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:

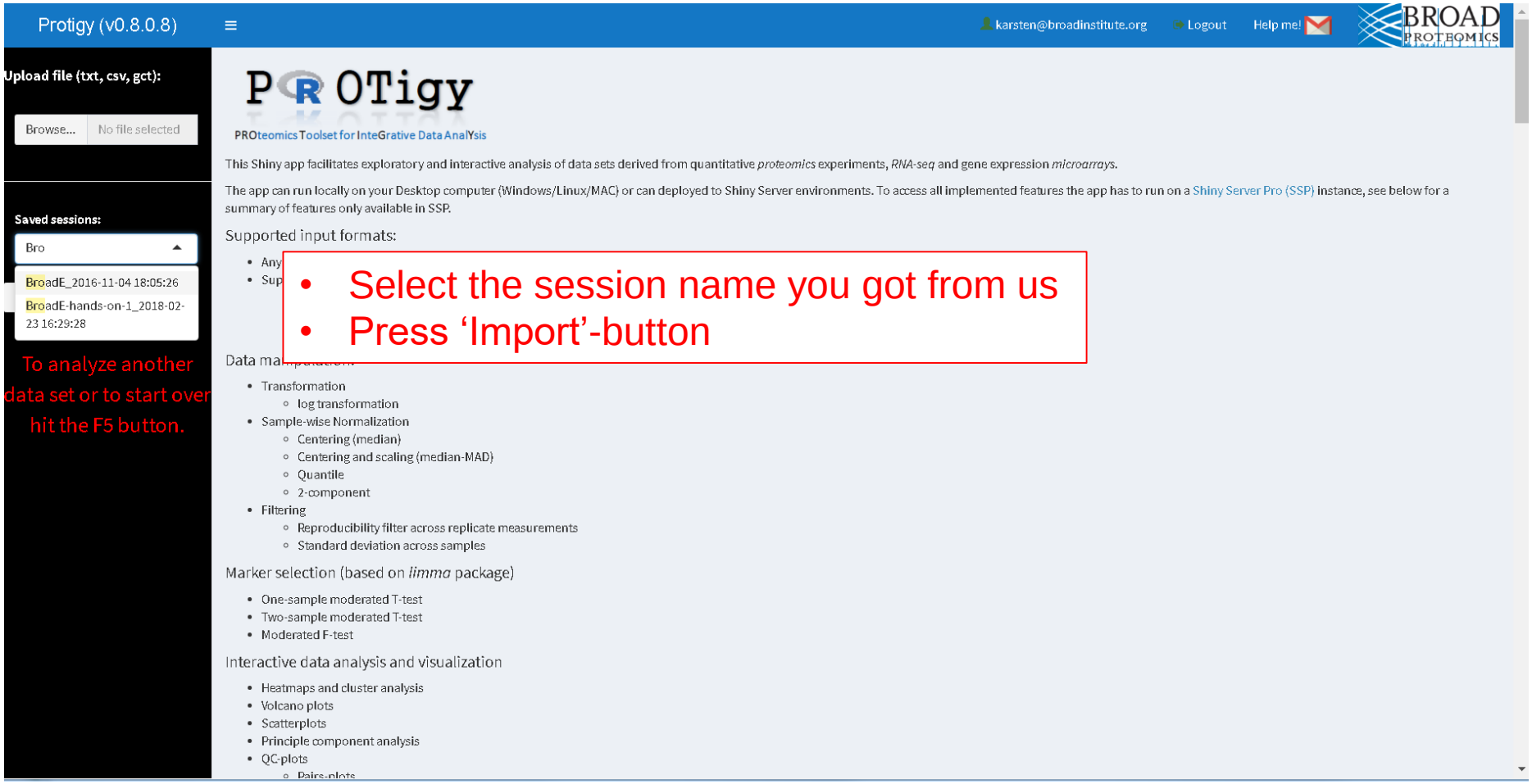
- just go ahead and upload the data file

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/



Data import: Option 2A (GCT 1.3)

