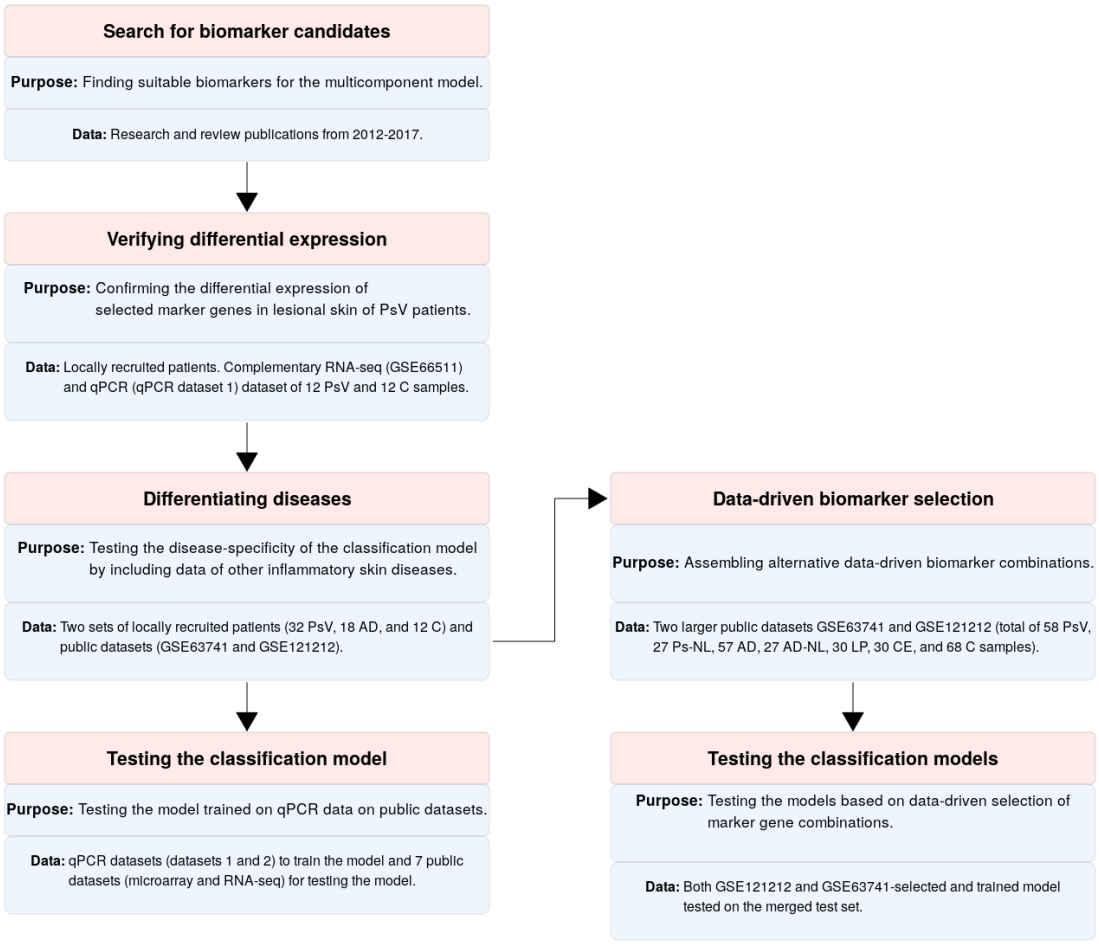

Fig. S1. Flowchart describing the study design. Abbreviations used in the flowchart: PsV – psoriasis vulgaris; qPCR - quantitative PCR; C – healthy control samples; AD – atopic dermatitis; CE – contact eczema; LP – lichen planus.

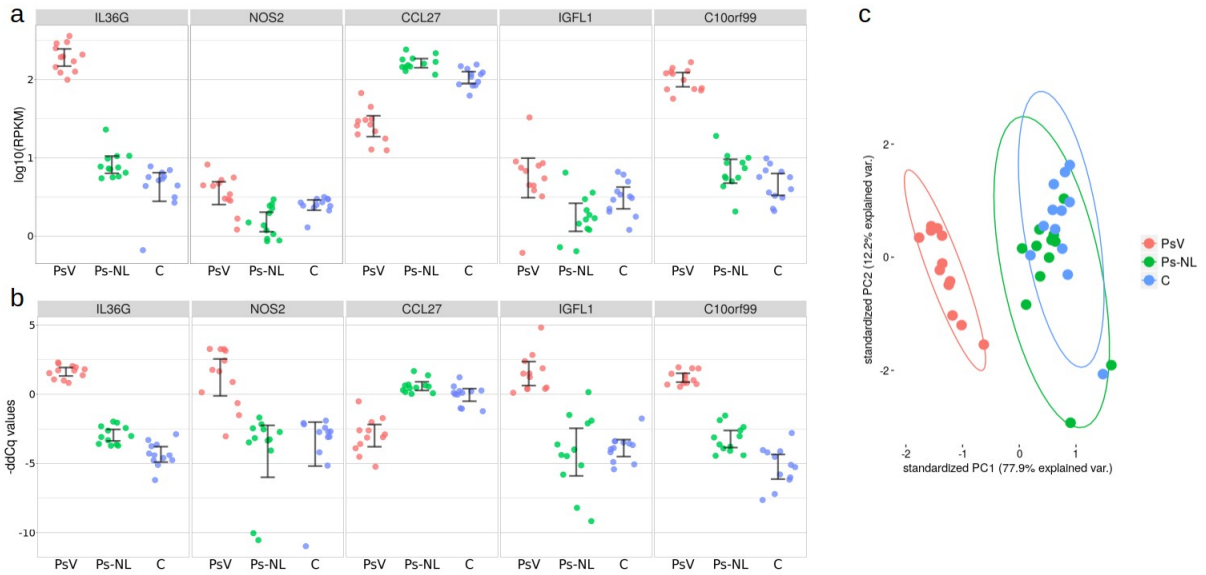

Fig. S2. Overview of expression levels of marker genes in psoriasis vulgaris lesions compared to non-lesional and healthy skin. (a) RNA-seq and (b) qPCR quantified expression levels of the selected marker genes in lesional psoriasis vulgaris skin (PsV), non-lesional skin of the same patients (Ps-NL) and the skin of healthy controls (C). Expression levels are presented as $\log_{10}$ of Reads Per Kilobase per Million mapped reads (RPKM) values (RNA-seq data) or -ddCq (ΔΔCq, quantitation cycle of the target gene adjusted for quantitation cycles of reference genes and calibration sample) values (qPCR data). The errorbars represent 95% confidence intervals. (c) The first two principle components of untransformed -ddCq values of qPCR derived gene expression data of the same samples and corresponding 95% confidence intervals.

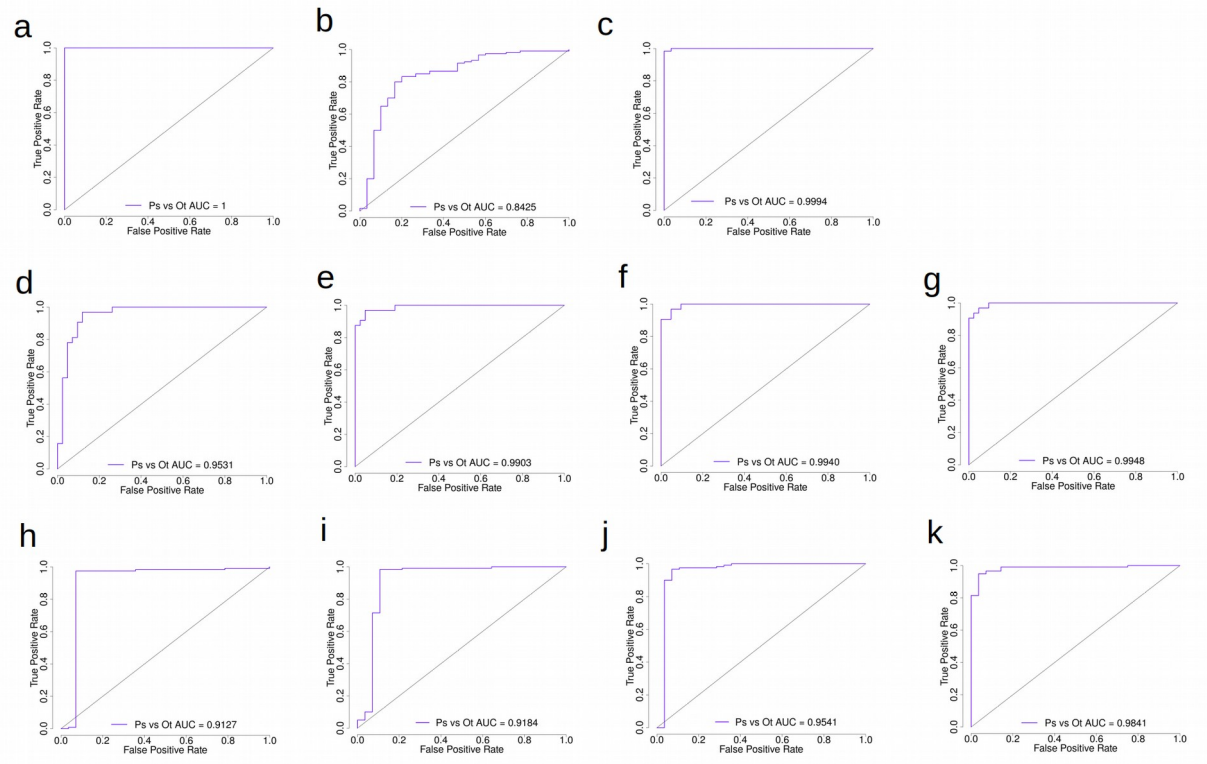

Fig. S3. Model performances on different datasets. Leave-one-out cross-validation (LOOCV) accuracies of (a) 1-gene (*IL36G*), (b) 2-gene (*NOS2* and *CCL27*), and (c) 4-gene (exluding *IGFL1*) model on the GSE63741 dataset. LOOCV accuracies of (d) 1-gene (*IL36G*), (e) 2-gene (*NOS2* and *CCL27*), (f) 4-gene (exluding *IGFL1*), and (g) 5-gene model on the merged qPCR datasets. LOOCV accuracies of (h) 1-gene (*IL36G*), (i) 2-gene (*NOS2* and *CCL27*), (j) 4-gene (exluding *IGFL1*), and (k) 5-gene model on the GSE121212 dataset.

