UNIVERSITÀ DI TRENTO

Project on

Exploring Transcriptomics data to Understand Human Skin Medical Condition, Psoriasis

Course: Network Data Analysis

University of Trento

Submitted to: Prof. Dr. Lauria Mario

Submitted by: Surya Hembrom

Matriculation Number: 229578

M.Sc. Quantitative and Computational Biology

Date: 25.7.22

1. Introduction

Psoriasis is a persistent medical condition affecting the human skin leading to disability and body disfeaturing, till now incurable and cause adverse impact on the quality of life. Those affected are predisposed to developing psoriatic arthritis and certain cardiovascular diseases later in life (WHO, 2016). Around 7.5 million adults in U.S. suffer from this skin condition (Armstrong et al., 2021) and approximately 2% people are affected worldwide (Rendon & Schäkel, 2019). The pathogenesis of disease involves the IL-23/Th17 axis that leads to inflammation and autoimmune response within the body. The inherited and acquired immune responses in the cutaneous layers of skin activate the autoimmune responses and autoinflammation with main effects and symptoms observed on the skin epidermis, constituting the keratinocytes. The various cells of the skin dermis i.e., inherited and acquired immune cells and vasculature interplay with the keratinocytes to induce inflammation and are severely affected by the complex disease pathogenesis (Rendon & Schäkel, 2019).

In this project, I aimed to investigate the genes concerning Psoriasis by analysis of high throughput RNA-Seq transcriptome data from the study by (Li et al., 2014). Their study involved RNA-Seq data generated from skin of normal individuals and psoriatic patients respectively. In this study their data was examined for standard gene expression analyses. I tried to apply various machine learning algorithms to predict important genes that were filtered based on standard statistical tests. The functional enrichment analysis was conducted for highly confident genes related to psoriasis. This study tries to understand the fine scale genomic differences between the non-disease and disease conditions and provides insight into models which might be employed to predict and identify the early prognosis of disease development in individuals who might show little symptoms at an early age and time-period.

2. Materials and Methods

2.1. Sample Data

The RNA-Seq data from the study (Li et al., 2014) was downloaded from recount3 (Wilks et al., 2021) database (recount3: uniformly processed RNA-seq). The data is indexed as SRP035988 in the database and comprised of high-quality single end reads. The sequencing data comprised of skin punch biopsy RNA of skin from 83 normal individuals and lesional skin from 95 psoriatic patients, refer (Li et al., 2014) and fetched from the database with *create_rse_manual*() function in the recount3 R package (Collado-Torres, 2022). The metadata was fetched with *read_metadata*() function (see file SRP035988_meta.csv). The read counts per sample was obtained from *compute_read_counts*().

2.2. Scaling and Normalisation of Read Counts

The scaling and normalisation of read counts per sample is necessary as the longer RNA transcripts will produce more reads and result in biased conclusions. This was done using *recount::getRPKM*() to obtain transformed normalised read counts known as RPKM (reads per kilobase of transcript) based on average mapped read length see Supplementary fig. 1.

A given gene might have different read coverage depth among different samples hence normalisation of read counts between samples is necessary. This normalisation step was done using the edgeR package (Chen et al., 2016; McCarthy et al., 2012; Robinson et al., 2010). This process comprised of two steps: firstly, filtering and retaining only those genes that are expressed in adequate number of samples and which contain defined annotation and recomputed the library sizes using *filterByExp*r(). Secondly, the normalisation of the filtered genes using *calcNormFactors*() and estimation of count-per-million (CPM) with *cpm*(). The dispersion of the genes expressed was calculated with *estimateDisp*() and differential gene expression GLM model was calculated with GLM quasi likelihood *glmQLFit*() function. The dispersion (biological CV) and mean deviance of the log CPM was estimated with *estimateDisp*() and *glmQLFit*() functions, respectively.

2.3. Differential Expression of Genes

Genes express differently, allowing to detect the highly expressed and suppressed genes responsible for a disease condition. The upregulated and downregulated differentially expressed genes were detected with *glmQLFTest*(). Only significant genes were selected with *topTags*() with cut-off p-value < 0.001 using edgeR package.

2.4. Unsupervised learning: PCA and Clustering

Unsupervised learning was conducted to detect downregulated and upregulated genes. PCA was done on significant differentially expressed genes data with *prcomp*() and *.scale* parameter set TRUE. Clustering was done on that data after scaling. K-means clustering of these genes was performed based on *clusGap*() with wss statistic and parameters *nstart =20*, *K.max*=20, *B*=30 (B - bootstrap) for detecting optimal clusters. K-means model was applied with optimal clusters set 9 (obtained from wss elbow method) and nstart=50 using *kmeans*(). Hierarchical clustering was conducted based on optimal number of clusters obtained from *clusGap*() with FUN = *hcut*, *nstart* =15, *K.max*=10, *B*=20. The best model was run with function hclust(), dist() set as Euclidean method and *method* = *"ward.D2"* (gives highest agglomerative coefficient) and the tree was cut at *k*=9.

2.5. Supervised learning: Random Forest, LDA, LASSO regression

For supervised learning methods, the significant differentially expressed genes data was split into training and tests with 0.75 and 0.25 proportions respectively. Random forests were run on training data and full data with *mtry* set for all predictors in each split and *ntree* =1000 with *randomForest*(). The variables with lesional psoriatic skin > 0.0 and those which intersected in both the training and full data were selected for Boruta algorithm. The algorithm detects the most important variables based on significant p-value. It was run with parameters *doTrace* = 2, *maxRuns* = 500 with *Boruta*() and further run with *TentativeRoughFix*() from Boruta R package (Kursa & Rudnicki, 2010). T-tests for each gene across all the samples were performed using *rowttests*() to filter out insignificant genes with p-value > 0.05. LDA was conducted on split and prepocessed data (i.e., centering and scaling of data to avoid biased interpretations) using *lda*(). Lasso regression was modelled on training data through cross validation *cv.glmnet*() yielding minimum λ. For selection of optimum number of folds to obtain minimum λ, cross validation was run in different number of folds ranging from 3 to 20, and final cross validation was done with 11 folds. The best λ was plugged to predict training data. The three learning methods were compared through cross validation with 15 folds and 10 repeats.

2.6. Signature-based clustering

Signature based clustering employs gene expression ranking and their subsequent supervised classification to determine gene expression clusters. For this, rScudo (Ciciani et al., 2019; Lauria, 2013)was used. It was trained on data split with 0.75 and 0.25 proportions and *nTop* = 25, *nBottom* = 25, *alpha* = 0.05 set using *scudoTrain*(). The network of training data was constructed with *scudoNetwork*() with *N* = 0.2 set.

scudoTest() was used on test data and the plots were created with *scudoPlot*(). For cross validation, scudoModel() at *N*=0.50 was used. *scudoClassify*() was used for test data classification with *N*=0.45, *alpha* = 0.05 and cross validated optimal *nTop* and *nBottom* values.

All the analyses were done in R (R Core Team, 2021) and the entire code script is available as Final_project.Rmd. The R packages used were: arsenal (Heinzen et al., 2021), ggplot2 (Wickham, 2016), statmod (Dunn & Smyth, 1996; Giner & Smyth, 2016; Hu & Smyth, 2009; Phipson & Smyth, 2010; Smyth, 2002, 2005a, 2005b), dplyr (Wickham et al., 2022), plotly (Sievert, 2020), ggfortify (Horikoshi & Tang, 2018; Tang et al., 2016), cluster (Maechler et al., 2021), factoextra (Kassambara & Mundt, 2020), reshape2 (Wickham, 2007), rsample (Silge et al., 2021), randomForest (Liaw & Wiener, 2002), genefilter (Gentleman et al., 2021), dendextend (Galili, 2015), MASS (Venables & Ripley, 2002), tidyverse (Wickham et al., 2019), caret (Kuhn, 2022), tidymodels (Kuhn & Wickham, 2020), e1071 (Meyer et al., 2021), ROCR (Sing et al., 2005), glmnet, (Friedman et al., 2010; Simon et al., 2011), biomaRt (Durinck et al., 2005, 2009), igraph (Csardi & Nepusz, 2006).

2.7. Functional Enrichment Analysis

Expression of highly confident genes leads to understanding their roles in important functional gene pathways related to disease development or various regular cellular functions. For this study, g:GOSt (for functional profiling) of g:profiler (Raudvere et al., 2019; Reimand et al., 2016) was run under default settings of g:SCS significance test and searched for only annotated genes with default p-value of 0.05 set.

2.8. Network Based Enrichment Analysis

The network based enrichment analysis was done to identify important reactome and GO based pathways related to the significant 79 genes with EnrichNet (Glaab et al., 2012) along with tissue specificity. Analysis with cytoscape (Pillich Rudolf T. and Chen, 2017; Pratt et al., 2015, 2017) was also done.

3. Results

3.1. Reads Counts and Reads Normalization

The reads ranged approx. between 8 to 46 million for different samples see Supplementary fig. 1. The RPKM ranged between 400,000 to 550,000 see Supplementary fig. 2. The filtering out of low expressed genes and genes that are not present in all samples, resulted in 24553 genes out of total 63856 genes. After the CPM normalisation between different samples, the CPM was around 1000 see Supplementary fig. 3. The log scaled CPM normalisation across samples resulted in CPM approx. 3 see Figure 1.The biological CV and mean deviance were around 0.5 and 1 respectively see Supplementary fig. 4.

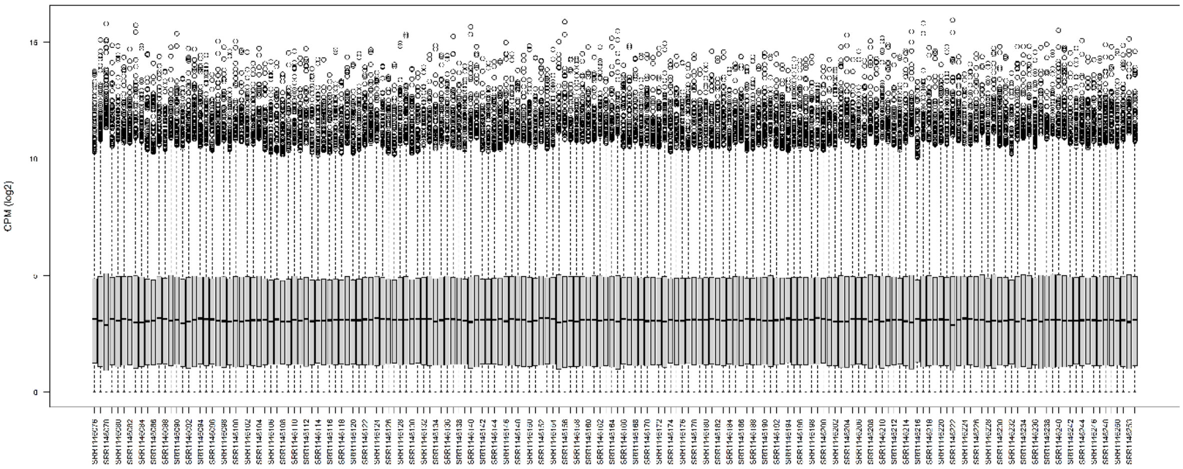


Figure 1. Log scaled CPM (y axis) per sample (x axis).

3.2. Differentially Expressed Genes

Out of 178 total samples (normal and psoriatic), 9497 genes were significantly downregulated and 9495 were upregulated whereas 5561 genes were insignificant and hence, were removed from further analyses see Table 1.

Table 1. Number of genes differentially expressed and regulated.

Genes regulated	Genes number
Downregulated	9497
Not Significant	5561
Upregulated	9495

3.3. PCA and Clustering

The PCA showed clear cut separation between the normal and psoriatic samples based on total upregulated and downregulated genes. One lesional psoriatic sample was assigned to normal samples, this is indicative of sample mislabelling. Around 44% of variance was well explained through PCA see Figure 2. The PCA of downregulated and upregulated genes separately showed similar trend see Supplementary fig. 5. For the clustering methods, optimal number of clusters were 9 see Supplementary fig. 6.

The k-means and hierarchical clustering generated 8 different clusters from all the samples and one psoriatic sample remained un-clustered. The normal samples formed 3 clusters and psoriatic samples formed 5 clusters see Figure 3 and Figure 4.

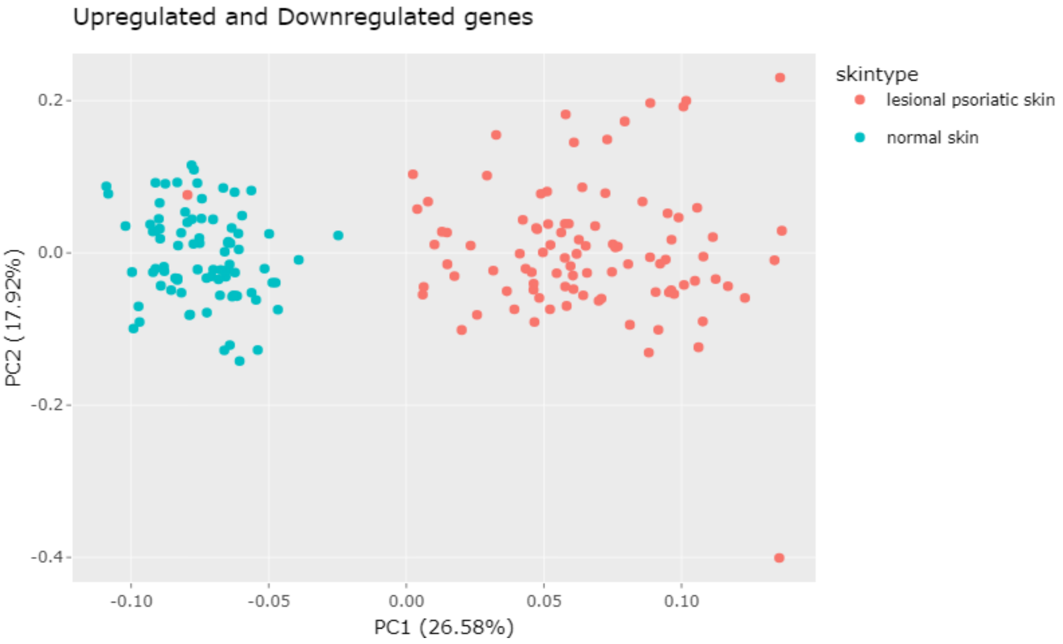


Figure 2. PCA analysis of total genes (upregulated and downregulated) for skin type conditions.

