Supplementary Information

Microbiomic Signatures of Psoriasis: Feasibility and Methodology Comparison

Alexander Statnikov[1,2,§], Alexander V. Alekseyenko[1,2], Zhiguo Li[1], Mikael Henaff[1], Guillermo I. Perez-Perez[2,3], Martin J. Blaser[2,3,5], Constantin F. Aliferis[1,4 §]

[1] Center for Health Informatics and Bioinformatics (CHIBI), New York University Langone Medical Center, New York, New York

[2] Department of Medicine, New York University School of Medicine, New York, New York

[3] Department of Microbiology, New York University School of Medicine, New York, New York

[4] Department of Pathology, New York University School of Medicine, New York, New York

[5] Medical Service, Department of Veterans Affairs New York Harbor Healthcare System, New York, New York

[§] Corresponding authors

Table S1: Feature/taxa selection methods/variants applied in this study*.

Method	*Abbreviations of variants and details*	*Reference*
Generalized Local Learning	• GLL: 5% alpha, Fisher's Z test, max-k = 1	1,2
SVM-based Recursive Feature Elimination	• SVM-RFE1: with stat. comp. • SVM-RFE2: w/o stat. comp.	3
Univ. based on the Kruskal-Wallis non-parametric one-way ANOVA test	• UAF-KW1: b.w., SVM eval. with stat. comp. • UAF-KW2: b.w., SVM eval. w/o stat. comp. • UAF-KW-FDR: using features that are significant at 5% FDR	4-7
Univ. base on the signal-to-noise ratio	• UAF-SN1: b.w., SVM eval. with stat. comp. • UAF-SN2: b.w., SVM eval. w/o stat. comp.	
Univ. based on the ratio of between-group to within-group sum of squares	• UAF-BW1: b.w., SVM eval. with stat. comp. • UAF-BW2: b.w., SVM eval. w/o stat. comp.	
Univ. based on the two-sample t-test	• UAF-T1: b.w., SVM eval. with stat. comp. • UAF-T2: b.w., SVM eval. w/o stat. comp. • UAF-T-FDR: using features that are significant at 5% FDR	
Univ. based on the X^2-test	• UAF-X21: b.w., SVM eval. with stat. comp. • UAF-X22: b.w., SVM eval. w/o stat. comp. • UAF-X2-FDR: using features that are significant at 5% FDR	
Minimum Redundancy and Maximum Relevancy	• MRMR1: b.w. on top-1000 ranking features, SVM eval. with stat. comp. • MRMR2: b.w. on top-1000 ranking features, SVM eval. w/o stat. comp. • MRMR3: b.w. on top-50 ranking features, SVM eval. with stat. comp. • MRMR4: b.w. on top-50 ranking features, SVM eval. w/o stat. comp. • MRMR5: b.w. on top-200 ranking features, SVM eval. with stat. comp. • MRMR6: b.w. on top-200 ranking features, SVM eval. w/o stat. comp.	8,9
Random Forest based Variable Selection	• RFVS1: b.w., with stat. comp. • RFVS2: b.w., w/o stat. comp.	10
Least-Angle Regression Elastic Net	• LARS-EN1: b.w. on regression coefficients, SVM eval. with stat. comp. • LARS-EN2: b.w. on regression coefficients, SVM eval. w/o stat. comp.	11
Soft Independent Modeling of Class Analogy	• SIMCA: original method • SIMCA-SVM1: b.w., SVM eval. with stat. comp. • SIMCA-SVM2: b.w., SVM eval. w/o stat. comp.	12
Principal Component Analysis	• PCA1: b.w. on the first principal component loadings of features, SVM eval. with stat. comp. • PCA2: b.w. on the first principal component loadings of features, SVM eval. w/o stat. comp.	13
Sparse Principal Component Analysis	• SPCA1: b.w. on the first sparse principal component loadings of features, SVM eval. with stat. comp. • SPCA2: b.w. on the first sparse principal component loadings of features, SVM eval. w/o stat. comp.	
Threshold Gradient Descent Regularization	• TGDR1: b.w. on regression coefficients, SVM eval. with stat. comp. • TGDR2: b.w. on regression coefficients, SVM eval. w/o stat. comp. • TGDR3: original method	14
No feature/taxa selection	• ALL: Using all features/taxa in the data	-

* 'Univ.' stands for univariate; 'b.w.' stands for backward wrapper; 'stat comp.' stands for statistical comparison of classification accuracy estimates of nested feature subsets; 'SVM eval.' stands for evaluation of the nested subsets of features by linear SVMs; 'FDR' stands for false discovery rate control.

