RA	DM
J48	CD34_P780_R	ABCA1_P45_F	IL4_P262_R
Naive Bayes	CD9_P585_R	ABCA1_P45_F	*CCNC_P132_R* *CDKN2B_E220_F* *IGFBP5_P9_R*
SMO	CD34_P780_R *IMPACT_P186_F* TES_E172_F	ABCA1_P45_F	*AATK_E63_R* *MYH11_P22_F* *NGFR_P355_F*

Table S8: Random forest based feature selection. *N* represents the selection frequencies of each gene during 100 trials.

SLE		RA		DM	
gene	N	gene	N	gene	N
S100A2_E36_R	18	ABCA1_P45_F	100	ESR1_P151_R	61
EMR3_E61_F	18	PTGS1_E80_F	42	IL4_P262_R	61
LMO2_E148_F	18	NQO1_E74_R	36	PRSS1_E45_R	51
SPI1_P48_F	16	PLAGL1_P334_F	34	NGFR_P355_F	23
CSF3R_P8_F	16	IGSF4C_P533_R	9	MCAM_P265_R	15
LMO2_P794_R	15	ONECUT2_P315_R	5	*LYN_P241_F*	13
CD9_P585_R	13	DLL1_P386_F	4	*SNRPN_seq_12_S127_F*	6
DHCR24_P652_R	13	GNMT_E126_F	4	**CARD15_P302_R**	6
EPHA2_P203_F	13	CDKN1C_P6_R	3	*FGF6_E294_F*	4
HOXB2_P99_F	13	PLAU_P176_R	2	*HLA-DOB_E432_R*	3
CD34_P780_R	12	TK1_E47_F	2	*IPF1_P750_F*	3
MAP3K8_P1036_F	12	CDC25B_E83_F	2	*MAD2L1_E93_F*	3
IL10_P85_F	11	IRAK3_P13_F	2	TES_P182_F	2
MMP14_P13_F	9	TES_E172_F	1	*FGF9_P862_R*	2
PMP22_P1254_F	8			*ARHGAP9_P518_R*	2
RARA_P1076_R	8			*MMP7_E59_F*	2
SYK_P584_F	8			BCAP31_P1072_F	2
IL10_P348_F	8			*PENK_E26_F*	1
PECAM1_P135_F	7			*PTHR1_P258_F*	1
IFNGR2_P377_R	7			*RASGRF1_P768_F*	1
SPI1_E205_F	5			*TGFB1_P833_R*	1
EPHA5_P66_F	5			*ELK3_P514_F*	1
SLC22A18_P216_R	3			*FGR_P39_F*	1
				IGFBP5_P9_R	1

Table S9: Genes selected in RA samples (red dots in Fig. S20) and SLE samples (red and green dots in Fig. S23). *N* is the number of probes selected and annotated to the gene.

RA (red)		SLE (red)				SLE (green)			
gene	*N*	gene	*N*	gene	*N*	gene	*N*	gene	*N*
ERCC3	2	ATP10A	4	CHFR	3	DCC	4	CASP8	3
IFNGR2	2	**HOXB2**	2	**MMP8**	2	**SLC22A18**	2	H19	2
RARA	2	DAPK1	2	DIRAS3	2	IL8	2	**AIM2**	1
FGR	2	GNAS	2	KCNQ1	2	**RARA**	1	**SPP1**	1
IL10	2	MGMT	2	PEG10	2	ABCG2	1	ACVR1B	1
TP73	2	WT1	2	**CARD15**	1	ATP10A	1	APC	1
DHCR24	1	**CSF3R**	1	**MPL**	1	BCAP31	1	BRCA1	1
HOXB2	1	**SLC22A18**	1	ABCC2	1	CALCA	1	CASP10	1
IFNGR2	1	ACVR1	1	AFP	1	CCND1	1	CD2	1
LCN2	1	AR	1	ASB4	1	CDKN2A	1	CHD2	1
LMO2	1	B3GALT5	1	BCL2A1	1	CHI3L2	1	COL6A1	1
MMP14	1	BLK	1	BRCA1	1	DIRAS3	1	DSP	1
PI3	1	CCKAR	1	CCNC	1	EPHA3	1	EPHA7	1
S100A12	1	CXCL9	1	DDR2	1	EPHB6	1	ERBB2	1
AGXT	1	DMP1	1	EDNRB	1	ERCC1	1	ERG	1
AREG	1	FANCG	1	FGR	1	ESR1	1	ETS2	1
CD9	1	FHL1	1	FMR1	1	FER	1	FGF1	1
ENC1	1	FRK	1	GABRA5	1	FGF2	1	FGFR4	1
EPHA2	1	GSTP1	1	H19	1	FLI1	1	GPATC3	1
EXT1	1	HLA-DOB	1	HSPA2	1	HDAC6	1	HFE	1
FER	1	HTR2A	1	LIG4	1	HLA-DPB1	1	IGF2AS	1
FLI1	1	IL1RN	1	IL3	1	IGFBP7	1	INSR	1
GFI1	1	IL10	1	KRT1	1	IRF7	1	ITK	1
GNMT	1	MAPK9	1	MC2R	1	KCNQ1	1	KLK11	1
GPX3	1	MGMT	1	MUC1	1	MAGEL2	1	MAP2K6	1
GSTP1	1	NID1	1	NOS3	1	MGMT	1	MLH1	1
HOXA9	1	SGCE	1	SNRPN	1	MME	1	MMP10	1
ISL1	1	TIMP3	1	TNK1	1	MYLK	1	PDE1B	1
NPR2	1	TP73	1	WNT8B	1	PLAGL1	1	PEG10	1
NR2F6	1	XRCC2	1	ZIM2	1	PLAGL1	1	PRKCDBP	1
NRG1	1					PTGS1	1	RAD50	1
PIK3R1	1					RAD54B	1	RIPK4	1
PRSS8	1					RUNX3	1	SEMA3B	1
SFTPD	1					SNRPN	1	SOX17	1
						TEK	1	TFRC	1
						TFPI2	1	TGFB2	1
						TJP2	1	THBS1	1
						TNC	1	TNFRSF10C	1
						TRIM29	1	VEGFB	1

