RNA-seq Report

Nathan Hughes (JIC)

May 22, 2019

Contents

1 Intro + example data **3**
1.1 Examples of data . . . 3
1.1.1 Normalised count data . . . 3
1.1.2 Expression difference hypothesis tests . . . 3
1.1.3 Data with extra descriptions . . . 3

2 Initial data checking **5**
2.1 Check sample counts . . . 5
2.2 Check for sample differences . . . 6

3 Analysis **7**
3.1 Comparing 05hr chitin to water treatments . . . 7
3.1.1 Clustermap of largest/smallest DE genes . . . 7
3.1.2 Boxplots of differential changes . . . 8
3.2 Comparing 6hr chitin to water treatments . . . 9
3.2.1 Clustermap of largest/smallest DE genes . . . 9
3.2.2 Boxplots of differential changes . . . 9
3.3 Comparing 05hr treatments to lym . . . 11
3.3.1 Clustermap of largest/smallest DE genes . . . 11
3.3.2 Boxplots of differential changes . . . 11
3.4 Comparing 6hr treatments to lym . . . 13
3.4.1 Clustermap of largest/smallest DE genes . . . 13
3.4.2 Boxplots of differential changes . . . 13
3.5 Comparing all treatments across time . . . 15
3.5.1 Clustermap of largest/smallest DE genes . . . 15
3.5.2 Boxplots of differential changes . . . 15
3.5.3 Lineplots of changes between samples for genes of interest . . . 15
3.5.4 Checking up and down data's largest . . . 18

1 Intro + example data

1.1 Examples of data

1.1.1 Normalised count data

	cer_c_05h_a37	cer_c_05h_b38	cer_c_05h_c39	cer_c_6h_a85
AT1G57560	6.66594	6.86016	6.71742	7.03654
AT2G03260	6.97298	7.08292	6.73242	6.89541
AT2G36355	7.53624	7.24664	7.22017	7.3226
AT1G11185	7.44287	7.37075	7.32766	6.51549
AT5G23030	6.06316	6.35515	6.21125	6.12698
AT5G16520	7.66076	7.63511	7.5679	7.40679
AT4G33150	9.03895	9.07082	9.05513	9.35499
AT4G14630	8.05476	8.24193	8.07485	7.42149
AT3G14560	7.08607	7.13215	7.1723	7.31465
AT1G18590	9.10513	8.96686	9.02159	9.35499
AT5G05365	8.09272	8.31546	7.93893	7.69591
AT5G11160	7.54439	7.81302	7.70158	7.61088
AT1G27420	7.12133	6.72751	6.53351	7.29047
AT5G18270	7.9986	7.83365	7.70791	7.24903
AT1G80440	6.8464	6.93494	6.83126	7.4786
AT1G07970	7.98129	8.0474	8.10328	8.20224
AT1G55450	10.1205	10.3351	10.2424	9.58924
AT5G24840	8.21536	8.27673	8.22432	8.19418
AT2G37460	8.14517	8.3438	8.19834	7.99419
AT2G07595	6.10621	6.02556	5.99113	5.97591

1.1.2 Expression difference hypothesis tests

gene	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	sample
AT5G39090	13.2487	0.9219	0.594288	1.55127	0.120838	0.335167	cer_w_05h
AT5G14380	8.87757	-1.17008	0.659602	-1.77391	0.0760777	0.257841	cer_w_6h
AT1G19730	119.519	-0.021061	0.203546	-0.10347	0.91759	0.968718	col_c_05h
AT1G05493	4.37256	0.191672	0.96016	0.199625	0.841774	0.933395	lym_c_05h
AT3G09032	22.053	0.399712	0.48984	0.816006	0.414497	0.676983	cer_c_05h
AT2G41375	45.7757	0.0301228	0.31699	0.0950276	0.924293	0.971808	cer_c_6h
AT4G36180	257.118	0.121669	0.189834	0.640924	0.521572	0.863141	col_w_6h
AT1G23210	5.9605	0.155948	0.988455	0.157769	0.874639	0.951012	col_c_05h
AT2G01750	254.113	-0.313262	0.140221	-2.23405	0.0254795	0.119981	col_c_05h
AT2G18780	17.807	0.149322	0.570296	0.261832	0.793451	0.905366	col_c_6h
AT1G17410	46.0602	-0.101394	0.267701	-0.378756	0.704869	0.858978	col_c_6h
AT4G39330	1441.72	-0.0939816	0.0989497	-0.949792	0.342218	0.863757	lym_w_05h
AT4G38130	1203.39	0.00369207	0.0794837	0.0464507	0.962951	0.998246	lym_w_05h
AT3G52190	273.452	0.55585	0.161152	3.44924	0.000562174	0.00753138	cer_w_6h
AT3G12270	323.137	0.377735	0.153265	2.46459	0.0137171	0.0821683	cer_w_05h
AT4G24805	199.389	0.6758	0.150516	4.4899	7.12563 (-06)	7.87341 (-05)	lym_c_05h
AT3G05430	32.4406	0.350837	0.362139	0.96879	0.33265	0.859876	lym_w_05h
AT3G01135	43.5689	-0.379412	0.268763	-1.4117	0.158039	0.421368	cer_c_6h
AT3G13404	6.12748	-0.193849	0.861789	-0.224938	0.822028	0.925835	cer_w_6h
AT2G24860	270.011	-0.174685	0.160757	-1.08664	0.277195	0.715961	col_w_6h

1.1.3 Data with extra descriptions

	incoming	converted	n_incoming	n_converted	name	description
0	AT2G47410	AT2G47410	1	1	AT2G47410	WD40/YVTN repeat-like-containing doma
1	AT1G16730	AT1G16730	2	1	UP6	F17F16.6 protein
2	AT2G17850	AT2G17850	3	1	AT2G17850	Rhodanese/Cell cycle control phosphatase
3	AT3G59410	AT3G59410	4	1	GCN2	Protein kinase family protein
4	AT1G02930	AT1G02930	5	1	GSTF6	Glutathione S-transferase F6
5	AT3G09970	AT3G09970	6	1	AT3G09970	Calcineurin-like metallo-phosphoesterase su
6	AT2G13840	AT2G13840	7	1	AT2G13840	Expressed protein
7	AT4G00355	AT4G00355	8	1	ATI2	ATG8-interacting protein 2
8	AT4G26380	AT4G26380	9	1	AT4G26380	Cysteine/Histidine-rich C1 domain family
9	AT5G01435	AT5G01435	10	1	AT5G01435	None
10	AT4G38825	AT4G38825	11	1	AT4G38825	SAUR-like auxin-responsive protein family
11	AT3G54220	AT3G54220	12	1	SCR	Protein SCARECROW
12	AT2G44020	AT2G44020	13	1	AT2G44020	Expressed protein
13	AT5G15170	AT5G15170	14	1	TDP1	Tyrosyl-DNA phosphodiesterase 1
14	AT2G36360	AT2G36360	15	1	AT2G36360	Galactose oxidase/kelch repeat superfamily
15	AT1G78780	AT1G78780	16	1	AT1G78780	pathogenesis-related family protein
16	AT4G23460	AT4G23460	17	1	BETAC-AD	Beta-adaptin-like protein C
17	AT4G00300	AT4G00300	18	1	AT4G00300	Receptor-like kinase
18	AT1G71130	AT1G71130	19	1	ERF070	Ethylene-responsive transcription factor ER
19	AT3G04000	AT3G04000	20	1	ChlADR2	NADPH-dependent aldehyde reductase 2, c

2 Initial data checking

2.1 Check sample counts

This shows that the normalisation of the count data has worked correctly, each sample is presented as having the same number of reads. This prevents different samples having different weights due to RNA-seq not producing uniform samples.

```
<Figure size 720x720 with 1 Axes>
```

Figure 1: Counting the number of reads found in each sample

2.2 Check for sample differences

Here, we do a naive check that there is variance within the samples; this matrix shows that a straight-forward euclidean distance of all counts in the samples are different. i.e. if there was a very small difference here it would be worrying and suggest that there isn't any significant changes. This figure is a simple data sanity check, **not of use for scientific purposes**.

