

University of Crete
Faculty of Applied and Technological Sciences, Department of Biology
School of Medicine
Institute of Molecular Biology & Biotechnology – Foundation for
Research and Technology
Computational Genomics Group

Graduate Programme in Molecular Biology & Biomedicine
Master thesis
Vasilis Ntasis

**COMPREHENSIVE STUDY OF GENE
EXPRESSION IN THE PERIPHERAL BLOOD OF
SYSTEMIC LUPUS ERYTHEMATOSUS (SLE)
PATIENTS**

Supervisor: Ass. Prof. Christoforos Nikolaou



19/09/2018

Table of Contents

1	ABSTRACT.....	1
2	INTRODUCTION.....	1
3	METHODS.....	4
3.1	SAMPLE COLLECTION AND RNA SEQUENCING.....	4
3.2	FRAGMENT SUMMARIZATION.....	4
3.3	GENE FILTERING.....	4
3.4	DIFFERENTIAL EXPRESSION ANALYSIS.....	5
3.5	FUNCTIONAL AND MODULAR ANALYSIS.....	5
3.5.1	<i>Functional Analysis with gProfileR.....</i>	<i>5</i>
3.5.2	<i>Modular analysis.....</i>	<i>6</i>
3.6	WEIGHTED GENE CO-EXPRESSION NETWORK ANALYSIS.....	7
3.7	TOPOLOGICAL ANALYSIS.....	7
3.7.1	<i>Normalization.....</i>	<i>7</i>
3.7.2	<i>Coordinates.....</i>	<i>7</i>
3.7.3	<i>Bin creation and Bin count calculation.....</i>	<i>7</i>
3.7.4	<i>Correlation matrix calculation.....</i>	<i>8</i>
3.7.5	<i>Permutations.....</i>	<i>8</i>
3.7.6	<i>COD definition and detection.....</i>	<i>9</i>
3.7.7	<i>COD comparison.....</i>	<i>11</i>
3.7.7.1	<i>Metrics.....</i>	<i>11</i>
3.7.7.2	<i>COD reorganization categories.....</i>	<i>12</i>
3.7.8	<i>Evaluate DEG inclusion in CODs.....</i>	<i>12</i>
3.8	CELL TYPE ESTIMATION AND ENTROPY ESTIMATION.....	13
3.9	SOURCE CODE.....	13
3.10	GRAPHS.....	14
4	RESULTS.....	14
4.1	DIFFERENTIAL EXPRESSION AND FUNCTIONAL ANALYSIS.....	14
4.2	MODULAR ANALYSIS.....	48
4.3	WGCNA ANALYSIS.....	52
4.4	TOPOLOGICAL ANALYSIS.....	54
4.4.1	<i>COD profiling.....</i>	<i>54</i>
4.4.2	<i>Structural COD changes.....</i>	<i>59</i>
4.4.3	<i>Interesting cases.....</i>	<i>62</i>
5	DISCUSSION.....	72
6	BIBLIOGRAPHY.....	73

List of Figures

Figure 1: Detailed information concerning SLE patient cohort.....	4
Figure 2: Histogram depicting the distribution of SLEDAI index in the studied patient cohort, and the implemented splitting based on that.....	6
Figure 3: The procedure of chromosomal bin count and correlation matrices calculation.....	8
Figure 4: Flowchart illustrating the reasoning of the COD detection algorithm.....	10
Figure 5: Boxplots of intra-COD average correlation values of CODs identified using the corresponding window size value indicated at x axis.....	11
Figure 6: Volcano plots illustrating differential expression resulted from different group comparisons.....	20
Figure 7: Treemaps depicting Gene Ontology terms enriched in the DEG list of SLE vs Healthy.....	23
Figure 8: Heatmaps illustrating Log ₂ FC values of DEGs derived from different comparisons. A. Patients have been grouped according to the state and the manifestations of the disease. B. Patients have been grouped according to disease activity. Genes depicted in each heatmap are differentially expressed in at least one of the comparisons indicated at x axis. Trees in each heatmap depict the results of hierarchical clustering.....	27
Figure 9: Volcano plots illustrating differential expression resulted from different group comparisons.....	40
Figure 10: Modular analysis for the differential expression of SLE patients. Modules are represented by pies. The proportion of module genes, which have been detected as DEGs, is indicated by the coloured portion of the pies. Red indicates overexpression and blue indicates underexpression.....	49
Figure 11: Modular analysis for the differential expression of DA groups. Modules are represented by pies. The proportion of module genes, which have been detected as DEGs, is indicated by the coloured portion of the pies. Red indicates overexpression and blue indicates underexpression.....	51
Figure 12: Heatmap representation of the correlations between WGCNA modules and a variety of traits recorded from all the patients. Red indicates positive correlation and green indicates negative correlation.....	52
Figure 13: Barplot depicting the total number of CODs detected in control healthy group and in the three DA patient groups.....	54
Figure 14: Barplot depicting the total number of CODs per chromosome detected in control healthy group and in the three DA patient groups.....	55
Figure 15: Violin plots illustrating the estimated distribution of COD sizes in each group, classic boxplots are included, the scale is logarithmic (log(bp)) and the results of Mann-Whitney-Wilcoxon tests comparing each one of patient groups with healthy group are represented by p-values.....	56
Figure 16: Average intra COD correlation estimated distribution summarized in violin plots. Classic boxplots are included and the results of Mann-Whitney-Wilcoxon tests comparing each one of patient groups with healthy group are represented by p-values.....	56
Figure 17: Average inter COD correlation estimated distribution summarized in violin plots. Classic boxplots are included and the results of Mann-Whitney-Wilcoxon tests comparing each one of patient groups with healthy group are represented by p-values.....	57
Figure 18: Violin plots illustrating the estimated distribution of chromosomal bins percentage, in which expressed genes are located and that reside inside CODs, classic boxplots are included and the results of Mann-Whitney-Wilcoxon tests comparing each one of patient groups with healthy group are represented by p-values.....	58
Figure 19: Violin plot representation of the entropy calculated using the different immune cell proportions per individual. Classic boxplots are included and the results of Mann-Whitney-Wilcoxon tests comparing each one of patient groups with healthy group are represented by p-values.....	59
Figure 20: Heatmap highlighting the number of CODs in healthy group, that are altered in a particular way when the corresponding genomic area is examined in	

patient groups, a normalization per change category (row) has been performed and change categories have been grouped based on a hierarchical clustering.....	61
Figure 21: A. Violin plots demonstrating the estimated distribution of a per chromosome distance score calculated for CODs in each patient group compared to CODs in healthy group. B. Violin plots demonstrating the estimated distribution of a per chromosome distance score calculated for randomized CODs of each patient group compared to CODs in healthy group.....	64
Figure 22: Violin plots demonstrating the estimated distribution of Jaccard similarity coefficients calculated for CODs in each patient group compared to CODs in healthy group.....	65
Figure 23: Barplots exhibiting the proportion of DEGs, derived from the comparison indicated at the horizontal axis, that reside inside CODs of the group indicated by colour code. Bootstrap p-values demonstrate statistical significance of the difference in DEG inclusion of each patient group and healthy control group.....	66
Figure 24: Cases where COD alterations, like a split (A) or a depletion (B), are associated with genes that are differentially expressed, when patient and healthy groups are compared.....	67
Figure 25: Cases where organization of gene expression alters in an almost gradient-like manner. In chromosome 14 a border emerges, which separates the IgH locus and the upstream neighbourhood (A), in chromosome 2 a border extension encloses the IgK locus (B), and in chromosome 22 CODs are merged and expanded encompassing that way more IgL genes.....	71
Figure 26: Variant rs1734787 in chromosome X is co-localized with a COD split region.....	71

List of Tables

Table 1: Description of the different categories of COD structural changes.....	13
Table 2: KEGG and Reactome pathways enriched in overexpressed (A) and underexpressed (B) genes in SLE patients.....	25
Table 3: KEGG and Reactome pathways enriched in overexpressed and underexpressed genes of DA1 (A,B), DA2 (C,D), DA3 (E,F) and pathways enriched in the overexpressed genes resulted from DA3 vs DA1 (G) and DA3 vs DA2 (H).....	33
Table 4: KEGG and Reactome pathways enriched in overexpressed and underexpressed genes of Active-Renal (A,B), Active-nonRenal (C,D), Inactive (E,F), pathways enriched in the overexpressed genes resulted from Active-Renal vs Active-nonRenal (G) and Active-Renal vs Inactive (H), and pathways enriched in the underexpressed genes resulted from Active-nonRenal vs Inactive (I).....	47

List of Abbreviations and Symbols

cpm	count per million
RLE	Relative log expression
SLE	Systemic Lupus Erythematosus
SLEDAI	SLE disease activity index
\log_2FC	base 2 logarithm of the fold change
DEGs	differentially expressed genes
KEGG	Kyoto Encyclopedia of Genes and Genomes
WGCNA	Weighted gene co-expression network analysis
COD	Coexpression domain
DA	Disease activity
GO	Gene Ontology
SNP	Single Nucleotide Polymorphism

1 Abstract

Systemic lupus erythematosus (SLE) is a complex autoimmune disease with patients presenting varying levels of disease activity and diverse clinical manifestations. Better comprehension of the molecular mechanisms that underlie the basis of the pathology, will promote a better diagnosis, prognosis and treatment of patients. In this study, RNASeq data from peripheral blood samples of 142 SLE patients and 58 healthy individuals were compared from multiple perspectives. A number of different analyses were implemented. Already known and novel molecular signatures were identified as differentially deregulated and associated with disease activity and/or renal manifestations of the disease. Among the most prominent of them were cell cycle, interferon and plasmablast signatures. Moreover, topological organization of gene expression was extensively studied. Genomic coexpression domains (CODs) were detected in patient subgroups and in healthy control group. Results suggest a more ‘fragmented’ topological profile for patient gene expression. At the same time, differences in the size and distribution of CODs were observed between different patient subgroups, suggesting a link between gene expression organization and disease development. Cross-correlation of the defined genomic regions with genetic data is likely to uncover probable origins of the gene expression aberrations associated with SLE progression.

2 Introduction

Systemic lupus erythematosus (SLE) is a multifaceted autoimmune disorder^{1,2}. It is highly heterogeneous and considered a prototype of the systemic autoimmune diseases. The aetiology of the disease includes the contribution of genetic, epigenetic, environmental and stochastic factors. Notably, SLE is incurable and may be life-threatening, with clinical manifestations involving essential organs and tissues, such as kidney, brain and blood. Moreover, SLE predominantly affects young female individuals and is characterized by an unpredictable disease course with flares interspersed among periods of remission. Patients are characterized by the production of autoantibodies, including antinuclear antibodies (ANA) and antiphospholipid antibodies (APA), impaired clearance of apoptotic debris and the formation of large amounts of immune complexes, that aggregate in tissues leading to damage. Accumulation of damage stems from the aforementioned progression, side-effects of treatment and comorbid conditions. In spite of the numerous studies concerning SLE, there are considerable unmet needs related to diagnosis, prognosis and therapy development. Thus, better comprehension of the disorder in a molecular level is required.

Several studies have been performed to investigate the transcriptional profile of SLE patients (reviewed in ^{3,4}) using high-throughput techniques. Early studies used

whole blood and microarray analysis to identify the so-called interferon (IFN) signature^{5,6}. Inflammatory and granulocyte signatures were also observed⁶. Subsequently, efforts focused on determining signatures in specific cell types and associating them with specific patient subgroups, in order to stratify patients and facilitate therapeutic targeting. However, some of them lack statistical power due to small sample sizes. Furthermore, it seems that there is a great diversity in the expression levels of specific signatures upon examination of different cell types or individuals of varying ancestry. These limitations aside, the high heterogeneity of the disorder ‘demands’ the stratification of patients. Additionally, there are recent studies that even follow a personalized strategy^{7,8}.

Besides expected functions and pathways related to the immune system (such as the IFN and granulocyte signatures mentioned above), metabolic pathways and oxidative stress functionalities have also been associated with the disease⁹. A plausible explanation stems from the fact that a physiological immune response is linked to metabolism. In this view, a recent line of research has been stably developing towards the definition of SLE-specific metabolic signatures, which could be linked to epigenetic¹⁰ or epitranscriptomic¹¹ abnormalities.

Research concerning the epigenetic landscape in SLE is also abundant (reviewed in ^{4,12}). Epigenetic mechanisms, that have been connected with SLE comprise DNA methylation, post-translational histone modifications, regulation by non-coding RNAs and, more recently, DNA hydroxymethylation. Aberrations of those mechanisms affect gene expression. More specifically, there is extensive DNA hypomethylation in the genome of T-cells from patients with active SLE, which leads to overexpression of various genes including those associated with the IFN signature¹³. Correspondingly, drugs that exhibit DNA methylation inhibitory activity are known to induce SLE-like features¹⁴. On the other hand, 5-hydroxymethylcytosine levels are higher in T-cells of SLE patients and are also positively correlated with increased gene expression¹⁵. An example of a post-translational histone modification associated with lupus is histone H4 hyperacetylation. Histone H4 hyperacetylation was detected throughout the genome of monocytes from lupus patients¹⁶. Lastly, there are studies that implicate the activity of diverse microRNA molecules with SLE pathogenesis. Altered microRNA expression have been detected in peripheral blood mononuclear cells, renal tissue and in the plasma from lupus patients. Some of them seem to influence significant to the disease processes, such as TLR signalling and IFN induced genes expression^{17,18}.

An interesting aspect of transcriptional regulation that is becoming increasingly relevant in light of recent technological advances is its relationship with genome structure¹⁹. In this respect, genomic organization may play a critical part in the disease and its exploration could be of great assistance for the comprehension of the pathogenesis. There are different approaches to define and study genome organization. For instance, one can study 3D chromatin interactions. The use of chromatin conformation capture assays facilitates the investigation of intra and inter-chromosomal contacts and the dis-

covery of ‘territories’, where the contact frequency is higher than average. However, there are no known studies exploiting contact data in the framework of lupus. Chromatin is also organized into ‘open’ and ‘closed’ regions. Accessibility assays can be used to identify those regions. Scharer and Blalock et al applied the Assay for Transposase Accessible Sequencing (ATAC-seq) to explore the accessibility landscape of naïve B cells from biobanked specimens of lupus patients under flare status²⁰. They detected alterations of genome accessibility located at regions encompassing genes related to B cell activation.

Another approach to study genomic architecture is to explore topological organization of gene expression. It is known that gene order is not random in eukaryotes and that genes with similar expression profiles tend to be clustered within the same genomic region²¹. In yeast, it has been demonstrated by studying differential expression upon topological stress that genes are organized in topologically co-regulated clusters²². In human, Soler-Oliva et al defined genomic coexpression domains (CODs) based on the correlation of gene expression levels in breast cancer and healthy specimens and tried to associate them with contact data²³. CODs are representative of the total expression coordination and hence it would be informative to study their aberrations in complex disorders, such as SLE.

In this study, gene expression was analysed in a big dataset, derived from whole blood RNA sequencing, of 142 SLE patients of varying levels of disease activity (DA) and diverse clinical traits (Figure 1), compared to 58 control healthy individuals. Several different analyses were applied, including differential expression, functional analysis and weighted gene co-expression network analysis, to explore the transcriptional profile of patients and correlate it with DA and clinical manifestations. At the level of genome organization, we devised a robust computational pipeline and used it to define, detect and compare CODs in different patient subgroups and the group of healthy individuals.

SLE characteristics	<i>n</i> = 148
Females	84%
Ethnicity (Caucasian)	99%
Age (years)	40 ± 14
No. ACR criteria	5.3 ± 1.5
Physician Global Assessment	
Inactive	32%
Mild activity	12%
Moderate/high activity	56%
Clinical SLEDAI-2K	
0	32%
1 – 5	23%
6 – 10	30%
>10	15%

Actively involved organs/domains	
General/constitutional	18%
Mucocutaneous	48%
Neurological	11%
Musculoskeletal	37%
Cardiorespiratory	6%
Vasculitis (skin/GI)	1%
Renal	24%
Hematology	18%

Figure 1: Detailed information concerning SLE patient cohort

3 Methods

3.1 Sample collection and RNA sequencing

Sample collection, RNA sequencing and mapping had already been performed. Detailed information regarding the patient cohort, sample collection, RNA sequencing, mapping and quality control are described by Panousis et al²⁴. Thus, the starting material of this work were the bam alignment files, produced by the mapping procedure.

3.2 Fragment summarization

We used FeatureCounts²⁵ to extract raw counts and quantify expression levels for a comprehensive set of human genes, as compiled under the latest GENCODE annotation v15²⁶. A fragment was counted in case of any overlap with an exon feature and the

counts were grouped based on the ‘gene_name’ attribute of the annotation entities. Only fragments with both ends successfully mapped were considered for summarization. Fragments that were chimeric, overlapping multiple metafeatures (genes), not uniquely mapped, or having any read marked as duplicate were discarded.

3.3 Gene filtering

The initial number of genes included in raw count table was 51716. A multi-step filtering approach was adopted. At first, the ‘type’ of each gene was extracted from the annotation GTF file used in fragment summarization. Then, genes belonging to any of the following types were filtered out: ‘pseudogene’, ‘processed transcript’, ‘poly-

morphic pseudogene', 'antisense', 'sense intronic', 'sense overlapping', 'IG_V pseudogene', 'IG_C pseudogene', 'TR_V pseudogene', 'TR_J pseudogene', 'IG_J pseudogene', 'non_coding', 'Mt-tRNA' and 'Mt-rRNA'. The total number of genes belonging to those categories were 20190. Subsequently, 167 genes, which had multiple entries in the annotation file, with the same 'gene_name', but different chromosome attribute, and that could generate errors in the fragment summarization process were removed from our dataset as well. . Lastly, genes with mean CPM value, in all samples, lower than 0.05 were also filtered out, though that was not applied for the topological analysis. The number of the remaining genes in the dataset were 18447.

3.4 Differential expression analysis

Differentially expressed genes (DEGs) were called with the implementation of MD-Seq²⁷. Firstly, raw counts were normalized using relative log expression (RLE)²⁸. Afterwards, a design matrix was constructed based on the groups to be compared. For the simple analysis, healthy individuals were the control group and SLE patients comprised the test group. For the activity analysis, the patient cohort was separated into three groups according to SLE activity index²⁹ (SLEDAI, Figure 2), low disease activity group (DA1, SLEDAI < 3), medium activity group (DA2, 2 < SLEDAI < 9) and high activity group (DA3, SLEDAI > 8). Furthermore, gender and any drug treatment have been taken under consideration as covariates in subsequent analyses. Finally, there was a third grouping relative to the manifestation of the disease. According to the status of the disease (Active or Inactive), and if a patient had any renal manifestation or not, the patients were split into three groups. Those are the inactive group, the renal active group and the non-renal active group. Gender and treatment have also been taken under consideration here. Last, for any significance assessment corrected p-value (q-value) and the base-2 logarithm of the fold change (\log_2FC) were considered. Genes were considered DEGs if they had a q-value lower or equal to 0.05 and an absolute \log_2FC value greater or equal to 0.5.

3.5 Functional and Modular analysis

To functionally interpret the results of the differential expression two different approaches were followed.

3.5.1 Functional Analysis with gProfileR

Functional Analysis was performed with the use of gProfileR³⁰, a tool that has access to data from different databases, and performs hypergeometric test and correction for multiple testing to find statistically significantly enriched functional ontologies in the provided gene lists. That analysis was restricted to pathways from Gene Ontology³¹, KEGG³² and Reactome databases³³ and was performed separately for overexpressed and underexpressed DEGs. Only pathways with q-value lower or equal to 0.05 were considered.

3.5.2 Modular analysis

Modular analysis was performed with the `tmod`³⁴ R package. `tmod` takes a pre-ranked gene list as input and implements a gene set enrichment analysis. In other words, it tries to find the gene sets, whose relative expression values tend to cluster towards the top (or bottom) of the ranked list in a way similar to Gene Set Enrichment Analysis (GSEA) methods³⁵. What differentiates `tmod` from standard GSEA approaches is that it implements the analysis of pre-defined, built-in gene sets (or modules as they are referred to), hence it performs a modular analysis. These modules are constructed and annotated in studies^{36,37}, relevant to blood tissue and immunity. Thus, they are more specific and applicable to this study than other general gene sets. The gene list (not only DEGs) was ranked according to absolute \log_2FC value in a decreasing manner. Subsequently, a ‘`tmodCERNOtest`’ was executed for all the built-in modules. From the derived statistically significant modules, only those with a minimal percentage of 15% being DEGs were retained.

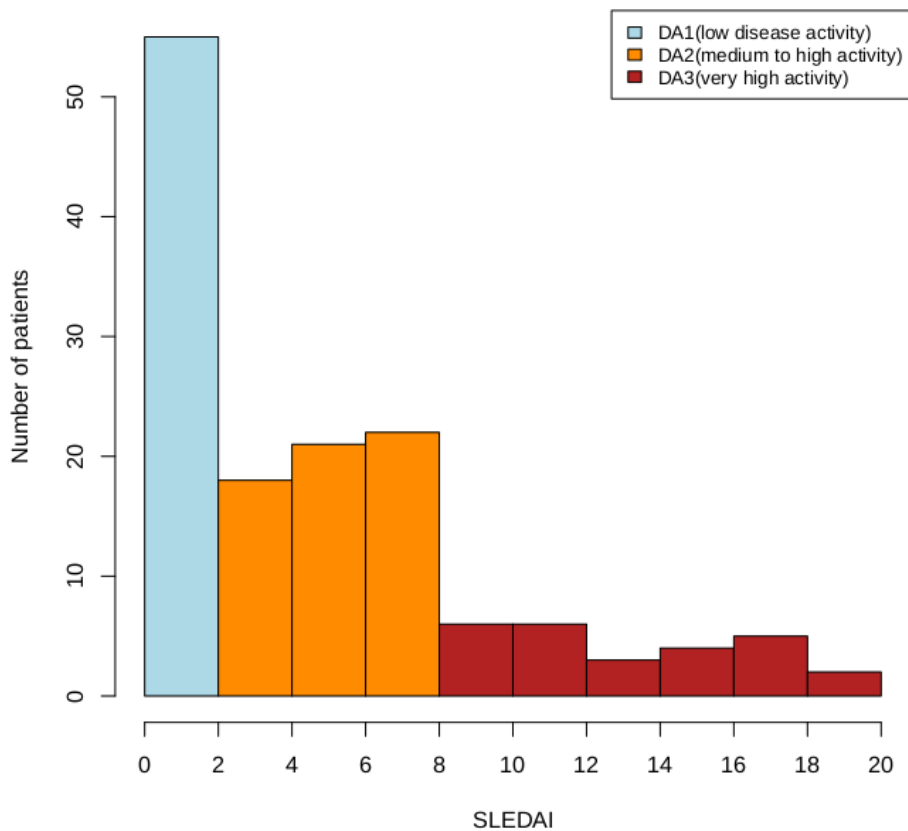


Figure 2: Histogram depicting the distribution of SLEDAI index in the studied patient cohort, and the implemented splitting based on that.

3.6 Weighted gene co-expression network analysis

To detect genes which are coexpressed in the patient cohort and form discrete modules, weighted gene co-expression network analysis (WGCNA)³⁸ was utilized. Raw counts were normalized using RLE²⁸ followed by normalization for gene length. Normalized gene counts of all patients were fed to the WGCNA algorithm. Dichotomizing information and hard-thresholding may result in information loss. The continuous nature of the co-expression information can be preserved by implementing soft thresholding. Particularly, WGCNA uses a power function, where the value of the power is the soft thresholding parameter. Here, the value assigned to that parameter was 10 according to scale free topology criterion, which amounts to choosing the lowest value of such that approximate scale free topology is reached. The resulting gene network was unassigned, meaning that the absolute value of correlation was used as a similarity measure. Modules were extracted based on a hierarchical clustering of the topology overlap matrix. Those initial modules were then merged using hierarchical clustering of their eigengene and by cutting the resulted tree at the height of 0.25. Pathway enrichment of the module genes was performed using gProfileR. Finally, the association of the modules with a list of clinical traits was estimated by calculating the Pearson correlations of the module eigengenes with the corresponding traits. Student asymptotic p-values were calculated to statistically assess the correlation values.

3.7 Topological analysis

3.7.1 Normalization

Normalized counts were utilized. A two step normalization was implemented on raw counts (filtered for the different irrelevant gene types), using RLE²⁸ followed by normalization for gene length.

3.7.2 Coordinates

Gene coordinates were isolated from the annotation file (gencode.v15²⁶). In the annotation file, there were some entries having the same 'gene_name', but different 'gene_id' and different chromosome attributes. For the subsequent analysis, those genes were discarded (167 genes).

3.7.3 Bin creation and Bin count calculation

Each chromosome was split in 10kb bins, starting from the start of the first gene, till the end of the last gene (Figure 3). Thus, there is a possibility for the last bin in each chromosome to be smaller than 10kb. To each bin were attributed the genes, which start inside the corresponding bin. Subsequently, using the normalized counts of genes belonging to a bin, the mean count was calculated for each individual. So, from

a matrix of gene normalized counts for each individual, a matrix of chromosomal bin counts for each individual was constructed (Figure 3).

3.7.4 Correlation matrix calculation

Using the bin counts, for each chromosomal bin the Spearman correlation coefficient was calculated in regard to each one of the rest of the bins that resided in the same chromosome (Figure 3). Chromosomal bins with zero expression were ignored for the rest of the analysis. This procedure produced a square correlation matrix for each chromosome.

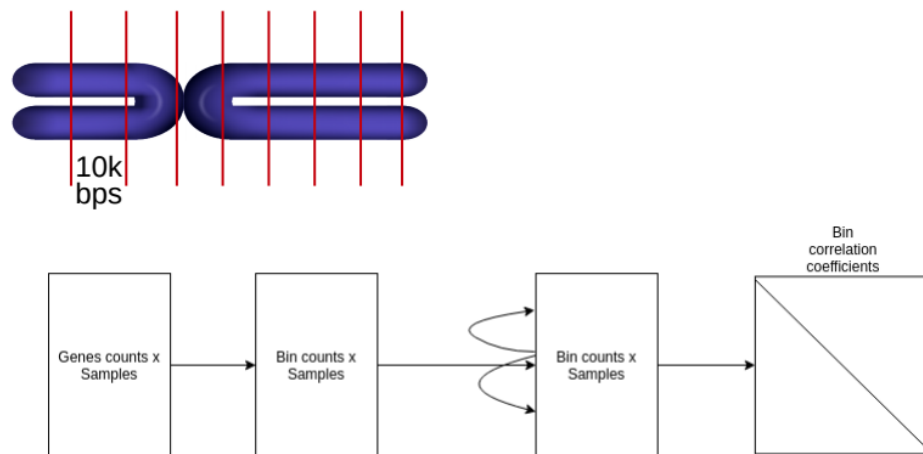


Figure 3: The procedure of chromosomal bin count and correlation matrices calculation

3.7.5 Permutations

To statistically evaluate the correlation coefficients, a Monte Carlo like approach was implemented. The bin counts, of each individual separately, were shuffled randomly and afterwards the correlation matrix was re-constructed. That procedure was repeated 1000 times for each chromosome. In every iteration the calculated correlation coefficients were compared to the actual (observed) correlation coefficients, that were calculated using the intact bin counts. P-value for each coefficient is equal to the fraction of those 1000 permutations, in which the corresponding coefficient had the same or more extreme value compared to the actual one. The correlation coefficients with p-value greater than 0.05 were discarded from the analysis (turned into 0s).

3.7.6 COD definition and detection

For the definition of coexpression domains (CODs), we followed a similar approach to the one suggested by Soler-Oliva et al²³. That analysis is influenced by methods designed for topologically associated domain identification. Roughly, CODs are defined as genomic regions of consecutive (having filtered outbins with zero expression) chromosomal bins with higher than average correlation among them, delimited by statistically significant boundaries. More specifically, COD detection is a two step procedure. First, for each bin an average correlation signal between its upstream and downstream regions (in a specified window) is computed. The exact formula for the binsignal calculation is as follows

$$\text{binsignal}(i) = 1/w^2 \cdot \sum_{l=1}^w \sum_{m=1}^w \text{correlation}(U_i(l), D_i(m))$$

where $U_i = \{i-w+1, \dots, i-1, i\}$, $D_i = \{i+1, i+2, \dots, i+w\}$, and w is the size of the window around i .

Subsequently, binsignal is used to infer CODs. CODs are detected sequentially as the algorithm reads the vector of binsignals. The complete reasoning of the implemented algorithm is represented in the flowchart depicted in Figure 4. So, CODs are regions containing bins with binsignal greater or equal to 0.25, the average genome binsignal of the healthy group. However, they can contain 2 bins (at maximum) with binsignal lower than 0.25, given that these bins are not statistically significant boundaries, in order to fuse small neighbouring CODs. The statistically significant boundaries are determined by computing for each bin a Student's t-test, between upstream and downstream binsignal values in the same window used for the binsignal calculation. By observing the flowchart, one can understand that the right boundary of the CODs produced by that algorithm is not necessarily statistically significant, because it is determined by the decrease in binsignal value but the low p-value is not required. For that reason, an additional filtering was implemented, so that those CODs without a significant right boundary are excluded.

