microarrays

July 28, 2020

0.1 Microarray analyses for C.glutamicum

This analyses aims for the following:

- find the groups of genes which cooperatively change their expression in a wide range of conditions
- find the groups of experiments with similar gene expression patterns
- find blocks of genes which behave similarly in particular sets of experiments

We have already processed the initial table in a way that: * it is filled by numeric values * formated to tab-separated version (.tsv)

Now we download the table as DataFrame:

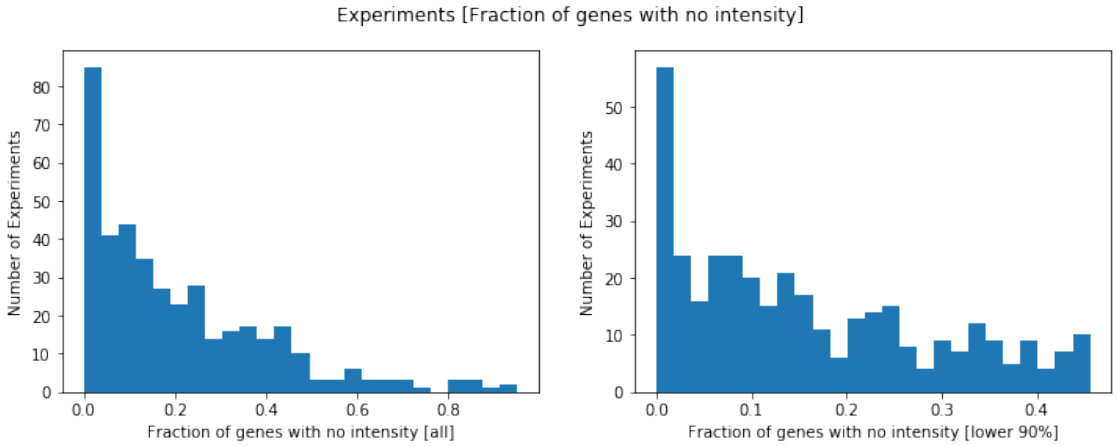
Basic statistics :

```
Number of experiments:  403
Number of genes:        3047
Number of experiments with no signal: 1
Number of genes with no signal: 0
```

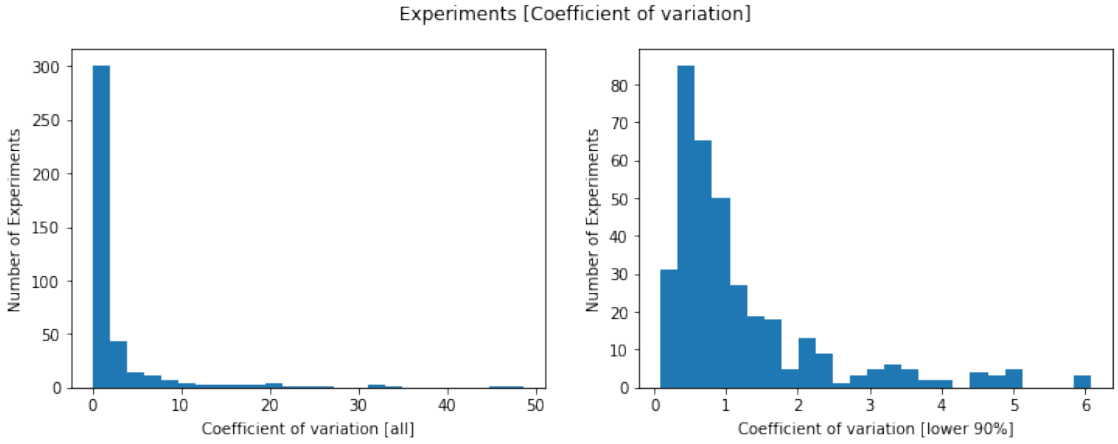
0.2 Remove genes and experiments without signal

```
Number of experiments:  402
Number of genes:        3047
```

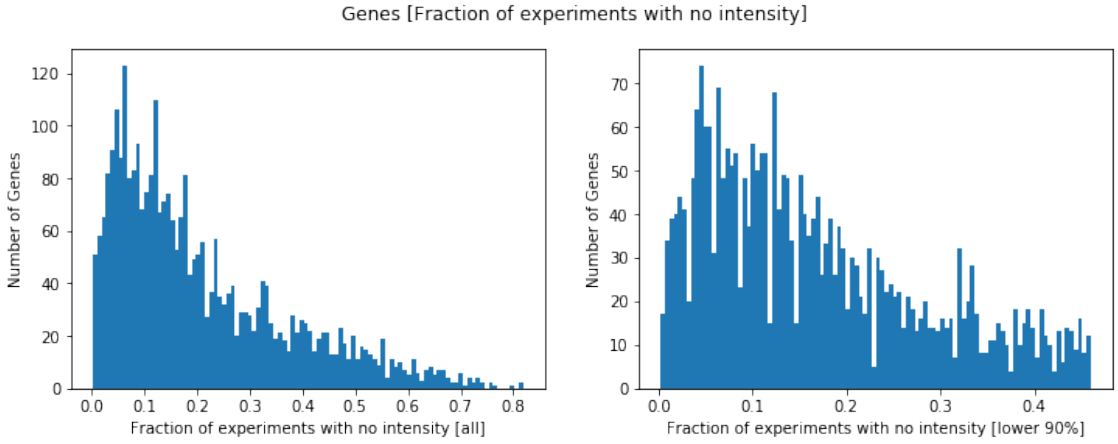
0.3 Experiments basic statistics:



Interpretation: we observe a significant fraction of experiments with zero expression for more than 50% percent of genes. Further these experiments will be deleted from the initial table to ensure better performance of clustering algorithms

