

LF97 - Hugo Germain - Nuclear envelope proteins

Sylvie Bourassa

22 January 2020

Samples

- 4 groups of Nuclear envelope proteins (CTRL, IFN, NL and IFNNL (VI)) x 3 replicates = 12 samples

Comparisons

- 3 IFN vs 3 CTRL
- 3 NL vs 3 CTRL
- 3 IFNNL (VI) vs 3 CTRL
- 3 IFN vs 3 NL
- 3 IFN vs 3 IFNNL (VI)
- 3 NL vs 3 IFNNL (VI)

Sample preparation and analysis

- Proteins were resuspended in Ammonium bicarbonate (50mM)/Deoxycholate(1%)
- S-S bridges were reduced by DTT and alkylated with Iodoacetamide
- Proteins were digested with Trypsin (1:50) overnight at 37 degree celsius
- Resulting peptides were purified using StageTips C18 and Speed Vac dried
- Samples were resuspended in LC loading solvant and eq. 1µg was injected
- Samples were analyzed by LC-MSMS on an Orbitrap Fusion Tribrid system (Thermo)
- 120 min LC runs (90 min gradients) were used and the MS operated in DDA mode
- Identification and LFQ quantification was done using MaxQuant software v1.6.10.43
- Database : Ref homo sapiens (74485 entries)
- Modifications: Fixed : Carbamidomethyl (Cys), Variable : Oxidation(M), Acetyl (Prot N-term)

DATA TREATMENT (R software)

1 - Import & filtering data

- Input file : "proteinGroups.txt" from the MaxQuant software
- Decoy proteins are removed
- For data processing, only the LFQ intensity columns are considered

2 - Data normalization

- For each sample, a median intensity is calculated

- For each sample, an intensity normalization factor is calculated by dividing the median intensity of said sample by the median of the median intensity of all samples.
- Each protein intensity is normalized by dividing their intensity by the normalization factor

3 - Missing data imputation

- For each sample, a noise value corresponding to the 0.01 percentile of the intensities of said sample is calculated
- This noise value is imputed for each samples when the intensity value is missing

4 - Comparisons

- Only proteins which present 100% of intensity values (not noise imputed value) in at least one group are considered. These proteins correspond to *quantifiable proteins*.
- Only proteins identified with at least 2 peptides are considered. They correspond to *quantified proteins*.

5 - Statistics

For each protein in each comparison, the following values are calculated :

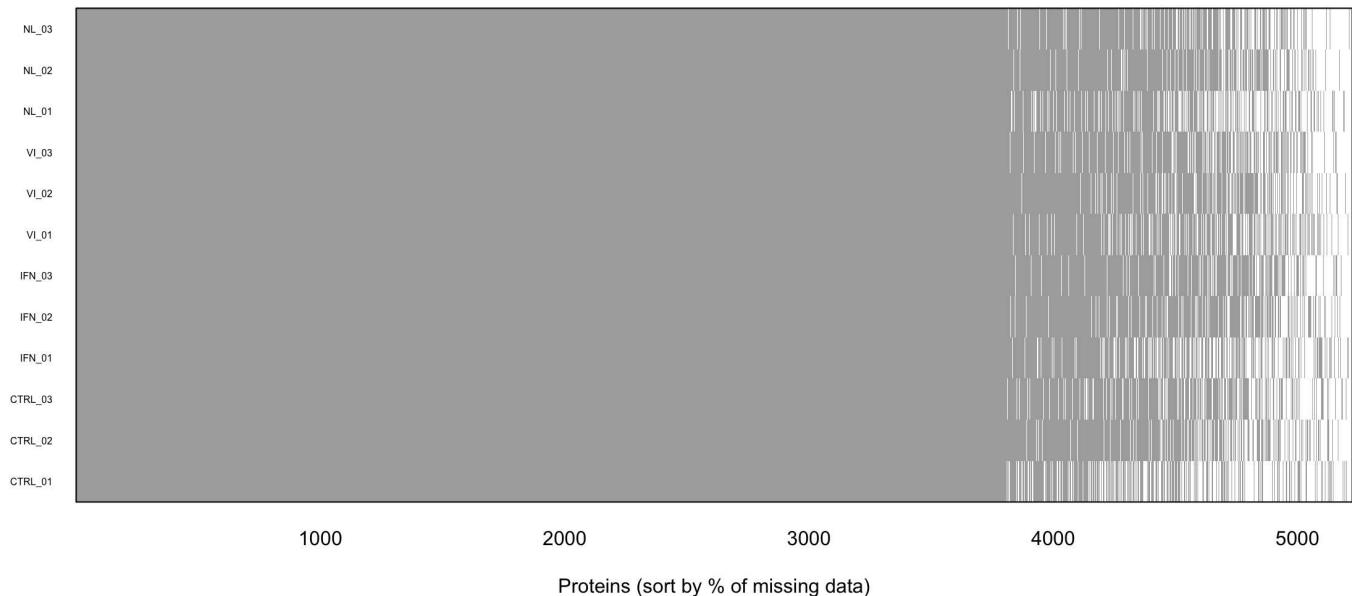
- the mean intensity in each group (“mean_g1” and “mean_g2”),
- the intensity ratio between each group (“Ratio g1/g2”),
- the log₂ of g1/g2 ratio (“log2Ratio””)
- the z-score $z = x - \text{mean} / \text{standard deviation}$ (“zscore””)
- Limma *p-value* and *q-value* (Benjamini Hochberg adjusted *p-value*) (“pval_Limma” and “qval_Limma”)
- The following filters were applied to define variant proteins : *q-value* < 0.05, $|z\text{-score}| > 1.96$ (“Sign_zscore_qval_Limma”)

DATA REPRESENTATION

1 - Check missing values

Low abundance proteins signal is usually not extracted in all samples. This graph shows the proportions of intensity values (grey) and missing values (white) for each sample of the analysis. We expect to observe low proportion of missing values (<15-20%) and a similar profile for all samples to compare.

Missingness pattern of proteins

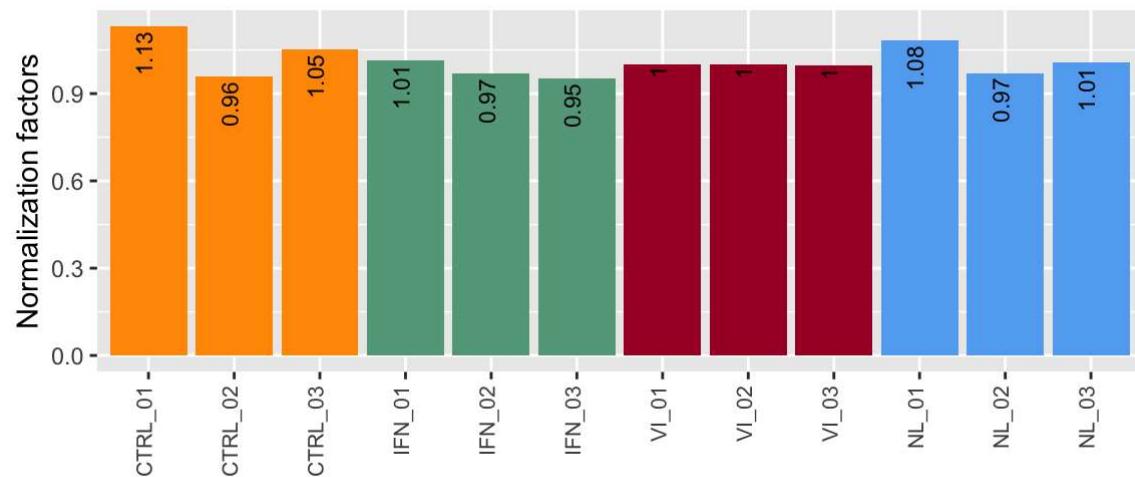


- Between 7.24% and 14.05% of missing values between the samples
- 9.29% of missing values for the whole data set

2 - Data normalization

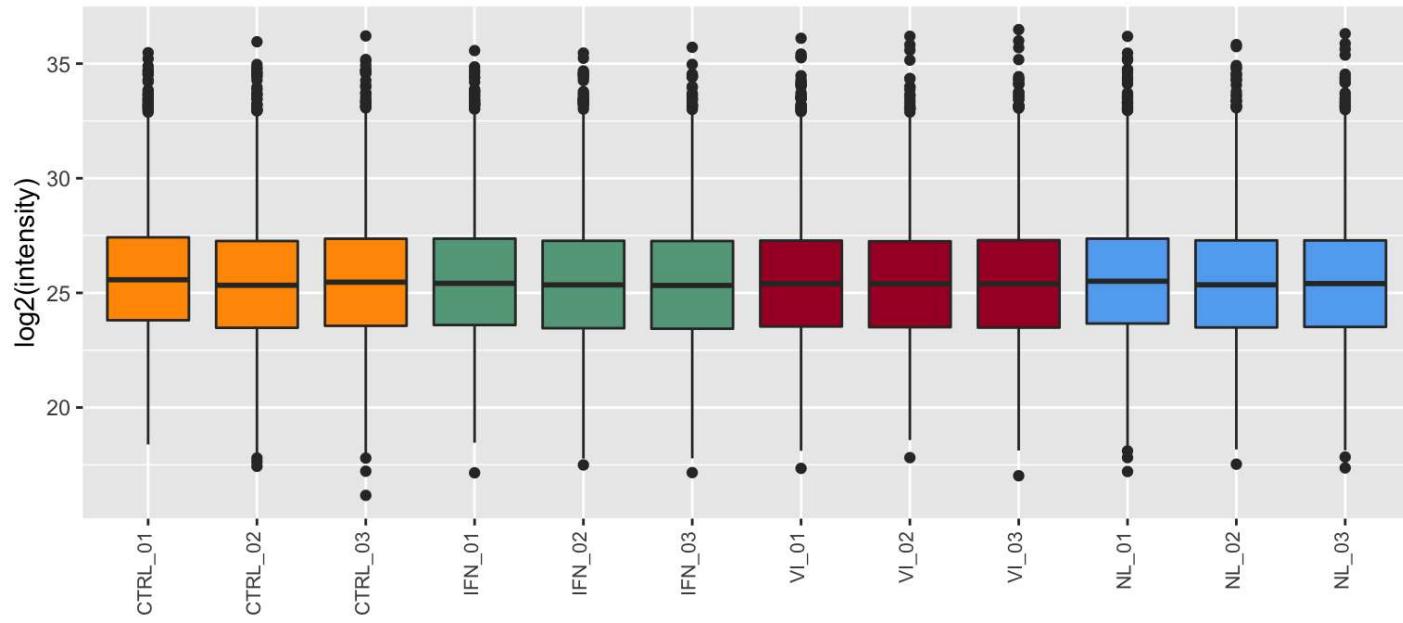
Normalization factors are expected to be close as possible to 1. After normalization, the median of intensities is identical for all samples and the box plots should display a similar repartition of the intensities across all samples.

