

# **NEU Proteomics Capstone Module: Hack-a-thon 1: Pick Your Own Probe Set**

You guys!

Everything you need:

[bit.ly/PCCSEData](http://bit.ly/PCCSEData)

The screenshot shows the PanoramaWeb interface for the LINCS PCCSE Overview. The top navigation bar includes the LINCS logo and the title "LINCS PCCSE Overview". A sidebar on the left contains a "LINCS" menu with a tree view of the data structure. The main content area displays a list of items under the "LINCS" heading, including "P100", "GCP", "LINCS\_Phase\_I", "P100 Template", "GCP Template", "Overview Information", "DIAdev", "AutoQCdev", "Abelin et al - Supplemental Skyline Documents", "Similarity and Connectivity Matrices", and "NEU Capstone Course 2017". A "SITORY" section on the right is partially visible. At the bottom, a timeline labeled "Time Resolved" shows three stages: "Acute stimulation (small molecule)", "P100", and "RNA". A yellow lightning bolt icon is positioned below the "Acute stimulation" label.

**Panorama** PanoramaWeb

**LINCS** LINCS PCCSE Overview

**LINCS**

- LINCS**
  - P100
    - GCP
  - LINCS\_Phase\_I
  - P100 Template
  - GCP Template
  - Overview Information
  - DIAdev
  - AutoQCdev
  - Abelin et al - Supplemental Skyline Documents
  - Similarity and Connectivity Matrices
  - NEU Capstone Course 2017

**SITORY**

nalizing and I  
tions (drug  
modificatio

**Acute stimulation**  
(small molecule)

**P100**

**RNA**

**Time Resolved**

# Goal for the hack-a-thon

- Pick a set of probes for your own P100 assay starting from the discovery proteomics data
  - Pick ~100 probes
  - For now, probes are at the level of the phosphosite, not the peptide itself
- Select a guiding principle for how you will pick probes
- Use R, Excel, or whatever you feel comfortable work with the data
- Provide a list for comparison with the other groups, and the original P100 probes

# What you will need

- Source data
  - 3 versions for your consideration:
    - Unfiltered (every site detected, even if only in 1 experiment)
    - Filtered (only sites detected in 75% of experiments are present)
    - Imputed (filtered + fill in missing values based on normal distribution)
- A method of reading source data
  - R, Excel
- A method of grouping and/or choosing probes
- A method for writing a “csv” file in a standard format