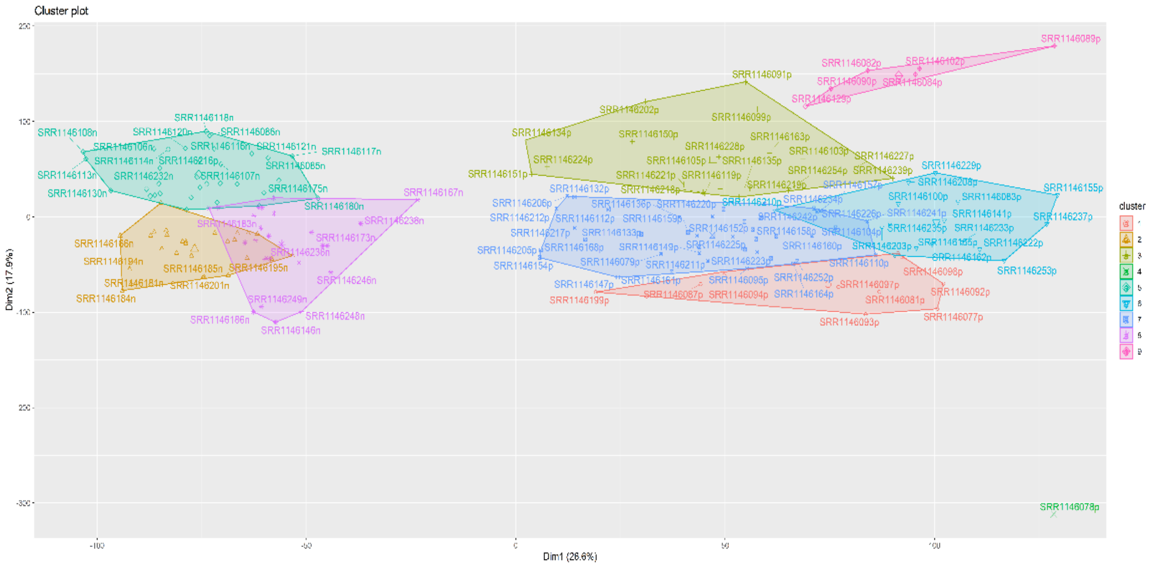


Figure 3. K-means clustering of normal samples and psoriatic samples. The 3 clusters on the left represent the normal samples and 5 clusters on the right represent the lesional psoriatic samples. One lesional psoriatic sample did not cluster (represented in green at bottom right

3.4. Supervised learning: Random Forest, LDA and Lasso Regression

Separate random forest runs on trained and full data resulted in total of 314 and 205 important features (or genes) respectively. Out of all genes obtained from both the runs, 157 genes were found in common. The Boruta algorithm run on 157 genes gave 69 important (significant) features see Supplementary fig. 9. The tentative fix on these features resulted in 79 significant features see Figure 5, Supplementary files-bor_genes_gprofiler.txt and hgnc_bor.txt. After the tentative fix, the total important features increased as the unassigned features from the Boruta algorithm were later assigned based on their importance calculated during the procedure. Hence, 79 genes from Boruta algorithm and its tentative fix were taken for functional enrichment analysis. The heatmap of these 79 genes represented several genes with high expression in psoriatic samples and low expression in normal samples. Nearly 10 genes were downregulated in the psoriatic samples as compared to the expression levels in normal skin see Figure 6. The T-tests retained all 157 genes with p-value < 0.05 cut-off. LDA trained on 75% of the full data, predicted the features on the test data accurately enough see Supplementary fig. 10. In Lasso regression, the cross validation at 4, 6 and 11 folds gave same minimum λ see Supplementary fig. 11. However, at two former folds values minimum λ fluctuated with each run, but this was not seen in 11 folds and was optimal see Supplementary fig. 12. Lasso regression model selected only 3 genes, viz., ENSG00000241794.1 (*SPRR2A*), ENSG00000206073.10 (*SERPINB4*), ENSG00000124102.4 (*PI3*) and filtered out 154 out of total 157 genes. The cross validated models of supervised learning indicated that random forest and lasso regression have higher prediction accuracy than LDA see Figure 7. The rScudo resulted in two well defined clusters separated according to the disease and normal conditions. Some samples from both types of skin condition remained unclustered. The igraph::cluster_spinglass() resulted in no connections formation between the normal and psoriatic clusters.

3.5. Functional enrichment analysis

The functional enrichment of the 79 genes from the random forest and Boruta coupled analyses, revealed that some genes are involved in most important pathways, viz., RAGE receptor binding, chemotaxis, bacterium response, cell migration, cell chemotaxis, cell motility, cell localisation, skin2 corneal layer, skin in granular layer and spinous layer and other important pathways see Figure 9.

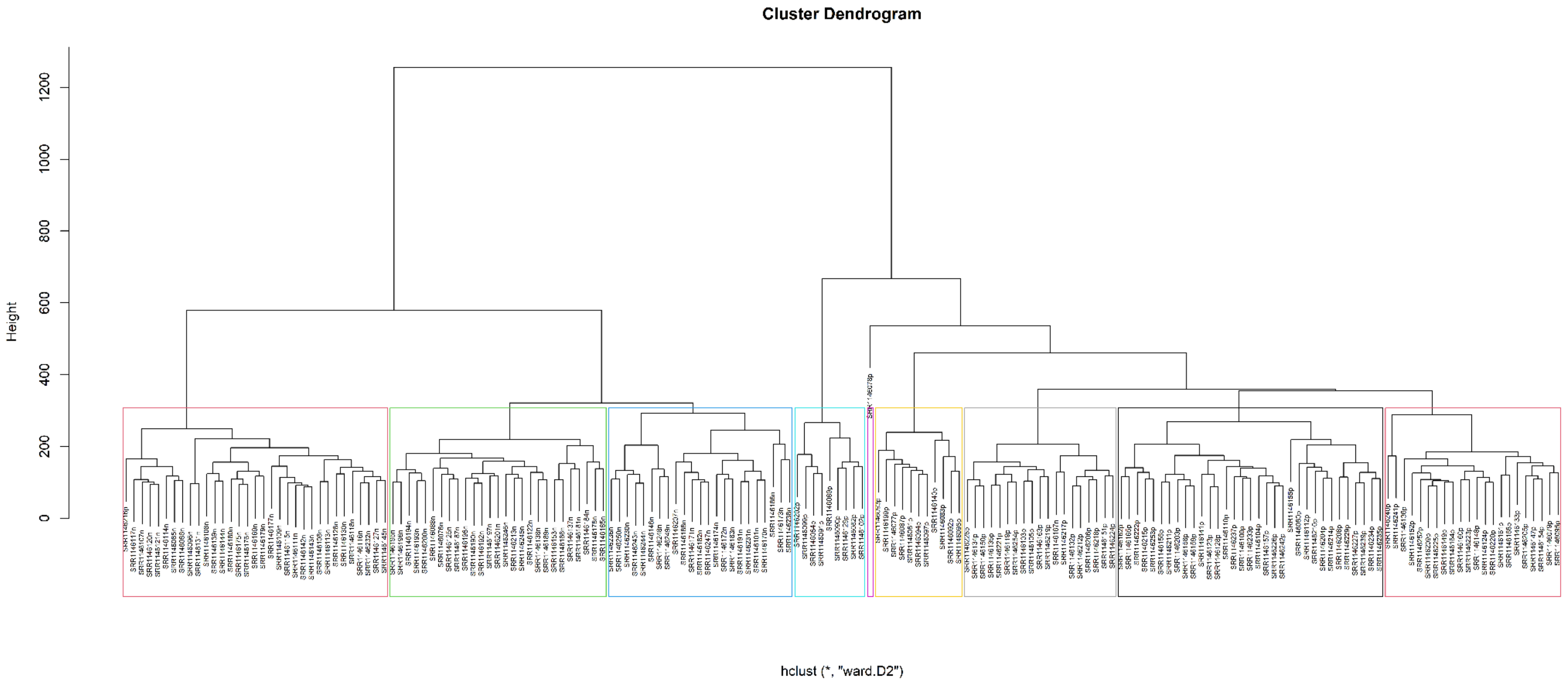


Figure 4. Hierarchical clustering of normal and psoriatic samples. 8 clusters are seen separating all the samples with one lesional psoriatic sample remained un-clustered.

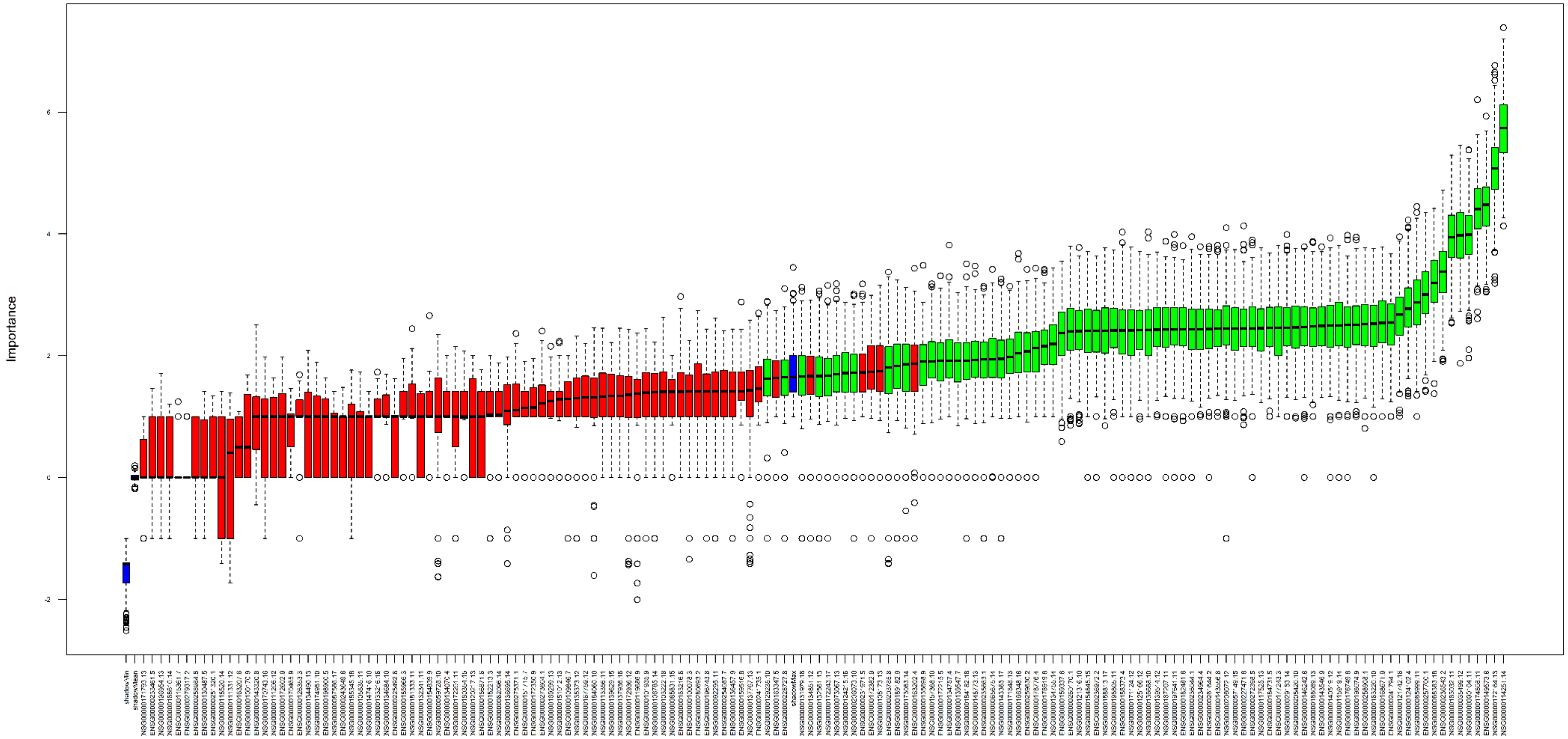


Figure 5. Important features (or genes) based on Boruta algorithm and tentative fix of genes assignment. 79 significantly important features (in green), 78 non-significant features (in red) out of total 157 genes represented on x axis.

