

MainMethod

March 25, 2021

Taking the output from DEG-SEQ2, the data “5LS_L2L3Combined.csv” contains the 5 life stages we are interested in: Embryo, L1 larva, Dauer Larva, L2L3 Larva and Adult, lets take a peek of that data

```
[1]: import csv
import os

#user configurable variables
number_of_lines_to_print=10
expressionCountFile=os.path.join(os.getcwd(),'csvs/5LS_L2L3Combined.csv')
#Code Chunk for printing the file
with open(os.path.join(os.getcwd(),expressionCountFile)) as csv_file:
    csv_reader = csv.reader(csv_file, delimiter=',')
    for row in csv_reader:
        print(row)
        number_of_lines_to_print-=1
        if number_of_lines_to_print<=0:
            break
```

```
['WBID', 'elongating embryo Ce', 'L1 larva Ce', 'dauer larva Ce', 'adult Ce',
'L2L3_larva']
['WBGene000000001', '4208', '12140', '5547', '2246', '2369']
['WBGene000000002', '12554', '7828', '831', '280', '2591']
['WBGene000000003', '7180', '11253', '570', '212', '2466']
['WBGene000000004', '33305', '26947', '3212', '576', '5391']
['WBGene000000005', '595', '132', '37', '281', '1410']
['WBGene000000006', '425', '12243', '3146', '228', '2446']
['WBGene000000007', '36', '314', '129', '197', '1719']
['WBGene000000008', '0', '19', '663', '19', '182']
['WBGene000000009', '71', '416', '193', '20', '64']
```

Lets look at some statistics about the data:

```
[2]: import pandas as pd
import numpy as np

exp_data = pd.read_csv(expressionCountFile)
```

```
print(exp_data.describe())
```

	elongating embryo Ce	L1 larva Ce	dauer larva Ce	adult Ce \
count	20361.000000	2.036100e+04	2.036100e+04	2.036100e+04
mean	3692.118364	1.097686e+04	3.908223e+03	2.065875e+03
std	12796.637118	5.382926e+04	2.111976e+04	2.269343e+04
min	0.000000	0.000000e+00	0.000000e+00	0.000000e+00
25%	8.000000	5.700000e+01	2.700000e+01	5.000000e+00
50%	201.000000	7.240000e+02	4.030000e+02	8.600000e+01
75%	2730.000000	4.645000e+03	2.375000e+03	8.690000e+02
max	355180.000000	1.890193e+06	1.303599e+06	2.253663e+06

	L2L3_larva
count	2.036100e+04
mean	4.178275e+03
std	1.964655e+04
min	0.000000e+00
25%	2.900000e+01
50%	3.910000e+02
75%	2.292000e+03
max	1.103229e+06

Now, we need to determine the genes that we consider to be life stage biased, here are the some criterias that must be fullfilled to be considered a life stage biased gene:

This gene has the highest expression in that life stage

This gene's expression at this life stage has at least a fold difference of 2 comparing the max expression in other life stages

At least one life stage has a count that is higher than at least 10% of of counts across all life stages.
 *This ensures we dont include genes that have high fold diff due to unbalanced low expression counts, for example, a gene has a count of 1 in one life stage and are not found in other life stages(0 counts),this gene is a uniformly lowly expressed gene in all life stages, however, using the criteria one, this gene would have a fold difference of infinity, by setting a lower bound filter, we exclude these extremely lowly expressed gene counts that are prone to sequencing uncertainties.

Let's process the expression file using above criterias:

```
[3]: from Code import LifeStageBiased as LSB
      #Speficy input and output
      LSB.inputFile= expressionCountFile
      outputPath=os.path.join(os.getcwd(), 'csvs/LSB.csv')
      LSB.outputFile= outputPath
      LSB.cutLowPercentile=0.15
      LSB.foldDiff=2
      LSB.fixedCutValue=0 #This overrides the percentil cut value, set to 0 disables
      ↪ it
      LSB.main()
```

The cutOff Value for the specified percentaile is: 5.0

*In the data we are analyzing, there are very few genes that have observed expression only in one life stage, in which case the max expression for other life stages is 0, this will yield infinity for the fold diff value, in theory, these are “life stage specific genes” rather than “life stage biased genes”, however, since the same reason we mentioned above about the sequencing uncertainties, we do not believe that a gene with a few counts only in one life stage is more likely to be a life stage biased gene comparing to a gene with high expression in one life stage and very low expression in other life stages, see example below:

```
[4]: ls_data = pd.read_csv(outputFilePath)

print(ls_data.loc[ls_data['GeneID'] == "WBGene00015845"])
print()
print(ls_data.loc[ls_data['GeneID'] == "WBGene00000609"])
```

	GeneID	LS	LS_EXP	SecondMax	RestMean	FoldDiff
6043	WBGene00015845	adult Ce	7.0	0.0	0.0	7.0

	GeneID	LS	LS_EXP	SecondMax	RestMean	FoldDiff
218	WBGene00000609	adult Ce	71952.0	148.0	76.75	486.162162

As shown above, Gene “WBGene00015845” is a relatively lowly expressed gene that is only expressed in adult stage, and “WBGene00000609” has significantly higher expression in adult stage comparing to other life stages, which one are we more confident to select as the life stage biased gene?

There is no sure way to know, to compensate that, instead of putting infinity as the fold difference value for these “life stage specific” genes and put more confidence in them above all other genes, we decided to use their expression value as their foldDiff value, in which case a highly expressed “life stage specific” gene will be given higher confidence comparing to a lowly expressed one.

We are aware that this is perhaps not the best way of handling these genes, but luckily, there are only 15 such genes out of the 10099 life stage biased genes(0.15%) we selected using above filter, so it is extremely unlikely that different handlings of these genes will make a significant difference.

Now the genes that fit into our criteria should be in the *outputFilePath* we set ealier, lets take a look at some basic statistics of these selected life stage biased genes:

```
[14]: ls_data.head()
print(ls_data.columns)

ls_data[["LS_EXP", "SecondMax", "FoldDiff"]]=ls_data[["LS_EXP", "SecondMax", "FoldDiff"]].
    ↪apply(pd.to_numeric)
ls_data.sort_values(by=['FoldDiff'],ascending=False)

ls_data['LS_EXP_LOG']=np.log((ls_data['LS_EXP']))
ls_data['FoldDiff_LOG']=np.log((ls_data['FoldDiff']))
ls_data['RestMean_LOG']=np.log((ls_data['RestMean']))
```

```

print(ls_data.describe())
print()
print("ls_data Summary By Life Stage Group")
print(ls_data.groupby("LS").describe())

ax=ls_data['LS_EXP_LOG'].hist(by=ls_data['LS'],range=[0,12])

```

```

Index(['GeneID', 'LS', 'LS_EXP', 'SecondMax', 'RestMean', 'FoldDiff',
      'LS_EXP_LOG', 'FoldDiff_LOG'],
      dtype='object')

```

	LS_EXP	SecondMax	RestMean	FoldDiff	LS_EXP_LOG \
count	1.009900e+04	10099.000000	10099.000000	10099.000000	10099.000000
mean	1.679783e+04	4102.327656	2191.947742	15.490207	7.073785
std	7.351532e+04	19035.148534	10168.396260	176.069930	2.538669
min	5.000000e+00	0.000000	0.000000	2.000000	1.609438
25%	1.560000e+02	28.000000	12.250000	2.617371	5.049856
50%	1.298000e+03	248.000000	108.750000	3.776471	7.168580
75%	8.224500e+03	1946.000000	988.500000	6.722003	9.014873
max	2.253663e+06	540124.000000	315851.750000	12412.000000	14.628067