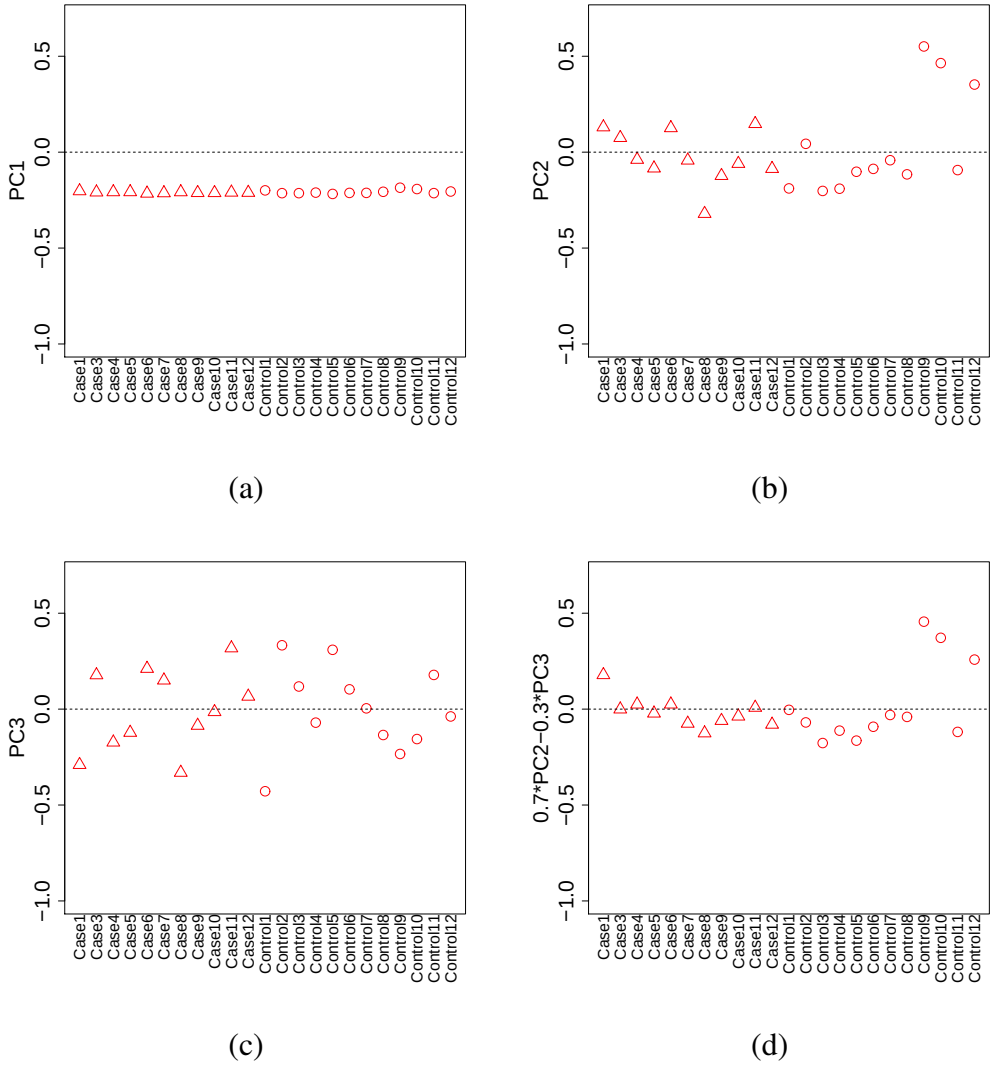


Figure S24: The first (a), second (b) and third (c) principal components. Red triangles and circles represent SLE patients and normal controls, respectively. (d) The combined PC, $0.7\text{PC2} - 0.3\text{PC3}$, represents the distinction between SLE patients and normal controls.