NameErrorTraceback (most recent call last) <ipython-input-11-16720fa86bb7> in <module> 17 #legend_TN = [mpatches.Patch(color=c, label=l) for (list(set([c[:3] for c in collapsed_counts.columns]))] 18 —> 19 distances = pdist(collapsed_counts.T.values, metric='euclidean') 20 dist_matrix = squareform(distances) 21 dist_df = pd.DataFrame(dist_matrix, columns = collapsed_counts.columns, index=collapsed_counts.columns)

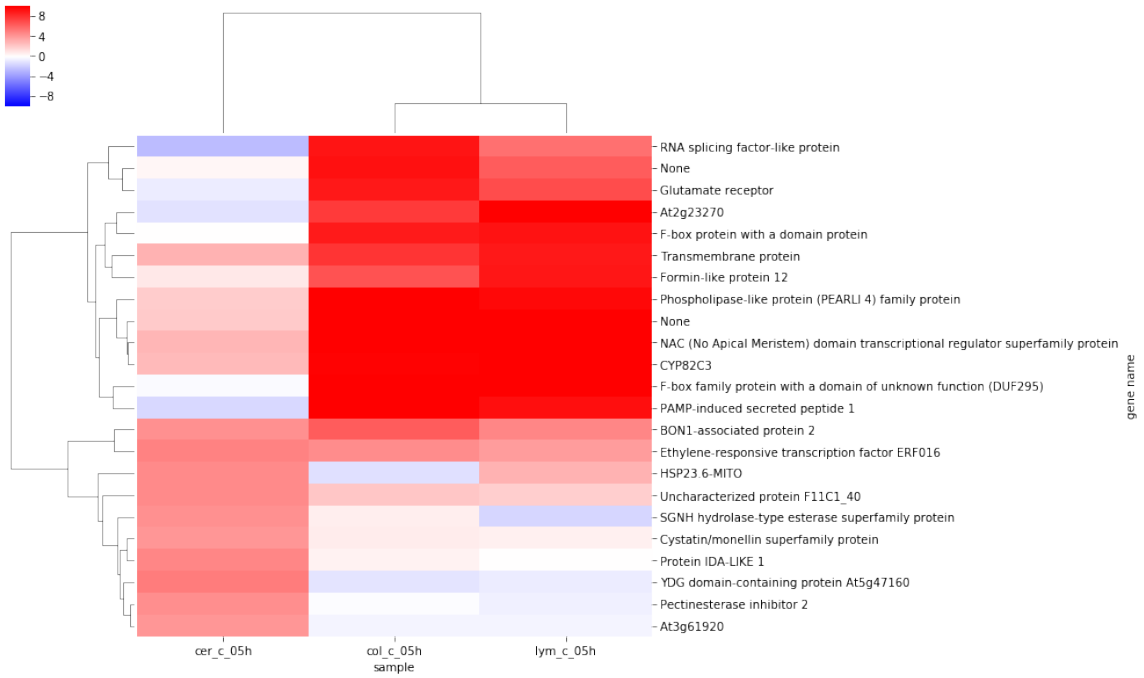
NameError: name 'pdist' is not defined

3 Analysis

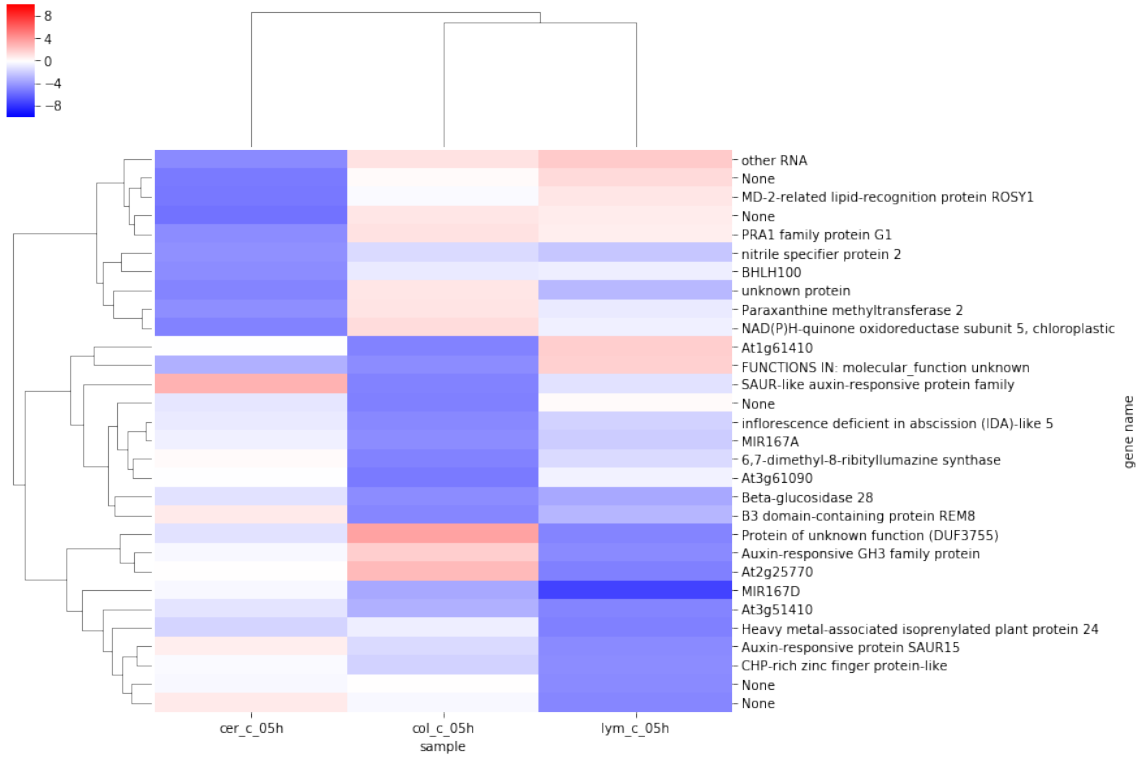
3.1 Comparing 05hr chitin to water treatments