Normalization factors

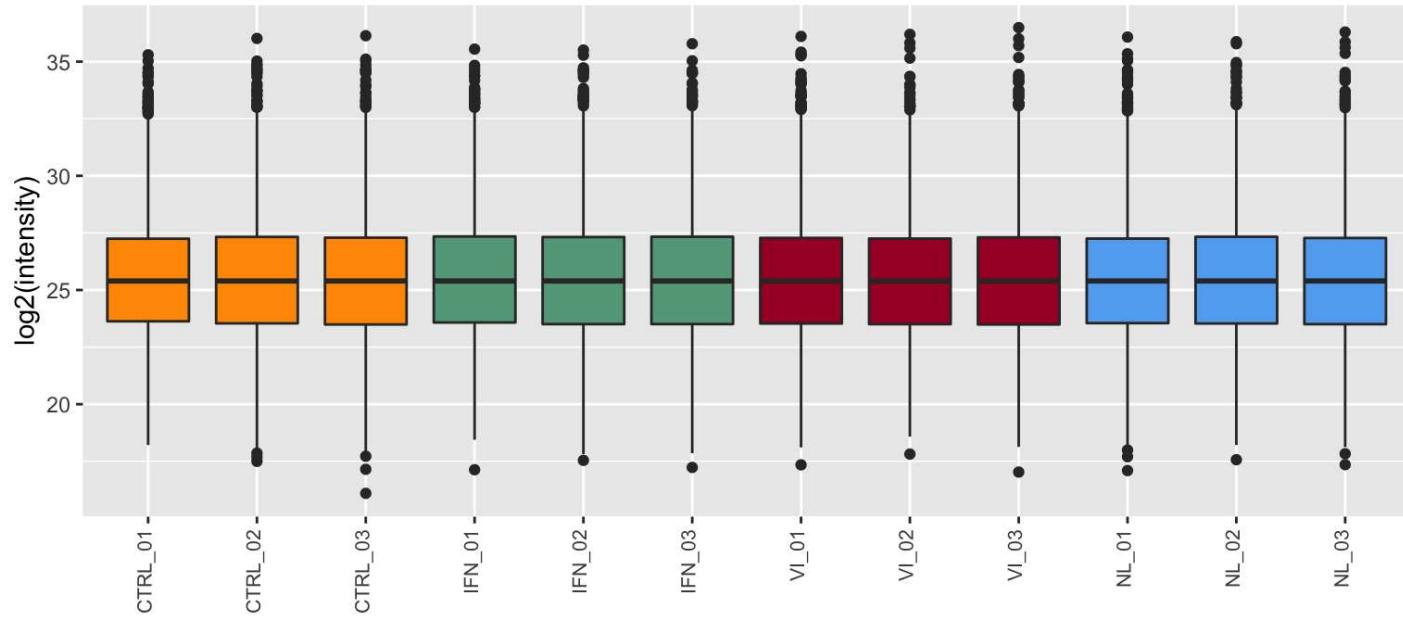


- Normalization factor are between a value of 0.95 and 1.13
- The mean normalization factor is 1.01

BoxPlot before normalization

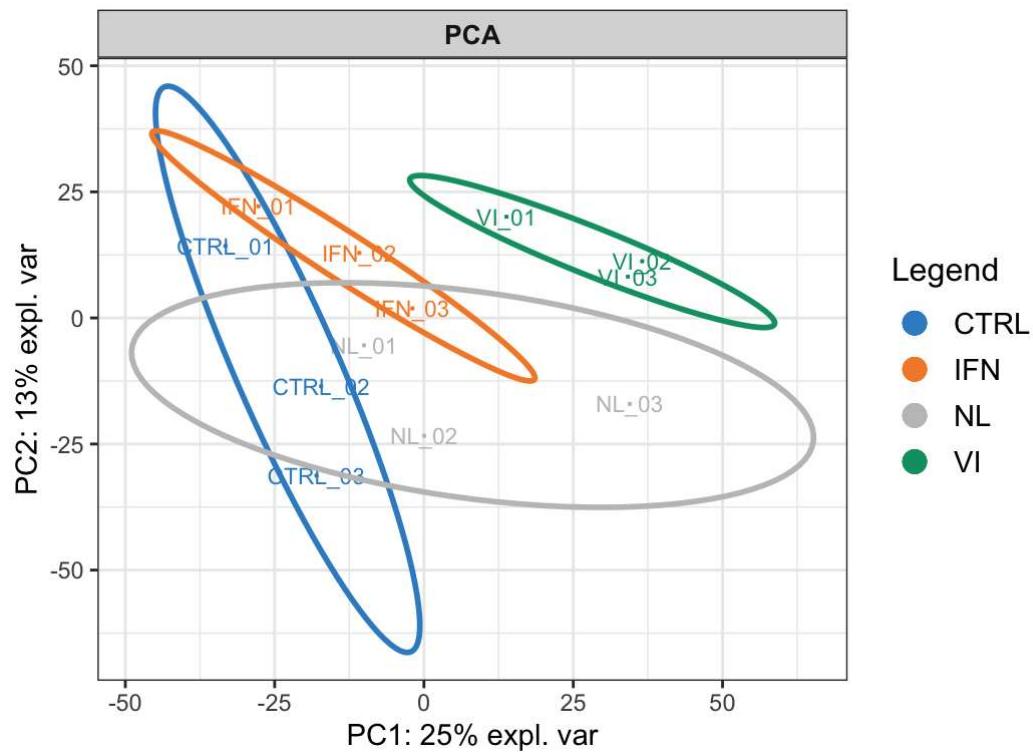


BoxPlot after normalization



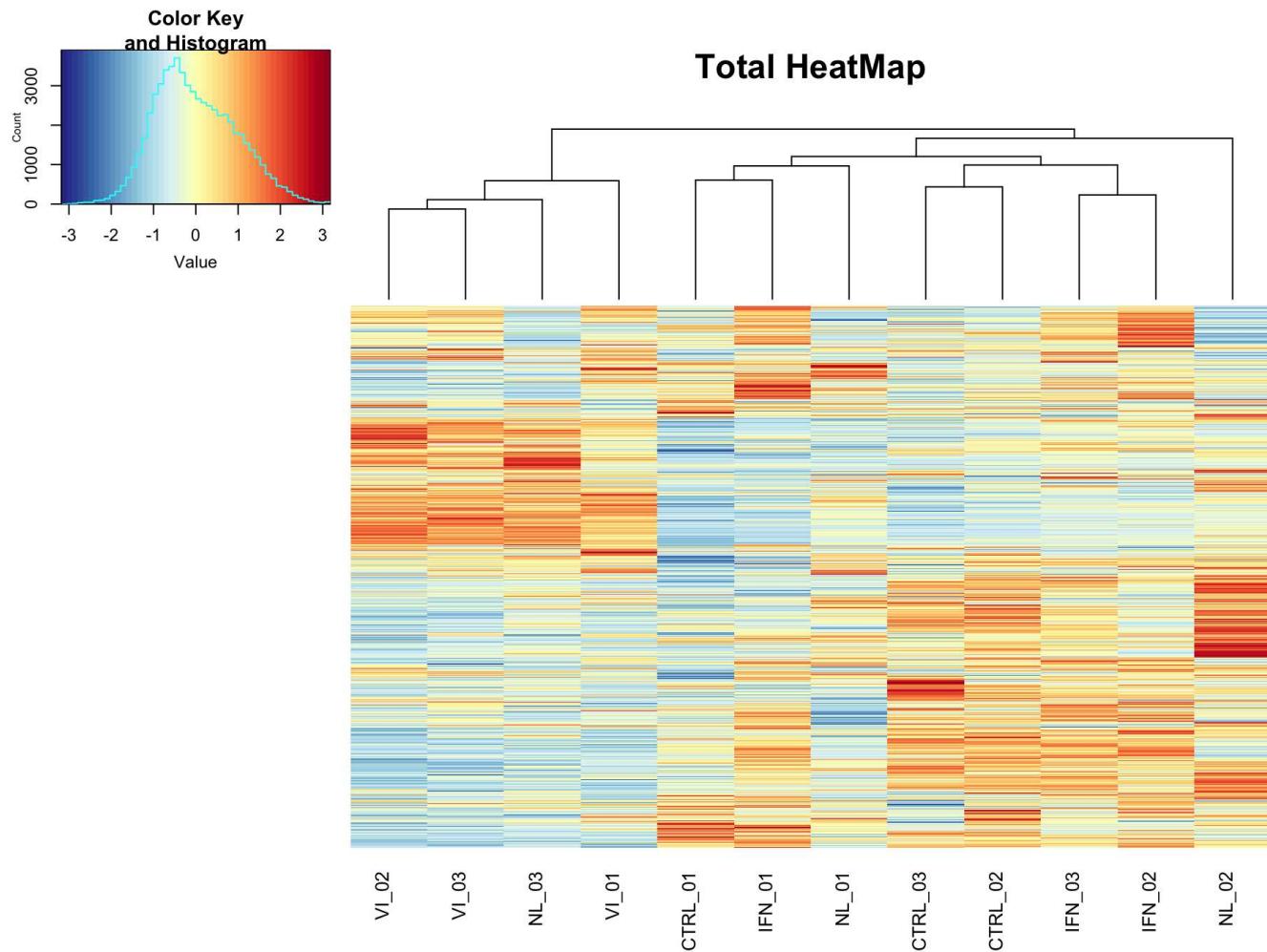
3 - PCA - Principal Component Analysis (proteins)

This representation uses an orthogonal transformation to convert all variables (here proteins) into new variables (principal component) explaining as much as possible the variability across the samples. The first component (PC1) always has the highest variance. Samples having similar proteomic profiles will appear close on the PCA plots. When the ellipses of two groups are not overlapping, we can consider that the sample groups can be distinguished based on their proteomic profiles.



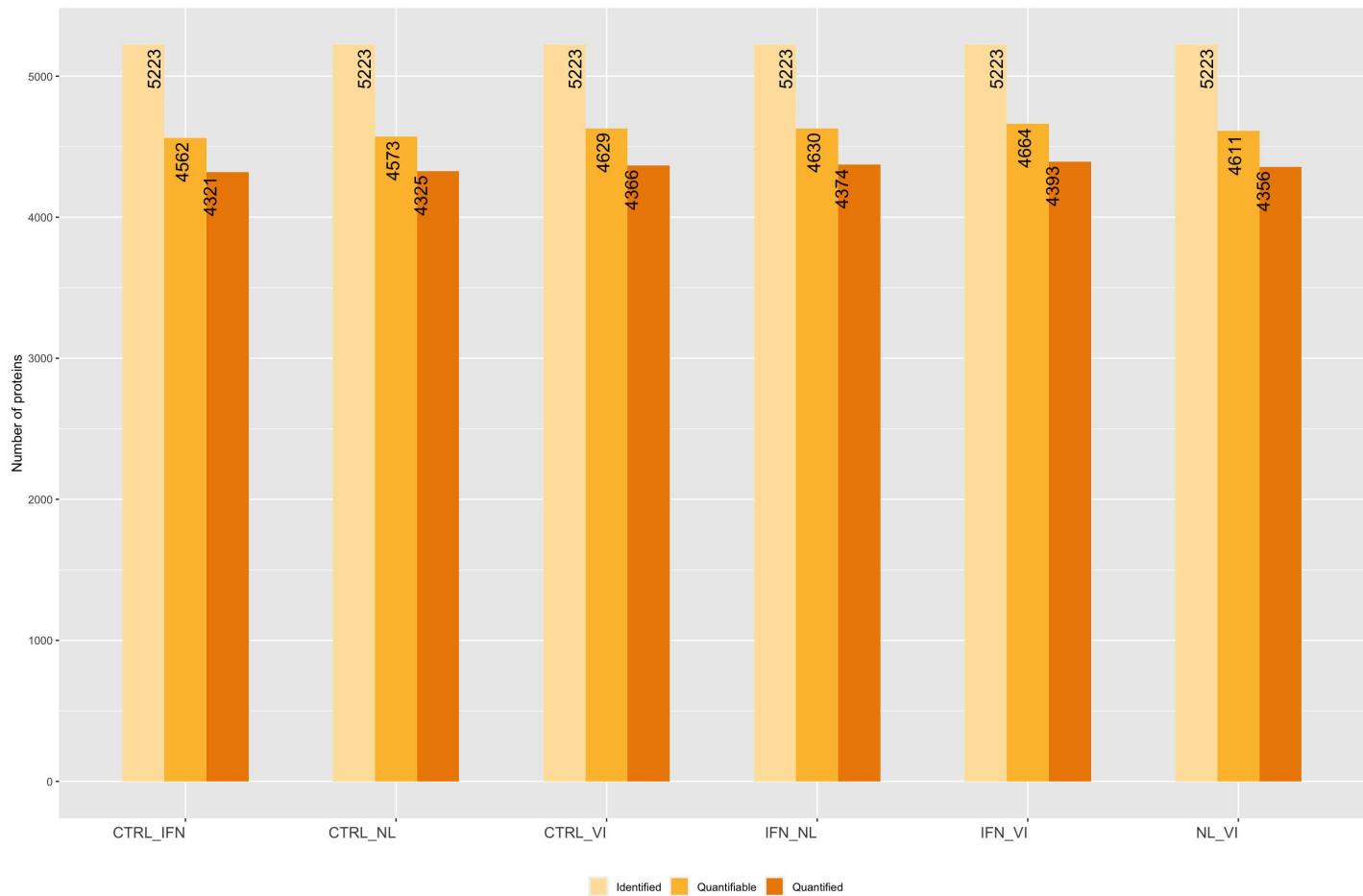
4 - Proteomic profiles heatmap

This heatmap represents the intensity values in all samples of all proteins of the analysis. A hierarchical clustering is applied both on rows and columns meaning that samples or proteins having similar profiles appear close to each other



4 - Number of quantifications for each comparison

Not all proteins of an analysis can be quantified, the signal of low abundance proteins might be difficult to extract resulting in missing values for certain sample. Proteins with a sufficient number of not missing values (100) are considered as quantifiable. Moreover, to add more confidence in the quantification value, we only retain proteins quantified with at least 2 peptides.



Identified proteins :

Total number of proteins identified in the analysis

5223 proteins identified in this analysis

Quantifiable proteins :

Proteins with at least 100% of observed intensities in each replicates in one of the two groups.

4612 quantifiable proteins on average for the whole analysis

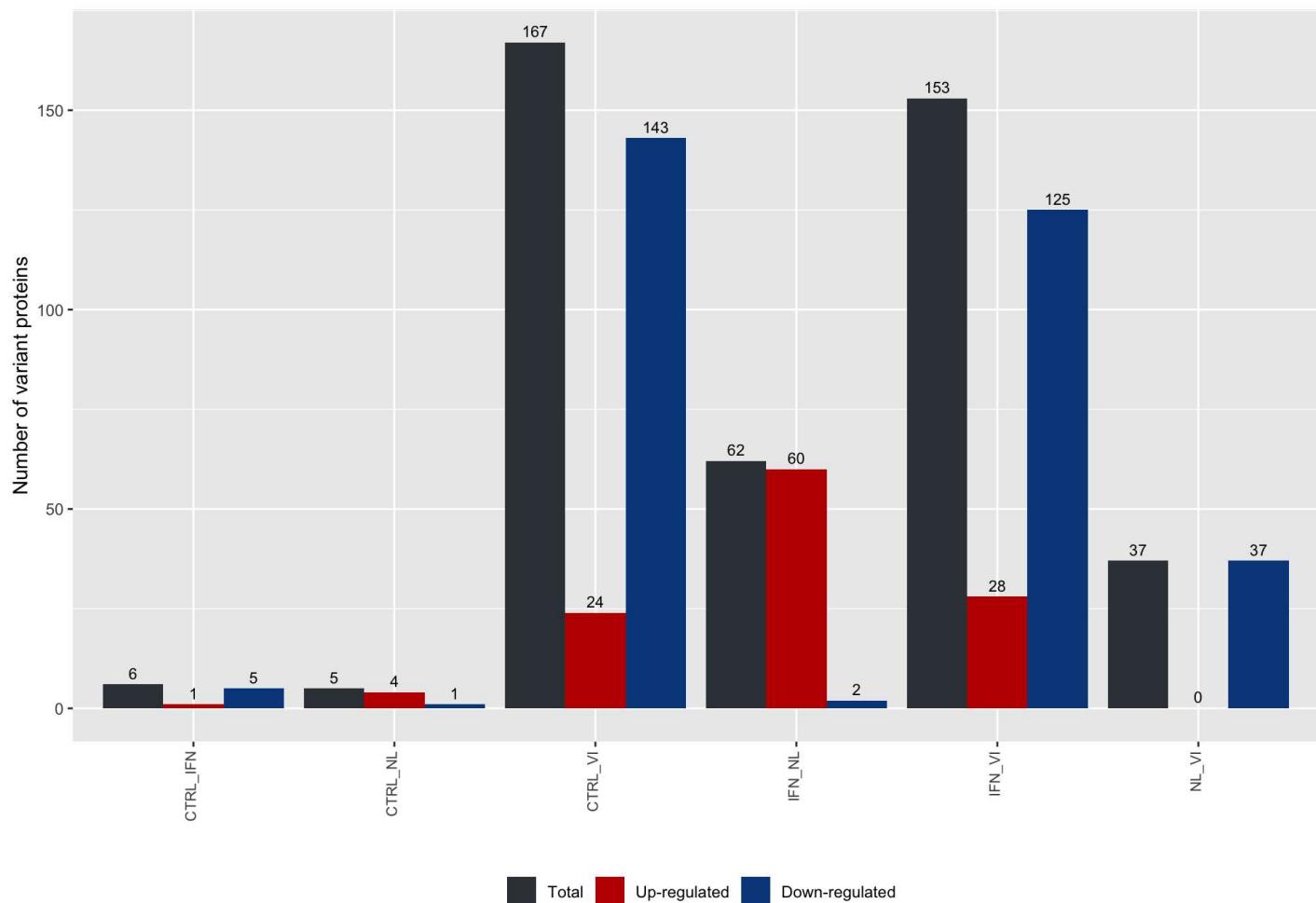
Quantified proteins :

Proteins identified with at least 2 peptides.

4356 quantified proteins on average for the whole analysis

5 - Number of variant proteins for each comparison

Proteins are considered as “variant” between two conditions/groups if they fulfill these criteria : q-value < 0.05 and $|z\text{-score}| > 1.96$. In a comparison annotated “group1_group2”: the proteins “up-regulated” (red) are more abundant in “group1” than in “group2”, the proteins “down-regulated” (blue) are more abundant in “group2” than in “group1”,

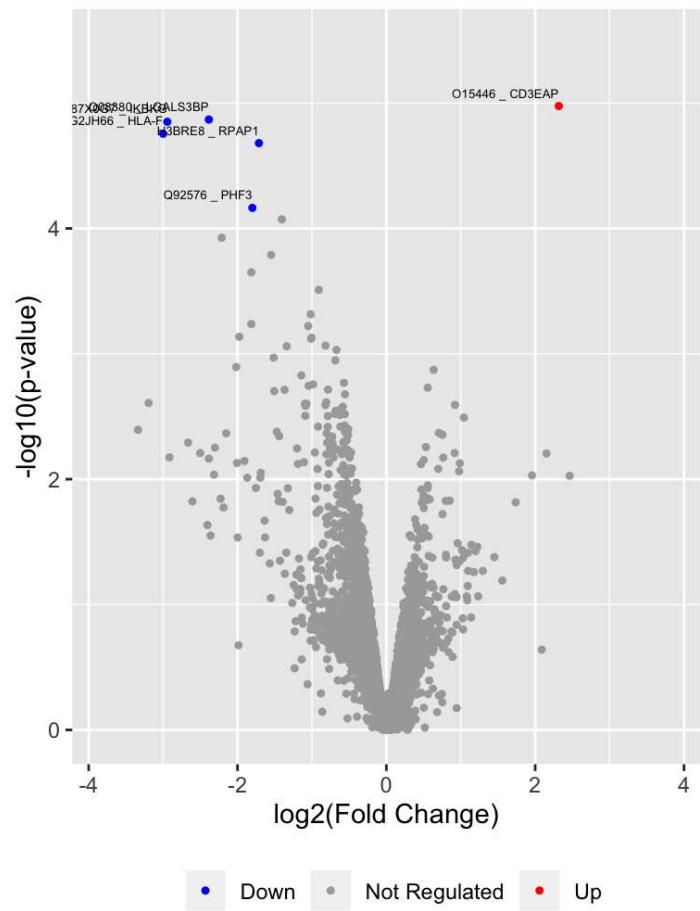
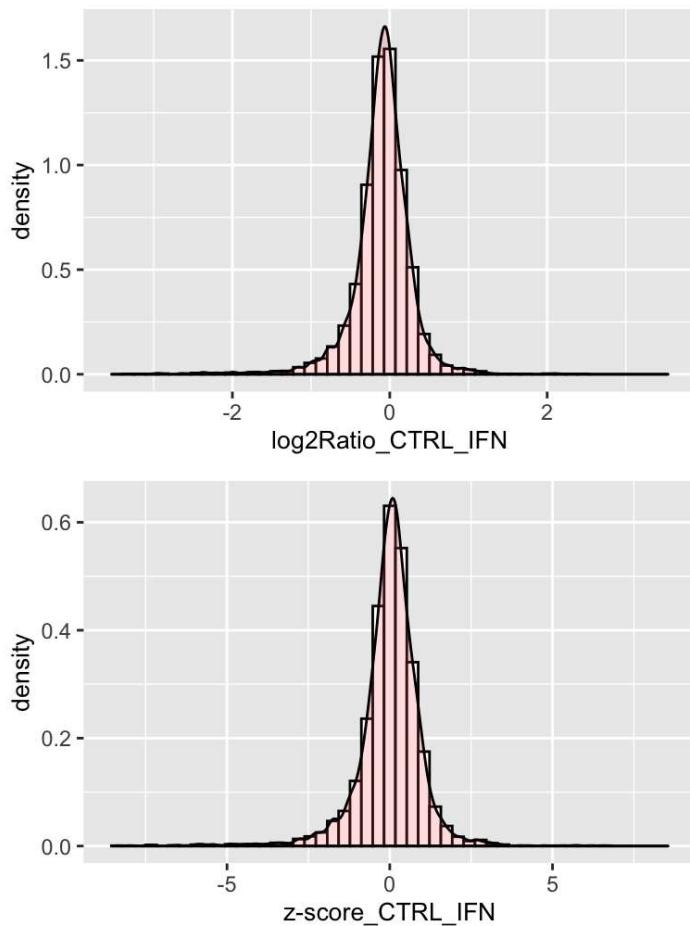