The window size used in binsignal calculation and in COD detection was three bins. That choice was based on the average intra-COD correlation of chromosome 1 of the healthy group. That was computed utilizing a range of values (three to twenty) for window size. Setting window size equal to three produced the higher average intra-COD correlation (Figure 5).

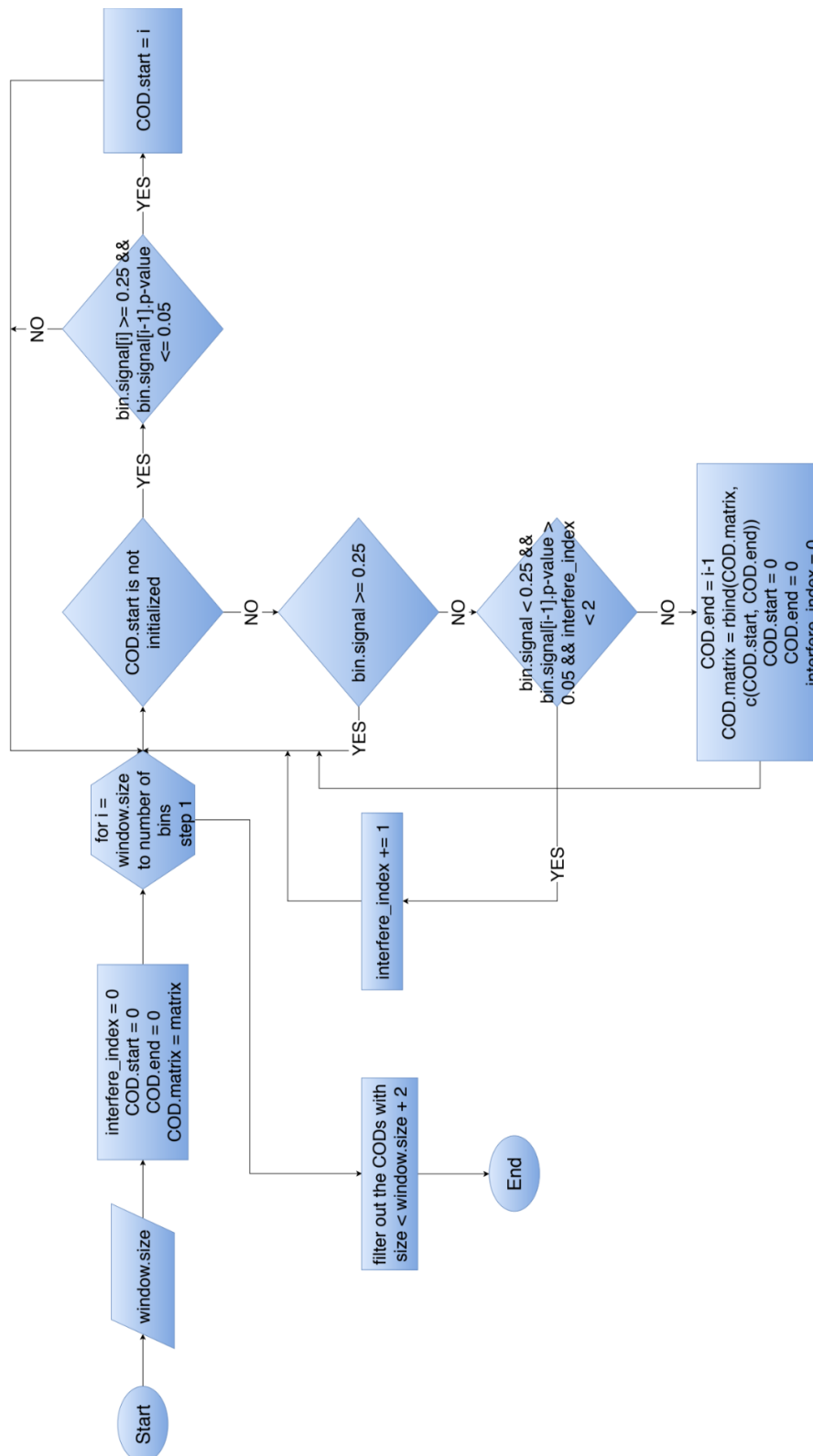


Figure 4: Flowchart illustrating the reasoning of the COD detection algorithm

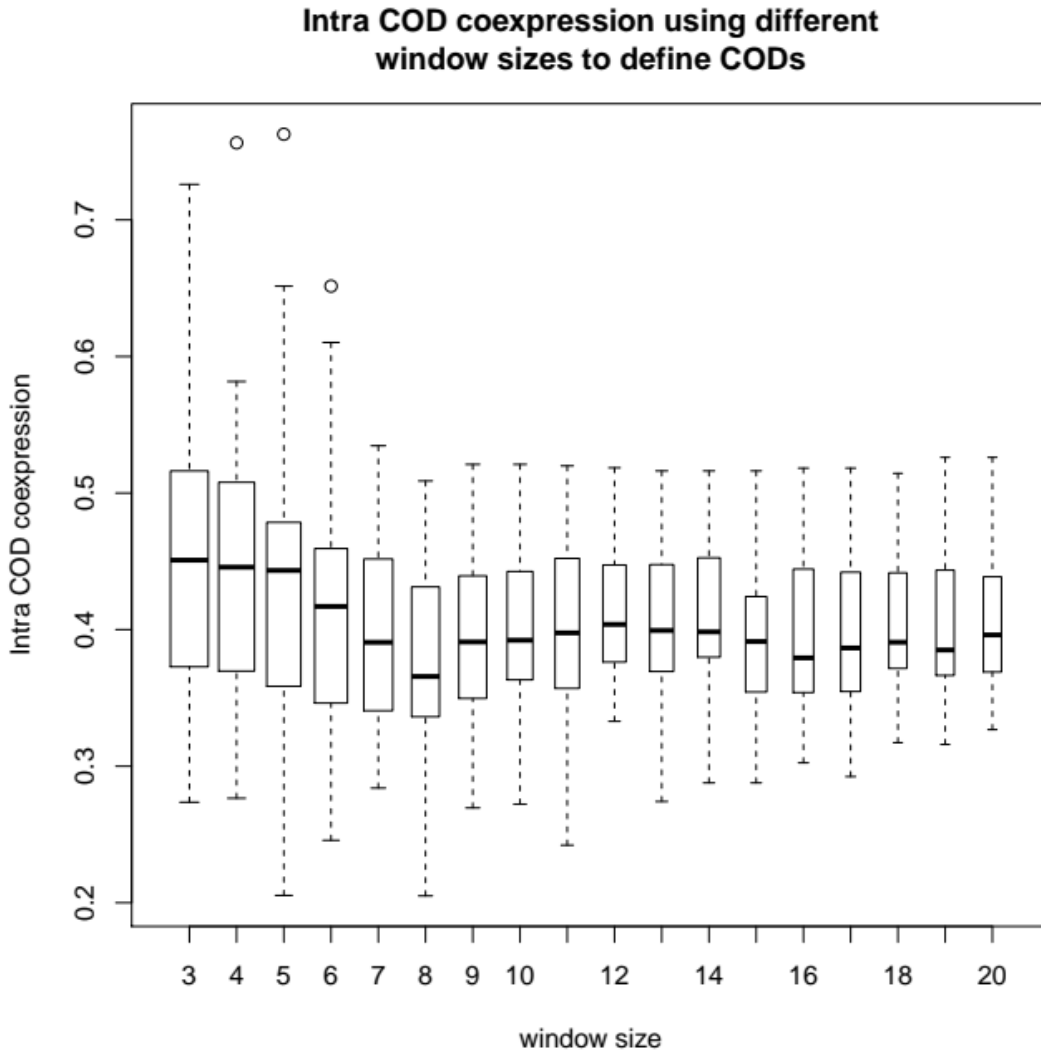


Figure 5: Boxplots of intra-COD average correlation values of CODs identified using the corresponding window size value indicated at x axis.

3.7.7 COD comparison

3.7.7.1 Metrics

Two different metrics were applied to explore the differences of COD sets of different groups. CODs were handled as a set of chromosomal intervals. The first metric used was the Jaccard similarity coefficient³⁹. COD pairs between two different groups (e.g. healthy and patient groups) with chromosomal coordinates that overlap were detected. For every pair the Jaccard index was calculated. The second metric used was the BP distance score⁴⁰. BPscore is more versatile and the authors of the study, which introduced it, recommend it as a more appropriate measure for comparisons between topologically associated domains, as it takes into account the relative chromosome

size. The formula used to calculate BPscore was slightly altered compared to the original work to better serve the purposes of the current study.

$$BP(A, B) = 1 - 1 / \sum_i^{|o|} \max(l(f_{A,B}(i)), l(f_{B,A}(i))) \cdot \sum_i^{|o|} \left(\frac{l(o[i])^2}{\max(l(f_{A,B}(i)), l(f_{B,A}(i)))} \right)$$

where A and B are two sets of CODs located in the same chromosome but of different groups, o is a vector containing all the intersections between A and B, $f_{A,B}(i)$ is a function mapping $o[i]$ to the exact COD in A, which induced $o[i]$ and $l(o[i])$ is a function mapping $o[i]$ to its length.

3.7.7.2 COD reorganization categories

When comparing two sets of CODs, a COD in one set can be further categorized based on its overlap ‘status’ against the other set (Table 1). In case a COD in the healthy group has no overlap with CODs of a patient group, it is referred to as ‘depleted’ from the relative patient group. On the other hand, if a COD present in a patient group has no overlap with CODs in the healthy group it is called ‘emerged’. If a COD in healthy group has any overlap with more than one CODs from a patient group, then it is categorized as ‘split’. In the opposite situation, where a COD of a patient group has more than one corresponding CODs in healthy group, these are assigned to the ‘merged’ category. Additionally, in case a COD pair has identical coordinates, it is characterised as ‘intact’. Finally, the remaining uncharacterised COD pairs are categorized based on which of the two borders (or even both) have been shifted.

3.7.8 Evaluate DEG inclusion in CODs

The proportion of DEGs, derived from a specific comparison, whose start reside inside a COD of the corresponding groups was calculated. For instance, inclusion of DEGs, emanated from a comparison of DA3 and healthy groups, was determined for DA3 and healthy CODs.

A bootstrap approach was adopted to statistically evaluate the difference of DEG inclusion in healthy and patient COD sets. A random gene sample, of same size with the corresponding DEG set and taking under consideration chromosomal gene density, was selected. For that gene set, COD inclusion was calculated as described above followed by the calculation of ratio X. X is equal to the inclusion in patient COD set over inclusion in healthy COD set. The described procedure was repeated 10000 times. The bootstrap p-value is equal to the amount of repeats, in which X was equal or more extreme than the corresponding ratio computed using the real DEG set, over the total amount of repeats.

COD reorganization category	Presence in healthy COD set	Presence in patient COD set	Overlap between a COD and a test COD set	Border possibly sifted
Depleted	+	-	No overlap	—
Emerged	-	+	No overlap	—
Split	+	+	Healthy COD overlaps with multiple patient CODs	Right or left or both or none
Merged	+	+	Patient COD overlaps with multiple healthy CODs	Right or left or both or none
Intact	+	+	There is 100% overlap between a healthy and a patient COD	none
Right border	+	+	0% < overlap < 100%	Right
Left border	+	+	0% < overlap < 100%	Left
Both Borders	+	+	0% < overlap < 100%	Both

Table 1: Description of the different categories of COD structural changes

3.8 Cell type estimation and entropy estimation

CIBERSORT⁴¹ was utilized to estimate the proportion of different immune cell types in whole blood. That analysis was performed by Panousis et al²⁴. Shannon entropy⁴² was used as a metric, in order to assess the variability/uncertainty in the proportions of the different cells types between healthy and SLE subjects.

$$H = - \sum_{i=1}^n p(x_i) \cdot \log_2(p(x_i))$$

where H is the Shannon (information) entropy, $p(x_i)$ is the estimated proportion of x_i cell type in whole blood and n is the total number of estimated cell types. Entropy was calculated for every individual in the dataset. Subsequently, the difference between the distribution of entropies of healthy and SLE groups were statistically evaluated by a non parametric Wilcoxon–Mann–Whitney test⁴³.

3.9 Source code

Most of the described analysis was implemented in the R programming language⁴⁴. Source code for any of the aforementioned pipelines is available upon request.

3.10 Graphs

In order to produce the plots presented in the current work, a number of different R packages were used. Those are the `ggplot2`⁴⁵, `gplots`⁴⁶, `graphics`⁴⁴, `tmod`³⁴ and `Sushi`⁴⁷ packages. The REVIGO⁴⁸ platform was used as well.

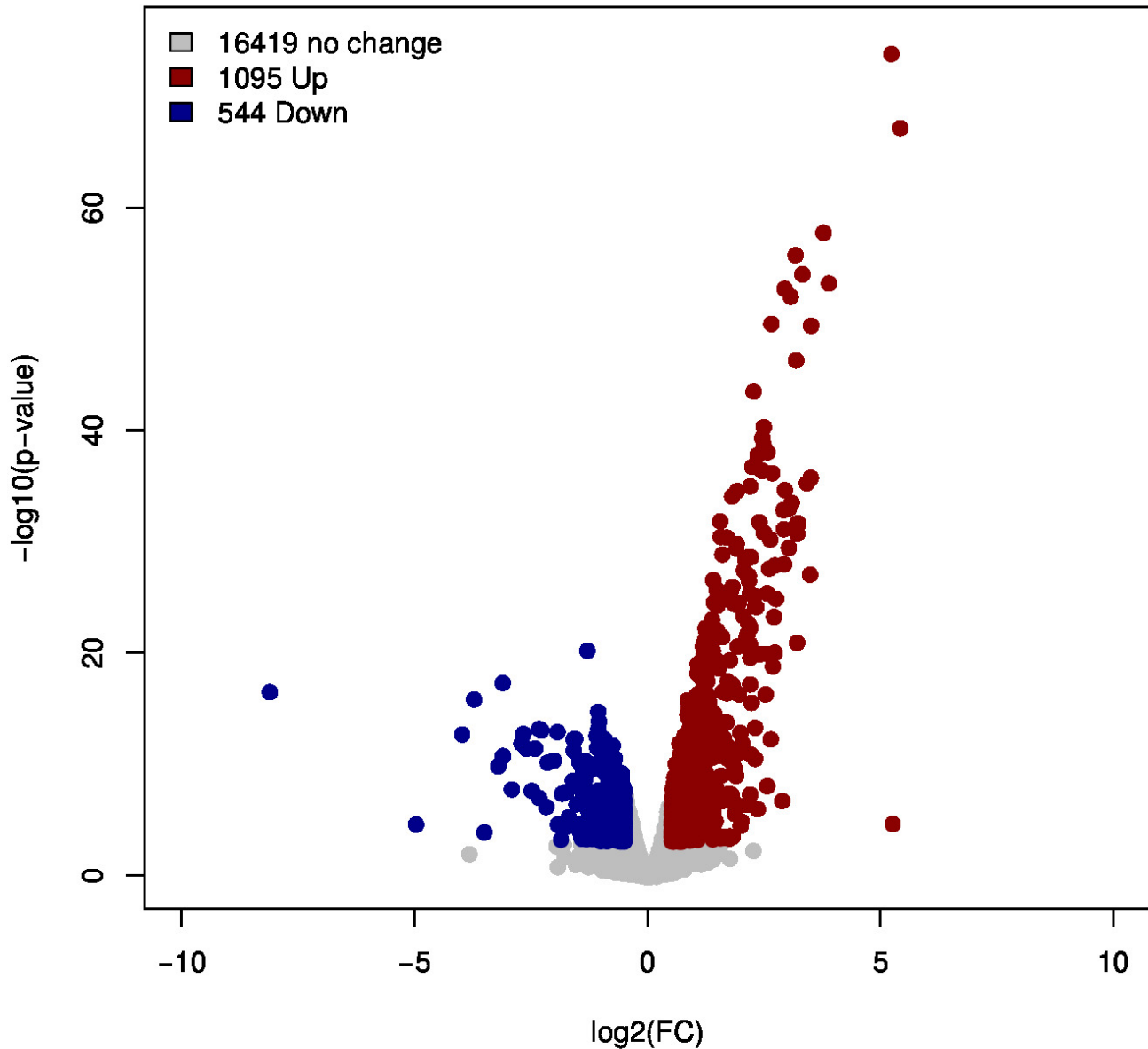
4 Results

4.1 Differential expression and Functional analysis

Differential expression analysis indicated a highly modified transcriptional profile for SLE patients. When SLE group was compared to healthy group, 1639 differentially expressed genes (DEGs, $FDR \leq 0.05$ & $|\log_2FC| \geq 0.5$) were identified, with 1095 being overexpressed and 544 being underexpressed (Figure 6A). A simple hypergeometric test revealed multiple statistically significantly enriched Gene Ontology terms (Figure 7). Enriched functions concerning immune response, cell activation and regulation of viral life cycle were identified. Consistently, enrichment analysis for KEGG and Reactome pathways uncovered a variety of enriched pathways in the overexpressed gene list, including previously identified terms, such as NOD-like receptor signalling and interferon signalling, and other unexpected pathways, such as cell cycle, oxidative stress-induced senescence and nucleosome assembly (Table 2A). Furthermore, the underexpressed DEG list was enriched for pathways, with PI3-Akt signalling and extracellular matrix organization acquiring the highest statistical significance (Table 2B). Thus, whole blood transcriptomic analysis illustrate a highly disrupted profile for SLE cohort, with specific signatures emanating.

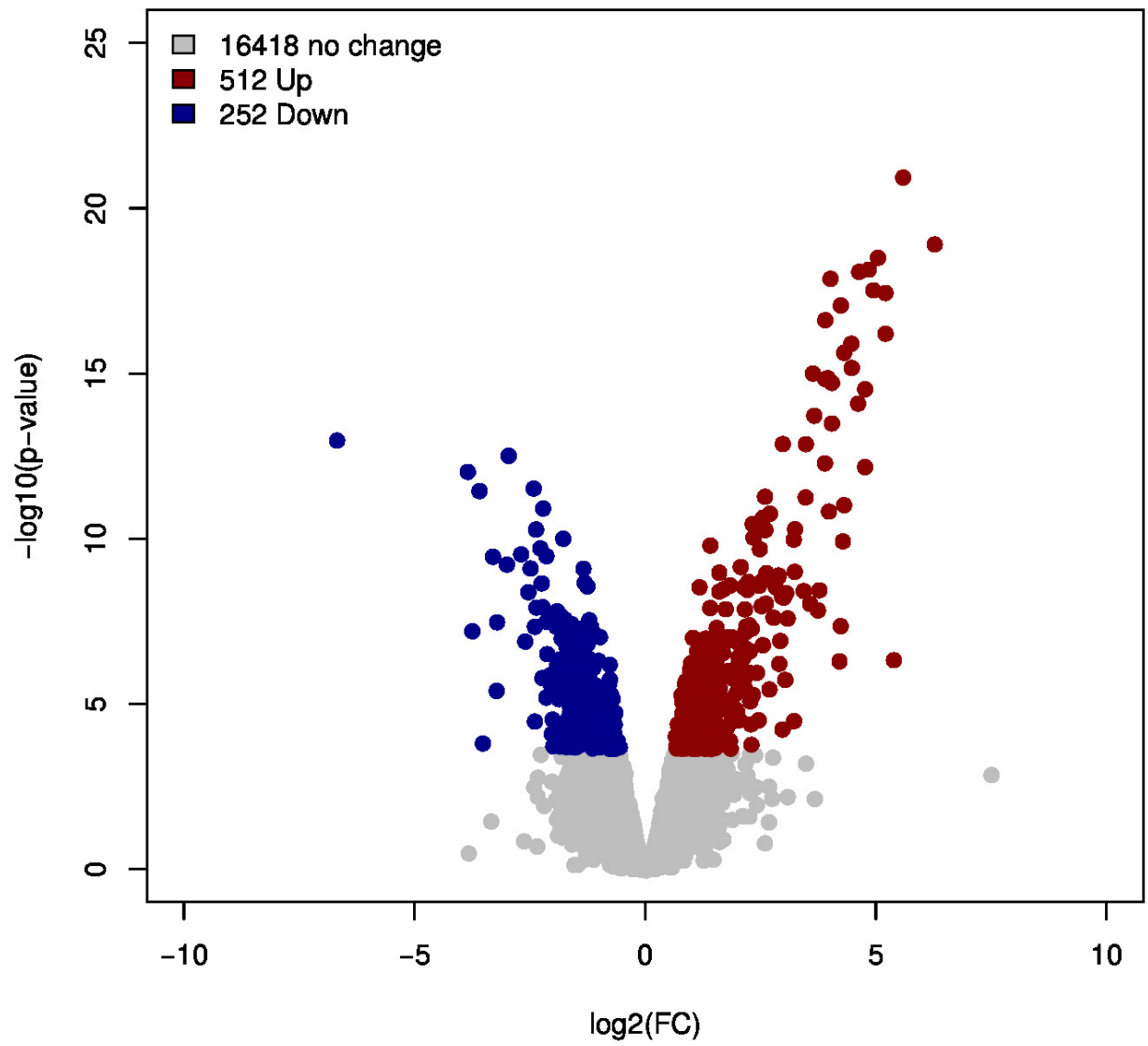
In an effort to uncover gene signatures associated with disease activity, the patient cohort was split into three subgroups of increasing activity, DA1, DA2 and DA3 (see Figure 2, above). The transcriptional profile of each DA group was analysed in comparison to control healthy expression levels, henceforth representing differential expression profile of the group unless stated otherwise, and to the rest of patient groups as well. The results of the different comparisons are illustrated in volcano plots of Figure 6 and in the heatmap depicted in Figure 8. Interestingly, patient groups with higher DA are linked with increased ratios of numbers of over- against under-expressed DEGs (which will be referred to as *r* ratio from here on). In other words, increased DA associates with a slight transcriptional ‘turnover’ to overexpression. However, the comparison between DA1 and DA2 groups resulted in only one statistically significant DEG. In general, the transcriptional profiles of DA1 and DA2 are closer to one another while DA3 differentiates significantly. That was verified by a hierarchical clustering, illustrated in Figure 8. Functional enrichment analysis was implemented to interpret these data. Differential expression of all three patient groups was enriched in immune system related functional terms (Table 3). Particularly, innate immunity pathways, such as NOD-like receptor signalling and IFN signalling, were enriched in over-

expressed DEGs. Nevertheless, differences were detected as well. Underexpressed DEGs of DA1 and DA2 but not DA3 are enriched in adaptive immunity pathways (signalling through Fc and B cell receptor signalling), something that can probably be explained by the lymphopenia that SLE patients endure. Moreover, cell cycle pathways seem to be enriched in DA3 overexpressed genes. Those differences were detected in the comparison between DA3 and DA1 (Table 3G). Overexpressed DEGs from that comparison were enriched in cell cycle and B cell receptor (BCR) signalling pathways.