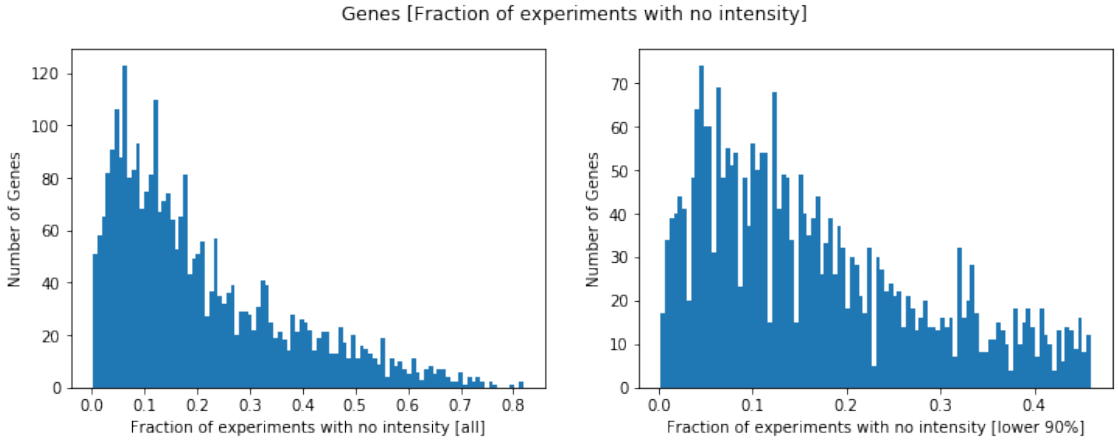


Interpretation: we observe a number of experiments with abnormal (far away from average) total intensity. Further the experiments will be deleted to avoid potential biases.

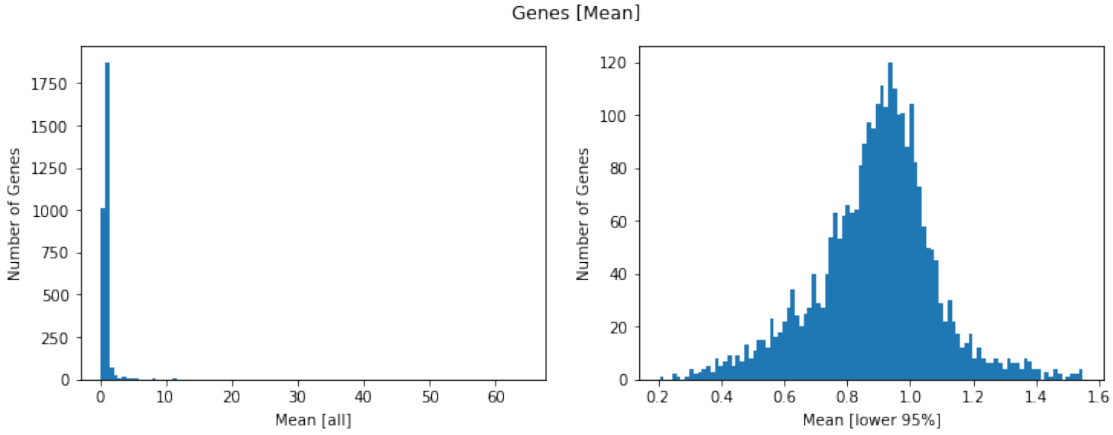


Interpretation: Experiments show pretty wide range of gene expression variation. However we will not delete outliers, since they may reflect strong experimental perturbations.

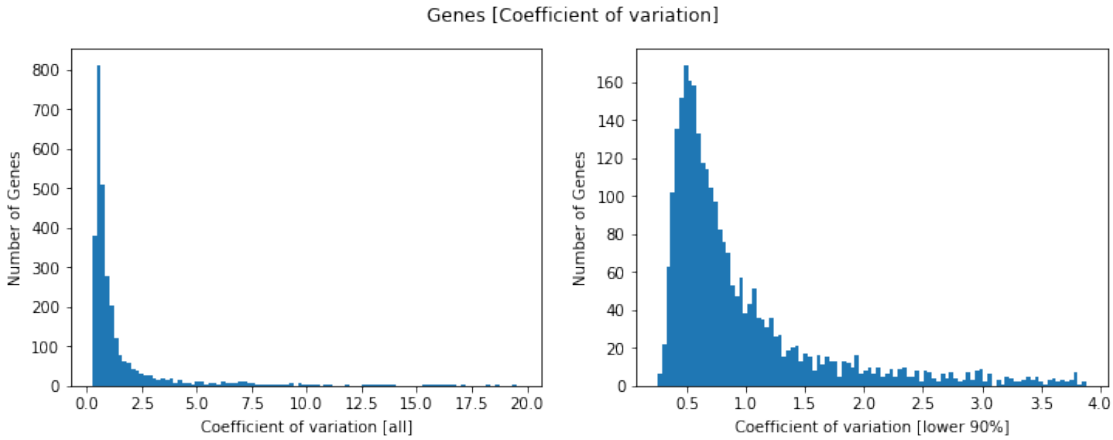
0.4 Genes basic statistics:



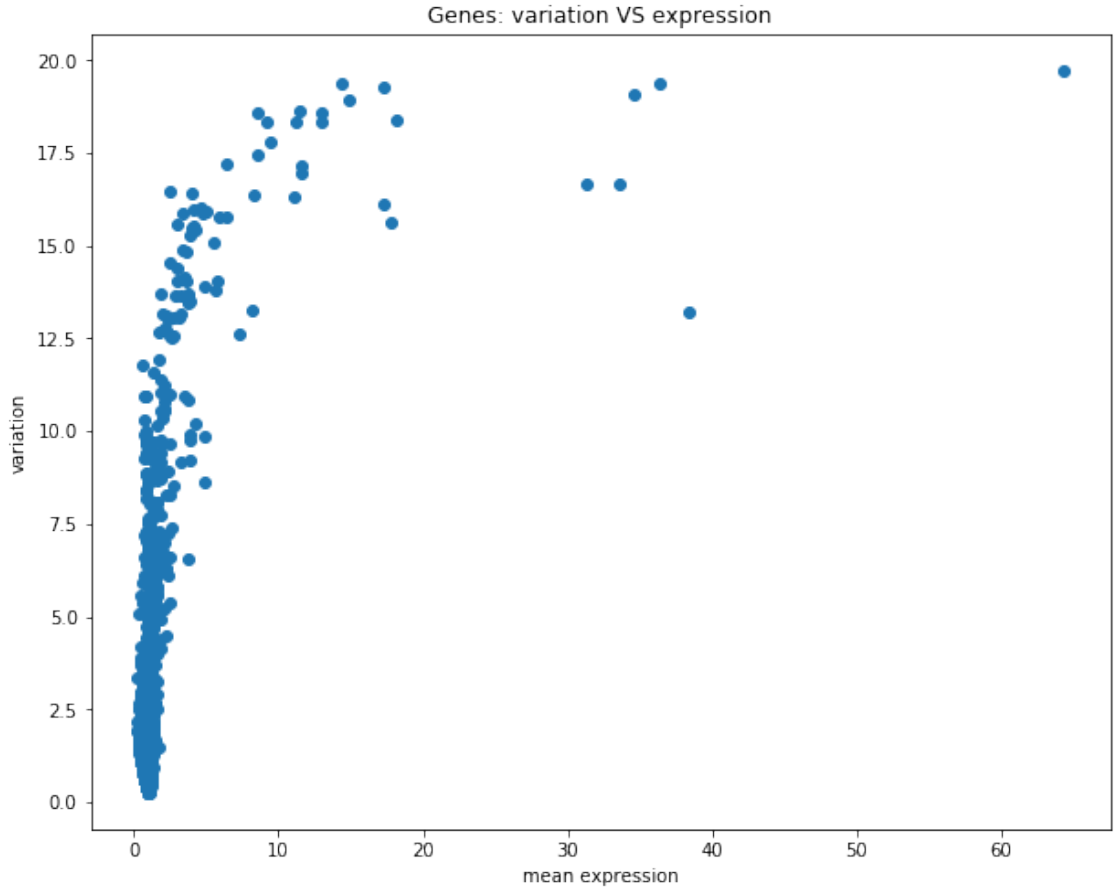
Interpretation: we observe a significant fraction of genes with zero expression in more than 50% percent of experiments. Further these genes will be deleted from the initial table to ensure better performance of clustering algorithms



Interpretation: Most of the genes are expressed with the similar intensity (it is quite counterintuitive?) with several outliers. Genes with low expression will deleted, as their expression may be strongly affected by noise.



Interpretation: Genes are expected to vary upon different conditions and genes with high variations are more important for succesfull clustering. The fact that we have a decent fraction of genes with high variation is promising.



```
Pearson correlation: 0.5472
```

Interpretation: In ideal case scenario variation should not depend on the intensity. We should be aware of potential experimental biases.

0.5 Data Clearance

Here we remove genes and experiments which may disturb/bias the downstream analyses

```
Filtering by nonzero fraction (75%):
Number of experiments:  333(402)
Number of genes:        2095(3047)
```

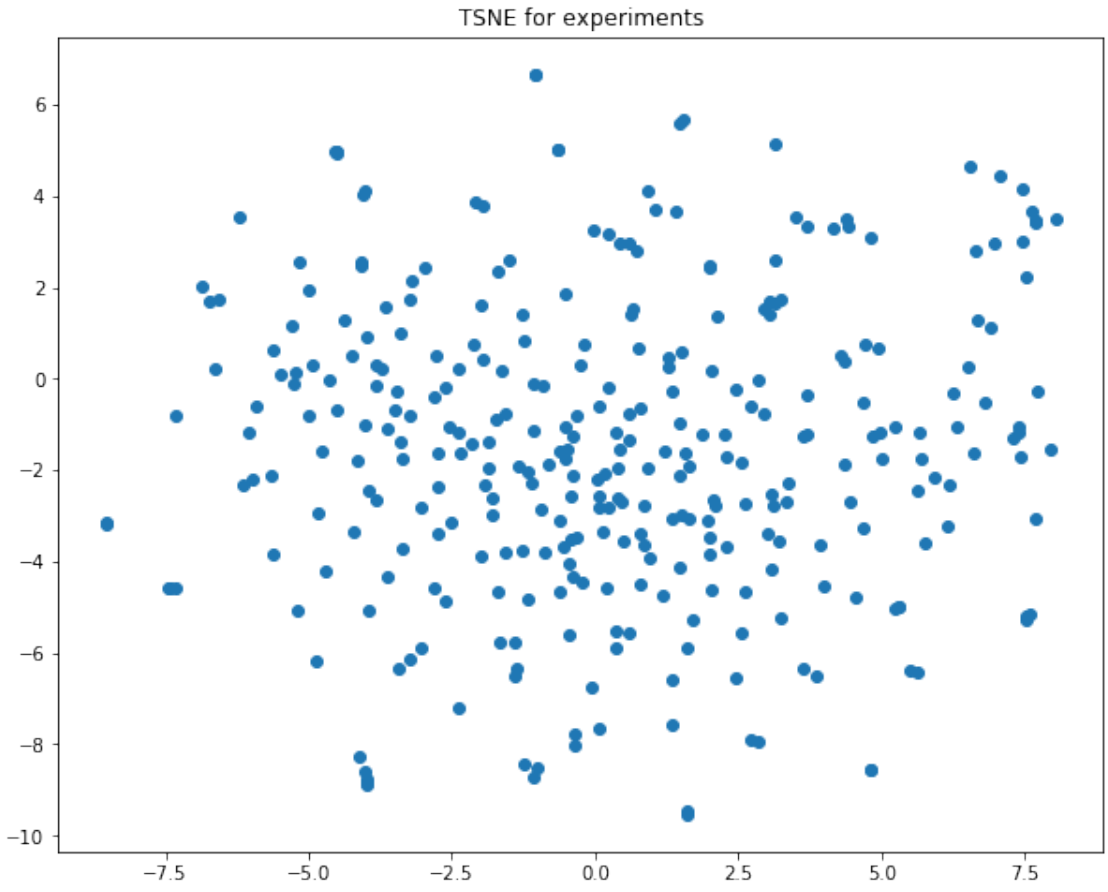
```
Filtering by abnormal intensity (genes: 0.8~1000000.0; experiments: 0.6~2.0):
Number of experiments:  324(402)
Number of genes:        2092(3047)
```

0.6 We split the DataFrame into:

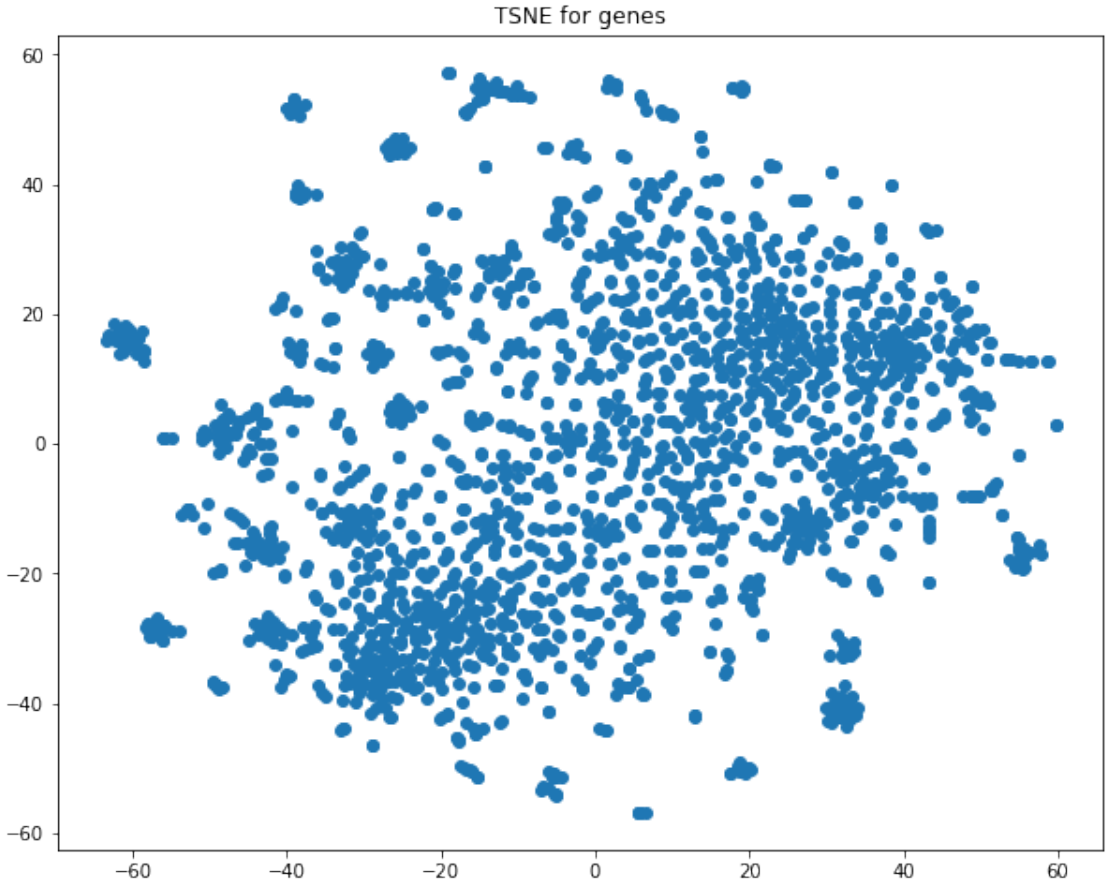
- microarray intensities (DataFrame values) as numpy array
- experiment names (DataFrame rows) as list
- gene names (DataFrame columns) as list

0.7 TSNE analyses

TSNE is a simple way to represent multidimensional data into 2D map. This is done to visually inspect the cluster structure of the data. Please find more here: https://www.displayr.com/using-t-sne-to-visualize-data-before-prediction/



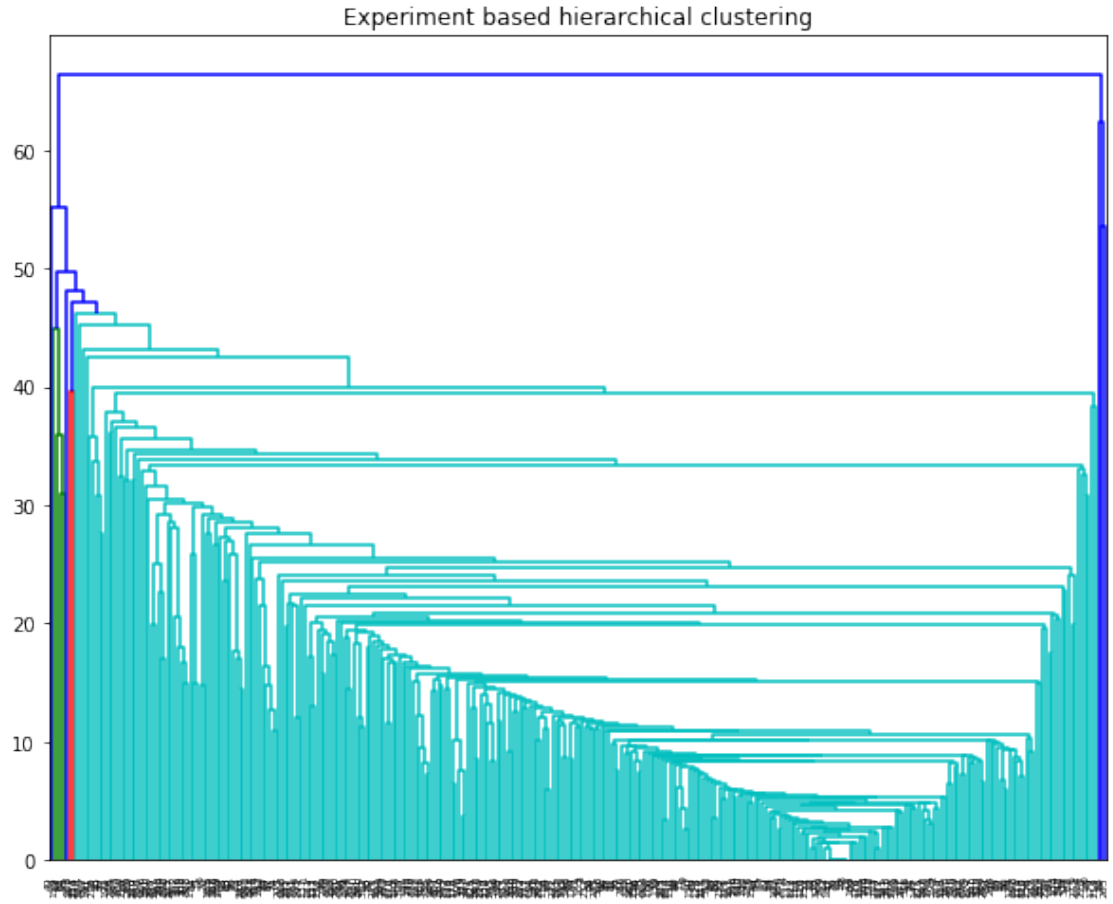
Interpretation: Comprehension of the multidimensional expression data (expression of each gene is a dimension for an experiment) does not show any obvious cluster. However it does not mean, that there are no clusters for the experiments.



Interpretation: Comprehension of the multidimensional expression data (expression in each experiment is a dimension for a gene) shows several small-size clusters. It gives a promise that groups of genes with similar expression patterns can be detected.

0.8 Hierarchical Clustering

Hierarchical Clustering is an interpretable method to find groups of similar objects (in our case: genes and experiments). Please find more here: https://www.displayr.com/what-is-hierarchical-clustering/



0.9 Table of similar Experiments

```
Conditions      distance
set_ACh_WT pEKEx2-GFP ind_vs_WT pEKEx2-GFP unind (18.11.2019)   0.000
set_ACh_WT pEKExL-cg0955-GFP ind_vs_WT pEKExL-cg0955 unind 30 min (18.11.2019)
0.000

Conditions      distance
set_ACh_WT pEKExL-cg0955-GFP ind_vs_WT pEKExL-cg0955-GFP unind (18.11.2019)
0.000
set_ACh_WT pEKEx2-GFP ind_vs_WT pEKEx2-GFP unind (18.11.2019)   0.000
set_ACh_WT pEKExL-cg0955-GFP ind_vs_WT pEKExL-cg0955 unind 30 min (18.11.2019)
0.000

Conditions      distance
set_ACh_WT pEKExL-cg3287-GFP ind_vs_WT pEKExL-cg3287-GFP unind (18.11.2019)
0.000
set_ACh_WT pEKExL-cg0955-GFP ind_vs_WT pEKExL-cg0955-GFP unind (18.11.2019)
0.000
```

```
set_ACh_WT pEKEx2-GFP ind_vs_WT pEKEx2-GFP unind (18.11.2019)   0.000
set_ACh_WT pEKExL-cg0955-GFP ind_vs_WT pEKExL-cg0955 unind 30 min (18.11.2019)
0.000

Conditions      distance
set_AH_Dcg2200 Dcg2201 2.5 uM FeSO4_vs_WT 2.5 uM FeSo4 (20.9.2019)      0.000
set_ACh_WT pEKExL-cg3287-GFP ind_vs_WT pEKExL-cg3287-GFP unind (18.11.2019)
0.000
set_ACh_WT pEKExL-cg0955-GFP ind_vs_WT pEKExL-cg0955-GFP unind (18.11.2019)
0.000
set_ACh_WT pEKEx2-GFP ind_vs_WT pEKEx2-GFP unind (18.11.2019)   0.000
set_ACh_WT pEKExL-cg0955-GFP ind_vs_WT pEKExL-cg0955 unind 30 min (18.11.2019)
0.000

Conditions      distance
set_AKoK_Dcg1300 Dcg1301_vs_WT (5.11.2019)      0.000
set_AH_Dcg2200 Dcg2201 2.5 uM FeSO4_vs_WT 2.5 uM FeSo4 (20.9.2019)      0.000
set_ACh_WT pEKExL-cg3287-GFP ind_vs_WT pEKExL-cg3287-GFP unind (18.11.2019)
0.000
set_ACh_WT pEKExL-cg0955-GFP ind_vs_WT pEKExL-cg0955-GFP unind (18.11.2019)
0.000
set_ACh_WT pEKEx2-GFP ind_vs_WT pEKEx2-GFP unind (18.11.2019)   0.000
set_ACh_WT pEKExL-cg0955-GFP ind_vs_WT pEKExL-cg0955 unind 30 min (18.11.2019)
0.000

Conditions      distance
set_MB_Dcg1083 Dcg1084_vs_WT (8.11.2019)        0.000
set_AKoK_Dcg1300 Dcg1301_vs_WT (5.11.2019)      0.000
set_AH_Dcg2200 Dcg2201 2.5 uM FeSO4_vs_WT 2.5 uM FeSo4 (20.9.2019)      0.000
set_ACh_WT pEKExL-cg3287-GFP ind_vs_WT pEKExL-cg3287-GFP unind (18.11.2019)
0.000
set_ACh_WT pEKExL-cg0955-GFP ind_vs_WT pEKExL-cg0955-GFP unind (18.11.2019)
0.000
set_ACh_WT pEKEx2-GFP ind_vs_WT pEKEx2-GFP unind (18.11.2019)   0.000
set_ACh_WT pEKExL-cg0955-GFP ind_vs_WT pEKExL-cg0955 unind 30 min (18.11.2019)
0.000

Conditions      distance
set_ACh_WT pEKEx2-TorA-cg2705_vs_WT pEKEx2 1 hour (18.11.2019)  1.000
set_MB_Dcg1083 Dcg1084_vs_WT (8.11.2019)        1.000
set_AKoK_Dcg1300 Dcg1301_vs_WT (5.11.2019)      1.000
set_AH_Dcg2200 Dcg2201 2.5 uM FeSO4_vs_WT 2.5 uM FeSo4 (20.9.2019)      1.000
set_ACh_WT pEKExL-cg3287-GFP ind_vs_WT pEKExL-cg3287-GFP unind (18.11.2019)
1.000
set_ACh_WT pEKExL-cg0955-GFP ind_vs_WT pEKExL-cg0955-GFP unind (18.11.2019)
1.000
set_ACh_WT pEKEx2-GFP ind_vs_WT pEKEx2-GFP unind (18.11.2019)   1.000
set_ACh_WT pEKExL-cg0955-GFP ind_vs_WT pEKExL-cg0955 unind 30 min (18.11.2019)
```

1.000

Conditions distance
set_AFK_Dcg1689_vs_WT_CGXII (25.9.2019) 1.000
set_AH_Dcg1890_vs_WT (28.10.2019) 1.000

Conditions distance
set_MBa_Dcg3210_vs_WT (2.10.2019) 1.000
set_MiB_Dcg0764_vs_WT CGXII (12.11.2019) 1.000

Conditions distance
set_AH_Dcg2040_vs_WT (28.10.2019) 1.414
set_CS_WT 500 ug/mL Ethambutol_vs_WT 0 h (3.12.2019) 1.414

Conditions distance
set_LK_WT 5 mM KNO3_vs_WT (7.11.2019) 1.414
set_AH_Dcg2040_vs_WT (28.10.2019) 1.414
set_CS_WT 500 ug/mL Ethambutol_vs_WT 0 h (3.12.2019) 1.414

Conditions distance
set_ACh_WT pEKEx2-TorA-cg2705_vs_WT pEKEx2 1 hour (18.11.2019) 1.414
set_MB_Dcg1083 Dcg1084_vs_WT (8.11.2019) 1.414
set_AKoK_Dcg1300 Dcg1301_vs_WT (5.11.2019) 1.414
set_AH_Dcg2200 Dcg2201 2.5 uM FeSO4_vs_WT 2.5 uM FeSo4 (20.9.2019) 1.414
set_ACh_WT pEKExL-cg3287-GFP ind_vs_WT pEKExL-cg3287-GFP unind (18.11.2019)
1.414
set_ACh_WT pEKExL-cg0955-GFP ind_vs_WT pEKExL-cg0955-GFP unind (18.11.2019)
1.414
set_ACh_WT pEKEx2-GFP ind_vs_WT pEKEx2-GFP unind (18.11.2019) 1.414
set_ACh_WT pEKExL-cg0955-GFP ind_vs_WT pEKExL-cg0955 unind 30 min (18.11.2019)
1.414
set_LK_WT 5 mM KNO3_vs_WT (7.11.2019) 1.414
set_AH_Dcg2040_vs_WT (28.10.2019) 1.414
set_CS_WT 500 ug/mL Ethambutol_vs_WT 0 h (3.12.2019) 1.414

Conditions distance
set_MBa_Dcg0741_vs_WT (2.10.2019) 1.414
set_ACh_WT pEKEx2-TorA-cg2705_vs_WT pEKEx2 1 hour (18.11.2019) 1.414
set_MB_Dcg1083 Dcg1084_vs_WT (8.11.2019) 1.414
set_AKoK_Dcg1300 Dcg1301_vs_WT (5.11.2019) 1.414
set_AH_Dcg2200 Dcg2201 2.5 uM FeSO4_vs_WT 2.5 uM FeSo4 (20.9.2019) 1.414
set_ACh_WT pEKExL-cg3287-GFP ind_vs_WT pEKExL-cg3287-GFP unind (18.11.2019)
1.414
set_ACh_WT pEKExL-cg0955-GFP ind_vs_WT pEKExL-cg0955-GFP unind (18.11.2019)
1.414
set_ACh_WT pEKEx2-GFP ind_vs_WT pEKEx2-GFP unind (18.11.2019) 1.414
set_ACh_WT pEKExL-cg0955-GFP ind_vs_WT pEKExL-cg0955 unind 30 min (18.11.2019)
1.414

set_LK_WT 5 mM KNO3_vs_WT (7.11.2019) 1.414
set_AH_Dcg2040_vs_WT (28.10.2019) 1.414
set_CS_WT 500 ug/mL Ethambutol_vs_WT 0 h (3.12.2019) 1.414

Conditions distance
set_XS_WT w/o Cu_vs_WT OD 5 (14.11.2019) 1.414
set_MBa_Dcg0741_vs_WT (2.10.2019) 1.414
set_ACh_WT pEKEx2-TorA-cg2705_vs_WT pEKEx2 1 hour (18.11.2019) 1.414
set_MB_Dcg1083 Dcg1084_vs_WT (8.11.2019) 1.414
set_AKoK_Dcg1300 Dcg1301_vs_WT (5.11.2019) 1.414
set_AH_Dcg2200 Dcg2201 2.5 uM FeSO4_vs_WT 2.5 uM FeSo4 (20.9.2019) 1.414
set_ACh_WT pEKExL-cg3287-GFP ind_vs_WT pEKExL-cg3287-GFP unind (18.11.2019)
1.414
set_ACh_WT pEKExL-cg0955-GFP ind_vs_WT pEKExL-cg0955-GFP unind (18.11.2019)
1.414
set_ACh_WT pEKEx2-GFP ind_vs_WT pEKEx2-GFP unind (18.11.2019) 1.414
set_ACh_WT pEKExL-cg0955-GFP ind_vs_WT pEKExL-cg0955 unind 30 min (18.11.2019)
1.414
set_LK_WT 5 mM KNO3_vs_WT (7.11.2019) 1.414
set_AH_Dcg2040_vs_WT (28.10.2019) 1.414
set_CS_WT 500 ug/mL Ethambutol_vs_WT 0 h (3.12.2019) 1.414

Conditions distance
set_AFK_Dcg1689_vs_WT_CGXII (25.9.2019) 1.732
set_AH_Dcg1890_vs_WT (28.10.2019) 1.732
set_XS_WT w/o Cu_vs_WT OD 5 (14.11.2019) 1.732
set_MBa_Dcg0741_vs_WT (2.10.2019) 1.732
set_ACh_WT pEKEx2-TorA-cg2705_vs_WT pEKEx2 1 hour (18.11.2019) 1.732
set_MB_Dcg1083 Dcg1084_vs_WT (8.11.2019) 1.732
set_AKoK_Dcg1300 Dcg1301_vs_WT (5.11.2019) 1.732
set_AH_Dcg2200 Dcg2201 2.5 uM FeSO4_vs_WT 2.5 uM FeSo4 (20.9.2019) 1.732
set_ACh_WT pEKExL-cg3287-GFP ind_vs_WT pEKExL-cg3287-GFP unind (18.11.2019)
1.732
set_ACh_WT pEKExL-cg0955-GFP ind_vs_WT pEKExL-cg0955-GFP unind (18.11.2019)
1.732
set_ACh_WT pEKEx2-GFP ind_vs_WT pEKEx2-GFP unind (18.11.2019) 1.732
set_ACh_WT pEKExL-cg0955-GFP ind_vs_WT pEKExL-cg0955 unind 30 min (18.11.2019)
1.732
set_LK_WT 5 mM KNO3_vs_WT (7.11.2019) 1.732
set_AH_Dcg2040_vs_WT (28.10.2019) 1.732
set_CS_WT 500 ug/mL Ethambutol_vs_WT 0 h (3.12.2019) 1.732

Conditions distance
set_CL_Dcg0313_vs_WT 50 mM Isoleucin (2.12.2019) 2.000
set_JP_Dcgb_03605_vs_WT (16.10.2019) 2.000

Conditions distance
set_MH_Dcg3315_vs_WT (7.10.2019) 2.000

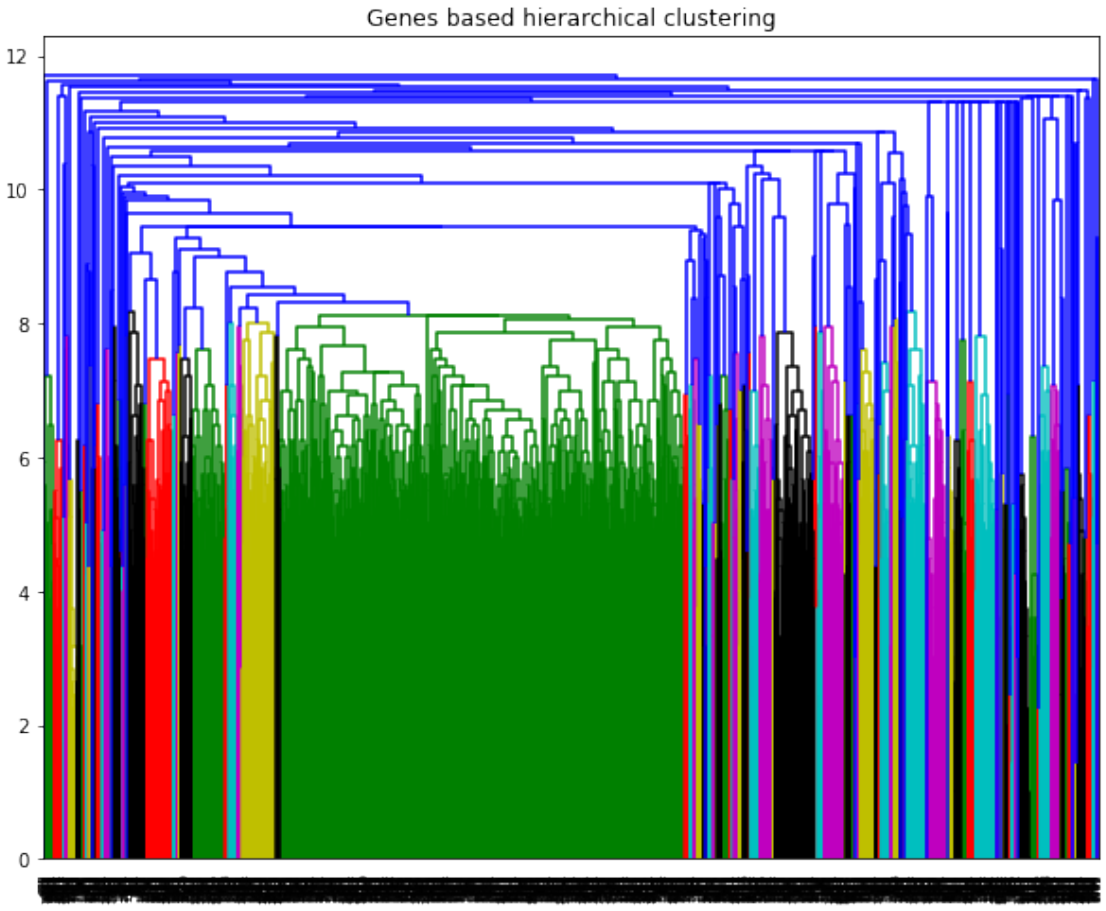
```
set_CL_Dcg0313_vs_WT 50 mM Isoleucin (2.12.2019)        2.000
set_JP_Dcgb_03605_vs_WT (16.10.2019)    2.000

Conditions      distance
set_AFK_Dcg1689_vs_WT_CGXII (25.9.2019) 2.000
set_AH_Dcg1890_vs_WT (28.10.2019)       2.000
set_XS_WT w/o Cu_vs_WT OD 5 (14.11.2019)        2.000
set_MBa_Dcg0741_vs_WT (2.10.2019)       2.000
set_ACh_WT pEKEx2-TorA-cg2705_vs_WT pEKEx2 1 hour (18.11.2019)  2.000
set_MB_Dcg1083 Dcg1084_vs_WT (8.11.2019)        2.000
set_AKoK_Dcg1300 Dcg1301_vs_WT (5.11.2019)      2.000
set_AH_Dcg2200 Dcg2201 2.5 uM FeSO4_vs_WT 2.5 uM FeSo4 (20.9.2019)      2.000
set_ACh_WT pEKExL-cg3287-GFP ind_vs_WT pEKExL-cg3287-GFP unind (18.11.2019)
2.000
set_ACh_WT pEKExL-cg0955-GFP ind_vs_WT pEKExL-cg0955-GFP unind (18.11.2019)
2.000
set_ACh_WT pEKEx2-GFP ind_vs_WT pEKEx2-GFP unind (18.11.2019)   2.000
set_ACh_WT pEKExL-cg0955-GFP ind_vs_WT pEKExL-cg0955 unind 30 min (18.11.2019)
2.000
set_LK_WT 5 mM KNO3_vs_WT (7.11.2019)   2.000
set_AH_Dcg2040_vs_WT (28.10.2019)       2.000
set_CS_WT 500 ug/mL Ethambutol_vs_WT 0 h (3.12.2019)    2.000
set_MH_Dcg3315_vs_WT (7.10.2019)        2.000
set_CL_Dcg0313_vs_WT 50 mM Isoleucin (2.12.2019)        2.000
set_JP_Dcgb_03605_vs_WT (16.10.2019)    2.000

Conditions      distance
set_CL_WT 100 mM Isoleucine_vs_WT (2.12.2019)   2.236
set_RM_Dcg2466 Dcg0791 100 mM HCO3-_vs_Dcg2466 Dcg0791_7 h (9.10.2019)  2.236

Conditions      distance
set_ACh_WT pEKEx2-TorA-cg2705_vs_WT pEKEx2 4 hours (18.11.2019) 2.236
set_MBa_Dcg3210_vs_WT (2.10.2019)       2.236
set_MiB_Dcg0764_vs_WT CGXII (12.11.2019)        2.236
```

0.10 Genes Hierarchical Clustering



0.11 Table of similar genes

Genes	distance
cg3237, cg2124, cg0970, cg1037	0.000
cg3396, cg3237, cg2124, cg0970, cg1037	0.000
cg2329, cg1848, cg1150, cg1426	0.000
cg2034, cg1907, cg1514, cg1905	0.000
cg2043, cg2034, cg1907, cg1514, cg1905	0.000
cg4005, cg2043, cg2034, cg1907, cg1514, cg1905	0.000
cg0204, cg0203, cg0199, cg0201	0.000
cg0223, cg0204, cg0203, cg0199, cg0201	0.000
cg1084, cg1083, cg1081, cg1082	0.000
cg2117, cg3396, cg3237, cg2124, cg0970, cg1037	1.000
cg1903, cg4005, cg2043, cg2034, cg1907, cg1514, cg1905	1.000
cg0205, cg0197, cg0202, cg0223, cg0204, cg0203, cg0199, cg0201	1.000
cg2630, cg2631, cg1226, cg2629	1.000
cg1705, cg2116, cg3285, cg3287	1.414

```
cg3138, cg3372, cg0842, cg3140  1.414
cg3360, cg3363, cg3361, cg3362  1.732
cg3216, cg1590, cg1643, cg2340  1.732
cg2298, cg2302, cg2299, cg2300  1.732
cg0254, cg2052, cg0751, cg1992  2.000
cg1266, cg2117, cg3396, cg3237, cg2124, cg0970, cg1037  2.000
cg2303, cg2304, cg2298, cg2302, cg2299, cg2300  2.000
cg0117, cg0283, cg1335, cg2691  2.236
cg1325, cg3321, cg3320, cg3322  2.236
cg0894, cg1266, cg2117, cg3396, cg3237, cg2124, cg0970, cg1037  2.236
cg1997, cg1903, cg4005, cg2043, cg2034, cg1907, cg1514, cg1905  2.236
cg3112, cg3138, cg3372, cg0842, cg3140  2.236
cg0119, cg1785, cg0115, cg0116  2.449
cg0894, cg1266, cg2117, cg3396, cg3237, cg2124, cg0970, cg1037, cg0172, cg1689
2.449
cg1322, cg2329, cg1848, cg1150, cg1426  2.449
cg1225, cg2630, cg2631, cg1226, cg2629  2.449
cg2132, cg1332, cg0149, cg0625  2.646
cg2232, cg3025, cg0628, cg2167  2.646
cg0878, cg3329, cg2349, cg2683  2.646
cg3359, cg3360, cg3363, cg3361, cg3362  2.646
cg3367, cg3368, cg3216, cg1590, cg1643, cg2340  2.646
cg2895, cg0894, cg1266, cg2117, cg3396, cg3237, cg2124, cg0970, cg1037, cg0172,
cg1689  2.646
cg1108, cg2132, cg1332, cg0149, cg0625  2.828
cg1325, cg3321, cg3320, cg3322, cg2811, cg2812  2.828
cg3227, cg2895, cg0894, cg1266, cg2117, cg3396, cg3237, cg2124, cg0970, cg1037,
cg0172, cg1689  2.828
cg1137, cg2100, cg0874, cg0282, cg0537  3.000
cg1878, cg0657, cg0674, cg2232, cg3025, cg0628, cg2167  3.000
cg2194, cg1482, cg3374, cg0796, cg3373  3.162
cg1392, cg0878, cg3329, cg2349, cg2683  3.162
cg3387, cg3385, cg3389, cg0211, cg3386  3.162
cg3227, cg2895, cg0894, cg1266, cg2117, cg3396, cg3237, cg2124, cg0970, cg1037,
cg0172, cg1689, cg1052, cg2925  3.162
cg0901, cg0117, cg0283, cg1335, cg2691  3.162
cg0206, cg0205, cg0197, cg0202, cg0223, cg0204, cg0203, cg0199, cg0201  3.162
cg3405, cg1705, cg2116, cg3285, cg3287  3.162
cg2750, cg1299, cg1300, cg1301  3.162
```

0.12 Further Plans:

- Iteratively adjust cluster parameters to achive better grouping.
- Apply more gentle methods like non-negative matrix factorization
- Optimize the parameters for data normalization