

## PIERRE WENSEL PROTEOMICS Task-2-dbsearch:

### Introduction

DYRK1A is a protein kinase whose dysregulation is linked to disease in humans. On the one side, its overexpression in trisomy 21 has been linked to certain Down syndrome pathological traits, while inactivating mutations in just one allele that hinders its function are responsible for a rare clinical syndrome.

In the study that we will use for the task, the researchers identify which proteins interact with DYRK1A in the cell by performing immunoprecipitation of DYRK1A in human cell lines with an antibody specific for DYRK1A and with an unspecific antibody as the negative control. . You can find the original article in the `Task-2-dbsearch` folder under the name `Exercise-2-article.pdf`.

### Instructions

- 1.) From the PRIDE repository database (<https://www.ebi.ac.uk/pride/>) search for Project PXD011925 and download the following files:

180208\_S\_CDSL\_13\_01\_45pto.raw

180208\_S\_CDSL\_14\_01\_45pto.raw

180208\_S\_CDSL\_15\_01\_45pto.raw

180208\_S\_CDSL\_04\_01\_45pto.raw

180208\_S\_CDSL\_05\_01\_45pto.raw

180208\_S\_CDSL\_06\_01\_45pto.raw

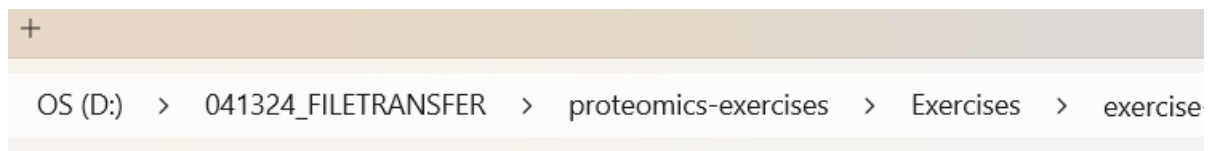
The 3 first files correspond to the negative control (unspecific antibody) and the 3 last files correspond to the DYRK1A specific antibody.

What is the size of each of the files?

- 2.) With these file a MaxQuant database search was performed and the LFQ values were calculated for each protein. You can find the `proteinGroups.txt` file in the exercise folder.
- 3.) In the `proteinGroups.txt` file generated in the MaxQuant use the LFQ values as quantitative measurement.
  - a. Transform the data using the function  $\log_2(x)$
  - b. Remove reverse sequences and contaminants
  - c. Annotate the samples with their corresponding condition
  - d. Keep only proteins that have at least 2 values in at least one group.

- e. For the proteins that you have 2 or 3 values in one condition and 0 values in the other, replace the missing values in the condition where the protein was not detected with the minimum value in the dataset.
- f. Calculate the Fold Change and an adjusted pvalue
- g. Create a Volcano plot
- h. Color the significant proteins in the Volcano Plot in your favorite color and label them with their gene names.
- i. What is the protein that has a bigger Fold change? What is the protein that has a lower adjusted pvalue?

