Experiment Data Depot (EDD) data quality analysis pipeline

Nurgul Kaplan Lease[1], Yan Chen[1], Jennifer Gin[1], Christopher Petzold[1]
[1]Lawrence Berkeley National Laboratory

Christopher Petzold
Lawrence Berkeley National Laboratory

ABSTRACT

This protocol details setting up and running a data quality analysis workflow for data available in the Experiment Data Depot (EDD).

https://www.protocols.io/private/A2B4B174CB7411EA9C520A58A9FEAC2A

EXTERNAL LINK

https://repo.jbei.org/users/nkaplan/repos/ese_automation/browse

THIS PROTOCOL ACCOMPANIES THE FOLLOWING PUBLICATION

Morrell, William C., et al. "The experiment data depot: a web-based software tool for biological experimental data storage, sharing, and visualization." ACS synthetic biology 6.12 (2017): 2248-2259.

EXTERNAL LINK

https://repo.jbei.org/users/nkaplan/repos/ese_automation/browse

PROTOCOL INFO

Nurgul Kaplan Lease, Yan Chen, Jennifer Gin, Christopher Petzold . Experiment Data Depot (EDD) data quality analysis pipeline. **protocols.io**
https://protocols.io/view/experiment-data-depot-edd-data-quality-analysis-pi-biuqkevw

MANUSCRIPT CITATION please remember to cite the following publication along with this protocol

Morrell, William C., et al. "The experiment data depot: a web-based software tool for biological experimental data storage, sharing, and visualization." ACS synthetic biology 6.12 (2017): 2248-2259.

KEYWORDS

Data quality, EDD, Jupyter Notebook, proteomics, metabolomics, omics data

CREATED

Jul 21, 2020

LAST MODIFIED

Nov 24, 2020

PROTOCOL INTEGER ID

39536

GUIDELINES

This python script takes data directly from a study deposited in the Experiment Data Depot (EDD).

MATERIALS TEXT

Access to a Jupyter Notebook

BEFORE STARTING

Set up EDD study and import your data. Users are referred to Morrell et al. in Step #1 for details about setting up a study in the EDD and uploading data.

EDD Study setup

1 Following are additional requests that are needed for successfully execute the script:

1. Line names should NOT include special characters shown below.
!"#$%&'()*+,./:;<=>?@[\]^`{|}~

2. LIne Name includes 2 parts:
Line-1-R1

Line-1 This is a descriptive line name and it is okay to use:
- "-" dash
- " " space
- "_"underscore

-R1 If assay has replicates, MUST append "-", letter "R or r", and number. Here is the list that is ok to use:
- -R1
- -r1
- -R001
- -r001

Line Name Examples:
- LK15_14500-R2
- LK15_14500-r2
- LK15_14500-R002
- LK15_14500-r002

Example EDD study:

https://public-edd.agilebiofoundry.org/s/example-data-quality-study/

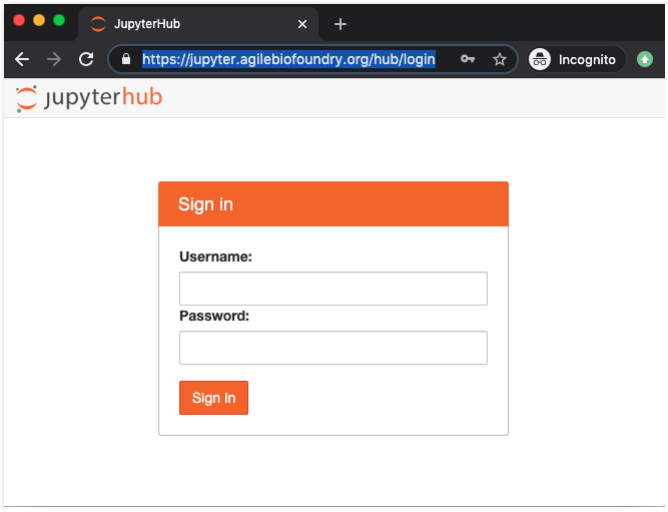
Morrell WC, Birkel GW, Forrer M, Lopez T, Backman TWH, Dussault M, Petzold CJ, Baidoo EEK, Costello Z, Ando D, Alonso-Gutierrez J, George KW, Mukhopadhyay A, Vaino I, Keasling JD, Adams PD, Hillson NJ, Garcia Martin H (2017). The Experiment Data Depot: A Web-Based Software Tool for Biological Experimental Data Storage, Sharing, and Visualization.. ACS synthetic biology. https://doi.org/10.1021/acssynbio.7b00204

Download and Open Jupyter Notebook

2 Click on Jupyter server URL link.
- jupyter.agilebiofoundry.org
- jupyter.jbei.org

Use your LBL LDAP account (lowercase username) to login.

If prompted, choose the 'ESE Data Analysis' Kernel.



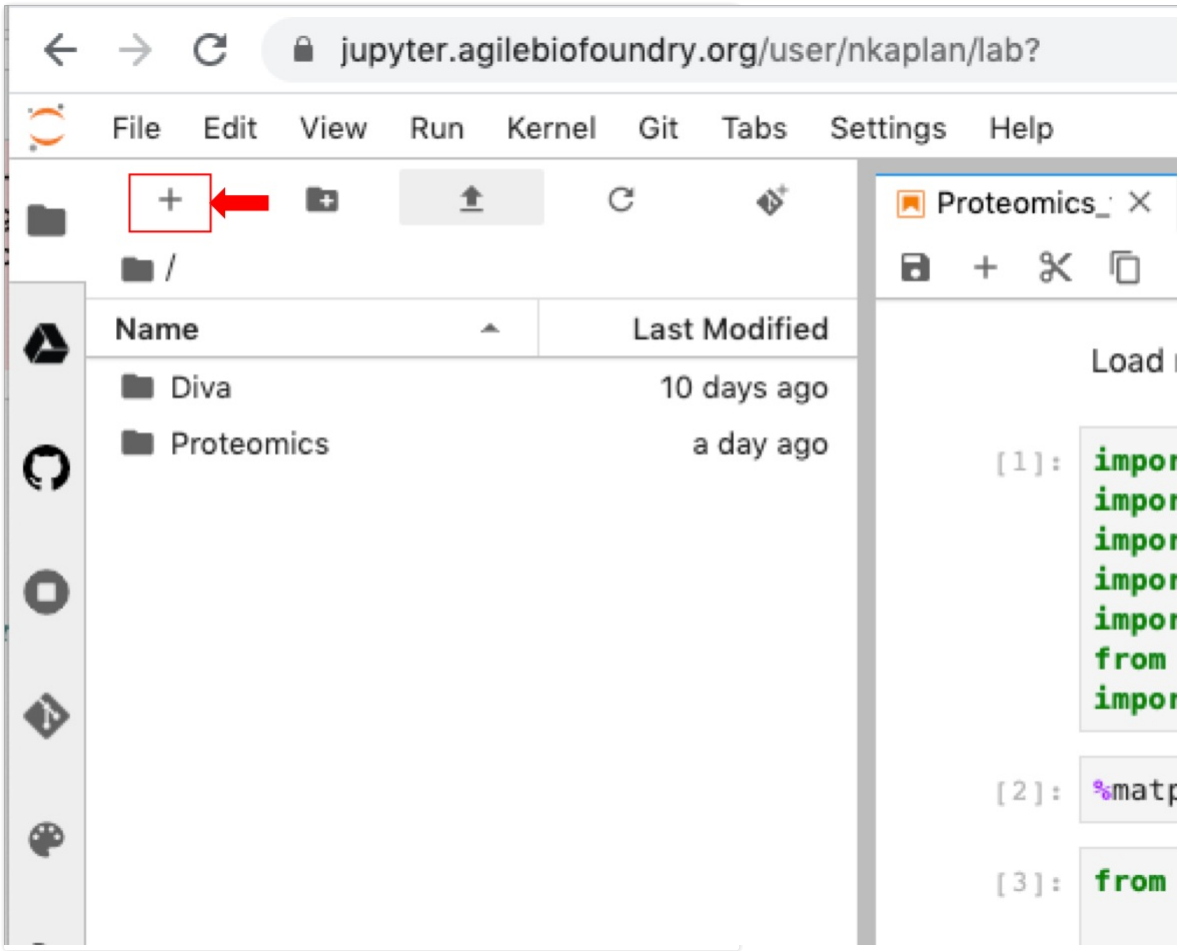
Note: If your account has not been activated, please reach out to Mark Kulawik. mkulawik@lbl.gov

3 Omics-20201123.ipynb is saved in repo.jbei.org

https://repo.jbei.org/users/nkaplan/repos/ese_automation/browse

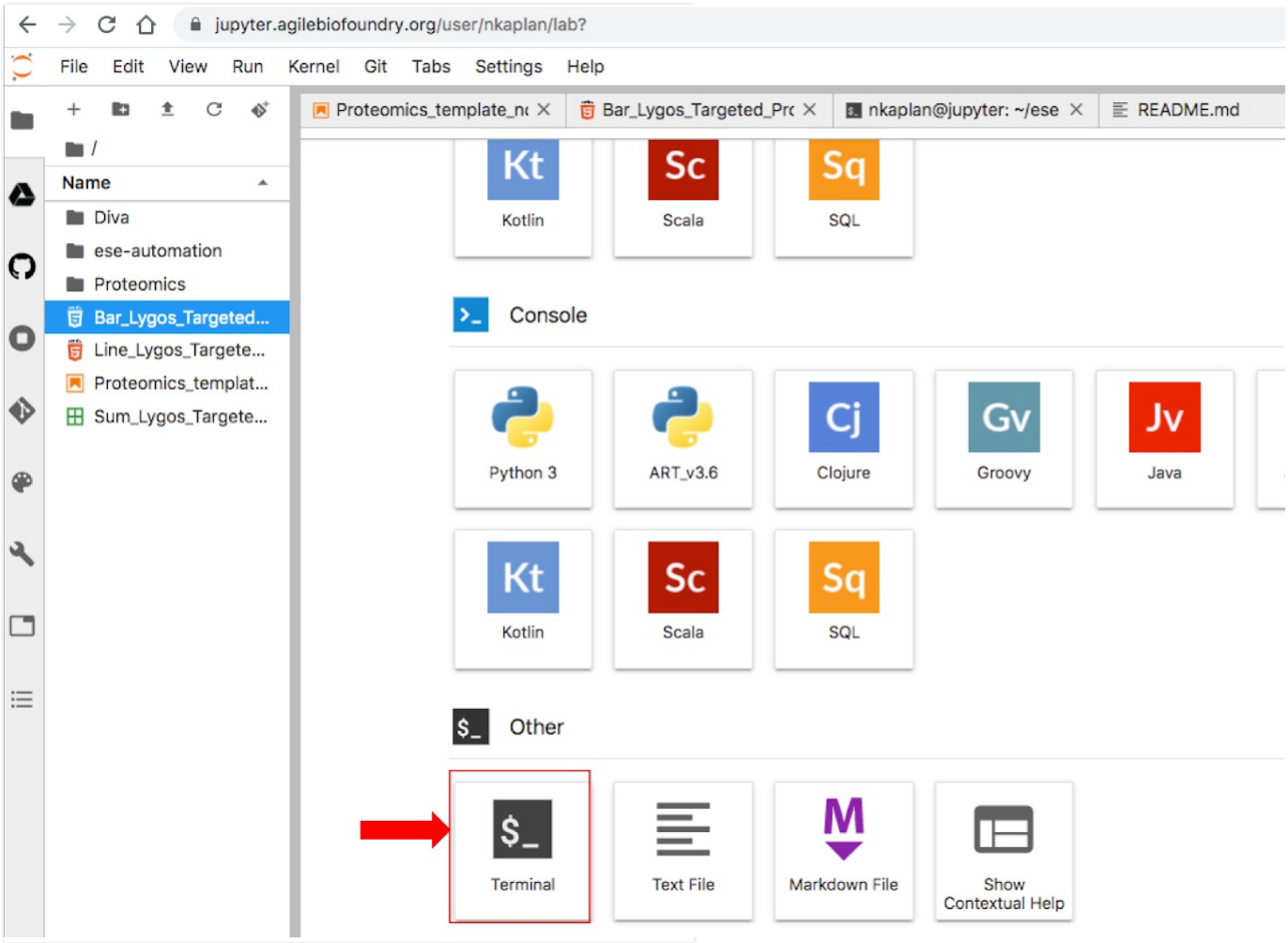
To clone updated repository for the Omics data analysis jupyter notebooks:

Click on the "+" sign on the upper left corner of jupyter server.



3.1 Or you can download this python notebook and upload it to the jupyter notebook server.
Omics-20201123.ipynb

4 Click on Terminal

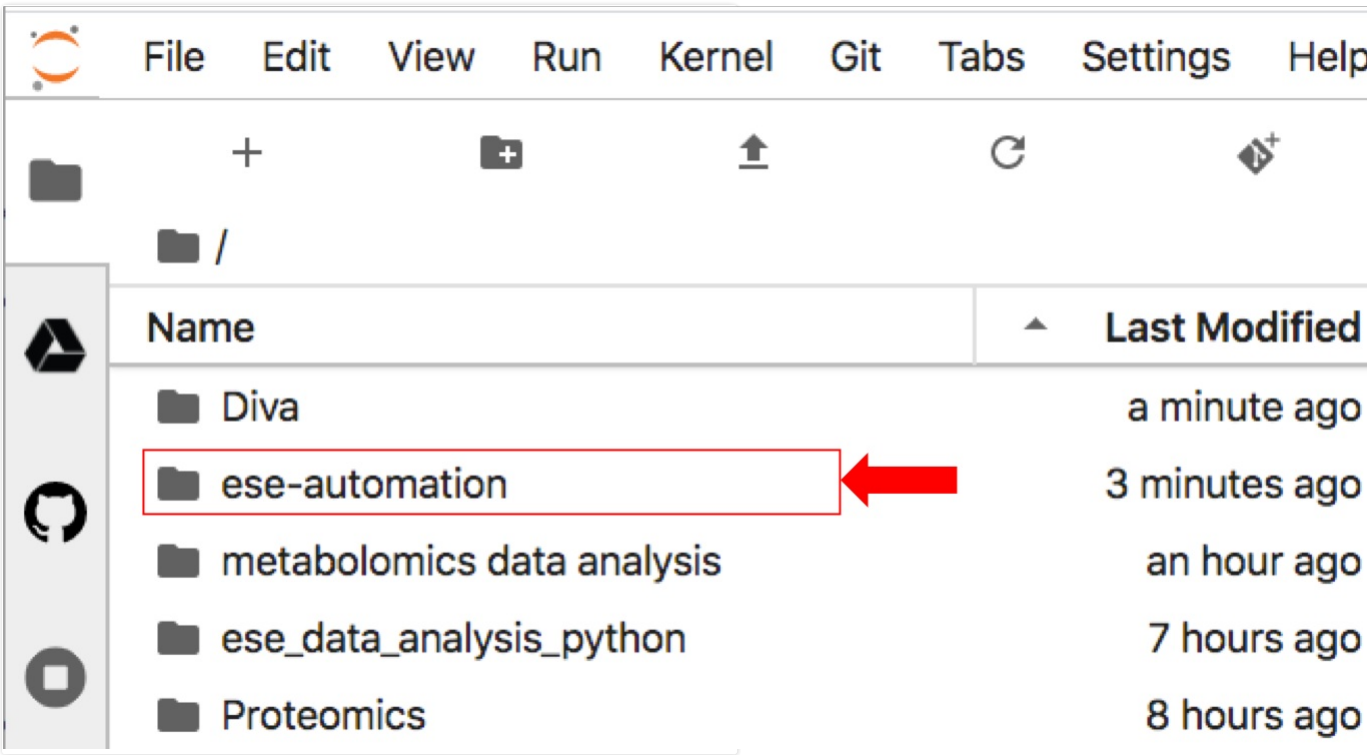


5 Copy the command shown below, and run it at your terminal window.

git clone https://repo.jbei.org/scm/~nkaplan/ese_automation.git

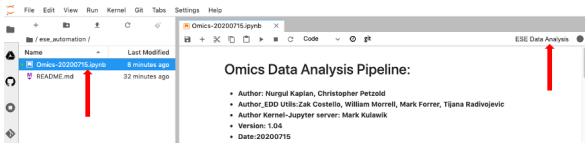
```
YourLDAPusername@jupyter:~$ git clone https://repo.jbei.org/scm/~nkaplan/ese_au
```

Now, you should be able to see the ese-automation folder at your jupyter server directory.



6

1. Double Click on ese_automation folder
2. Double Click on the Omics-20201113.ipynb
3. Choose the Kernel (on the top right side): ESE Data Analysis (if necessary)

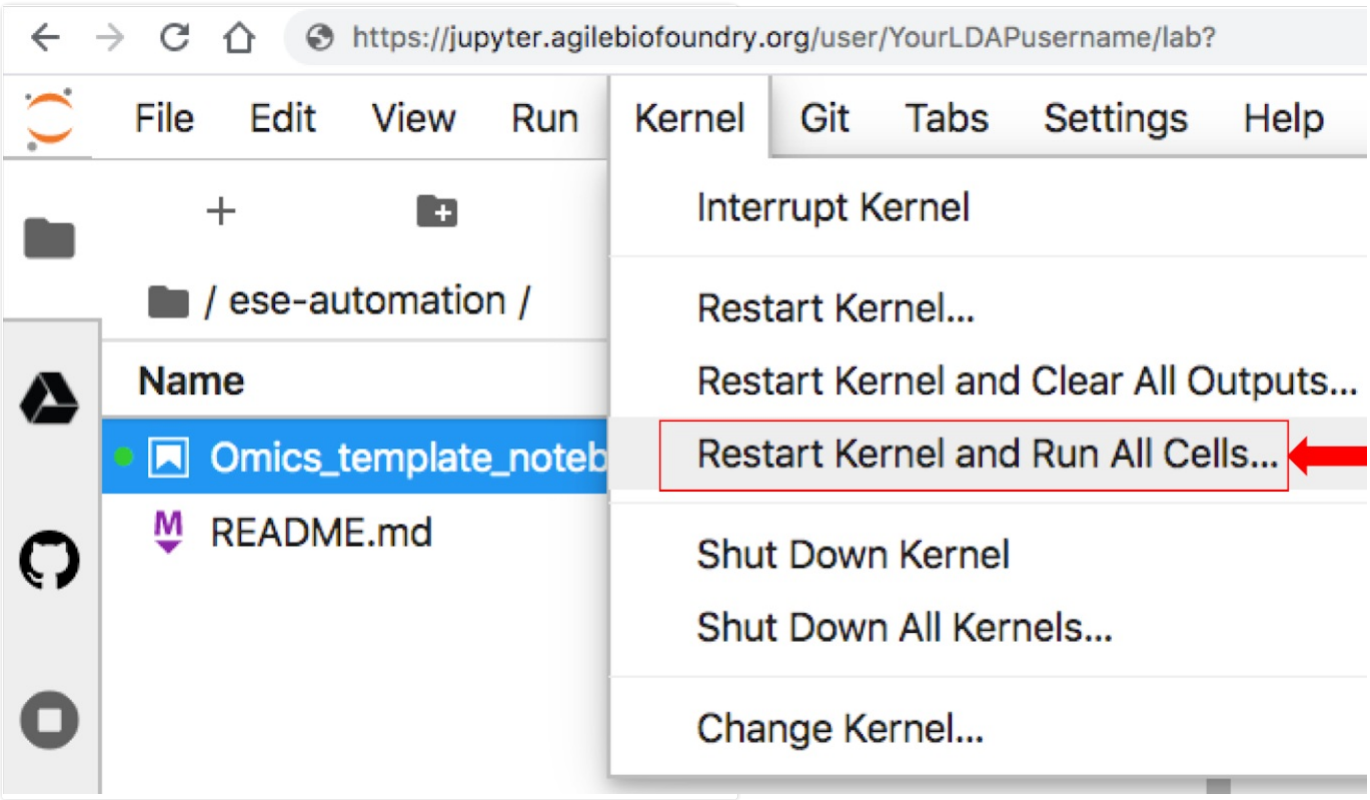


Running the Data Quality Notebook

7 From the 'Kernel' dropdown menu:

Select 'Restart Kernel and Run All Cells'

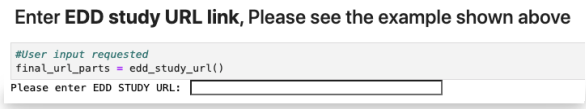
Select 'Restart' at the prompt



8 Copy the EDD url link and paste into code line properly.

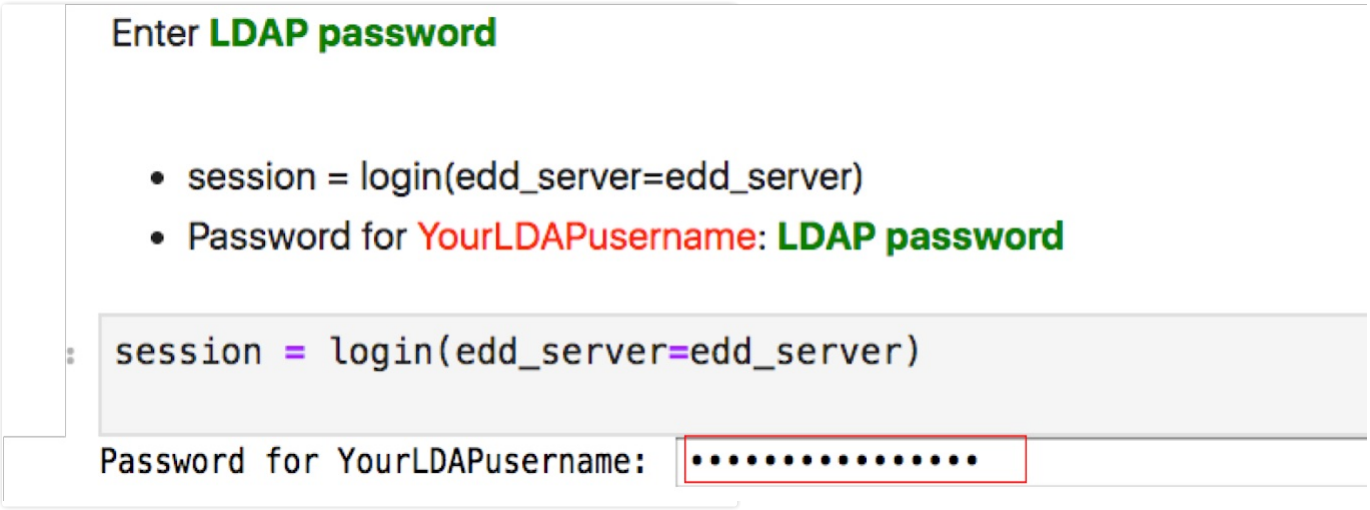
Example:

https://public-edd.agilebiofoundry.org/s/example-data-quality-study/



Input EDD study URL

9 You will need to enter your LBNL LDAP password in the box, and click on the return button.



Once the progress bar reaches 100%, that means your data has been uploaded successfully.

```
[6]: session = login(edd_server=edd_server, user='nkaplan')

Password for nkaplan: ··· ··· ··· ···

[7]: df = export_study(session, study_slug, edd_server=edd_server)

100% 4410/4410 [00:01<00:00, 3119.75it/s]
```

10 By default, the script will analyze the first protocol in the dataset. If there are multiple protocols, please identify the protocol in dropdown menu.

```
OPTIONAL:SPECIFY THE PROTOCOL FROM DROPDOWN MENU
If protocol is NOT specified, ONLY the FIRST protocol will be ANALYZED
```

Protocol: Targeted Proteomics

	Protocol	Assay Name	Formal Type	Measurement Type	Units	Value	Hours
0	Targeted Proteomics	LPK15_14588-R1	tr|A0A1V2LGF2|A0A1V2LGF2_PICKU	Fructose-bisphosphate aldolase	counts	4600.0	10.0
1	Targeted Proteomics	LPK15_14588-R1	tr|A0A1V2LH63|A0A1V2LH63_PICKU	Succinate--CoA ligase [ADP-forming] subunit be...	counts	5124.0	10.0
2	Targeted Proteomics	LPK15_14588-R1	tr|A0A1V2LHP2|A0A1V2LHP2_PICKU	Glucose-6-phosphate isomerase	counts	3673.0	10.0
3	Targeted Proteomics	LPK15_14588-R1	tr|A0A1V2LHT6|A0A1V2LHT6_PICKU	Alpha,alpha-trehalose-phosphate synthase [UDP-...	counts	1006.0	10.0
4	Targeted Proteomics	LPK15_14588-R1	tr|A0A1V2LI03|A0A1V2LI03_PICKU	Transketolase	counts	1905.0	10.0

```
AFTER PROTOCOL SELECTION, YOU MUST GO TO THE NEXT CELL
Run --> Run Selected Cell and All Below
```

Example:
Protocol: Metabolomics

```
OPTIONAL:SPECIFY THE PROTOCOL FROM DROPDOWN MENU
If protocol is NOT specified, ONLY the FIRST protocol will be ANALYZED
```

Protocol: Metabolomics

	Protocol	Assay Name	Formal Type	Measurement Type	Units	Value	Hours
0	Metabolomics	LPK15_11260-R1	cid:6140	L-phenylalanine	counts	661.0	52.0
1	Metabolomics	LPK15_11260-R1	cid:6140	L-phenylalanine	counts	157.0	60.0
2	Metabolomics	LPK15_11260-R1	cid:6140	L-phenylalanine	counts	250.0	72.0
3	Metabolomics	LPK15_11260-R1	cid:6140	L-phenylalanine	counts	174.0	10.0
4	Metabolomics	LPK15_11260-R1	cid:6140	L-phenylalanine	counts	587.0	22.0

```
AFTER PROTOCOL SELECTION, YOU MUST GO TO THE NEXT CELL
Run --> Run Selected Cell and All Below
```

11 By default, the script will analyze ALL of the assays in the dataset. Specific assays can be excluded from the analysis via selection if desired.

```
OPTIONAL: EXCLUDE ASSAY(S)
None means ALL the ASSAYS will be ANALYZED
```

Exclude Assay(s): None, LINE-1, LINE-2, LINE-3

	Protocol	Assay_Name	Replicate_Num	Type_ID	Type_Abb	Measurement Type	Hours	Units	Value
0	TARGETED PROTEOMICS	LINE-3	R1	P00761	TRYP_PIG	TRYPSIN	24.0	COUNTS	27148.0
1	TARGETED PROTEOMICS	LINE-3	R1	P02769	ALBU_BOVIN	SERUM ALBUMIN	24.0	COUNTS	8305.0
2	TARGETED PROTEOMICS	LINE-3	R1	P00698	LYSC_CHICK	LYSOZYME C	24.0	COUNTS	8508.0
3	TARGETED PROTEOMICS	LINE-3	R2	P00761	TRYP_PIG	TRYPSIN	24.0	COUNTS	33504.0
4	TARGETED PROTEOMICS	LINE-3	R2	P02769	ALBU_BOVIN	SERUM ALBUMIN	24.0	COUNTS	10277.0

```
AFTER ASSAY(S) SELECTION, YOU MUST GO TO THE NEXT CELL
Run --> Run Selected Cell and All Below
```

Output Files and Plots

12 Summary report files and plots will be generated in the same folder as your jupyter notebook. The list of files includes:

- Summary report (Average, Standard deviation, Coefficient of variation)
- Full data report (Full data export)
- Bar charts
- Line charts (if there are multiple time points)
- Data quality charts and tables (Scatter plot, Violin plots, Data quality metrics)
- To view the plots in jupyter server, click on "Trust". Alternatively, download the plots and open them in your browser.

Name	Last Modified
Bar_Example_Data_Quality_Study_TARGETED_PROTEOMICS_20200722-121017.html	a minute ago
DataQuality_Example_Data_Quality_Study_TARGETED_PROTEOMICS_20200722-121019.html	a minute ago
Full_data_Example_Data_Quality_Study_TARGETED_PROTEOMICS_20200722-121017.csv	a minute ago
Line_Example_Data_Quality_Study_TARGETED_PROTEOMICS_20200722-121017.html	a minute ago
Omics-20200715.ipynb	seconds ago
README.md	an hour ago
Sum_data_Example_Data_Quality_Study_TARGETED_PROTEOMICS_20200722-121017.csv	a minute ago

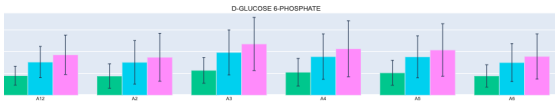
Full data report:

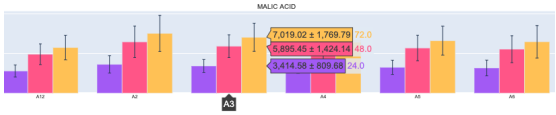
Protocol	Assay_Name	Replicate_Num	Type_ID	Type_Abb	Measurement Type	Hours	Units	Value
METABOLOMICS	A1	R1	CID:33032	L-GLUTAMIC ACID	L-GLUTAMIC ACID	24.0	COUNTS	15492.0
METABOLOMICS	A1	R1	CID:33032	L-GLUTAMIC ACID	L-GLUTAMIC ACID	48.0	COUNTS	25097.04
METABOLOMICS	A1	R1	CID:33032	L-GLUTAMIC ACID	L-GLUTAMIC ACID	72.0	COUNTS	29899.56
METABOLOMICS	A1	R1	CID:3614358	GLYOXYLATE	GLYOXYLATE	24.0	COUNTS	59589.0
METABOLOMICS	A1	R1	CID:3614358	GLYOXYLATE	GLYOXYLATE	48.0	COUNTS	107856.09
METABOLOMICS	A1	R1	CID:3614358	GLYOXYLATE	GLYOXYLATE	72.0	COUNTS	115006.77
METABOLOMICS	A1	R1	CID:439958	OSE 6-PHOSPHATE	OSE 6-PHOSPHATE	24.0	COUNTS	3291.0
METABOLOMICS	A1	R1	CID:439958	OSE 6-PHOSPHATE	OSE 6-PHOSPHATE	48.0	COUNTS	5331.42
METABOLOMICS	A1	R1	CID:439958	OSE 6-PHOSPHATE	OSE 6-PHOSPHATE	72.0	COUNTS	6944.01
METABOLOMICS	A1	R1	CID:525	MALIC ACID	MALIC ACID	24.0	COUNTS	2772.0
METABOLOMICS	A1	R1	CID:525	MALIC ACID	MALIC ACID	48.0	COUNTS	5017.32
METABOLOMICS	A1	R1	CID:525	MALIC ACID	MALIC ACID	72.0	COUNTS	5349.96
METABOLOMICS	A1	R1	CID:5793	D-GLUCOSE	D-GLUCOSE	24.0	COUNTS	8820.0
METABOLOMICS	A1	R1	CID:5793	D-GLUCOSE	D-GLUCOSE	48.0	COUNTS	15964.2
METABOLOMICS	A1	R1	CID:5793	D-GLUCOSE	D-GLUCOSE	72.0	COUNTS	17022.6

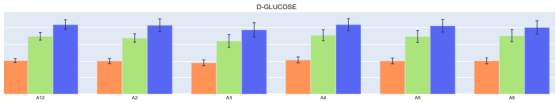
Summary data report:

Protocol	Assay_Name	Type_ID	Type_Abb	Units	Hours	mean	std	CV
METABOLOMICS	A1	CID:33032	L-GLUTAMIC ACID	COUNTS	24.0	13412.25	1876.68	0.14
METABOLOMICS	A1	CID:33032	L-GLUTAMIC ACID	COUNTS	48.0	23426.58	3697.09	0.16
METABOLOMICS	A1	CID:33032	L-GLUTAMIC ACID	COUNTS	72.0	27513.87	3484.74	0.13
METABOLOMICS	A1	CID:3614358	GLYOXYLATE	COUNTS	24.0	51646.42	3702.2	0.07
METABOLOMICS	A1	CID:3614358	GLYOXYLATE	COUNTS	48.0	90042.09	8205.66	0.09
METABOLOMICS	A1	CID:3614358	GLYOXYLATE	COUNTS	72.0	104685.51	6537.44	0.06
METABOLOMICS	A1	CID:439958	OSE 6-PHOSPHATE	COUNTS	24.0	1763.0	961.8	0.55
METABOLOMICS	A1	CID:439958	OSE 6-PHOSPHATE	COUNTS	48.0	3014.36	1612.17	0.53
METABOLOMICS	A1	CID:439958	OSE 6-PHOSPHATE	COUNTS	72.0	3653.98	2049.82	0.56
METABOLOMICS	A1	CID:525	MALIC ACID	COUNTS	24.0	3525.25	1073.94	0.3
METABOLOMICS	A1	CID:525	MALIC ACID	COUNTS	48.0	6176.33	1917.4	0.31
METABOLOMICS	A1	CID:525	MALIC ACID	COUNTS	72.0	7247.93	2229.91	0.31
METABOLOMICS	A1	CID:5793	D-GLUCOSE	COUNTS	24.0	9555.75	1064.36	0.11
METABOLOMICS	A1	CID:5793	D-GLUCOSE	COUNTS	48.0	16346.4	1862.77	0.11
METABOLOMICS	A1	CID:5793	D-GLUCOSE	COUNTS	72.0	19371.27	2684.6	0.14

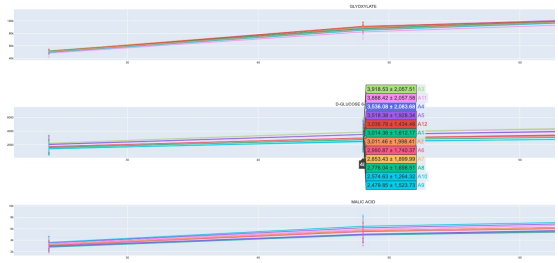
Bar charts:







Line charts:



13 Data Quality:

If there are no replicates in the study, the Data Quality script will not generate Scatter or Violin plots.

If there are replicates in the dataset, scatter and violin plots will be generated with the complete dataset.

Users have an option to input an intensity threshold to exclude data with values less than the cutoff line. There is a live feedback of the action by showing the number and the percentage of remaining data points.

OPTIONAL: SPECIFY CUTOFF PEAK AREA , Default is ZERO for cutoff peak area

Cutoff Peak: 0

Subset Data	Full Data	Percentage %
0	27	0.0%

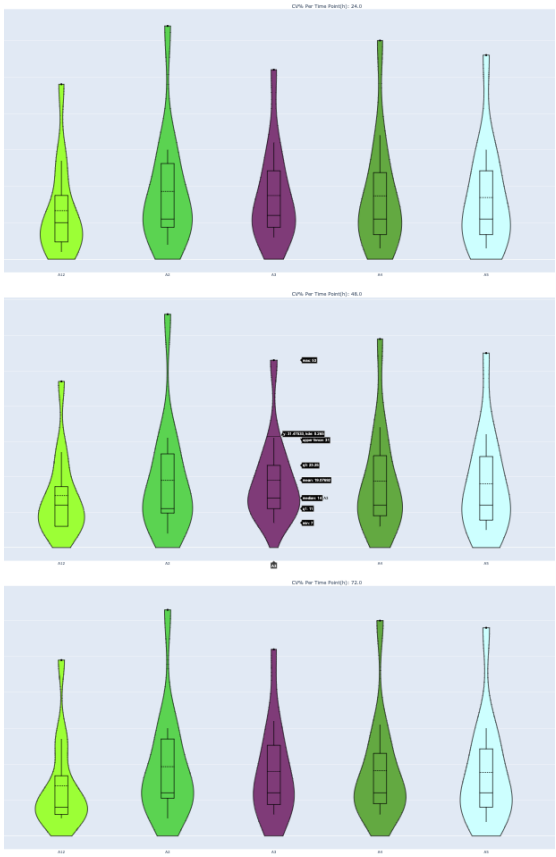
	Protocol	Assay_Name	Type_ID	Type_Abb	Units	Time_Point	mean	std	CV	CV%	type_units
0	TARGETED PROTEOMICS	LINE-1	P00698	LYSC_CHICK	COUNTS	24.0	3257.00	145.66	0.04	4.0	LYSC_CHICK*counts
1	TARGETED PROTEOMICS	LINE-1	P00698	LYSC_CHICK	COUNTS	48.0	4885.50	218.50	0.04	4.0	LYSC_CHICK*counts
2	TARGETED PROTEOMICS	LINE-1	P00698	LYSC_CHICK	COUNTS	72.0	6351.15	284.04	0.04	4.0	LYSC_CHICK*counts
3	TARGETED PROTEOMICS	LINE-1	P00761	TRYP_PIG	COUNTS	24.0	9861.00	3750.49	0.38	38.0	TRYP_PIG*counts
4	TARGETED PROTEOMICS	LINE-1	P00761	TRYP_PIG	COUNTS	48.0	14791.50	5625.74	0.38	38.0	TRYP_PIG*counts

AFTER CHANGING THE INTENSITY, YOU MUST GO TO THE NEXT CELL
Run --> Run Selected Cell and All Below

Scatter Plot:

DataQuality%: The percentage of the data is below 20% CV (coefficient variation) for each set of replicates.
The red line in the scatter plot corresponds to 20% CV, a commonly used cutoff for data quality. This may or may not be suitable for your application..

Violin Plot:



Data Quality metrics:
DataQuality%: The percentage of the data is below 20% CV (coefficient variation) for each set of replicates

Assay_Name	Time_Point	DataPointCounts	DataQuality%
A1	24	13	46.15
A1	48	13	46.15
A1	72	13	46.15
A10	24	13	69.23
A10	48	13	69.23
A10	72	13	69.23
A11	24	13	69.23
A11	48	13	69.23
A11	72	13	69.23
A12	24	13	76.92
A12	48	13	76.92
A12	72	13	76.92
A2	24	13	69.23
A2	48	13	69.23
A2	72	13	69.23
A3	24	13	69.23
A3	48	13	61.54

Overall Data Quality%: The percentage of the overall data is below 20% CV (coefficient variation).

Overall Data Quality%
66.03