A**SLE vs Healthy**

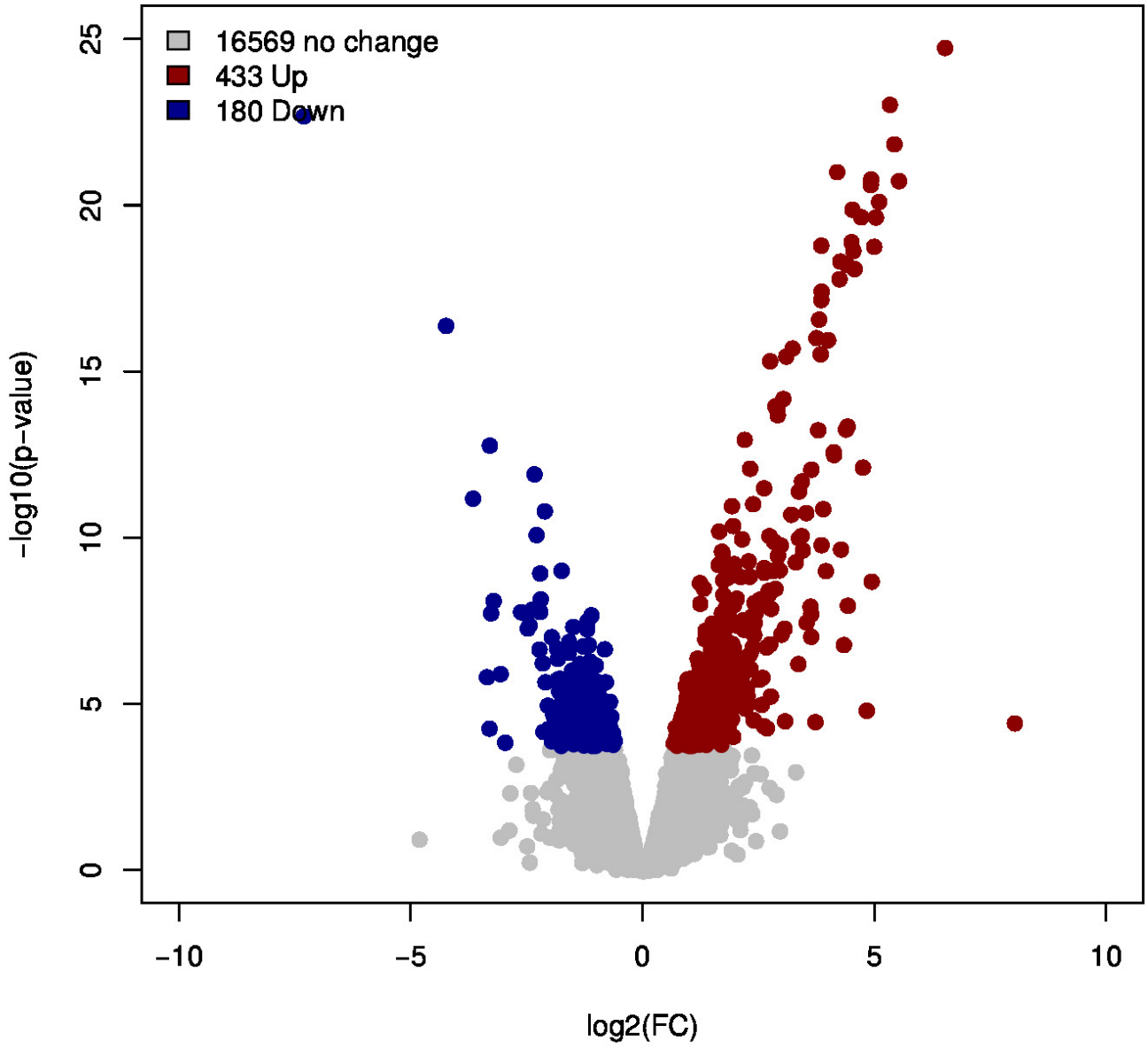
B

DA1 vs Healthy



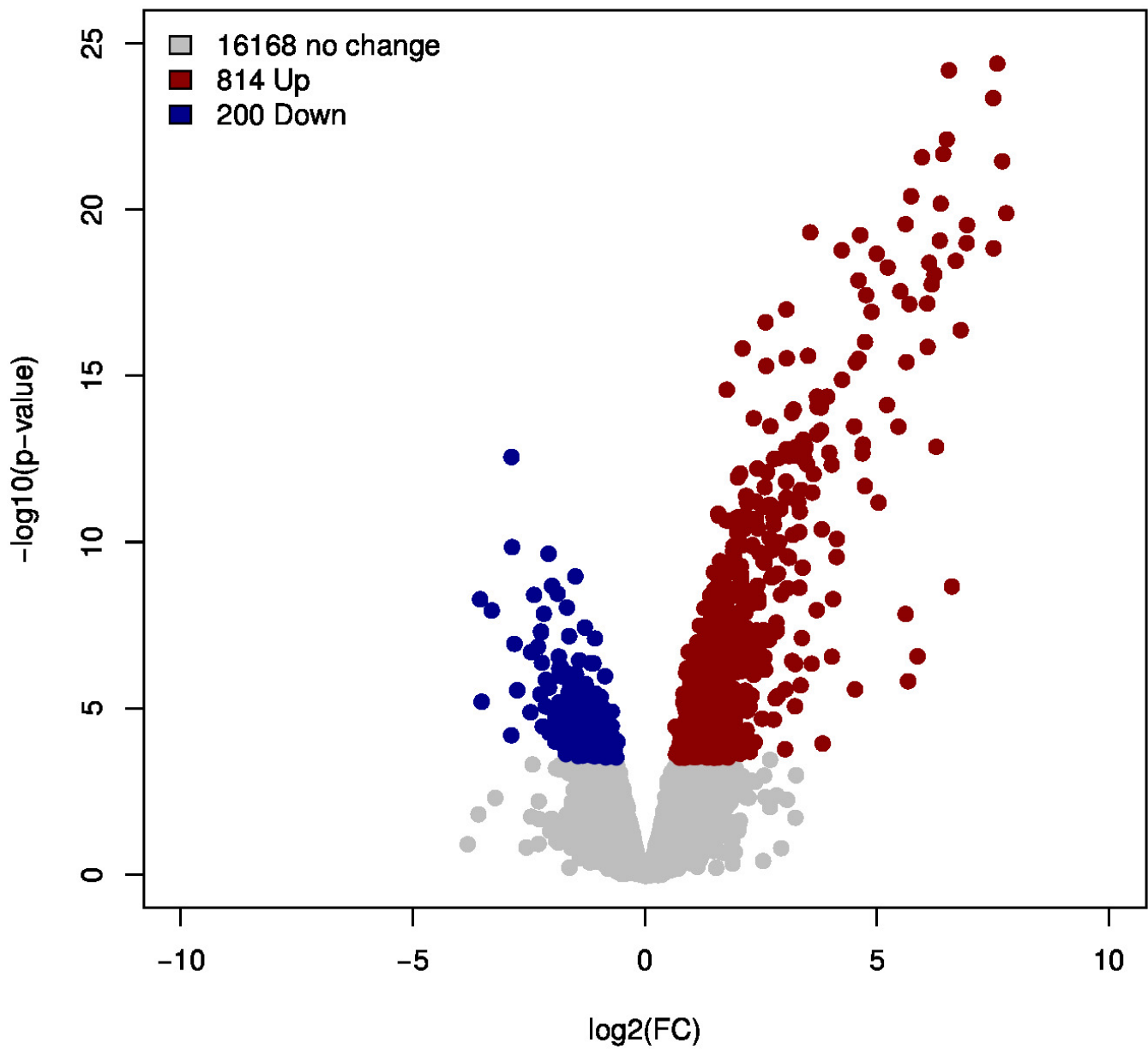
C

DA2 vs Healthy



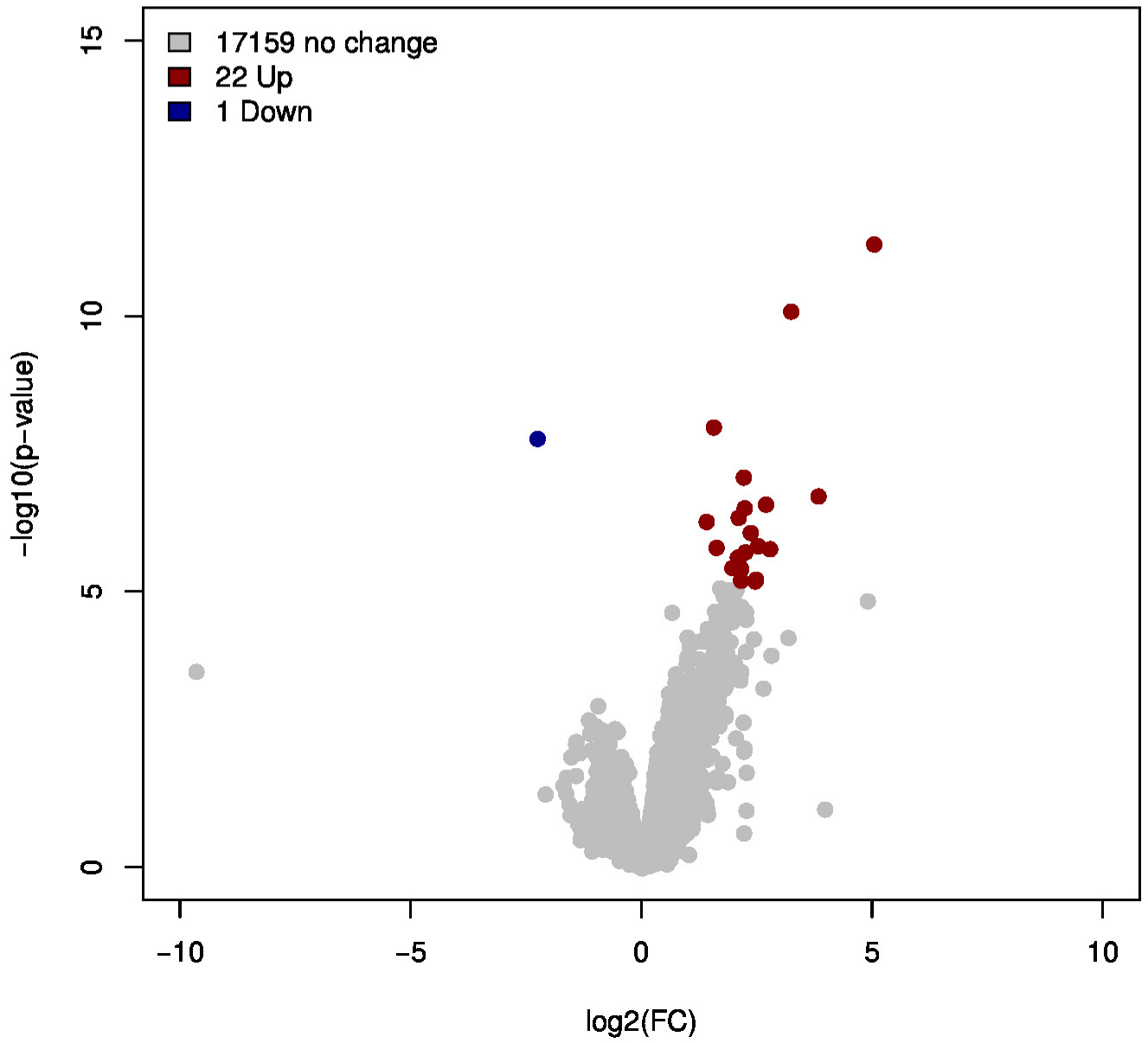
D

DA3 vs Healthy



III

DA3 vs DA2



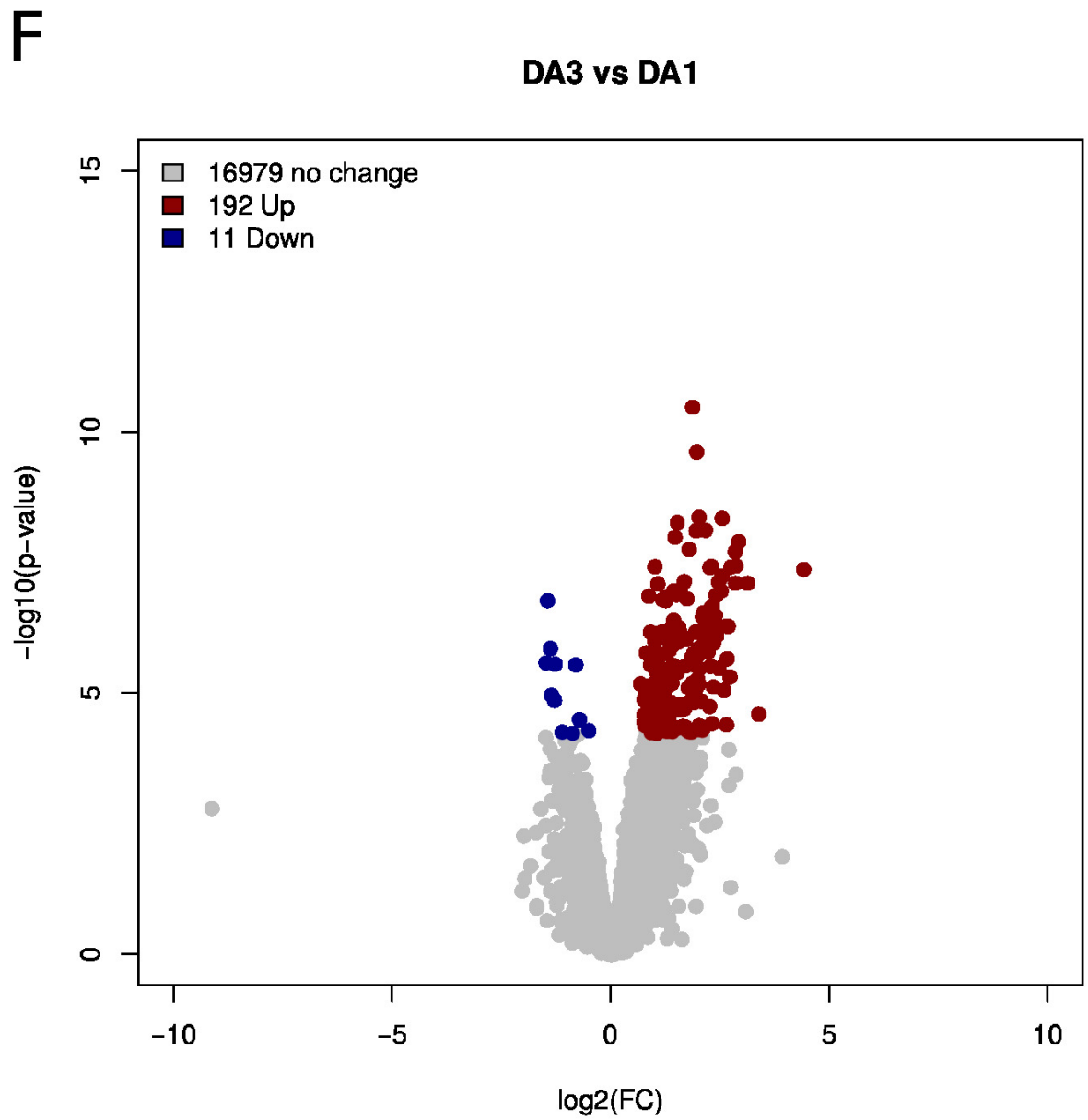
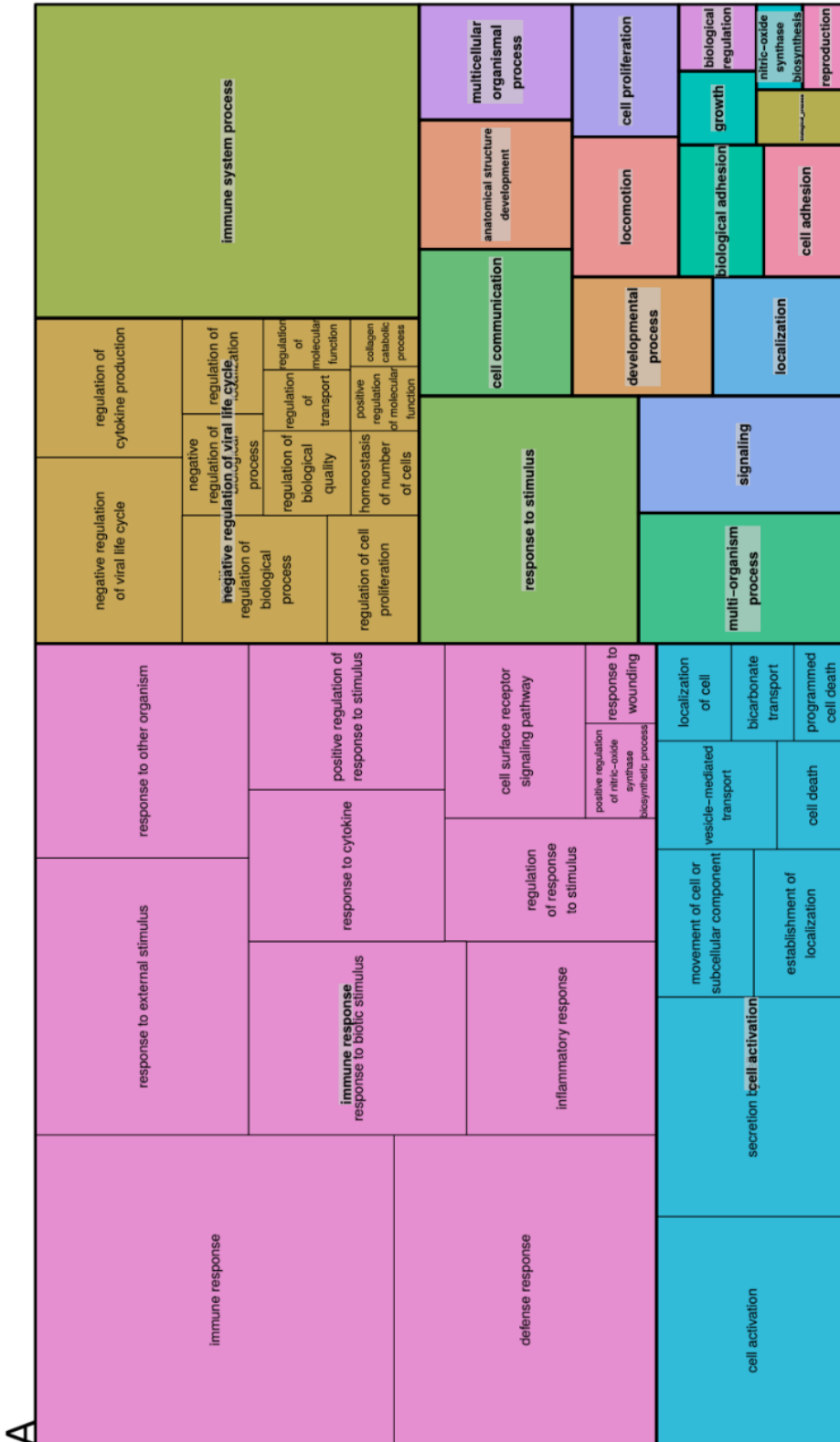
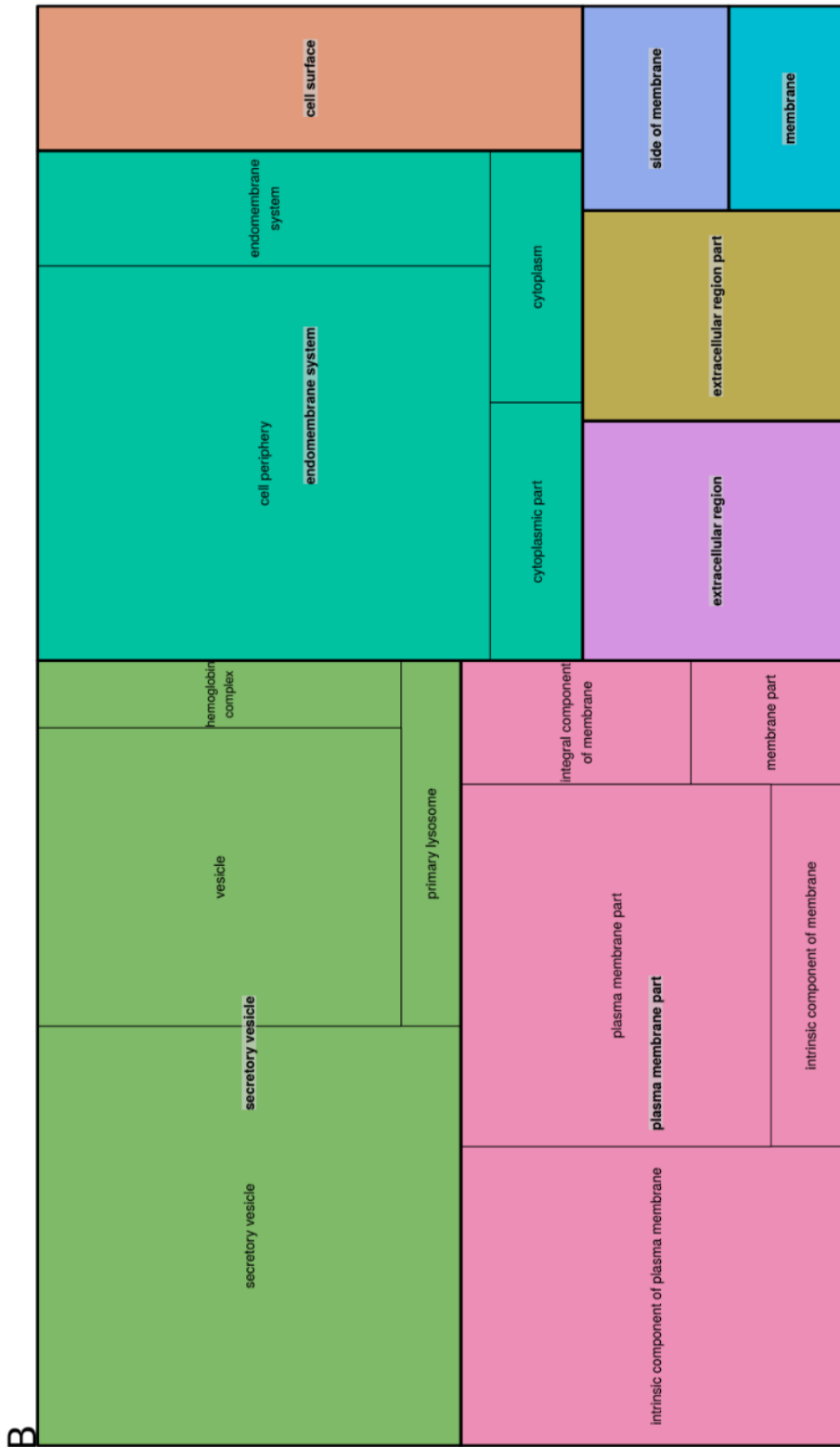


Figure 6: Volcano plots illustrating differential expression resulted from different group comparisons.





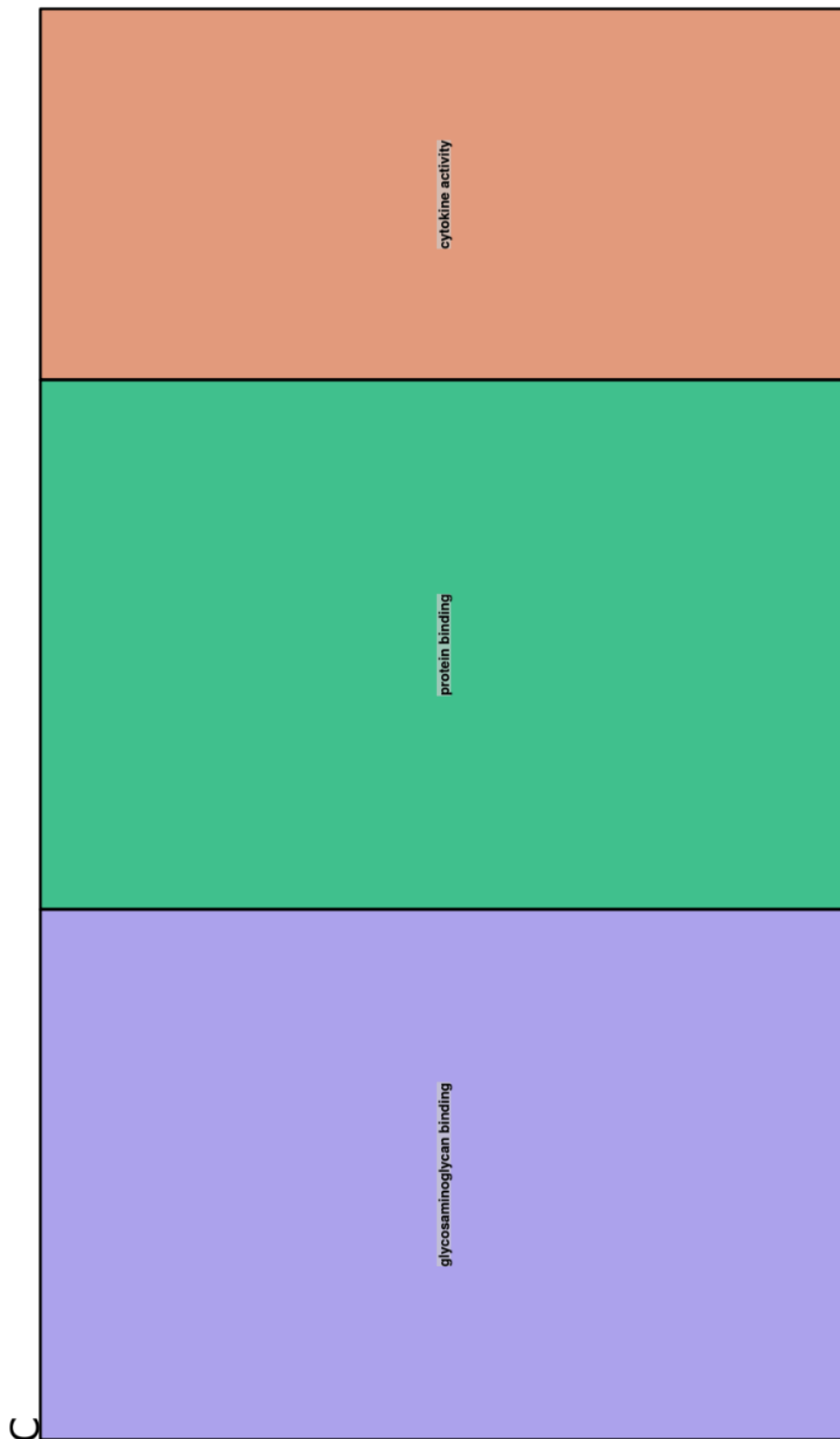


Figure 7: Treemaps depicting Gene Ontology terms enriched in the DEG list of SLE vs Healthy

4. Results

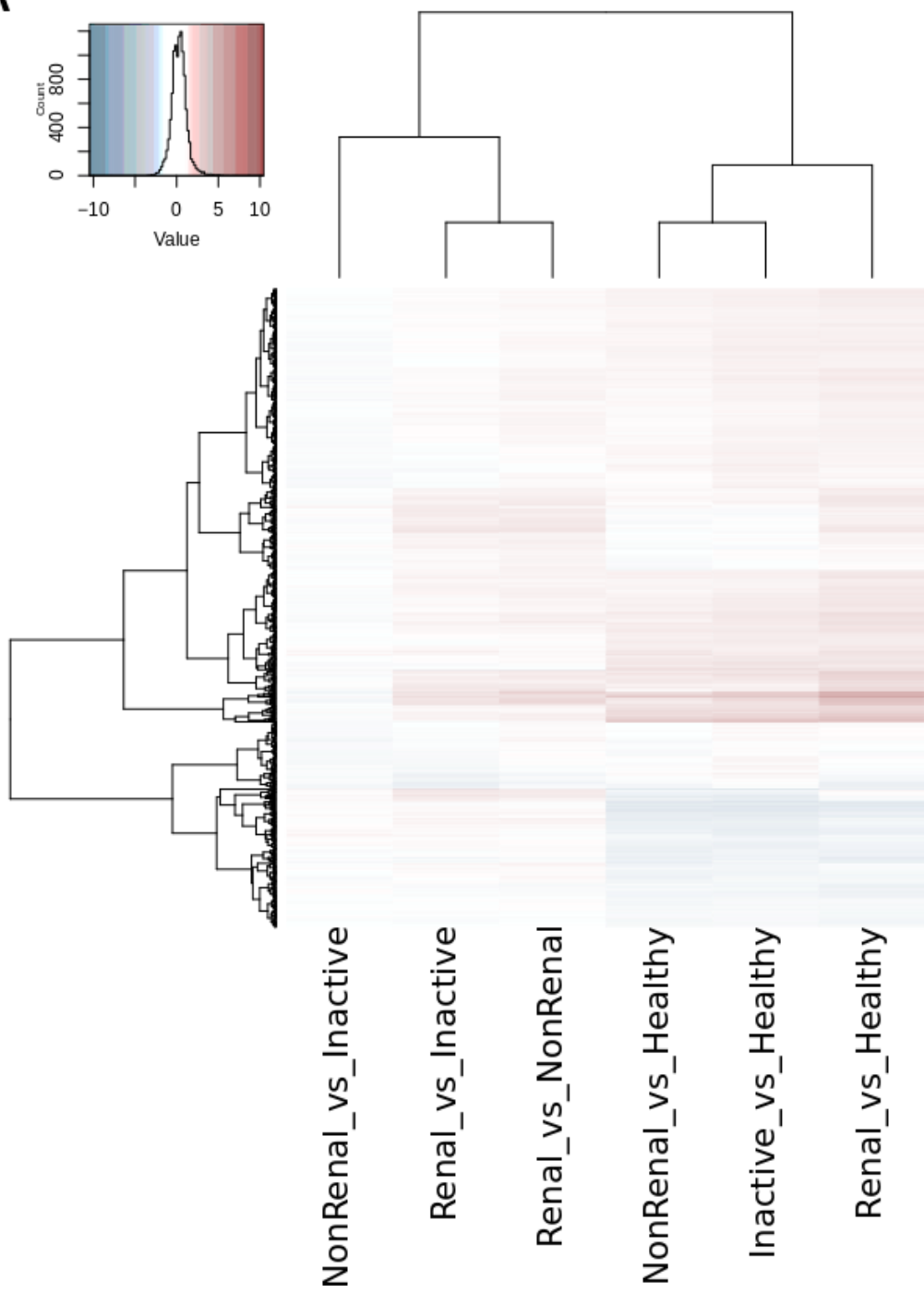
A						
source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
Biological pathways (KEGG)						
keg	NOD-like receptor signaling pathway	KEGG:04621	167	414	32	9.78e-08
keg	Cell cycle	KEGG:04110	124	414	21	6.43e-04
keg	Systemic lupus erythematosus	KEGG:05322	128	414	22	3.00e-04
keg	Influenza A	KEGG:05164	171	414	28	3.74e-05
keg	Malaria	KEGG:05144	48	414	12	1.16e-03
keg	Transcriptional misregulation in cancer	KEGG:05202	184	414	28	1.73e-04
keg	Osteoclast differentiation	KEGG:04380	125	414	20	2.55e-03
source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
Biological pathways (Reactome)						
rea	O ₂ /CO ₂ exchange in erythrocytes	REAC:1480926	13	621	7	1.11e-03
rea	Erythrocytes take up carbon dioxide and release oxygen	REAC:1237044	13	621	7	1.11e-03
rea	Erythrocytes take up oxygen and release carbon dioxide	REAC:1247673	9	621	7	2.85e-05
rea	Cell Cycle	REAC:1640170	632	621	66	7.29e-04
rea	Cell Cycle Checkpoints	REAC:69620	293	621	37	2.49e-03
rea	G2/M Checkpoints	REAC:69481	169	621	24	1.65e-02
rea	Cell Cycle, Mitotic	REAC:69278	525	621	56	2.47e-03
rea	G0 and Early G1	REAC:1538133	27	621	8	4.16e-02
rea	G1/S-Specific Transcription	REAC:69205	29	621	11	1.24e-04
rea	Cellular Senescence	REAC:2559583	196	621	28	3.41e-03
rea	Senescence-Associated Secretory Phenotype (SASP)	REAC:2559582	110	621	19	6.66e-03
rea	DNA Damage/Telomere Stress Induced Senescence	REAC:2559586	80	621	17	1.03e-03
rea	Oxidative Stress Induced Senescence	REAC:2559580	126	621	20	1.45e-02
rea	Meiotic synapsis	REAC:1221632	79	621	15	1.65e-02
rea	Nucleosome assembly	REAC:774815	73	621	15	6.15e-03
rea	Deposition of new CENPA-containing nucleosomes at the centromere	REAC:606279	73	621	15	6.15e-03
rea	Immune System	REAC:168256	2297	621	257	1.98e-28
rea	Cytokine Signaling in Immune system	REAC:1280215	826	621	100	1.84e-10
rea	Interferon Signaling	REAC:913531	197	621	52	1.56e-18
rea	Interferon gamma signaling	REAC:877300	91	621	28	3.78e-11
rea	Interferon alpha/beta signaling	REAC:909733	70	621	29	1.28e-15
rea	Innate Immune System	REAC:168249	1450	621	157	1.99e-13
rea	Packaging Of Telomere Ends	REAC:171306	52	621	12	1.29e-02
rea	RHO GTPases activate PKNs	REAC:5625740	95	621	16	4.20e-02

B

source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (KEGG)					
keg	Primary immunodeficiency	KEGG:05340	35	176	7	1.47e-03
keg	Toxoplasmosis	KEGG:05145	111	176	10	3.03e-02
keg	Viral myocarditis	KEGG:05416	56	176	8	4.78e-03
keg	Cytokine-cytokine receptor interaction	KEGG:04060	268	176	17	2.33e-02
keg	Protein digestion and absorption	KEGG:04974	90	176	9	2.76e-02
keg	ECM-receptor interaction	KEGG:04512	82	176	9	1.35e-02
keg	Intestinal immune network for IgA production	KEGG:04672	46	176	7	9.35e-03
keg	Small cell lung cancer	KEGG:05222	93	176	10	6.94e-03
keg	Hematopoietic cell lineage	KEGG:04640	94	176	9	3.82e-02
keg	Graft-versus-host disease	KEGG:05332	37	176	6	2.05e-02
keg	Hippo signaling pathway	KEGG:04390	154	176	12	3.08e-02
keg	Pathways in cancer	KEGG:05200	395	176	24	2.43e-03
keg	PI3K-Akt signaling pathway	KEGG:04151	351	176	24	3.31e-04
keg	Focal adhesion	KEGG:04510	199	176	15	8.22e-03
	Biological pathways (Reactome)					
rea	Assembly of collagen fibrils and other multimeric structures	REAC:2022090	59	261	8	3.05e-02
rea	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	REAC:198933	199	261	16	1.02e-02
rea	Extracellular matrix organization	REAC:1474244	296	261	28	2.49e-07
rea	Non-integrin membrane-ECM interactions	REAC:3000171	42	261	8	2.32e-03
rea	Collagen biosynthesis and modifying enzymes	REAC:1650814	67	261	9	1.18e-02

Table 2: KEGG and Reactome pathways enriched in overexpressed (A) and underexpressed (B) genes in SLE patients.