1. The size of the downloaded files are as follows:



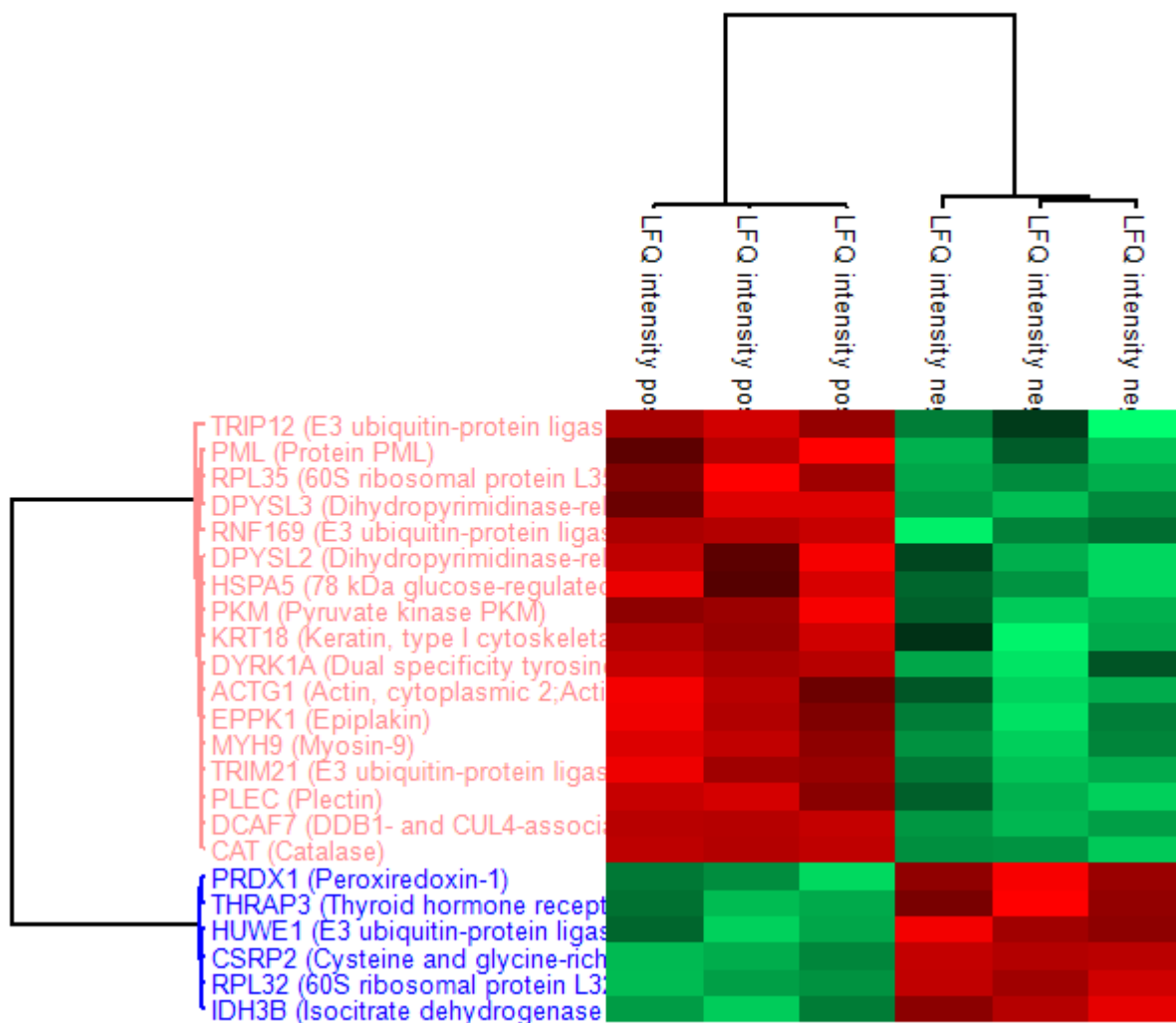
Name	Date modified	Type	Size
combined	4/30/2024 5:10 PM	File folder	
180208_S_CDSL_04_01_45pto	4/29/2024 10:33 PM	RAW File	532,544 KB
180208_S_CDSL_05_01_45pto	4/29/2024 10:33 PM	RAW File	534,969 KB
180208_S_CDSL_06_01_45pto	4/29/2024 10:33 PM	RAW File	538,281 KB
180208_S_CDSL_13_01_45pto	4/29/2024 10:07 PM	RAW File	563,211 KB
180208_S_CDSL_14_01_45pto	4/29/2024 10:10 PM	RAW File	567,260 KB
180208_S_CDSL_15_01_45pto	4/29/2024 10:11 PM	RAW File	523,971 KB