Table S2: Classification accuracy (AUC) for 37 feature selection methods for each of the four classification tasks in data from the *V3-V5* 16S rRNA locus*.

Tasks	GLL	ALL	SVM_RFE1	SVM_RFE2	UAF_KW1	UAF_KW2	UAF_KW_FDR	UAF_SN1	UAF_SN2	UAF_BW1	UAF_BW2	UAF_T1	UAF_T2	UAF_T_FDR	UAF_X21	UAF_X22	UAF_X2_FDR
PN vs. CC	**0.854**	**0.771**	**0.821**	**0.835**	**0.858**	**0.896**	**0.911**	**0.847**	**0.871**	**0.858**	**0.831**	**0.850**	**0.867**	**0.807**	**0.787**	**0.821**	**0.894**
PL vs. CC	**0.806**	**0.752**	**0.779**	**0.799**	**0.825**	**0.824**	**0.832**	**0.800**	**0.824**	**0.799**	**0.788**	**0.799**	**0.807**	0.561	**0.768**	**0.804**	**0.817**
PL vs. PN	**0.754**	0.323	0.589	0.544	**0.669**	0.631	0.500	**0.683**	0.618	0.503	0.450	**0.703**	**0.660**	0.500	0.454	0.419	0.447
CC vs. PL and PN	**0.894**	**0.845**	**0.820**	**0.872**	**0.829**	**0.877**	**0.874**	**0.854**	**0.875**	**0.857**	**0.852**	**0.860**	**0.877**	**0.884**	**0.789**	**0.844**	**0.861**
Average	0.827	0.673	0.752	0.762	0.795	0.807	0.779	0.796	0.797	0.754	0.730	0.803	0.802	0.688	0.700	0.722	0.755

Tasks	MRMR1	MRMR2	MRMR3	MRMR4	MRMR5	MRMR6	RFVS1	RFVS2	LARS_EN1	LARS_EN2	SIMCA	SIMCA_SVM1	SIMCA_SVM2	PCA1	PCA2	SPCA1	SPCA2	TGDR1	TGDR2	TGDR3
PN vs. CC	**0.777**	**0.806**	**0.774**	**0.817**	**0.776**	**0.813**	**0.923**	**0.906**	**0.850**	**0.854**	**0.818**	**0.838**	**0.838**	**0.782**	**0.776**	**0.673**	**0.774**	**0.834**	**0.838**	**0.848**
PL vs. CC	**0.767**	**0.801**	**0.767**	**0.802**	**0.767**	**0.804**	**0.851**	**0.849**	**0.800**	**0.813**	**0.782**	**0.806**	**0.800**	0.638	**0.700**	0.636	**0.727**	**0.781**	**0.787**	**0.842**
PL vs. PN	0.478	0.432	0.474	0.459	0.478	0.442	**0.751**	**0.685**	**0.685**	0.645	0.638	0.634	0.611	0.468	0.412	0.454	0.440	0.564	0.568	0.509
CC vs. PL and PN	**0.799**	**0.829**	**0.798**	**0.824**	**0.799**	**0.826**	**0.906**	**0.899**	**0.860**	**0.876**	**0.815**	**0.842**	**0.843**	**0.715**	**0.808**	**0.698**	**0.826**	**0.790**	**0.849**	**0.876**
Average	0.705	0.717	0.703	0.725	0.705	0.721	0.857	0.835	0.799	0.797	0.763	0.780	0.773	0.651	0.674	0.615	0.692	0.742	0.761	0.769

* The results shown with bold underlined font have statistically significant classification accuracy (at 5% alpha-level adjusted for multiple comparisons). For a detailed description of the feature selection methods, see **Table S1**.

Table S3: Number of features/taxa selected by 37 feature selection methods for each of the four classification tasks in data from the *V3-V5* 16S rRNA locus*.