6 - Plots for each comparison

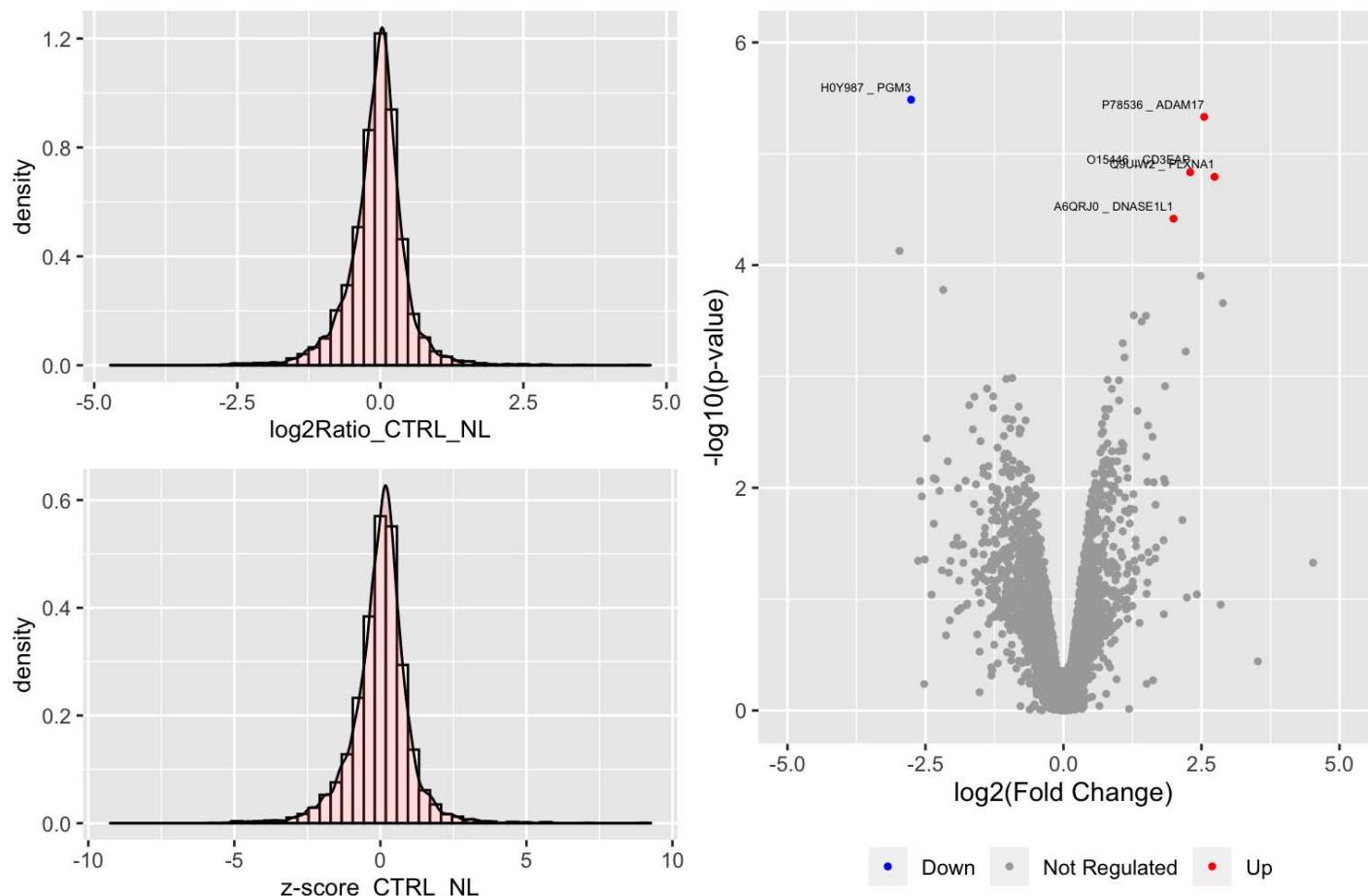
Repartition Histograms and Volcano plots

- The repartition histogram of all protein ratios ($\log_2(\text{ratio})$), expected to be centered as much as possible to 0 (ratio $g_1/g_2 = 1$)
 - The repartition histogram of z-scores (centering), must be centered to 0
 - The volcano plot, where x-axis represents the protein ratio ($\log_2(\text{ratio})$) and the y-axis the probability (- $\log_{10}(\text{p-value})$)
- The proteins are variant if they have a high ratio (in absolute value) and a good probability, i.e proteins in the upper right and upper left corners of the graph.

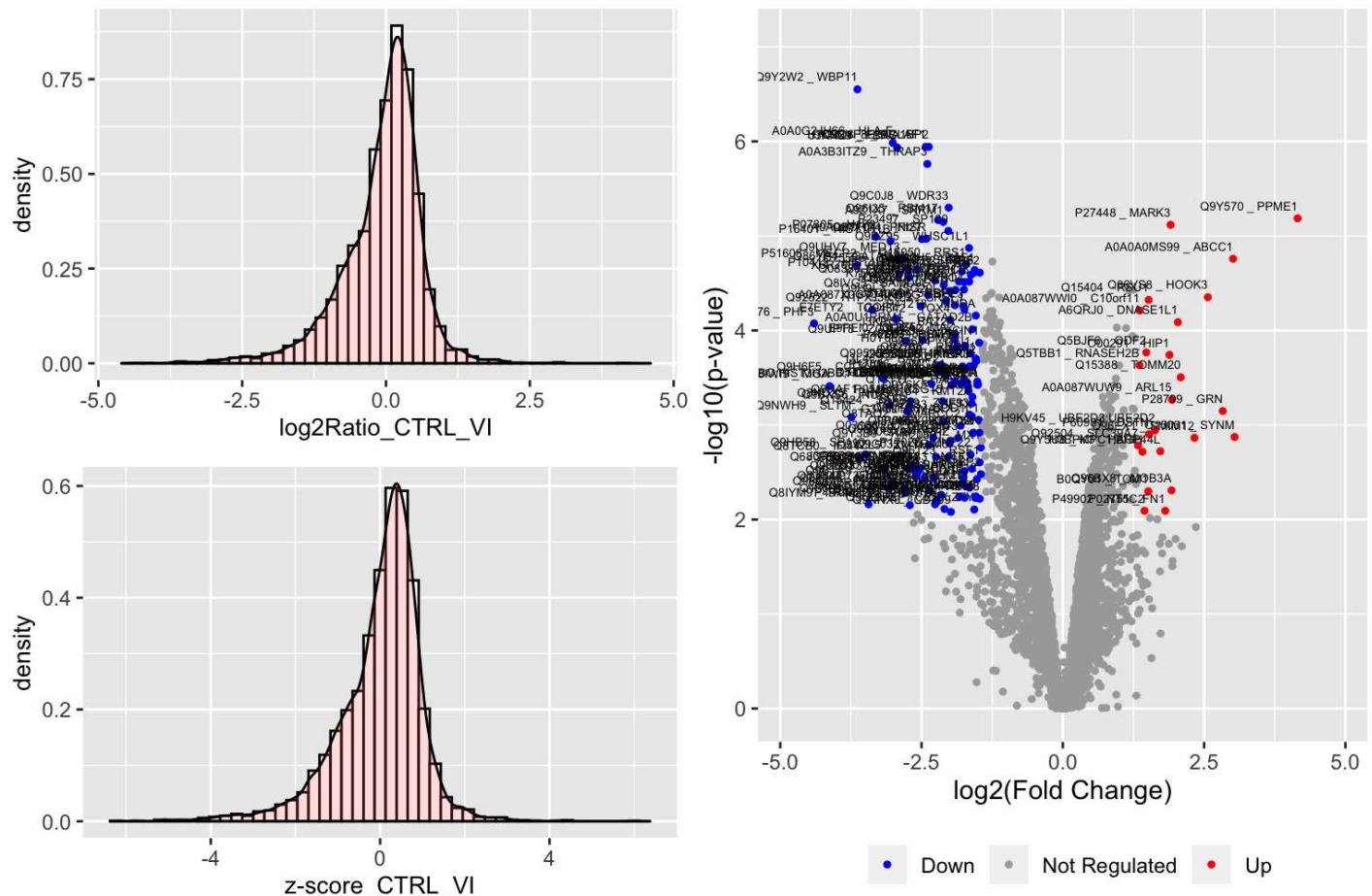
CTRL_IFN



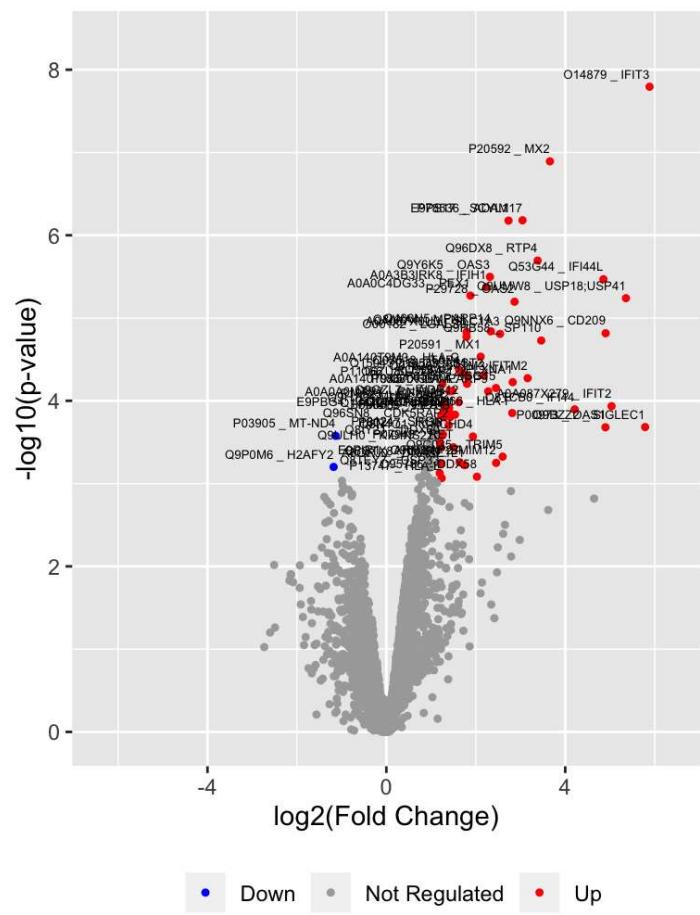
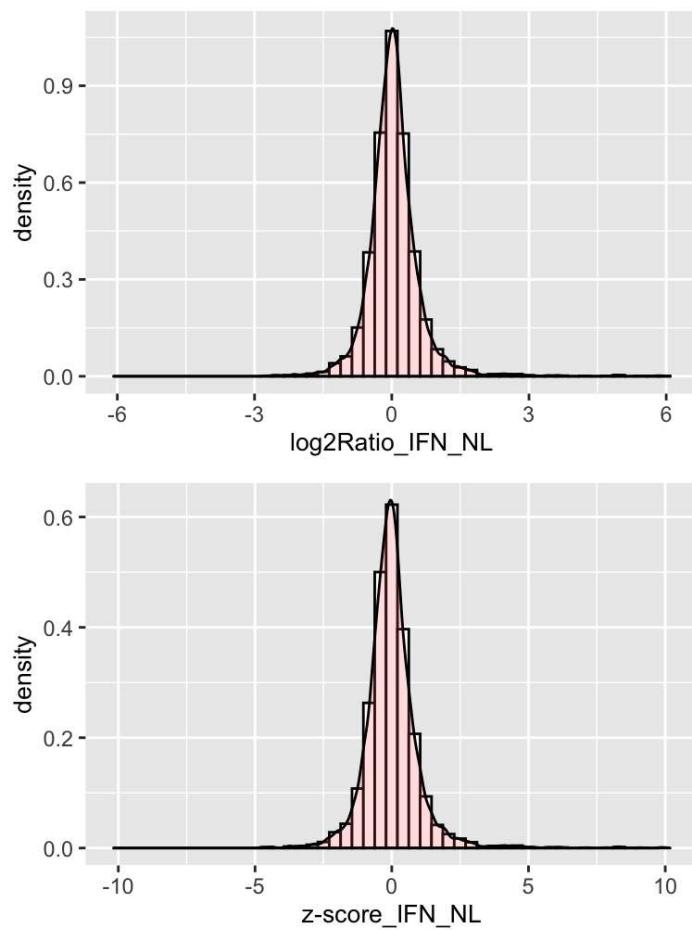
CTRL_NL