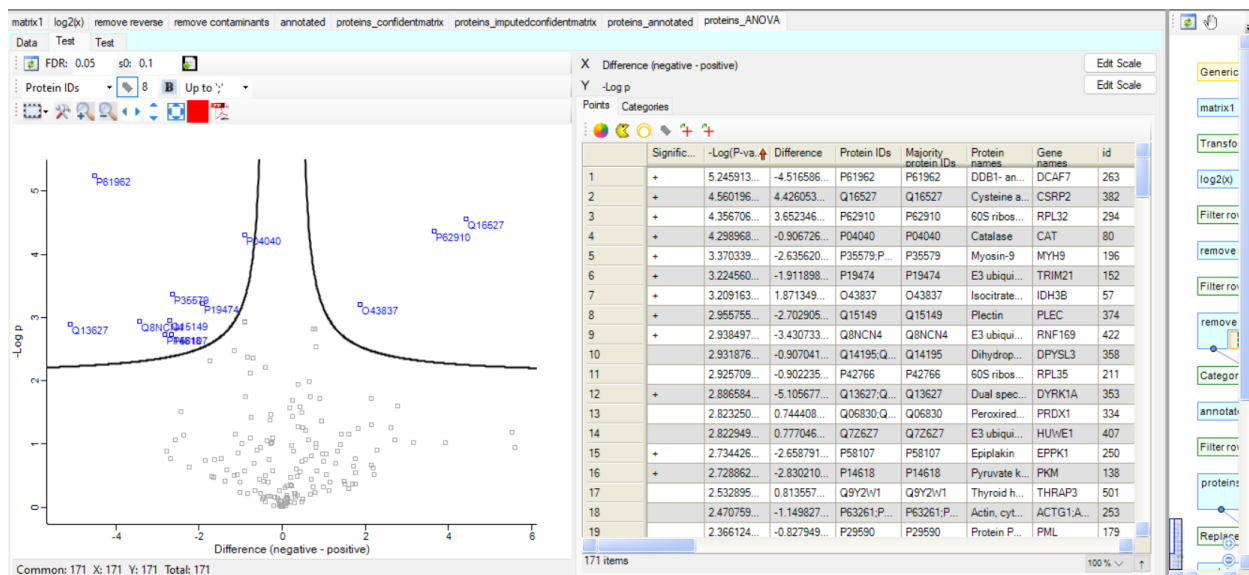
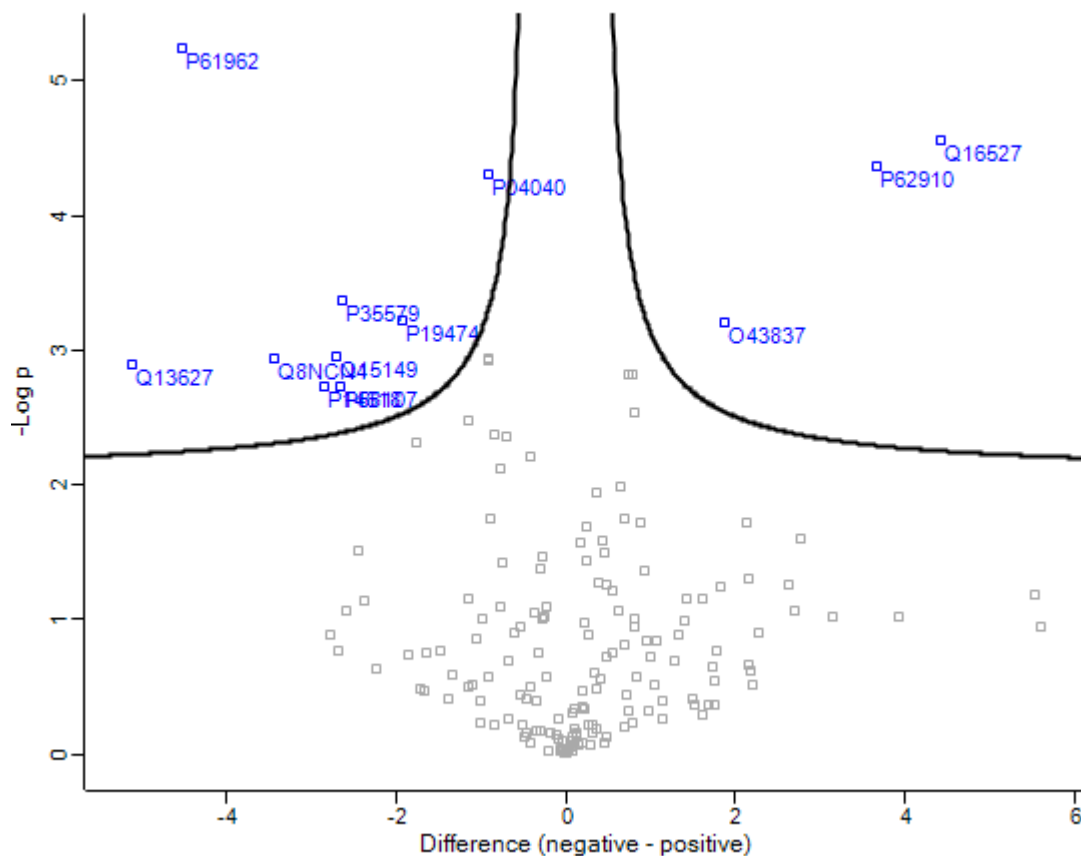
	File	Exists	Size	Data format	Parameter group	Experiment	Fraction	PTM	Refere
1	D:\041324_FILETRANSFER\proteomics-exercises\Exercises\exercise-2-db-search\180208_S_CDSL_04_01_45pto.raw	True	520 MB	Thermo raw...	Group 0			False	
2	D:\041324_FILETRANSFER\proteomics-exercises\Exercises\exercise-2-db-search\180208_S_CDSL_05_01_45pto.raw	True	522.4 MB	Thermo raw...	Group 0			False	
3	D:\041324_FILETRANSFER\proteomics-exercises\Exercises\exercise-2-db-search\180208_S_CDSL_06_01_45pto.raw	True	525.6 MB	Thermo raw...	Group 0			False	
4	D:\041324_FILETRANSFER\proteomics-exercises\Exercises\exercise-2-db-search\180208_S_CDSL_13_01_45pto.raw	True	550 MB	Thermo raw...	Group 0			False	
5	D:\041324_FILETRANSFER\proteomics-exercises\Exercises\exercise-2-db-search\180208_S_CDSL_14_01_45pto.raw	True	553.9 MB	Thermo raw...	Group 0			False	
6	D:\041324_FILETRANSFER\proteomics-exercises\Exercises\exercise-2-db-search\180208_S_CDSL_15_01_45pto.raw	True	511.6 MB	Thermo raw...	Group 0			False	