3.1.1 Clustermap of largest/smallest DE genes

<Figure size 720x720 with 4 Axes>



<Figure size 720x720 with 4 Axes>



3.1.2 Boxplots of differential changes

```
<Figure size 1080x360 with 2 Axes>
```

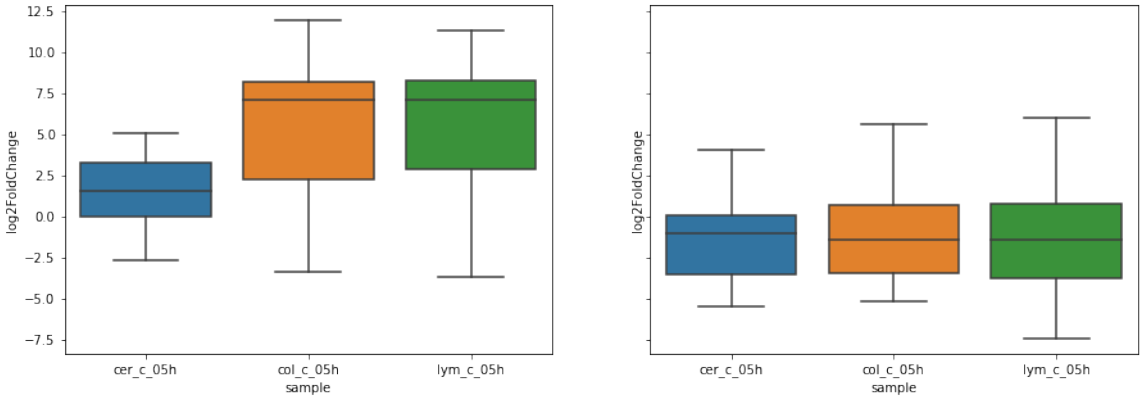


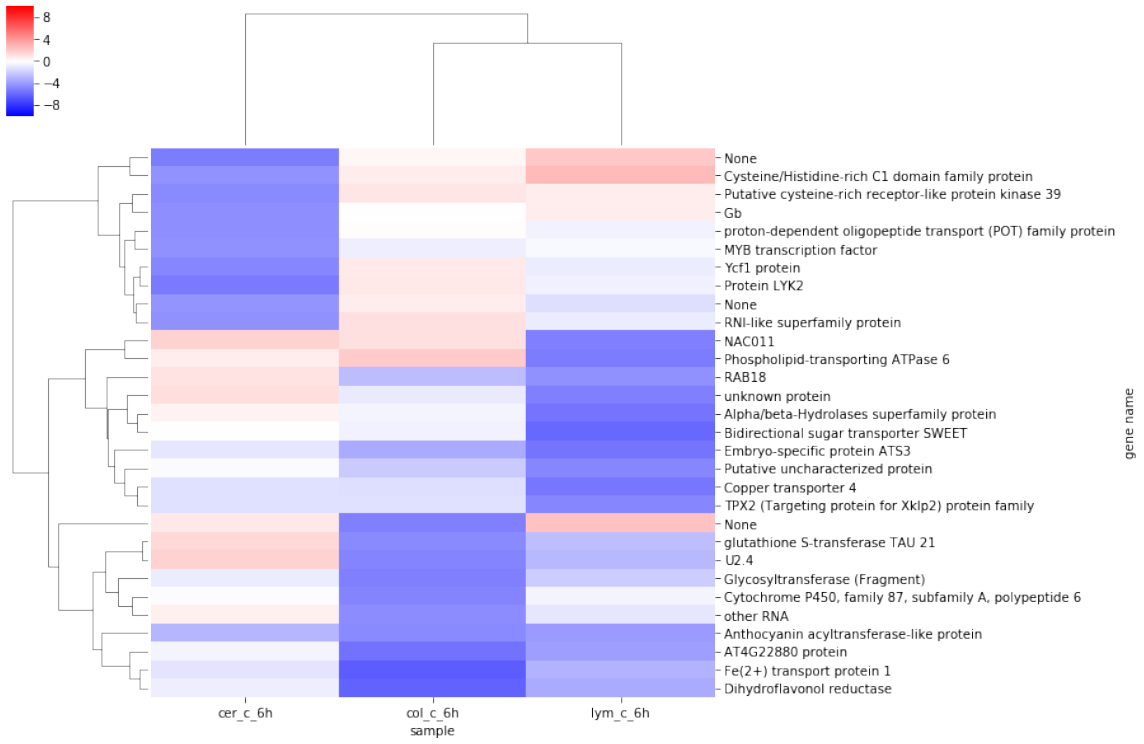
Figure 2: Boxplots of differential expressions from 50 largest (left) and 50 lowe05hst (right) DE genes

3.2 Comparing 6hr chitin to water treatments

3.2.1 Clustermap of largest/smallest DE genes

<Figure size 720x720 with 4 Axes>

<Figure size 720x720 with 4 Axes>



3.2.2 Boxplots of differential changes

<Figure size 1080x360 with 2 Axes>

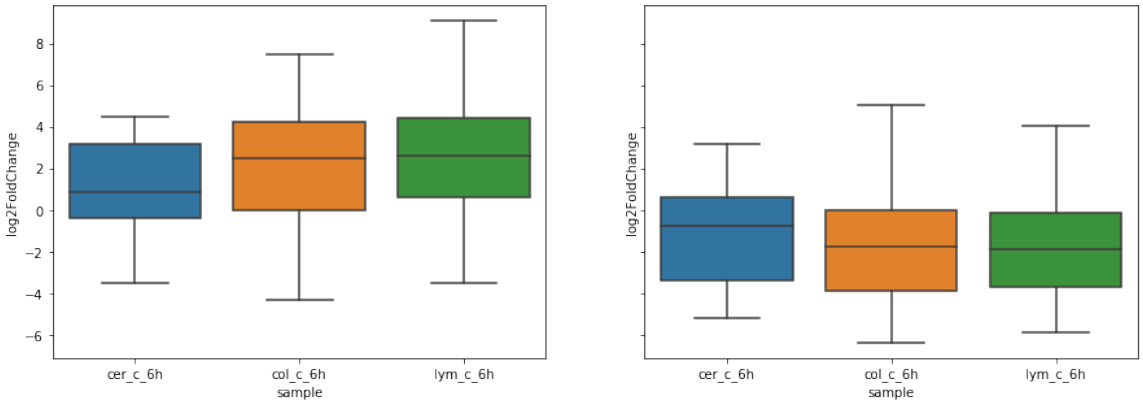
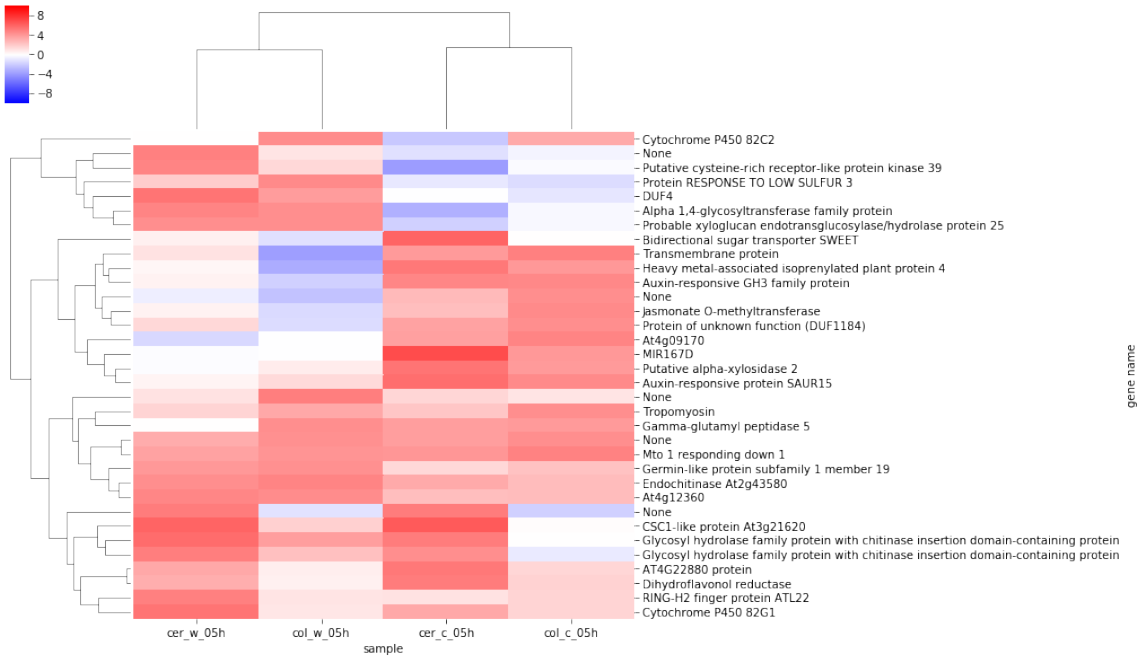