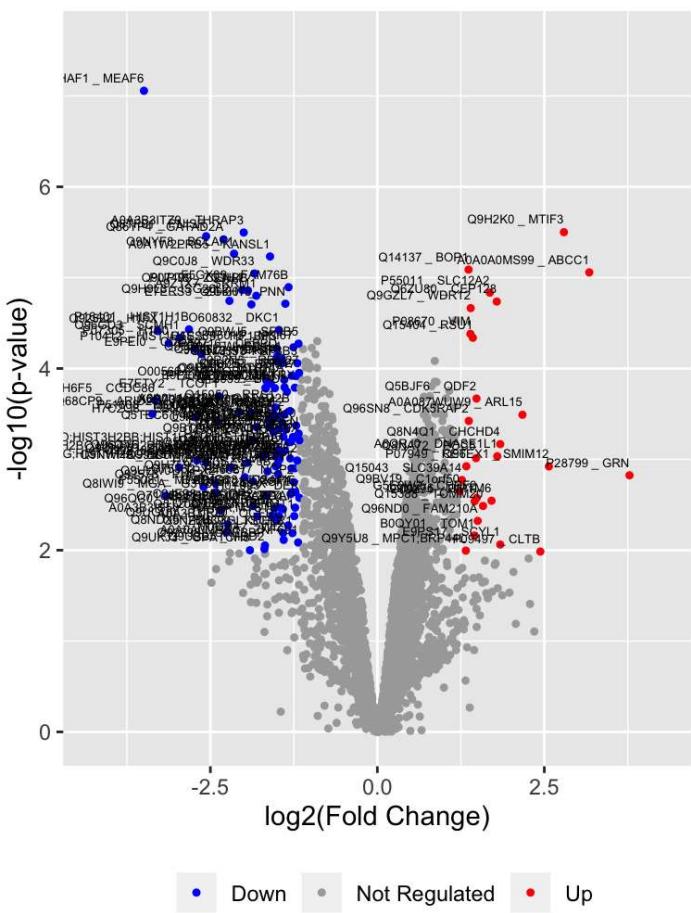
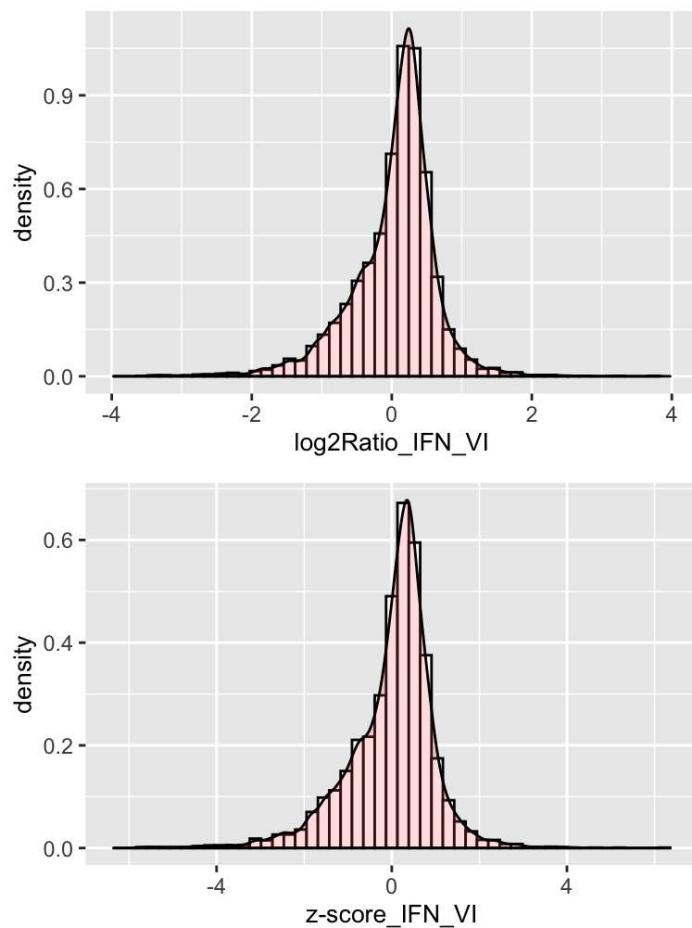
CTRL_VI