GCT1.3 files enable storing of 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	pr_gene_symbol	pr_is_lmark	LPROT001_A375_6H_X1_B20:B03	LPROT001_A375_6H_X1_B20:B05	LPROT001_A375_6H_X1_B20:B07	LPROT001_A375_6H_X1_B20:B09	LPROT001_A375_6H_X1_B20:B11	LPROT001_A375_6H_X1_B20:B13
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	6	6	6	6	6	6
8	pert_dose	-666	-666	-666	-666	2	2	2	2
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

column metadata

rids row metadata data matrix

https://clue.io/connectopedia/gct_format

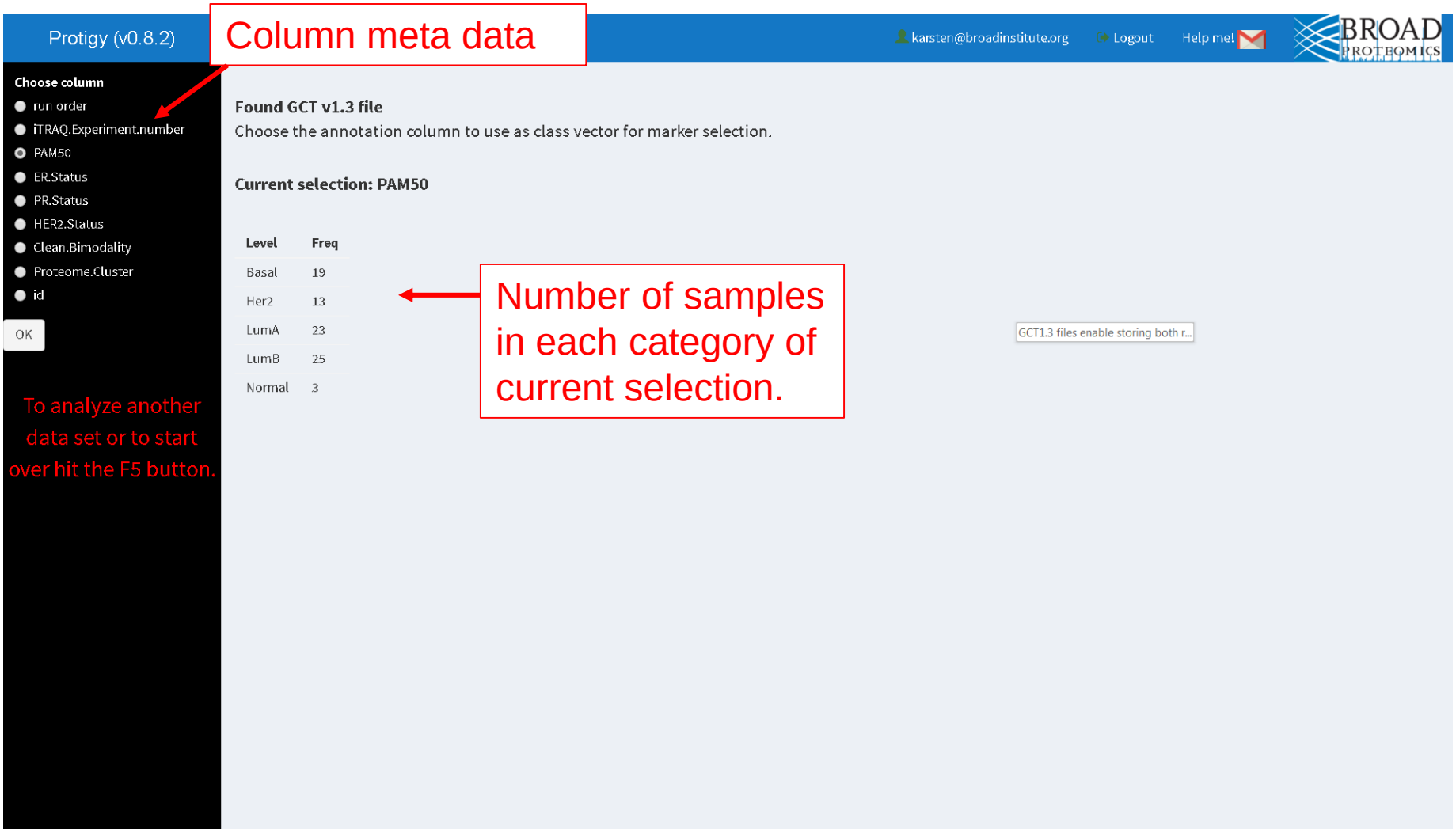
Data import: Option 2A (GCT 1.3)