Figure 3: Boxplots of differential expressions from 50 largest (left) and 50 lowest (right) DE genes

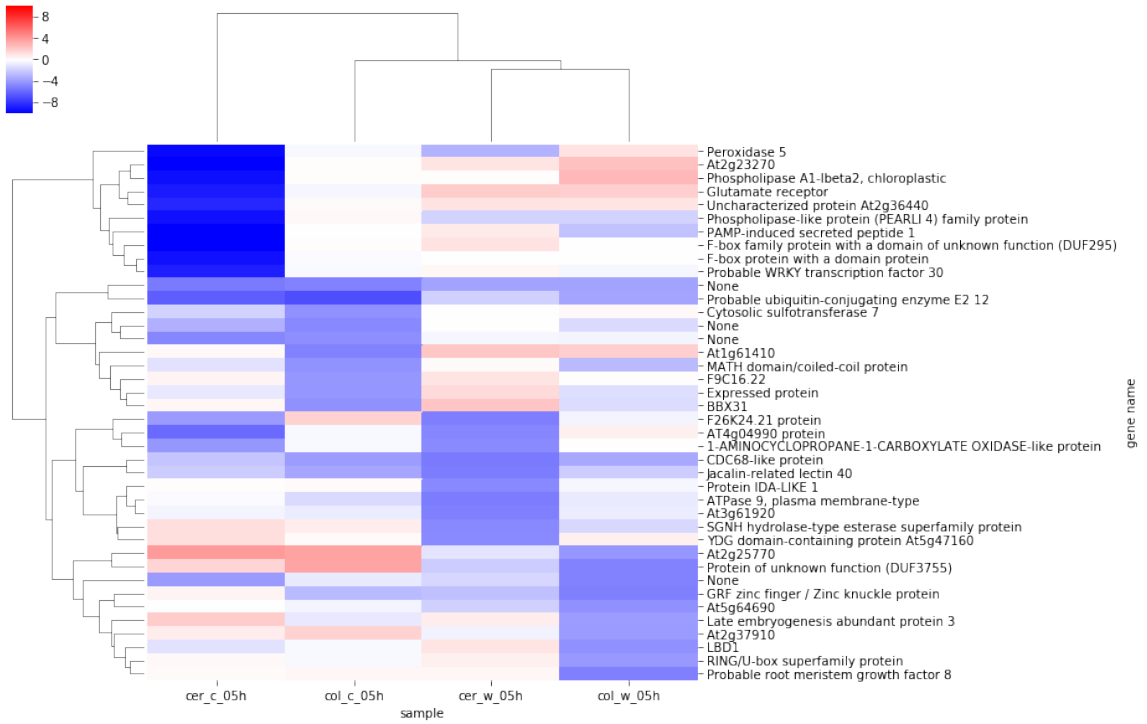
3.3 Comparing 05hr treatments to lym

3.3.1 Clustermap of largest/smallest DE genes

```
<Figure size 720x720 with 4 Axes>
```




```
<Figure size 720x720 with 4 Axes>
```



3.3.2 Boxplots of differential changes

```
<Figure size 1080x360 with 2 Axes>
```

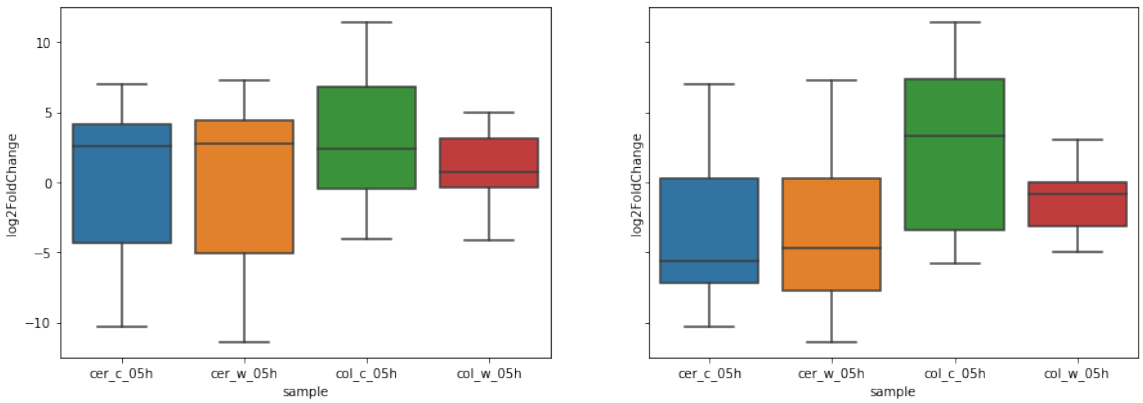
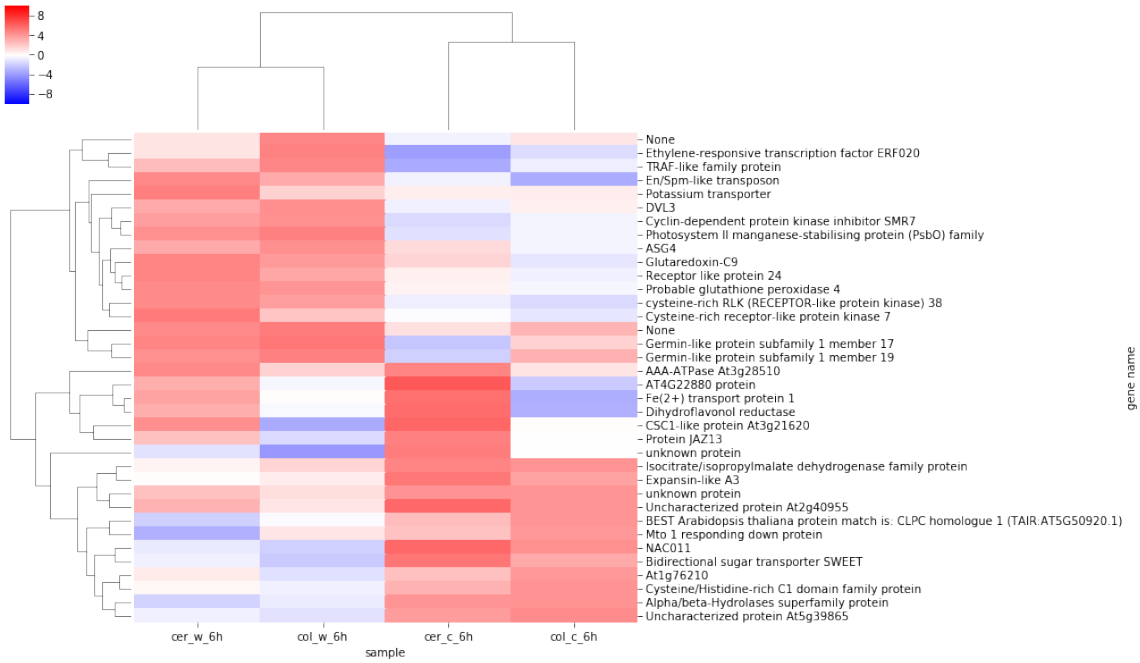


Figure 4: Boxplots of differential expressions from 50 largest (left) and 50 lowest (right) DE genes

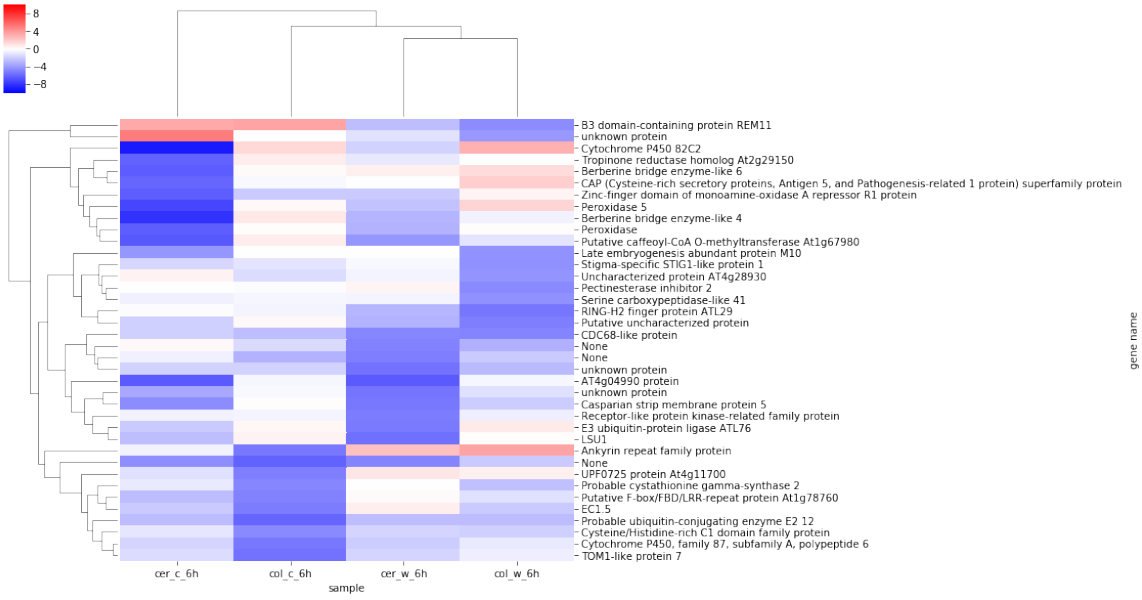
3.4 Comparing 6hr treatments to lym

3.4.1 Clustermap of largest/smallest DE genes

```
<Figure size 720x720 with 4 Axes>
```




```
<Figure size 720x720 with 4 Axes>
```



3.4.2 Boxplots of differential changes

```
<Figure size 1080x360 with 2 Axes>
```

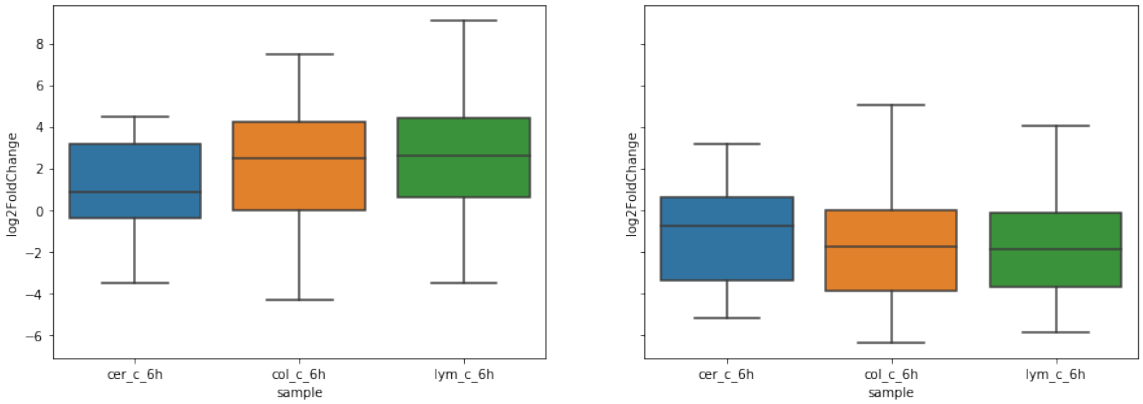
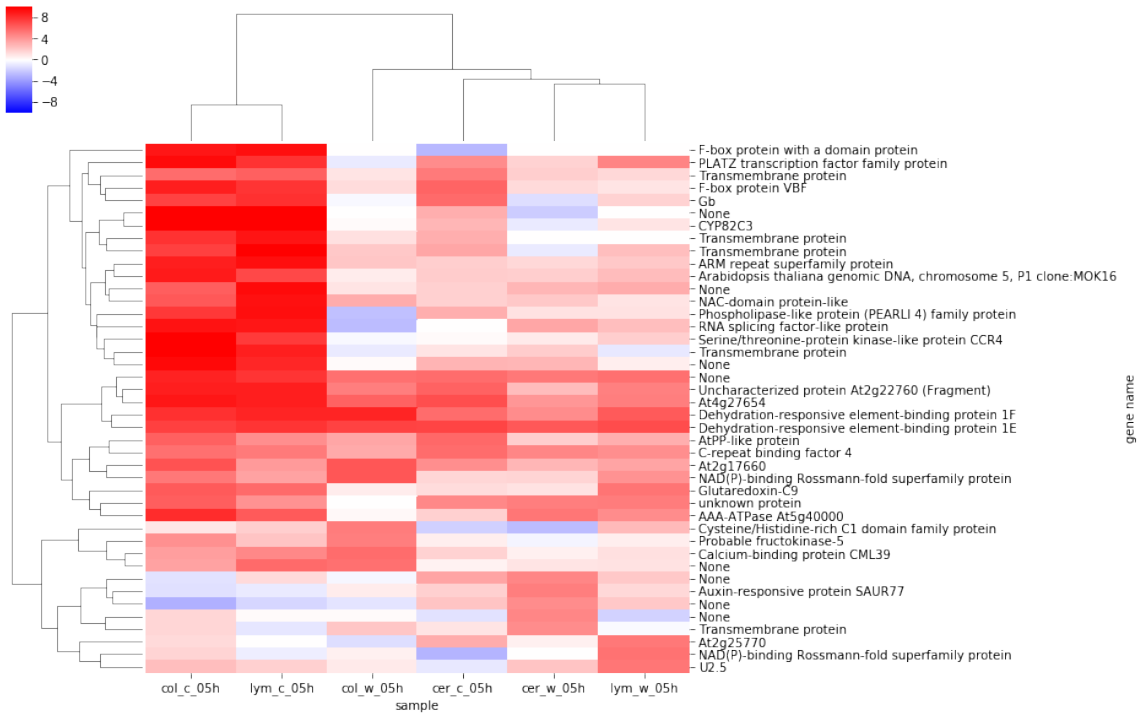


Figure 5: Boxplots of differential expressions from 50 largest (left) and 50 lowest (right) DE genes

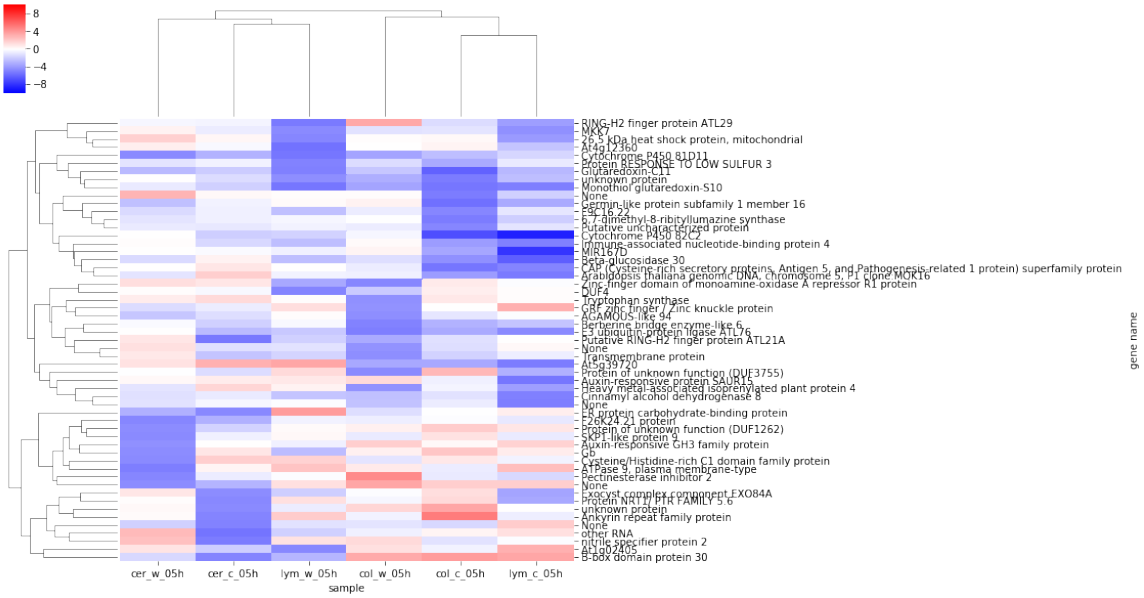
3.5 Comparing all treatments across time

3.5.1 Clustermap of largest/smallest DE genes

<Figure size 720x720 with 4 Axes>



<Figure size 720x720 with 4 Axes>



3.5.2 Boxplots of differential changes

<Figure size 1080x360 with 2 Axes>

3.5.3 Lineplots of changes between samples for genes of interest

<Figure size 1440x720 with 10 Axes>

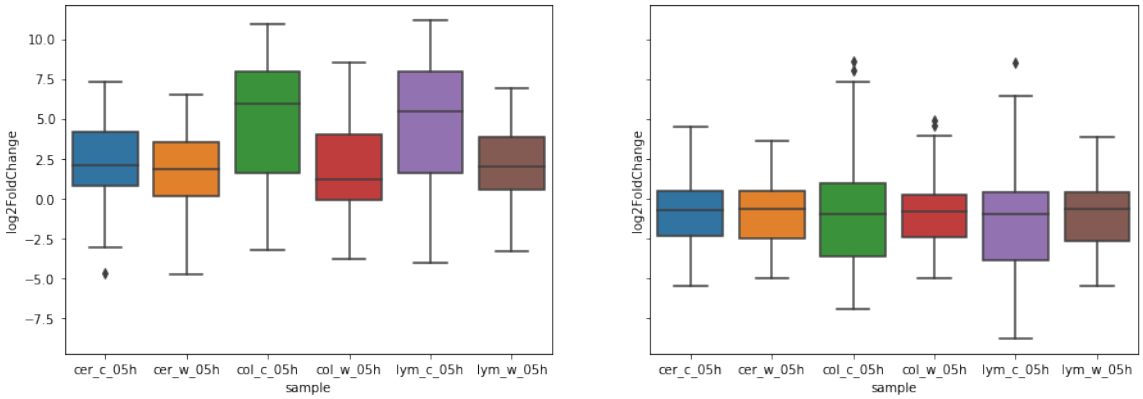
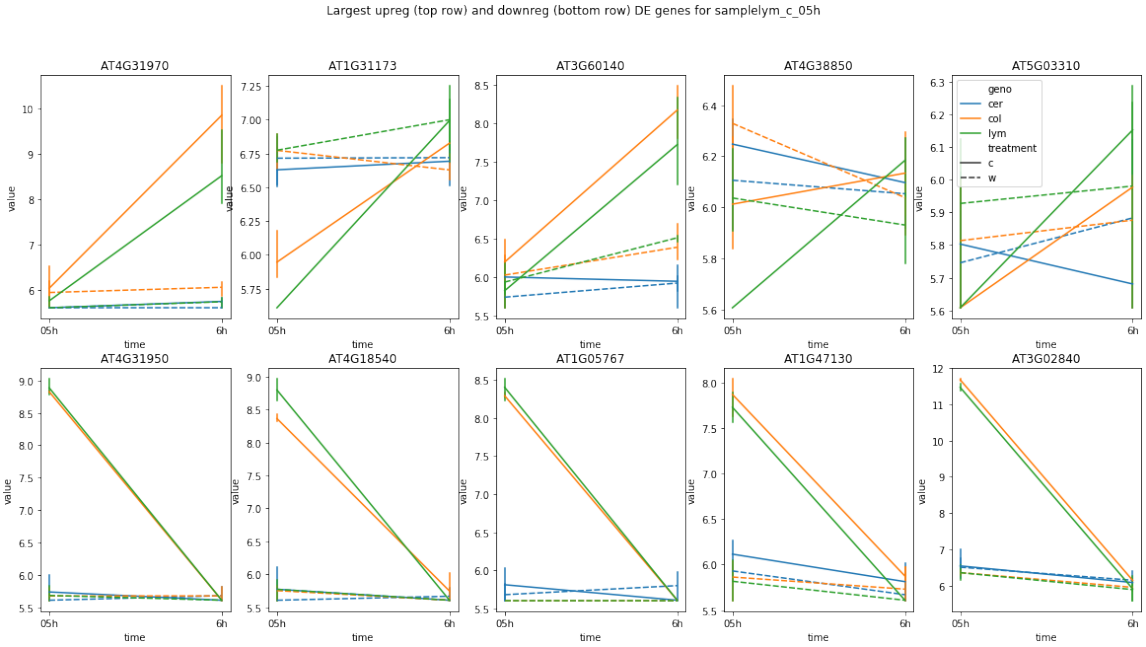
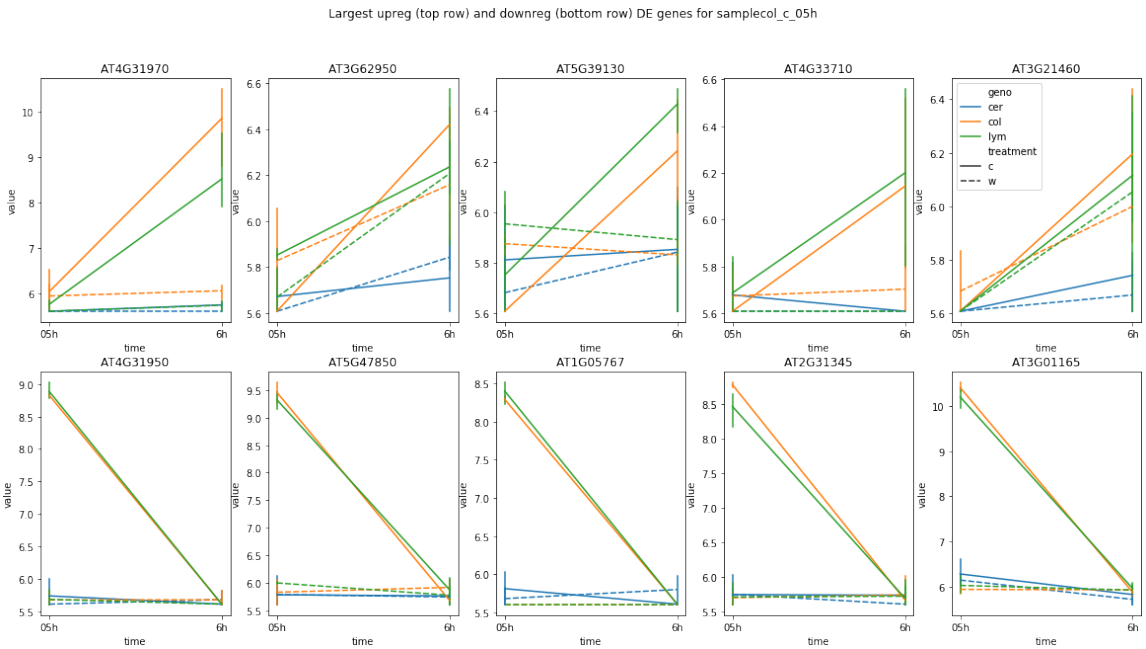


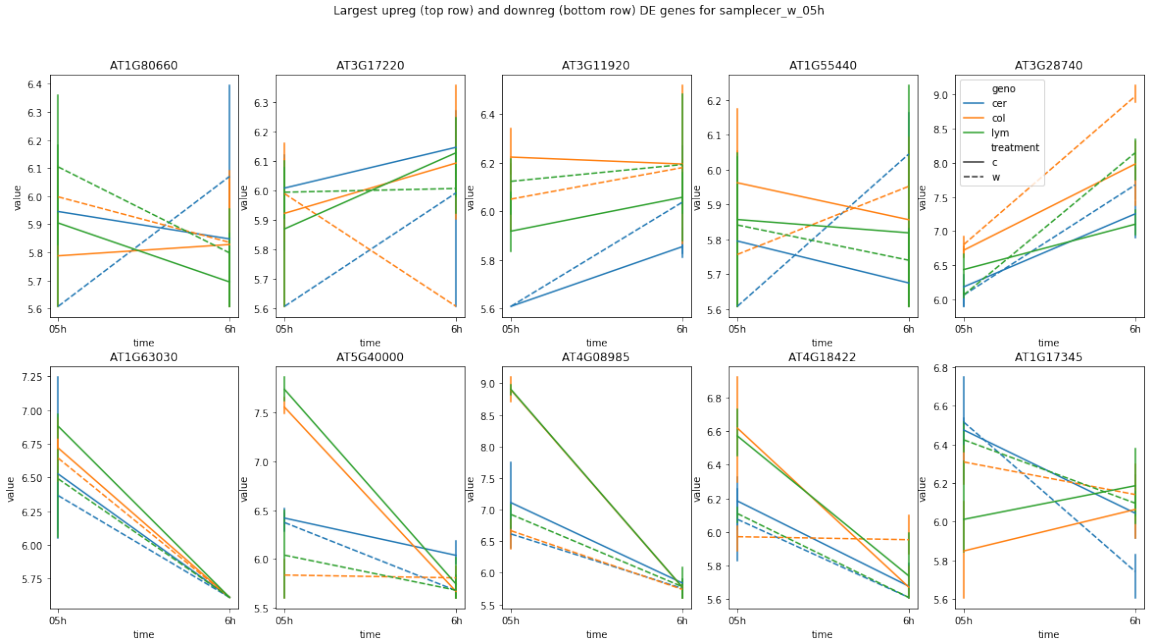
Figure 6: Boxplots of differential expressions from 50 largest (left) and 50 lowest (right) DE genes



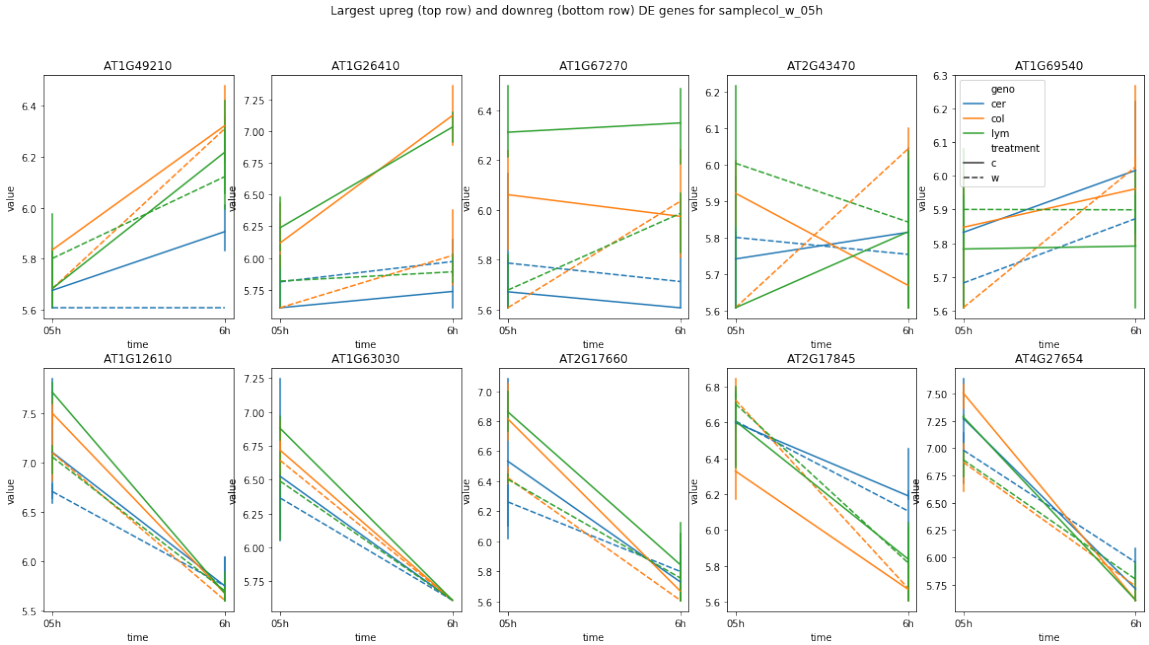
<Figure size 1440x720 with 10 Axes>



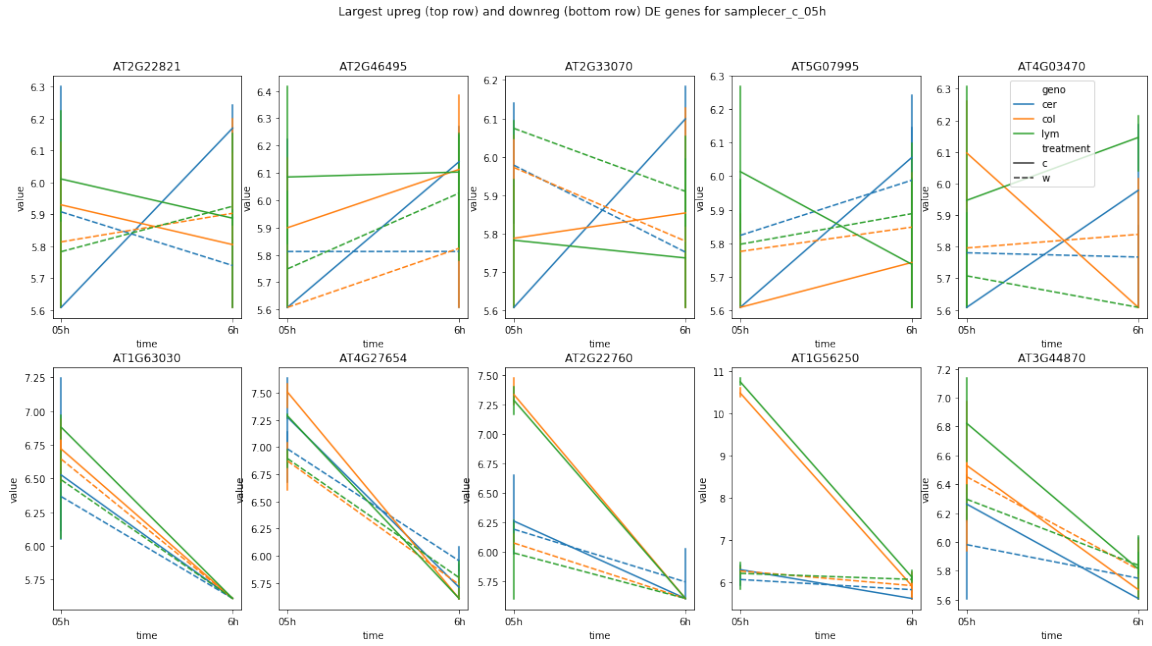
<Figure size 1440x720 with 10 Axes>



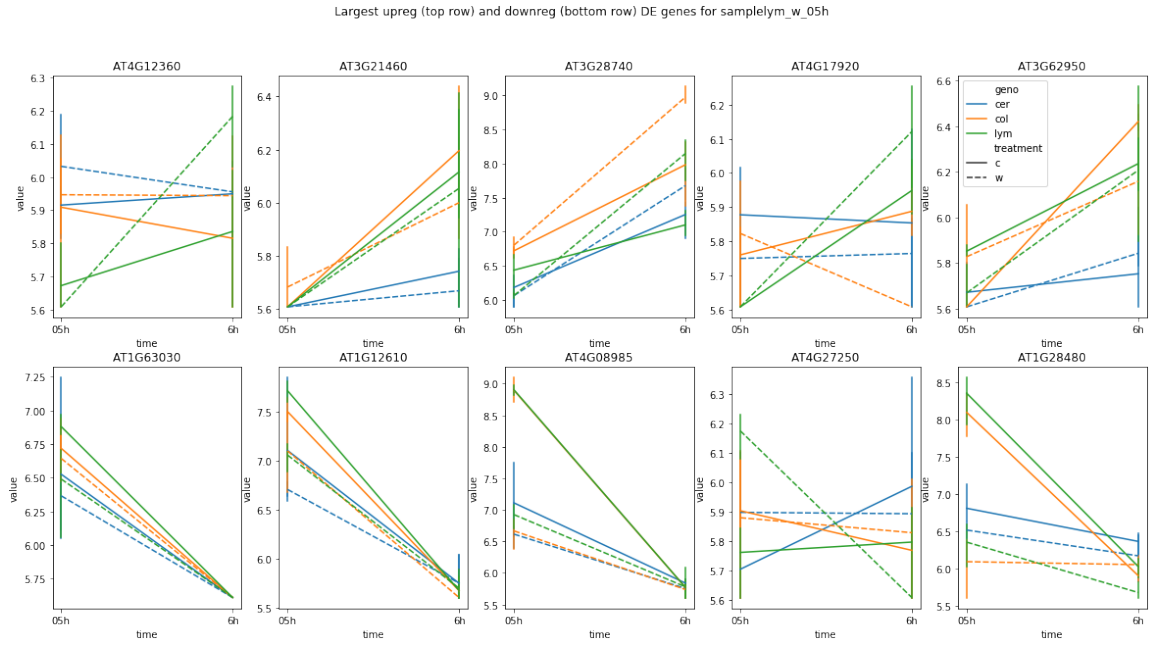
<Figure size 1440x720 with 10 Axes>



<Figure size 1440x720 with 10 Axes>



<Figure size 1440x720 with 10 Axes>



3.5.4 Checking up and down data's largest

	sample	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	lym__c__
AT4G31950	lym__c__05h	62.134	11.1474	1.19661	9.31583	1.21003 (-20)	3.04446 (-19)	
AT4G18540	lym__c__05h	49.8177	11.0383	2.57567	4.28561	1.82236 (-05)	0.000143129	
AT1G05767	lym__c__05h	39.595	10.5272	1.20735	8.7192	2.80169 (-18)	6.36071 (-17)	
AT1G47130	lym__c__05h	24.3826	9.55227	1.23862	7.71201	1.23856 (-14)	2.32962 (-13)	
AT3G02840	lym__c__05h	498.838	9.44554	0.726166	13.0074	1.11048 (-38)	4.82548 (-37)	
AT4G31970	lym__c__05h	107.475	-8.72719	1.60424	-5.44009	5.32543 (-08)	6.00415 (-07)	
AT1G31173	lym__c__05h	29.224	-7.94445	1.221	-6.50649	7.69248 (-11)	1.12916 (-09)	
AT3G60140	lym__c__05h	34.3907	-6.2782	1.08995	-5.76011	8.40611 (-09)	1.03509 (-07)	
AT4G38850	lym__c__05h	6.02148	-5.34424	1.36622	-3.91169	9.16526 (-05)	0.000633583	
AT5G03310	lym__c__05h	2.43405	-5.16479	1.7501	-2.95114	0.00316608	0.0146648	