	FoldDiff_LOG	RestMean_LOG
count	10099.000000	1.009900e+04
mean	1.586584	-inf
std	0.907085	NaN
min	0.693147	-inf
25%	0.962170	2.505526e+00
50%	1.328790	4.689052e+00
75%	1.905386	6.896189e+00
max	9.426419	1.266303e+01

ls_data Summary By Life Stage Group

	LS_EXP					
	count	mean	std	min	25%	\
LS						
L1 larva Ce	5426.0	25057.749355	89726.448364	5.0	187.00	
L2L3_larva	1018.0	4984.386051	13889.489577	5.0	124.00	
adult Ce	695.0	13820.099281	111338.218970	5.0	131.00	
dauer larva Ce	1550.0	3776.957419	15036.754400	5.0	83.00	
elongating embryo Ce	1410.0	9322.402128	25125.246231	5.0	517.75	

				SecondMax		...	\
	50%	75%	max	count	mean	...	
LS							
L1 larva Ce	2159.5	14543.50	1890193.0	5426.0	6524.922226	...	
L2L3_larva	547.0	2903.50	173036.0	1018.0	1200.904715	...	
adult Ce	619.0	3951.00	2253663.0	695.0	842.952518	...	
dauer larva Ce	300.0	1844.00	235883.0	1550.0	691.027097	...	

```
elongating embryo Ce  2472.5   7250.25   328012.0    1410.0   2231.002837 ...
```

	FoldDiff_LOG		RestMean_LOG \			
	75%	max	count	mean	std	min
LS						
L1 larva Ce	1.807424	5.210326	5426.0	-inf	NaN	-inf
L2L3_larva	1.843245	5.637077	1018.0	-inf	NaN	-inf
adult Ce	1.925177	6.186542	695.0	-inf	NaN	-inf
dauer larva Ce	2.483414	9.426419	1550.0	-inf	NaN	-inf
elongating embryo Ce	1.930634	5.998820	1410.0	-inf	NaN	-inf

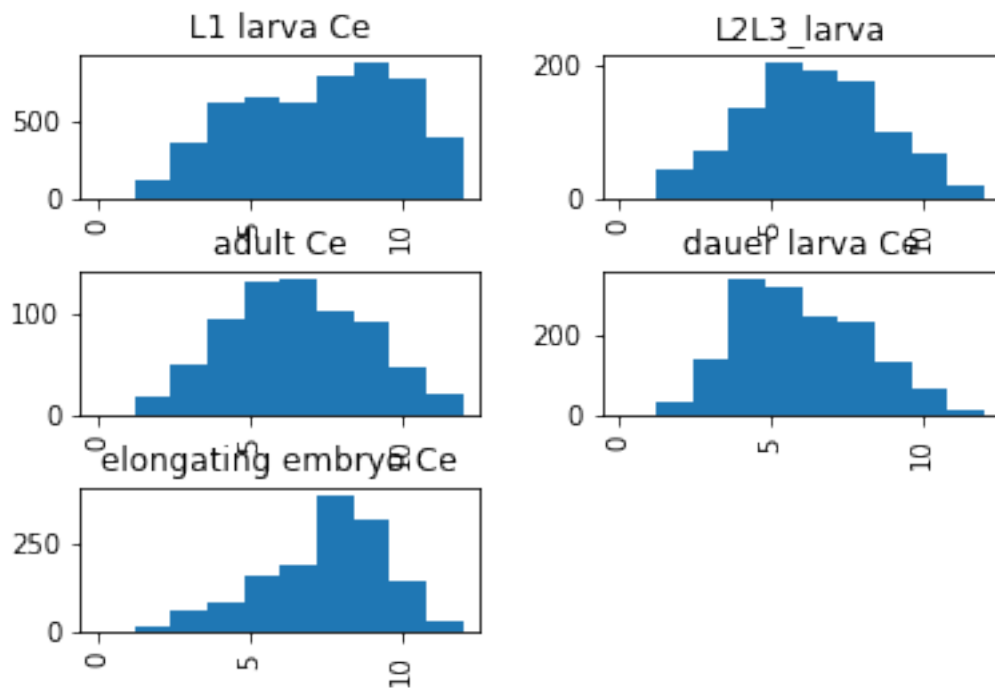
	25%	50%	75%	max
LS				
L1 larva Ce	2.791910	5.493061	7.677081	12.663028
L2L3_larva	2.490061	3.963188	5.468584	10.355311
adult Ce	2.611864	3.936716	5.523903	9.268963
dauer larva Ce	1.386294	2.803360	4.901099	10.365506
elongating embryo Ce	3.459854	5.441335	6.927496	10.556119

[5 rows x 56 columns]

/home/lu/.local/lib/python3.8/site-packages/pandas/core/series.py:726:

RuntimeWarning: divide by zero encountered in log

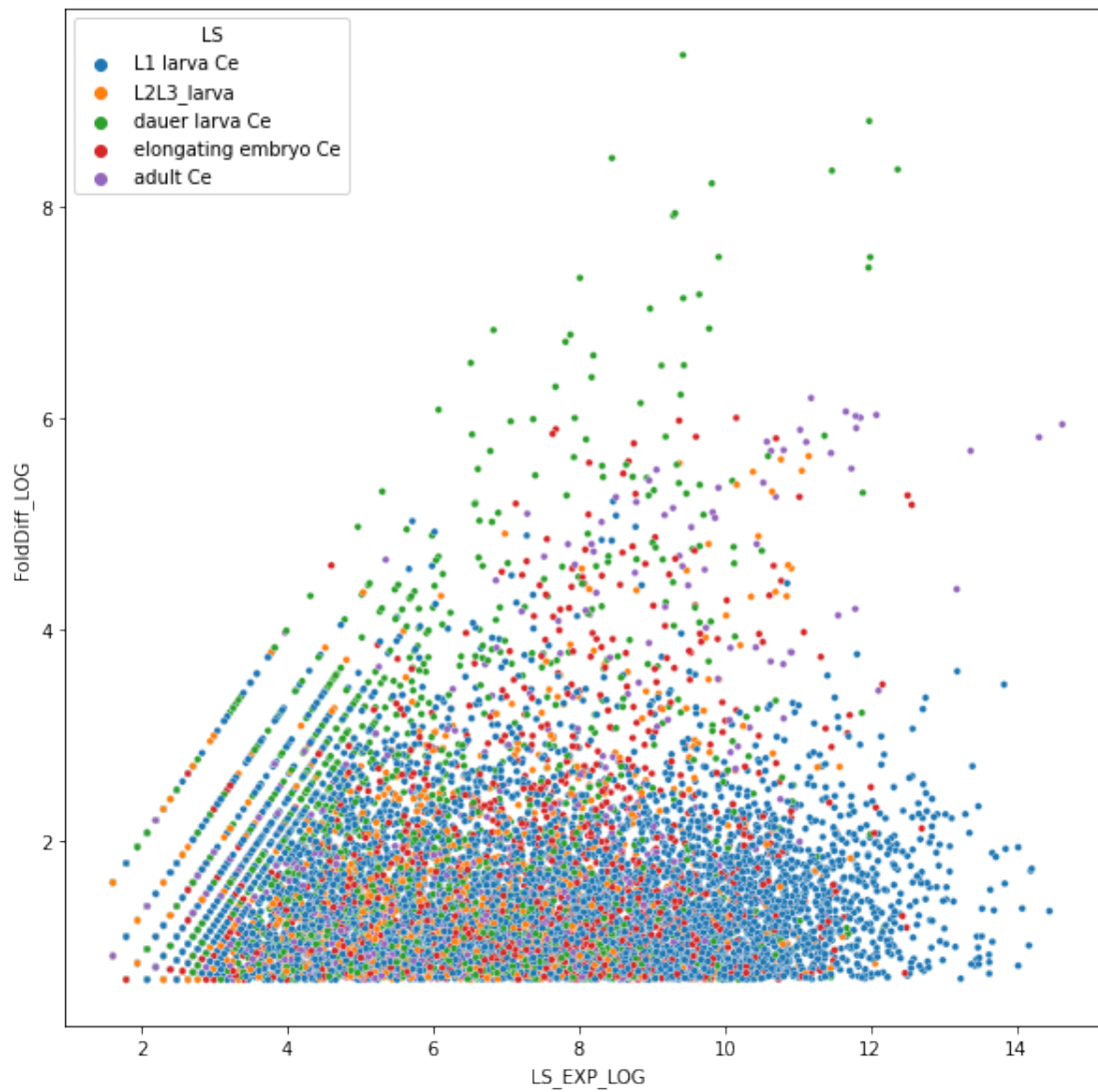
```
result = getattr(ufunc, method)(*inputs, **kwargs)
```



```
[6]: from matplotlib import pyplot as plt
import seaborn as sns
```

```
ax2=plt.figure(figsize=[10,10])
sns.scatterplot(x='LS_EXP_LOG',y='FoldDiff_LOG',hue='LS', data=ls_data,s=15)
```

```
[6]: <AxesSubplot:xlabel='LS_EXP_LOG', ylabel='FoldDiff_LOG'>
```