A



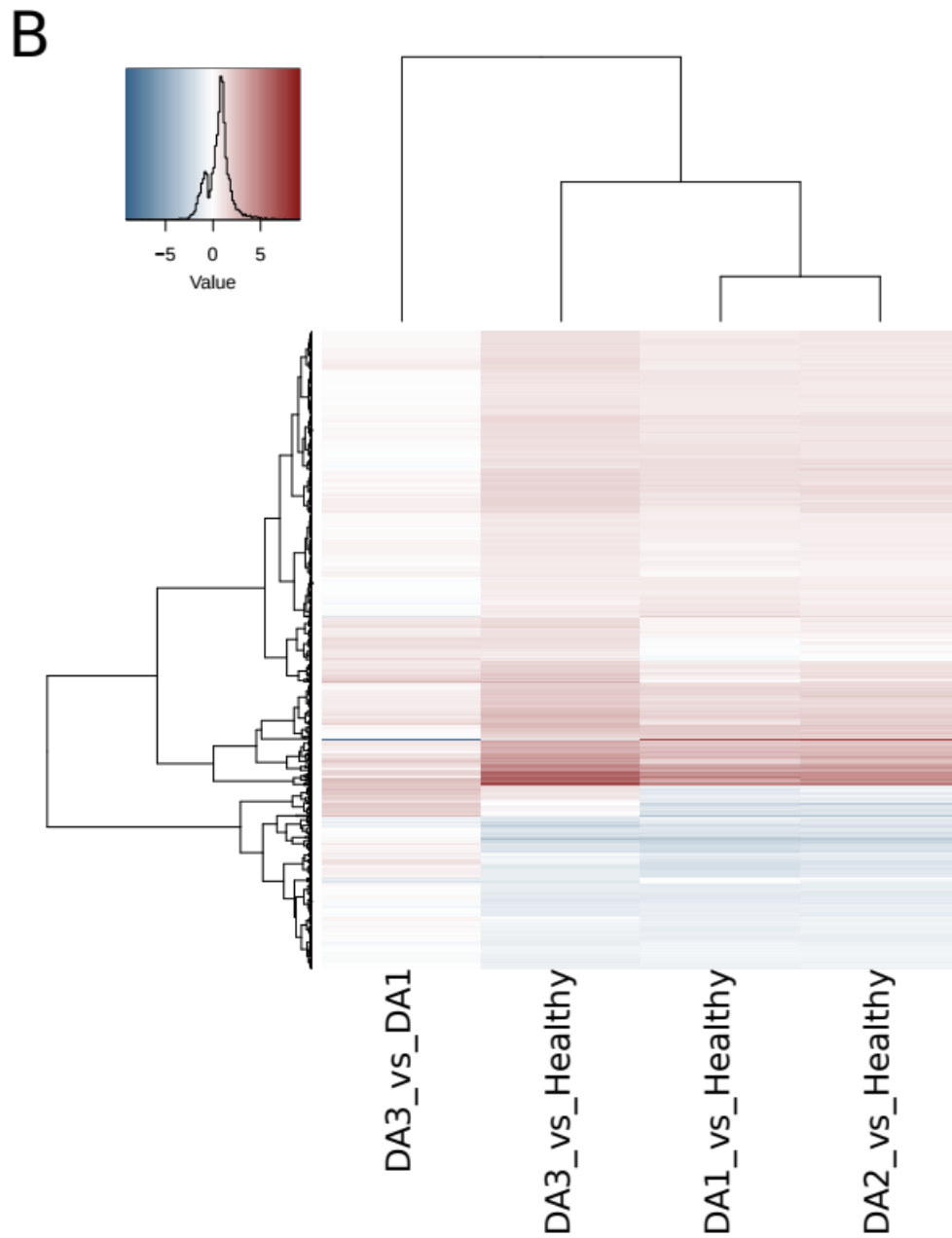


Figure 8: Heatmaps illustrating Log_2FC values of DEGs derived from different comparisons. **A.** Patients have been grouped according to the state and the manifestations of the disease. **B.** Patients have been grouped according to disease activity. Genes depicted in each heatmap are differentially expressed in at least one of the comparisons indicated at x axis. Trees in each heatmap depict the results of hierarchical clustering.

4. Results

A

source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (KEGG)					
keg	RIG-I-like receptor signaling pathway	KEGG:04622	70	217	9	2.16e-02
keg	Influenza A	KEGG:05164	171	217	25	3.20e-09
keg	Herpes simplex infection	KEGG:05168	182	217	17	2.92e-03
keg	Leishmaniasis	KEGG:05140	70	217	10	3.86e-03
keg	NOD-like receptor signaling pathway	KEGG:04621	167	217	23	6.99e-08
keg	Transcriptional misregulation in cancer	KEGG:05202	184	217	18	8.94e-04
keg	Osteoclast differentiation	KEGG:04380	125	217	14	1.99e-03
keg	Hepatitis C	KEGG:05160	131	217	13	1.39e-02
keg	Cytosolic DNA-sensing pathway	KEGG:04623	62	217	8	4.71e-02
keg	Measles	KEGG:05162	133	217	16	1.96e-04
source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (Reactome)					
rea	Immune System	REAC:168256	2297	326	188	8.23e-45
rea	Innate Immune System	REAC:168249	1450	326	128	1.97e-29
rea	Toll-Like Receptors Cascades	REAC:168898	152	326	16	5.99e-03
rea	RIG-I/MDA5 mediated induction of IFN-alpha/beta pathways	REAC:168928	56	326	9	1.70e-02
rea	TRAF6 mediated IRF7 activation	REAC:933541	33	326	7	1.93e-02
rea	TRAF6 mediated NF-kB activation	REAC:933542	26	326	7	3.58e-03
rea	TRAF3-dependent IRF activation pathway	REAC:918233	14	326	5	1.60e-02
rea	Cytokine Signaling in Immune system	REAC:1280215	826	326	68	6.21e-12
rea	Interferon Signaling	REAC:913531	197	326	39	1.11e-18
rea	Interferon gamma signaling	REAC:877300	91	326	19	8.94e-09
rea	Antiviral mechanism by IFN-stimulated genes	REAC:1169410	73	326	11	4.57e-03
rea	ISG15 antiviral mechanism	REAC:1169408	73	326	11	4.57e-03
rea	Interferon alpha/beta signaling	REAC:909733	70	326	25	3.63e-18

B

source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (KEGG)					
keg	Primary immunodeficiency	KEGG:05340	35	79	4	4.22e-02
keg	B cell receptor signaling pathway	KEGG:04662	70	79	7	7.79e-04
keg	Allograft rejection	KEGG:05330	35	79	4	4.22e-02
keg	Cytokine-cytokine receptor interaction	KEGG:04060	268	79	10	4.96e-02
source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (Reactome)					
rea	Cell surface interactions at the vascular wall	REAC:202733	212	120	12	1.43e-03
rea	Fc epsilon receptor (FCERI) signaling	REAC:2454202	542	120	17	3.60e-02
rea	FCERI mediated MAPK activation	REAC:2871796	338	120	16	3.59e-04
rea	FCERI mediated NF-kB activation	REAC:2871837	154	120	9	1.87e-02
rea	FCERI mediated Ca ²⁺ mobilization	REAC:2871809	105	120	9	8.39e-04
rea	Fcgamma receptor (FCGR) dependent phagocytosis	REAC:2029480	165	120	10	5.19e-03
rea	FCGR activation	REAC:2029481	90	120	10	1.92e-05
rea	Regulation of actin dynamics for phagocytic cup formation	REAC:2029482	140	120	10	1.21e-03
rea	Role of phospholipids in phagocytosis	REAC:2029485	103	120	10	7.02e-05
rea	Binding and Uptake of Ligands by Scavenger Receptors	REAC:2173782	118	120	10	2.52e-04
rea	Scavenging of heme from plasma	REAC:2168880	89	120	9	2.06e-04
rea	Adaptive Immune System	REAC:1280218	996	120	28	1.23e-03
rea	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	REAC:198933	199	120	15	1.83e-06
rea	Signaling by the B Cell Receptor (BCR)	REAC:983705	399	120	18	1.45e-04
rea	Antigen activates B Cell Receptor (BCR) leading to generation of second me ...	REAC:983695	100	120	16	4.56e-12
rea	CD22 mediated BCR regulation	REAC:5690714	66	120	13	9.40e-11
rea	Complement cascade	REAC:166658	134	120	10	8.11e-04
rea	Initial triggering of complement	REAC:166663	99	120	9	5.11e-04
rea	Creation of C4 and C2 activators	REAC:166786	92	120	9	2.74e-04
rea	Classical antibody-mediated complement activation	REAC:173623	84	120	9	1.25e-04
rea	Regulation of Complement cascade	REAC:977606	123	120	10	3.70e-04

4. Results

C

source	term name	term ID	n. of genes	n. of query genes	n. of common genes	corrected p-value
Biological pathways (KEGG)						
keg	Measles	KEGG:05162	133	167	11	2.20e-02
keg	Transcriptional misregulation in cancer	KEGG:05202	184	167	16	4.37e-04
keg	Influenza A	KEGG:05164	171	167	19	1.03e-06
keg	NOD-like receptor signaling pathway	KEGG:04621	167	167	18	4.15e-06
Biological pathways (Reactome)						
source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
rea	Immune System	REAC:168256	2297	266	148	2.24e-32
rea	Cytokine Signaling in Immune system	REAC:1280215	826	266	58	7.20e-11
rea	Interferon Signaling	REAC:913531	197	266	35	6.06e-18
rea	Interferon alpha/beta signaling	REAC:909733	70	266	21	4.18e-15
rea	Antiviral mechanism by IFN-stimulated genes	REAC:1169410	73	266	11	5.84e-04
rea	ISG15 antiviral mechanism	REAC:1169408	73	266	11	5.84e-04
rea	Interferon gamma signaling	REAC:877300	91	266	16	2.34e-07
rea	Innate Immune System	REAC:168249	1450	266	94	2.24e-17

D

source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
Biological pathways (KEGG)						
keg	Cytokine-cytokine receptor interaction	KEGG:04060	268	53	9	9.24e-03
keg	PI3K-Akt signaling pathway	KEGG:04151	351	53	10	1.46e-02
keg	Focal adhesion	KEGG:04510	199	53	7	4.28e-02
keg	B cell receptor signaling pathway	KEGG:04662	70	53	5	1.11e-02
Biological pathways (Reactome)						
source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
rea	Fc epsilon receptor (FCERI) signaling	REAC:2454202	542	82	14	1.79e-02
rea	FCERI mediated MAPK activation	REAC:2871796	338	82	11	1.59e-02
rea	FCERI mediated Ca ²⁺ mobilization	REAC:2871809	105	82	6	4.73e-02
rea	FCGR activation	REAC:2029481	90	82	6	2.01e-02
rea	Role of phospholipids in phagocytosis	REAC:2029485	103	82	7	4.19e-03
rea	Chemokine receptors bind chemokines	REAC:380108	48	82	5	9.88e-03
rea	Immune System	REAC:168256	2297	82	34	1.05e-02
rea	Adaptive Immune System	REAC:1280218	996	82	23	2.97e-04
rea	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	REAC:198933	199	82	13	9.92e-07
rea	Signaling by the B Cell Receptor (BCR)	REAC:983705	399	82	14	5.54e-04
rea	CD22 mediated BCR regulation	REAC:5690714	66	82	9	4.82e-07
rea	Antigen activates B Cell Receptor (BCR) leading to generation of second ...	REAC:983695	100	82	10	1.26e-06

4. Results

E

source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (KEGG)					
keg	Cell cycle	KEGG:04110	124	315	15	2.98e-02
keg	Herpes simplex infection	KEGG:05168	182	315	20	1.19e-02
keg	Influenza A	KEGG:05164	171	315	25	7.95e-06
keg	NOD-like receptor signaling pathway	KEGG:04621	167	315	26	1.14e-06
keg	Transcriptional misregulation in cancer	KEGG:05202	184	315	24	1.25e-04
source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (Reactome)					
rea	Cell Cycle	REAC:1640170	632	466	51	5.66e-03
rea	Cell Cycle, Mitotic	REAC:69278	525	466	44	9.25e-03
rea	Mitotic G1-G1/S phases	REAC:453279	148	466	18	3.31e-02
rea	Cell Cycle Checkpoints	REAC:69620	293	466	29	1.34e-02
rea	G2/M Checkpoints	REAC:69481	169	466	20	1.98e-02
rea	Protein repair	REAC:5676934	6	466	5	3.95e-04
rea	G1/S-Specific Transcription	REAC:69205	29	466	10	9.40e-05
rea	Immune System	REAC:168256	2297	466	218	5.39e-33
rea	Innate Immune System	REAC:168249	1450	466	140	8.06e-19
rea	Cytokine Signaling in Immune system	REAC:1280215	826	466	85	1.20e-11
rea	Interferon Signaling	REAC:913531	197	466	51	2.00e-23
rea	Interferon alpha/beta signaling	REAC:909733	70	466	28	7.94e-18
rea	Interferon gamma signaling	REAC:877300	91	466	27	2.77e-13
rea	Antiviral mechanism by IFN-stimulated genes	REAC:1169410	73	466	12	2.93e-02
rea	ISG15 antiviral mechanism	REAC:1169408	73	466	12	2.93e-02

F

source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (KEGG)					
keg	Cytokine-cytokine receptor interaction	KEGG:04060	268	72	13	1.49e-04
source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (Reactome)					
rea	Chemokine receptors bind chemokines	REAC:380108	48	103	6	1.89e-03
rea	GPCR ligand binding	REAC:500792	459	103	14	3.77e-02
rea	G alpha (i) signalling events	REAC:418594	246	103	11	7.29e-03

4. Results

G

source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (KEGG)					
keg	Cell cycle	KEGG:04110	124	55	11	1.17e-07
keg	Oocyte meiosis	KEGG:04114	121	55	6	2.18e-02
keg	Progesterone-mediated oocyte maturation	KEGG:04914	96	55	6	6.08e-03
source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (Reactome)					
rea	RHO GTPases Activate Formins	REAC:5663220	123	126	9	5.02e-03
rea	Cell Cycle	REAC:1640170	632	126	38	4.44e-15
rea	Cell Cycle, Mitotic	REAC:69278	525	126	34	3.67e-14
rea	M Phase	REAC:68886	383	126	20	6.83e-06
rea	Mitotic Prometaphase	REAC:68877	186	126	15	1.56e-06
rea	Resolution of Sister Chromatid Cohesion	REAC:2500257	108	126	12	1.62e-06
rea	Condensation of Prometaphase Chromosomes	REAC:2514853	12	126	4	3.05e-03
rea	Mitotic Metaphase and Anaphase	REAC:2555396	185	126	12	6.30e-04
rea	Mitotic Anaphase	REAC:68882	184	126	12	5.95e-04
rea	Separation of Sister Chromatids	REAC:2467813	173	126	12	3.09e-04
rea	Mitotic G1-G1/S phases	REAC:453279	148	126	9	2.18e-02
rea	G0 and Early G1	REAC:1538133	27	126	5	4.96e-03
rea	Mitotic G2-G2/M phases	REAC:453274	187	126	10	2.51e-02
rea	G2/M Transition	REAC:69275	185	126	10	2.30e-02
rea	Polo-like kinase mediated events	REAC:156711	16	126	5	2.97e-04
rea	Cell Cycle Checkpoints	REAC:69620	293	126	21	8.71e-09
rea	G2/M Checkpoints	REAC:69481	169	126	11	1.68e-03
rea	G2/M DNA replication checkpoint	REAC:69478	5	126	3	5.67e-03
rea	G2/M DNA damage checkpoint	REAC:69473	95	126	7	4.49e-02
rea	Mitotic Spindle Checkpoint	REAC:69618	111	126	9	2.17e-03
rea	Amplification of signal from the kinetochores	REAC:141424	95	126	9	5.90e-04
rea	Amplification of signal from unattached kinetochores via a MAD2 in ...	REAC:141444	95	126	9	5.90e-04
rea	Immune System	REAC:168256	2297	126	63	3.95e-10
rea	Adaptive Immune System	REAC:1280218	996	126	32	3.44e-05
rea	Signaling by the B Cell Receptor (BCR)	REAC:983705	399	126	25	1.69e-09
rea	CD22 mediated BCR regulation	REAC:5690714	66	126	25	1.23e-29
rea	Antigen activates B Cell Receptor (BCR) leading to generation of second ...	REAC:983695	100	126	25	2.08e-24
rea	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	REAC:198933	199	126	30	9.18e-23
rea	Innate Immune System	REAC:168249	1450	126	56	6.11e-15
rea	Complement cascade	REAC:166658	134	126	33	8.45e-33
rea	Regulation of Complement cascade	REAC:977606	123	126	33	3.59e-34
rea	Initial triggering of complement	REAC:166663	99	126	32	5.10e-36
rea	Creation of C4 and C2 activators	REAC:166786	92	126	32	3.16e-37
rea	Classical antibody-mediated complement activation	REAC:173623	84	126	32	9.32e-39
rea	Fcgamma receptor (FCGR) dependent phagocytosis	REAC:2029480	165	126	32	4.11e-28
rea	Role of phospholipids in phagocytosis	REAC:2029485	103	126	32	2.24e-35
rea	Regulation of actin dynamics for phagocytic cup formation	REAC:2029482	140	126	32	1.42e-30
rea	FCGR activation	REAC:2029481	90	126	32	1.36e-37
rea	Fc epsilon receptor (FCERI) signaling	REAC:2454202	542	126	30	2.11e-10
rea	FCERI mediated MAPK activation	REAC:2871796	338	126	30	5.63e-16
rea	Role of LAT2/NTAL/LAB on calcium mobilization	REAC:2730905	357	126	29	2.56e-14
rea	FCERI mediated Ca ²⁺ mobilization	REAC:2871809	105	126	29	4.25e-30
rea	FCERI mediated NF-κB activation	REAC:2871837	154	126	29	7.87e-25
rea	Vesicle-mediated transport	REAC:5653656	737	126	35	1.46e-10
rea	Binding and Uptake of Ligands by Scavenger Receptors	REAC:2173782	118	126	31	1.33e-31
rea	Scavenging of heme from plasma	REAC:2168880	89	126	30	3.12e-34
rea	Hemostasis	REAC:109582	708	126	39	2.98e-14
rea	Cell surface interactions at the vascular wall	REAC:202733	212	126	35	1.83e-28
rea	G1/S-Specific Transcription	REAC:69205	29	126	6	3.25e-04

H

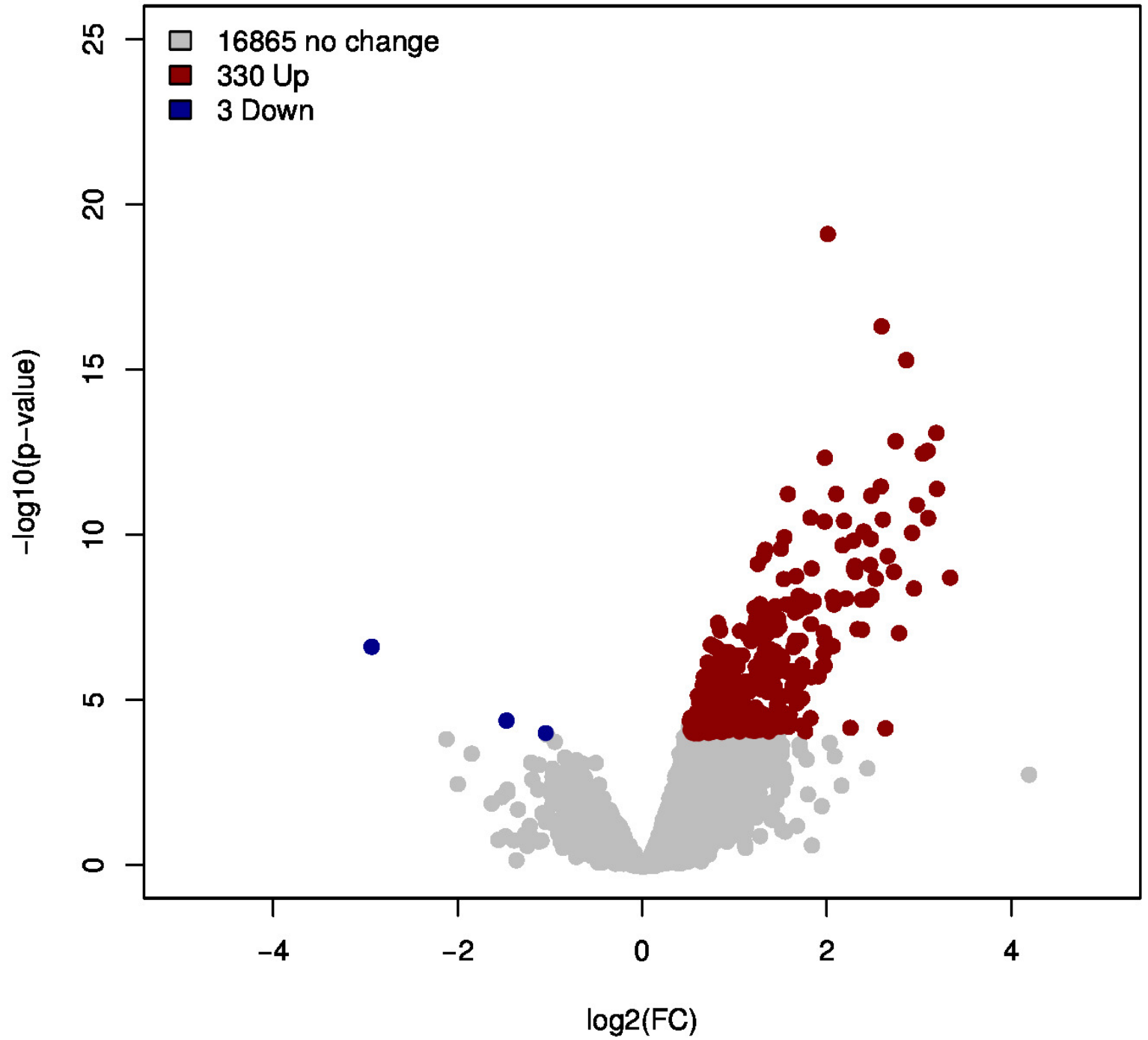
source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
keg	Biological pathways (KEGG)					
keg	ECM-receptor interaction	KEGG:04512	82	8	3	3.82e-03
source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
rea	Biological pathways (Reactome)					
rea	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	REAC:198933	199	18	6	1.43e-04
rea	Immune System	REAC:168256	2297	18	13	1.37e-03
rea	Innate Immune System	REAC:168249	1450	18	13	5.26e-06
rea	Fc epsilon receptor (FCERI) signaling	REAC:2454202	542	18	7	3.89e-03
rea	Role of LAT2/NTAL/LAB on calcium mobilization	REAC:2730905	357	18	6	4.23e-03
rea	FCERI mediated Ca ²⁺ mobilization	REAC:2871809	105	18	6	3.16e-06
rea	FCERI mediated MAPK activation	REAC:2871796	338	18	7	1.68e-04
rea	FCERI mediated NF-κB activation	REAC:2871837	154	18	6	3.14e-05
rea	Complement cascade	REAC:166658	134	18	8	4.67e-09
rea	Regulation of Complement cascade	REAC:977606	123	18	8	2.33e-09
rea	Initial triggering of complement	REAC:166663	99	18	8	3.95e-10
rea	Creation of C4 and C2 activators	REAC:166786	92	18	8	2.16e-10
rea	Classical antibody-mediated complement activation	REAC:173623	84	18	8	1.02e-10
rea	Fcγ receptor (FCGR) dependent phagocytosis	REAC:2029480	165	18	8	2.50e-08
rea	FCGR activation	REAC:2029481	90	18	8	1.80e-10
rea	Regulation of actin dynamics for phagocytic cup formation	REAC:2029482	140	18	8	6.65e-09
rea	Role of phospholipids in phagocytosis	REAC:2029485	103	18	8	5.46e-10
rea	Vesicle-mediated transport	REAC:5653656	737	18	7	2.81e-02
rea	Binding and Uptake of Ligands by Scavenger Receptors	REAC:2173782	118	18	7	1.15e-07
rea	Scavenging of heme from plasma	REAC:2168880	89	18	6	1.16e-06
rea	Hemostasis	REAC:109582	708	18	8	2.06e-03
rea	Cell surface interactions at the vascular wall	REAC:202733	212	18	8	1.86e-07
rea	Signaling by the B Cell Receptor (BCR)	REAC:983705	399	18	6	7.94e-03
rea	Antigen activates B Cell Receptor (BCR) leading to generation of second mess ...	REAC:983695	100	18	6	2.35e-06
rea	CD22 mediated BCR regulation	REAC:5690714	66	18	6	1.85e-07

Table 3: KEGG and Reactome pathways enriched in overexpressed and underexpressed genes of DA1 (A,B), DA2 (C,D), DA3 (E,F) and pathways enriched in the overexpressed genes resulted from DA3 vs DA1 (G) and DA3 vs DA2 (H).

Because there is not yet any simple, universal and absolute method to define and quantify SLE disease activity, a more relevant, and clear to detect, characteristic that may assist in a stratified tackling of the disease, is any clinical manifestation. One of the most severe manifestations of SLE is renal damage. To analyse the expressional data through that reasoning, the patient cohort was split again to three groups, this time according to disease status (Active or Inactive) and if a patient had any renal manifestation or not. We coin these as the inactive group, the renal active group and the non-renal active group. The transcriptional profile of each group was analysed in regard to control healthy expression, henceforth representing the expressional profile of the group unless stated otherwise, and to the rest of patient groups as well (Figure 9 and Figure 8A). Once more, there seems to be an association between the groups and the r ratio, with the renal active group having the highest r value and the non-renal active the lowest. Generally, non-renal active and inactive expression profiles cocluster first, in a hierarchical clustering (Figure 8A). Similar functional terms are enriched compared to former analysis (Table 4), which is expected, as The DA3 group includes mostly patients with nephritis and the DA1 one includes all the Inactive patients. Enriched pathways in overexpressed genes include cytokine signalling, IFN signalling and NOD-like receptor signalling. BCR signalling is enriched in the list of underexpressed DEGs of inactive and non-renal active groups and cell cycle-related pathways emerge from the analysis overexpressed DEGs in the renal active group compared to the rest of patient groups. Finally, DEGs from the non-renal active group compared to the inactive group are almost exclusively underexpressed and enriched for innate immune system and mitochondrion related ontologies, i.e. oxidative phosphorylation and tRNA processing in mitochondrion (Table 4I).