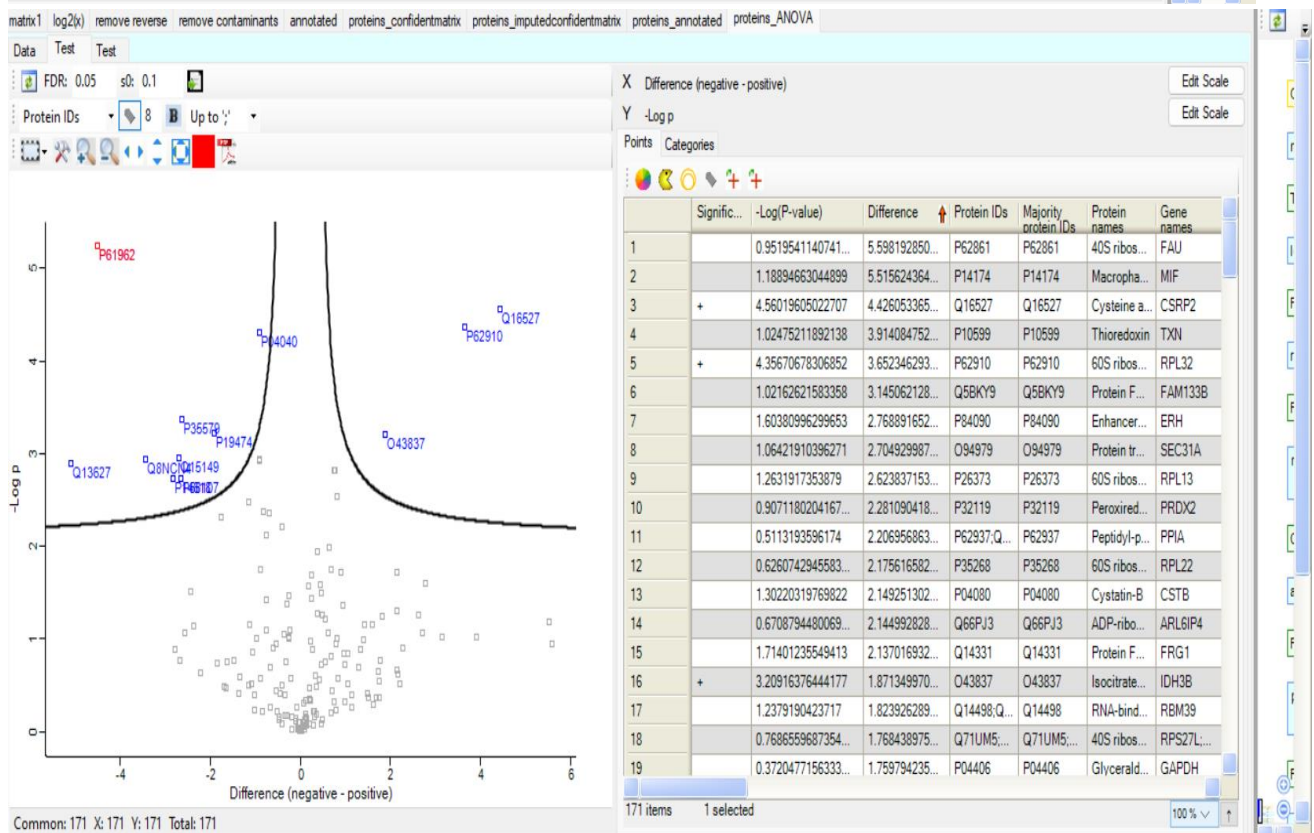
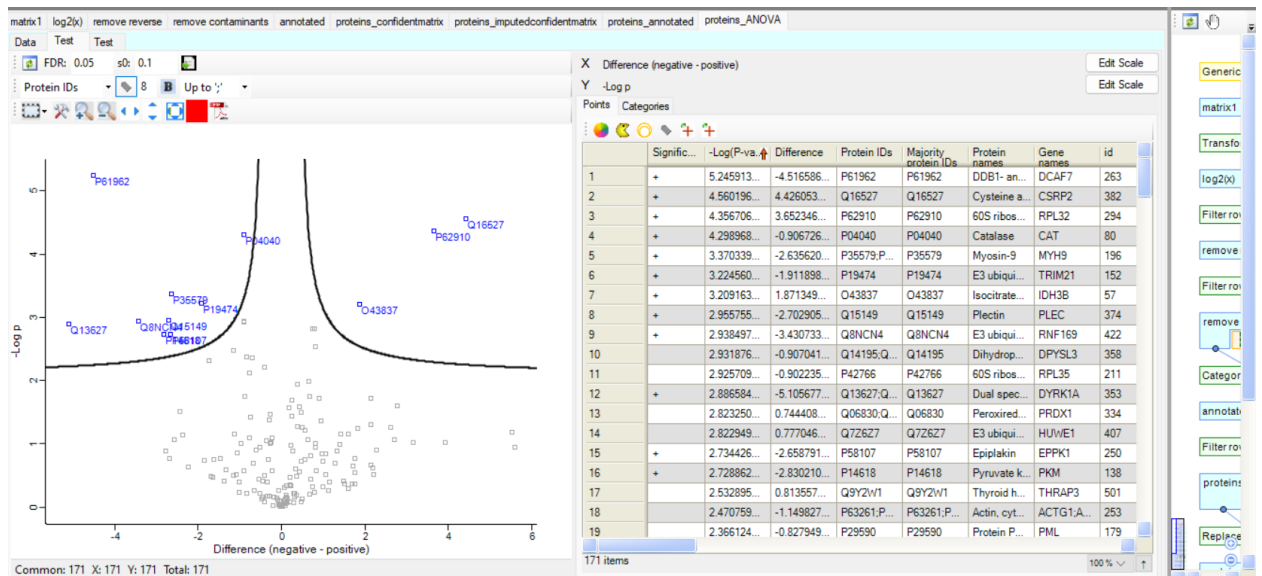
	LFQ intensity negativ...	LFQ intensity negativ...	LFQ intensity negativ...	LFQ intensity positive-1	LFQ intensity positive-2	LFQ intensity positive-3	C: Only identified by site	C: Reverse	C: Potential contam...	C: Taxono... IDs	N: Peptides	N: Razor + unique peptides	N: Unique peptides	N: Sequen... coverage	N: Unique + razor	N: Unique sequence	N: Mol. weight [kDa]	N: Q-value	N
Type	Main	Main	Main	Main	Main	Main	Category	Category	Category	Category	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	N
1	NaN	26.8564	NaN	NaN	NaN	26.6429				-1	1	1	1	13.9	13.9	13.9	12.537	0	1
2	NaN	NaN	NaN	NaN	NaN	NaN				-1	1	1	1	2.9	2.9	2.9	44.601	0	6
3	24.0148	23.8538	23.6302	25.4591	24.9855	23.9659				-1	2	2	2	12	12	12	29.173	0	1
4	NaN	22.212	23.6055	NaN	21.1828	22.8581				-1	6	6	6	5.6	5.6	5.6	293.88	0	4
5	NaN	NaN	NaN	NaN	NaN	NaN				-1	1	1	1	1.6	1.6	1.6	71.695	0	6
6	NaN	NaN	NaN	NaN	NaN	NaN				-1	1	1	1	5.4	5.4	5.4	53.747	0	7