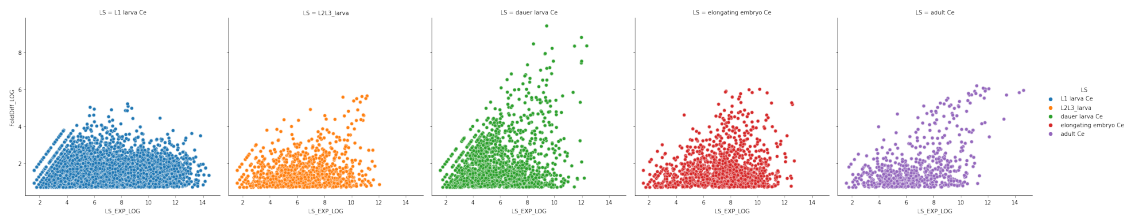


```
[7]: ax3=plt.figure(figsize=[20,20])
sns.relplot(
    data=ls_data,x='LS_EXP_LOG', y="FoldDiff_LOG",
```

```
col="LS", hue="LS",
kind="scatter"
)
```

```
[7]: <seaborn.axisgrid.FacetGrid at 0x7f7aecfcab20>
```

```
<Figure size 1440x1440 with 0 Axes>
```



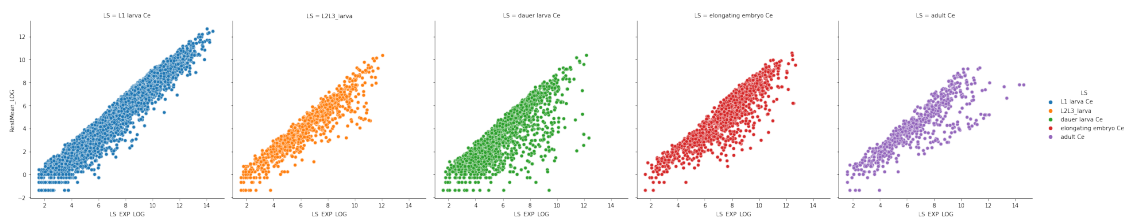
```
[ ]:
```

Look at the relationship between the max expression vs mean of expression in other life stages

```
[15]: ax4=plt.figure(figsize=[20,20])

sns.relplot(
    data=ls_data,x='LS_EXP_LOG', y="RestMean_LOG",
    col="LS", hue="LS",
    kind="scatter"
)
```

```
<Figure size 1440x1440 with 0 Axes>
```



Now, lets look at the number of genes from each life stage selected when we change the threshold:

```
[63]: sorted_ls_data=ls_data.sort_values(['LS','FoldDiff'],ascending=False)

thresholds=[2**i for i in range(1,11)]

for threshold in thresholds:
```

```

df_filtered=sorted_ls_data.loc[sorted_ls_data['FoldDiff'] >= threshold]
df_count=df_filtered.groupby("LS").count()
ax=plt.figure(figsize=[8,6])
text=("Threshold of FoldDiff: "+ str(threshold))
sns.histplot(df_filtered, x="LS",hue="LS").set_title(text)
print(text)
print(df_filtered.describe())

```

Threshold of FoldDiff: 2

	LS_EXP	SecondMax	RestMean	FoldDiff	LS_EXP_LOG \
count	1.009900e+04	10099.000000	10099.000000	10099.000000	10099.000000
mean	1.679783e+04	4102.327656	2191.947742	15.490207	7.073785
std	7.351532e+04	19035.148534	10168.396260	176.069930	2.538669
min	5.000000e+00	0.000000	0.000000	2.000000	1.609438
25%	1.560000e+02	28.000000	12.250000	2.617371	5.049856
50%	1.298000e+03	248.000000	108.750000	3.776471	7.168580
75%	8.224500e+03	1946.000000	988.500000	6.722003	9.014873
max	2.253663e+06	540124.000000	315851.750000	12412.000000	14.628067

	FoldDiff_LOG	RestMean_LOG
count	10099.000000	1.009900e+04
mean	1.586584	-inf
std	0.907085	NaN
min	0.693147	-inf
25%	0.962170	2.505526e+00
50%	1.328790	4.689052e+00
75%	1.905386	6.896189e+00
max	9.426419	1.266303e+01

Threshold of FoldDiff: 4

	LS_EXP	SecondMax	RestMean	FoldDiff	LS_EXP_LOG \
count	4.736000e+03	4736.000000	4736.000000	4736.000000	4736.000000
mean	1.874109e+04	2476.493454	1307.303262	29.896106	7.027185
std	8.095115e+04	10934.194589	5513.457001	256.362229	2.558391
min	5.000000e+00	0.000000	0.000000	4.000000	1.609438
25%	1.450000e+02	15.000000	6.500000	5.108633	4.976734
50%	1.014000e+03	95.000000	41.625000	7.072728	6.921658
75%	7.755500e+03	829.250000	387.875000	12.500000	8.956157
max	2.253663e+06	261525.000000	118162.250000	12412.000000	14.628067

	FoldDiff_LOG	RestMean_LOG
count	4736.000000	4736.000000
mean	2.253014	-inf
std	0.934246	NaN
min	1.386294	-inf
25%	1.630932	1.871802
50%	1.956246	3.728696

75% 2.525729 5.960682
max 9.426419 11.679814

Threshold of FoldDiff: 8

	LS_EXP	SecondMax	RestMean	FoldDiff	LS_EXP_LOG \
count	2.065000e+03	2065.000000	2065.000000	2065.000000	2065.000000
mean	1.737960e+04	998.086199	542.603995	61.463707	6.996054
std	8.397990e+04	4394.859318	2509.116448	386.007561	2.438047
min	8.000000e+00	0.000000	0.000000	8.000000	2.079442
25%	1.510000e+02	8.000000	3.750000	10.000000	5.017280
50%	9.560000e+02	40.000000	19.500000	14.112252	6.862758
75%	6.565000e+03	291.000000	129.250000	27.111111	8.789508
max	2.253663e+06	69368.000000	34733.500000	12412.000000	14.628067

	FoldDiff_LOG	RestMean_LOG
count	2065.000000	2065.000000
mean	2.989515	-inf
std	0.995191	NaN
min	2.079442	-inf
25%	2.302585	1.321756
50%	2.647043	2.970414
75%	3.299944	4.861749
max	9.426419	10.455460