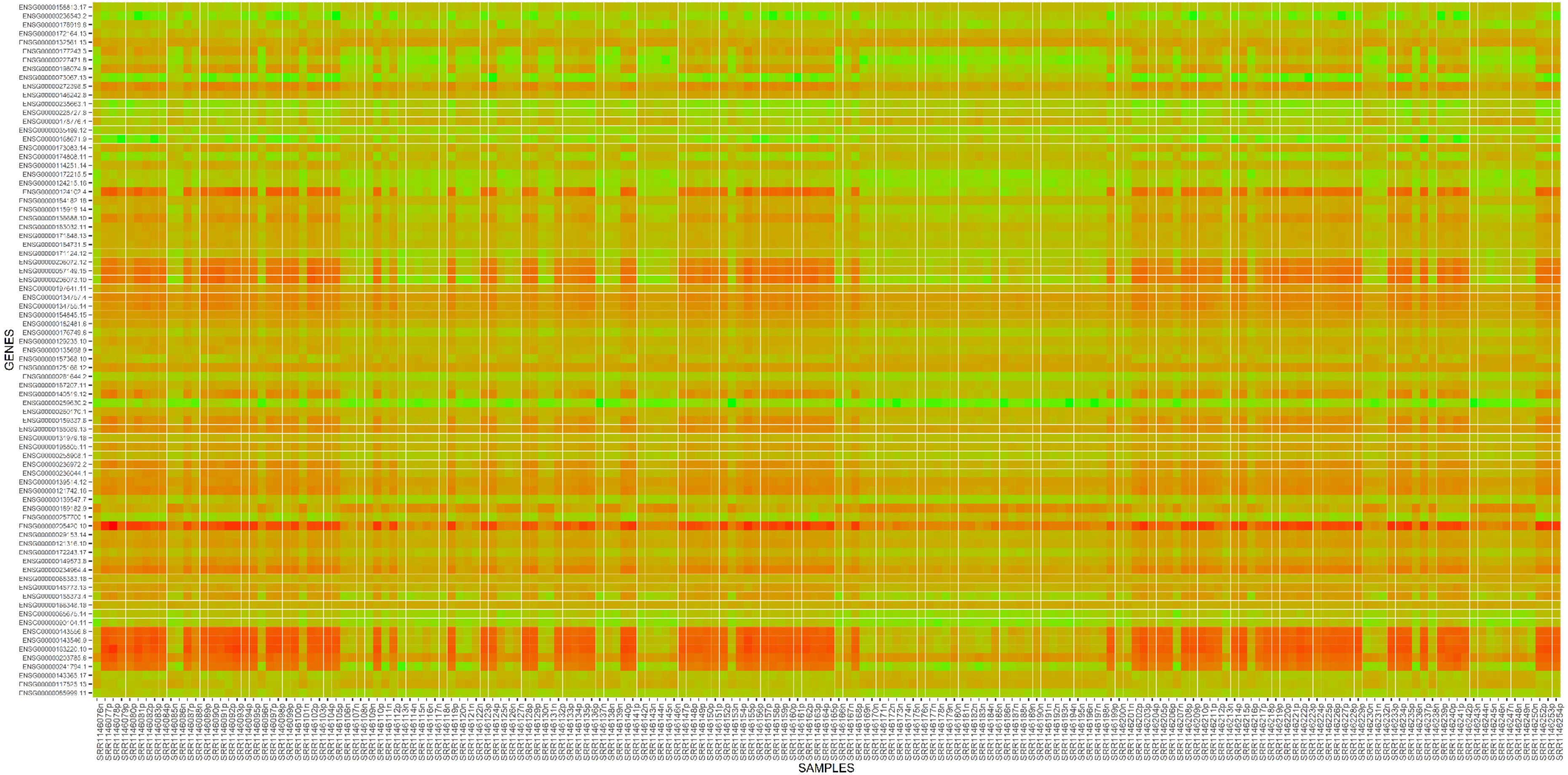


Figure 6. Heatmap of 79 genes selected from the Boruta algorithm and tentative fix. The x axis represents samples and y axis represents genes. Red indicates high expression and green indicates low expression of genes.

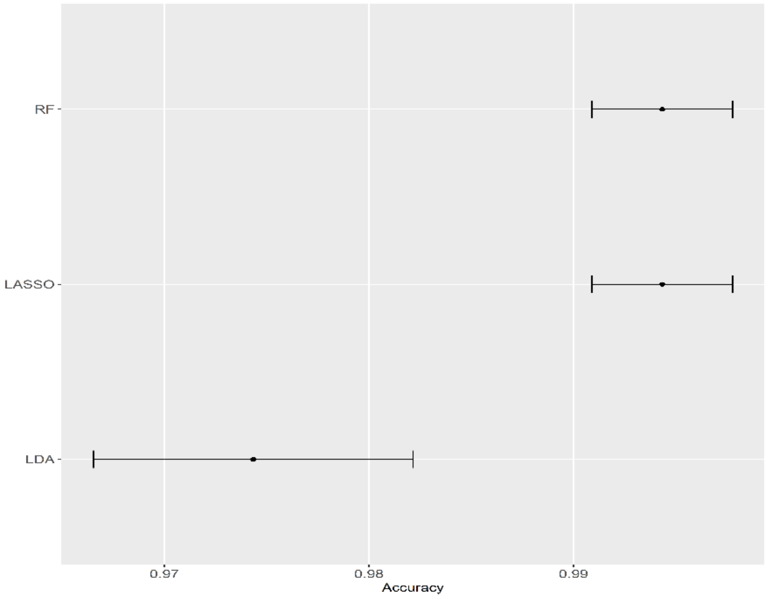


Figure 7. . Comparison between different supervised learning models based on their prediction accuracy.

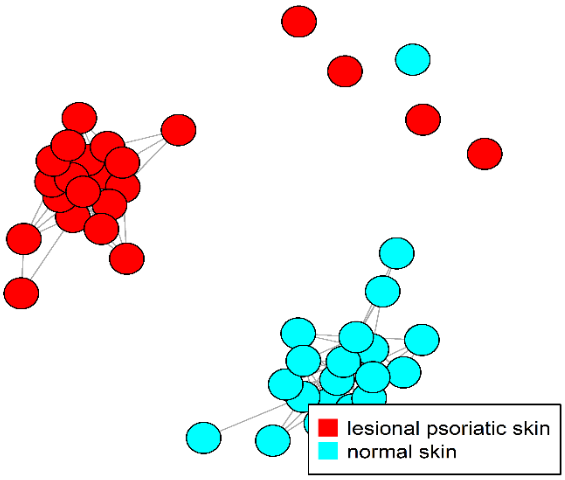
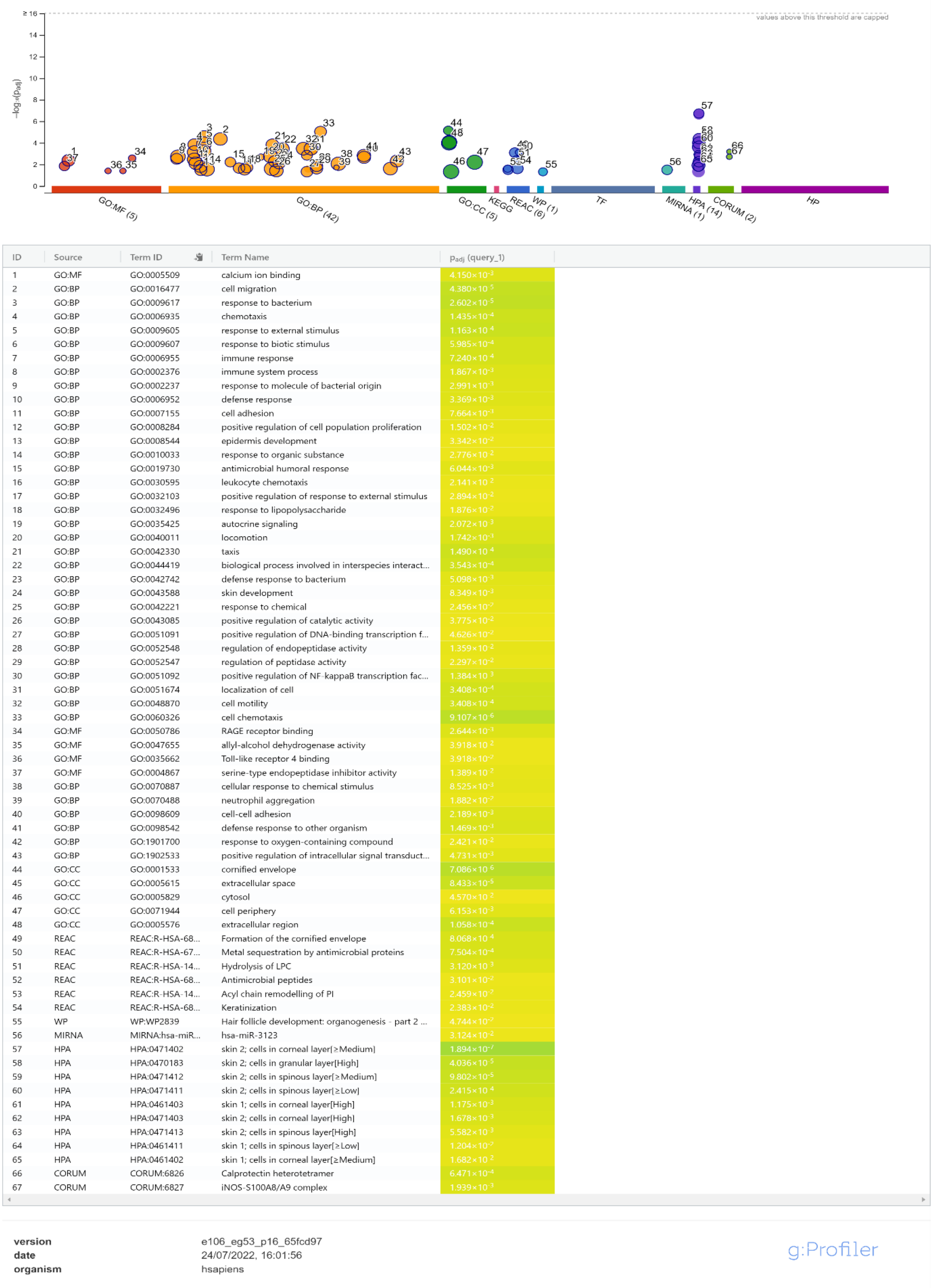


Figure 8. The clustering of normal and psoriatic samples by rScudo.

The genes which act in gene sets or independently, affecting several crucial cellular functions, thereby, inducing psoriasis were even detected through functional profiling. These genes were associated to different reaction pathways, viz., metal sequestration by antimicrobial proteins, formation of cornified envelope, hydrolysis of LPC, keratinisation, acyl chain modelling of PI and antimicrobial peptides see Supplementary fig. 13.



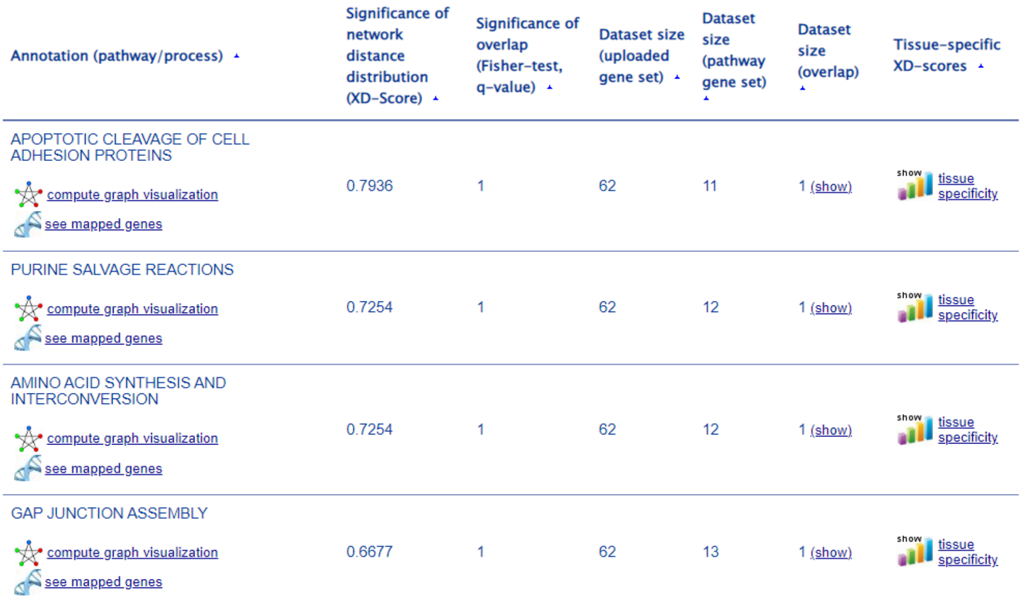
ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0005509	calcium ion binding	4.150×10^{-3}
2	GO:BP	GO:0016477	cell migration	4.380×10^{-5}
3	GO:BP	GO:0009617	response to bacterium	2.602×10^{-5}
4	GO:BP	GO:0006935	chemotaxis	1.435×10^{-4}
5	GO:BP	GO:0009605	response to external stimulus	1.163×10^{-4}
6	GO:BP	GO:0009607	response to biotic stimulus	5.985×10^{-4}
7	GO:BP	GO:0006955	immune response	7.240×10^{-4}
8	GO:BP	GO:0002376	immune system process	1.867×10^{-3}
9	GO:BP	GO:0002237	response to molecule of bacterial origin	2.991×10^{-3}
10	GO:BP	GO:0006952	defense response	3.369×10^{-3}
11	GO:BP	GO:0007155	cell adhesion	7.664×10^{-3}
12	GO:BP	GO:0008284	positive regulation of cell population proliferation	1.502×10^{-2}
13	GO:BP	GO:0008544	epidermis development	3.342×10^{-2}
14	GO:BP	GO:0010033	response to organic substance	2.776×10^{-2}
15	GO:BP	GO:0019730	antimicrobial humoral response	6.044×10^{-3}
16	GO:BP	GO:0030595	leukocyte chemotaxis	2.141×10^{-2}
17	GO:BP	GO:0032103	positive regulation of response to external stimulus	2.894×10^{-2}
18	GO:BP	GO:0032496	response to lipopolysaccharide	1.876×10^{-2}
19	GO:BP	GO:0035425	autocrine signaling	2.072×10^{-3}
20	GO:BP	GO:0040011	locomotion	1.742×10^{-3}
21	GO:BP	GO:0042330	taxis	1.490×10^{-4}
22	GO:BP	GO:0044419	biological process involved in interspecies interact...	3.543×10^{-4}
23	GO:BP	GO:0042742	defense response to bacterium	5.098×10^{-3}
24	GO:BP	GO:0043588	skin development	8.349×10^{-3}
25	GO:BP	GO:0042221	response to chemical	2.456×10^{-2}
26	GO:BP	GO:0043085	positive regulation of catalytic activity	3.775×10^{-2}
27	GO:BP	GO:0051091	positive regulation of DNA-binding transcription f...	4.626×10^{-2}
28	GO:BP	GO:0052548	regulation of endopeptidase activity	1.359×10^{-2}
29	GO:BP	GO:0052547	regulation of peptidase activity	2.297×10^{-2}
30	GO:BP	GO:0051092	positive regulation of NF-kappaB transcription fac...	1.384×10^{-3}
31	GO:BP	GO:0051674	localization of cell	3.408×10^{-4}
32	GO:BP	GO:0048870	cell motility	3.408×10^{-4}
33	GO:BP	GO:0060326	cell chemotaxis	9.107×10^{-6}
34	GO:MF	GO:0050786	RAGE receptor binding	2.644×10^{-3}
35	GO:MF	GO:0047655	allyl-alcohol dehydrogenase activity	3.918×10^{-2}
36	GO:MF	GO:0035662	Toll-like receptor 4 binding	3.918×10^{-2}
37	GO:MF	GO:0004867	serine-type endopeptidase inhibitor activity	1.389×10^{-2}
38	GO:BP	GO:0070887	cellular response to chemical stimulus	8.525×10^{-3}
39	GO:BP	GO:0070488	neutrophil aggregation	1.882×10^{-2}
40	GO:BP	GO:0098609	cell-cell adhesion	2.189×10^{-3}
41	GO:BP	GO:0098542	defense response to other organism	1.469×10^{-3}
42	GO:BP	GO:1901700	response to oxygen-containing compound	2.421×10^{-2}
43	GO:BP	GO:1902533	positive regulation of intracellular signal transduct...	4.731×10^{-3}
44	GO:CC	GO:0001533	cornified envelope	7.086×10^{-6}
45	GO:CC	GO:0005615	extracellular space	8.433×10^{-5}
46	GO:CC	GO:0005829	cytosol	4.570×10^{-2}
47	GO:CC	GO:0071944	cell periphery	6.153×10^{-3}
48	GO:CC	GO:0005576	extracellular region	1.058×10^{-4}
49	REAC	REAC:R-HSA-68...	Formation of the cornified envelope	8.068×10^{-4}
50	REAC	REAC:R-HSA-67...	Metal sequestration by antimicrobial proteins	7.504×10^{-4}
51	REAC	REAC:R-HSA-14...	Hydrolysis of LPC	3.120×10^{-3}
52	REAC	REAC:R-HSA-68...	Antimicrobial peptides	3.101×10^{-2}
53	REAC	REAC:R-HSA-14...	Acyl chain remodelling of PI	2.459×10^{-2}
54	REAC	REAC:R-HSA-68...	Keratinization	2.383×10^{-2}
55	WP	WP:WP2839	Hair follicle development: organogenesis - part 2 ...	4.744×10^{-2}
56	MIRNA	MIRNA:hsa-miR...	hsa-miR-3123	3.124×10^{-2}
57	HPA	HPA:0471402	skin 2; cells in corneal layer[≥Medium]	1.894×10^{-7}
58	HPA	HPA:0470183	skin 2; cells in granular layer[High]	4.036×10^{-5}
59	HPA	HPA:0471412	skin 2; cells in spinous layer[≥Medium]	9.802×10^{-5}
60	HPA	HPA:0471411	skin 2; cells in spinous layer[≥Low]	2.415×10^{-4}
61	HPA	HPA:0461403	skin 1; cells in corneal layer[High]	1.175×10^{-3}
62	HPA	HPA:0471403	skin 2; cells in corneal layer[High]	1.678×10^{-3}
63	HPA	HPA:0471413	skin 2; cells in spinous layer[High]	5.582×10^{-3}
64	HPA	HPA:0461411	skin 1; cells in spinous layer[≥Low]	1.204×10^{-2}
65	HPA	HPA:0461402	skin 1; cells in corneal layer[≥Medium]	1.682×10^{-2}
66	CORUM	CORUM:6826	Calprotectin heterotetramer	6.471×10^{-4}
67	CORUM	CORUM:6827	iNOS-S100A8/A9 complex	1.939×10^{-3}

version	e106_eg53_p16_65fcd97
date	24/07/2022, 16:01:56
organism	hsapiens

Figure 9. Functional enrichment analysis of important gene functional pathways and reactions for 79 genes from Boruta tentative fix. The dark green values represent cellular reactions which are more significant.

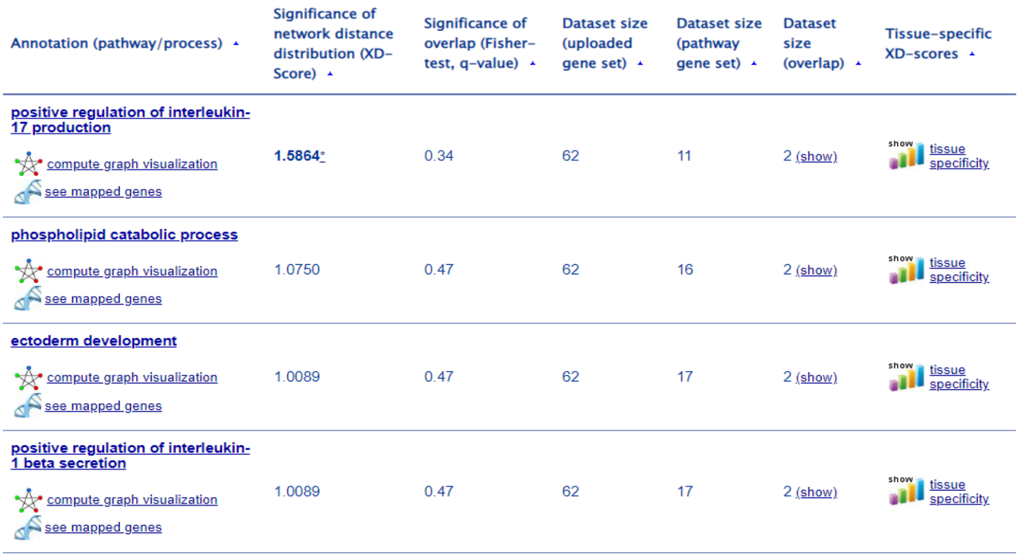
3.6. Network Based Enrichment Analysis

The enrichment analysis based on reactome pathways and processes based on our 79 gene set gave some important pathways affecting cell adhesion proteins and amino acid synthesis and purine salvage reactions and gap junction, but statistical significance was low for these genes see Figure 10. Only 6 gene was responsible for these pathways with the significance XD score above 0.50 see Table 2. The GO pathways showed only 1 significant positive regulation of IL-17 pathway controlled by genes *NOD2* and *PRKCQ* see Figure 11.



Annotation (pathway/process) ▲	Significance of network distance distribution (XD-Score) ▲	Significance of overlap (Fisher-test, q-value) ▲	Dataset size (uploaded gene set) ▲	Dataset size (pathway gene set) ▲	Dataset size (overlap) ▲	Tissue-specific XD-scores ▲
APOPTOTIC CLEAVAGE OF CELL ADHESION PROTEINS compute graph visualization see mapped genes	0.7936	1	62	11	1 (show)	show tissue specificity
PURINE SALVAGE REACTIONS compute graph visualization see mapped genes	0.7254	1	62	12	1 (show)	show tissue specificity
AMINO ACID SYNTHESIS AND INTERCONVERSION compute graph visualization see mapped genes	0.7254	1	62	12	1 (show)	show tissue specificity
GAP JUNCTION ASSEMBLY compute graph visualization see mapped genes	0.6677	1	62	13	1 (show)	show tissue specificity

Figure 10. Network based gene enrichment by EnrichNet based on reactome database for pathways

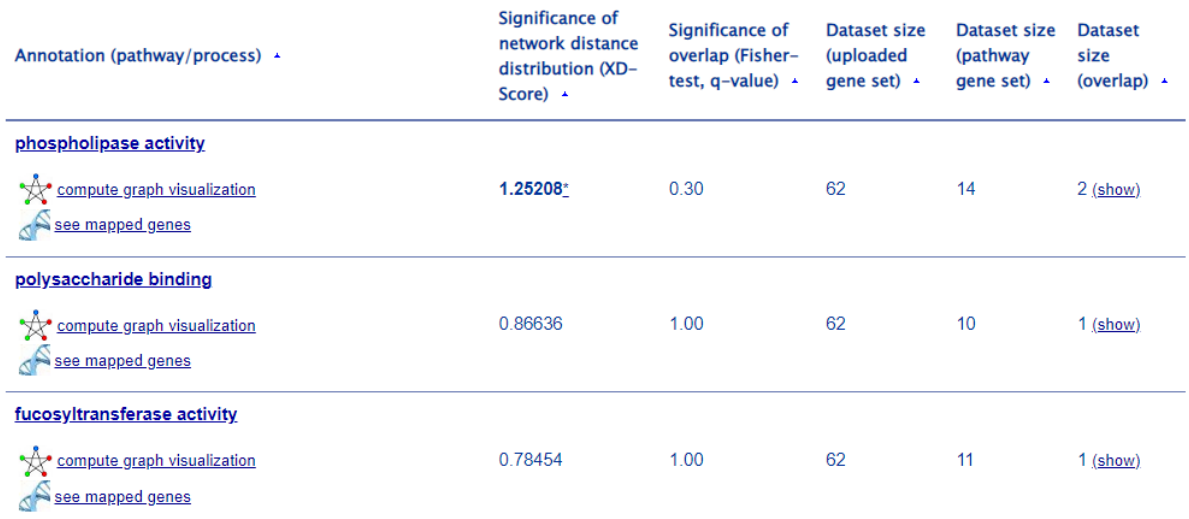


Annotation (pathway/process) ▲	Significance of network distance distribution (XD-Score) ▲	Significance of overlap (Fisher-test, q-value) ▲	Dataset size (uploaded gene set) ▲	Dataset size (pathway gene set) ▲	Dataset size (overlap) ▲	Tissue-specific XD-scores ▲
positive regulation of interleukin-17 production compute graph visualization see mapped genes	**1.5864***	0.34	62	11	2 (show)	show tissue specificity
phospholipid catabolic process compute graph visualization see mapped genes	1.0750	0.47	62	16	2 (show)	show tissue specificity
ectoderm development compute graph visualization see mapped genes	1.0089	0.47	62	17	2 (show)	show tissue specificity
positive regulation of interleukin-1 beta secretion compute graph visualization see mapped genes	1.0089	0.47	62	17	2 (show)	show tissue specificity

Figure 11. Network based gene enrichment by EnrichNet based on GO database for pathways

Table 2. Enriched Pathways and the genes involved found by EnrichNet.

Pathways	Overlap genes
Apoptotic cleavage of cell adhesion proteins	*DSG3*
Purine Salvage Reactions	*PNP*
Amino acid synthesis and interconversion	*GOT2*
Gap Junction Assembly	*GJB6*
CRMPS in SEMA3A signalling	*CDK5R1*
Synthesis and interconversion of nucleotide di and triphosphates	*RRM2*



Annotation (pathway/process) ▲	Significance of network distance distribution (XD-Score) ▲	Significance of overlap (Fisher-test, q-value) ▲	Dataset size (uploaded gene set) ▲	Dataset size (pathway gene set) ▲	Dataset size (overlap) ▲
phospholipase activity compute graph visualization see mapped genes	**1.25208***	0.30	62	14	2 (show)
polysaccharide binding compute graph visualization see mapped genes	0.86636	1.00	62	10	1 (show)
fucosyltransferase activity compute graph visualization see mapped genes	0.78454	1.00	62	11	1 (show)

Figure 12. Network based enrichment analysis with EnrichNet using GO database for molecular functions.

Analysis with Cytoscape did not detect major skin-affecting pathways except, hair follicle development (*EDA*, *GJB6*, *WNT5A*) and Endogenous TLR signalling (*S100A8*, *S100A8*) and these genes were already found during functional profiling, hence the analysis tool was further aborted.

4. Discussion

PCA is a strong tool to detect sample mislabelling see Figure 2. In PCA, K-means and hierarchical clustering, one psoriatic samples did not cluster, probably due to low to no expression of some important genes or maybe due to contamination see Figure 2, Figure 3, Figure 4. The 8 different clusters obtained from k-means and hierarchical clustering suggests that expression of certain gene sets in a group of individuals, with normal or psoriatic condition, might express at different degrees. If all the genes expressed uniformly amongst all the normal individuals and likewise even in the diseased individuals, then only two main clusters would have been evident, however, this was not observed. Random forest provided high number of genes with differing importance when run on training and full data see section 3.4 and Supplementary fig. 7 and Supplementary fig. 8. This might be due to extraction of more genes under limited training data observations. However, for extraction of high confidence genes it might prove beneficial to retain the genes common to training data-run and full data-run modes. The Boruta algorithm might be an alternative for feature (or important genes) selection rather than arbitrary top features' number set based on random forest's variable importance see Figure 5. As noticed in the heatmap see Figure 6, the low expressed genes in the psoriatic samples might be genes related to normal cell functions which might function abnormally in the given disease condition, but these genes have high gene expression in normal condition. Whereas the highly expressed genes in the psoriatic sample might be related to overactivation of immune response pathways.

In this study, LDA was not the optimal selection for gene expression analysis as it suffered multicollinearity of features. This could be because certain psoriasis-activating genes might be paralogous and LDA might treat it as technical artifact and not deal them as biological complexity. Lasso regression can overcome such multicollinearity issues, however, it selected only 3 features and shrunk coefficients of 154 features to 0. It proved too stringent for this study. It might fail to discover valid disease-associated rare genes which might not be present in all the samples or expressed only in high detectable amounts at advanced stage of the disease and henceforth, not get functionally enriched. In signature-based clustering, the unclustered samples apart from psoriatic disease and normal clusters could be due to some specific genes which might have nearly similar expression levels in disease and normal samples see Figure 8. There was no connection between the two clusters. This indicates that the networks formed by the two skin conditions are very distinct from each other.

Among the common genes obtained from Boruta algorithm with fix and lasso regression namely, ENSG00000241794.1 (*SPRR2A:*keratinization), ENSG00000206073.10 (*SERPINB4*:inflammatory response), ENSG00000124102.4

(*PI3*), *SERPINB4* and *SPRR2B* are concordant with the findings of (Li et al., 2014) (see Table1 (*SPRR2A* instead of *SPRR2B*)) see Supplementary files: hgnc_bor.txt (contains 76 annotated genes), hgnc_lasso.txt (3 genes), bor_genes_gprofiler.txt (79 genes). In this study, *C10orf99* gene was found and is similar to *C1orf68* mentioned by Li et al., 2014. The *C10orf99* is an important psoriasis marker gene (Guo et al., 2014). Other genes, viz., *S100A8*, *S100A7*, *S100A9*, *KRT77*, *KRT6A*, *SPRR2E*, *SPRR2A* detected here, are concordant to previous findings by (Krishnan & Kõks, 2022). Some genes from the studies by Li et al., 2014, (Xu et al., 2022) were not detected in this study, possibly due to different R packages and methods used here.

From the network enrichment analysis of 79 genes, around 6 genes were above 0.5 significance XD score in the reactome pathway see Table 2 and Figure 10. These genes might play a role in psoriasis however were not marked significant by EnrichNet. But these 6 genes have already been stated for their important role in connection to psoriasis diagnosis and treatment in the past findings (NIETSCHE, 1978; Omura & others, 2000; Ramos et al., 2021; Suárez-Fariñas et al., 2011; Zhang et al., 2019, Ramos et al., 2021) . This could be due to the significance score calculated for these genes among all the genes set involved in a given pathway are based on mapping to reference gene set pathways from the GO or reactome database. And if these databases have few genes sets reported for certain complex disease pathways for instance, here psoriasis, then the target genes might get low significance XD score or if the target genes are found in multiple pathways then, they might be given same overrepresentation score, and low XD score. However, these low significance genes might be interesting as they might depict new functional networks discoveries not yet reported (Glaab et al., 2012). The other genes were not significantly enriched even in EnrichNet GO pathways, except for IL-17 production regulated by *PRKCQ* gene, see Figure 11 and section 3.6, and was supported by findings of (Kim et al., 2016; Lin et al., 2021). However, *NOD2* has been reported to play no role in psoriasis in the past (Zhu et al., 2012) but one study stresses on its effect on psoriasis (Shehata et al., 2022). The *PLA2G4E* and *PLA2G4D* genes were found significant (Figure 12) for the phospholipase activity when enriched with GO pathway for molecular function and supported by findings of (Liang et al., 2022; Shao et al., 2021). Overall, this study revealed many genes from functional profiling as compared to few from the network enrichment for pathways. These genes are not well studied in past findings and invites additional research. Moreover, it also gives a detailed insight into psoriasis responsible genes which could help develop precision medicine for this incurable skin condition.

References

Armstrong, A. W., Mehta, M. D., Schupp, C. W., Gondo, G. C., Bell, S. J., & Griffiths, C. E. M. (2021). Psoriasis prevalence in adults in the United States. *JAMA Dermatology*, *157*(8), 940–946.

Chen, Y., Lun, A. A. T., & Smyth, G. K. (2016). From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research*, *5*, 1438. https://doi.org/10.12688/f1000research.8987.2

Ciciani, M., Cantore, T., & Lauria, M. (2019). rScudo: an R package for classification of molecular profiles using rank-based signatures. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btaa296

Collado-Torres, L. (2022). *Explore and download data from the recount3 project*. https://doi.org/10.18129/B9.bioc.recount3

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Systems*, 1695. https://igraph.org

Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *J. Comput. Graph. Statist*, *5*, 236–244.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., de Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, *21*, 3439–3440.

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, *4*, 1184–1191.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Galili, T. (2015). dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btv428

Gentleman, R., Carey, V. J., Huber, W., & Hahne, F. (2021). *genefilter: genefilter: methods for filtering genes from high-throughput experiments*.

Giner, G., & Smyth, G. K. (2016). statmod: probability calculations for the inverse Gaussian distribution. *R Journal*, *8*(1), 339–351.

Glaab, E., Baudot, A., Krasnogor, N., Schneider, R., & Valencia, A. (2012). EnrichNet: network-based gene set enrichment analysis. *Bioinformatics*, *28*(18), i451.

Guo, P., Luo, Y., Mai, G., Zhang, M., Wang, G., Zhao, M., Gao, L., Li, F., & Zhou, F. (2014). Gene expression profile based classification models of psoriasis.

Genomics, *103*(1), 48–55. https://doi.org/https://doi.org/10.1016/j.ygeno.2013.11.001

Heinzen, E., Sinnwell, J., Atkinson, E., Gunderson, T., & Dougherty, G. (2021). *arsenal: An Arsenal of “R” Functions for Large-Scale Statistical Summaries*. https://CRAN.R-project.org/package=arsenal

Horikoshi, M., & Tang, Y. (2018). *ggfortify: Data Visualization Tools for Statistical Analysis Results*. https://CRAN.R-project.org/package=ggfortify

Hu, Y., & Smyth, G. K. (2009). ELDA: extreme limiting dilution analysis for comparing depleted and enriched populations in stem cell and other assays. *Journal of Immunological Methods*, *347*(1), 70–78.

Kassambara, A., & Mundt, F. (2020). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. https://CRAN.R-project.org/package=factoextra

Kim, J., Bissonnette, R., Lee, J., da Rosa, J. C., Suárez-Fariñas, M., Lowes, M. A., & Krueger, J. G. (2016). The spectrum of mild to severe psoriasis vulgaris is defined by a common activation of IL-17 pathway genes, but with key differences in immune regulatory genes. *Journal of Investigative Dermatology*, *136*(11), 2173–2182.

Krishnan, V. S., & Kõks, S. (2022). Transcriptional Basis of Psoriasis from Large Scale Gene Expression Studies: The Importance of Moving towards a Precision Medicine Approach. *International Journal of Molecular Sciences*, *23*(11). https://doi.org/10.3390/ijms23116130

Kuhn, M. (2022). *caret: Classification and Regression Training*. https://CRAN.R-project.org/package=caret

Kuhn, M., & Wickham, H. (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.* https://www.tidymodels.org

Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software*, *36*(11), 1–13. http://www.jstatsoft.org/v36/i11/

Lauria, M. (2013). Rank-based transcriptional signatures. *Systems Biomedicine*, *1*(4), 228–239. https://www.tandfonline.com/doi/abs/10.4161/sysb.25982

Li, B., Tsoi, L. C., Swindell, W. R., Gudjonsson, J. E., Tejasvi, T., Johnston, A., Ding, J., Stuart, P. E., Xing, X., Kochkodan, J. J., Voorhees, J. J., Kang, H. M., Nair, R. P., Abecasis, G. R., & Elder, J. T. (2014). Transcriptome analysis of psoriasis in a large case-control sample: RNA-seq provides insights into disease mechanisms. *Journal of Investigative Dermatology*, *134*(7). https://doi.org/10.1038/jid.2014.28

Liang, L., Takamiya, R., Miki, Y., Heike, K., Taketomi, Y., Sugimoto, N., Yamaguchi, M., Shitara, H., Nishito, Y., Kobayashi, T., & others. (2022). Group IVE cytosolic phospholipase A2 limits psoriatic inflammation by mobilizing the anti-inflammatory lipid N-acylethanolamine. *The FASEB Journal*, *36*(5), e22301.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22. https://CRAN.R-project.org/doc/Rnews/

Lin, J., Li, X., Zhang, F., Zhu, L., & Chen, Y. (2021). Transcriptome wide analysis of long non-coding RNA-associated ceRNA regulatory circuits in psoriasis. *Journal of Cellular and Molecular Medicine*, *25*(14), 6925–6935.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2021). *cluster: Cluster Analysis Basics and Extensions*. https://CRAN.R-project.org/package=cluster

McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, *40*(10), 4288–4297. https://doi.org/10.1093/nar/gks042

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2021). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. https://CRAN.R-project.org/package=e1071

NIETSCHE, U.-B. (1978). Photochemotherapy of psoriasis. *International Journal of Dermatology*, *17*(2), 149–157.

Omura, G. A., & others. (2000). Inhibitors of the enzyme purine nucleoside phosphorylase as potential therapy for psoriasis. *Current Pharmaceutical Design*, *6*(9), 943–959.

Phipson, B., & Smyth, G. K. (2010). Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, *9*(1), Article 39.

Pillich Rudolf T. and Chen, J. and R. V. and W. D. and P. D. (2017). NDEx: A Community Resource for Sharing and Publishing of Biological Networks. In C. N. and R. K. E. Wu Cathy H. and Arighi (Ed.), *Protein Bioinformatics: From Protein Modifications and Networks to Proteomics* (pp. 271–301). Springer New York. https://doi.org/10.1007/978-1-4939-6783-4_13

Pratt, D., Chen, J., Pillich, R., Rynkov, V., Gary, A., Demchak, B., & Ideker, T. (2017). NDEx 2.0: A Clearinghouse for Research on Cancer Pathways. *Cancer Research*, *77*(21), e58–e61. https://doi.org/10.1158/0008-5472.CAN-17-0606

Pratt, D., Chen, J., Welker, D., Rivas, R., Pillich, R., Rynkov, V., Ono, K., Miello, C., Hicks, L., Szalma, S., Stojmirovic, A., Dobrin, R., Braxenthaler, M., Kuentzer, J., Demchak, B., & Ideker, T. (2015). NDEx, the Network Data Exchange. *Cell Systems*, *1*(4), 302–305. https://doi.org/https://doi.org/10.1016/j.cels.2015.10.001

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/

Ramos, W., Gutierrez, E. L., Jiménez, G., Lazarte, J. S., Ronceros, G., & Ortega-Loayza, A. G. (2021). Simultaneous endemic pemphigus foliaceus and psoriasis vulgaris in Peru–immunogenetic or environmental factors? *Dermatology Review/Przeglad Dermatologiczny*, *108*(2), 153–159.

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., & Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, *47*(W1), W191–W198. https://doi.org/10.1093/nar/gkz369

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., & Vilo, J. (2016). g:Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Research*, *44*(W1), W83–W89. https://doi.org/10.1093/nar/gkw199

Rendon, A., & Schäkel, K. (2019). Psoriasis pathogenesis and treatment. In *International Journal of Molecular Sciences* (Vol. 20, Issue 6). https://doi.org/10.3390/ijms20061475

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Shao, S., Chen, J., Swindell, W. R., Tsoi, L. C., Xing, X., Ma, F., Uppala, R., Sarkar, M. K., Plazyo, O., Billi, A. C., & others. (2021). Phospholipase A2 enzymes represent a shared pathogenic pathway in psoriasis and pityriasis rubra pilaris. *JCI Insight*, *6*(20).

Shehata, W. A., Shoeib, M., Shoeib, M. M., Shokhba, H., & Shams, A. (2022). Nucleotide binding and oligomerization domain 2 in psoriasis: a clinical and immunohistochemical study. *Journal of Immunoassay and Immunochemistry*, *43*(1), 43–53.

Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC. https://plotly-r.com

Silge, J., Chow, F., Kuhn, M., & Wickham, H. (2021). *rsample: General Resampling Infrastructure*. https://CRAN.R-project.org/package=rsample

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, *39*(5), 1–13. https://doi.org/10.18637/jss.v039.i05

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, *21*(20), 7881. http://rocr.bioinf.mpi-sb.mpg.de

Smyth, G. K. (2002). An efficient algorithm for REML in heteroscedastic regression. *Journal of Computational and Graphical Statistics*, *11*, 836–847.

Smyth, G. K. (2005a). Numerical integration. *Encyclopedia of Biostatistics*, 3088–3095.

Smyth, G. K. (2005b). Optimization and nonlinear equations. *Encyclopedia of Biostatistics*, 3088–3095.

Suárez-Fariñas, M., Fuentes-Duculan, J., Lowes, M. A., & Krueger, J. G. (2011). Resolved psoriasis lesions retain expression of a subset of disease-related genes. *Journal of Investigative Dermatology*, *131*(2), 391–400.

Tang, Y., Horikoshi, M., & Li, W. (2016). ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages. *The R Journal*, *8*(2), 474–485. https://doi.org/10.32614/RJ-2016-060

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth). Springer. https://www.stats.ox.ac.uk/pub/MASS4/

WHO. (2016, October 26). *Global report on Psoriasis*.

Wickham, H. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software*, *21*(12), 1–20. http://www.jstatsoft.org/v21/i12/

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., & Müller, K. (2022). *dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr

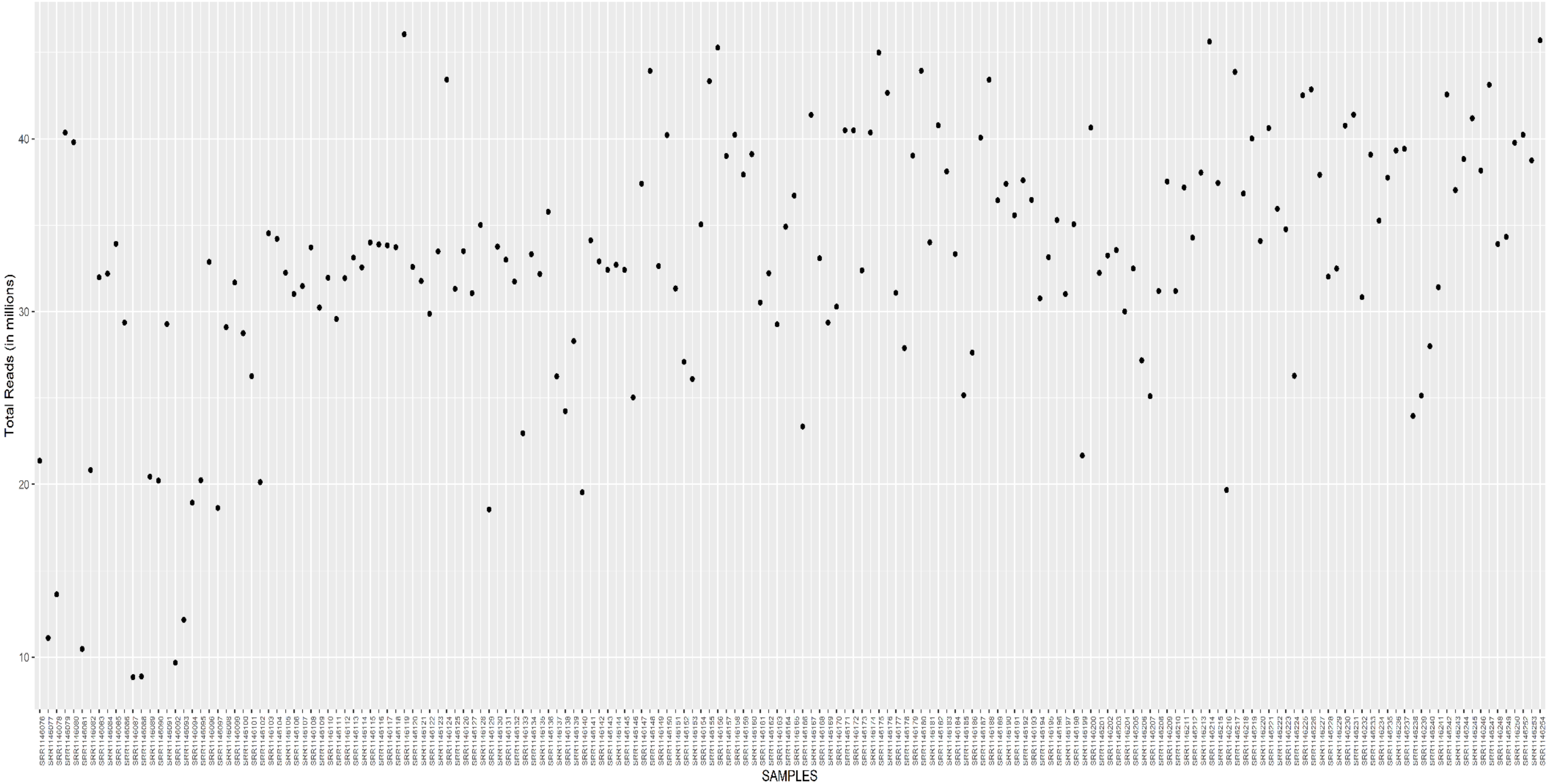
Wilks, C., Zheng, S. C., Chen, F. Y., Charles, R., Solomon, B., Ling, J. P., Imada, E. L., Zhang, D., Joseph, L., Leek, J. T., Jaffe, A. E., Nellore, A., Collado-Torres, L., Hansen, K. D., & Langmead, B. (2021). recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biol*. https://doi.org/10.1186/s13059-021-02533-6

Xu, Q., Zheng, X., Mao, Y., Chen, W., Chen, S., Zhang, H., Zhen, Q., Li, B., Yong, L., Ge, H., Yu, Y., Zhang, R., Cao, L., Cheng, H., Wang, W., & Sun, L. (2022). Gene interaction analysis of psoriasis in Chinese Han population. *Molecular Genetics & Genomic Medicine*, *10*(5), e1858. https://doi.org/https://doi.org/10.1002/mgg3.1858

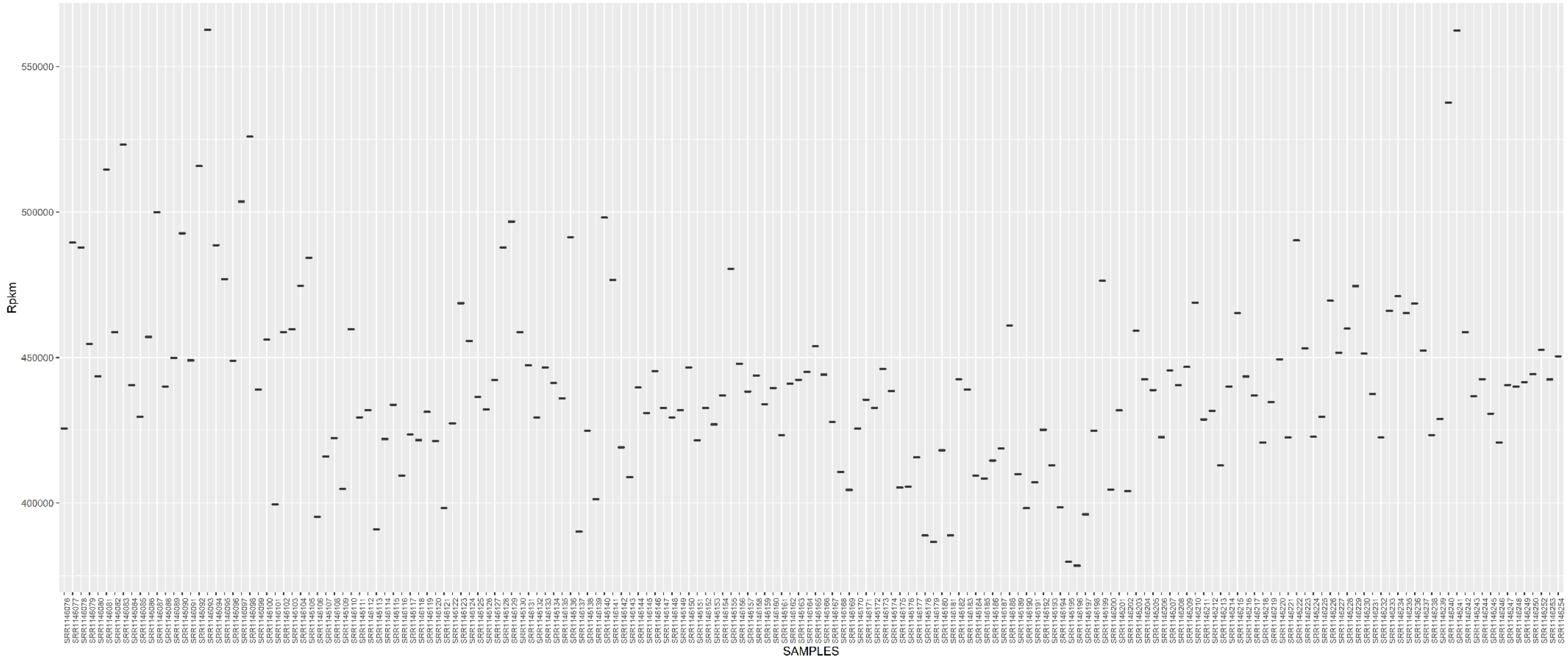
Zhang, Y.-J., Sun, Y.-Z., Gao, X.-H., & Qi, R.-Q. (2019). Integrated bioinformatic analysis of differentially expressed genes and signaling pathways in plaque psoriasis. *Molecular Medicine Reports*, *20*(1), 225–235.

Zhu, K., Yin, X., Tang, X., Zhang, F., Yang, S., & Zhang, X. (2012). Meta-analysis of NOD2/CARD15 polymorphisms with psoriasis and psoriatic arthritis. *Rheumatology International*, *32*(7), 1893–1900. https://doi.org/10.1007/s00296-011-1813-2

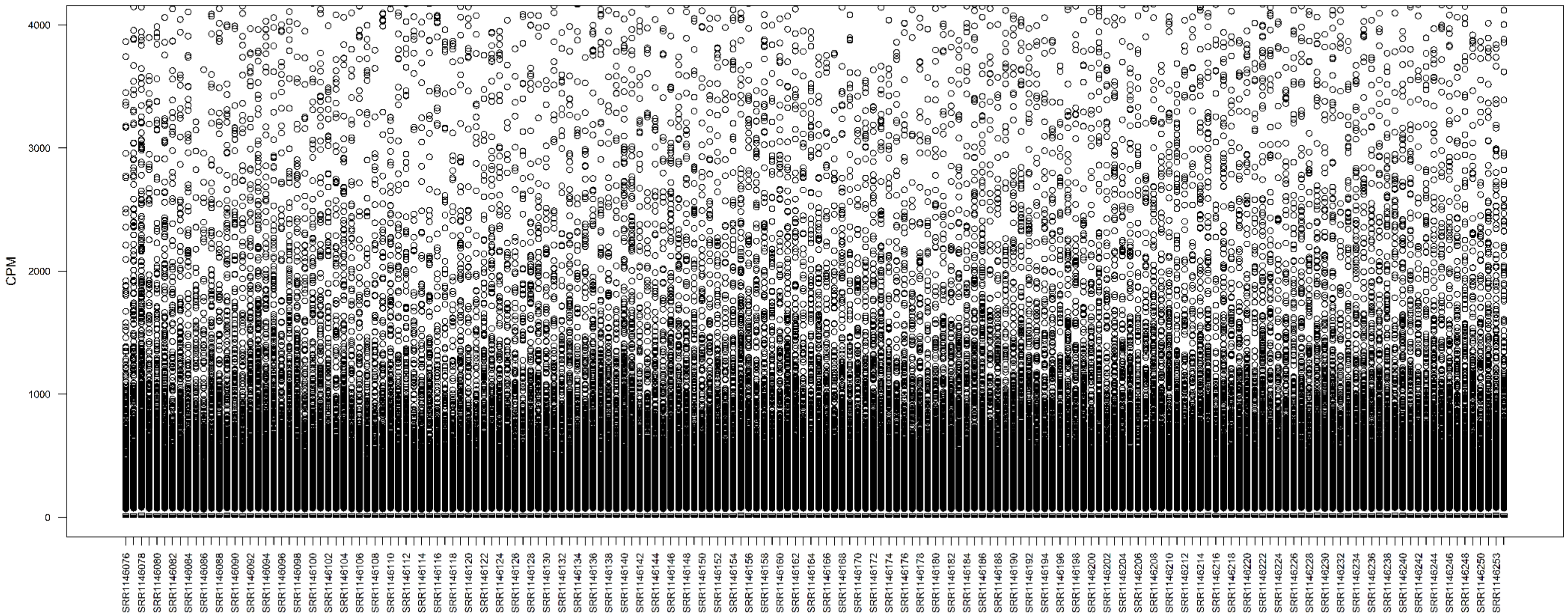
Supplementary



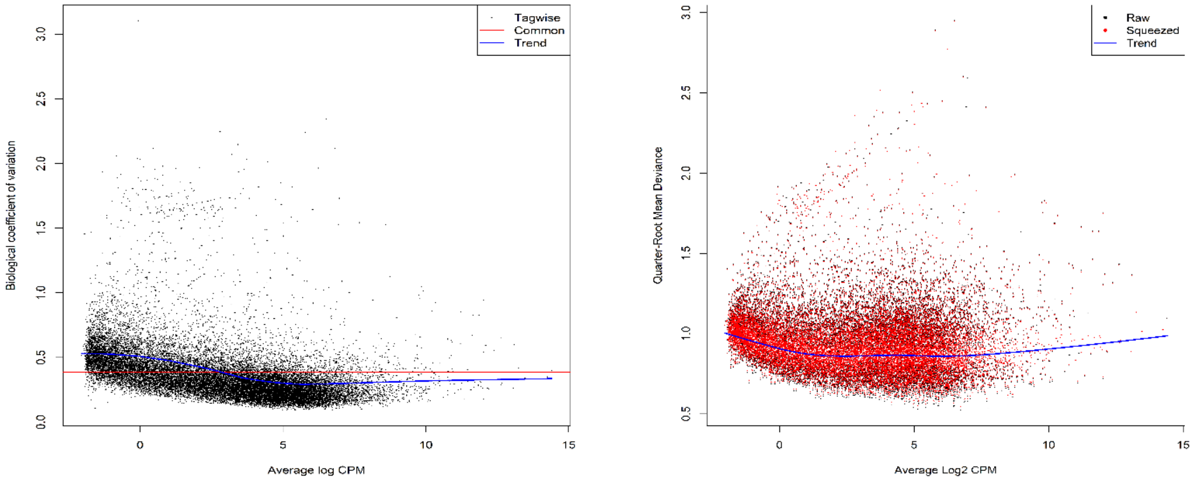
Supplementary fig. 1. Total reads per sample.



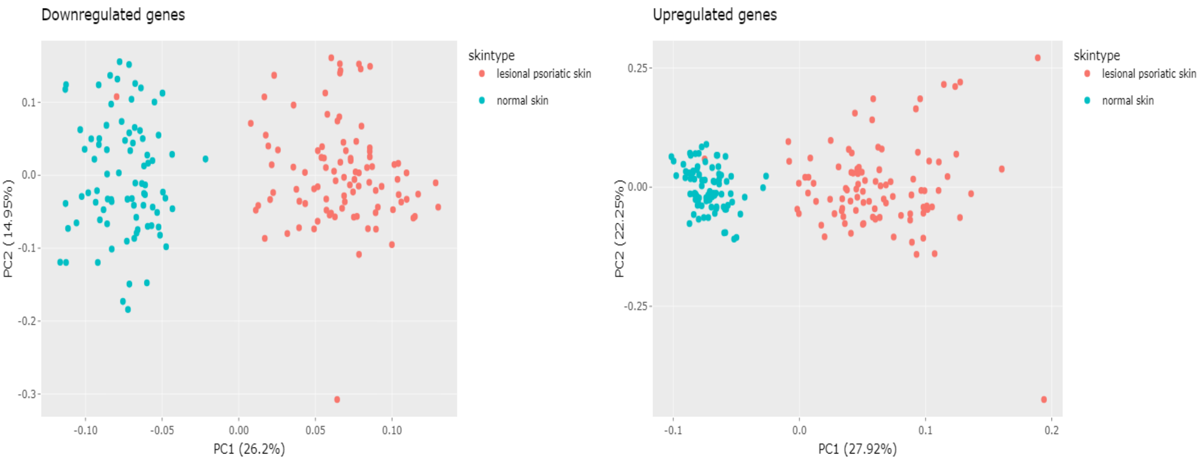
Supplementary fig. 2. RPKM before normalisation between samples.