7	NaN	NaN	NaN	NaN	24.6257	26.846			+	-1	4	4	1	15.3	15.3	4.5	46.543	0	6
8	NaN	NaN	NaN	NaN	NaN	NaN			+	-1	4	1	1	22.4	11.2	11.2	43.9	0	6
9	NaN	NaN	NaN	NaN	NaN	NaN			+	-1	1	1	1	15.1	15.1	15.1	14.759	0	7
10	24.222	NaN	23.4354	NaN	NaN	23.5885			+	-1	2	2	2	19.5	19.5	19.5	14.186	0	1
11	31.013	31.1215	31.1346	31.141	31.168	30.9367			+	-1	5	5	5	25.1	25.1	25.1	24.409	0	1
12	NaN	NaN	NaN	NaN	NaN	NaN			+	-1	10	3	0	22	9.7	0	51.561	0	1
13	NaN	NaN	NaN	NaN	NaN	NaN			+	-1	11	1	0	20.2	1.8	0	60.024	0	7
14	27.3819	27.539	27.6106	27.3987	28.4785	26.6862			+	-1	4	4	4	28.1	28.1	28.1	22.975	0	9
15	26.3318	NaN	NaN	NaN	NaN	25.561			+	-1	3	3	3	19.8	19.8	19.8	24.348	0	3
16	NaN	NaN	NaN	NaN	NaN	NaN			+	-1	2	2	2	15.8	15.8	15.8	23.583	0	5
17	NaN	25.8074	25.2639	26.1691	26.2992	NaN			+	-1	4	4	4	32	32	32	18.974	0	2
18	25.9178	25.7371	25.7393	25.6641	25.7928	25.8652			+	-1	7	7	7	59.9	59.9	59.9	18.281	0	1
19	NaN	NaN	26.0096	25.9704	NaN	NaN			+	-1	2	2	2	3.6	3.6	3.6	69.366	0	3