Threshold of FoldDiff: 16

	LS_EXP	SecondMax	RestMean	FoldDiff	LS_EXP_LOG \
count	8.960000e+02	896.000000	896.000000	896.000000	896.000000
mean	2.031709e+04	389.906250	201.989397	127.457483	7.345930
std	1.092277e+05	1657.088913	1000.941265	579.578809	2.296912
min	1.600000e+01	0.000000	0.000000	16.000000	2.772589
25%	2.440000e+02	6.000000	2.750000	21.266304	5.497168
50%	1.468000e+03	28.500000	13.250000	30.992188	7.291656
75%	8.493000e+03	143.250000	67.312500	68.844626	9.046993
max	2.253663e+06	31253.000000	21759.250000	12412.000000	14.628067

	FoldDiff_LOG	RestMean_LOG
count	896.000000	896.000000
mean	3.801458	-inf
std	1.033031	NaN
min	2.772589	-inf
25%	3.057124	1.011601
50%	3.433735	2.583998
75%	4.231837	4.209345
max	9.426419	9.987794

Threshold of FoldDiff: 32

	LS_EXP	SecondMax	RestMean	FoldDiff	LS_EXP_LOG \
count	4.350000e+02	435.000000	435.000000	435.000000	435.000000
mean	3.073044e+04	296.452874	151.092529	238.972031	8.016322
std	1.519043e+05	1758.203487	1122.729190	817.602619	2.146121
min	3.300000e+01	0.000000	0.000000	32.125000	3.496508

25%	6.830000e+02	6.500000	3.000000	44.006400	6.526456
50%	3.270000e+03	31.000000	14.250000	74.452830	8.092545
75%	1.319700e+04	123.500000	55.875000	161.126961	9.487694
max	2.253663e+06	31253.000000	21759.250000	12412.000000	14.628067

	FoldDiff_LOG	RestMean_LOG
count	435.000000	435.000000
mean	4.564515	-inf
std	1.011964	NaN
min	3.469635	-inf
25%	3.784335	1.098612
50%	4.310166	2.656757
75%	5.082192	4.023095
max	9.426419	9.987794

Threshold of FoldDiff: 64

	LS_EXP	SecondMax	RestMean	FoldDiff	LS_EXP_LOG \
count	2.430000e+02	243.000000	243.000000	243.000000	243.000000
mean	4.192780e+04	184.621399	80.042181	392.949462	8.639722
std	1.882614e+05	686.621839	282.602225	1069.972609	1.927644
min	7.500000e+01	0.000000	0.000000	64.000000	4.317488
25%	1.528500e+03	7.000000	3.500000	90.302020	7.331232
50%	5.698000e+03	32.000000	16.250000	135.000000	8.647871
75%	1.771950e+04	121.000000	52.875000	275.230000	9.782410
max	2.253663e+06	6635.000000	2439.000000	12412.000000	14.628067

	FoldDiff_LOG	RestMean_LOG
count	243.000000	243.000000
mean	5.192860	-inf
std	0.954800	NaN
min	4.158883	-inf
25%	4.503147	1.252763
50%	4.905275	2.788093
75%	5.617553	3.967794
max	9.426419	7.799343

Threshold of FoldDiff: 128

	LS_EXP	SecondMax	RestMean	FoldDiff	LS_EXP_LOG \
count	1.280000e+02	128.000000	128.000000	128.000000	128.000000
mean	6.679294e+04	198.757812	84.787109	665.200762	9.205471
std	2.526545e+05	719.674442	312.341413	1422.451233	1.893452
min	1.440000e+02	0.000000	0.000000	128.200000	4.969813
25%	2.808000e+03	6.750000	3.250000	190.966121	7.940212
50%	9.453000e+03	26.500000	13.625000	263.228205	9.153855
75%	3.939975e+04	121.250000	51.312500	405.291451	10.581488
max	2.253663e+06	5944.000000	2439.000000	12412.000000	14.628067

	FoldDiff_LOG	RestMean_LOG
count	128.000000	128.000000
mean	5.833815	-inf

std	0.909382	NaN
min	4.853592	-inf
25%	5.252095	1.178655
50%	5.573018	2.611527
75%	6.004585	3.937566
max	9.426419	7.799343

Threshold of FoldDiff: 256

	LS_EXP	SecondMax	RestMean	FoldDiff	LS_EXP_LOG \
count	6.800000e+01	68.000000	68.000000	68.000000	68.000000
mean	1.041002e+05	262.176471	116.525735	1089.632088	9.667749
std	3.396591e+05	948.722172	418.548195	1855.841604	1.882576
min	4.350000e+02	0.000000	0.000000	256.062500	6.075346
25%	3.520250e+03	5.750000	2.500000	321.370004	8.166179
50%	1.222350e+04	21.000000	8.500000	402.690848	9.410997
75%	6.791150e+04	141.250000	59.562500	894.000000	11.125859
max	2.253663e+06	5944.000000	2439.000000	12412.000000	14.628067

	FoldDiff_LOG	RestMean_LOG
count	68.000000	68.000000
mean	6.395987	-inf
std	0.921161	NaN
min	5.545422	-inf
25%	5.772592	0.916291
50%	5.998169	2.140066
75%	6.795521	4.086894
max	9.426419	7.799343

Threshold of FoldDiff: 512

	LS_EXP	SecondMax	RestMean	FoldDiff	LS_EXP_LOG \
count	25.000000	25.000000	25.000000	25.000000	25.000000
mean	39358.52000	16.680000	6.890000	2371.482954	9.341956
std	66170.68322	25.450475	9.392783	2626.937144	1.612699
min	677.00000	0.000000	0.000000	540.750000	6.517671
25%	3550.00000	3.000000	1.250000	826.666667	8.174703
50%	11140.00000	6.000000	3.250000	1293.500000	9.318298
75%	18451.00000	19.000000	8.000000	2785.000000	9.822874
max	235883.00000	95.000000	33.250000	12412.000000	12.371091

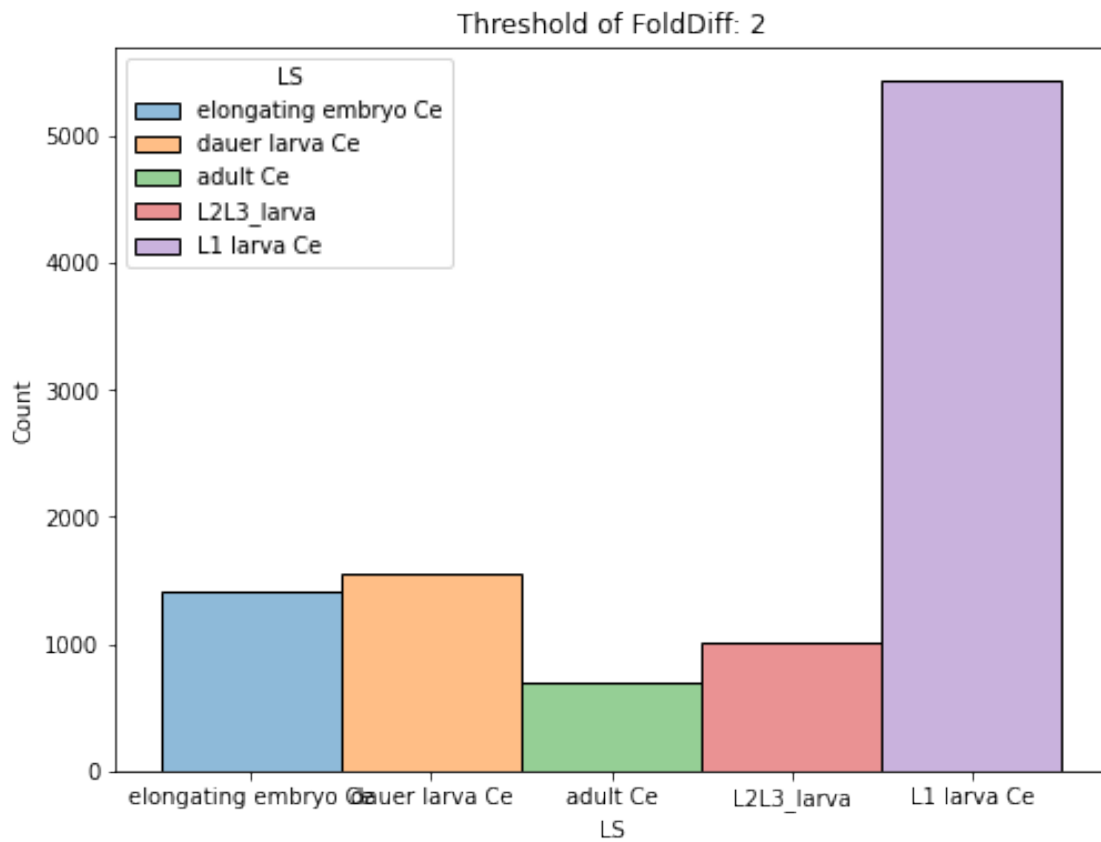
	FoldDiff_LOG	RestMean_LOG
count	25.000000	25.000000
mean	7.378386	-inf
std	0.848040	NaN
min	6.292957	-inf
25%	6.717402	0.223144
50%	7.165107	1.178655
75%	7.932003	2.079442
max	9.426419	3.504055