a

		PsV	C	LP	AD	CE
Predicted	PsV	29	0	0	0	0
	Ot	1	30	30	30	30

b

$$Accuracy = \frac{N_{\text{true pos.}} + N_{\text{true neg.}}}{N_{\text{total pred.}}} =$$

$$= \frac{29 + 30 + 30 + 30 + 30}{150} = \frac{149}{150} = 0.993$$

c

		PsV_2	PsV_1	AD	Ps-NL	C
Predicted	PsV	19	11	1	0	0
	Ot	1	1	17	12	12

d

$$Accuracy = \frac{N_{\text{true pos.}} + N_{\text{true neg.}}}{N_{\text{total pred.}}} =$$

$$= \frac{19 + 11 + 17 + 12 + 12}{74} = \frac{71}{74} = 0.959$$

e

		PsV	Ps-NL	AD	AD-NL	C
Predicted	PsV	26	2	0	0	0
	Ot	2	25	27	27	38

f

$$Accuracy = \frac{N_{\text{true pos.}} + N_{\text{true neg.}}}{N_{\text{total pred.}}} =$$

$$= \frac{26 + 25 + 27 + 27 + 38}{147} = \frac{143}{147} = 0.973$$

Fig. S4. 4-gene model (*IL36G*, *CCL27*, *NOS2*, and *C10orf99*) performances on different datasets. (a) Confusion matrix and (b) the calculated accuracy of the 4-gene model leave-one-out cross-validation LOOCV on the GSE63741 dataset. (c) Confusion matrix and (d) the calculated accuracy of the 4-gene model LOOCV on the qPCR datasets. (e) Confusion matrix and (f) the calculated accuracy of the 4-gene model LOOCV on the GSE121212 dataset.

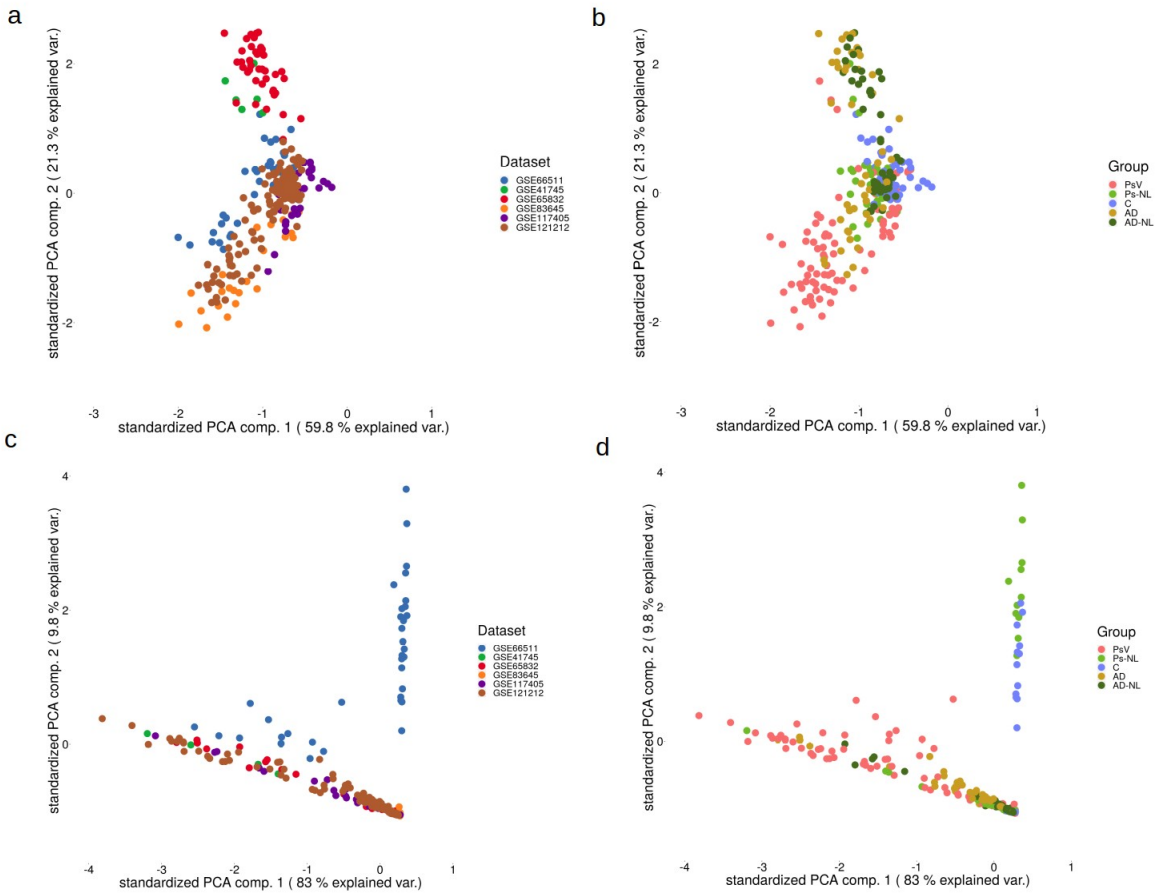

Fig. S5. Gene expression profiles of RNA-seq datasets. The first two principle components of Reads Per Kilobase per Million mapped reads (RPKM) values of gene annotations overlapping UCSC RefGene annotations for hg19 and GENCODE v26 annotations for hg38 with different colors representing different (a) datasets or (b) inflammatory skin diseases. These PCA plots represent the whole transcriptome expression patterns of the samples in the public datasets used in the analysis. RPKM values were normalized as z-scores prior to principal component calculations. (c) The first two principle components of RPKM values calculated based on the 4 marker genes (*IL36G*, *CCL27*, *NOS2*, and *C10orf99*) used for classification model construction with different colors representing different datasets. (d) The first two principle components of RPKM values of the 4 marker genes with different colors representing different inflammatory skin diseases.

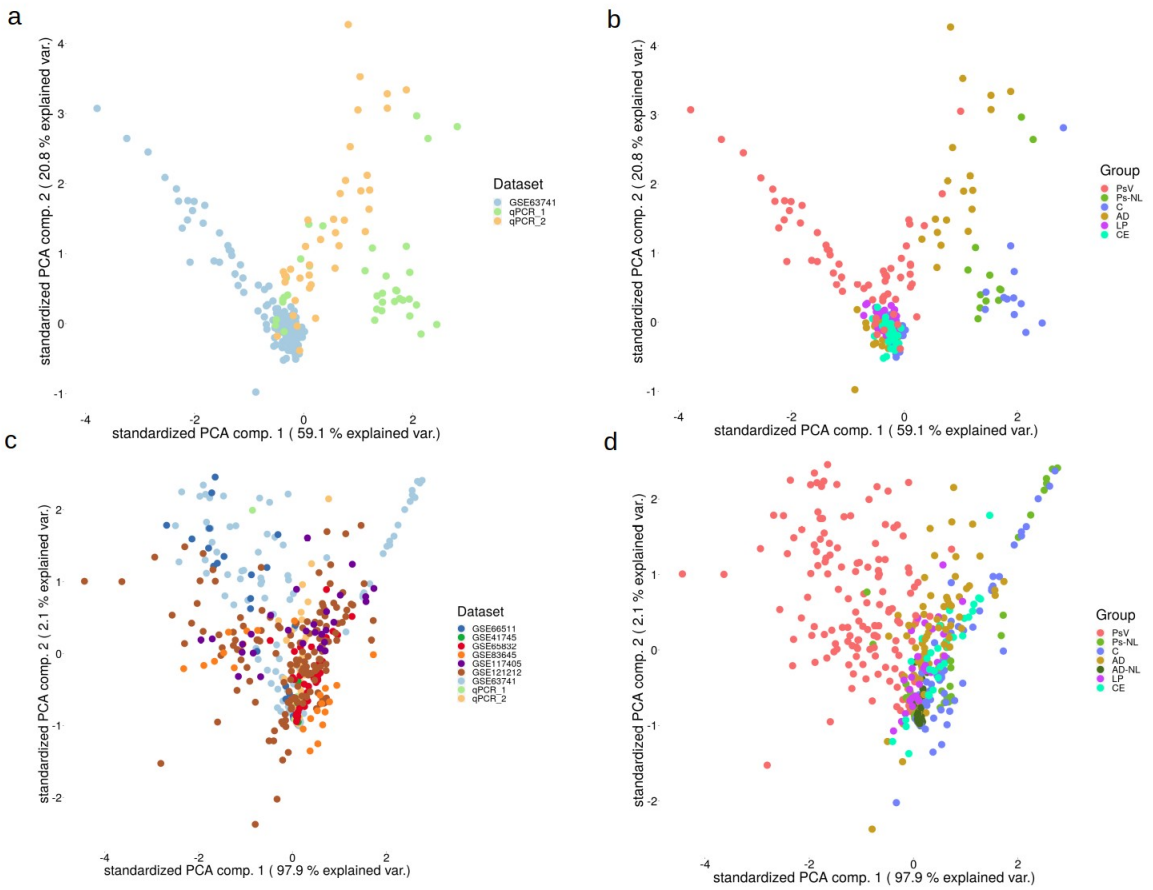

Fig. S6. Marker gene expression profiles in microarray, qPCR and transformed datasets. The first two principle components representing expression values of the 4 marker genes (*IL36G*, *CCL27*, *NOS2*, and *C10orf99*) used for classification model construction in the microarray (GSE63741) and qPCR datasets with different colors representing different (a) datasets or (b) inflammatory skin diseases. Gene expression values in the form of Cy5/Cy3 intensity ratios (microarray) or -ΔΔCq (qPCR) were normalized as z-scores prior to principal component calculations. -ΔΔCq values corresponded to quantitation cycle of the target gene adjusted for quantitation cycles of reference genes and calibration sample. (c) The first two principle components of transformed expression values of the 4 marker genes in all datasets with different colors representing different datasets. (d) The first two principle components of transformed expression values of the 4 marker genes in all datasets with different colors representing different inflammatory skin diseases. Expression values for the four marker genes were transformed into new features: *intercept*, *x*, *x2* and *x3* as in $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon_1$. The polynome line equation was obtained by coding the four marker genes as integers and using a linear model to fit a degree 3 polynomial line to connect expression values along the x-axis. The new features were further standardized as z-scores, where standard deviation (SD) and mean were calculated separately for each gene and expression quantification method group (RNA-seq, microarray or qPCR).