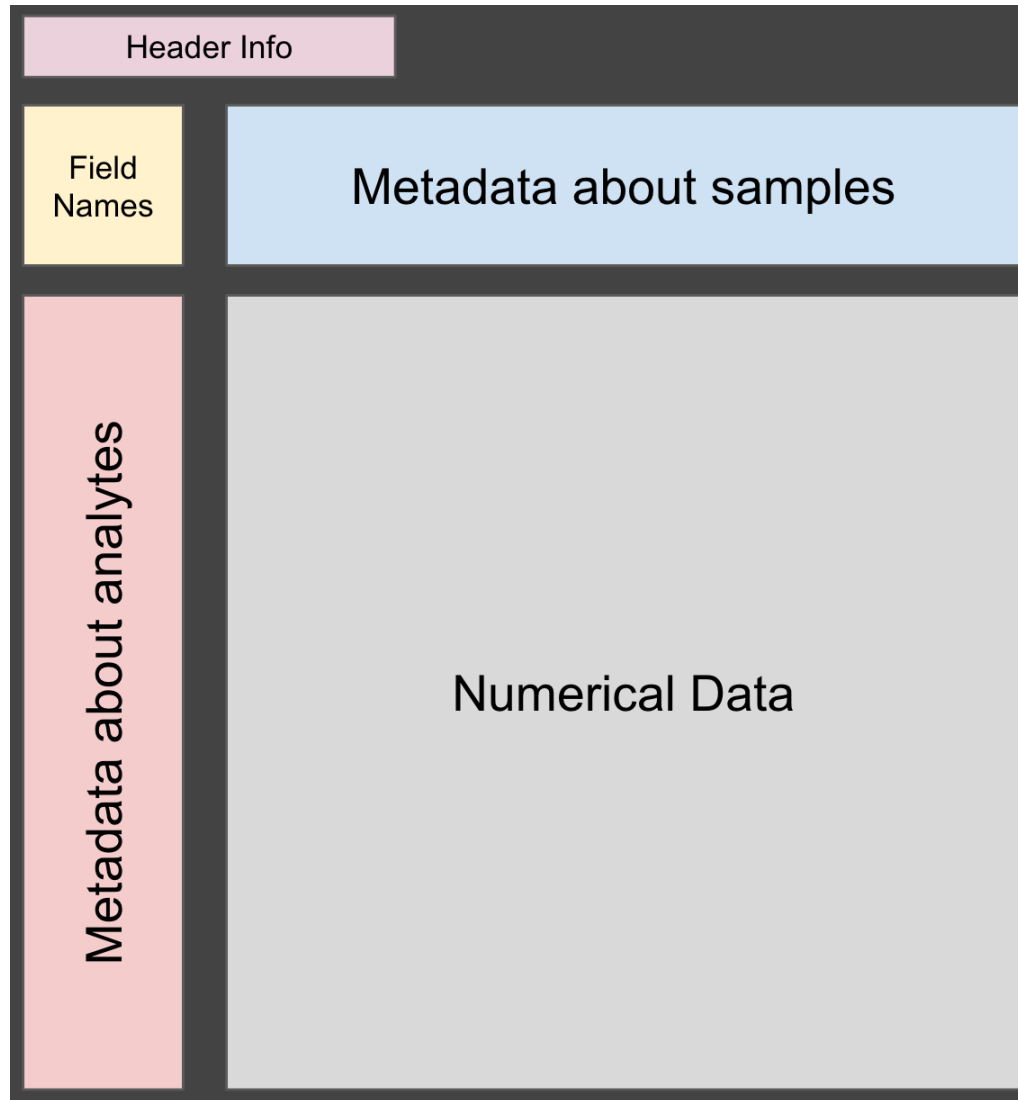
# Evaluation Metrics

- Evaluation of separation of replicates from non-replicates for each team
- Comparison to current P100 probes
- Comparison among all teams' solutions

## Additional notes on the data

- All data are ratios of treatment to DMSO control
- Ratios have been  $\log_2$  transformed
- Consider row median normalizing and/or z-scoring