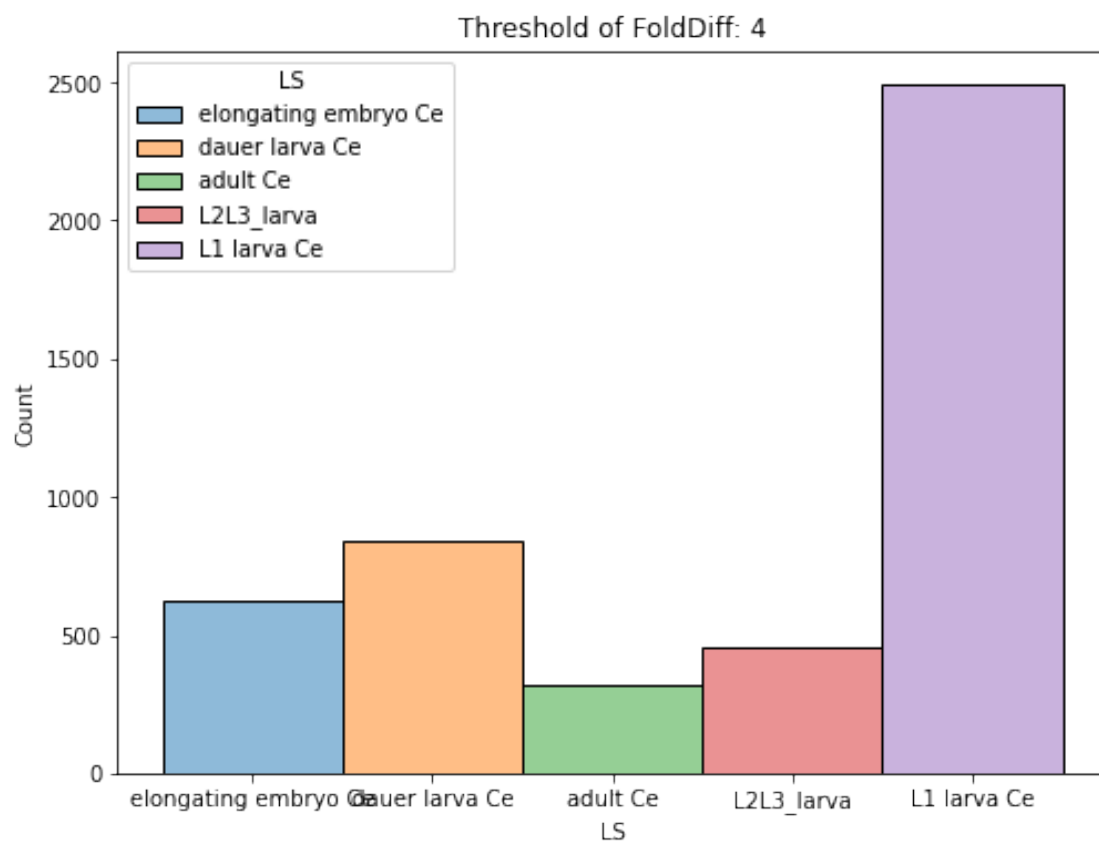
Threshold of FoldDiff: 1024

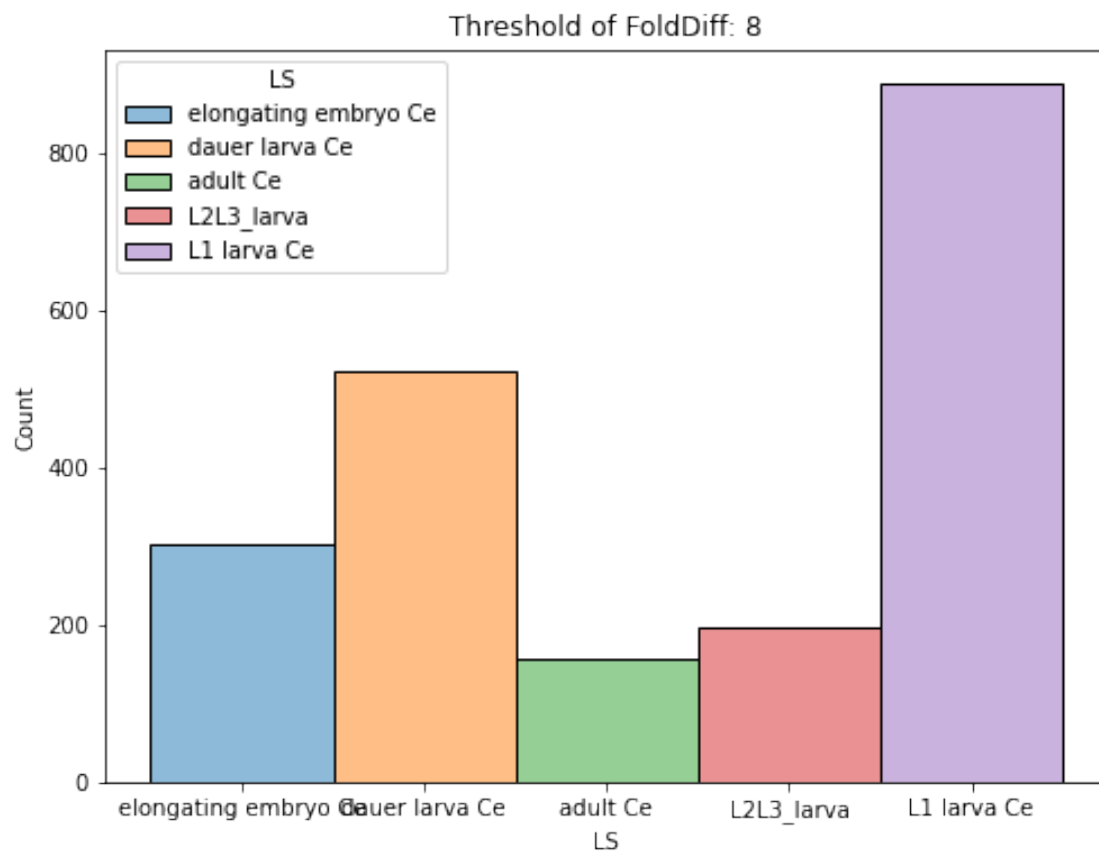
	LS_EXP	SecondMax	RestMean	FoldDiff	LS_EXP_LOG \
--	--------	-----------	----------	----------	--------------