## Log2X transformation

	LFQ intensity negativ...	LFQ intensity negativ...	LFQ intensity negativ...	LFQ intensity positive-1	LFQ intensity positive-2	LFQ intensity positive-3	C: Only identified by site	C: Reverse	C: Potential contam...	C: Taxono... IDs	N: Peptides	N: Razor + unique peptides	N: Unique peptides	N: Sequen... coverage	N: Unique + razor	N: Unique sequence	N: Mol. weight [kDa]	N: Q-value	N
Type	Main	Main	Main	Main	Main	Main	Category	Category	Category	Category	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	N
1	NaN	26.8564	NaN	NaN	NaN	26.6429				-1	1	1	1	13.9	13.9	13.9	12.537	0	1
2	NaN	NaN	NaN	NaN	NaN	NaN				-1	1	1	1	2.9	2.9	2.9	44.601	0	6
3	24.0148	23.8538	23.6302	25.4591	24.9855	23.9659				-1	2	2	2	12	12	12	29.173	0	1
4	NaN	22.212	23.6055	NaN	21.1828	22.8581				-1	6	6	6	5.6	5.6	5.6	293.88	0	4
5	NaN	NaN	NaN	NaN	NaN	NaN				-1	1	1	1	1.6	1.6	1.6	71.695	0	6
6	NaN	NaN	NaN	NaN	NaN	NaN				-1	1	1	1	5.4	5.4	5.4	53.747	0	7
7	NaN	NaN	NaN	NaN	24.6257	26.846			+	-1	4	4	1	15.3	15.3	4.5	46.543	0	6
8	NaN	NaN	NaN	NaN	NaN	NaN			+	-1	4	1	1	22.4	11.2	11.2	43.9	0	6
9	NaN	NaN	NaN	NaN	NaN	NaN			+	-1	1	1	1	15.1	15.1	15.1	14.759	0	7
10	24.222	NaN	23.4354	NaN	NaN	23.5885			+	-1	2	2	2	19.5	19.5	19.5	14.186	0	1
11	31.013	31.1215	31.1346	31.141	31.168	30.9367			+	-1	5	5	5	25.1	25.1	25.1	24.409	0	1
12	NaN	NaN	NaN	NaN	NaN	NaN			+	-1	10	3	0	22	9.7	0	51.561	0	1
13	NaN	NaN	NaN	NaN	NaN	NaN			+	-1	11	1	0	20.2	1.8	0	60.024	0	7
14	27.3819	27.539	27.6106	27.3987	28.4785	26.6862			+	-1	4	4	4	28.1	28.1	28.1	22.975	0	9
15	26.3318	NaN	NaN	NaN	NaN	25.561			+	-1	3	3	3	19.8	19.8	19.8	24.348	0	3
16	NaN	NaN	NaN	NaN	NaN	NaN			+	-1	2	2	2	15.8	15.8	15.8	23.583	0	5
17	NaN	25.8074	25.2639	26.1691	26.2992	NaN			+	-1	4	4	4	32	32	32	18.974	0	2
18	25.9178	25.7371	25.7393	25.6641	25.7928	25.8652			+	-1	7	7	7	59.9	59.9	59.9	18.281	0	1
19	NaN	NaN	26.0096	25.9704	NaN	NaN			+	-1	2	2	2	3.6	3.6	3.6	69.366	0	3







The proteinID with smallest adjusted p-value of  $-\log p\text{-value}=5.2459$  is P61962  
 The proteinID with biggest positive difference of 5.5981 (largest positive fold-change) is P62861

The proteinID with biggest adjusted p-value of  $-\log p\text{-value}=0.0058$  is P13639

The protein ID with largest negative Difference of  $-5.1056$  (largest negative fold-change) is Q13627:Q