IFN_NL

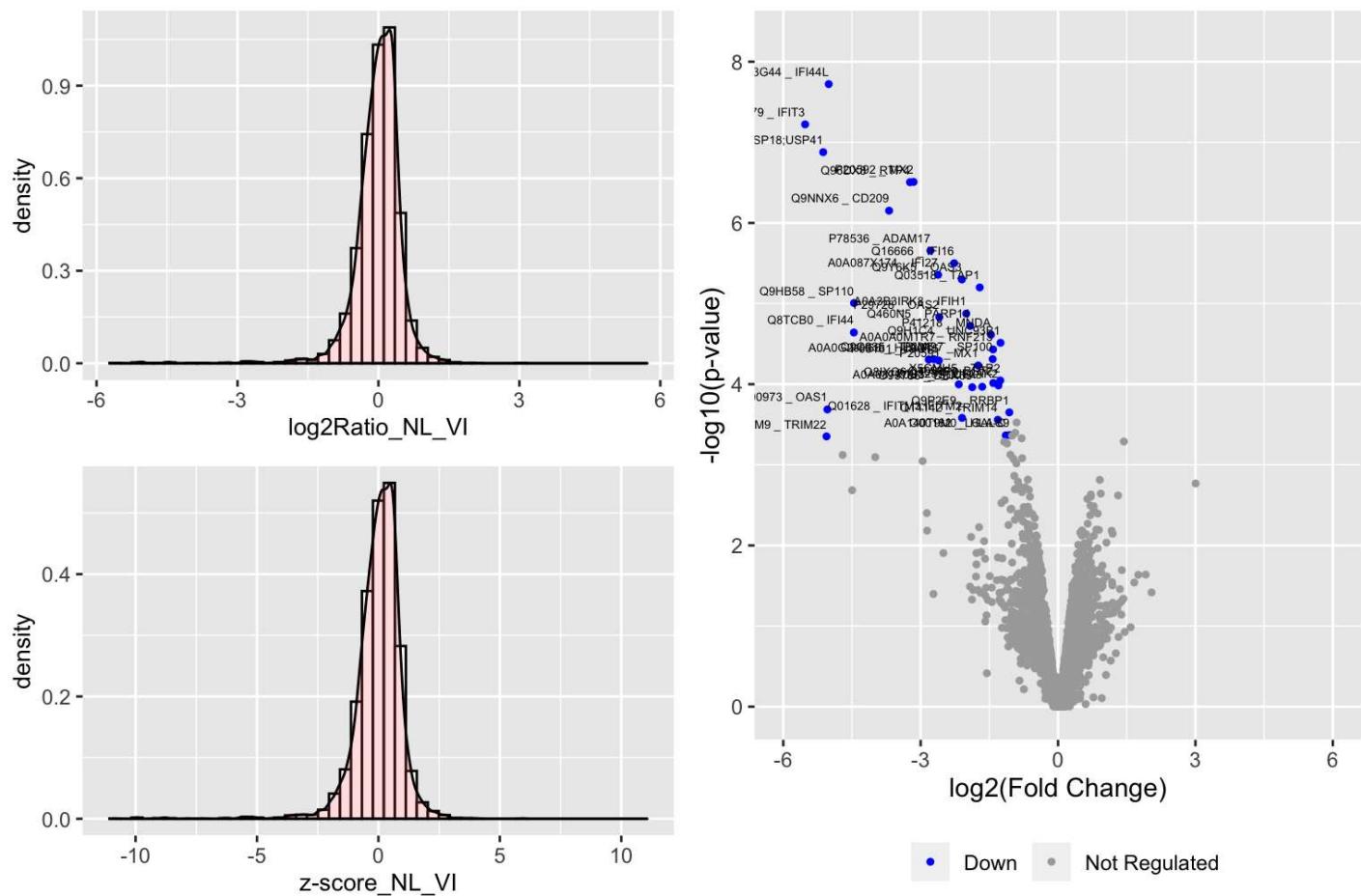


IFN_VI



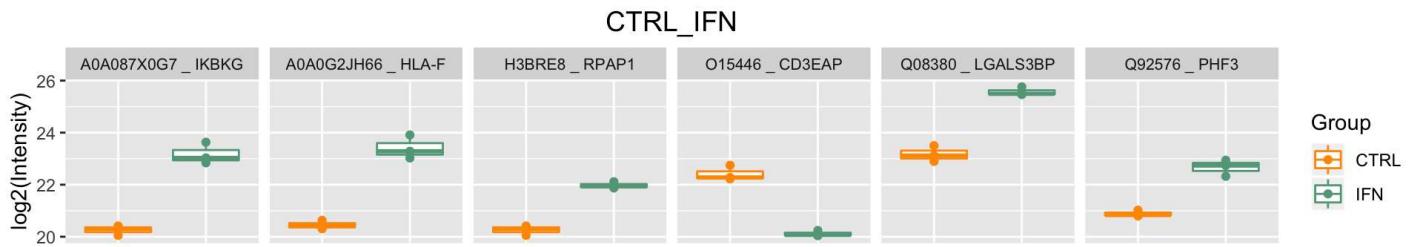
● Down ● Not Regulated ● Up

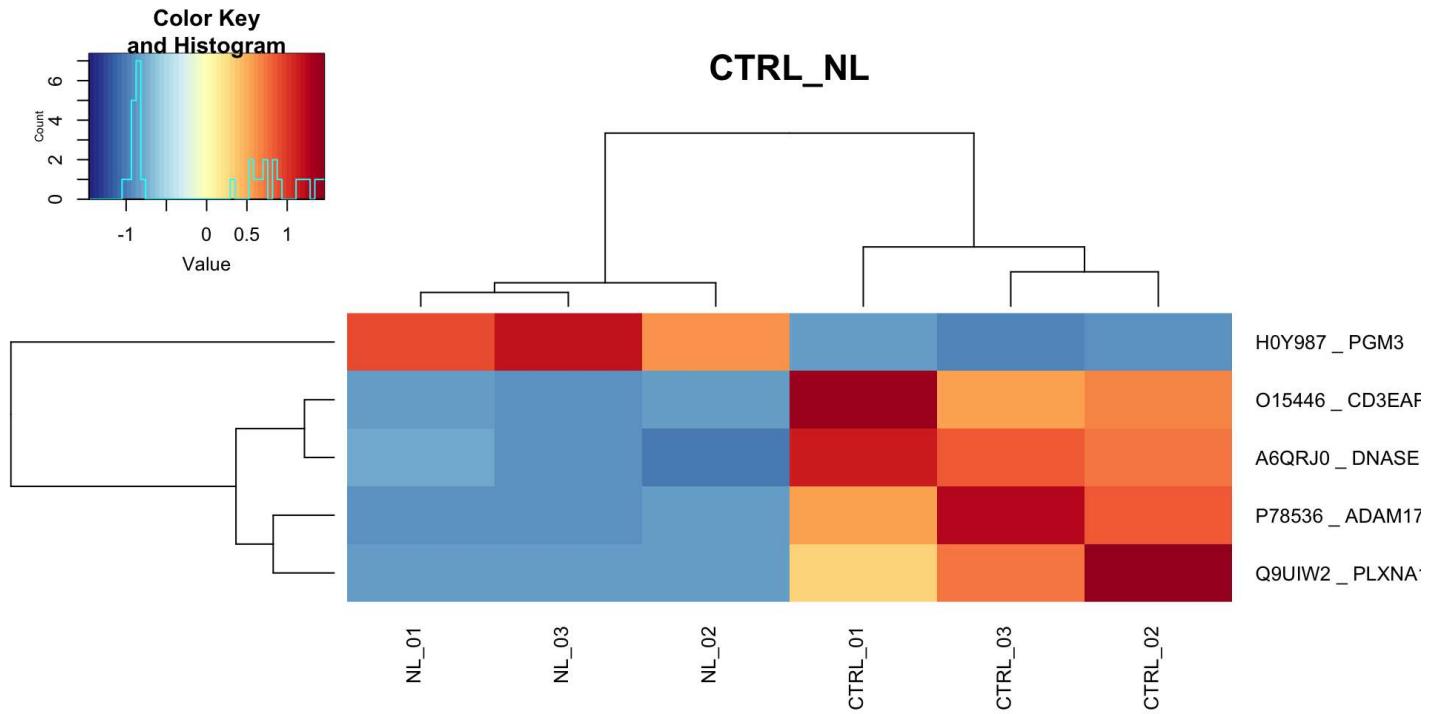
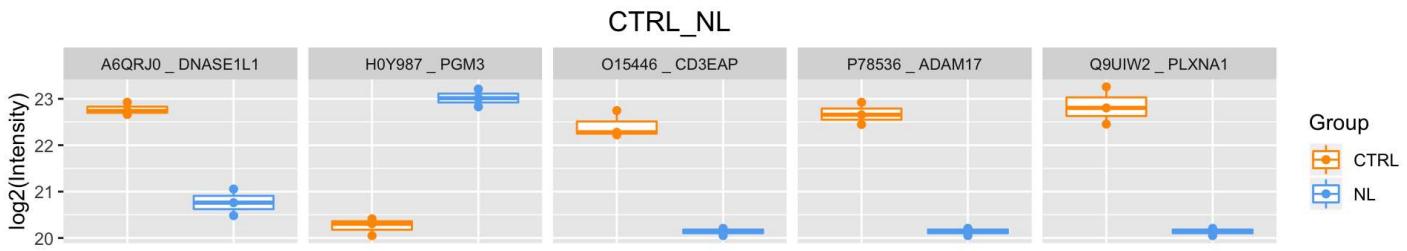
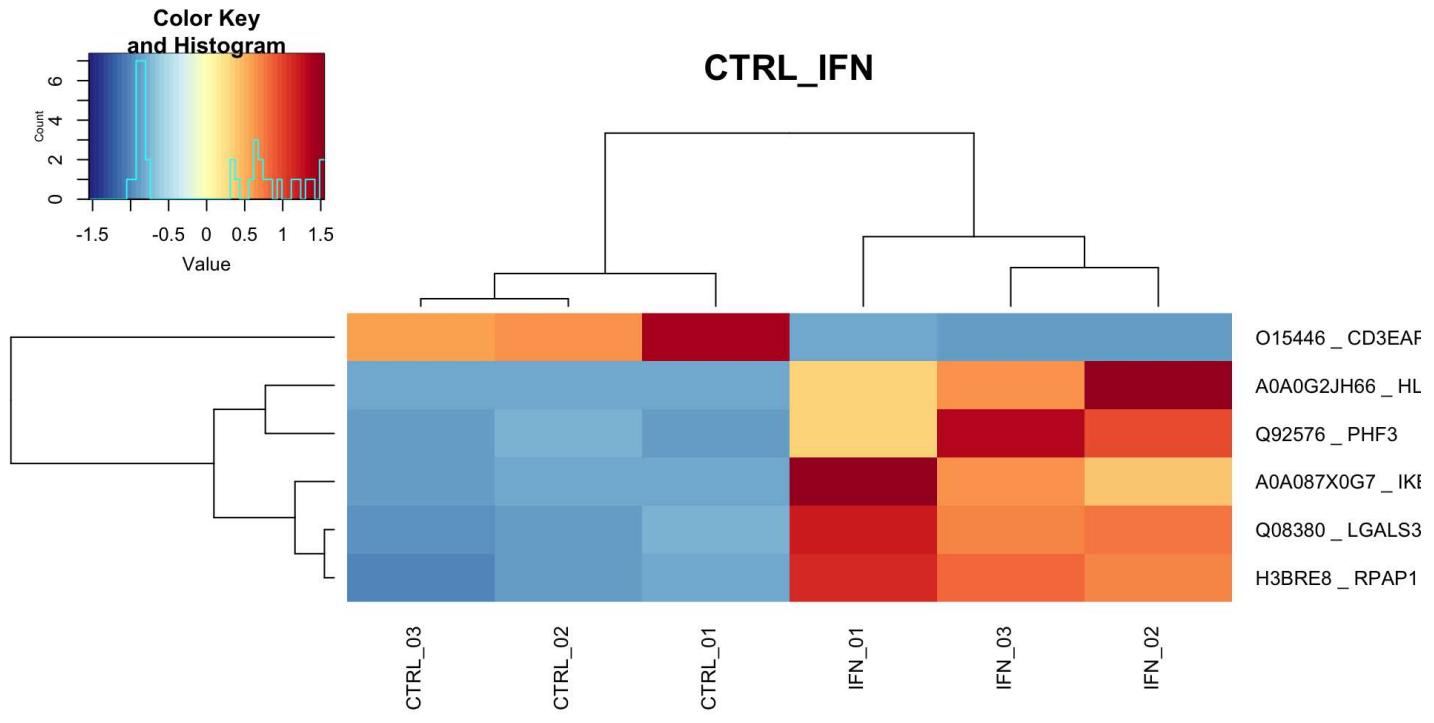
NL_VI



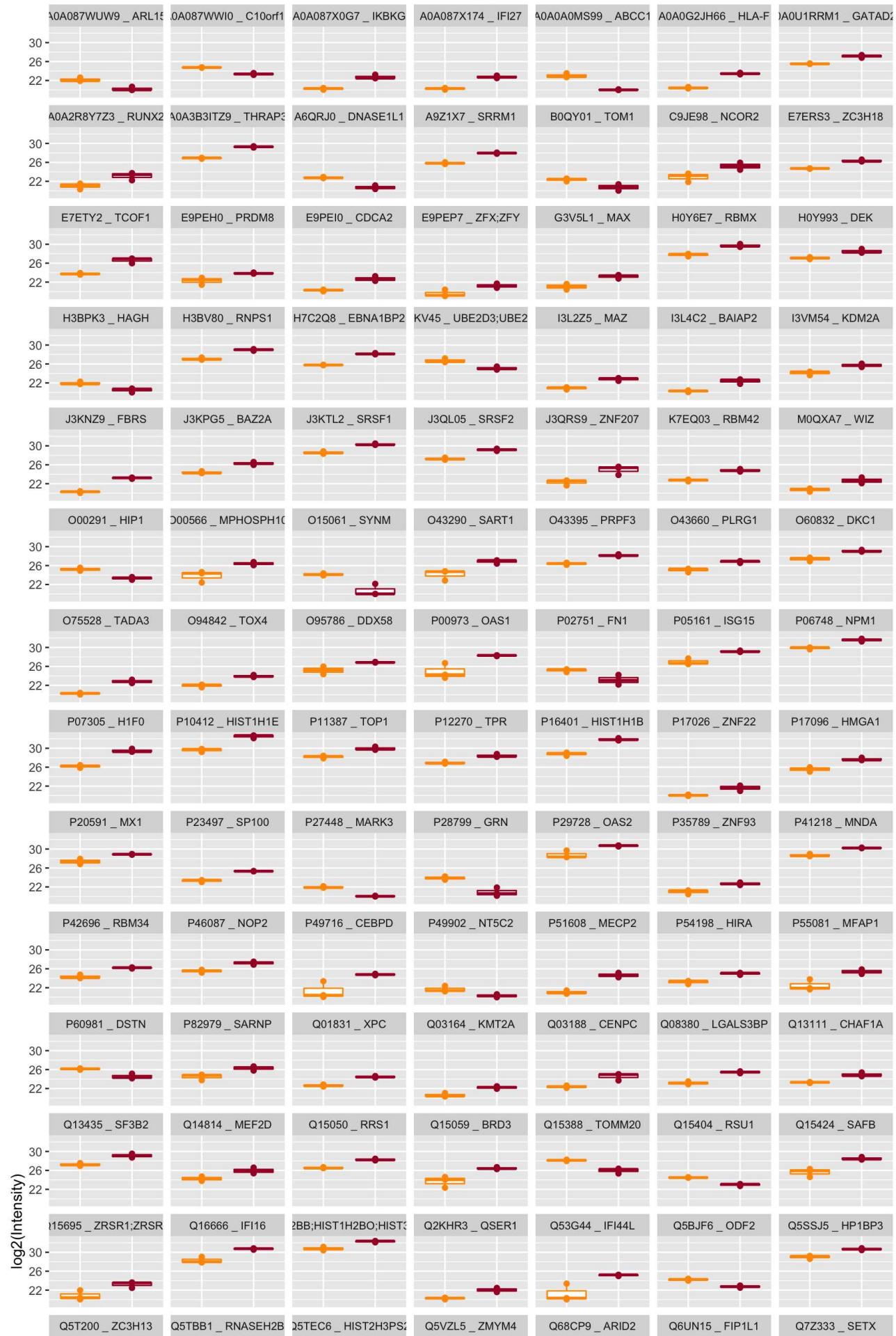
Box-plots and HeatMap of variant proteins

- The box-plots shows the proteins intensity value for all replicates of each group
- The heatmap shows the scaled intensity values for all variant proteins in all replicates, a hierarchical clustering is applied both on rows and columns meaning that samples or proteins having similar profiles appear close to each other.