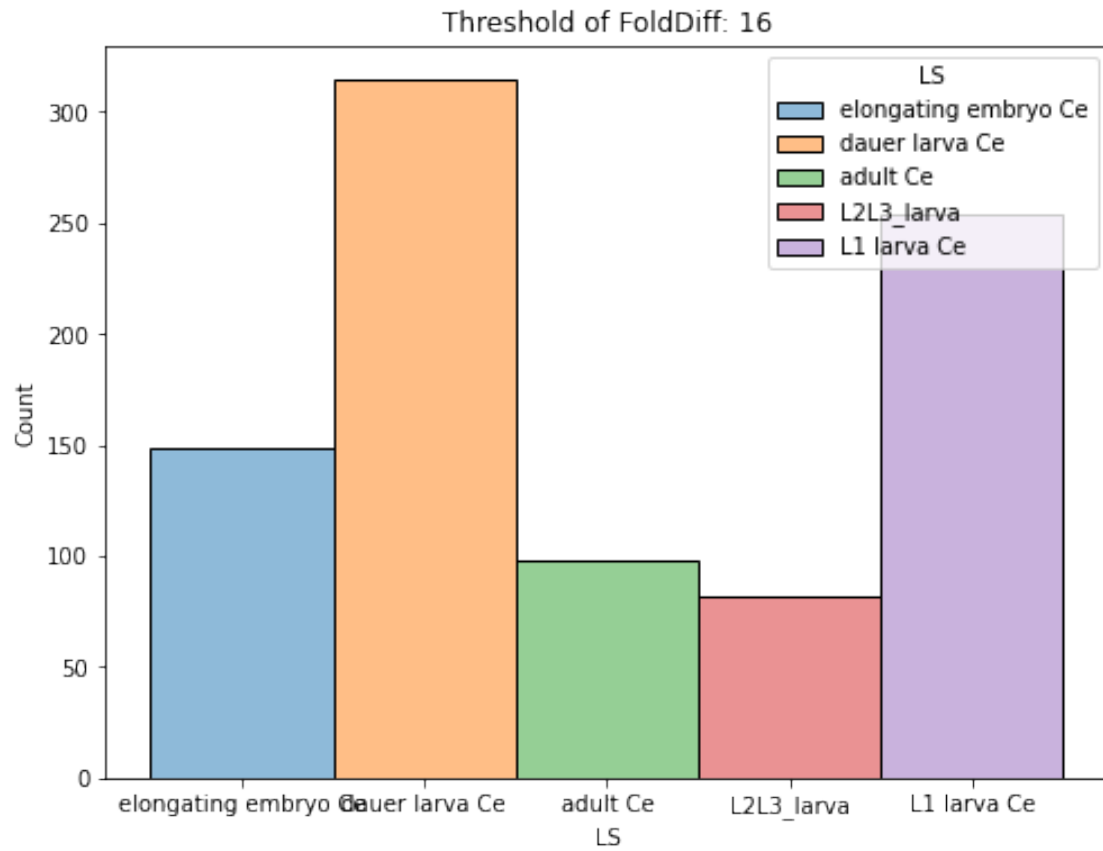
count	15.000000	15.000000	15.000000	15.000000	15.000000
mean	61883.400000	22.866667	9.433333	3457.096160	10.139472
std	78268.637283	31.217822	11.307972	2942.840547	1.425323
min	3023.000000	1.000000	0.250000	1130.571429	8.014005
25%	11010.500000	4.000000	1.375000	1588.897368	9.306535
50%	15522.000000	10.000000	4.250000	2720.250000	9.650014
75%	127015.000000	23.500000	10.750000	4187.228649	11.720771
max	235883.000000	95.000000	33.250000	12412.000000	12.371091

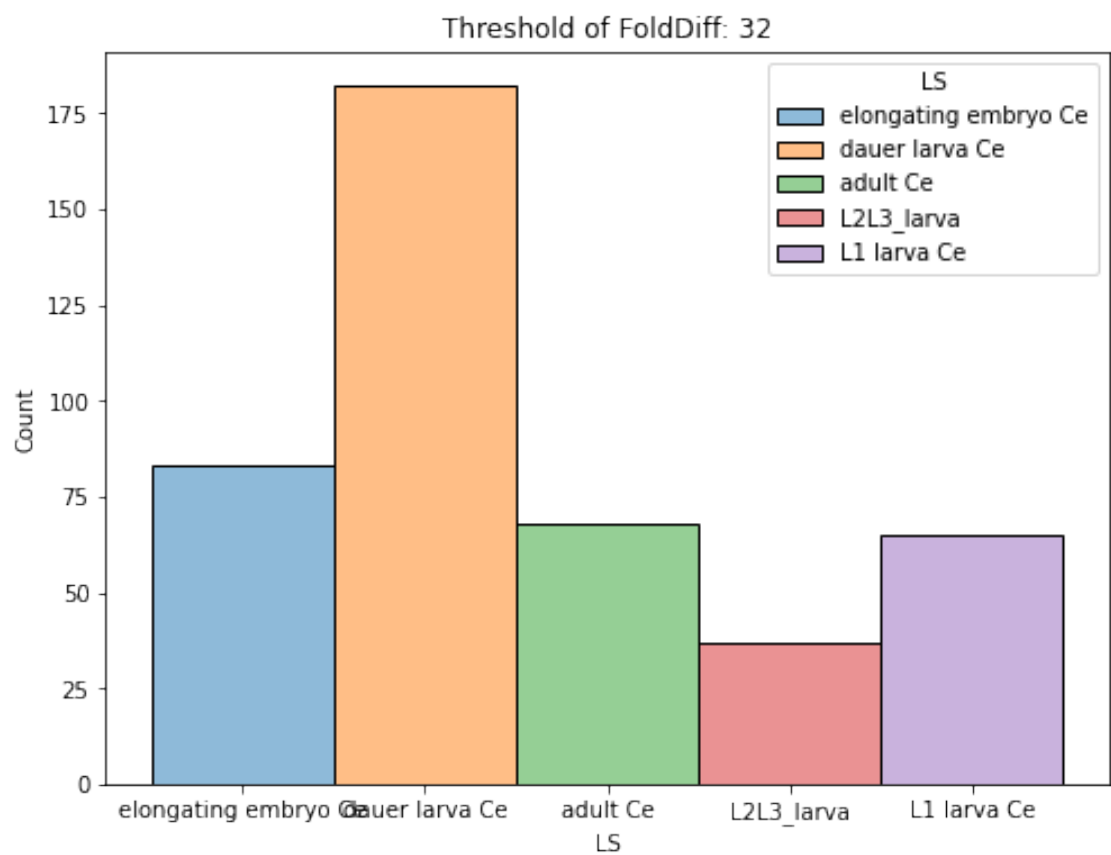
	FoldDiff_LOG	RestMean_LOG
count	15.000000	15.000000
mean	7.901033	1.405217
std	0.691138	1.512399
min	7.030478	-1.386294
25%	7.369608	0.314304
50%	7.908479	1.446919
75%	8.339777	2.361477
max	9.426419	3.504055