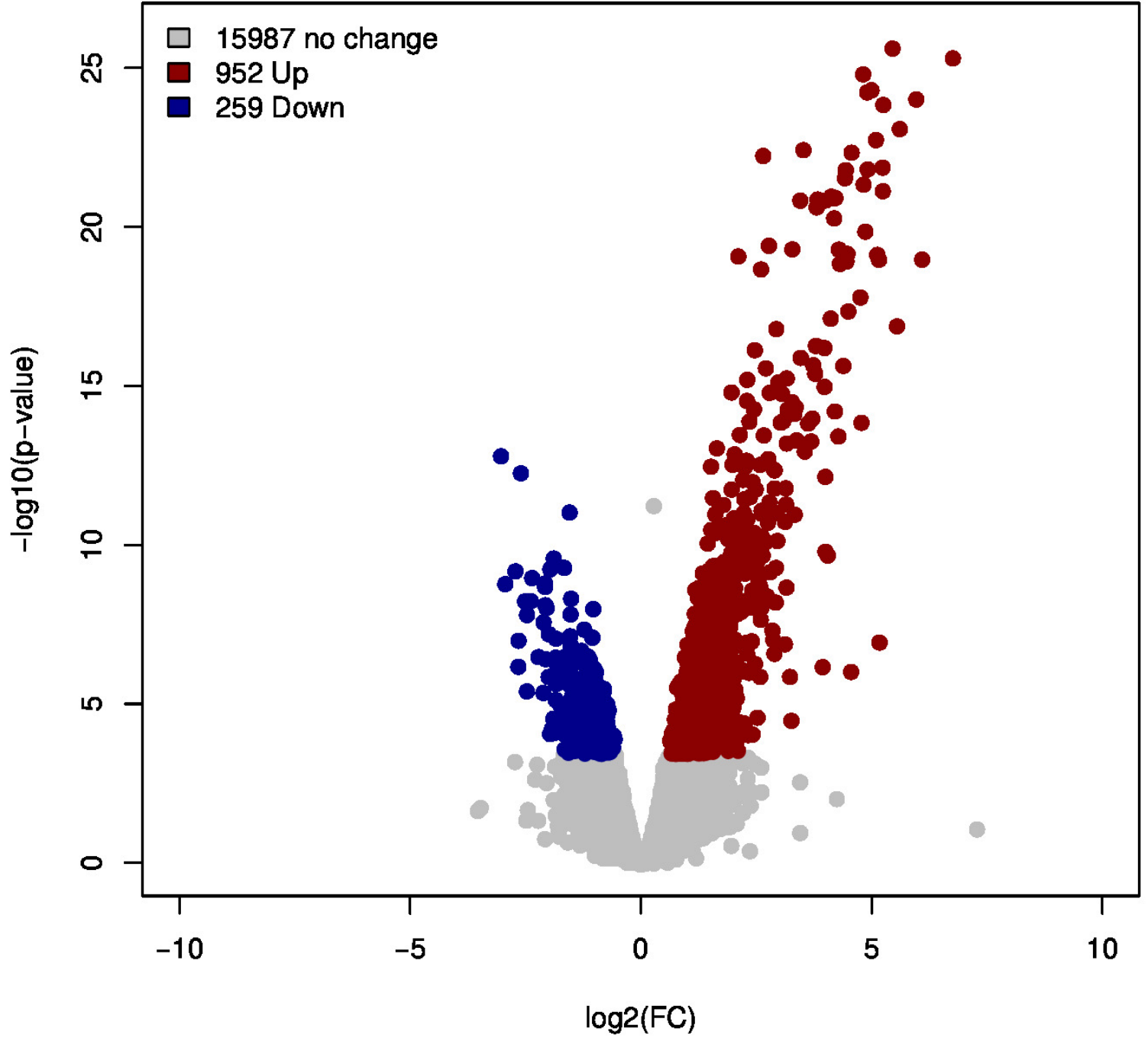
A

Active – Renal Disease vs Active – Non Renal Disease



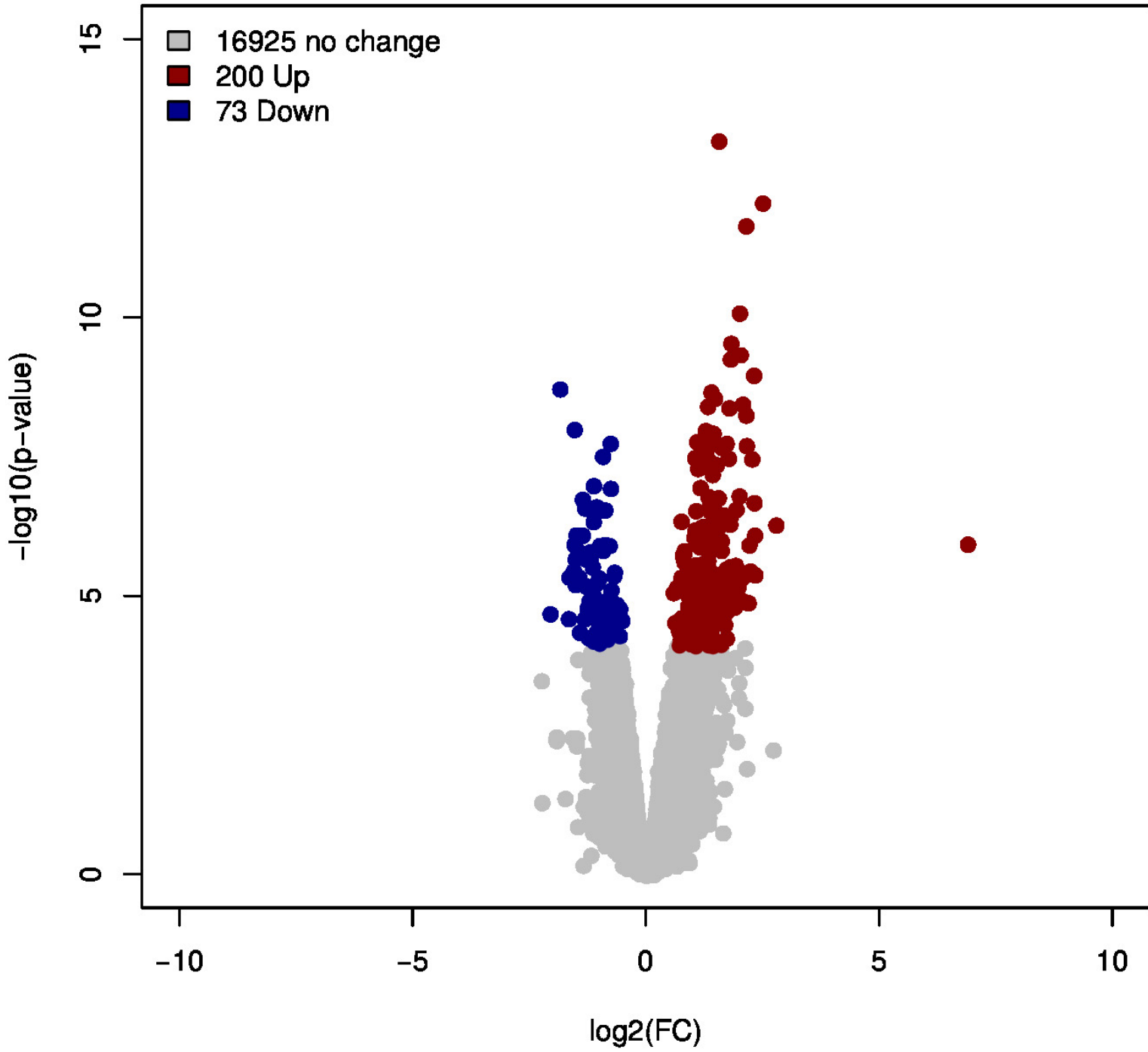
B

Active – Renal Disease vs Healthy



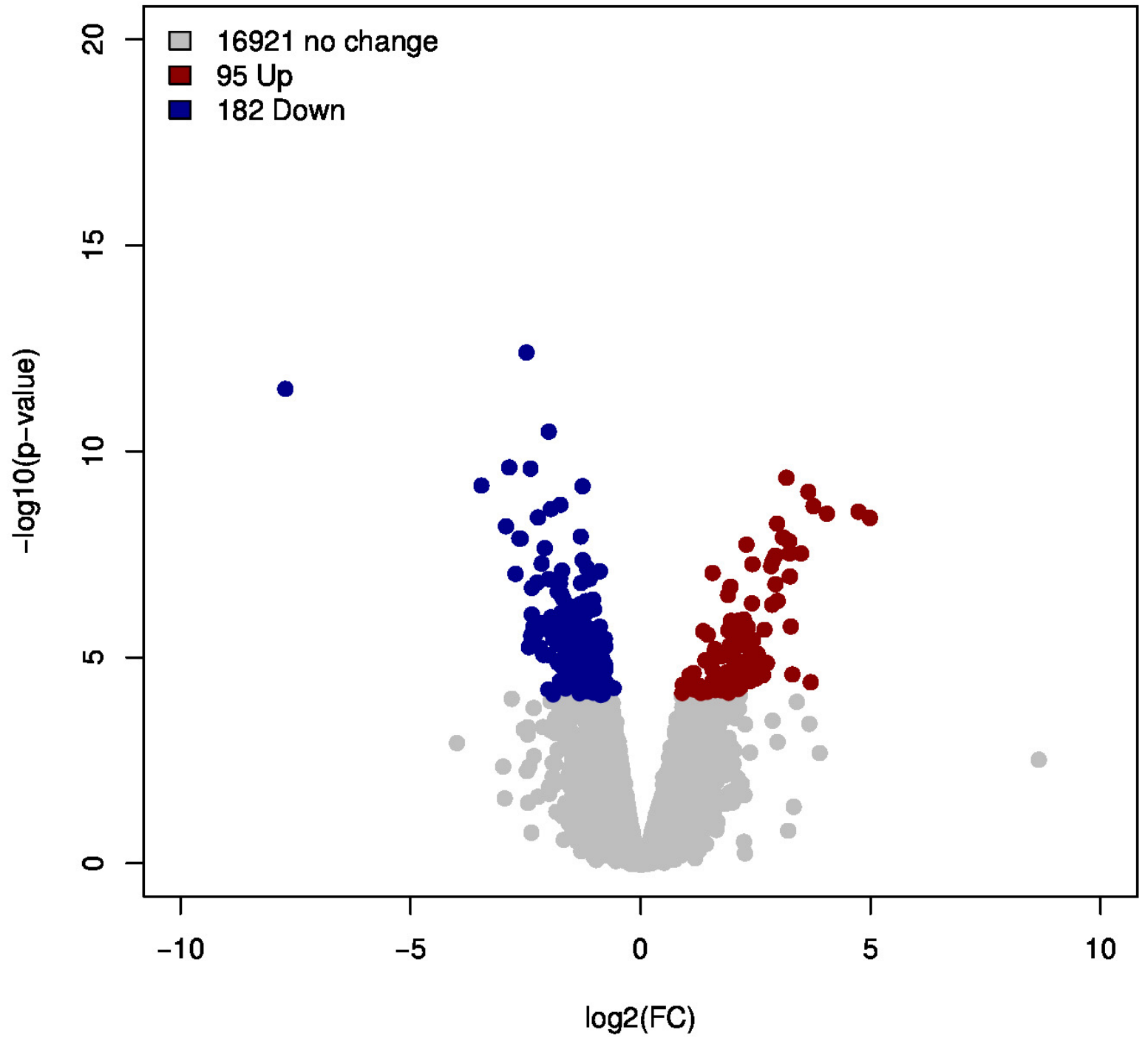
C

Active – Renal Disease vs Inactive



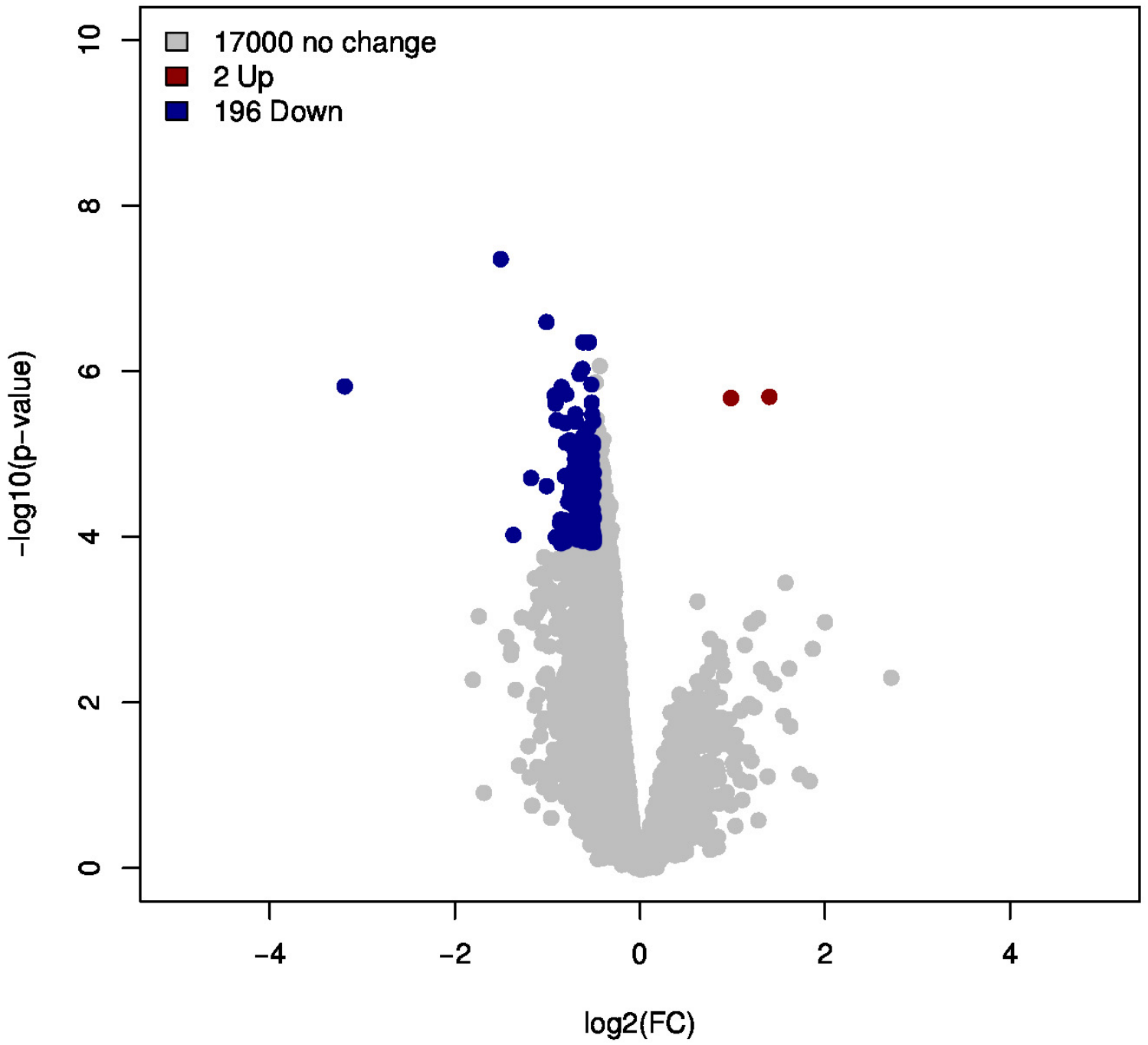
D

Active – Non Renal Disease vs Healthy



E

Active – Non Renal Disease vs Inactive



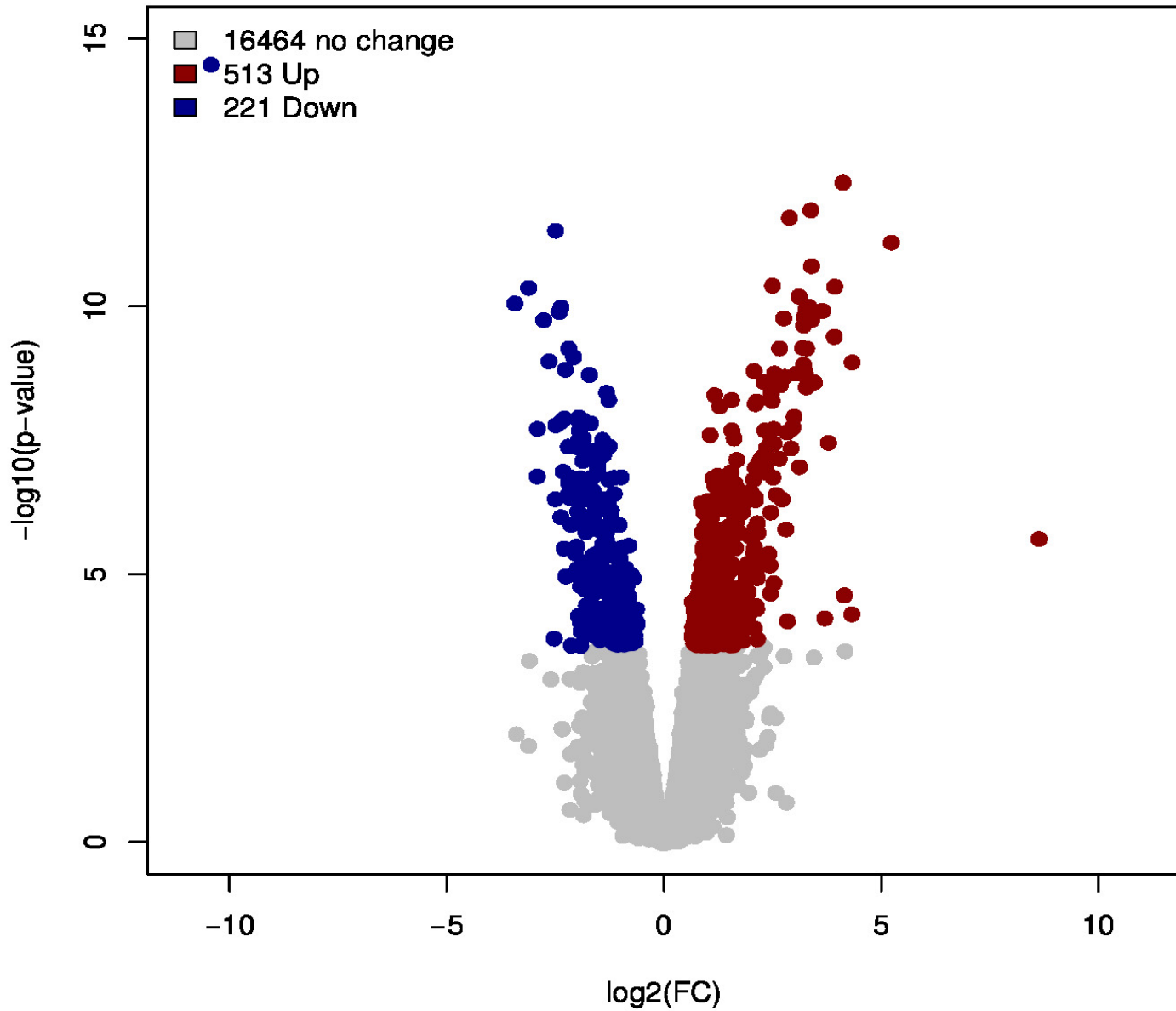
F**Inactive vs Healthy**

Figure 9: Volcano plots illustrating differential expression resulted from different group comparisons.

A

source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (KEGG)					
keg	Influenza A	KEGG:05164	171	376	26	6.59e-05
keg	NOD-like receptor signaling pathway	KEGG:04621	167	376	33	1.66e-09
keg	Osteoclast differentiation	KEGG:04380	125	376	17	2.39e-02
keg	Herpes simplex infection	KEGG:05168	182	376	21	4.74e-02
keg	Transcriptional misregulation in cancer	KEGG:05202	184	376	26	2.74e-04
source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (Reactome)					
rea	Immune System	REAC:168256	2297	544	253	2.99e-38
rea	Cytokine Signaling in Immune system	REAC:1280215	826	544	95	5.00e-12
rea	Interferon Signaling	REAC:913531	197	544	50	1.81e-19
rea	Interferon alpha/beta signaling	REAC:909733	70	544	28	4.72e-16
rea	Interferon gamma signaling	REAC:877300	91	544	28	1.38e-12
rea	Innate Immune System	REAC:168249	1450	544	162	1.06e-21
rea	Protein repair	REAC:5676934	6	544	5	8.04e-04
rea	G1/S-Specific Transcription	REAC:69205	29	544	10	3.70e-04

B

source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (KEGG)					
keg	Chemokine signaling pathway	KEGG:04062	183	94	9	4.81e-02
keg	PI3K-Akt signaling pathway	KEGG:04151	351	94	14	1.31e-02
keg	Cytokine-cytokine receptor interaction	KEGG:04060	268	94	16	2.31e-05
keg	Th17 cell differentiation	KEGG:04659	105	94	7	3.33e-02
keg	ABC transporters	KEGG:02010	44	94	5	1.99e-02
keg	Inflammatory bowel disease (IBD)	KEGG:05321	63	94	6	1.26e-02
keg	Phospholipase D signaling pathway	KEGG:04072	145	94	9	8.43e-03
source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (Reactome)					
rea	Chemokine receptors bind chemokines	REAC:380108	48	129	6	7.61e-03
rea	GPCR ligand binding	REAC:500792	459	129	19	7.94e-04
rea	Class A/1 (Rhodopsin-like receptors)	REAC:373076	325	129	15	2.80e-03
rea	G alpha (i) signalling events	REAC:418594	246	129	13	2.76e-03

4. Results

C

source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (KEGG)					
keg	Influenza A	KEGG:05164	171	41	9	2.50e-05
keg	Transcriptional misregulation in cancer	KEGG:05202	184	41	6	3.88e-02
keg	Amoebiasis	KEGG:05146	94	41	5	1.25e-02
source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (Reactome)					
rea	Immune System	REAC:168256	2297	68	48	1.42e-15
rea	Innate Immune System	REAC:168249	1450	68	26	9.69e-05
rea	Cytokine Signaling in Immune system	REAC:1280215	826	68	24	2.59e-08
rea	Interferon Signaling	REAC:913531	197	68	16	1.90e-11
rea	Interferon alpha/beta signaling	REAC:909733	70	68	13	9.32e-14
rea	Antiviral mechanism by IFN-stimulated genes	REAC:1169410	73	68	6	2.04e-03
rea	ISG15 antiviral mechanism	REAC:1169408	73	68	6	2.04e-03
rea	Extracellular matrix organization	REAC:1474244	296	68	9	3.13e-02
rea	Fibronectin matrix formation	REAC:1566977	6	68	3	1.54e-03

D


source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (KEGG)					
keg	Th17 cell differentiation	KEGG:04659	105	65	6	2.47e-02
keg	Cytokine-cytokine receptor interaction	KEGG:04060	268	65	9	4.42e-02
keg	B cell receptor signaling pathway	KEGG:04662	70	65	6	2.58e-03
keg	Intestinal immune network for IgA production	KEGG:04672	46	65	6	2.18e-04
source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (Reactome)					
rea	Chemokine receptors bind chemokines	REAC:380108	48	88	5	1.45e-02
rea	Antigen activates B Cell Receptor (BCR) leading to generation of second messengers	REAC:983695	100	88	7	5.72e-03

4. Results

E

source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (KEGG)					
keg	NOD-like receptor signaling pathway	KEGG:04621	167	224	23	1.27e-07
keg	Leishmaniasis	KEGG:05140	70	224	10	4.87e-03
keg	Transcriptional misregulation in cancer	KEGG:05202	184	224	19	3.45e-04
keg	Legionellosis	KEGG:05134	55	224	8	2.43e-02
keg	Toll-like receptor signaling pathway	KEGG:04620	102	224	11	2.77e-02
keg	Osteoclast differentiation	KEGG:04380	125	224	15	6.07e-04
keg	Measles	KEGG:05162	133	224	16	2.87e-04
keg	Cytosolic DNA-sensing pathway	KEGG:04623	62	224	9	1.02e-02
keg	Tuberculosis	KEGG:05152	175	224	17	2.55e-03
keg	RIG-I-like receptor signaling pathway	KEGG:04622	70	224	9	2.64e-02
keg	Herpes simplex infection	KEGG:05168	182	224	17	4.22e-03
keg	Influenza A	KEGG:05164	171	224	25	6.20e-09
source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (Reactome)					
rea	TRAF6 mediated IRF7 activation	REAC:933541	33	336	7	2.42e-02
rea	Immune System	REAC:168256	2297	336	196	6.74e-48
rea	Cytokine Signaling in Immune system	REAC:1280215	826	336	71	8.02e-13
rea	Interferon Signaling	REAC:913531	197	336	39	3.45e-18
rea	Interferon alpha/beta signaling	REAC:909733	70	336	24	1.42e-16
rea	Antiviral mechanism by IFN-stimulated genes	REAC:1169410	73	336	11	6.27e-03
rea	ISG15 antiviral mechanism	REAC:1169408	73	336	11	6.27e-03
rea	Interferon gamma signaling	REAC:877300	91	336	20	1.71e-09
rea	Innate Immune System	REAC:168249	1450	336	131	8.64e-30
rea	Toll-Like Receptors Cascades	REAC:168898	152	336	16	8.97e-03
rea	TRAF3-dependent IRF activation pathway	REAC:918233	14	336	5	1.90e-02

F

source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
Biological pathways (KEGG)						
keg	B cell receptor signaling pathway	KEGG:04662	70	70	7	3.50e-04
keg	Type I diabetes mellitus	KEGG:04940	41	70	4	4.98e-02
keg	Graft-versus-host disease	KEGG:05332	37	70	4	3.35e-02
keg	Amoebiasis	KEGG:05146	94	70	6	2.22e-02
keg	Primary immunodeficiency	KEGG:05340	35	70	5	1.53e-03
keg	Intestinal immune network for IgA production	KEGG:04672	46	70	7	1.88e-05
keg	Hematopoietic cell lineage	KEGG:04640	94	70	7	2.48e-03
keg	Allograft rejection	KEGG:05330	35	70	4	2.69e-02
keg	Cytokine-cytokine receptor interaction	KEGG:04060	268	70	11	3.70e-03
Biological pathways (Reactome)						
rea	Chemokine receptors bind chemokines	REAC:380108	48	101	5	2.59e-02
rea	 Adaptive Immune System	REAC:1280218	996	101	22	3.68e-02
rea	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	REAC:198933	199	101	10	5.36e-03
rea	Signaling by the B Cell Receptor (BCR)	REAC:983705	399	101	13	2.82e-02
rea	Antigen activates B Cell Receptor (BCR) leading to generation of second me ...	REAC:983695	100	101	12	4.16e-08
rea	CD22 mediated BCR regulation	REAC:5690714	66	101	9	3.02e-06
rea	G alpha (i) signalling events	REAC:418594	246	101	10	3.24e-02
rea	GPCR ligand binding	REAC:500792	459	101	14	3.01e-02

4. Results

G

source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (KEGG)					
keg	Oocyte meiosis	KEGG:04114	121	133	11	1.06e-03
keg	Cell cycle	KEGG:04110	124	133	23	2.07e-15
keg	Progesterone-mediated oocyte maturation	KEGG:04914	96	133	9	5.30e-03
source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (Reactome)					
rea	RHO GTPase Effectors	REAC:195258	299	216	19	3.50e-03
rea	RHO GTPases Activate Formins	REAC:5663220	123	216	14	6.11e-05
rea	Golgi Cisternae Pericentriolar Stack Reorganization	REAC:162658	14	216	4	4.99e-02
rea	Immune System	REAC:168256	2297	216	75	1.45e-03
rea	Innate Immune System	REAC:168249	1450	216	55	6.06e-04
rea	DNA Replication	REAC:69306	109	216	13	9.84e-05
rea	Cell Cycle	REAC:1640170	632	216	55	1.41e-18
rea	Cell Cycle, Mitotic	REAC:69278	525	216	50	2.57e-18
rea	M Phase	REAC:68886	383	216	26	1.94e-05
rea	Mitotic Prometaphase	REAC:68877	186	216	20	3.45e-07
rea	Condensation of Prometaphase Chromosomes	REAC:2514853	12	216	4	2.55e-02
rea	Resolution of Sister Chromatid Cohesion	REAC:2500257	108	216	18	1.54e-09
rea	Mitotic Metaphase and Anaphase	REAC:2555396	185	216	19	2.00e-06
rea	Mitotic Anaphase	REAC:68882	184	216	19	1.83e-06
rea	Separation of Sister Chromatids	REAC:2467813	173	216	19	6.47e-07
rea	S Phase	REAC:69242	140	216	15	4.92e-05
rea	Synthesis of DNA	REAC:69239	101	216	12	2.99e-04
rea	DNA strand elongation	REAC:69190	32	216	7	9.80e-04
rea	Unwinding of DNA	REAC:176974	12	216	5	8.16e-04
rea	Mitotic G1-G1/S phases	REAC:453279	148	216	19	4.34e-08
rea	G0 and Early G1	REAC:1538133	27	216	8	1.39e-05
rea	G1/S Transition	REAC:69206	130	216	16	2.50e-06
rea	G1/S-Specific Transcription	REAC:69205	29	216	11	1.59e-09
rea	Activation of the pre-replicative complex	REAC:68962	32	216	6	1.33e-02
rea	Regulation of mitotic cell cycle	REAC:453276	87	216	11	4.63e-04
rea	APC/C-mediated degradation of cell cycle proteins	REAC:174143	87	216	11	4.63e-04
rea	Activation of APC/C and APC/C:Cdc20 mediated degradation of mitotic p ...	REAC:176814	77	216	10	1.11e-03
rea	APC/C:Cdc20 mediated degradation of mitotic proteins	REAC:176409	76	216	9	7.50e-03
rea	Mitotic G2-G2/M phases	REAC:453274	187	216	16	4.12e-04
rea	G2/M Transition	REAC:69275	185	216	15	1.75e-03
rea	Polo-like kinase mediated events	REAC:156711	16	216	8	9.79e-08
rea	Cell Cycle Checkpoints	REAC:69620	293	216	30	5.73e-11
rea	G2/M Checkpoints	REAC:69481	169	216	14	2.86e-03
rea	Activation of ATR in response to replication stress	REAC:176187	37	216	7	2.75e-03
rea	G2/M DNA replication checkpoint	REAC:69478	5	216	4	2.87e-04
rea	Mitotic Spindle Checkpoint	REAC:69618	111	216	15	2.05e-06
rea	Amplification of signal from the kinetochores	REAC:141424	95	216	14	2.13e-06
rea	Amplification of signal from unattached kinetochores via a MAD2 in ...	REAC:141444	95	216	14	2.13e-06
rea	Phosphorylation of Emi1	REAC:176417	6	216	4	8.49e-04
rea	Activation of NIMA Kinases NEK9, NEK6, NEK7	REAC:2980767	7	216	4	1.95e-03