# Data file format: GCT



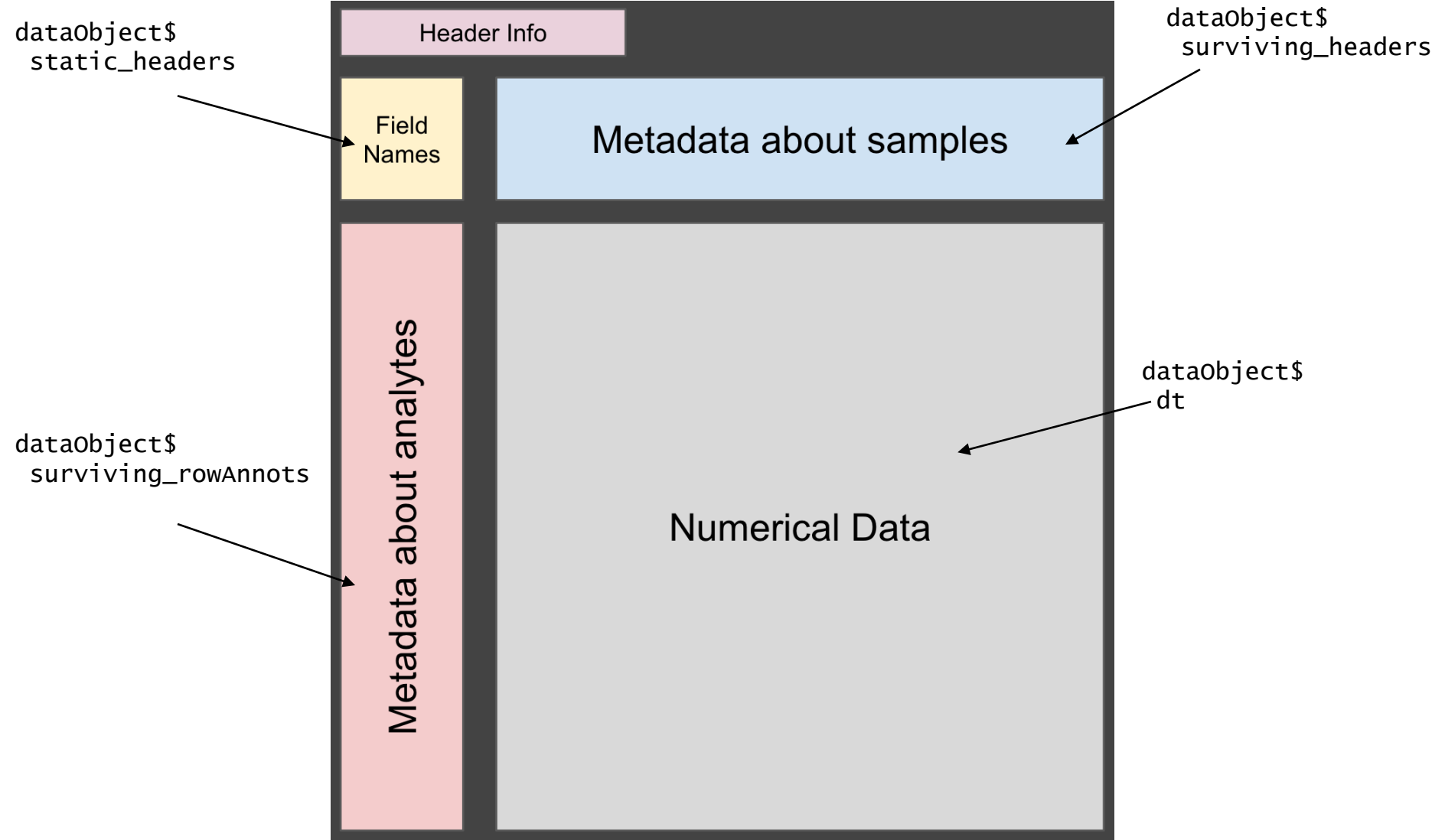
- This is a tab-delimited text file; you can drag/drop into Excel

## Reading the data using the P100 code-base in R

```
#download or copy the R code and source data to a directory on  
your machine  
#create a blank Rstudio project in this directory (or setwd to  
here if just using R)  
  
install.packages('jsonlite') #if not installed already  
  
source('p100_processing.R')  
  
dataObject<-  
P100provideGCTlistObjectFromFile('p100_sourcedata_...gct')
```



# Reading the data using the P100 code-base in R



# Explanation of the dataObject

- dataObject\$dt – matrix data. colnames() are sample IDs, rownames() are probe IDs.

- Sample ID format:

*cellLine\_treatment\_repeat#*            i.e., MCF7\_captopril\_1

- Probe ID format:

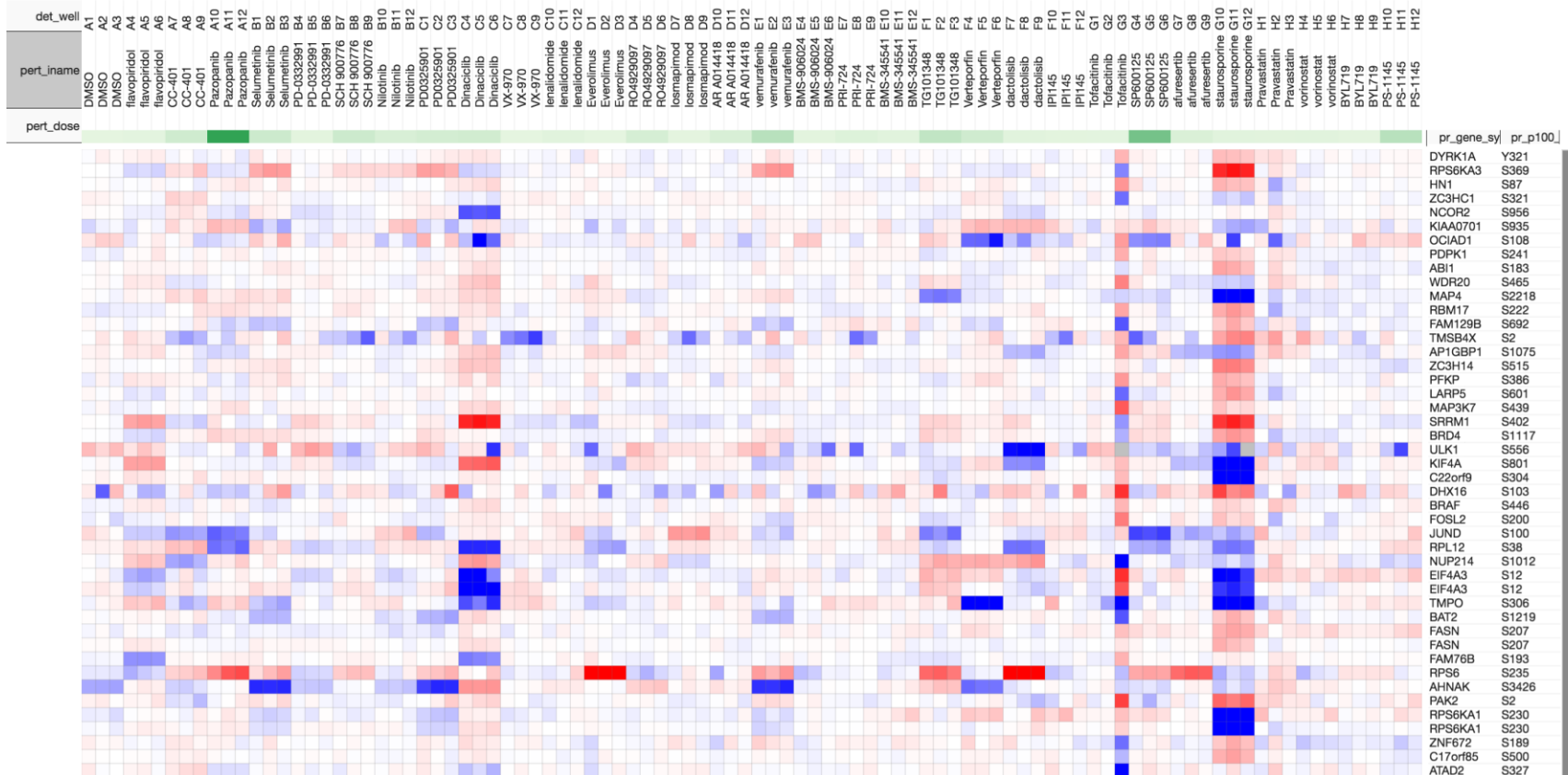
*Id#\_genename\_phosphosites*            i.e., 1234\_RPS6\_S123  
(meaning RPS6 phosphorylated on Ser123).

# A tool to visualize GCT and other matrix data

## ■ Morpheus:

<https://software.broadinstitute.org/morpheus/>

(quick tour)



## Desired output format (.csv file):

probes

495\_HDAC2\_S394

538\_ESYT1\_S830

569\_CDA02\_S506

580\_API5\_S464

617\_FUSIP1\_S133

620\_FUSIP1\_S131

622\_KIAA0055\_S718

656\_DNAJB6\_S277

684\_NSUN2\_S743

685\_NSUN2\_S751

745\_EIF3C\_S39

774\_hCG\_32198\_S811

796\_KIAA1823\_S155

800\_KIAA1823\_S199

824\_C2orf49\_S189

848\_PRPSAP2\_S227

914\_BAT2D1\_S2107

928\_EML3\_S177

...

# Reading the data using the P100 code-base in R

```
source('p100_processing.R')

dataObject<-
P100provideGCTlistObjectFromFile('p100_sourcedata_....gct')

numericData<-dataObject$dt

myProbeData<-doSomethingToPickProbes(numericData)
  #return a matrix with fewer rows than input matrix

myProbeList<-list(probes=rownames(myProbeData))

write.csv(myProbeList,file='probelist.csv',row.names=FALSE)
```

# Some ideas to get you started...

- Clustering
  - There are many flavors...
- High variance probes
- Marker selection
  - I.e., pick probes that typify response to certain drugs
- Use prior knowledge
  - I.e., I know what that gene does
- Pick by eye
- Pick randomly