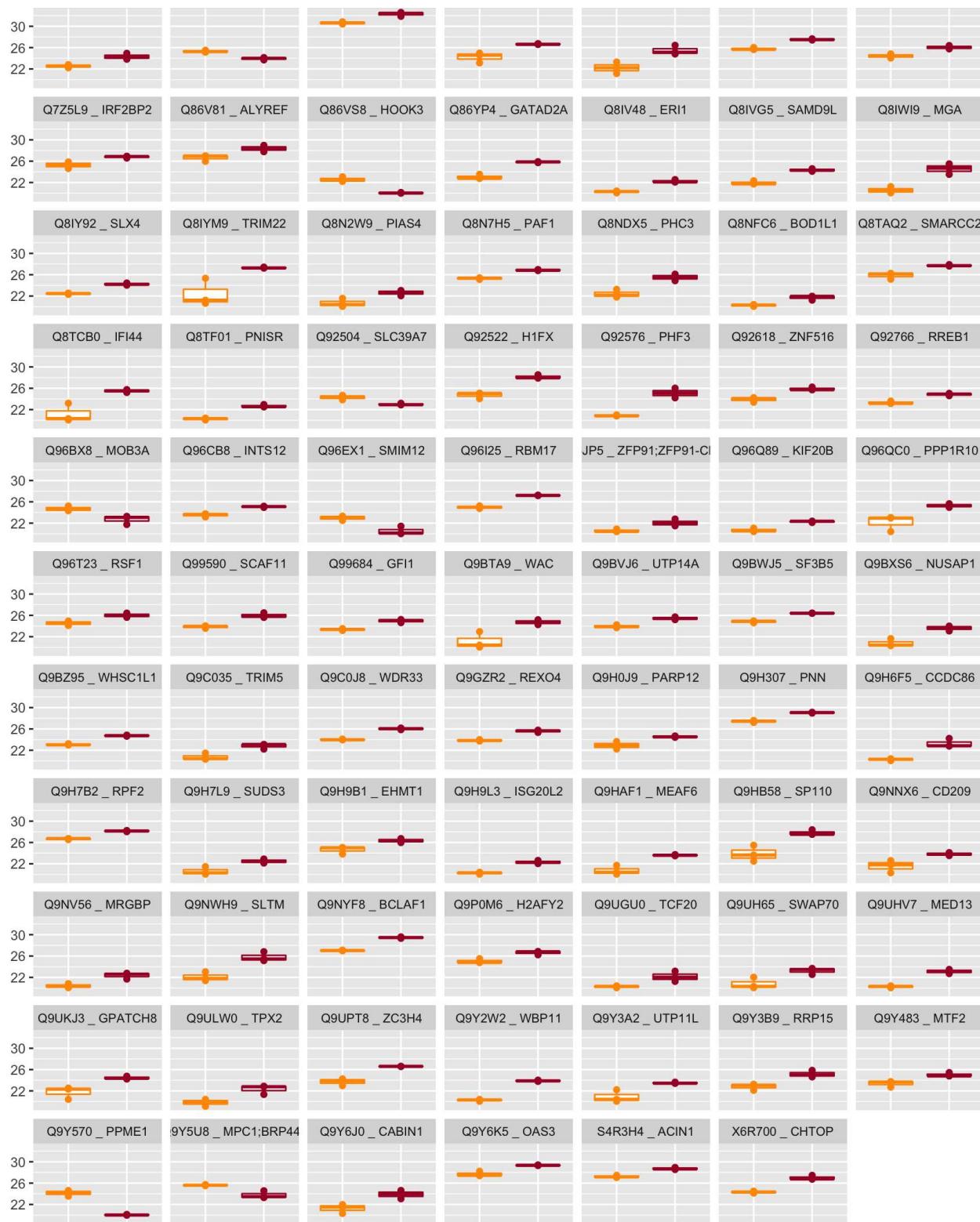


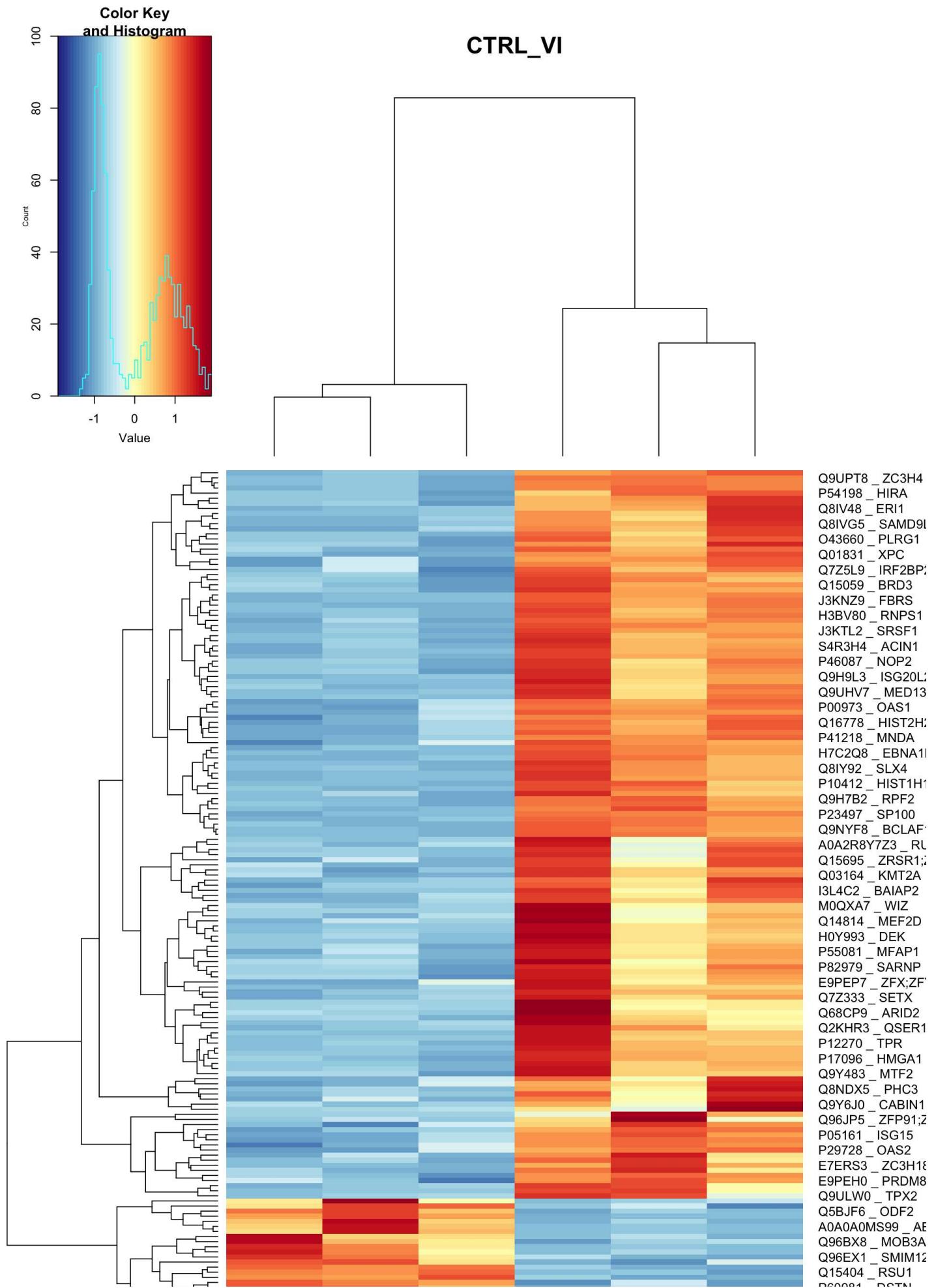


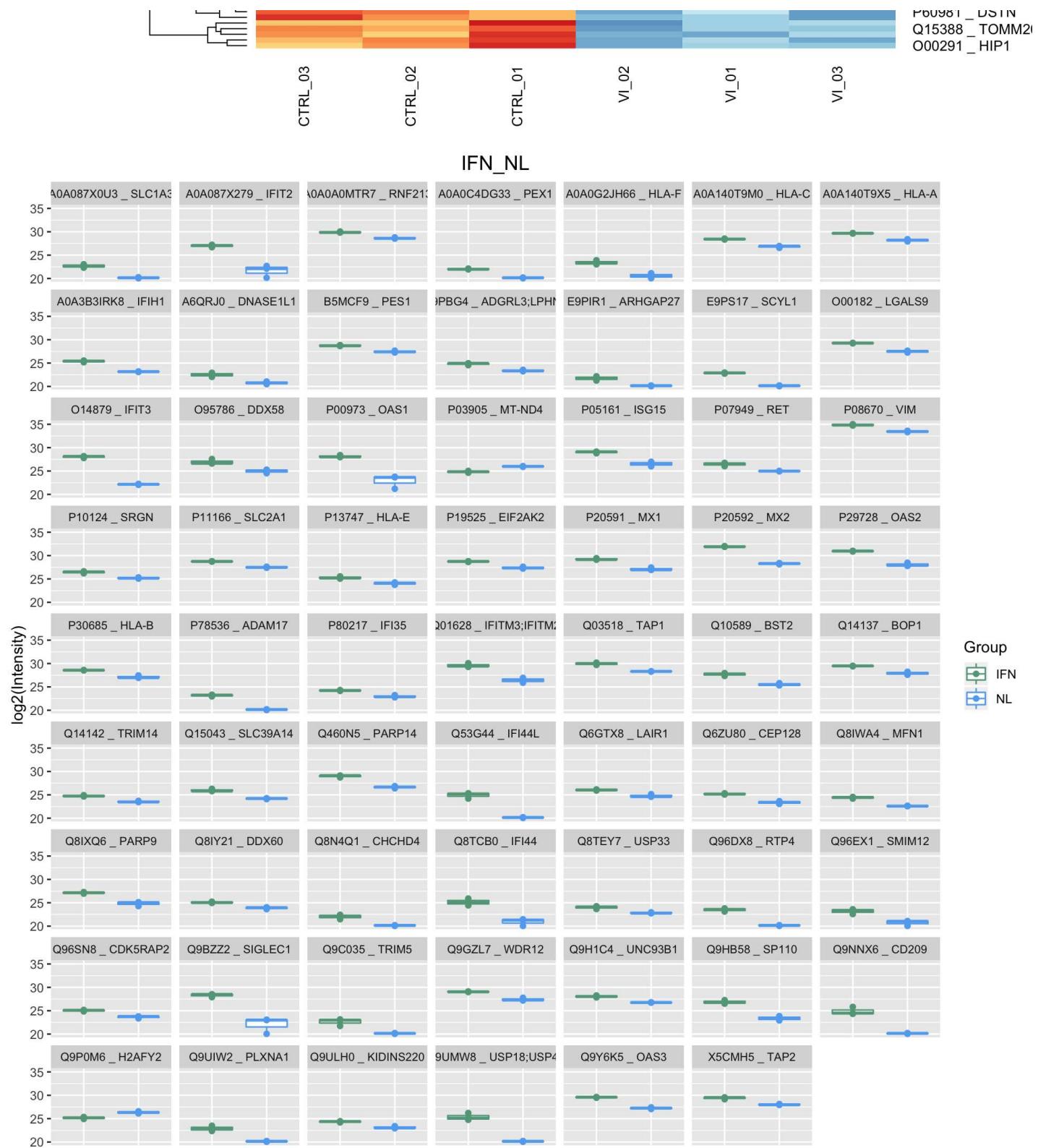
CTRL_VI

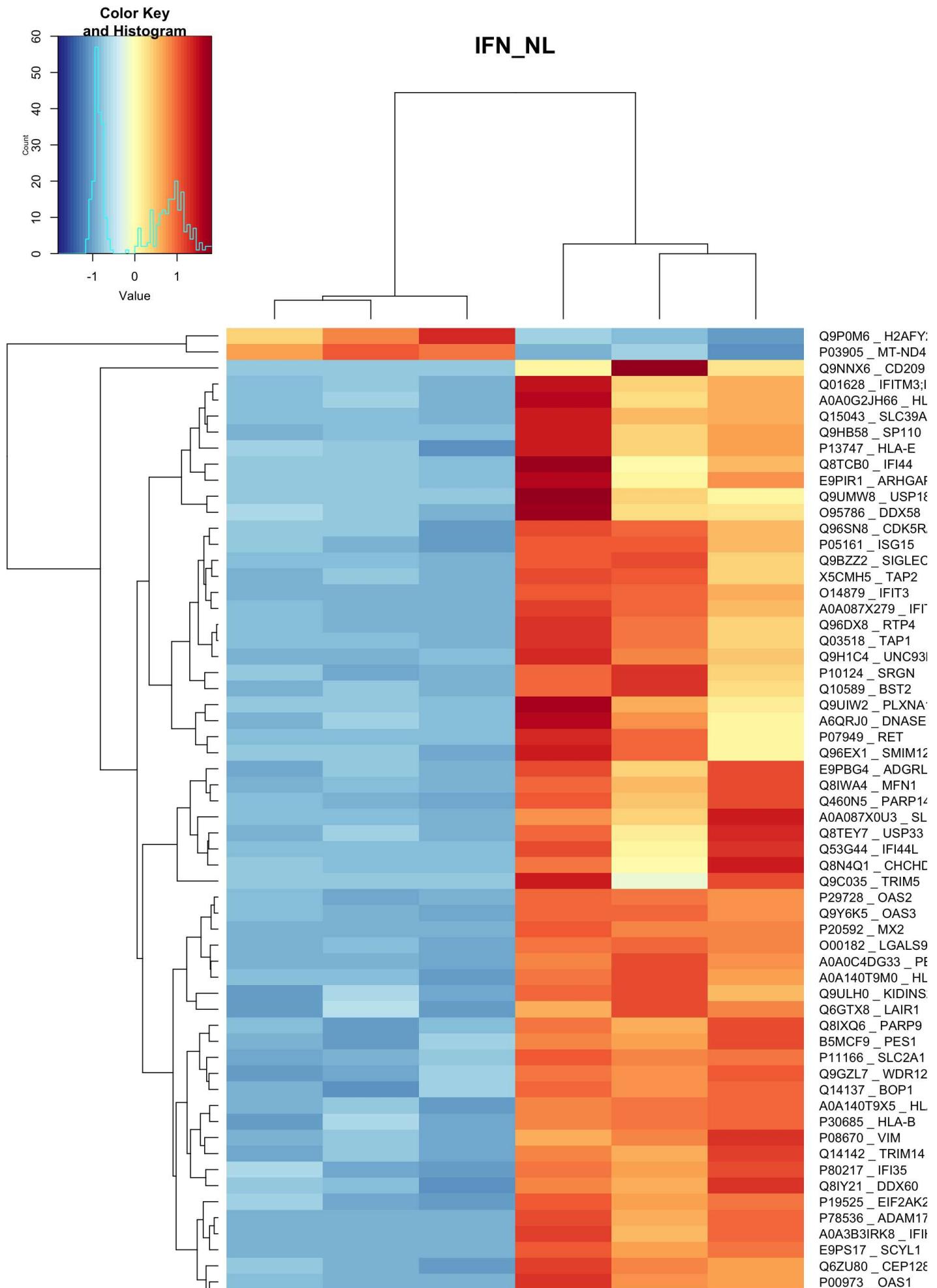


Group
■ CTRL
■ VI



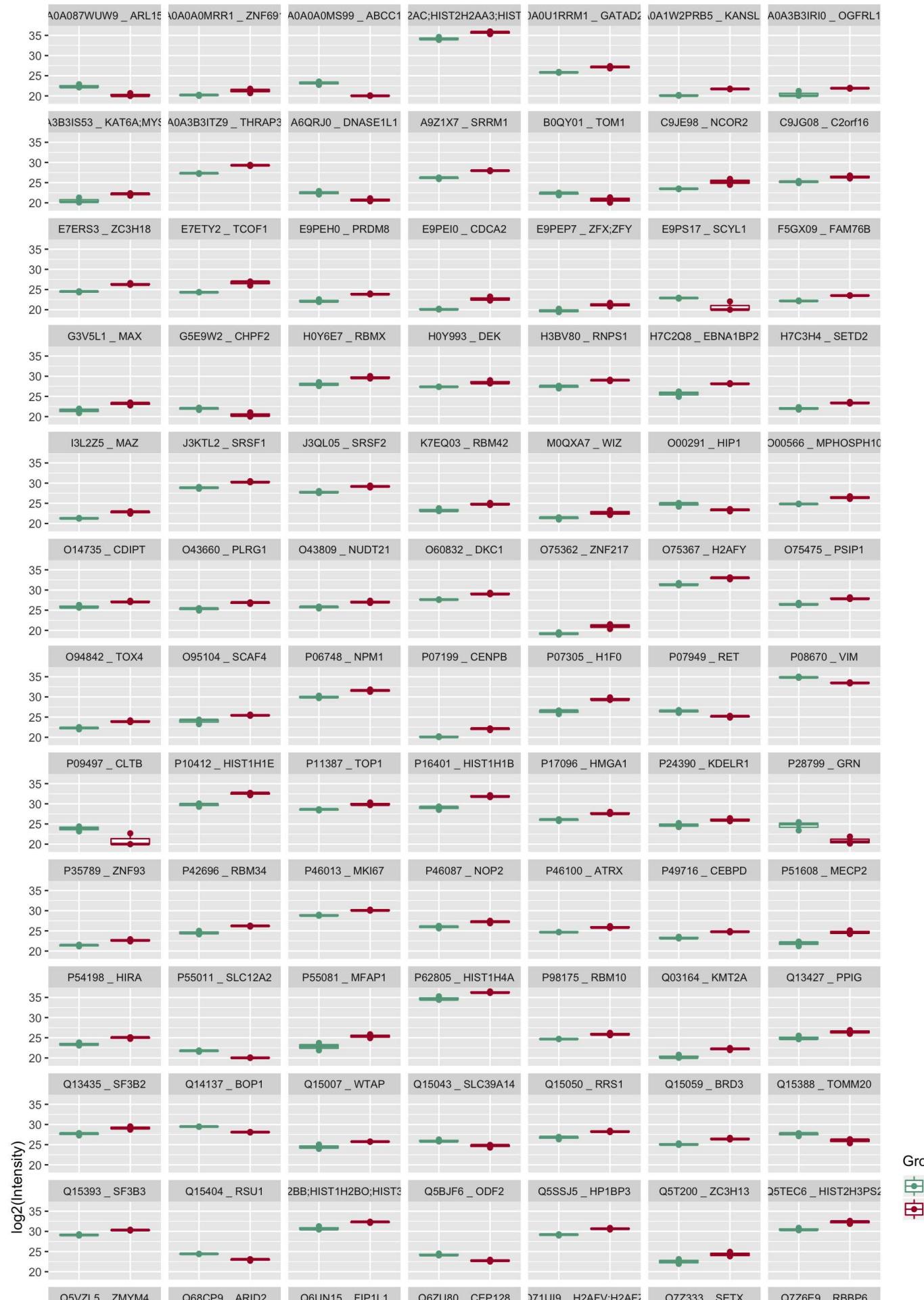




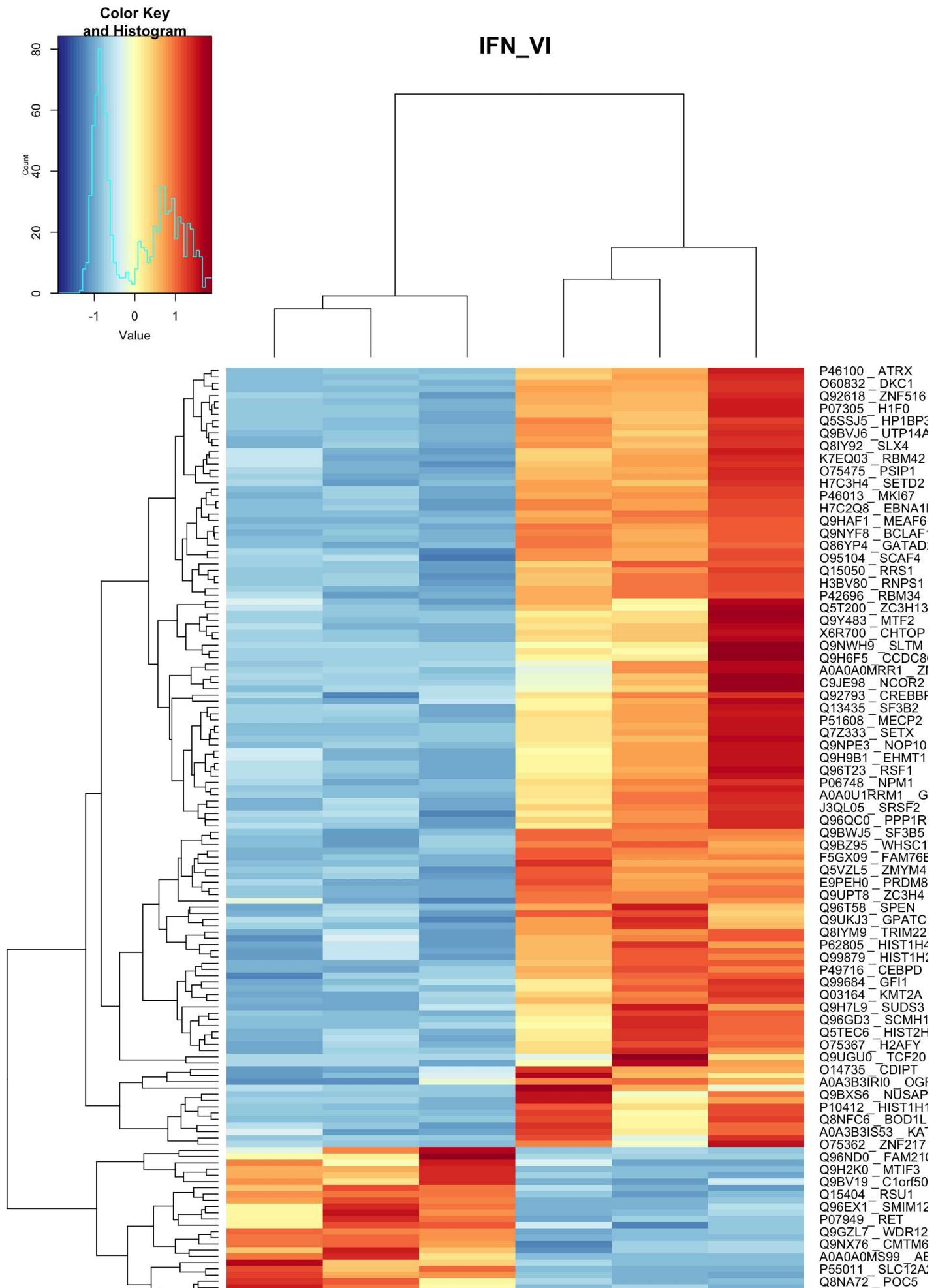


P20591_MX1
A0A0A0MTR7_RI

IFN_VI







O00291_HIP1
G5E9W2_CHPF2
E9PS17_SCYL1