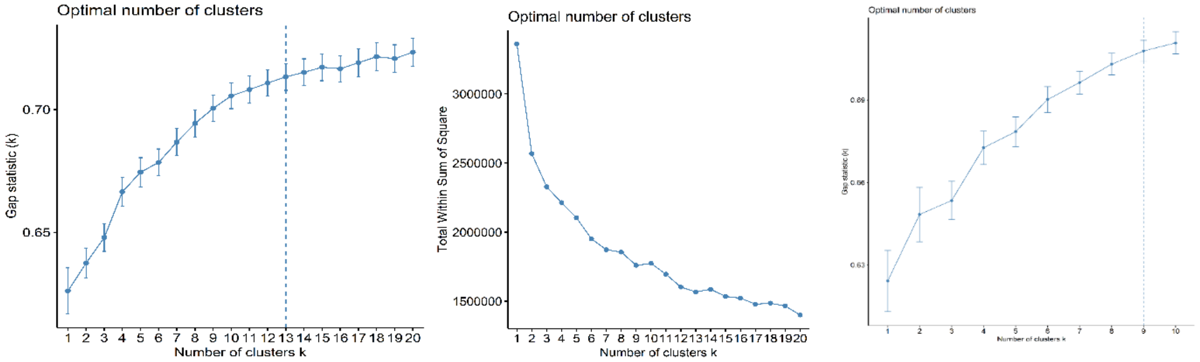
Supplementary fig. 3. CPM filtered and normalised for each sample (represented on x axis).



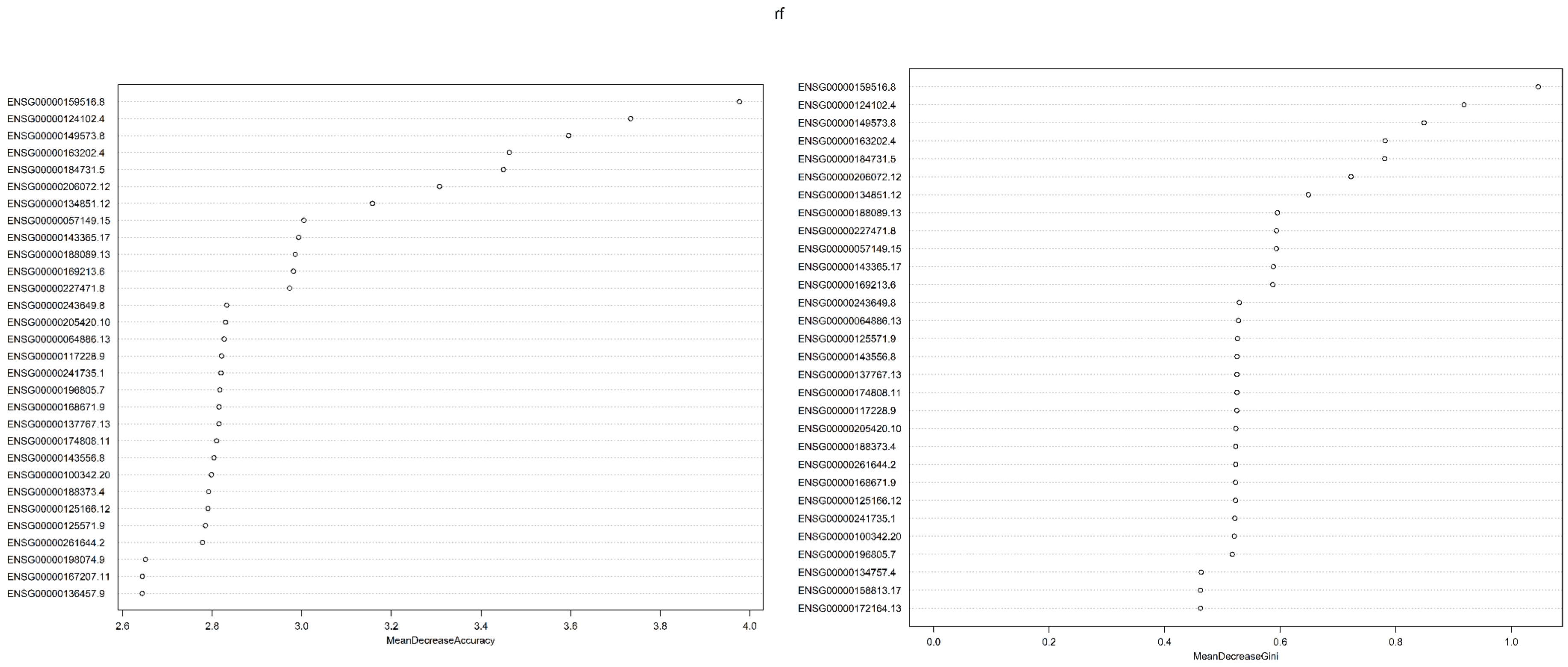
Supplementary fig. 4. Biological coefficient of variation (left) and quarter root deviance (right) versus average log CPM.



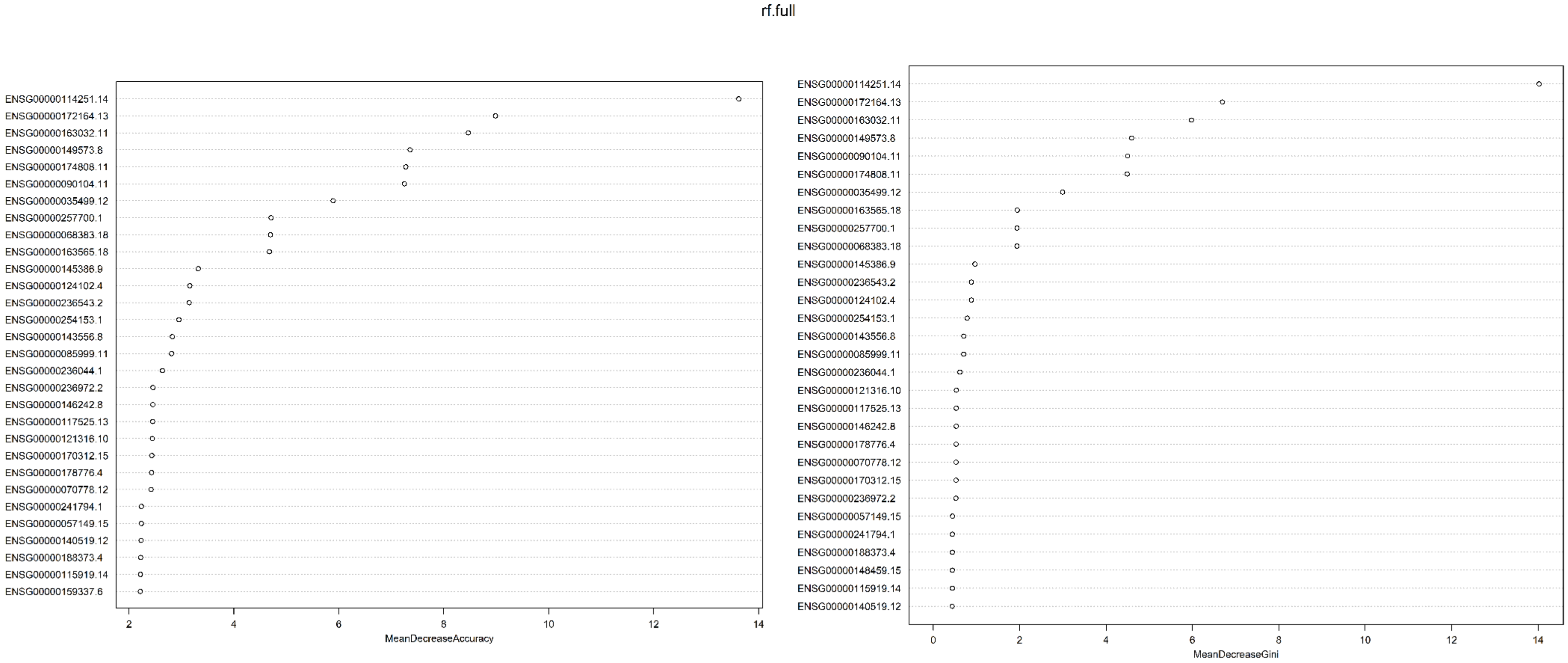
Supplementary fig. 5. PCA of downregulated genes (left) and upregulated genes (right).



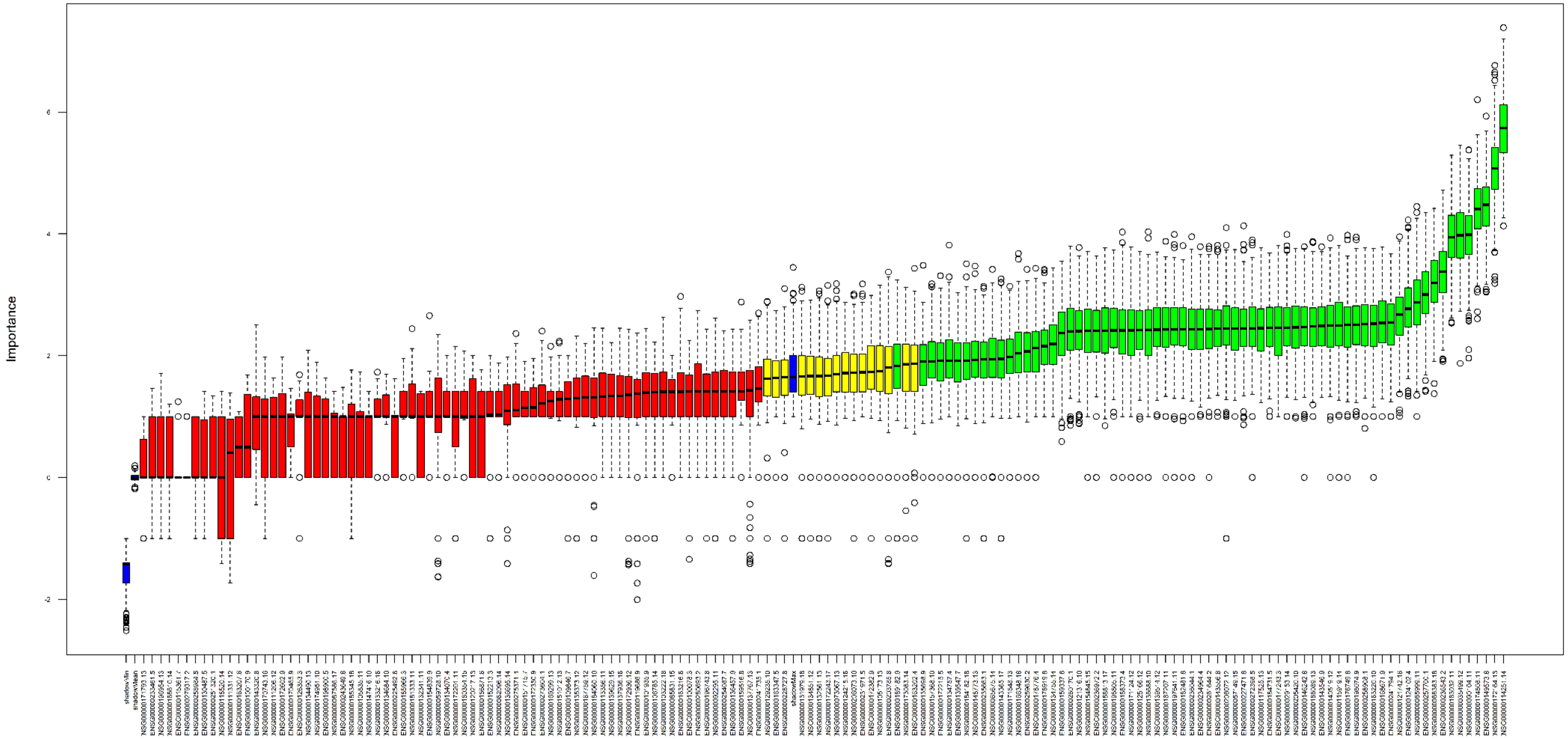
Supplementary fig. 6. Optimal number of clusters for k-means using gap statistic (left), within sum of squares i.e., wss method (middle) and hierarchical clustering (right).



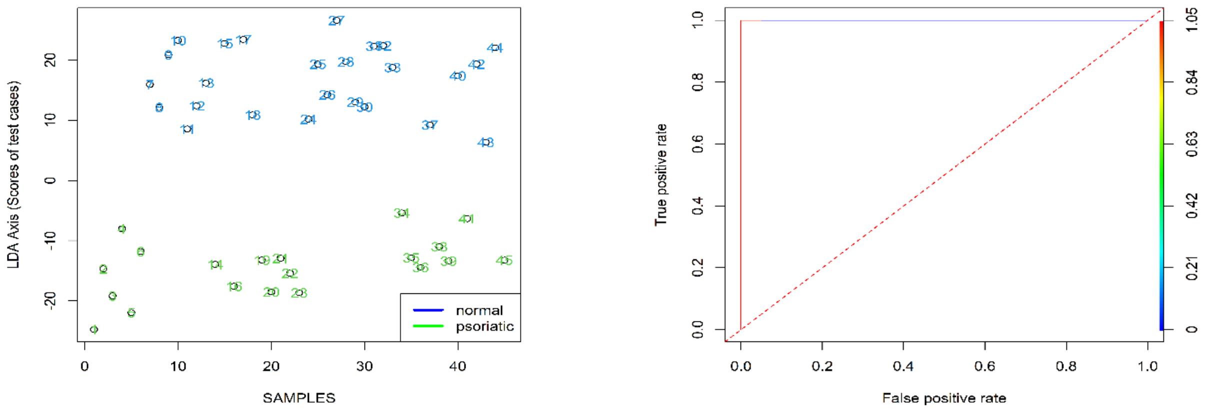
Supplementary fig. 7. Variable importance from Random Forest on training data. Most important genes on the y axis and Mean Decrease Accuracy (left) and Mean Decrease Gini (right).