4. Results

H

source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (KEGG)					
keg	Cell cycle	KEGG:04110	124	67	13	5.14e-09
keg	p53 signaling pathway	KEGG:04115	68	67	5	2.89e-02
keg	Progesterone-mediated oocyte maturation	KEGG:04914	96	67	7	1.97e-03
keg	Oocyte meiosis	KEGG:04114	121	67	7	8.70e-03
source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (Reactome)					
rea	Immune System	REAC:168256	2297	131	50	2.95e-03
rea	Innate Immune System	REAC:168249	1450	131	42	1.15e-05
rea	Fcγ receptor (FCγR) dependent phagocytosis	REAC:2029480	165	131	11	1.80e-03
rea	Regulation of actin dynamics for phagocytic cup formation	REAC:2029482	140	131	11	3.60e-04
rea	Role of phospholipids in phagocytosis	REAC:2029485	103	131	11	1.55e-05
rea	FCγR activation	REAC:2029481	90	131	11	3.71e-06
rea	Complement cascade	REAC:166658	134	131	12	2.70e-05
rea	Initial triggering of complement	REAC:166663	99	131	11	1.02e-05
rea	Creation of C4 and C2 activators	REAC:166786	92	131	11	4.69e-06
rea	Classical antibody-mediated complement activation	REAC:173623	84	131	11	1.76e-06
rea	Regulation of Complement cascade	REAC:977606	123	131	12	1.03e-05
rea	Binding and Uptake of Ligands by Scavenger Receptors	REAC:2173782	118	131	12	6.45e-06
rea	Scavenging of heme from plasma	REAC:2168880	89	131	10	4.01e-05
rea	DNA Replication	REAC:69306	109	131	8	1.84e-02
rea	FCER1 mediated NF-κB activation	REAC:2871837	154	131	9	3.72e-02
rea	Nucleosome assembly	REAC:774815	73	131	7	9.77e-03
rea	Deposition of new CENPA-containing nucleosomes at the centromere	REAC:606279	73	131	7	9.77e-03
rea	Cell surface interactions at the vascular wall	REAC:202733	212	131	15	1.45e-05
rea	Association of licensing factors with the pre-replicative complex	REAC:69298	15	131	4	8.83e-03
rea	Signaling by Rho GTPases	REAC:194315	429	131	18	1.52e-03
rea	RHO GTPase Effectors	REAC:195258	299	131	17	4.31e-05
rea	RHO GTPases Activate Formins	REAC:5663220	123	131	13	9.94e-07
rea	CD22 mediated BCR regulation	REAC:5690714	66	131	7	5.01e-03
rea	E2F-enabled inhibition of pre-replication complex formation	REAC:113507	9	131	3	4.78e-02
rea	FCER1 mediated Ca ²⁺ mobilization	REAC:2871809	105	131	9	1.75e-03
rea	Cell Cycle	REAC:1640170	632	131	48	1.08e-23
rea	Cell Cycle Checkpoints	REAC:69620	293	131	26	3.84e-13
rea	G2/M Checkpoints	REAC:69481	169	131	12	3.39e-04
rea	G2/M DNA replication checkpoint	REAC:69478	5	131	3	5.90e-03
rea	Mitotic Spindle Checkpoint	REAC:69618	111	131	13	2.75e-07
rea	Amplification of signal from the kinetochores	REAC:141424	95	131	13	3.77e-08
rea	Amplification of signal from unattached kinetochores via a MAD2 in ...	REAC:141444	95	131	13	3.77e-08
rea	Cell Cycle, Mitotic	REAC:69278	525	131	43	4.31e-22
rea	Mitotic G1-G1/S phases	REAC:453279	148	131	13	9.54e-06
rea	G0 and Early G1	REAC:1538133	27	131	6	2.41e-04
rea	G1/S Transition	REAC:69206	130	131	11	1.71e-04
rea	G1/S-Specific Transcription	REAC:69205	29	131	9	1.27e-08
rea	Mitotic G2-G2/M phases	REAC:453274	187	131	11	5.93e-03
rea	G2/M Transition	REAC:69275	185	131	11	5.36e-03
rea	Polo-like kinase mediated events	REAC:156711	16	131	5	3.34e-04
rea	S Phase	REAC:69242	140	131	10	2.67e-03
rea	M Phase	REAC:68886	383	131	24	1.05e-08
rea	Mitotic Prometaphase	REAC:68877	186	131	20	2.79e-11
rea	Condensation of Prometaphase Chromosomes	REAC:2514853	12	131	4	3.30e-03
rea	Resolution of Sister Chromatid Cohesion	REAC:2500257	108	131	16	6.57e-11
rea	Mitotic Metaphase and Anaphase	REAC:2555396	185	131	16	2.71e-07
rea	Mitotic Anaphase	REAC:68882	184	131	16	2.50e-07
rea	Separation of Sister Chromatids	REAC:2467813	173	131	16	9.98e-08

4. Results

source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
Biological pathways (KEGG)						
keg	Oxidative phosphorylation	KEGG:00190	132	118	12	1.10e-04
keg	Parkinson's disease	KEGG:05012	141	118	12	2.22e-04
keg	Carbon metabolism	KEGG:01200	116	118	9	8.55e-03
keg	Fc gamma R-mediated phagocytosis	KEGG:04666	89	118	7	4.70e-02
Biological pathways (Reactome)						
rea	The citric acid (TCA) cycle and respiratory electron transport	REAC:1428517	171	141	13	1.21e-04
rea	Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and ...	REAC:163200	126	141	10	1.93e-03
rea	Respiratory electron transport	REAC:611105	103	141	8	1.99e-02
rea	The role of Nef in HIV-1 replication and disease pathogenesis	REAC:164952	29	141	5	1.10e-02
rea	Hemostasis	REAC:109582	708	141	23	1.54e-02
rea	Platelet activation, signaling and aggregation	REAC:76002	281	141	17	5.03e-05
rea	Response to elevated platelet cytosolic Ca ²⁺	REAC:76005	136	141	11	5.39e-04
rea	Platelet degranulation	REAC:114608	131	141	11	3.72e-04
rea	Immune System	REAC:168256	2297	141	55	4.67e-04
rea	Innate Immune System	REAC:168249	1450	141	39	2.15e-03
rea	tRNA processing	REAC:72306	146	141	10	7.09e-03
rea	tRNA processing in the mitochondrion	REAC:6785470	42	141	10	3.77e-08

Table 4: KEGG and Reactome pathways enriched in overexpressed and underexpressed genes of Active-Renal (A,B), Active-nonRenal (C,D), Inactive (E,F), pathways enriched in the overexpressed genes resulted from Active-Renal vs Active-nonRenal (G) and Active-Renal vs Inactive (H), and pathways enriched in the underexpressed genes resulted from Active-nonRenal vs Inactive (I).

4.2 Modular Analysis

Another approach was followed for a more comprehensive functional interpretation of expression deregulation in the different patient groups. That is the so called ‘modular analysis’. It involves a gene set enrichment – like analysis, which may overcome restrictions posed by an analysis based on the use of p-value threshold for DEG identification, a hypergeometric test for functional enrichment and correction for multiple testing. The gene sets used were the ‘blood modules’, constructed by computational analysis based on expression data derived from blood tissue samples³⁶. Blood modules comprise genes, whose expression co-cluster in multiple experiments related to immune system. The size of those modules in terms of number of genes is smaller than pathway ontologies used so far. Hence, blood modules may be more appropriate for the current study and could result in more specific signatures.

The results of the modular analysis are illustrated in detail in Supplemented Figure 1. They seem to agree with the results described in the previous section. A plethora of deregulated modules was identified, comprising an SLE related ‘module profile’. The majority of them were overexpressed, with the most prominent ones including cell cycle, cell death and extracellular matrix signatures and innate immunity related signatures, namely IFN, inflammation, neutrophil, dendritic cell and cytosolic DNA sensing signatures (Figure 10). There are underexpressed modules as well, mainly associated with B cell, plasma cell and NK cell signatures. It is really interesting to examine the profile of the aforementioned signature in the different patient subgroups (Figure 11). There are those that are very similar between the different groups and those that diversify, in terms of complete presence or absence from a group, or the amount of the gene members of the module which are deregulated. The profile of inflammation, dendritic cell and cytosolic DNA sensing is approximately the same in the different DA groups. Neutrophil and extracellular matrix signature profiles are also similar between the DA groups. However, they are overexpressed in DA3 compared to DA1 as well, which imply that while the corresponding genes are deregulated independently of disease activity, the levels of deregulation are greater in patients with increased activity. Regarding IFNs, there are multiple blood modules, from which some remain stable throughout DA and some are enhanced, as in higher DA groups. Cell death signature fluctuate, with DA1 being the most extreme and DA2 the least extreme. Intriguingly, B cell module underexpression is ‘weakened’ as DA increases and plasma cell signature is overexpressed in DA3 compared to DA1. A plasmablast signature was identified as the most robust biomarker of DA in a study of longitudinal blood transcriptomic data in pediatric lupus cases⁷. Finally, cell cycle modules are deregulated only in the DA3 group and are deregulated even when DA3 is compared to DA1. Modular analysis of the data split according to renal manifestation provided similar results, with renal active group resembling DA3 group (Supplemented Figure 1).

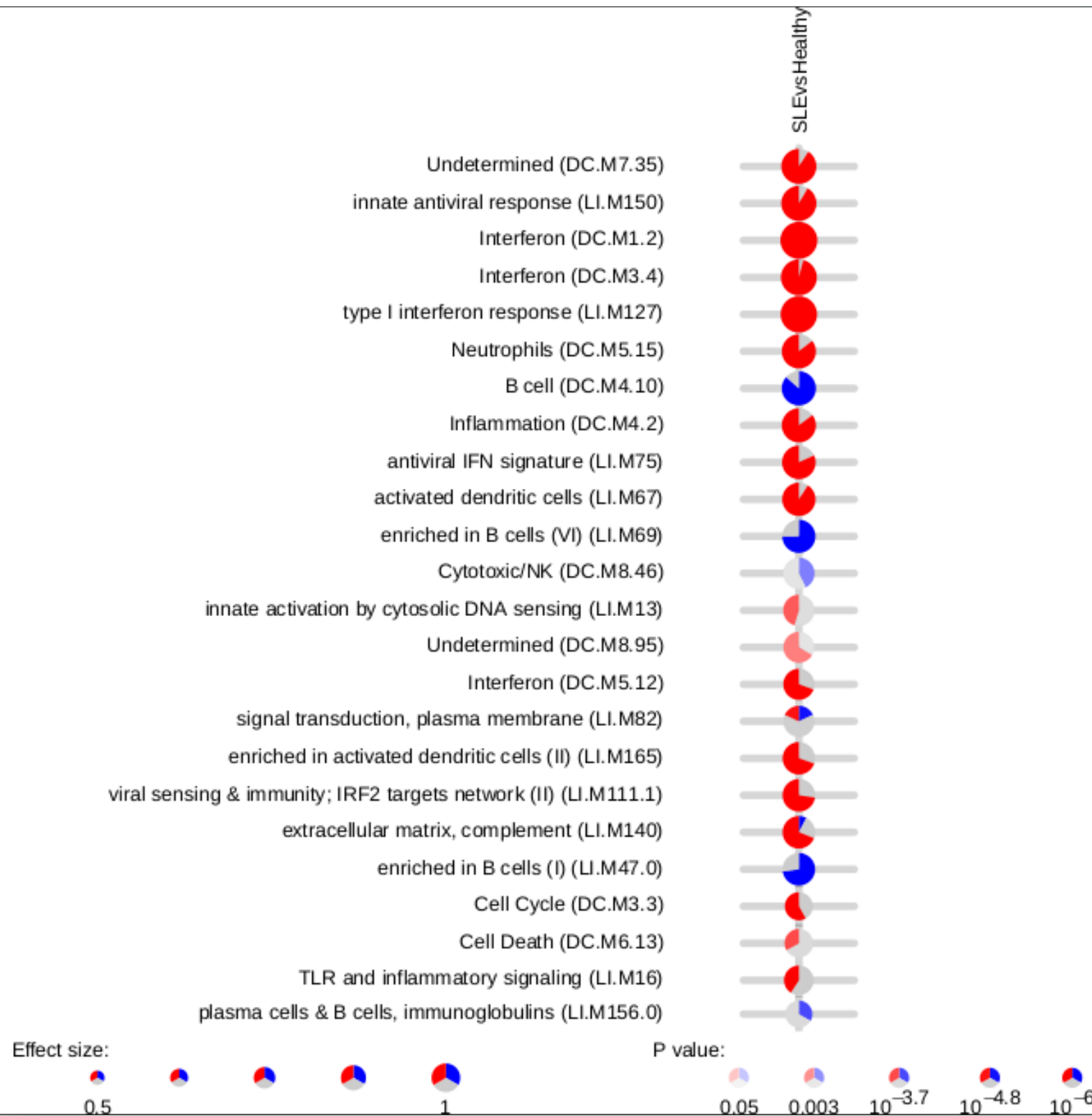


Figure 10: Modular analysis for the differential expression of SLE patients. Modules are represented by pies. The proportion of module genes, which have been detected as DEGs, is indicated by the coloured portion of the pies. Red indicates overexpression and blue indicates underexpression.

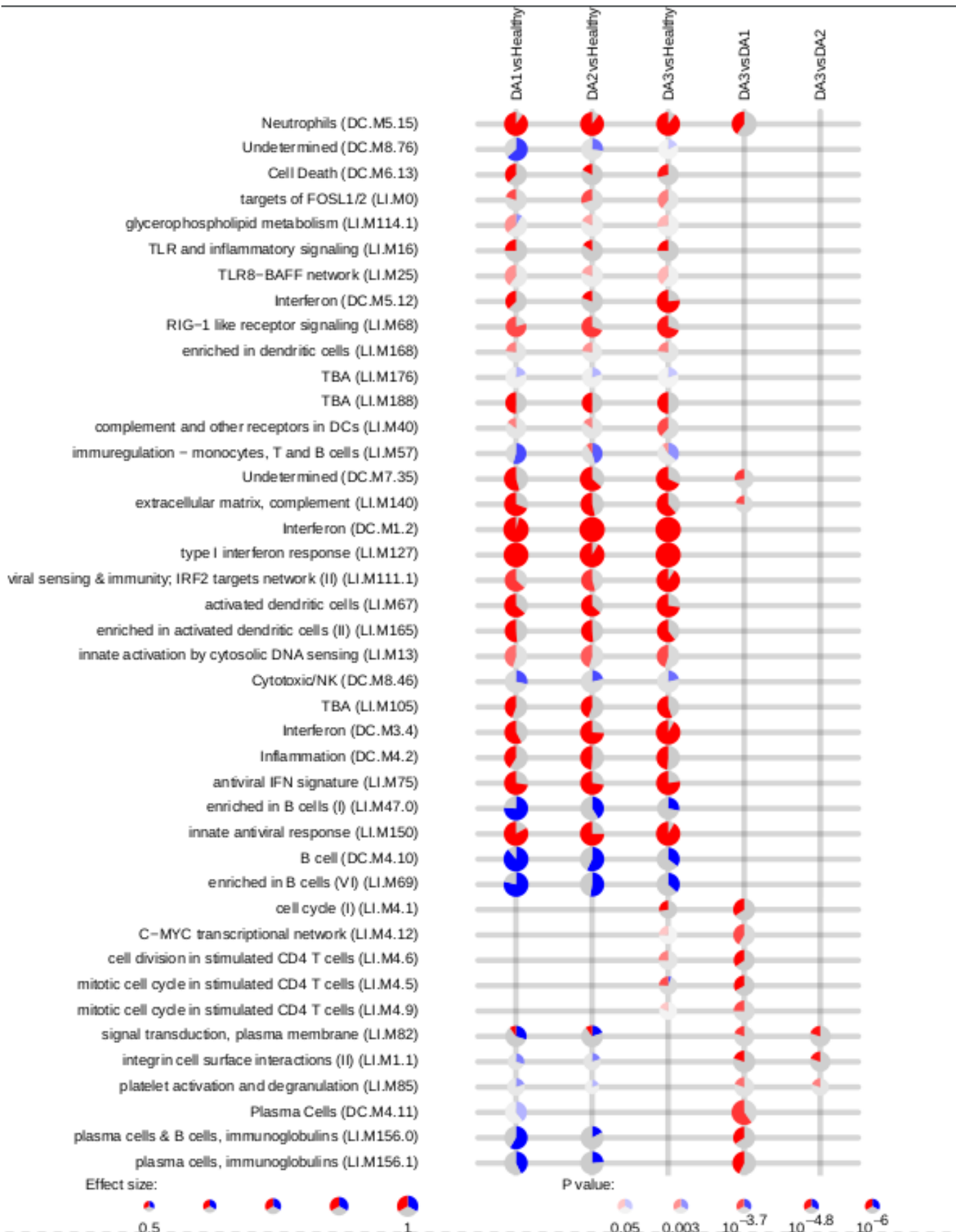


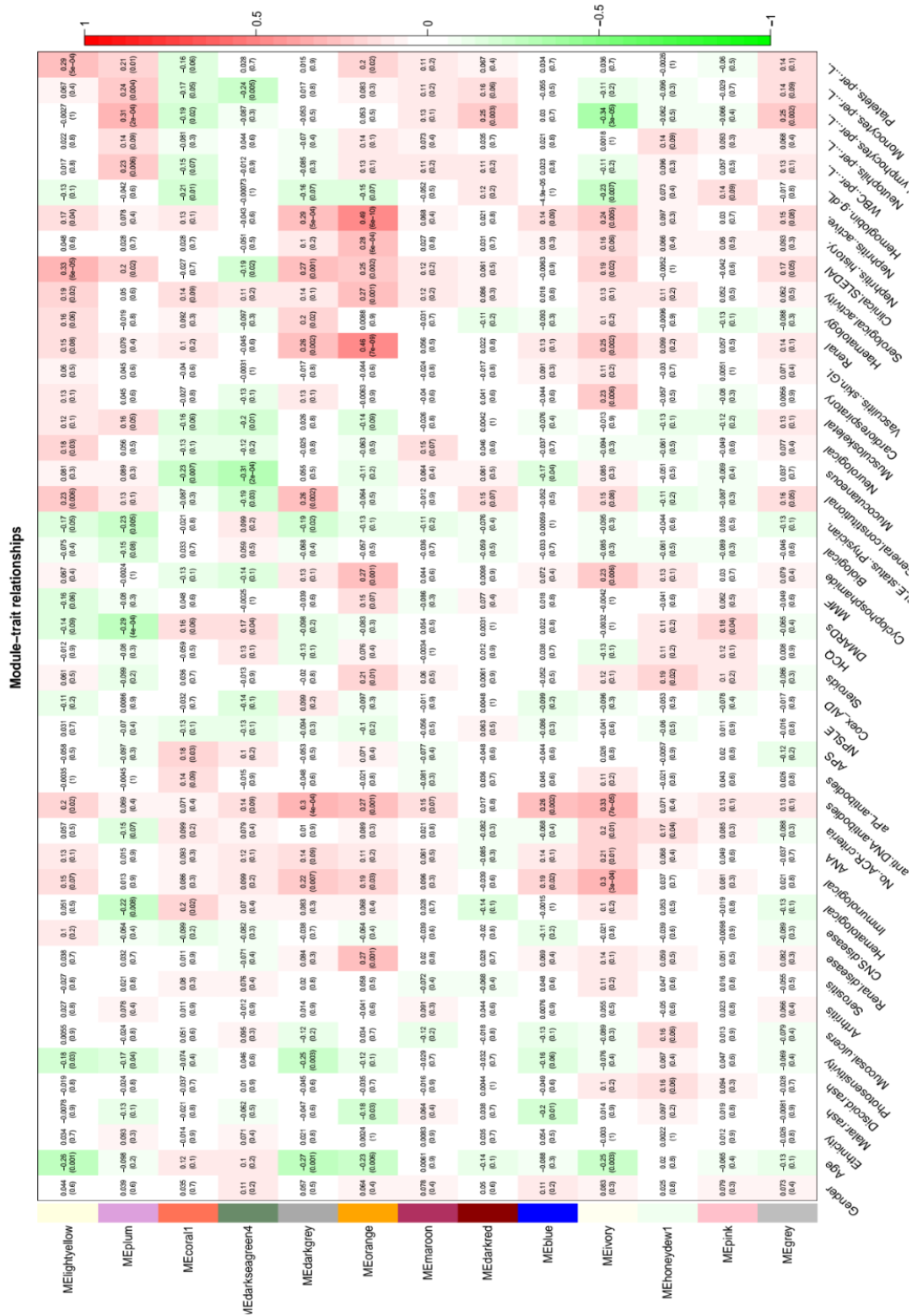
Figure 11: Modular analysis for the differential expression of DA groups. Modules are represented by pies. The proportion of module genes, which have been detected as DEGs, is indicated by the coloured portion of the pies. Red indicates overexpression and blue indicates underexpression.

4.3 WGCNA analysis

Weighted gene co-expression network analysis is an alternative to classical differential expression analysis for investigation and interpretation of expression data. WGCNA modules are identified *de novo*, based on the clustering of a network constructed by exploiting weighted correlation of gene expression in the studied dataset. Here, 13 modules were detected in the patient cohort, which were named after a colour randomly assigned to them. The significance of each module was assessed on the basis of the correlation between the module eigengene, used as a representative variable for the module, and a number of traits recorded for every patient, such as age, SLEDAI activity index, different manifestations and given treatments (Figure 12, Supplemented Figure 2). At the same time, WGCNA modules were also explored for any enriched pathway. The ‘Lightyellow’ module, the one with the highest significant correlation with SLEDAI, comprising 224 genes, was enriched for innate and adaptive immune pathways, mainly signalling through FC and BCR. Interestingly, that module intersected with the B cell signature identified during modular analysis. The ‘Orange’ module comprising 184 genes was pinpointed as the most highly correlated with renal manifestations and active nephritis and was enriched for neutrophil activation and neutrophil degranulation biological processes. Consistently, neutrophil activity has been previously associated with nephritis in the literature. The ‘Ivory’ module was highly linked with antinuclear and anti-DNA antibodies presence. It included 282 genes enriched for IFN signalling. The ‘Darkgrey’ module had the second greatest association with anti-DNA antibodies and active nephritis. It contained 245 genes enriched for cell cycle, P53 signalling, DNA repair and cellular senescence pathways. In general, modules that appeared to be linked with a particular trait comprised at most a few hundred genes, whereas modules that demonstrated no correlation involved a few thousand genes.

Figure 12: Heatmap representation of the correlations between WGCNA modules and a variety of traits recorded from all the patients. Red indicates positive correlation and green indicates negative correlation.

4. Results



4.4 Topological analysis

4.4.1 COD profiling

Topological organization of gene expression was explored in the form of domain co-expression analyses and at the focus of our study. Co-expression domains (CODs) were defined as regions of consecutive, gene containing chromosomal bins, which have higher than average correlation of expression among them, delimited by statistically significant borders. The exact pipeline implemented to define, detect, and analyse CODs is described in detail in ‘Methods’ section. COD organization was studied in the three DA groups and compared to CODs of the healthy group. First, CODs were identified in each group. Approximately, 460 CODs were detected in healthy control group (Figure 13). The two chromosomes with the greatest COD abundance were chromosome 1, the largest chromosome and chromosome 19, that is, expectedly, the most gene-dense chromosome (Figure 14).

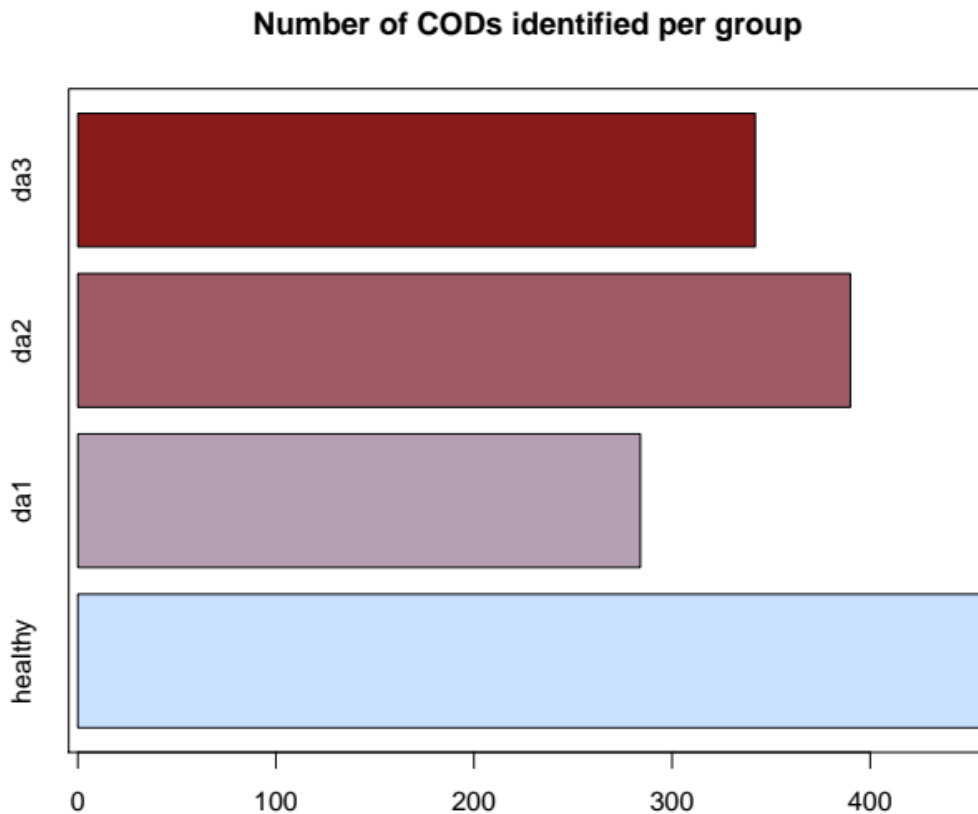


Figure 13: Barplot depicting the total number of CODs detected in control healthy group and in the three DA patient groups.

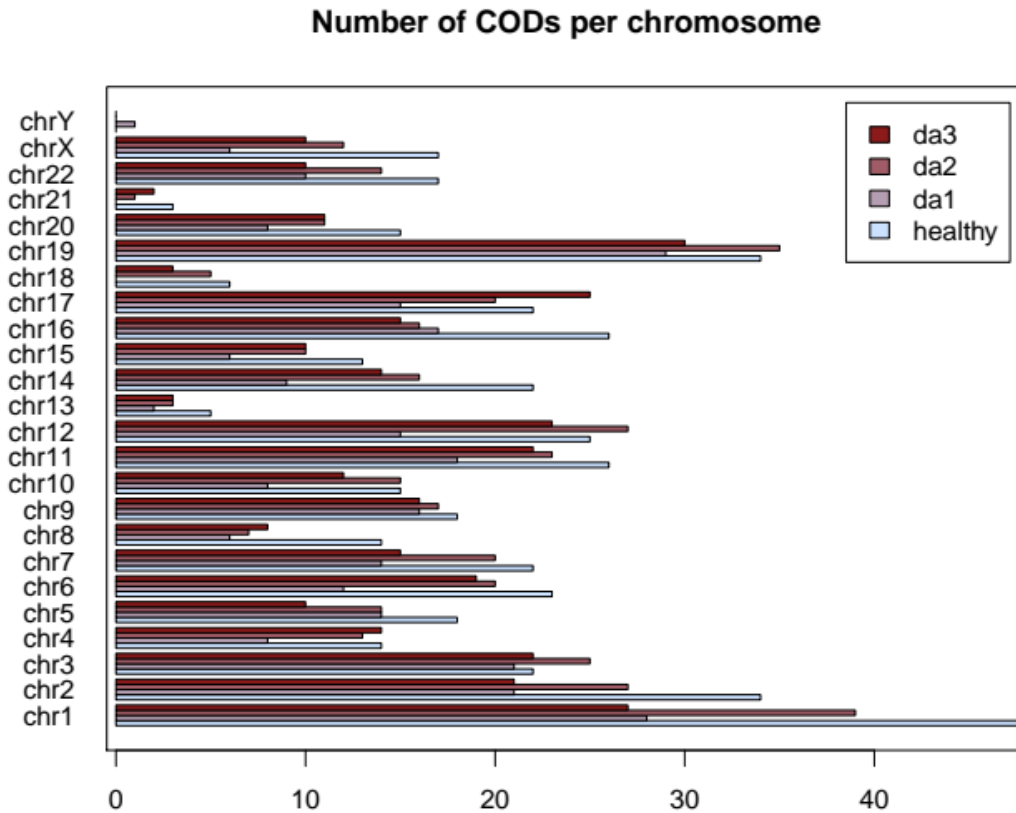


Figure 14: Barplot depicting the total number of CODs per chromosome detected in control healthy group and in the three DA patient groups.

Subsequently, we reported a number of different COD characteristics, in order to monitor, characterize and compare their distribution. Particularly, these variables included the total number of CODs (Figure 13), the size of CODs (Figure 15) and the average intra (Figure 16) and inter COD co-expression (Figure 17). Moreover, the proportion of chromosomal bins, in which expressed genes are located, and that reside inside CODs was documented (Figure 18). A smaller number and mean size of CODs in patients vs controls, strongly suggested different topological organization of gene expression, with patient expression profiles being more "fragmented". This was also supported by the smaller percentage of chromosomal bins that resided inside CODs (Figure 18). In other words, regions that are transcriptionally active in patient profiles are less topologically correlated. Interestingly, the greatest extent of changes is observed in patients with the lowest disease activity, an observation that could prove valuable in pinpointing factors underlying early disease onset. It should be noted that inactive patients were previously active and went into remission as a result of treatment. Hence

despite the lack of disease manifestation, their transcriptional profile remains disrupted.

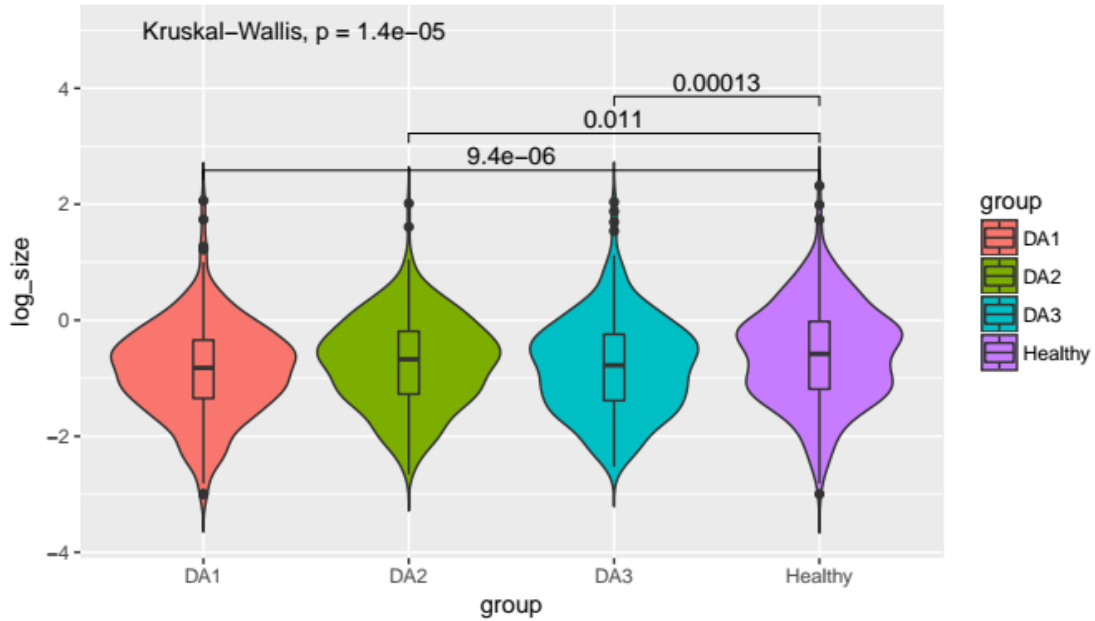


Figure 15: Violin plots illustrating the estimated distribution of COD sizes in each group, classic boxplots are included, the scale is logarithmic ($\log(\text{bp})$) and the results of Mann-Whitney-Wilcoxon tests comparing each one of patient groups with healthy group are represented by p-values.

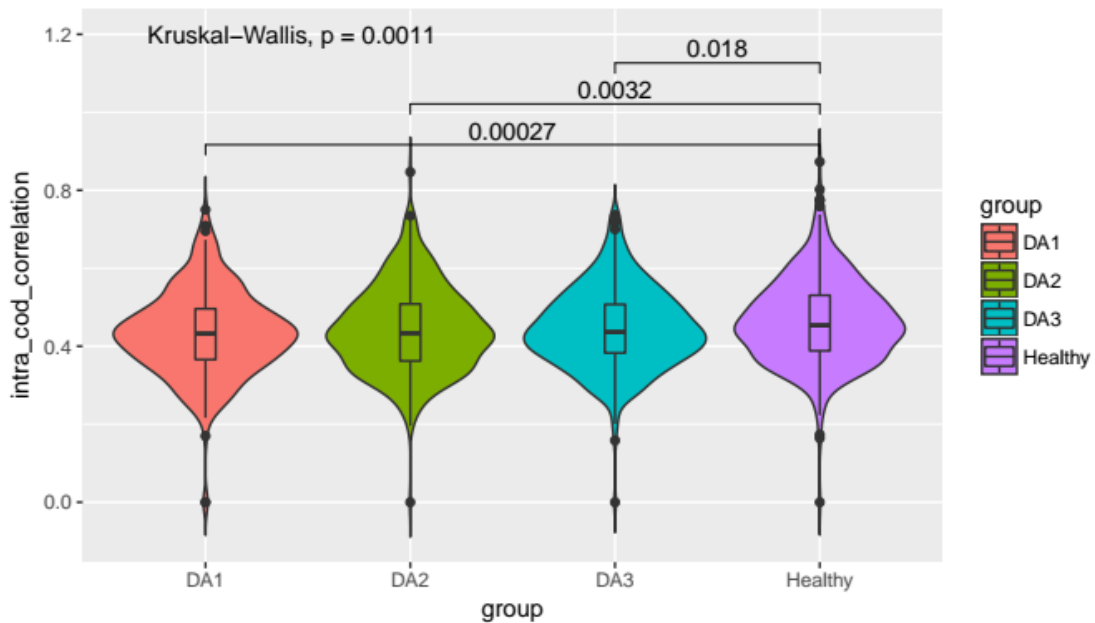


Figure 16: Average intra COD correlation estimated distribution summarized in violin plots. Classic boxplots are included and the results of Mann-Whitney-Wilcoxon tests comparing each one of patient groups with healthy group are represented by p-values.

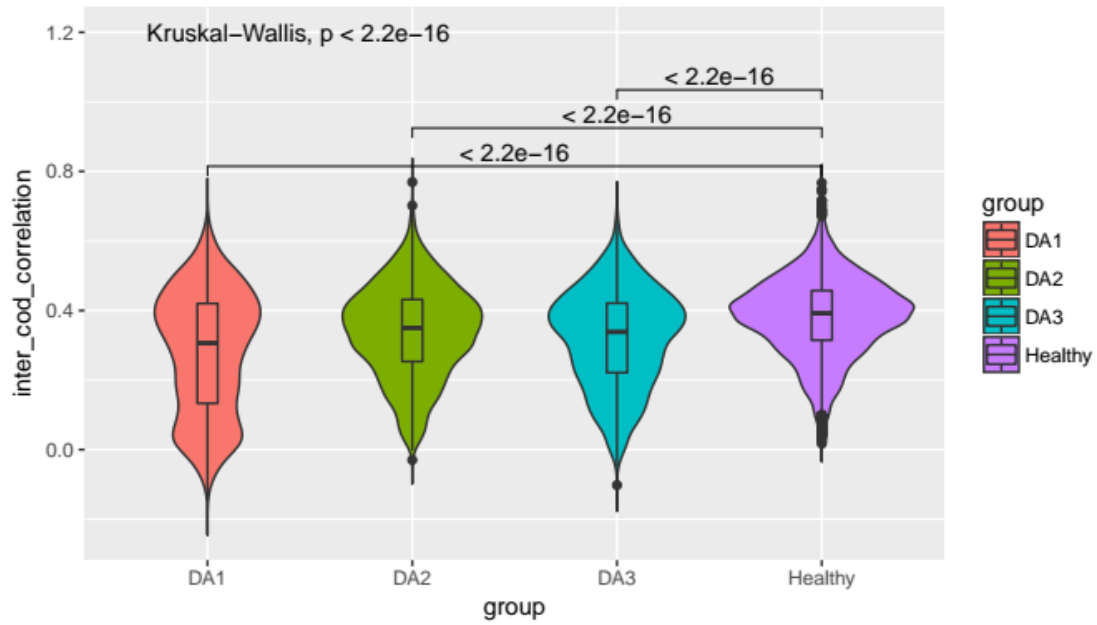


Figure 17: Average inter COD correlation estimated distribution summarized in violin plots. Classic boxplots are included and the results of Mann-Whitney-Wilcoxon tests comparing each one of patient groups with healthy group are represented by p-values.

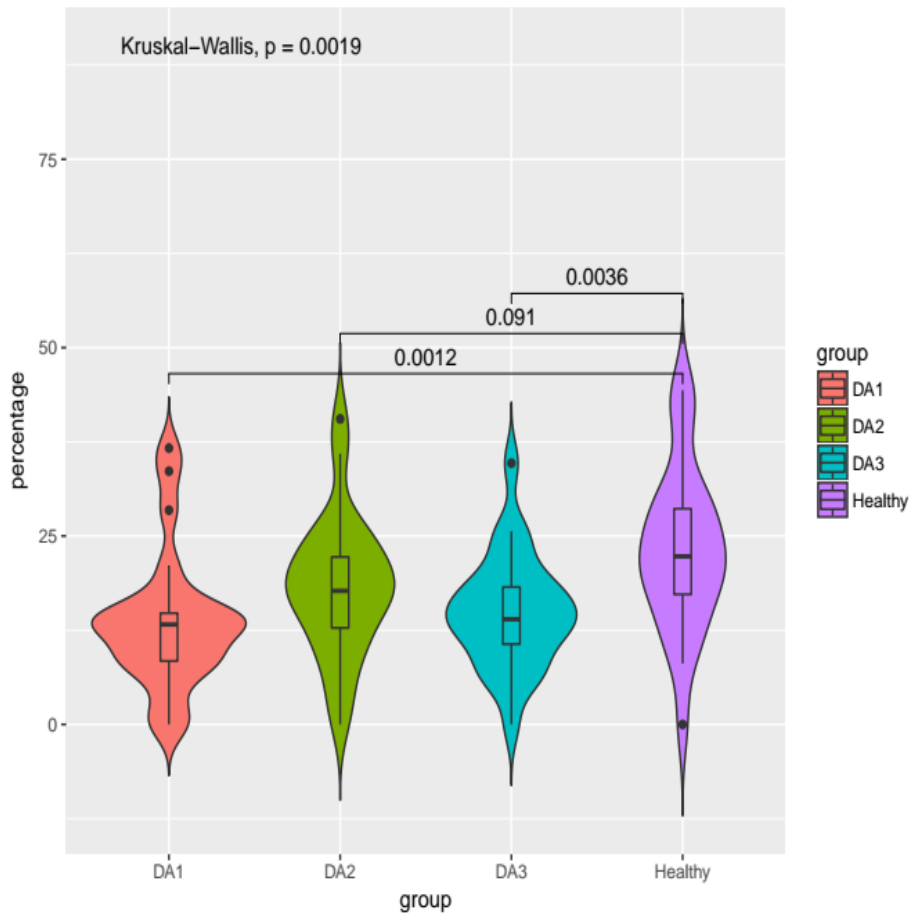


Figure 18: Violin plots illustrating the estimated distribution of chromosomal bins percentage, in which expressed genes are located and that reside inside CODs, classic boxplots are included and the results of Mann-Whitney-Wilcoxon tests comparing each one of patient groups with healthy group are represented by p-values.

Concerning the levels of co-expression, intra-COD average correlation values also appear to differentiate in patient groups. Surprisingly, though the latter statement is true when examined in a genome scale (Figure 16), if intra – COD correlation distribution is examined in a per chromosome basis, only in four chromosomes seems to be significantly lower (Supplemented Figure 3). In two of these cases the group that discriminate from healthy is DA1. On the other hand, average correlation of expression between chromosomal bins contained in two different patient CODs is significantly different in most chromosomes from the corresponding variable in healthy group (Supplemented Figure 4). Soler-Oliva et al made the same observation when comparing CODs from breast cancer tissue to control healthy CODs²³. However, it should be mentioned that the total amount of average inter – COD correlation values is greater than the corresponding intra – COD values, because every possible intra-chromosomal COD pair is checked, and though statistical significance is obvious, further study is re-

quired to verify biological significance. At the same time, differences are observed among the different DA groups concerning the aforementioned measured variables, suggesting a link between gene expression organization and disease development.

The observed aberrations could be partially explained by a higher variability in blood cell composition of patients. Since whole blood samples were used, a greater variability in cell type populations could be the culprit of lower correlations at transcriptional level. In order to test that, the proportion of different immune cell types in whole blood was estimated for every individual (see Methods section for details). Subsequently, proportions were handled as probabilities and entropy was calculated for every subject as a measure of the variability/uncertainty of blood cell composition. The distribution of entropy values differed significantly only between DA1 and healthy groups. Nevertheless, DA1 had smaller median entropy value (Figure 19). Consequently, the observed abnormalities cannot be attributed to cell-type dependent expression variability, as in fact greater fragmentation in CODs is derived from samples with comparable cell-type population variance.

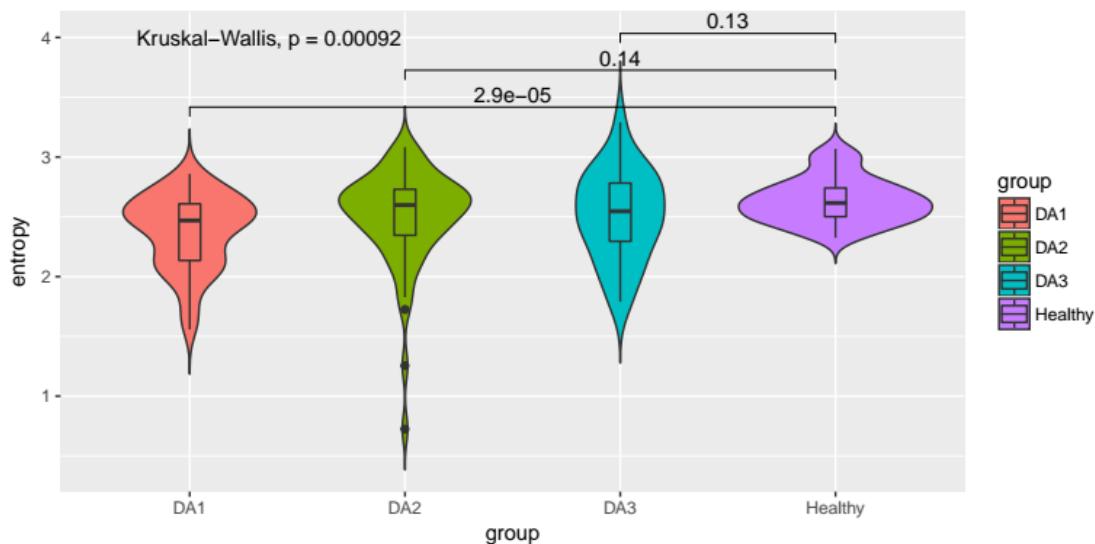


Figure 19: Violin plot representation of the entropy calculated using the different immune cell proportions per individual. Classic boxplots are included and the results of Mann-Whitney-Wilcoxon tests comparing each one of patient groups with healthy group are represented by p -values.

4.4.2 Structural COD changes

A more detailed inspection revealed different cases of COD alterations. That was performed with the healthy COD set designated as a reference. When the corresponding coordinates of a healthy COD are explored in a DA group, the possible outcomes are for a COD or a segment of a COD to be encountered, or not. That is because a COD of the healthy group could have been completely absent from a patient group or its borders could have been altered resulting in a COD that retain an overlap with the healthy control COD. In addition, a COD could have been identified in a patient

group, which had not existed in the healthy COD profile. Indeed, all those cases were discovered (Supplemented Figure 5). COD changes were systematically classified (Table 1). Specifically, COD alteration categories are ‘depleted’, ‘emerged’, ‘intact’, and left, right or both ‘borders shifted’. More complex phenomena were described as well. A COD could also split to more, or multiple CODs could merge to one. The normalized number of CODs altered in a particular way in each DA group are demonstrated in the heatmap of Figure 20. DA1 had relatively the biggest amount of split CODs and DA2 had the biggest amount of intact CODs. As far as the depleted CODs are concerned, DA2 had the least amount and DA1 and DA3 had approximately the same number. In an endeavour to quantify COD rearrangements for patient CODs that had an overlap with healthy CODs, two different metrics were used. Both metrics agreed in DA2 being ‘closer’ to the healthy group relatively to the rest of patient groups and absence of any statistically significant difference between DA1 and DA3 (Figures 21, 22).

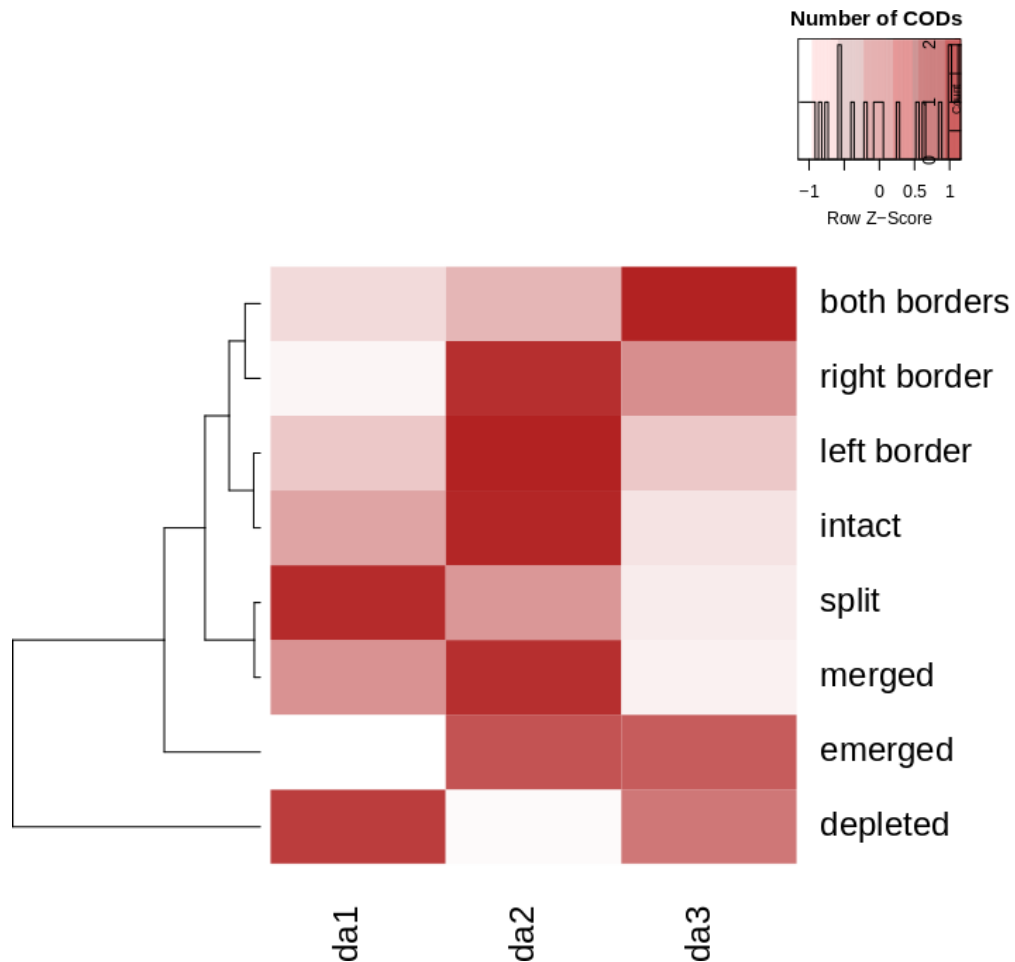


Figure 20: Heatmap highlighting the number of CODs in healthy group, that are altered in a particular way when the corresponding genomic area is examined in patient groups, a normalization per change category (row) has been performed and change categories have been grouped based on a hierarchical clustering.

Results so far suggested that gene expression correlation, even though quite extensive, may be sharply disrupted in ways that are associated with disease development. In this respect we next set out to investigate whether changes in the co-expression patterns could be attributed to particular genes with differential expression status. The hypothesis here is that differentially expressed genes may be drivers of aberrant expression patterns extending in much wider areas and thus act as disruptive agents for co-expression domains. In order to do this we calculated the proportion of DEGs residing inside a COD region for all studied groups. So as to statistically evaluate the differences in DEG inclusion between a patient COD set and healthy COD set, a bootstrap based approach was implemented with 10000 iterations. DA1 and DA3 groups were significantly distinctive from healthy control group with a bootstrap p-value of 0,0028 and 0 respectively (Figure 23). Therefore, at those groups the observed disruptions in the organization of gene expression is definitely linked to and may be partially explained by the activity of DEGs. Consistently, there were cases, in which CODs are

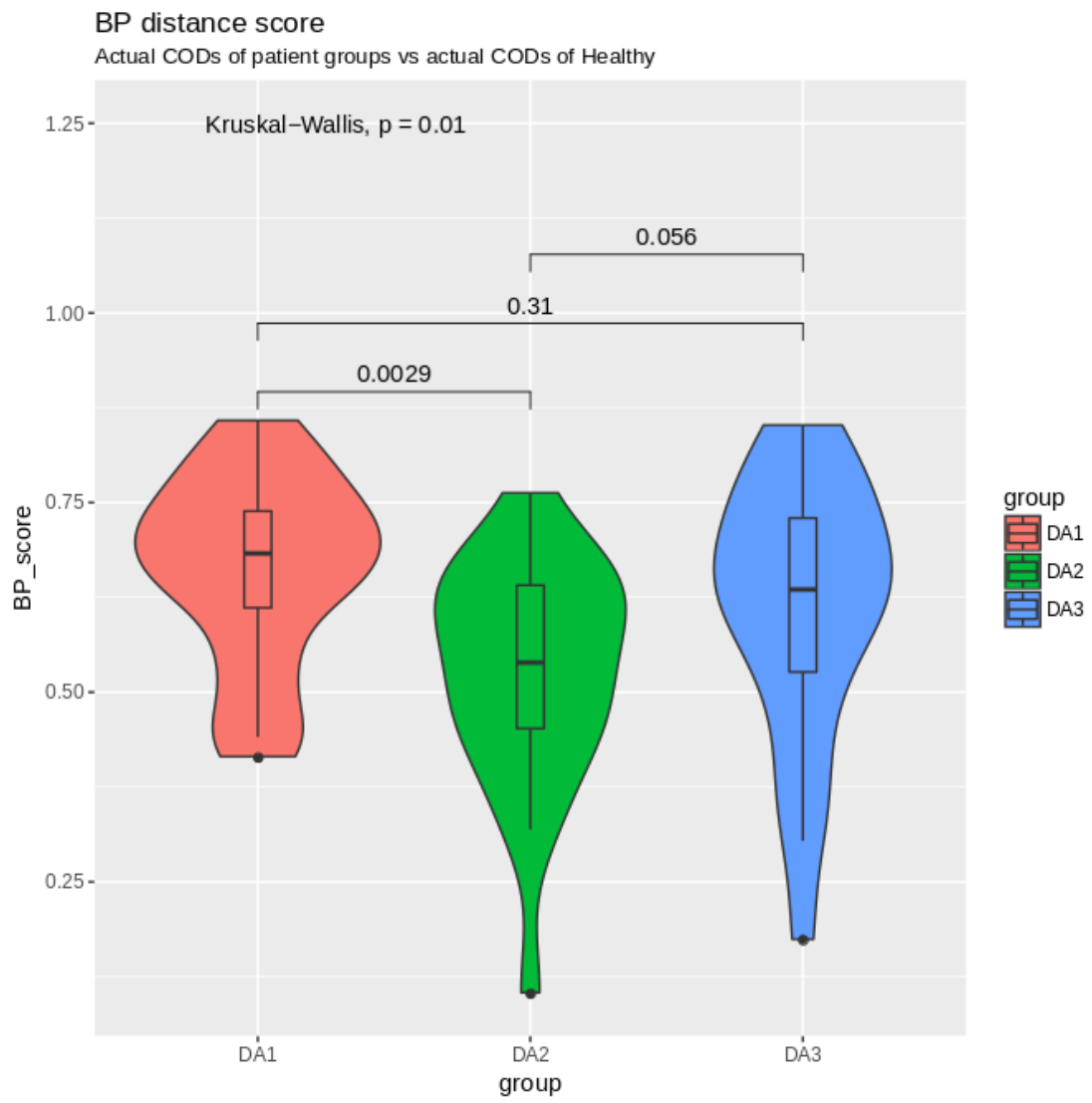
disrupted in a patient group and DEGs are associated with the region of the disruption (Figure 24).

4.4.3 *Interesting cases*

A number of very interesting cases of COD alteration were identified. These include reorganizations that progress in an almost gradient-like manner across DA groups. For instance, in chromosome 14 a border emerges, which separates the IgH locus and the upstream neighbourhood (Figure 25A), in chromosome 2 a border extension encloses the IgK locus (Figure 25B) and in chromosome 22 CODs are merged and expanded encompassing that way more IgL genes (Figure 25C). In contrast to previous observations, in those examples co-expression is enhanced in patient groups and remarkably is getting more robust as DA increases. Immunoglobulin expression is substantial for SLE progression and hence the latter findings support the importance of topological organization of gene expression for the development of this complex disease. Noteworthy, the aforementioned genes correspond to the underexpressed B cell signature identified during modular analysis, which ‘fades away’ with increase in DA and though gene expression levels become less differential, topological architecture of gene-expression does not develop to ‘healthy-like’.

Another interesting example of COD disruption involves a SNP. Variant rs1734787 in chromosome X has been implicated in SLE development and is co-localized with a COD split region in DA3 (Figure 26). Accordingly, genes reported to be affected by that variant, IRAK1 and MECP2, are not included in the post-split CODs.