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

https://clue.io/connectopedia/gct_format

Data import: Option 2A (GCT 1.3)

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



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:

of samples

Third column onwards are sample names. These must be UNIQUE

Always "#1.2"

The # of rows (i.e probe sets)

Microsoft Excel - ExampleA_dataset_HG_U133A

	A	B	C	D	E	F	G
1	#1.2						
2	1000	130					
3	NAME	Description	DLBCL 205	DLBCL 206	DLBCL 232	DLBCL 239	DLBCL 240
4	1007_s_at	U48705 /FEATURE=mRNA /DEFINITION=HS	280.53	271.48	113.57	124.91	124.91
5	1053_at	M87338 /FEATURE= /DEFINITION=HUMA1S	32.13	91.6	117.43	41.29	33.66
6	117_at	X51757 /FEATURE=cds /DEFINITION=HSP7	51.27	61.12	24.1	41.44	43.56
7	121_at	X69699 /FEATURE= /DEFINITION=HSPAX8A	738.32	330.59	249.89	394.55	329.55
8	1255_g_at	L36861 /FEATURE=expanded_cds /DEFINITI	88.45	12.94	18.46	29.96	39
9	1294_at	L13852 /FEATURE= /DEFINITION=HUME1UR	85.57	88.06	62.24	96.59	81.01
10	1316_at	X55005 /FEATURE=mRNA /DEFINITION=HS	106.87	45.11	30.05	46.65	36.5
11	1320_at	X79510 /FEATURE=cds /DEFINITION=HSPTF	58.49	27.95	17.6	27.87	26.52
12	1405_i_at	M21121 /FEATURE= /DEFINITION=HUMTCS	10.83	135.24	13.43	203.16	85.74
13	1431_at	J02843 /FEATURE=cds /DEFINITION=HUMC	41.88	24.09	16.07	26.68	25.4
14	1438_at	X75208 /FEATURE=cds /DEFINITION=HSPT	80.87	9.77	15.33	11.18	44.59
15	1487_at	L38487 /FEATURE=mRNA /DEFINITION=HU	64.26	80.61	102.9	59.77	105.72
16	1494_f_at	M33318 /FEATURE=mRNA /DEFINITION=HU	213.37	96.88	65.06	96.14	78.77
17	1598_g_at	L13720 /FEATURE= /DEFINITION=HUMGAS	458.88	215.59	186.72	187.36	237.69
18	160020_at	Z48481 /FEATURE=cds /DEFINITION=HSMM	411.94	171.16	130	234.76	266.96
19	1729_at	L41690 /FEATURE= /DEFINITION=HUMTRAD	81.59	83.94	74.75	110.9	126.98
20	1773_at	L00635 /FEATURE= /DEFINITION=HUMFPTE	62.82	45.96	41.15	23.1	28.41
21	177_at	U38545 /FEATURE= /DEFINITION=HSU3854	57.04	28.05	16.74	29.66	53.29
22	179_at	U38980 /FEATURE= /DEFINITION=U38980 H	333.96	254.15	241.24	350.58	193.53

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).

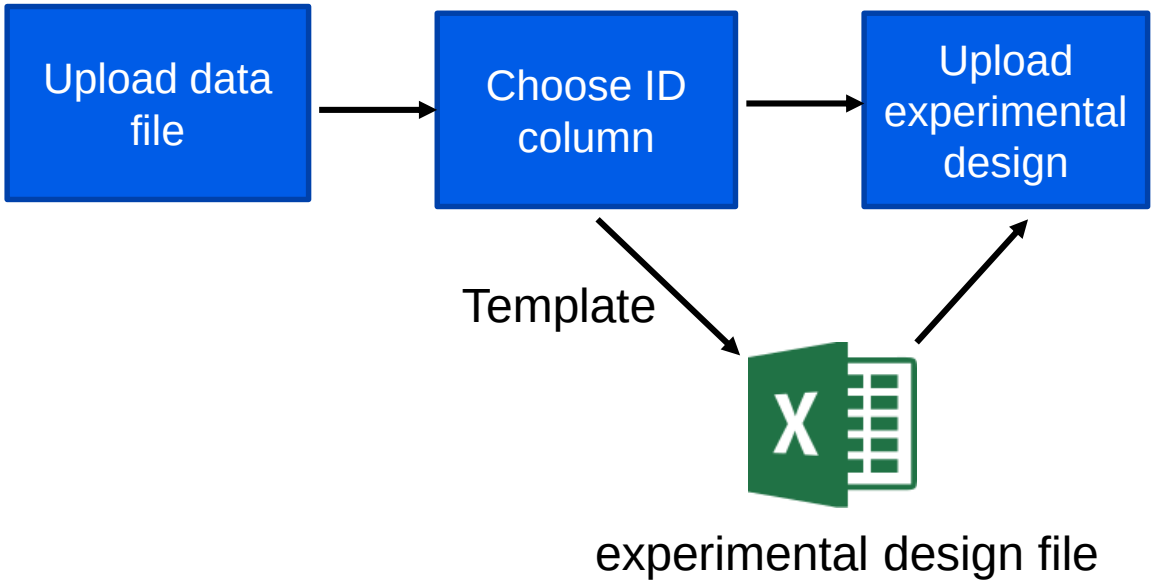
If editing, in Excel, make sure to save your data as "tab delimited text"

https://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats#GCT:_Gene_Cluster_Text_file_format_.28.2A.gct.29

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)

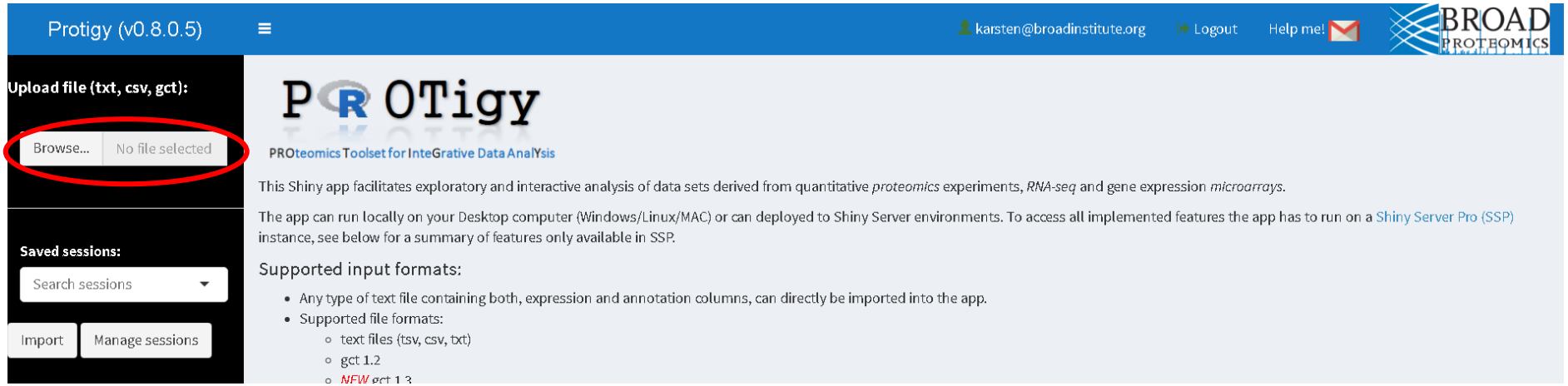
- Requires upload of an ‘experimental design’-file in order to:
  - separate expression columns (e.g. log2-ratios) from annotation columns (e.g. protein description)