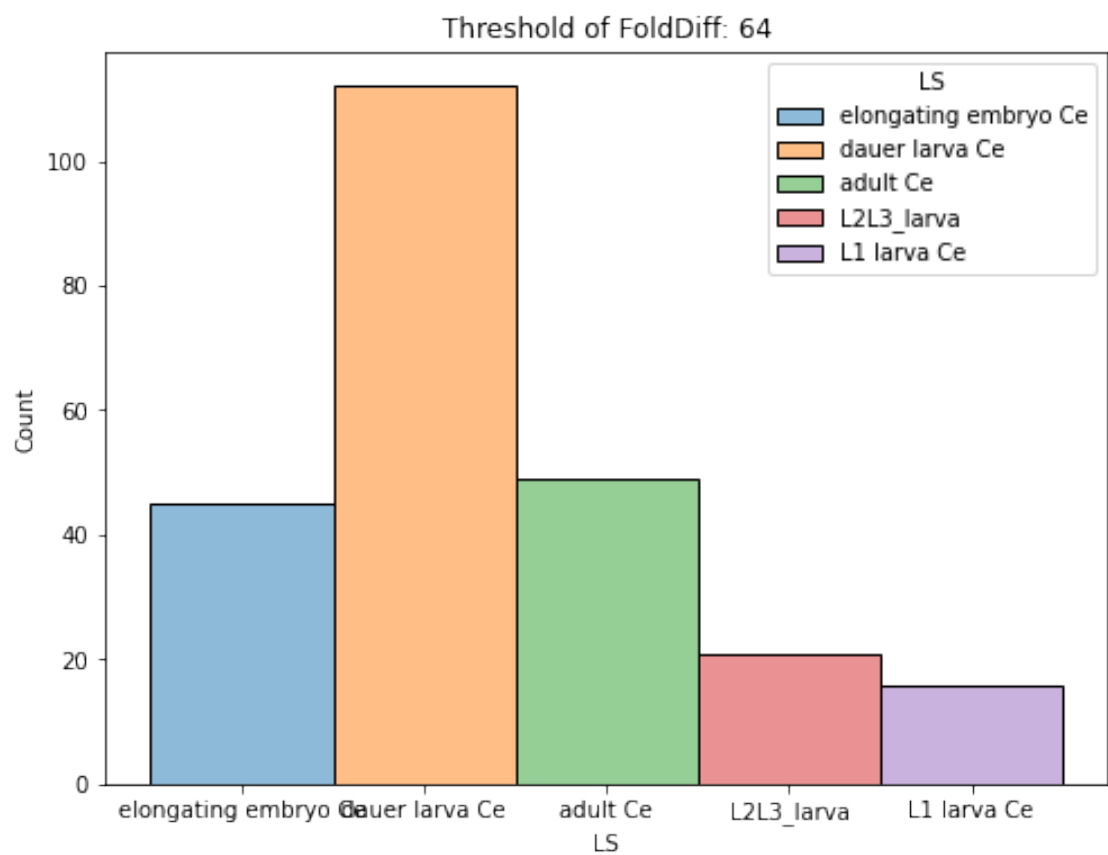


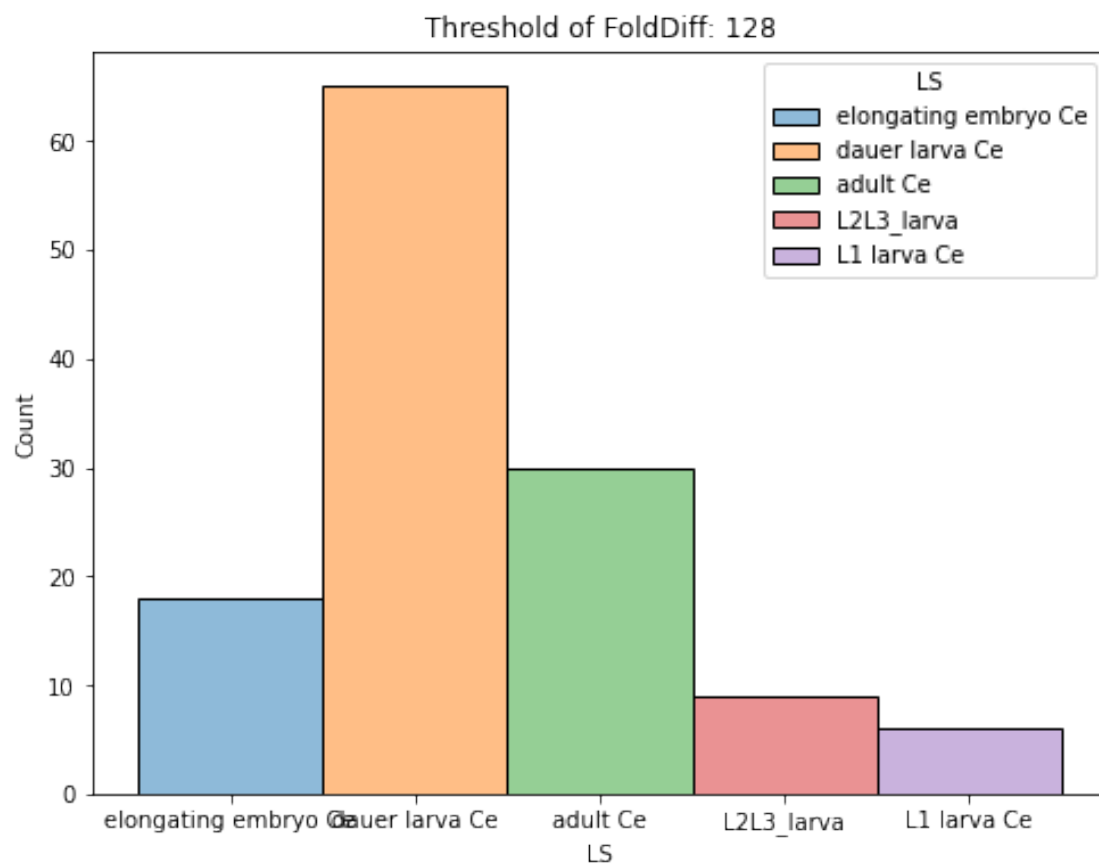


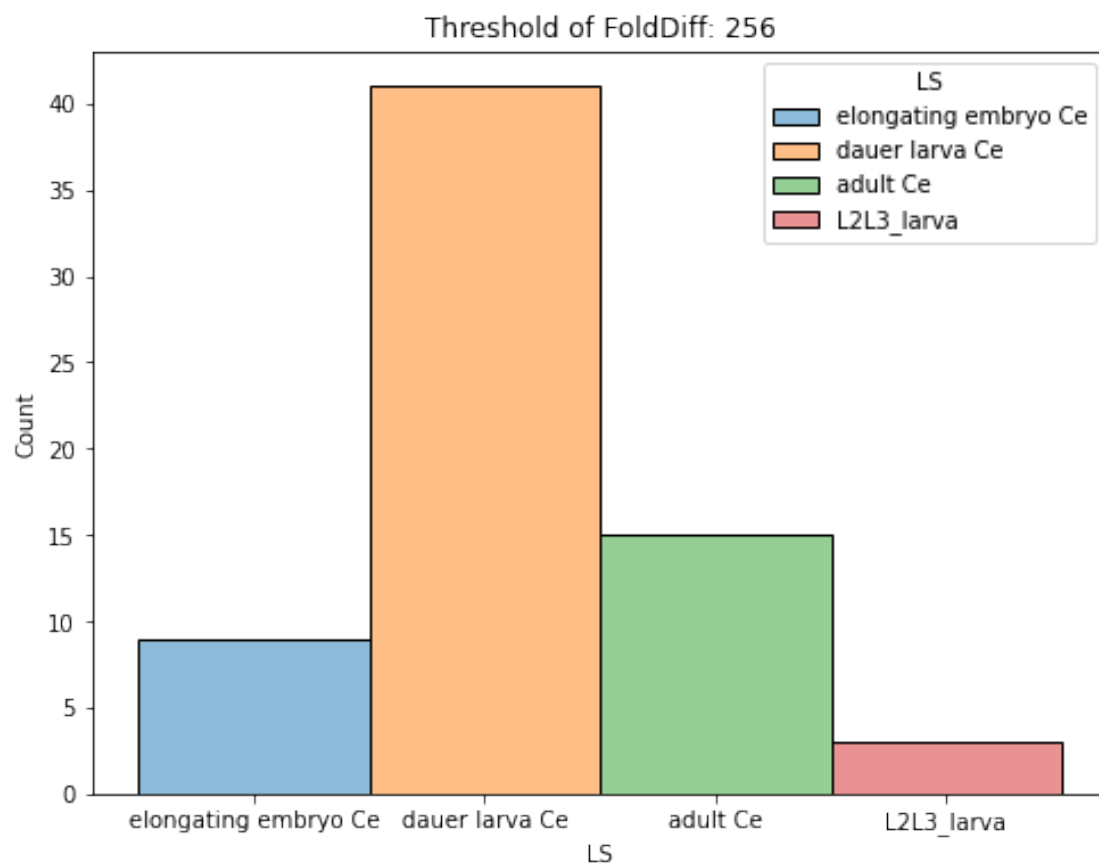


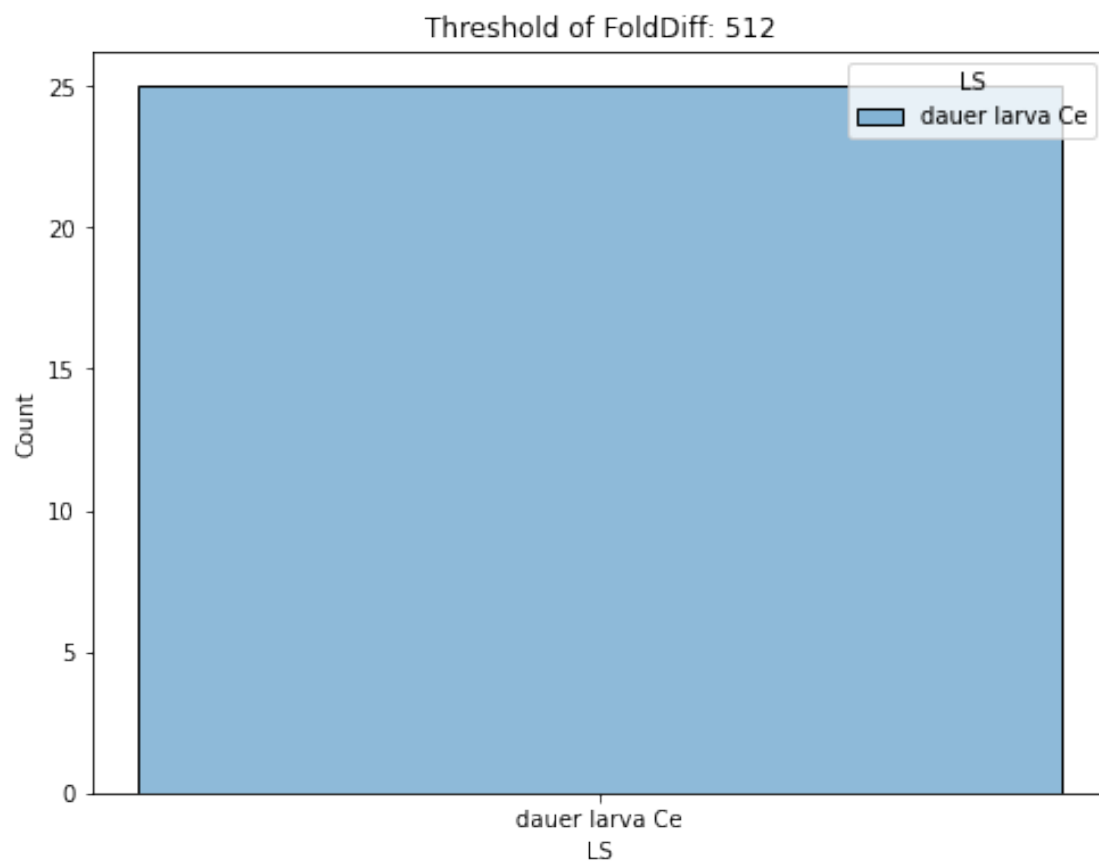


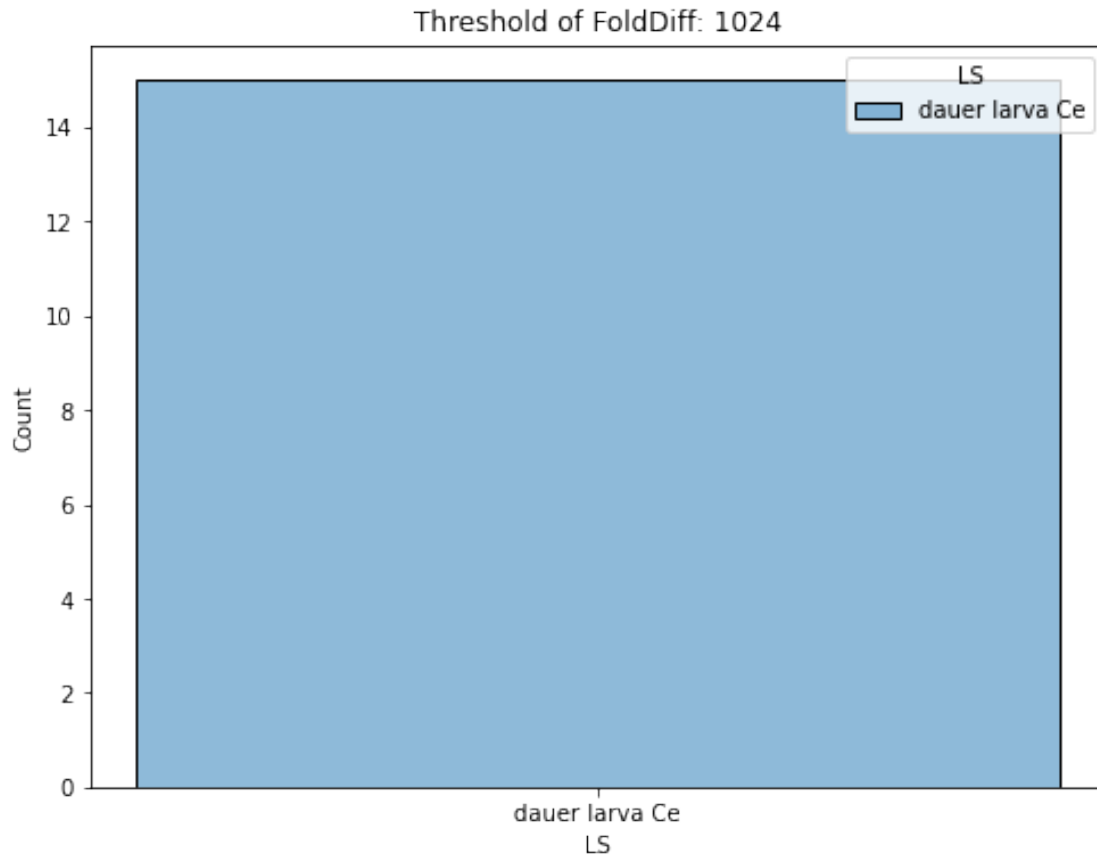












[30]:

```
[30]:
```

	GeneID	LS_EXP	SecondMax	RestMean	FoldDiff	\
LS						
L1 larva Ce	5426	5426	5426	5426	5426	
L2L3_larva	1018	1018	1018	1018	1018	
adult Ce	695	695	695	695	695	
dauer larva Ce	1550	1550	1550	1550	1550	
elongating embryo Ce	1410	1410	1410	1410	1410	

	LS_EXP_LOG	FoldDiff_LOG	RestMean_LOG
LS			
L1 larva Ce	5426	5426	5426
L2L3_larva	1018	1018	1018
adult Ce	695	695	695
dauer larva Ce	1550	1550	1550
elongating embryo Ce	1410	1410	1410

[]: