

Protigy

How-to: Data upload and file formats

BroadE Proteomics Workshop
February 28, 2018

How do I get my data into Protigy?

Option 1: You are collaborating with us:

- We upload the data for you and create a Protigy session
- You are all set to play with your data

Option 2: You want to upload your own data and

A) the data is in GCT1.3 format:

- just go ahead and upload the data file

B) your data is NOT in GCT 1.3 format

- Protigy supports multiple text formats
 - tab/comma/semicolon-separated (txt, tsv, csv, ssv, GCT1.2)
- Experimental design file to assign phenotype labels to sample columns

Data import: Option 1

`http://shiny-proteomics.broadinstitute.org:3838/protigy/`

The screenshot shows the PROtigy Shiny app interface. The top navigation bar is blue and contains the version 'Protigy (v0.8.0.8)', a menu icon, the user email 'karsten@broadinstitute.org', and links for 'Logout', 'Help me!', and the BROAD PROTEOMICS logo.

On the left sidebar, there is an 'Upload file (txt, csv, gct):' section with a 'Browse...' button and 'No file selected' text. Below it is a 'Saved sessions:' section with a dropdown menu showing 'Bro' and a list of sessions: 'BroadE_2016-11-04 18:05:26' and 'BroadE-hands-on-1_2018-02-23 16:29:28'.

The main content area features the 'PROtigy' logo and the subtitle 'PROteomics Toolset for InteGrative Data Analysis'. It includes a description of the app's purpose and deployment options. A red box highlights the 'Supported input formats:' section, which contains the following instructions:

- Select the session name you got from us
- Press 'Import'-button

Below the red box, the 'Data manipulation' section lists various processing steps: Transformation (log transformation), Sample-wise Normalization (Centering (median), Centering and scaling (median-MAD), Quantile, 2-component), and Filtering (Reproducibility filter across replicate measurements, Standard deviation across samples). The 'Marker selection (based on limma package)' section lists: One-sample moderated T-test, Two-sample moderated T-test, and Moderated F-test. The 'Interactive data analysis and visualization' section lists: Heatmaps and cluster analysis, Volcano plots, Scatterplots, Principle component analysis, and QC-plots (Pairs-plots).

On the far left, there is a red text overlay that reads: 'To analyze another data set or to start over hit the F5 button.'

Data import: Option 2A (GCT 1.3)

GCT1.3 files enable storing of both row and column metadata.

The diagram illustrates the structure of a GCT 1.3 file. It shows a table with columns labeled A through I and rows numbered 1 through 18. Annotations with arrows point to specific parts of the table:

- chds**: Points to cell A1 containing "#1.3".
- version**: Points to cell B1 containing "10".
- dimensions**: Points to cell C1 containing "6".
- rhds**: Points to cell D1 containing "2".
- cids**: Points to cell I1 containing "5".
- column metadata**: Points to the header row (row 3), which contains identifiers for each column: "id", "pr_gene_symbol", "pr_is_lmark", and then six "LPROT001_A375" entries followed by "BRD-K52313696" and "BRD-K77908580".
- rids**: Points to the first column of data (column A, rows 9-18), which contains gene identifiers like "5720", "55847", "7416", etc.
- row metadata**: Points to the second column of data (column B, rows 9-18), which contains gene symbols like "PSME1", "CISD1", "VDAC1", etc.
- data matrix**: Points to the numerical data values in columns D through I, rows 9-18.

	A	B	C	D	E	F	G	H	I
1	#1.3								
2	10	6	2	5					
3	id	pr_gene_symbol	pr_is_lmark	LPROT001_A375	LPROT001_A375	LPROT001_A375	LPROT001_A375	LPROT001_A375	LPROT001_A375
4	pert_id	-666	-666	DMSO	DMSO	BRD-K52313696	BRD-K52313696	BRD-K52313696	BRD-K77908580
5	pert_iname	-666	-666	DMSO	DMSO	tacedinaline	tacedinaline	tacedinaline	entinostat
6	cell_id	-666	-666	A375	A375	A375	A375	A375	A375
7	pert_time	-666	-666						
8	pert_dose	-666	-666						
9	5720	PSME1	1	8.7980	8.9395	8.9561	9.4491	9.1994	9.2937
10	55847	CISD1	1	9.8349	9.5334	9.7543	9.9203	9.8904	10.0666
11	7416	VDAC1	1	12.5431	12.3479	12.4662	12.5892	12.7088	12.6397
12	10174	SORBS3	1	8.1017	8.5660	8.3563	8.6225	8.5800	8.3666
13	25803	SPDEF	1	11.0651	11.0922	11.3566	11.1527	11.0557	10.7427
14	466	ATF1	1	6.6887	6.8780	6.6684	7.1662	6.7492	6.7492
15	6676	SPAG4	1	3.5195	3.8840	3.6936	3.4822	3.2684	3.7300
16	1870	E2F2	1	4.5798	4.6260	4.4156	4.2644	4.4611	4.4611
17	6009	RHEB	1	11.7969	12.3234	12.0216	11.9772	12.0580	12.0216
18	3480	IGF1R	1	8.9370	10.0307	9.5279	10.0616	8.9231	8.6343

Data import: Option 2A (GCT 1.3)

GCT1.3 files enable storing both row and column metadata.

chds

version

dimensions

rhds

cids

	A	B	C	D	E	F	G	H	I
1	#1.3								
2	10		6	2	5				
3	id	proteome symbol	proteome id	LPROT001_A375	LPROT001_A375	LPROT001_A375	LPROT001_A375	LPROT001_A375	LPROT001_A375
4				EH_X1_P20-P03	EH_X1_P20-P05	EH_X1_P20-P07	EH_X1_P20-P09	EH_X1_P20-P11	EH_X1_P20-P13

column

Protigy (v0.8.0.5)

karsten@broadinstitute.org Logout Help me!

Upload file (txt, csv, gct):

Browse...

No file selected

Saved sessions:

Search sessions

Import

Manage sessions

PROTigy

PROteomics Toolset for InteGrative Data Analysis

This Shiny app facilitates exploratory and interactive analysis of data sets derived from quantitative *proteomics* experiments, *RNA-seq* and gene expression *microarrays*.

The app can run locally on your Desktop computer (Windows/Linux/MAC) or can be deployed to Shiny Server environments. To access all implemented features the app has to run on a [Shiny Server Pro \(SSP\)](#) instance, see below for a summary of features only available in SSP.

Supported input formats:

- Any type of text file containing both, expression and annotation columns, can directly be imported into the app.
- Supported file formats:
 - text files (tsv, csv, txt)
 - gct 1.2
 - NEW* gct 1.3

rids

row metadata

data matrix

Data import: Option 2A (GCT 1.3)

Experimental condition to compare can be selected from column meta data:

Column meta data

Protigy (v0.8.2)

karsten@broadinstitute.org

Logout

Help me!



Choose column

- ☐ run order
- ☒ iTRAQ.Experiment.number
- ☐ PAM50
- ☐ ER.Status
- ☐ PR.Status
- ☐ HER2.Status
- ☐ Clean.Bimodality
- ☐ Proteome.Cluster
- ☐ id

OK

To analyze another data set or to start over hit the F5 button.

Found GCT v1.3 file

Choose the annotation column to use as class vector for marker selection.

Current selection: PAM50

Level	Freq
Basal	19
Her2	13
LumA	23
LumB	25
Normal	3

Number of samples in each category of current selection.

GCT1.3 files enable storing both r...

Data import: Option 2B

- Supports upload of data tables in various text-based formats
 - tab-separated (txt, tsv, GCT 1.2), comma-separated (csv), semicolon-separated (ssv)

GCT 1.2:

Always "#1.2"

The # of rows (i.e probe sets)

of samples

Third column onwards are sample names. These must be UNIQUE

Data starts on line 4

Column 1: Row identifiers. Typically probe set ids or clone ids. These must be UNIQUE

Column 2: Row descriptions. Ignored by the program – can be dummy values (e.g. "na")

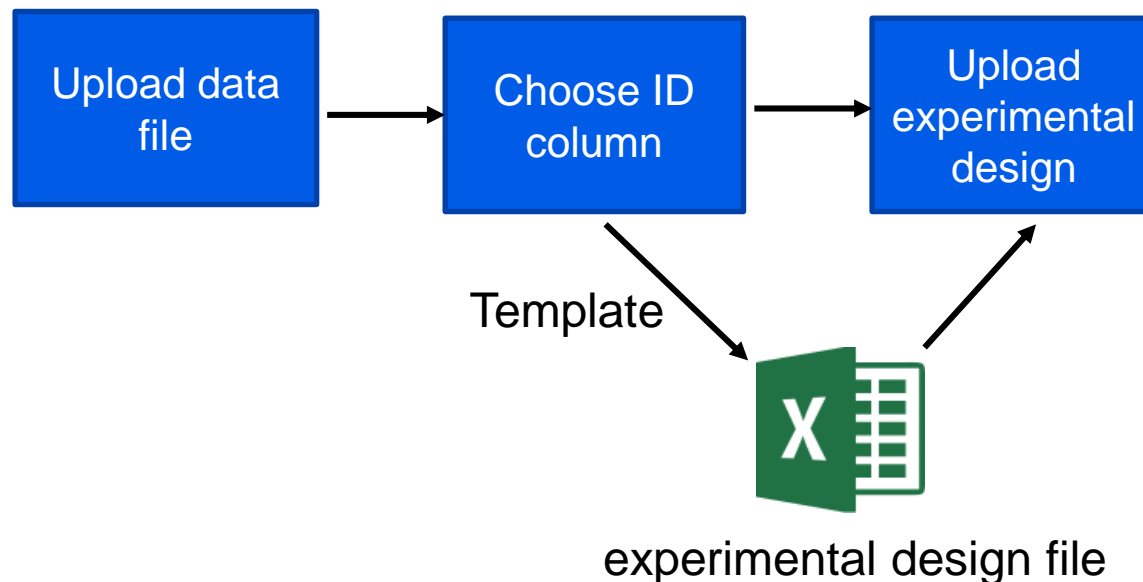
Each column contains expression values from 1 sample. Missing values are allowed (leave empty).

If editing, in Excel, make sure to save your data as "tab delimited text"

1	A	B	C	D	E	F	G
2	#1.2						
3	1000	130					
4	NAME	Description	DLBCL 205	DLBCL 206	DLBCL 232	DLBCL 239	DLBCL 240
5	1007_s_at	U48705 /FEATURE=mRNA /DEFINITION=HSP	280.53	271.48	113.57	124.91	124.91
6	1053_at	M87338 /FEATURE= /DEFINITION=HUMA1S	32.13	91.6	117.43	41.29	33.66
7	117_at	X51757 /FEATURE=cds /DEFINITION=HSP70	51.27	61.12	24.1	41.44	43.56
8	121_at	X69699 /FEATURE= /DEFINITION=HSPAX8A	738.32	330.59	249.89	394.55	329.55
9	1255_g_at	L36861 /FEATURE=expanded_cds /DEFINIT	88.45	12.94	18.46	29.96	39
10	1294_at	L13852 /FEATURE= /DEFINITION=HUME1UR	85.57	88.06	62.24	96.59	81.01
11	1316_at	X55005 /FEATURE=mRNA /DEFINITION=HSC	106.87	45.11	30.05	46.65	36.5
12	1320_at	X79510 /FEATURE=cds /DEFINITION=HSPTR	58.49	27.95	17.6	27.87	26.52
13	1405_i_at	M21121 /FEATURE= /DEFINITION=HUMTCS	10.83	135.24	13.43	203.16	85.74
14	1431_at	J02843 /FEATURE=cds /DEFINITION=HUMC	41.88	24.09	16.07	26.68	25.4
15	1438_at	X75208 /FEATURE=cds /DEFINITION=HSPTR	80.87	9.77	15.33	11.18	44.59
16	1487_at	L38487 /FEATURE=mRNA /DEFINITION=HUI	64.26	80.61	102.9	59.77	105.72
17	1494_f_at	M33318 /FEATURE=mRNA /DEFINITION=HU	213.37	96.88	65.06	96.14	78.77
18	1598_g_at	L13720 /FEATURE= /DEFINITION=HUMGAS	458.88	215.59	186.72	187.36	237.69
19	160020_at	Z48481 /FEATURE=cds /DEFINITION=HSMM	411.94	171.16	130	234.76	266.96
20	1729_at	L41690 /FEATURE= /DEFINITION=HUMTRAD	81.59	83.94	74.75	110.9	126.98
21	1773_at	L00635 /FEATURE= /DEFINITION=HUMFPTE	62.82	45.96	41.15	23.1	28.41
22	177_at	U38545 /FEATURE= /DEFINITION=HSU3854	57.04	28.05	16.74	29.66	53.29
23	179_at	U38980 /FEATURE= /DEFINITION=U38980 H	333.96	254.15	241.24	350.58	193.53

Data import: Option 2B






- Supports upload of data tables in various text-based formats
 - tab-separated (txt, tsv, GCT 1.2), comma-separated (csv), semicolon-separated (ssv)
- Requires upload of an 'experimental design'-file in order to:
 - separate expression columns (e.g. log2-ratios) from annotation columns (e.g. protein description)



Data import: Option 2B

Upload local data set stored in text format.


- tab or comma-separated (csv, tsv, txt)
- gct 1.2

Protigy (v0.8.0.5)   karsten@broadinstitute.org  Logout Help me!  

Upload file (txt, csv, gct):

Browse... No file selected

Saved sessions:

Search sessions 

Import **Manage sessions**

PROTigy

PROteomics Toolset for InteGrative Data Analysis

This Shiny app facilitates exploratory and interactive analysis of data sets derived from quantitative *proteomics* experiments, *RNA-seq* and gene expression *microarrays*.

The app can run locally on your Desktop computer (Windows/Linux/MAC) or can be deployed to Shiny Server environments. To access all implemented features the app has to run on a [Shiny Server Pro \(SSP\)](#) instance, see below for a summary of features only available in SSP.

Supported input formats:

- Any type of text file containing both, expression and annotation columns, can directly be imported into the app.
- Supported file formats:
 - text files (tsv, csv, txt)
 - gct 1.2
 - **NEW** gct 1.3

Data import: Option 2B - select unique identifiers

Protigy (v0.8.0.6)

Export experimental design file.

Export

Next

Choose ID column

- ☒ id
- ☐ LumosMS20p4..126.130 ...
- ☐ LumosMS20p4..126.130 ...
- ☐ LumosMS20p4..126.131 ...
- ☐ LumosMS20p4..127N.12 ...
- ☐ LumosMS20p4..127N.13 ...
- ☐ LumosMS20p4..127N.13 ...
- ☐ LumosMS20p4..127N.13 ...
- ☐ LumosMS20p4..127C.12 ...
- ☐ LumosMS20p4..127C.13 ...
- ☐ LumosMS20p4..127C.13 ...
- ☐ LumosMS20p4..127C.13 ...
- ☐ LumosMS20p4..127C.13 ...
- ☐ LumosMS20p4..128N.12 ...
- ☐ LumosMS20p4..128N.13 ...
- ☐ LumosMS20p4..128N.13 ...
- ☐ LumosMS20p4..128N.13 ...

Group assignment

Here you can download a template of an experimental design file. You can open this file in Excel and define the groups you want to compare. Replicate measurements have to be grouped under a single name in the 'Experiment'-column. Please don't use special characters, like blanks or any punctuation, when defining these names!

Select ID column

Choose a column from the list on the left that contains **unique** identifiers for the features in the data table. If the enentries are not unique, uniqueness will enforces by appending "_1". Preferably, IDs should be unique protein accession numbers (e.g. NP_073737) or a combination of protein accession and residue number in case of PTM analysis (e.g. NP_073737_S544s_1_1_544_544).

Automatic

If the ID colu symbols. Cu

- human
- mouse (*Mus musculus*)
- rat (*Rattus norvegicus*)
- zebrafish (*Danio rerio*)

Download a template file to define experimental design.

Lists the header line of the uploaded table. Column names containing 'id' are moved to the top of the list.

Data import: Option 2B - define the experimental design

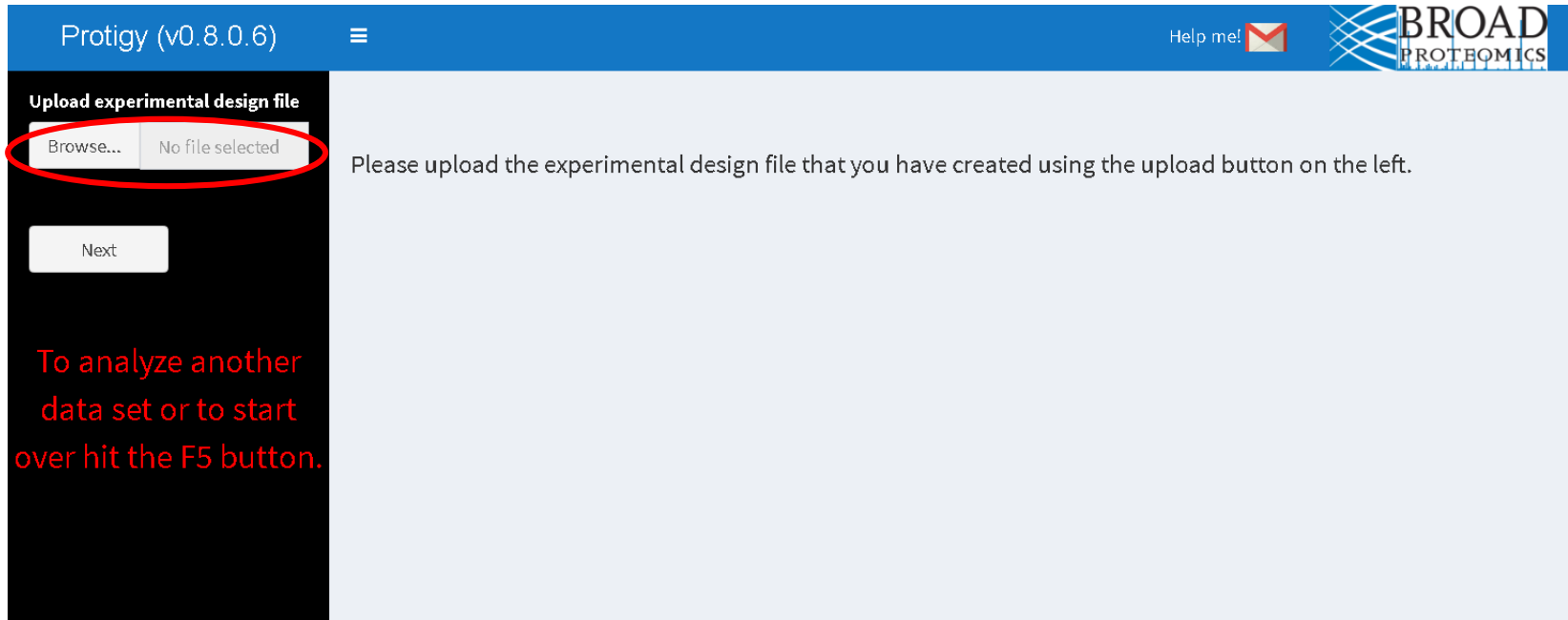
header line
of data table




	A	B
1	Column.Name	Experiment
2	LumosMS20p4..126.130C	T1
3	LumosMS20p4..126.130N	
4	LumosMS20p4..126.131..4	
5	LumosMS20p4..127N.126	
6	LumosMS20p4..127N.130	T2
7	LumosMS20p4..127N.130	
8	LumosMS20p4..127N.131	
9	LumosMS20p4..127C.126	
10	LumosMS20p4..127C.130	T3
11	LumosMS20p4..127C.130	
12	LumosMS20p4..127C.131	
13	LumosMS20p4..128N.126	T4
14	LumosMS20p4..128N.130	
15	LumosMS20p4..128N.130	
16	LumosMS20p4..128N.131	
17	LumosMS20p4..128C.126	
18	LumosMS20p4..128C.130	
19	LumosMS20p4..128C.130	T3
20	LumosMS20p4..128C.131	
21	LumosMS20p4..129N.126	
22	LumosMS20p4..129N.130	
23	LumosMS20p4..129N.130	
24	LumosMS20p4..129N.131	T4
25	LumosMS20p4..129C.126	
26	LumosMS20p4..129C.130	
27	LumosMS20p4..129C.130	T2
28	LumosMS20p4..129C.131	
29	LumosMS20p4..130N.126	
30	LumosMS20p4..130N.130	
31	LumosMS20p4..130N.131	
32	LumosMS20p4..130C.126	
33	LumosMS20p4..130C.130	
34	LumosMS20p4..130C.131	
35	LumosMS20p4..131.126..4	
36	LumosMS20p4..131.130C	
37	LumosMS20p4..131.130N	T1
38	id	
39	accession_number	
40	numColumnsProteinObserved	
41	numSpectraProteinObserved	
42	numPepsUnique	
43	accession_numbers	
44	species	
45	protein_mw	
46	subunitNum	

- Assign phenotype labels to expression columns
- Group replicate measurements under same label
- Columns without label will get carried through the analysis as meta-data columns
- Save as tab-delimited text file

Row-meta data

Data import : Option 2B - upload experimental design file



Protigy (v0.8.0.6)  [Help me!](#)  

Upload experimental design file

Browse... No file selected

Next

To analyze another data set or to start over hit the F5 button.

Please upload the experimental design file that you have created using the upload button on the left.