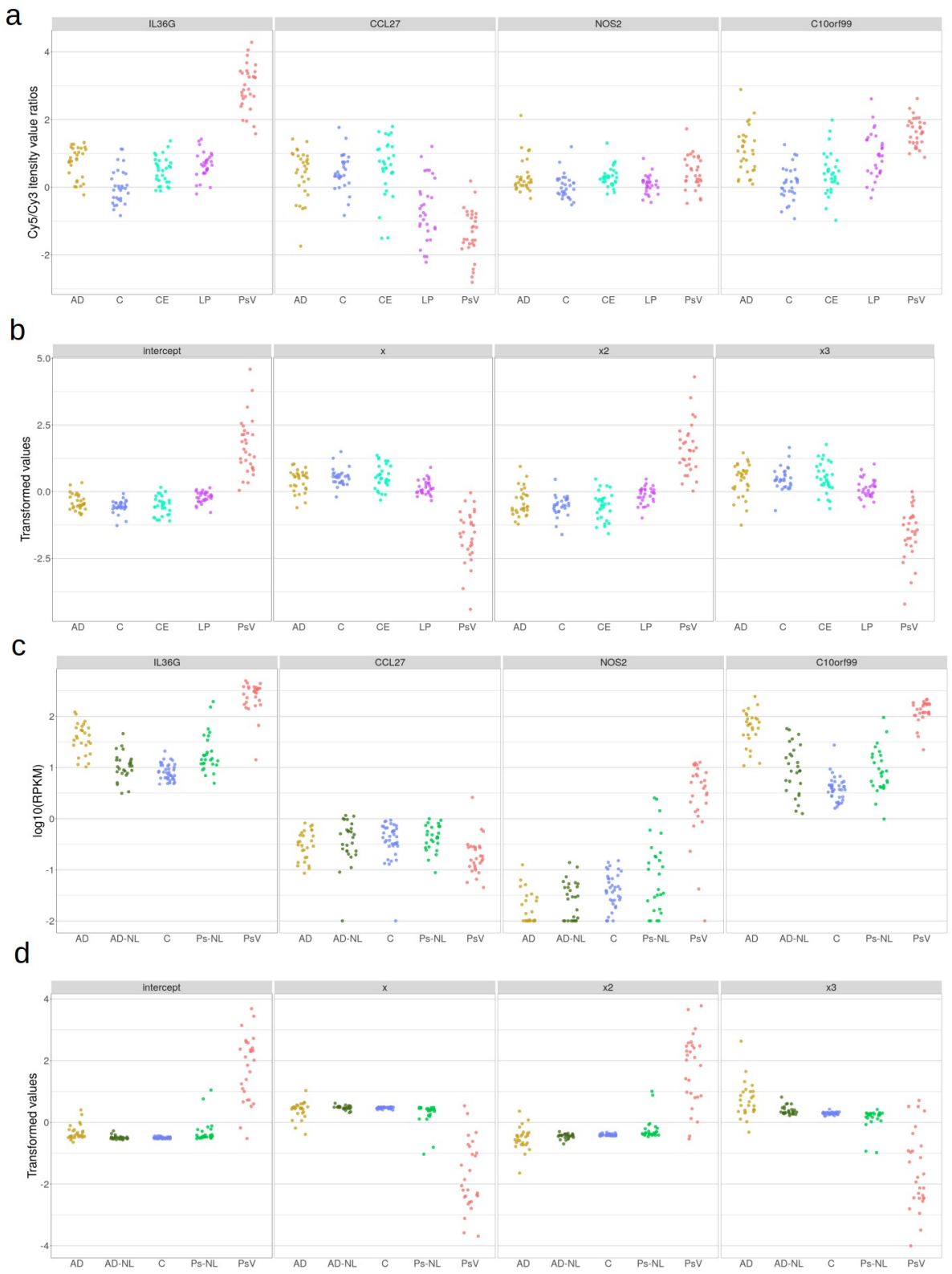

Fig. S7. Transformation of gene expression values in the GSE63741 and GSE121212 datasets. (a) Original expression values of the *IL36G*, *NOS2*, *CCL27*, and *C10orf99* marker genes in the GSE63741 dataset and (b) resulting transformed values. (c) Original expression values of the *IL36G*, *CCL27*, *NOS2*, and *C10orf99* marker genes in the GSE121212 dataset and (d) resulting transformed values.

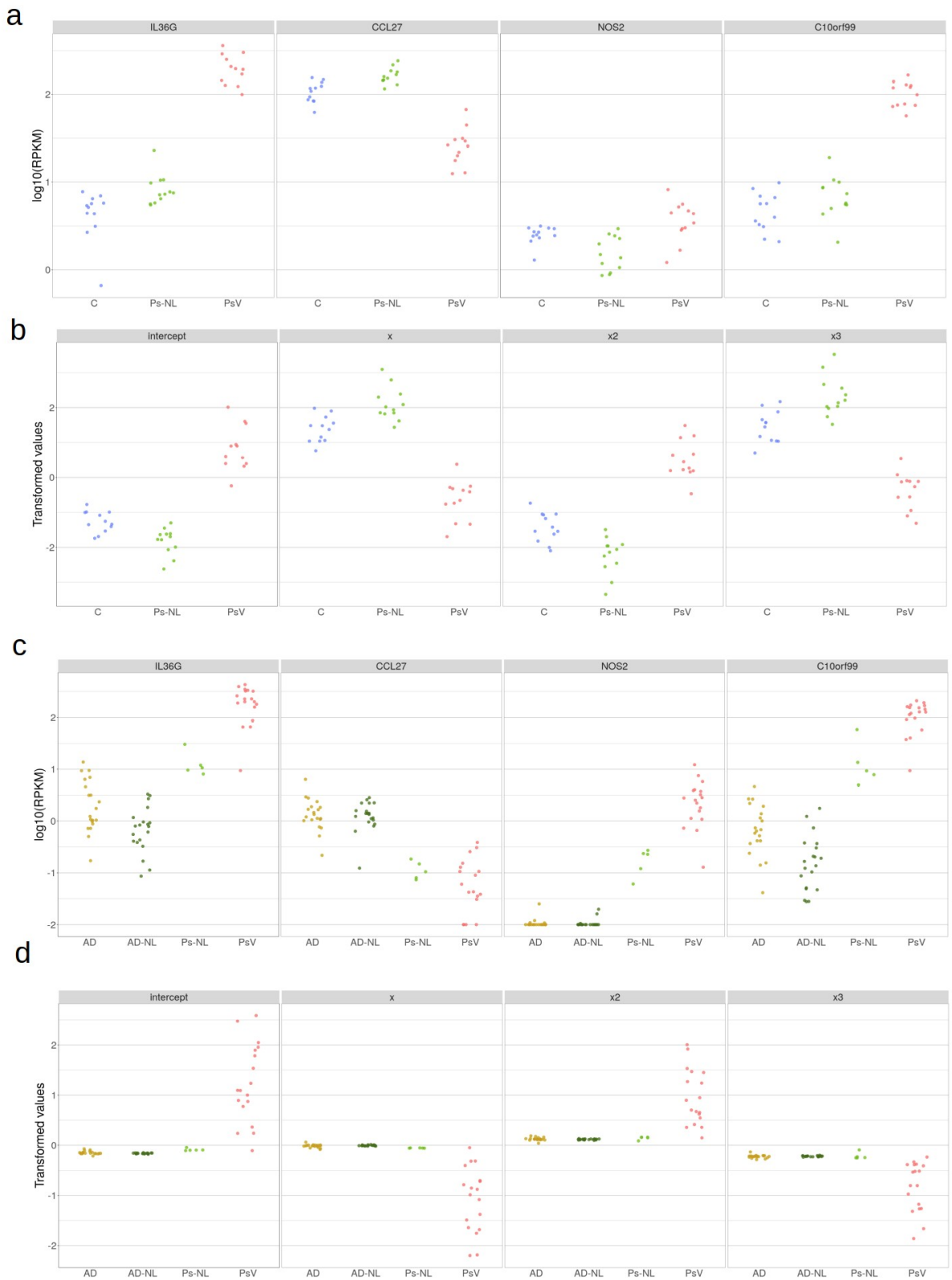

Fig. S8. Transformation of gene expression values in the GSE66511, GSE65832, and GSE83645 datasets. (a) Original expression values of the *IL36G*, *NOS2*, *CCL27*, and *C10orf99* marker genes in the GSE66511 dataset and (b) resulting transformed values. (c) Original expression values of the *IL36G*, *CCL27*, *NOS2*, and *C10orf99* marker genes in the GSE65832 (AD and AD-NL samples) and GSE83645 (Ps-NL and PsV samples) datasets and (d) resulting transformed values.

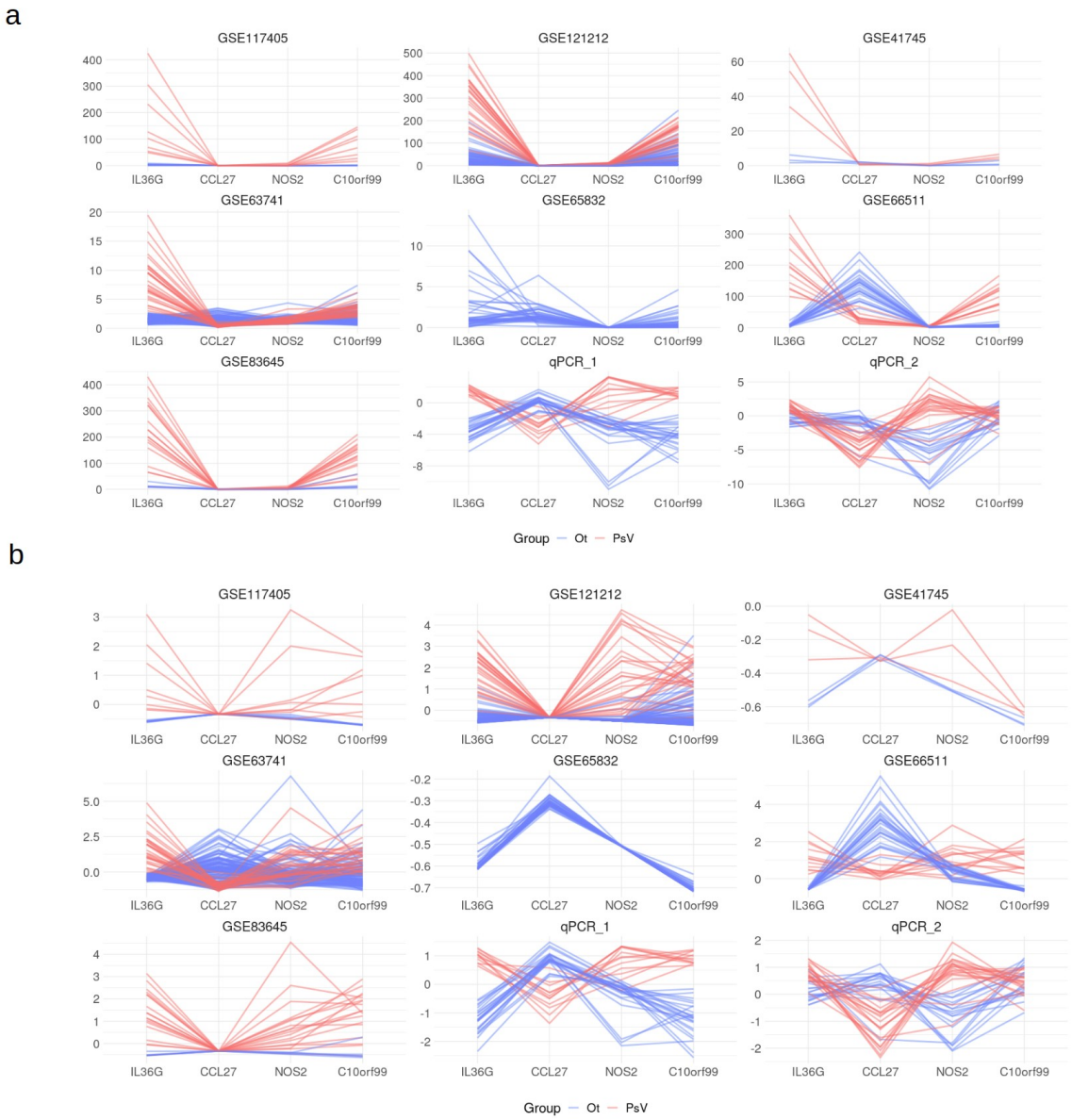

Fig. S9. Expression values of marker genes normalized but not transformed to new features. (a) Untransformed expression values for the four marker genes as Reads Per Kilobase of transcript per Million mapped reads (RPKM) values for RNA-seq datasets (GSE121212, GSE117405, GSE41745, GSE63741, and GSE65832), $2^{(Cy5/Cy3)}$ intensity value ratios for the GSE66511 microarray dataset or -ΔΔCq values (quantitation cycle of the target gene adjusted for quantitation cycles of reference genes and calibration sample) for qPCR datasets (qPCR_1 and qPCR_2). Samples are colored on the graphs as lesional psoriasis vulgaris skin samples (PsV - red) or other samples (Ot - blue). The latter consists of non-lesional skin samples of psoriasis vulgaris patients, atopic dermatitis, non-lesional atopic dermatitis, contact eczema, lichen planus and healthy control skin samples. (b) Expression values for the four marker genes standardized as z-scores, where standard deviation (SD) and mean were calculated separately for each gene and expression quantification method group (RNA-seq, microarray or qPCR).

a

		PsV	AD	Ps-NL	C
Predicted	PsV	29	0	0	0
	Ot	3	18	12	12

$$Accuracy = \frac{N_{\text{true pos.}} + N_{\text{true neg.}}}{N_{\text{total pred.}}} = \frac{29 + 18 + 12 + 12}{74} = \frac{71}{74} = 0.959$$

b

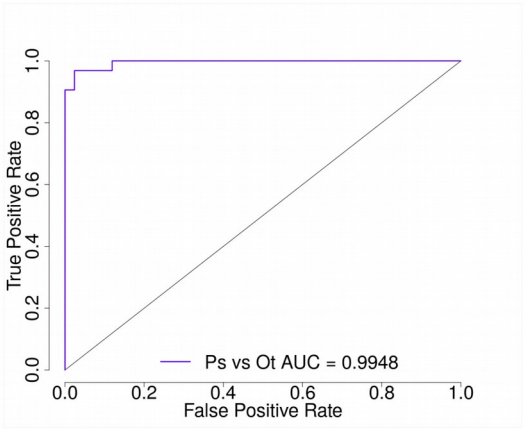

Fig. S10. SVM classification model leave-one-out cross-validation (LOOCV) on transformed datasets. (a) Predictions resulting LOOCV of the support vector machine (SVM) classification model based on four marker genes (*IL36G*, *NOS2*, *CCL27*, and *C10orf99*) using the transformed qPCR datasets of 32 psoriasis (PsV), 20 atopic dermatitis (AD), 12 non-lesional psoriasis (Ps-NL), and 12 healthy control (C) skin samples. (b) Receiver operating characteristic (ROC) plot and corresponding area under the curve (AUC) value based on the prediction scores obtained from LOOCV of the SVM classification model.

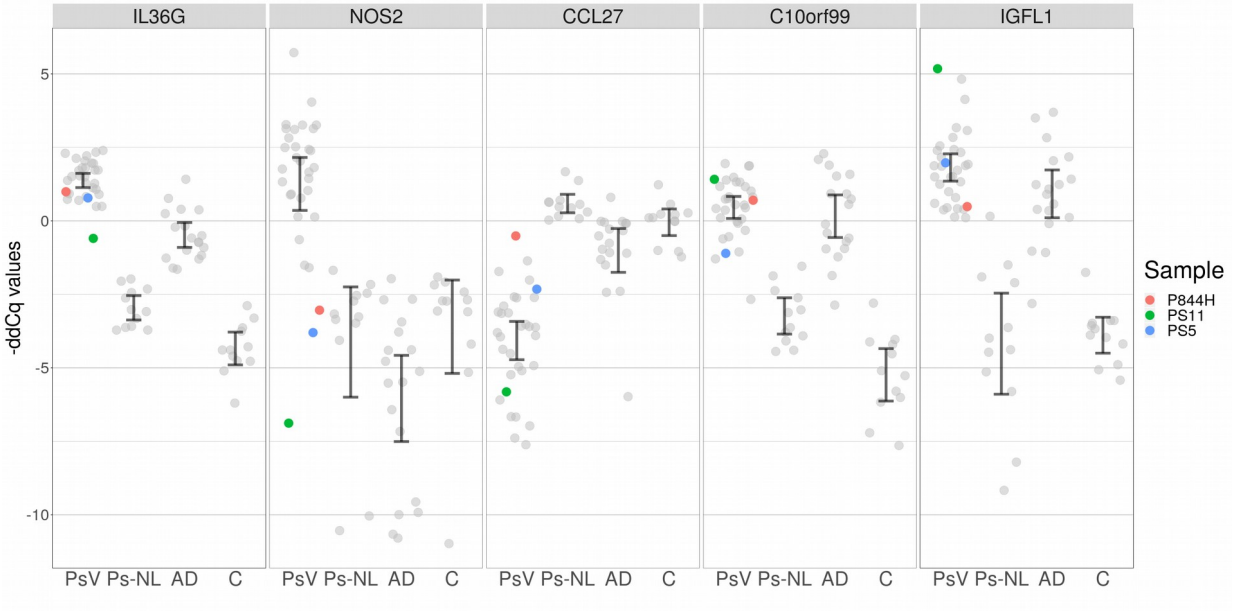

Fig. S11. Expression levels of marker genes in specific samples. Gene expression levels quantified as -ΔΔCq (quantitation cycle of the target gene adjusted for quantitation cycles of reference genes and calibration sample) values of the 4 marker genes for 32 psoriasis vulgaris (PsV), 18 atopic dermatitis (AD), 12 non-lesional psoriasis (Ps-NL) and 12 healthy control (C) skin samples. The 3 samples (P844, PS5, and PS11) resulting in incorrect predictions during leave-one-out cross-validation (LOOCV) of the 4-gene model on the test set of transformed data are colored differently from the samples predicted correctly, highlighting the differences in the expression levels of these samples. 95% confidence intervals are displayed as errorbars.

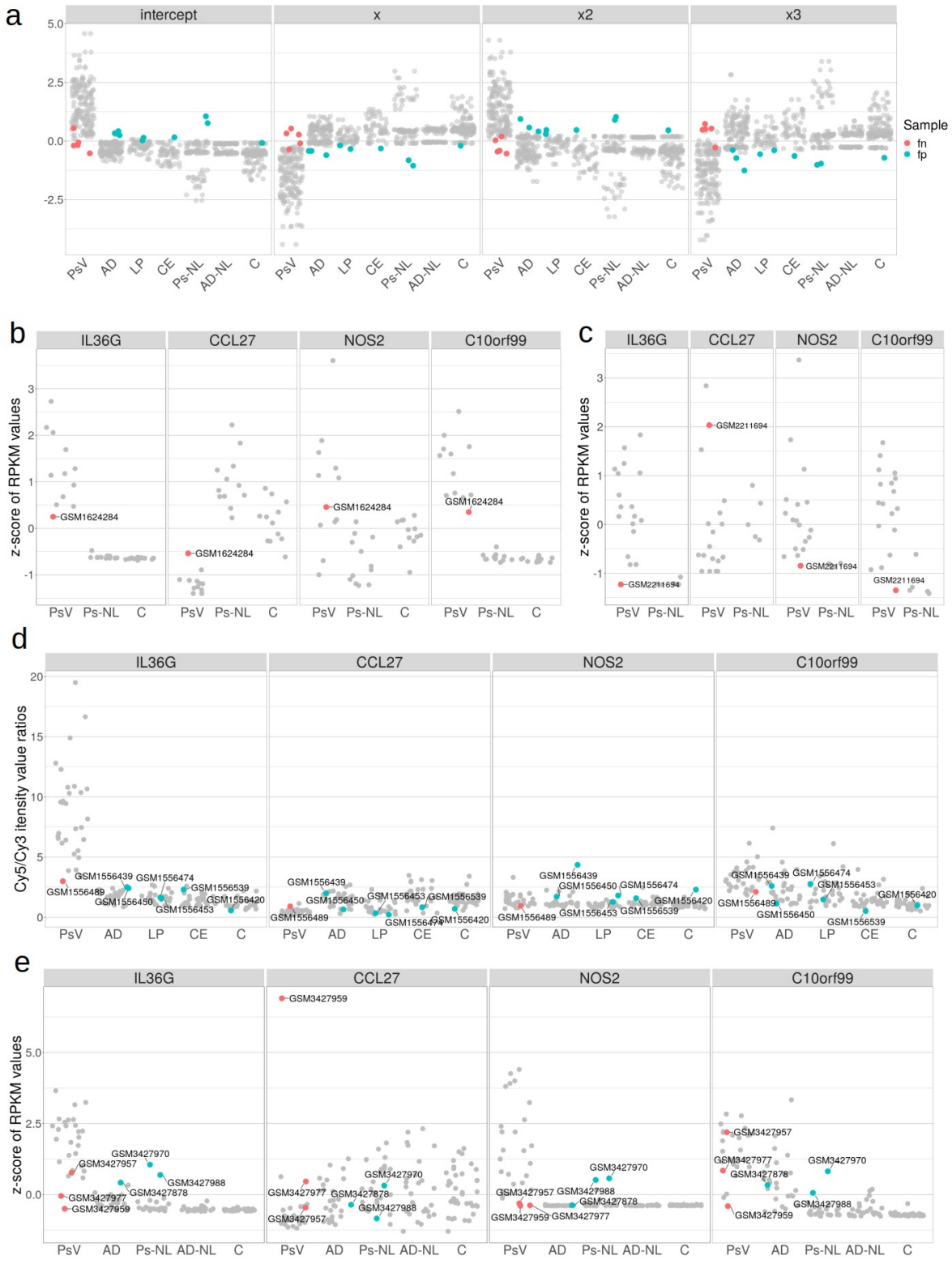

Fig. S12. Expression levels of marker genes in test set samples with prediction errors. (a) Values of the transformed features representing gene expression values in all test set samples. False negative (fn) and false positive (fp) predictions are highlighted with different colors. Marker gene expression values in the (b) GSE6651, (c) GSE83645, (d) GSE63741, and (e) GSE121212 datasets with fn and fp predictions highlighted.

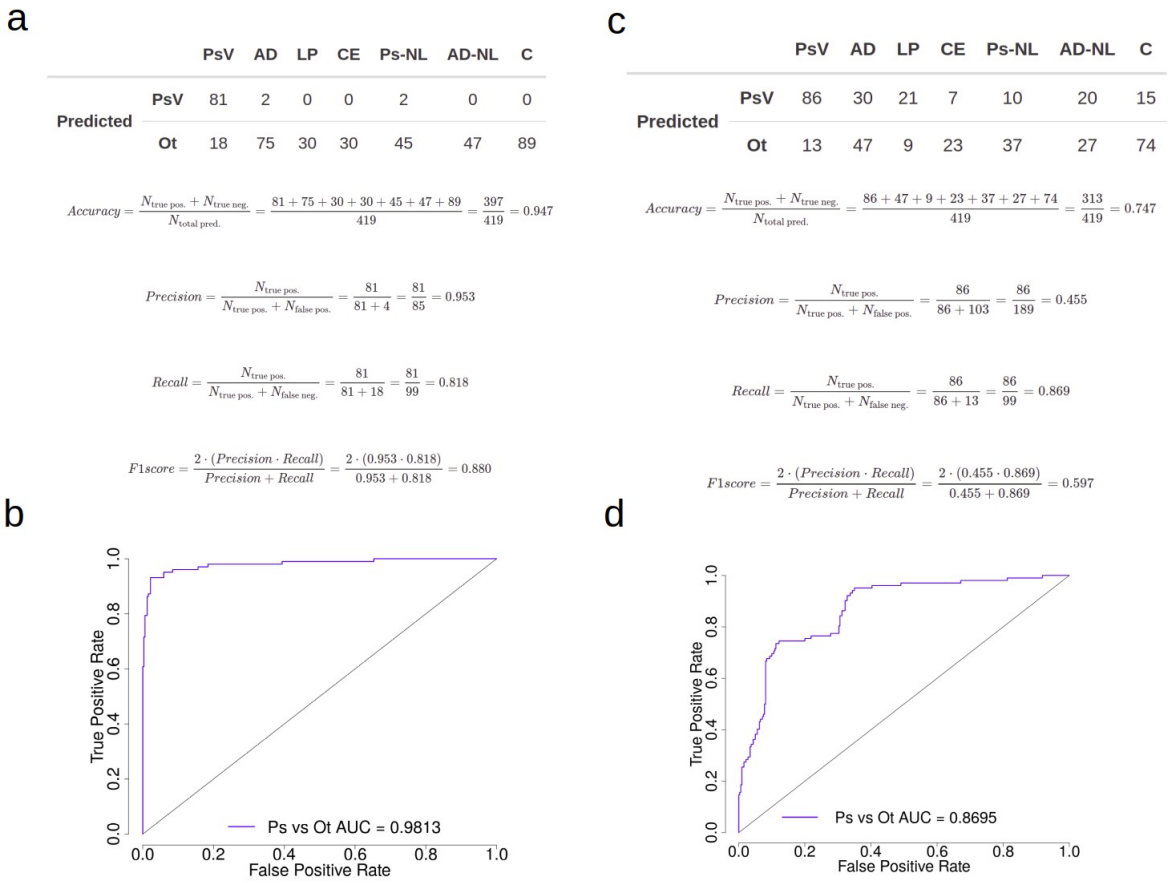

Fig. S13. SVM classification model performance on transformed datasets using alternative gene sets. All models were trained on transformed qPCR datasets and tested on 6 transformed RNA-seq and 1 transformed microarray datasets of 71 PsV, 20 Ps-NL, 50 AD, 20 non-lesional atopic dermatitis (AD-NL), 30 lichen planus (LP), 30 contact eczema (CE), and 51 C skin samples. (a) Predictions of the SVM classification model based solely on *IL36G* as the marker gene and resulting performance metrices. (b) Receiver operating characteristic (ROC) plot and corresponding area under the curve (AUC) value based on the prediction scores obtained from testing the *IL36G* model trained on qPCR data on the assembled test set. (c) Predictions of the SVM classification model based on *NOS2* and *CCL27* as the marker genes and resulting performance metrices. (d) Receiver operating characteristic (ROC) plot and corresponding area under the curve (AUC) value based on the prediction scores obtained from testing the *NOS2* and *CCL27* model trained on qPCR data on the assembled test set.

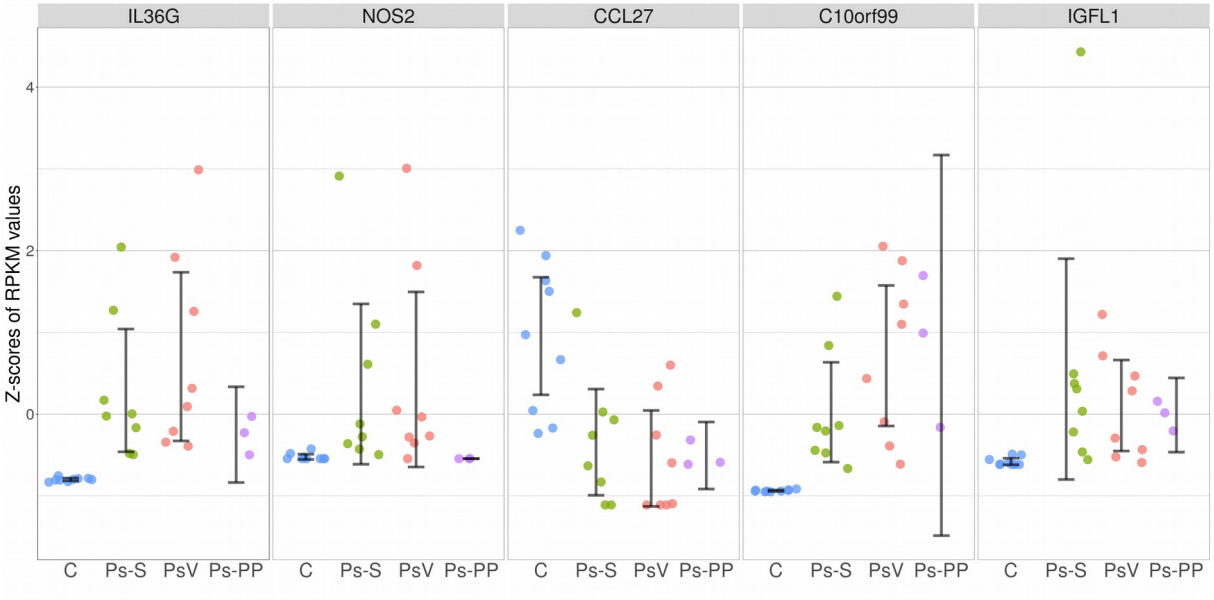

Fig. S14. Expression levels of marker genes in the GSE117405 dataset. Gene expression levels quantified as z-scored Reads Per Kilobase per Million mapped reads (RPKM) values of the 4 marker genes for 8 psoriasis vulgaris (PsV), 3 plaque-type palmoplantar psoriasis (Ps-PP), 8 scalp psoriasis (Ps-S) and 9 healthy control (C) skin samples in the GSE117405 dataset. 95% confidence intervals are displayed as errorbars.

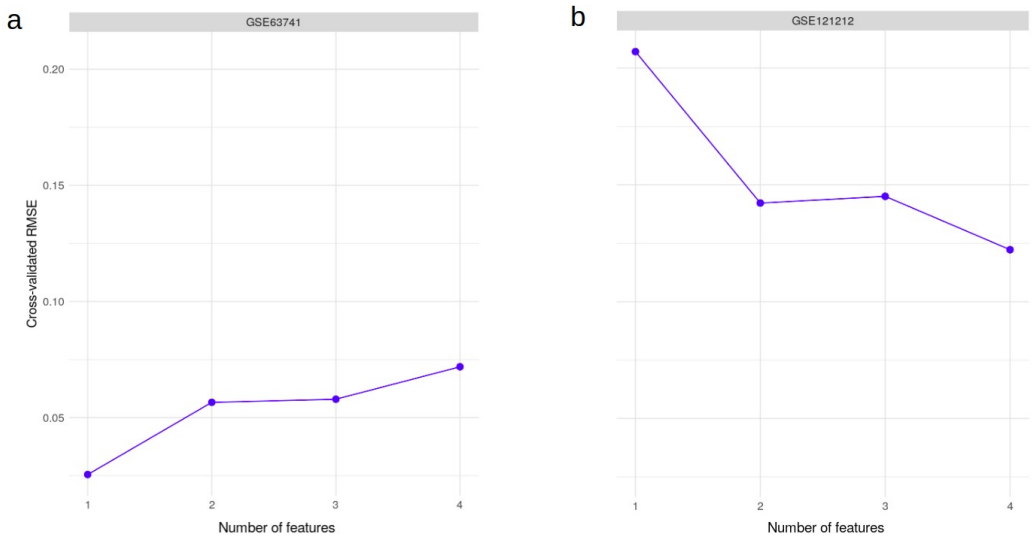

Fig. S15. Prediction error of gene (feature) sets obtained by Recursive Feature Elimination (RFE). Prediction errors quantified as Root Mean Squared Error (RMSE) resulting from RFE implemented with Random Forest function and 10-fold cross-validation on the (a) GSE63741 and (b) GSE121212 datasets.

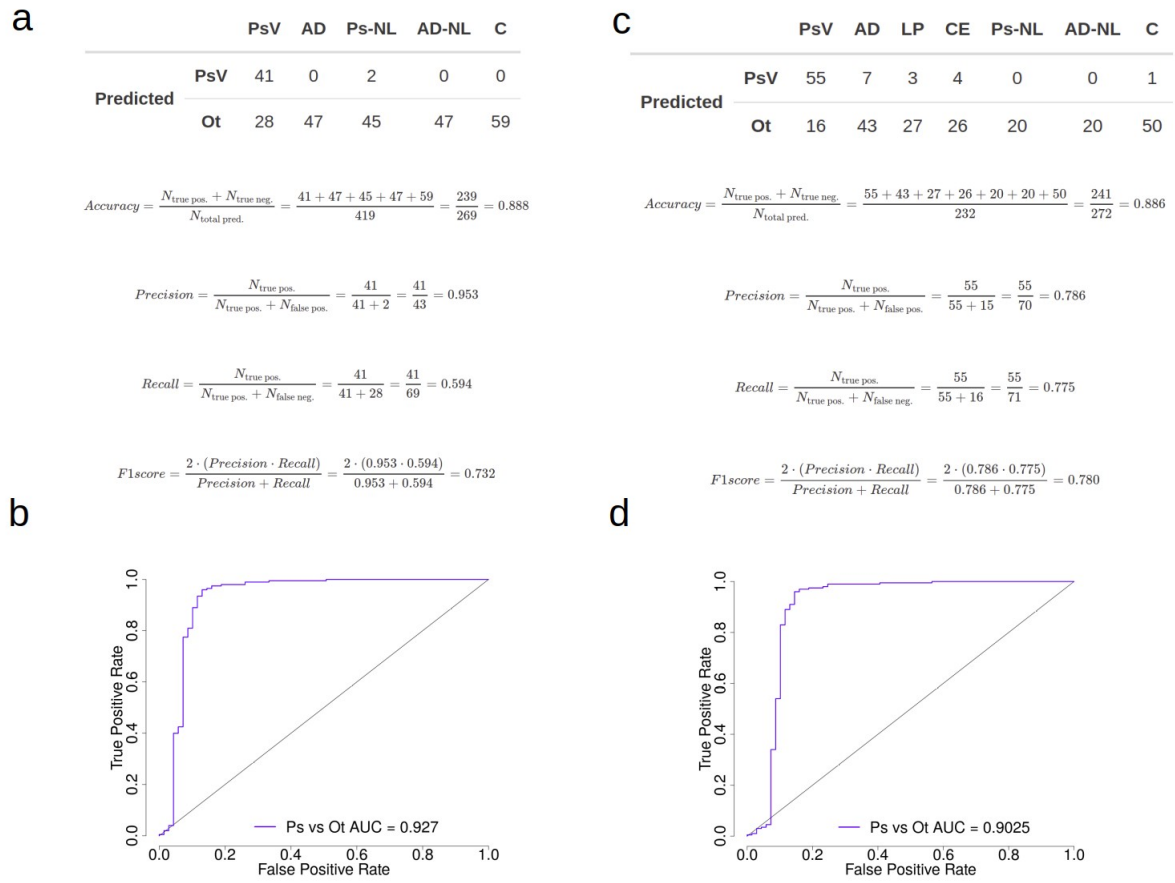

Fig. S16. SVM classification model performance on transformed datasets using gene sets resulting from Recursive Feature Elimination (RFE). The model based on RFE-selected genes (*IL36G*, *CRABP2*, *S100A7A*, and *IL36RN)* from the GSE63741 dataset was trained of the transformed GSE63741 dataset and tested on the other transformed public datasets. (a) Confusion matrix and derived classification performance metrices resulting from testing the GSE63741-derived model. (b) Receiver operating characteristic (ROC) plot and corresponding area under the curve (AUC) value resulting from testing the GSE63741-derived model. The model based on RFE-selected genes (*SPRR2A*, *PRELP*, *ARG1* and *KYNU)* from the GSE121212 dataset was trained of the transformed GSE121212 dataset and tested on the other transformed public datasets. (c) Confusion matrix and derived classification performance metrices resulting from testing the GSE121212-derived model. (d) Receiver operating characteristic (ROC) plot and corresponding area under the curve (AUC) value resulting from testing the GSE121212-derived model.

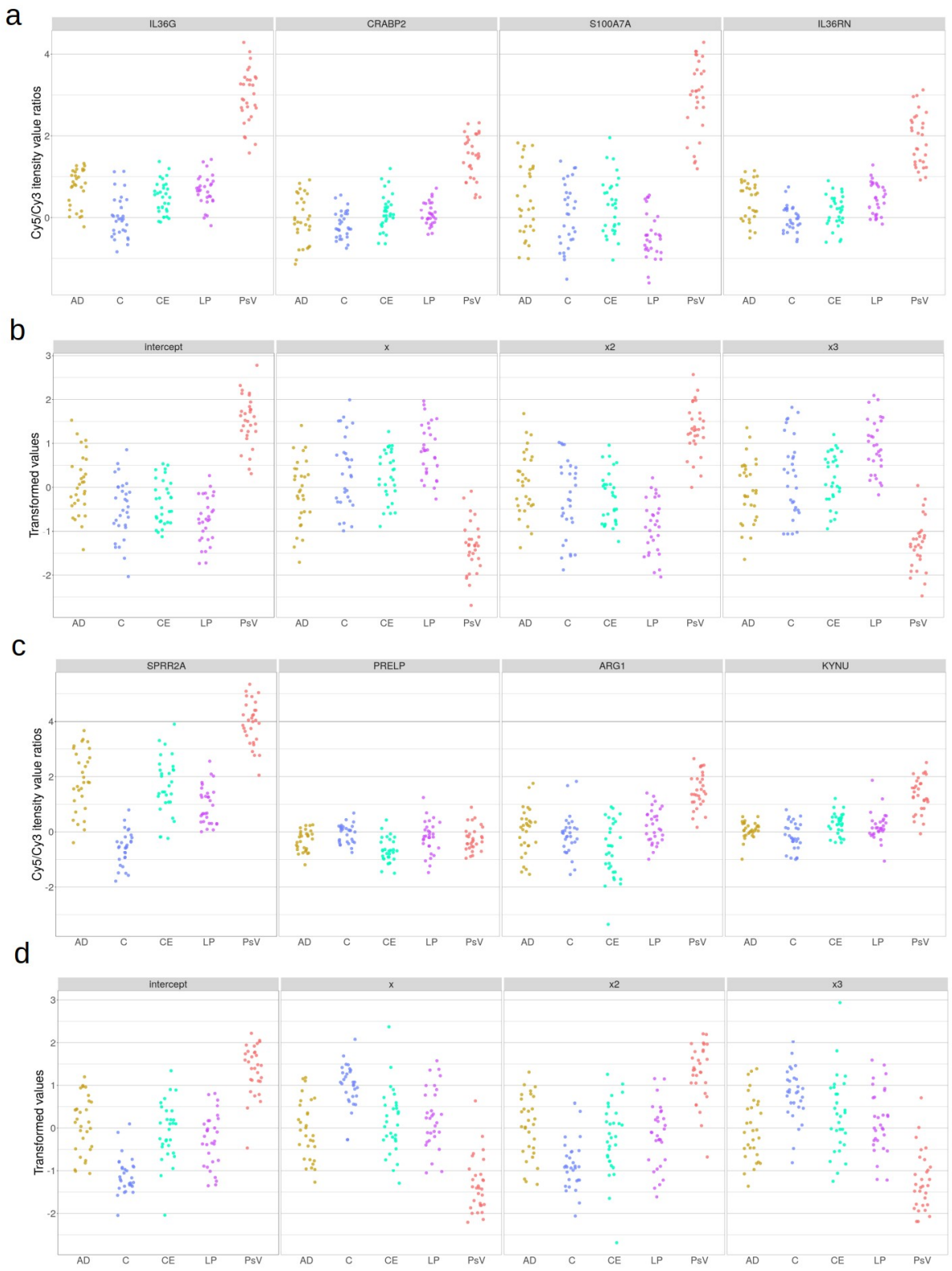

Fig. S17. Transformation of gene expression values in the GSE63741 dataset (a) Original expression values of the *IL36G*, *CRABP2*, *S100A7A*, and *IL36RN* marker genes in the GSE63741 dataset and (b) resulting transformed values. (c) Original expression values of the *SPRR2A*, *PRELP*, *ARG1* and *KYNU* marker genes in the GSE63741 dataset and (d) resulting transformed values

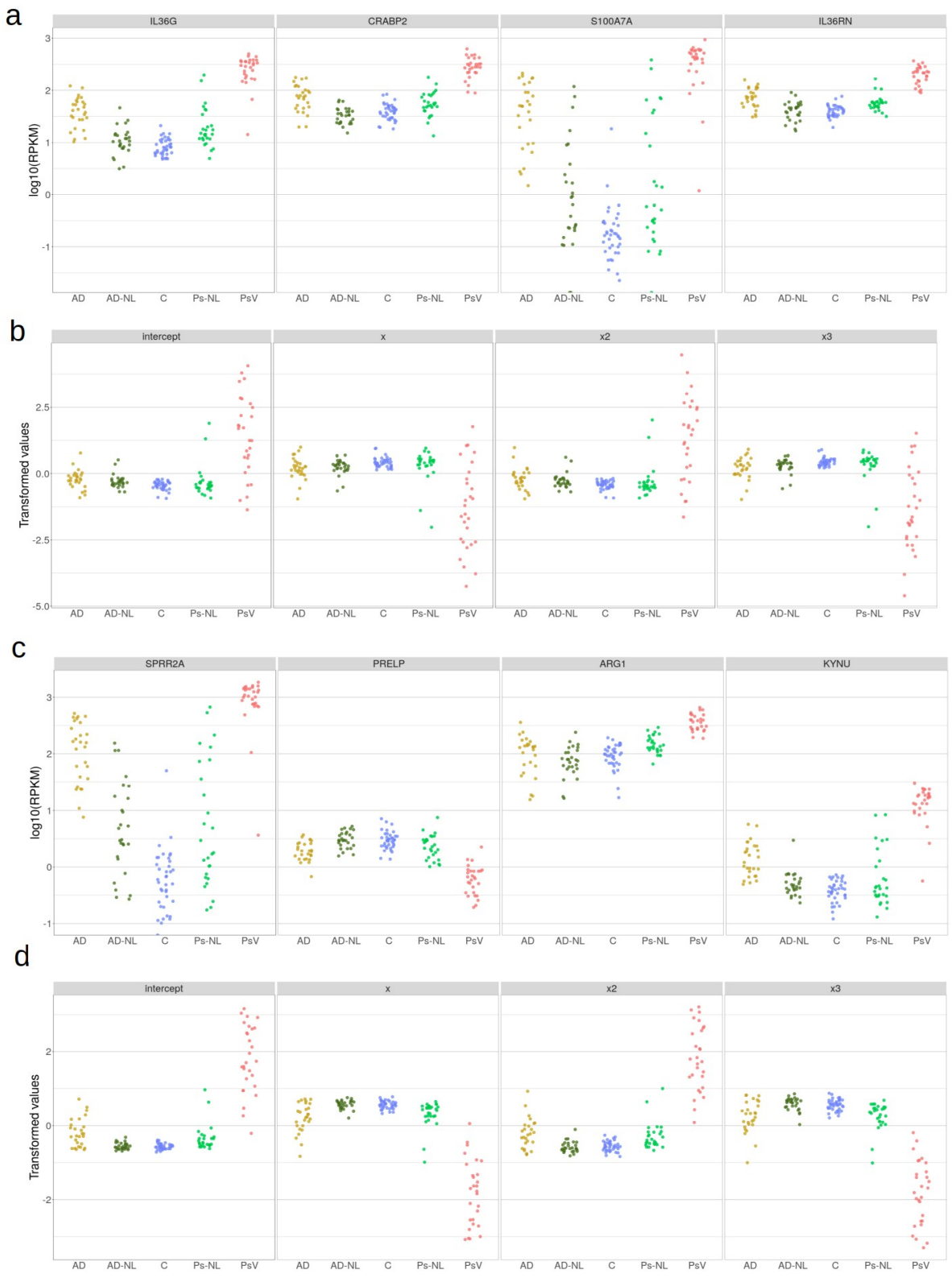

Fig. S18. Transformation of gene expression values in the GSE121212 dataset (a) Original expression values of the *IL36G*, *CRABP2*, *S100A7A*, and *IL36RN* marker genes in the GSE121212 dataset and (b) resulting transformed values. (c) Original expression values of the *SPRR2A*, *PRELP*, *ARG1* and *KYNU* marker genes in the GSE121212 dataset and (d) resulting transformed values

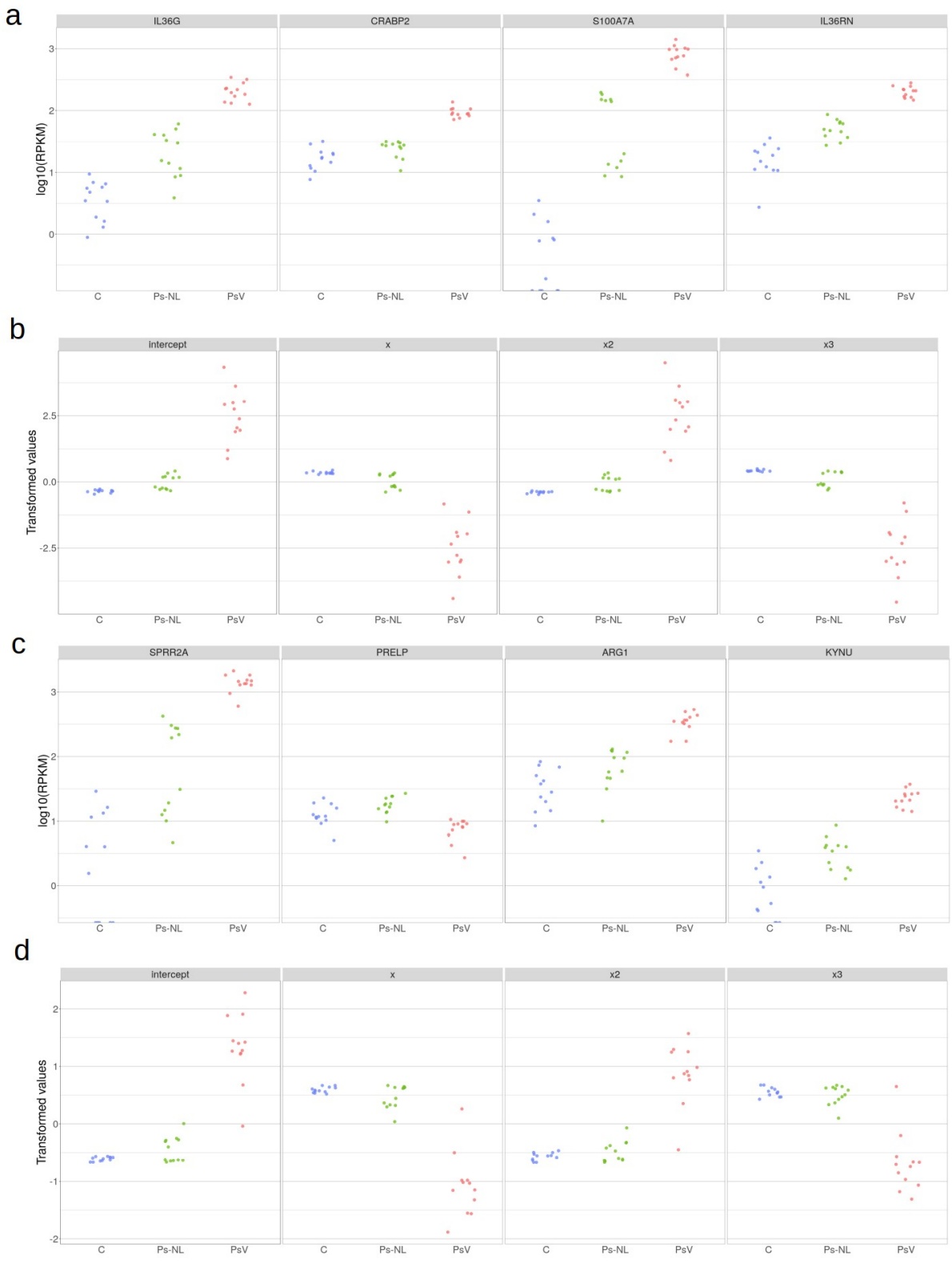

Fig. S19. Transformation of gene expression values in the GSE66511 dataset (a) Original expression values of the *IL36G*, *CRABP2*, *S100A7A*, and *IL36RN* marker genes in the GSE66511 dataset and (b) resulting transformed values. (c) Original expression values of the *SPRR2A*, *PRELP*, *ARG1* and *KYNU* marker genes in the GSE66511 dataset and (d) resulting transformed values