A



B

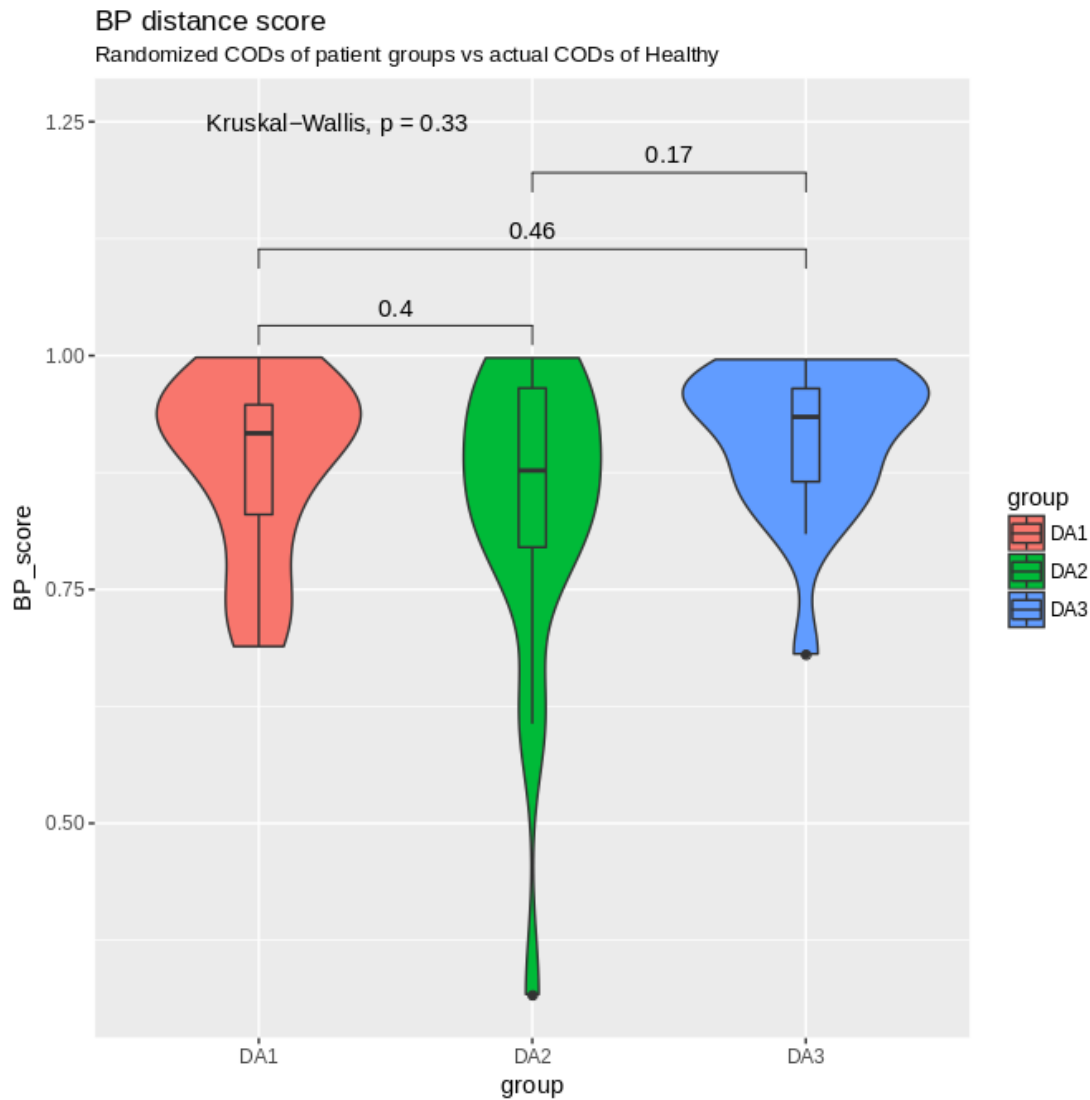


Figure 21: A. Violin plots demonstrating the estimated distribution of a per chromosome distance score calculated for CODs in each patient group compared to CODs in healthy group. B. Violin plots demonstrating the estimated distribution of a per chromosome distance score calculated for randomized CODs of each patient group compared to CODs in healthy group

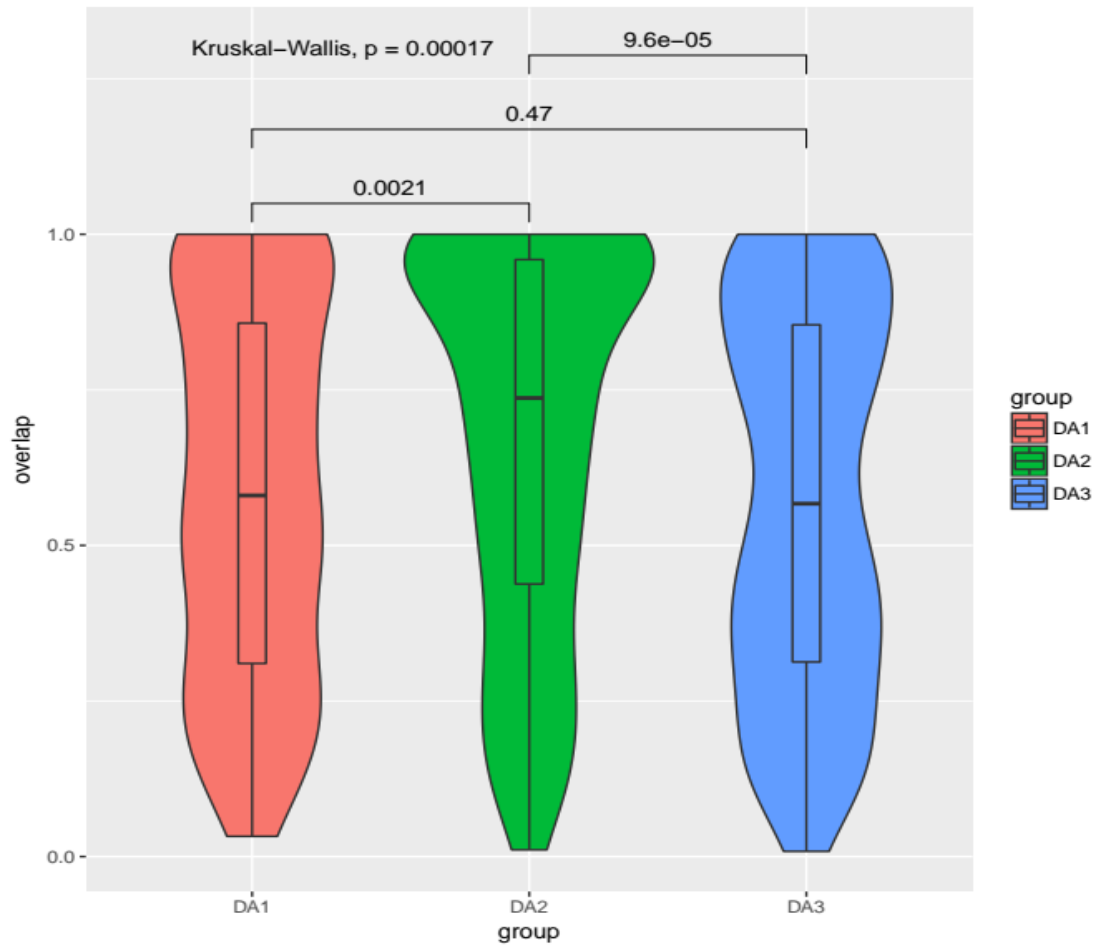


Figure 22: Violin plots demonstrating the estimated distribution of Jaccard similarity coefficients calculated for CODs in each patient group compared to CODs in healthy group.

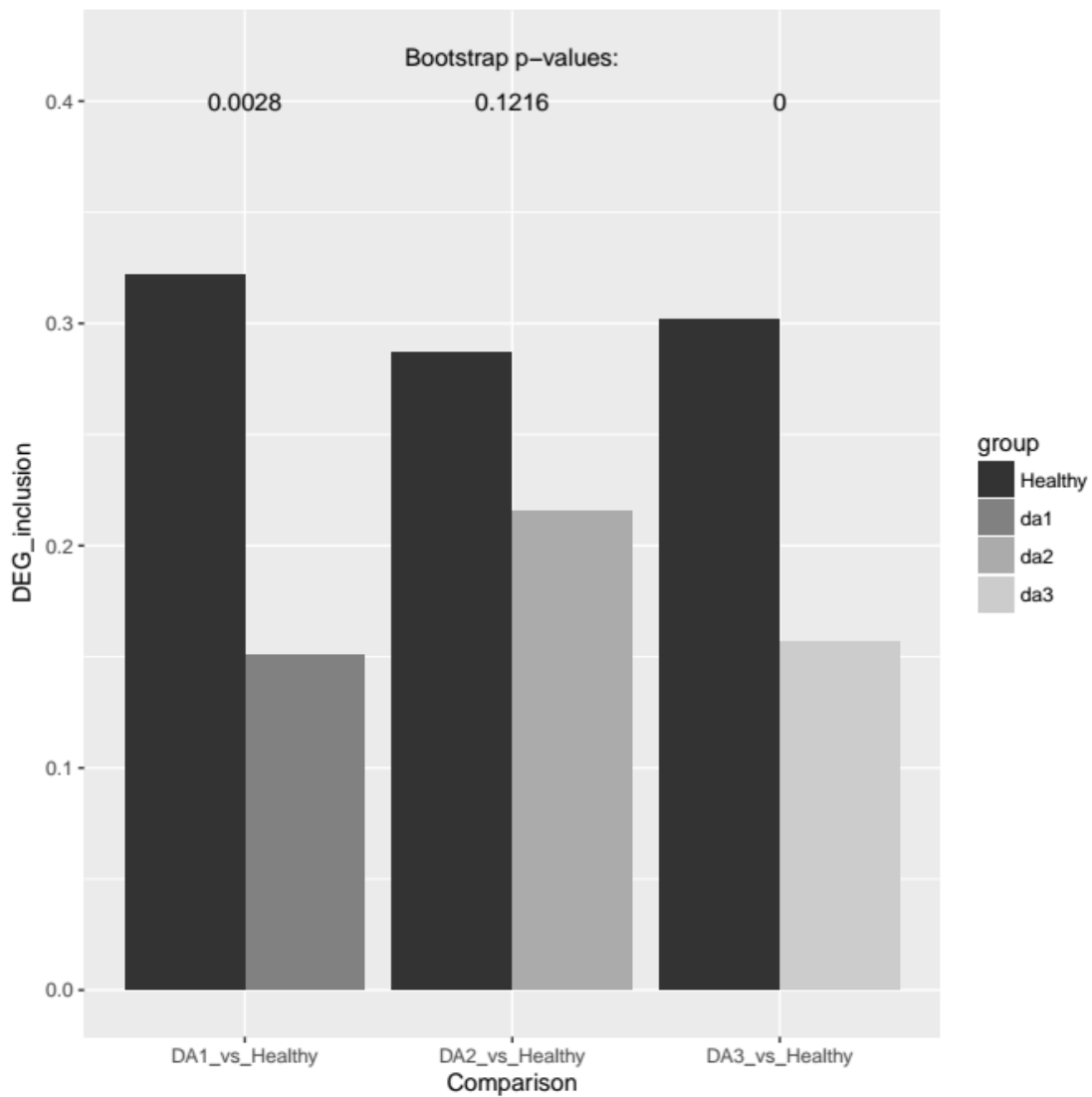


Figure 23: Barplots exhibiting the proportion of DEGs, derived from the comparison indicated at the horizontal axis, that reside inside CODs of the group indicated by colour code. Bootstrap p-values demonstrate statistical significance of the difference in DEG inclusion of each patient group and healthy control group.

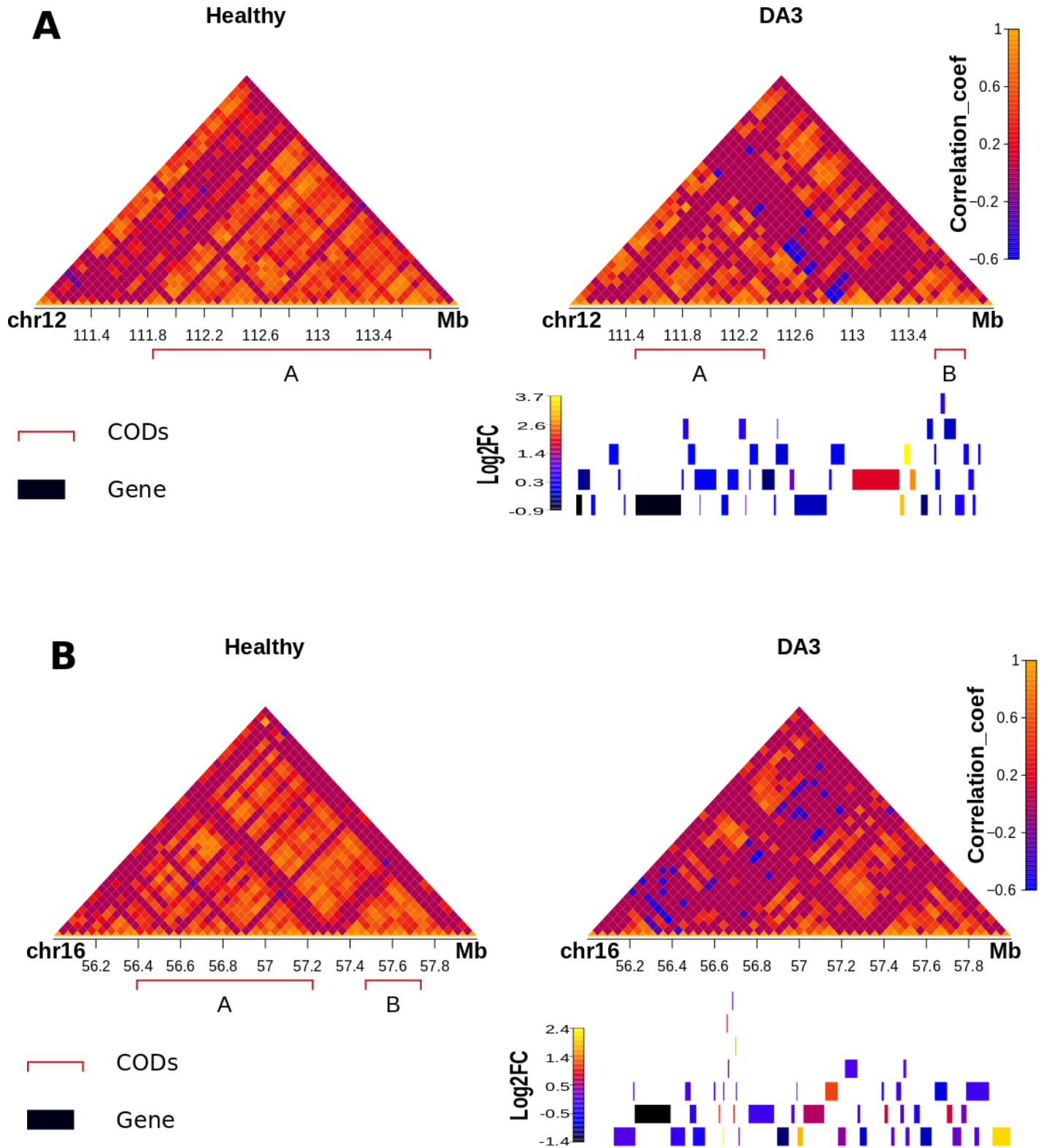
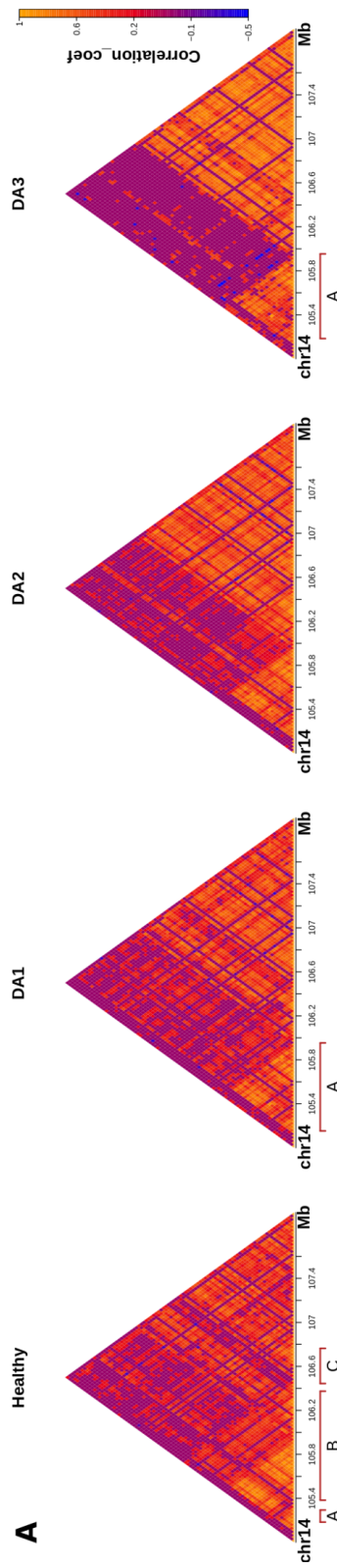
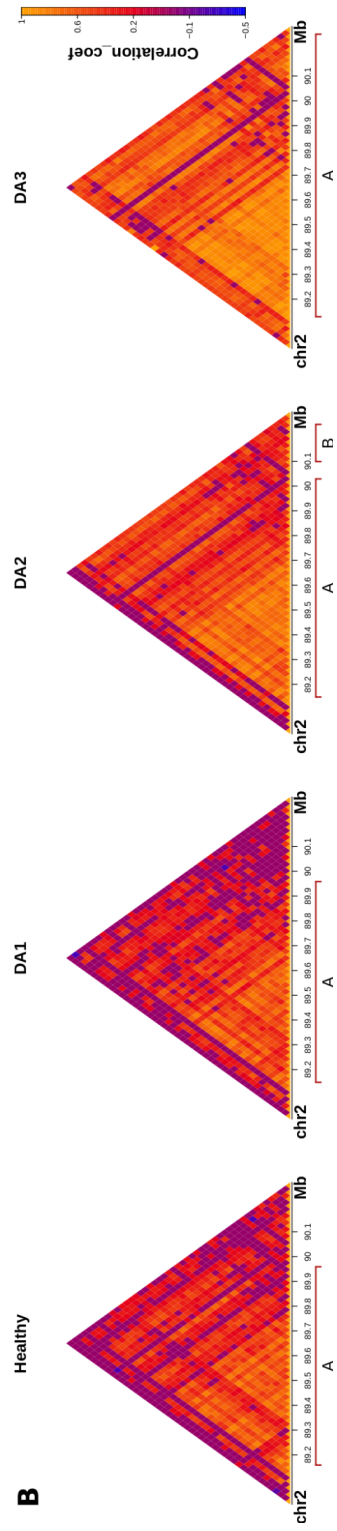


Figure 24: Cases where COD alterations, like a split (A) or a depletion (B), are associated with genes that are differentially expressed, when patient and healthy groups are compared.





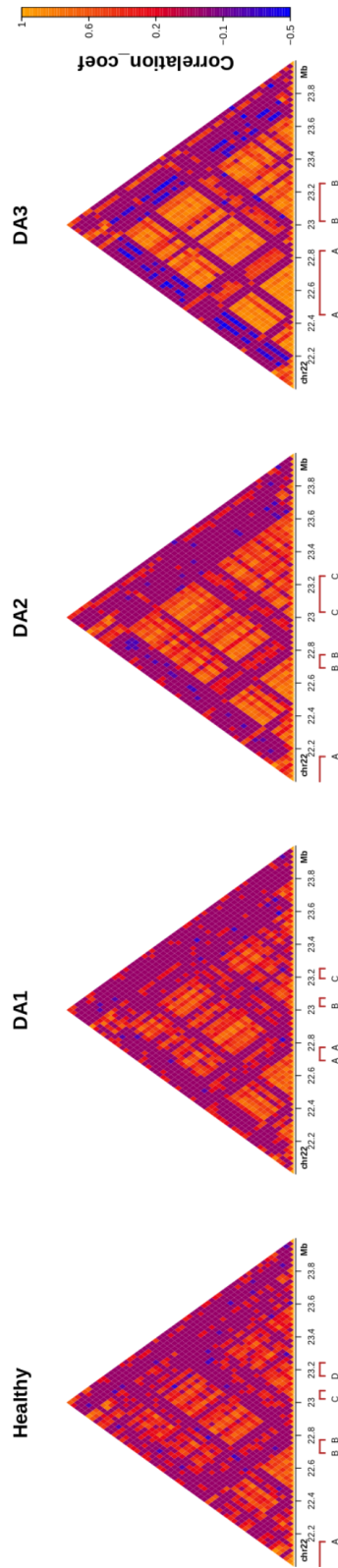


Figure 25: Cases where organization of gene expression alters in an almost gradient-like manner. In chromosome 14 a border emerges, which separates the *IgH* locus and the upstream neighbourhood (A), in chromosome 2 a border extension encloses the *IgK* locus (B), and in chromosome 22 CODs are merged and expanded encompassing that way more *IgL* genes.

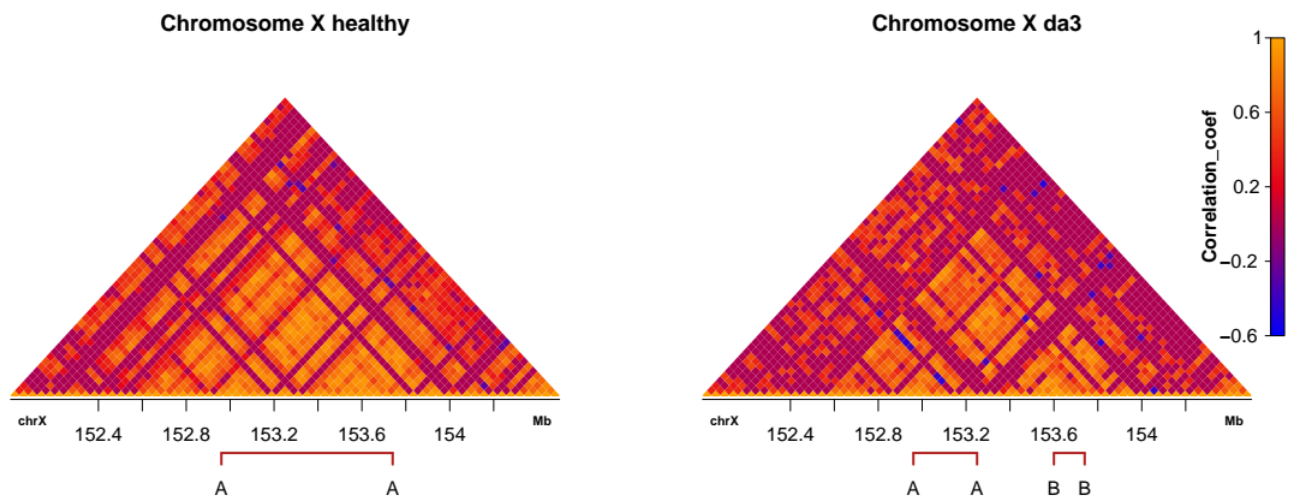


Figure 26: Variant rs1734787 in chromosome X is co-localized with a COD split region

5 Discussion

SLE is an incurable systemic autoimmune disease. Better comprehension of the molecular basis of the disorder would lead to improved diagnosis, prognosis and unveil new therapeutic targets. Here, gene expression profiles from whole blood samples of patients with diverse disease activity and manifestations have been studied. That was accomplished using various methodologies, in an effort to extract as much information as possible from the data. Through differential expression analysis coupled with modular analysis, previously reported and novel signatures were identified. Intriguingly, some signatures were common among the different patient cohorts, while others could be attributed to a specific patient subgroup or differentially modified in patient subgroups. Comparable signatures were identified *de novo* from the data using WGCNA, which were also correlated with varying clinical traits of patients. In addition, coordination of expression was investigated. Disease activity is associated with greater fragmentation of topologically defined co-expressed gene domains, suggesting possible implications of the genome architecture in SLE development.

It should be noted that the current work has limitations. Though the cellular functional units are predominantly proteins, expression was studied through monitoring mRNA levels and hence a significant part of transcriptional regulation is inaccessible. Moreover, whole blood samples were utilized and as a result the RNA abundance observed is an average of all the different cell types included. Although this could be useful for biomarker test development, it renders the interpretation of the results rather complex when specific mechanisms are studied. Finally, it should be mentioned that the profile studied is a ‘snapshot’ in time. Longitudinal data could be proved beneficial for disease development and even personalized studies.

However, this study has a lot of perspectives. Due to the abundance of information contained in the results, a thorough inspection of them combined with new experiments could aid the formation of new hypothesis concerning specific aspects of the disease. Incorporation of the different signatures currently identified into machine learning based approaches could lead to novel biomarker discovery. Lastly, concerning the topological analysis, further incorporation of genetic association data is likely to uncover mechanistic links for the observed gene expression aberrations associated with the progression of this complex disease. At the same time, integration of other sources of information linked to genome structure (epigenetically marked, chromatin domains, enhancer-promoter pairing etc) could lead to the development of additional working hypothesis towards an enhanced understanding of SLE onset and progression mechanisms.

6 Bibliography

1. Bertsias, G. K., Salmon, J. E. & Boumpas, D. T. Therapeutic opportunities in systemic lupus erythematosus: State of the art and prospects for the new decade. *Ann. Rheum. Dis.* **69**, 1603–1611 (2010).
2. Tsokos, G. C. Systemic Lupus Erythematosus. *N. Engl. J. Med.* **365**, 2110–2121 (2011).
3. Barturen, G. & Alarcón-Riquelme, M. E. SLE redefined on the basis of molecular pathways. *Best Pract. Res. Clin. Rheumatol.* **31**, 291–305 (2017).
4. Tsokos, G. C., Lo, M. S., Reis, P. C. & Sullivan, K. E. New insights into the immunopathogenesis of systemic lupus erythematosus. *Nat. Rev. Rheumatol.* **12**, 716–730 (2016).
5. Baechler, E. C. *et al.* Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. *Proc. Natl. Acad. Sci.* **100**, 2610 LP-2615 (2003).
6. Bennett, L. *et al.* Interferon and Granulopoiesis Signatures in Systemic Lupus Erythematosus Blood. *J. Exp. Med.* **197**, 711 LP-723 (2003).
7. Banchereau, R. *et al.* Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients. *Cell* **165**, 551–565 (2016).
8. Ding, Y. *et al.* Identification of a gene-expression predictor for diagnosis and personalized stratification of lupus patients. *PLoS One* **13**, e0198325 (2018).
9. Lightfoot, Y. L., Blanco, L. P. & Kaplan, M. J. Metabolic abnormalities and oxidative stress in lupus. *Curr. Opin. Rheumatol.* **29**, 442–449 (2017).
10. Oaks, Z. & Perl, A. Metabolic control of the epigenome in systemic Lupus erythematosus. *Autoimmunity* **47**, 256–264 (2014).
11. Li, L.-J., Fan, Y.-G., Leng, R.-X., Pan, H.-F. & Ye, D.-Q. Potential link between m(6)A modification and systemic lupus erythematosus. *Mol. Immunol.* **93**, 55–63 (2018).
12. Hedrich, C. M. Epigenetics in SLE. *Curr. Rheumatol. Rep.* **19**, (2017).
13. Sawalha, A. H. *et al.* Defective T-cell ERK signaling induces interferon-regulated gene expression and overexpression of methylation-sensitive genes similar to lupus patients. *Genes Immun.* **9**, 368 (2008).
14. Richardson, B. *et al.* Evidence for impaired T cell DNA methylation in systemic lupus erythematosus and rheumatoid arthritis. *Arthritis Rheum.* **33**, 1665–1673 (1990).
15. Zhao, M. *et al.* Increased 5-hydroxymethylcytosine in CD4(+) T cells in systemic lupus erythematosus. *J. Autoimmun.* **69**, 64–73 (2016).
16. Zhang, Z., Song, L., Maurer, K., Petri, M. A. & Sullivan, K. E. Global H4 acetylation analysis by ChIP-chip in systemic lupus erythematosus monocytes. *Genes Immun.* **11**, 124–133 (2010).

17. Yan, S., Yim, L. Y., Lu, L., Lau, C. S. & Chan, V. S.-F. MicroRNA Regulation in Systemic Lupus Erythematosus Pathogenesis. *Immune Network* **14**, 138–148 (2014).
18. Costa-Reis, P. *et al.* The Role of MicroRNAs and Human Epidermal Growth Factor Receptor 2 in Proliferative Lupus Nephritis. *Arthritis Rheumatol. (Hoboken, N.J.)* **67**, 2415–2426 (2015).
19. Rada Iglesias, A., Grosveld, F. G. & Papantonis, A. Forces driving the three dimensional folding of eukaryotic genomes. *Molecular Systems Biology* **14**, (2018).
20. Scharer, C. D. *et al.* ATAC-seq on biobanked specimens defines a unique chromatin accessibility structure in naïve SLE B cells. *Sci. Rep.* **6**, 1–9 (2016).
21. Michalak, P. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics* **91**, 243–248 (2008).
22. Tsochatzidou, M., Malliarou, M., Papanikolaou, N., Roca, J. & Nikolaou, C. Genome urbanization: Clusters of topologically co-regulated genes delineate functional compartments in the genome of *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **45**, 5818–5828 (2017).
23. Soler-Oliva, M. E., Guerrero-Martínez, J. A., Bachetti, V. & Reyes, J. C. Analysis of the relationship between coexpression domains and chromatin 3D organization. *PLoS Comput. Biol.* **13**, 1–25 (2017).
24. Panousis, N. I. *et al.* Genomic dissection of Systemic Lupus Erythematosus: Distinct Susceptibility, Activity and Severity Signatures. *bioRxiv* 255109 (2018). doi:<https://doi.org/10.1101/255109>
25. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
26. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
27. Ran, D. & Daye, Z. J. Gene expression variability and the analysis of large-scale RNA-seq studies with the MDSeq. *Nucleic Acids Res.* **45**, (2017).
28. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
29. Bombardier, C., Gladman, D. D., Urowitz, M. B., Caron, D. & Chang, C. H. Derivation of the SLEDAI. A disease activity index for lupus patients. The Committee on Prognosis Studies in SLE. *Arthritis Rheum.* **35**, 630–640 (1992).
30. Reimand, J. *et al.* g:Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* **44**, W83–W89 (2016).
31. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45**, D331–D338 (2017).
32. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
33. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2018).

34. Weiner, J. & Domaszewska, T. tmod: an R package for general and multivariate enrichment analysis. *PeerJ Prepr.* **4**, e2420v1 (2016).
35. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545 LP-15550 (2005).
36. Chaussabel, D. *et al.* A Modular Analysis Framework for Blood Genomics Studies: Application to Systemic Lupus Erythematosus. *Immunity* **29**, 150–164 (2008).
37. Li, S. *et al.* Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nat. Immunol.* **15**, 195–204 (2014).
38. Langfelder, P. & Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, (2008).
39. Jaccard, P. *Distribution de la Flore Alpine dans le Bassin des Dranses et dans quelques régions voisines. Bulletin de la Societe Vaudoise des Sciences Naturelles* **37**, (1901).
40. Zaborowski, R. & Wilczynski, B. BPscore: an effective metric for meaningful comparisons of structural chromosome segmentations. *bioRxiv* (2018).
41. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453 (2015).
42. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
43. Mann, H. B. & Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Stat.* **18**, 50–60 (1947).
44. R Core Team. R: A Language and Environment for Statistical Computing. (2018).
45. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2009).
46. Warnes, G. R. *et al.* gplots: Various R Programming Tools for Plotting Data. (2016).
47. Phanstiel, D. H. Sushi: Tools for visualizing genomics data. (2015).
48. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS One* **6**, e21800 (2011).

Module-trait relationships