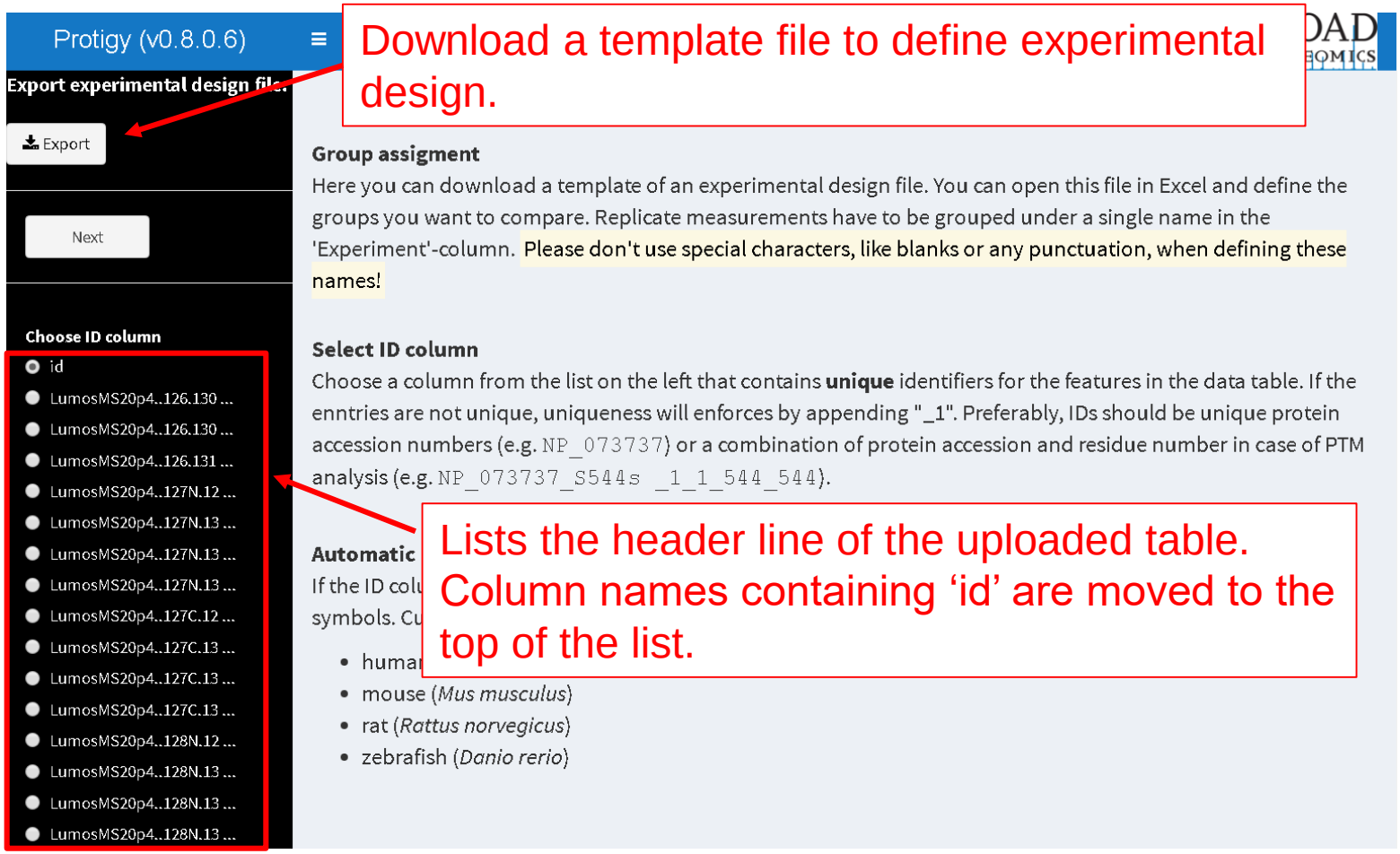
Data import: Option 2B

Upload local data set stored in text format.

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



Data import: Option 2B - select unique identifiers



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	subgroupNum	

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

Row-meta data

Data import : Option 2B - upload experimental design file