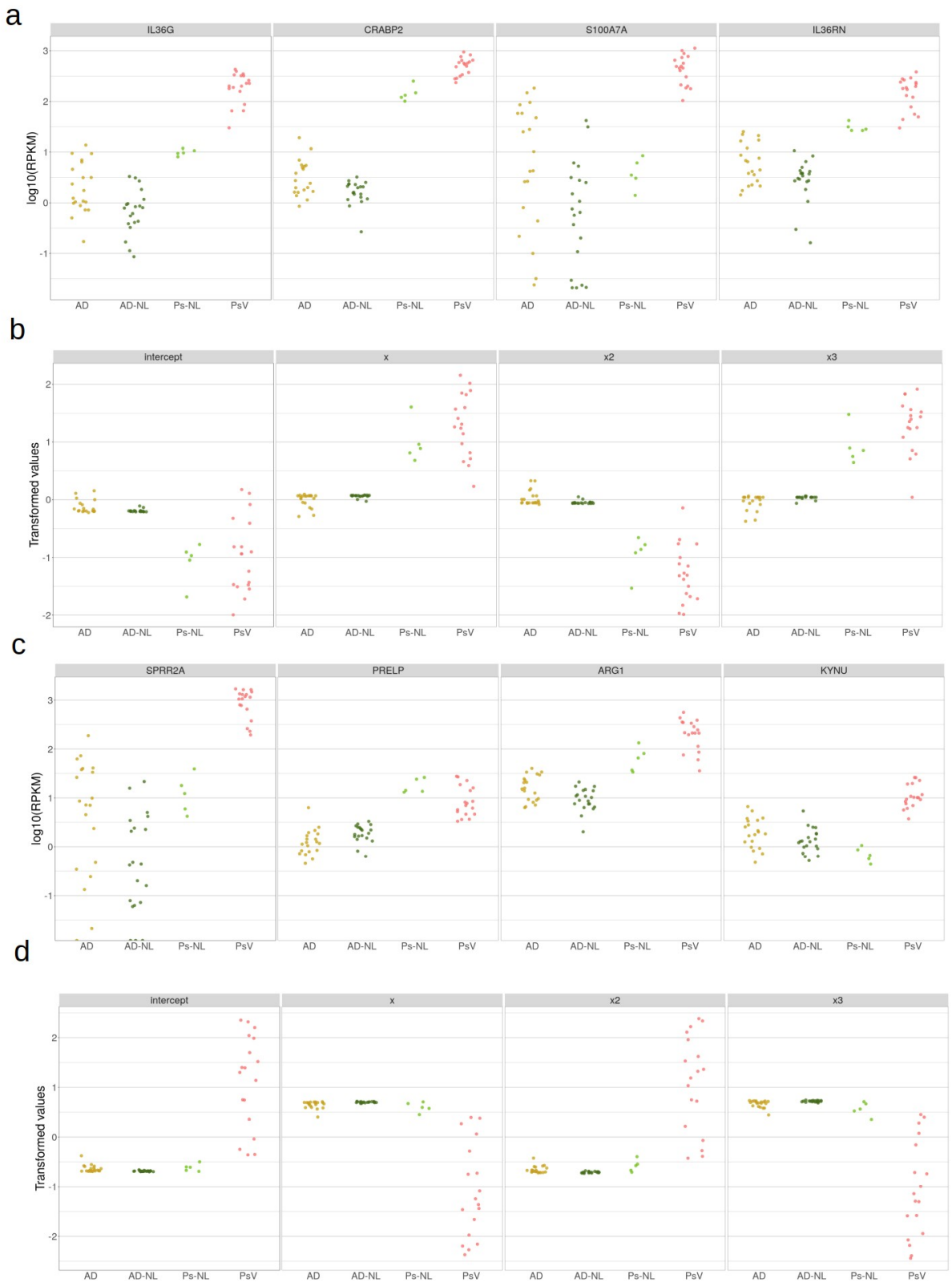

Fig. S20. Transformation of gene expression values in the GSE65832 (AD and AD-NL samples) and GSE83645 (Ps-NL and PsV samples) datasets (a) Original expression values of the *IL36G*, *CRABP2*, *S100A7A*, and *IL36RN* marker genes in the and (b) resulting transformed values. (c) Original expression values of the *SPRR2A*, *PRELP*, *ARG1* and *KYNU* marker genes and (d) resulting transformed values

Table SII. *Genes selected for constructing the classification model.*

Gene Symbol	Gene Name	Publication
IGFL1	IGF Like Family Member 1	Guo, P. *et al.* (2014)
C10orf99	Chromosome 10 Open Reading Frame 99	
IL36G	Interleukin 36, Gamma	D'Erme, A. M. *et al.* (2015)
NOS2	Nitric Oxide Synthase 2	Quaranta, M. *et al.* (2014)
CCL27	C-C Motif Chemokine Ligand 27	

Table SIII. *Spearman's correlation coefficients between measured gene expression values and known phenotypes of the recruited psoriasis vulgaris (PsV) and atopic dermatitis (AD) patients. Sex has been coded as "0" for males and "1" for females – in this case, positive correlation coefficient values should be interpreted as higher expression levels in females and negative correlation coefficient values as higher expression levels in males. Nail involvement ("Nails" in the table), diagnosis of psoriatic arthritis ("PsA" in the table), presence of atopic dermatitis in the family ("Family" in the table) and diagnosis of asthma were codes as binary traits ("1" if trait exists and "0" if not) – in this case, positive correlation coefficient values should be interpreted as higher expression levels if the trait exists and negative correlation coefficient values as lower expression levels when the trait exists. P-values are presented only in the case of significant correlations ($p < 0.05$). Traits with continuous values (age, duration and PASI/EASI) were not re-coded. Patients from both qPCR datasets were included in the correlation analysis ($N_{PsV} = 32$; $N_{AD} = 18$).*

	PsV						**AD**				
	C10orf99	**CCL27**	**IGFL1**	**IL36G**	**NOS2**		**C10orf99**	**CCL27**	**IGFL1**	**IL36G**	**NOS2**
Sex	-0.12	0	-0.02	-0.08	0.31		0.11	0.16	-0.20	0.11	0.30
Age	0.09	0.19	0.28	-0.19	-0.27		-0.23	0.24	0.09	0.03	-0.06
Duration	-0.01	-0.22	0.19	0	-0.16		0.05	0	-0.16	0.25	-0.09
PASI	0.23	-0.04	-0.41	0.45	0.15	***EASI***	-0.30	0.32	0.45	0.06	-0.40
Nails	-0.06	-0.35	-0.08	0.27	0.40	***Family***	0.25	0.11	0.23	**0.51 (0.029)**	-0.01
PsA	0.26	0.24	0.19	0.21	-0.16	***Asthma***	0.18	-0.03	0.18	0.21	0.08

Table SIV. *Samples resulting in incorrect predictions during leave-one-out cross-validation (LOOCV) or validation of the 4-gene model. The column "Dataset" specifies whether the predictions were obtained during LOOCV (Train) or validation (Validation) step. The accession number of the GEO dataset and accession number of the sample are listed in the "GEO" and "GSM" columns, respectively. GEO and GSM accession numbers are not available for the qPCR datasets and dataset or sample name is used instead. The type of error – false negative (fn) or false positive (fp) with psoriasis vulgaris as the positive outcome and other group as the negative outcome - is listed in the "Error" column. The "Actual" column specifies the actual disease group of the sample as follows: PsV for psoriasis vulgaris, C for healthy control, AD for atopic dermatitis and LP for lichen planus. It should be noted that sample GSM1624284 from GEO dataset GSE66511 and sample P844H from qPCR dataset 1 correspond to the same biopsy, but the method of gene expression quantification is different. GSE66511 dataset was generated by RNA-seq and qPCR dataset 1 by qPCR.*

Dataset	GEO	GSM	Error	Actual
Test	GSE66511	GSM1624284	fn	PsV
Test	GSE83645	GSM2211694	fn	PsV
Test	GSE121212	GSM3427988	fp	Ps-NL
Test	GSE121212	GSM3427977	fn	PsV
Test	GSE121212	GSM3427970	fp	Ps-NL
Test	GSE121212	GSM3427959	fn	PsV
Test	GSE121212	GSM3427957	fn	PsV
Test	GSE121212	GSM3427878	fp	AD
Test	GSE63741	GSM1556489	fn	PsV
Test	GSE63741	GSM1556420	fp	C
Test	GSE63741	GSM1556439	fp	AD
Test	GSE63741	GSM1556539	fp	CE
Test	GSE63741	GSM1556450	fp	AD
Test	GSE63741	GSM1556453	fp	LP
Test	GSE63741	GSM1556474	fp	LP
Train	qPCR dataset 1	P844H	fn	PsV
Train	qPCR dataset 2	PS5	fn	PsV
Train	qPCR dataset 2	PS11	fn	PsV