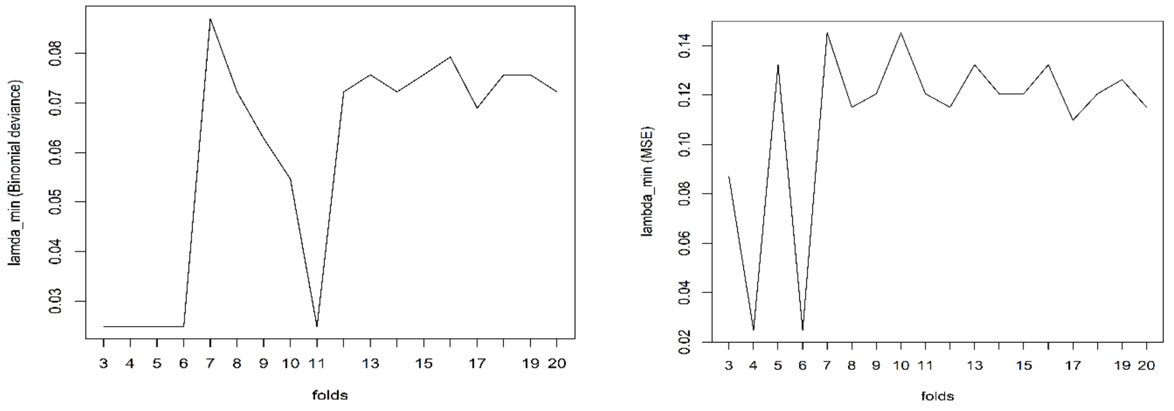
Supplementary fig. 8. Variable importance from Random Forest on full data. Most important genes on the y axis and Mean Decrease Accuracy (left) and Mean Decrease Gini (right).



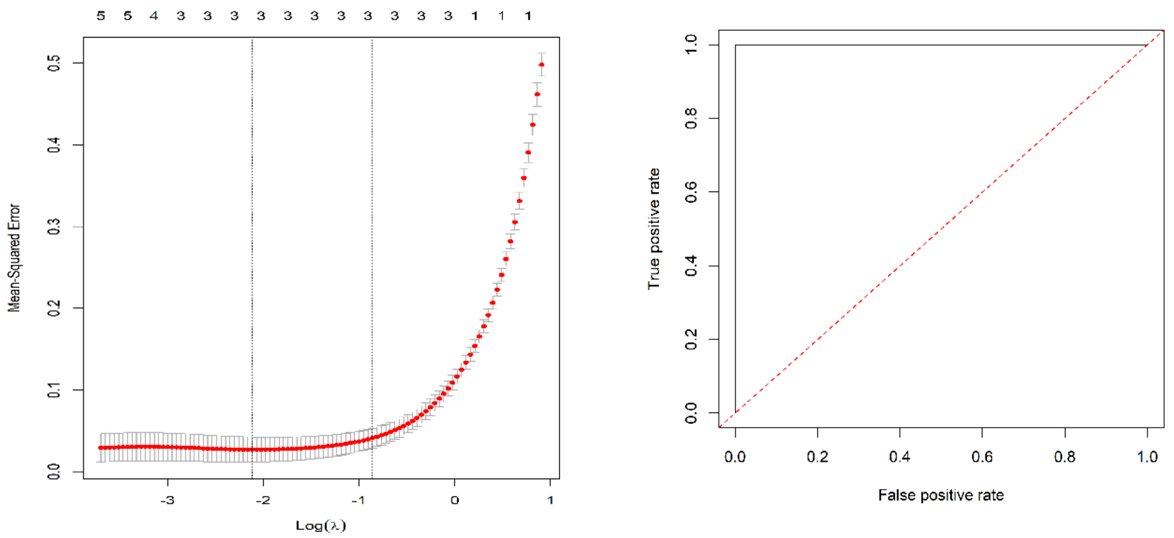
Supplementary fig. 9. Important genes after Boruta algorithm: 69 features were important (in green) and 72 unimportant (in red) and 16 tentative (in yellow) out of 157 genes represented on x axis.



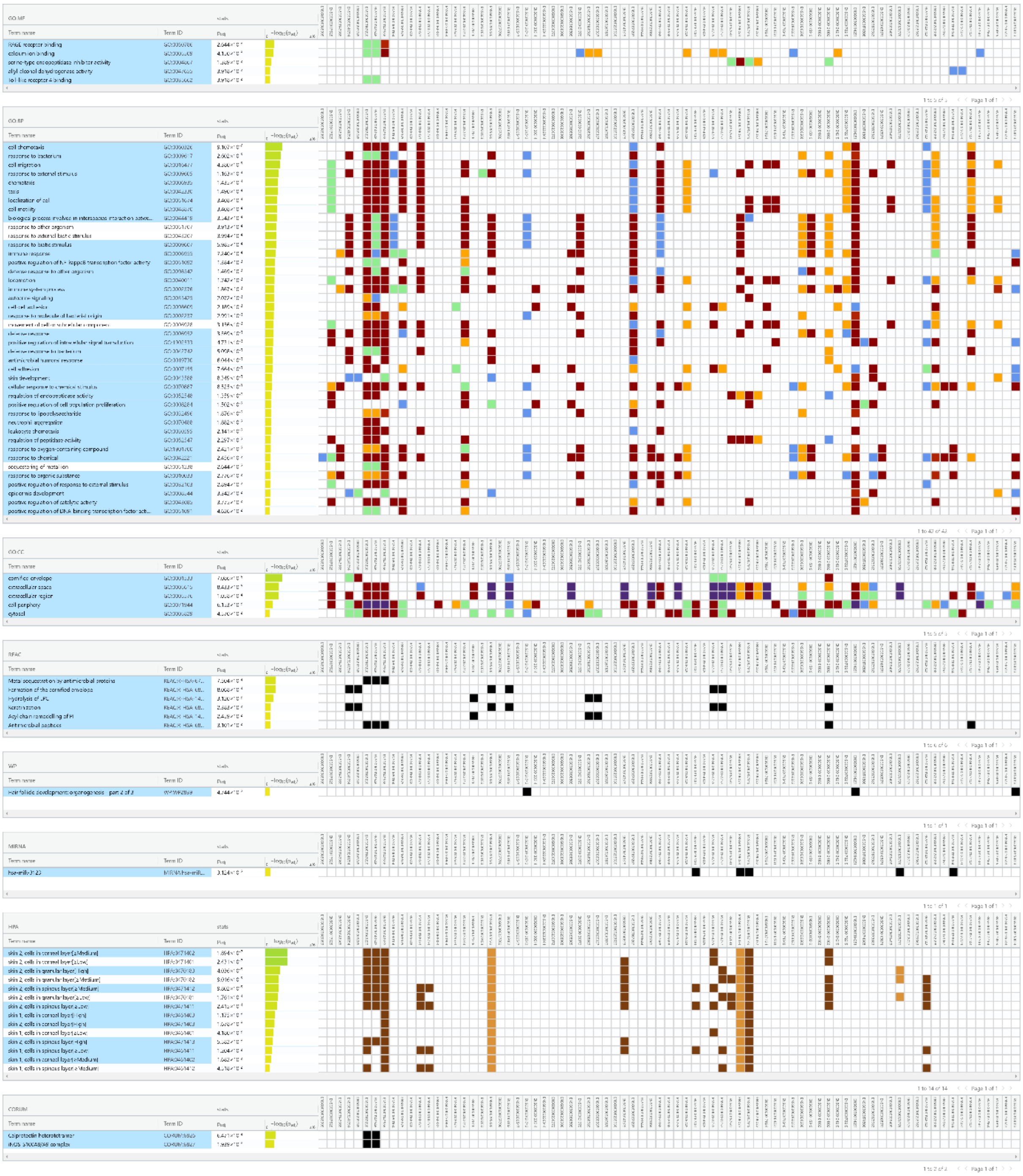
Supplementary fig. 10. LDA prediction on test data (left) and ROC curve (right).



Supplementary fig. 11. Lambda minimum based on binomial deviance (left) and MSE (right) obtained through cross validation at different folds.



Supplementary fig. 12. Log lambda of Lasso regression cross validated model (left) and the ROC curve (right) on test data prediction by the model.



GO:MF

Term name	Term ID	Padj
RAGE receptor binding	GO:0050786	2.644×10^{-3}
calcium ion binding	GO:0005509	4.130×10^{-3}
serine-type endopeptidase inhibitor activity	GO:0004867	1.389×10^{-2}
allyl-alcohol dehydrogenase activity	GO:0047655	3.918×10^{-2}
Toll-like receptor 4 binding	GO:0035662	3.918×10^{-2}

1 to 5 of 5 Page 1 of 1

GO:BP

Term name	Term ID	Padj
cell chemotaxis	GO:0060326	9.107×10^{-6}
response to bacterium	GO:0009617	2.602×10^{-5}
cell migration	GO:0016477	4.380×10^{-5}
response to external stimulus	GO:0009605	1.163×10^{-4}
chemotaxis	GO:0006935	1.443×10^{-4}
taxis	GO:0042330	1.490×10^{-4}
localization of cell	GO:0051674	3.408×10^{-4}
cell motility	GO:0048870	3.408×10^{-4}
biological process involved in interspecies interaction betwe...	GO:0044419	3.548×10^{-4}
response to other organism	GO:0051707	3.913×10^{-4}
response to external biotic stimulus	GO:0043207	3.994×10^{-4}
response to biotic stimulus	GO:0009607	5.985×10^{-4}
immune response	GO:0006955	7.240×10^{-4}
positive regulation of NF-kappaB transcription factor activity	GO:0051092	1.384×10^{-3}
defense response to other organism	GO:0098542	1.469×10^{-3}
locomotion	GO:0040011	1.742×10^{-3}
immune system process	GO:0002376	1.867×10^{-3}
autocrine signaling	GO:0035425	2.072×10^{-3}
cell-cell adhesion	GO:0098609	2.180×10^{-3}
response to molecule of bacterial origin	GO:0002237	2.991×10^{-3}
movement of cell or subcellular component	GO:0006928	3.156×10^{-3}
defense response	GO:0006952	3.560×10^{-3}
positive regulation of intracellular signal transduction	GO:1902533	4.731×10^{-3}
defense response to bacterium	GO:0042742	5.098×10^{-3}
antimicrobial humoral response	GO:0019730	6.044×10^{-3}
cell adhesion	GO:0007155	7.664×10^{-3}
skin development	GO:0043588	8.349×10^{-3}
cellular response to chemical stimulus	GO:0070887	8.525×10^{-3}
regulation of endopeptidase activity	GO:0052548	1.359×10^{-2}
positive regulation of cell population proliferation	GO:0008284	1.502×10^{-2}
response to lipopolysaccharide	GO:0032496	1.676×10^{-2}
neutrophil aggregation	GO:0070488	1.882×10^{-2}
leukocyte chemotaxis	GO:0030595	2.141×10^{-2}
regulation of peptidase activity	GO:0052547	2.297×10^{-2}
response to oxygen-containing compound	GO:1901700	2.421×10^{-2}
response to chemical	GO:0042221	2.456×10^{-2}
sequestering of metal ion	GO:0051238	2.644×10^{-2}
response to organic substance	GO:0010033	2.776×10^{-2}
positive regulation of response to external stimulus	GO:0032103	2.894×10^{-2}
epidermis development	GO:0008544	3.342×10^{-2}
positive regulation of catalytic activity	GO:0043085	3.773×10^{-2}
positive regulation of DNA-binding transcription factor acti...	GO:0051091	4.606×10^{-2}

1 to 42 of 42 Page 1 of 1

GO:CC

Term name	Term ID	Padj
cornified envelope	GO:0001533	7.086×10^{-5}
extracellular space	GO:0005615	8.433×10^{-5}
extracellular region	GO:0005576	1.058×10^{-4}
cell periphery	GO:0071944	6.133×10^{-3}
cytosol	GO:0005829	4.570×10^{-2}

1 to 5 of 5 Page 1 of 1

REAC

Term name	Term ID	Padj
Metal sequestration by antimicrobial proteins	REAC:R-HSA-67...	7.504×10^{-4}
Formation of the cornified envelope	REAC:R-HSA-68...	8.008×10^{-4}
Hydrolysis of LPC	REAC:R-HSA-14...	3.130×10^{-3}
Keratinization	REAC:R-HSA-68...	2.383×10^{-2}
Acyl chain remodelling of PI	REAC:R-HSA-14...	2.439×10^{-2}
Antimicrobial peptides	REAC:R-HSA-68...	3.101×10^{-2}

1 to 6 of 6 Page 1 of 1

WP

Term name	Term ID	Padj
Hair follicle development: organogenesis - part 2 of 3	WP:WP2839	4.744×10^{-2}

1 to 1 of 1 Page 1 of 1

MIRNA

Term name	Term ID	Padj
hsa-miR-3123	MIRNA:hsa-miR...	3.124×10^{-2}

1 to 1 of 1 Page 1 of 1

HPA

Term name	Term ID	Padj
skin 2; cells in cornified layer[≥Medium]	HPA:0471402	1.894×10^{-7}
skin 2; cells in cornified layer[≥Low]	HPA:0471401	2.431×10^{-7}
skin 2; cells in granular layer[High]	HPA:0470183	4.036×10^{-6}
skin 2; cells in granular layer[≥Medium]	HPA:0470182	9.016×10^{-6}
skin 2; cells in spinous layer[≥Medium]	HPA:0471412	9.802×10^{-6}
skin 2; cells in granular layer[≥Low]	HPA:0470181	1.761×10^{-5}
skin 2; cells in spinous layer[≥Low]	HPA:0471411	2.415×10^{-5}
skin 1; cells in cornified layer[High]	HPA:0461403	1.175×10^{-4}
skin 2; cells in cornified layer[High]	HPA:0471403	1.678×10^{-4}
skin 1; cells in cornified layer[≥Low]	HPA:0461401	4.180×10^{-4}
skin 2; cells in spinous layer[High]	HPA:0471413	5.383×10^{-4}
skin 1; cells in spinous layer[≥Low]	HPA:0461411	1.204×10^{-3}
skin 1; cells in cornified layer[≥Medium]	HPA:0461402	1.682×10^{-3}
skin 1; cells in spinous layer[≥Medium]	HPA:0461412	4.318×10^{-3}

1 to 14 of 14 Page 1 of 1

CORUM

Term name	Term ID	Padj
Calprotectin heterotetramer	CORUM:6925	6.471×10^{-4}
iNOS-S100A8/A9 complex	CORUM:6927	1.939×10^{-3}

1 to 2 of 2 Page 1 of 1

Supplementary fig. 13. The functional enrichment profiling of 79 genes and their involvement in different pathways individually or in groups.