Tasks	GLL	ALL	SVM_RFE1	SVM_RFE2	UAF_KW1	UAF_KW2	UAF_KW_FDR	UAF_SN1	UAF_SN2	UAF_BW1	UAF_BW2	UAF_T1	UAF_T2	UAF_T_FDR	UAF_X21	UAF_X22	UAF_X2_FDR
PN vs. CC	2.80	660.00	1.10	33.84	1.06	14.26	12.63	1.09	19.37	212.26	315.25	1.02	19.07	1.59	1.23	29.75	124.11
PL vs. CC	2.50	660.00	1.22	52.03	1.21	38.98	10.20	1.21	36.08	239.18	342.26	1.12	43.42	0.31	1.26	38.25	155.31
PL vs. PN	2.10	660.00	1.83	26.18	1.54	23.37	0.00	1.86	33.66	173.17	248.30	1.21	24.23	0.00	2.72	66.94	113.95
CC vs. PL and PN	3.70	660.00	1.21	52.20	1.13	38.32	22.07	1.06	47.29	157.17	283.64	1.06	39.72	5.56	1.23	59.88	74.42
Average	2.78	660.00	1.34	41.06	1.24	28.73	11.23	1.31	34.10	195.45	297.36	1.10	31.61	1.87	1.61	48.71	116.95

Tasks	MRMR1	MRMR2	MRMR3	MRMR4	MRMR5	MRMR6	RFVS1	RFVS2	LARS_EN1	LARS_EN2	SIMCA	SIMCA_SVM1	SIMCA_SVM2	PCA1	PCA2	SPCA1	SPCA2	TGDR1	TGDR2	TGDR3
PN vs. CC	2.85	60.68	2.62	14.15	2.77	20.99	41.07	146.99	1.03	21.29	8.28	8.16	11.12	1.06	6.08	54.60	242.87	1.05	4.09	18.60
PL vs. CC	2.12	19.77	2.06	10.91	2.07	13.65	11.94	17.33	1.06	33.74	2.66	14.65	18.61	1.14	7.93	39.26	242.06	1.12	5.54	18.85
PL vs. PN	2.86	53.18	2.34	11.45	2.59	24.20	10.18	37.08	1.38	11.96	4.75	10.80	19.28	1.36	9.24	28.34	119.12	1.24	3.63	12.96
CC vs. PL and PN	2.58	36.28	2.58	14.15	2.60	22.36	7.54	25.58	1.06	33.55	9.11	19.29	27.95	1.29	16.30	46.82	339.78	1.24	6.56	19.30
Average	2.60	42.48	2.40	12.67	2.51	20.30	17.68	56.75	1.13	25.14	6.20	13.23	19.24	1.21	9.89	42.26	235.96	1.16	4.96	17.43

* For a detailed description of the feature selection methods, see **Table S1**.

Table S4: Classification accuracy (AUC) for 37 feature selection methods for each of the four classification tasks in data from the *V1-V3* 16S rRNA locus*.

Tasks	GLL	ALL	SVM_RFE1	SVM_RFE2	UAF_KW1	UAF_KW2	UAF_KW_FDR	UAF_SN1	UAF_SN2	UAF_BW1	UAF_BW2	UAF_T1	UAF_T2	UAF_T_FDR	UAF_X21	UAF_X22	UAF_X2_FDR
PN vs. CC	0.405	0.421	0.454	0.414	0.458	0.400	0.500	0.436	0.389	0.439	0.430	0.445	0.376	0.500	0.478	0.424	0.460
PL vs. CC	**0.751**	0.622	**0.677**	**0.662**	**0.718**	**0.693**	**0.712**	**0.676**	**0.667**	0.634	0.621	**0.682**	**0.674**	0.502	**0.642**	0.635	**0.675**
PL vs. PN	0.576	0.497	0.576	0.545	**0.643**	0.612	0.533	0.579	0.556	0.562	0.537	0.588	0.564	0.499	**0.663**	0.602	**0.664**
CC vs. PL and PN	0.482	0.508	0.488	0.472	0.566	0.525	0.513	0.500	0.470	0.474	0.472	0.525	0.481	0.500	0.514	0.478	0.492
Average	*0.553*	*0.512*	*0.549*	*0.524*	*0.596*	*0.558*	*0.565*	*0.548*	*0.520*	*0.527*	*0.515*	*0.560*	*0.524*	*0.500*	*0.574*	*0.535*	*0.573*

Tasks	MRMR1	MRMR2	MRMR3	MRMR4	MRMR5	MRMR6	RFVS1	RFVS2	LARS_EN1	LARS_EN2	SIMCA	SIMCA_SVM1	SIMCA_SVM2	PCA1	PCA2	SPCA1	SPCA2	TGDR1	TGDR2	TGDR3
PN vs. CC	0.454	0.418	0.461	0.426	0.455	0.412	0.424	0.401	0.449	0.404	0.484	0.460	0.423	0.446	0.421	0.462	0.452	0.451	0.446	0.418
PL vs. CC	0.625	0.621	0.627	**0.639**	0.627	0.633	**0.721**	**0.687**	**0.679**	**0.671**	**0.648**	**0.662**	**0.656**	0.508	0.577	0.493	0.578	**0.661**	**0.652**	**0.639**
PL vs. PN	0.616	0.580	0.615	0.594	0.611	0.590	**0.679**	**0.654**	0.590	0.564	0.570	0.574	0.563	0.495	0.512	0.492	0.537	0.596	0.576	0.560
CC vs. PL and PN	0.501	0.484	0.516	0.480	0.510	0.465	0.585	0.552	0.530	0.497	0.533	0.531	0.497	0.500	0.480	0.487	0.485	0.534	0.520	0.519
Average	*0.549*	*0.525*	*0.555*	*0.535*	*0.551*	*0.525*	*0.602*	*0.573*	*0.562*	*0.534*	*0.559*	*0.557*	*0.535*	*0.487*	*0.498*	*0.483*	*0.513*	*0.560*	*0.548*	*0.534*

* The results shown with bold underlined font have statistically significant classification accuracy (at 5% alpha-level adjusted for multiple comparisons). For a detailed description of the feature selection methods, see **Table S1**.

Table S5: Number of features/taxa selected by 37 feature selection methods for each of the four classification tasks in data from the *V1-V3* 16S rRNA locus*.

Tasks	GLL	ALL	SVM_RFE1	SVM_RFE2	UAF_KW1	UAF_KW2	UAF_KW_FDR	UAF_SN1	UAF_SN2	UAF_BW1	UAF_BW2	UAF_T1	UAF_T2	UAF_T_FDR	UAF_X21	UAF_X22	UAF_X2_FDR
PN vs. CC	2.00	791.00	1.85	59.42	2.49	74.80	0.00	4.11	87.78	94.48	164.47	2.37	86.85	0.00	2.44	66.37	98.50
PL vs. CC	3.80	791.00	1.44	49.62	1.21	38.85	10.01	1.57	40.52	162.88	310.39	1.27	38.21	0.07	2.33	60.48	100.38
PL vs. PN	3.10	791.00	1.84	50.79	1.61	50.51	0.82	2.52	77.16	154.18	287.93	1.70	64.16	0.01	1.95	49.96	84.20
CC vs. PL and PN	4.20	791.00	6.08	88.67	1.51	63.50	0.62	3.87	110.47	93.89	203.94	2.74	85.53	0.01	2.00	77.62	42.81
Average	*3.28*	*791.00*	*2.80*	*62.13*	*1.71*	*56.92*	*2.86*	*3.02*	*78.98*	*126.36*	*241.68*	*2.02*	*68.69*	*0.02*	*2.18*	*63.61*	*81.47*

Tasks	MRMR1	MRMR2	MRMR3	MRMR4	MRMR5	MRMR6	RFVS1	RFVS2	LARS_EN1	LARS_EN2	SIMCA	SIMCA_SVM1	SIMCA_SVM2	PCA1	PCA2	SPCA1	SPCA2	TGDR1	TGDR2	TGDR3
PN vs. CC	7.43	169.05	4.34	16.63	5.59	37.94	11.77	34.85	1.84	39.95	5.58	6.86	27.80	1.18	14.58	12.17	76.18	1.36	2.62	3.92
PL vs. CC	9.90	283.83	4.96	22.22	6.29	59.50	11.57	189.17	1.22	19.28	5.59	11.68	38.25	1.72	20.93	50.76	292.74	1.44	5.19	14.16
PL vs. PN	4.13	70.93	3.52	13.79	4.17	25.32	10.69	61.02	1.66	35.53	4.33	15.37	44.11	1.51	21.69	29.62	167.76	1.53	4.64	13.71
CC vs. PL and PN	14.05	331.43	5.28	19.75	6.14	57.34	6.42	26.60	2.78	38.78	6.92	8.55	40.27	1.29	14.91	23.54	160.19	1.52	2.70	7.04
Average	*8.88*	*213.81*	*4.53*	*18.10*	*5.55*	*45.03*	*10.11*	*77.91*	*1.88*	*33.39*	*5.61*	*10.62*	*37.61*	*1.43*	*18.03*	*29.02*	*174.22*	*1.46*	*3.79*	*9.71*

* For a detailed description of the feature selection methods, see **Table S1**.

Table S6: Comparison of the original *univariate* analysis of the V1-V3 dataset [15] with the present *multivariate* analyses of the V1-V3 and V3-V5 datasets for development of molecular signatures[*].

	PL vs. CC		PL. vs. PN		PN vs. CC		CC vs. PN and PL	
Cupriavidus	ORIG	+	ORIG	+	ORIG	-	ORIG	N/A
	V1-V3	++	V1-V3	++	V1-V3	--	V1-V3	--
	V3-V5	++	V3-V5	--	V3-V5	++	V3-V5	++
Methylobacterium	ORIG	+	ORIG	+	ORIG	-	ORIG	N/A
	V1-V3	++	V1-V3	--	V1-V3	--	V1-V3	++
	V3-V5	--	V3-V5	--	V3-V5	--	V3-V5	++
Schlegelella	ORIG	+	ORIG	+	ORIG	+	ORIG	N/A
	V1-V3	--	V1-V3	++	V1-V3	--	V1-V3	--
	V3-V5	++	V3-V5	--	V3-V5	++	V3-V5	++

[*] The results of the original univariate analysis of the V1-V3 dataset [15] are denoted as "ORIG", and the results of the present multivariate analyses of the V1-V3 and V3-V5 datasets are denoted as "V1-V3" and "V3-V5", respectively. The following notation is used to summarize results: "+" means that the genera was statistically significant in the prior analysis [15]; "-" means that the genera was not statistically significant in the prior analysis [15]; "++" means that that the genera was either often included in molecular signatures constructed by cross-validation or in the molecular signature constructed on the entire dataset; and "--" means that that the genera was neither often included in molecular signatures constructed by cross-validation nor in the molecular signature constructed on the entire dataset (see **Supplementary files 1 and 2**).

References

1 Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S. & Koutsoukos, X. D. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification. Part II: Analysis and Extensions. *Journal of Machine Learning Research* **11**, 235-284 (2010).

2 Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S. & Koutsoukos, X. D. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification. Part I: Algorithms and Empirical Evaluation. *Journal of Machine Learning Research* **11**, 171-234 (2010).

3 Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Machine Learning* **46**, 389-422 (2002).

4 Hollander, M. & Wolfe, D. *Nonparametric statistical methods*. Vol. 2nd (Wiley, 1999).

5 Statnikov, A., Aliferis, C. F., Hardin, D. P. & Guyon, I. *A Gentle Introduction to Support Vector Machines in Biomedicine, Volume 2: Case Studies & Benchmarks*. (World Scientific Publishing (in press), 2013).

6 Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537 (1999).

7 Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D. & Levy, S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics* **21**, 631-643 (2005).

8 Ding, C. & Peng, H. Minimum redundancy feature selection from microarray gene expression data. *J Bioinform.Comput.Biol.* **3**, 185-205 (2005).

9 Peng, H., Long, F. & Ding, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE transactions on pattern analysis and machine intelligence* **27**, 1226-1238, doi:10.1109/TPAMI.2005.159 (2005).

10 Diaz-Uriarte, R. & Alvarez de Andres, S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**, 3 (2006).

11 Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B(Statistical Methodology)* **67**, 301-320 (2005).

12 Bicciato, S., Luchini, A. & Di, B. C. PCA disjoint models for multiclass cancer analysis using gene expression data. *Bioinformatics* **19**, 571-578 (2003).

13 Zou, H., Hastie, T. & Tibshirani, R. Sparse principal component analysis. *Journal of Computational and Graphical Statistics* **15**, 265-286 (2006).

14 Ma, S. & Huang, J. Clustering threshold gradient descent regularization: with applications to microarray studies. *Bioinformatics* **23**, 466-472, doi:10.1093/bioinformatics/btl632 (2007).

15 Alekseyenko, A. V. *et al.* Population differentiation of the cutaneous microbiota in psoriasis. *Submitted* (2013).