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

You guys!

Everything you need:

bit.ly/PCCSEData

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

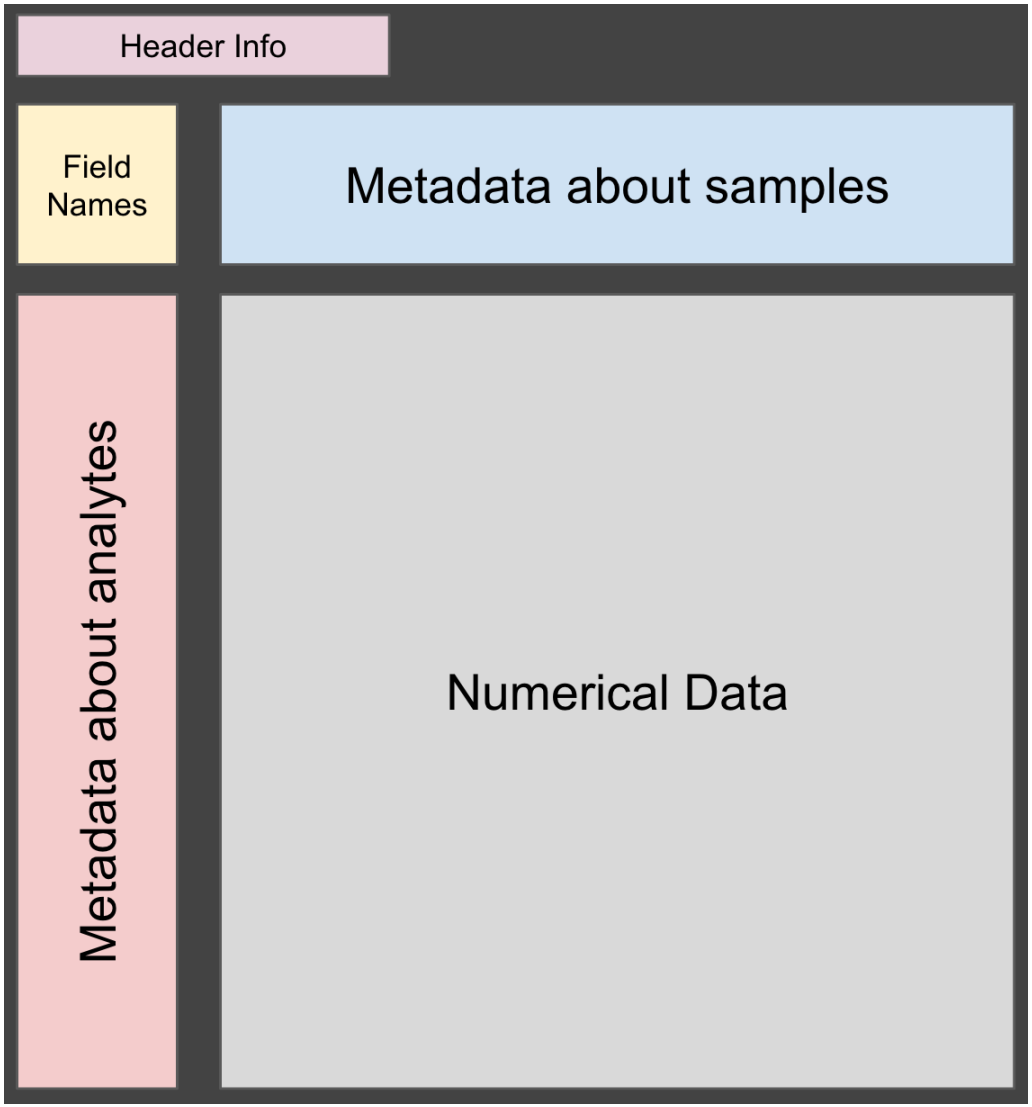
Evaluation Metrics

- 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

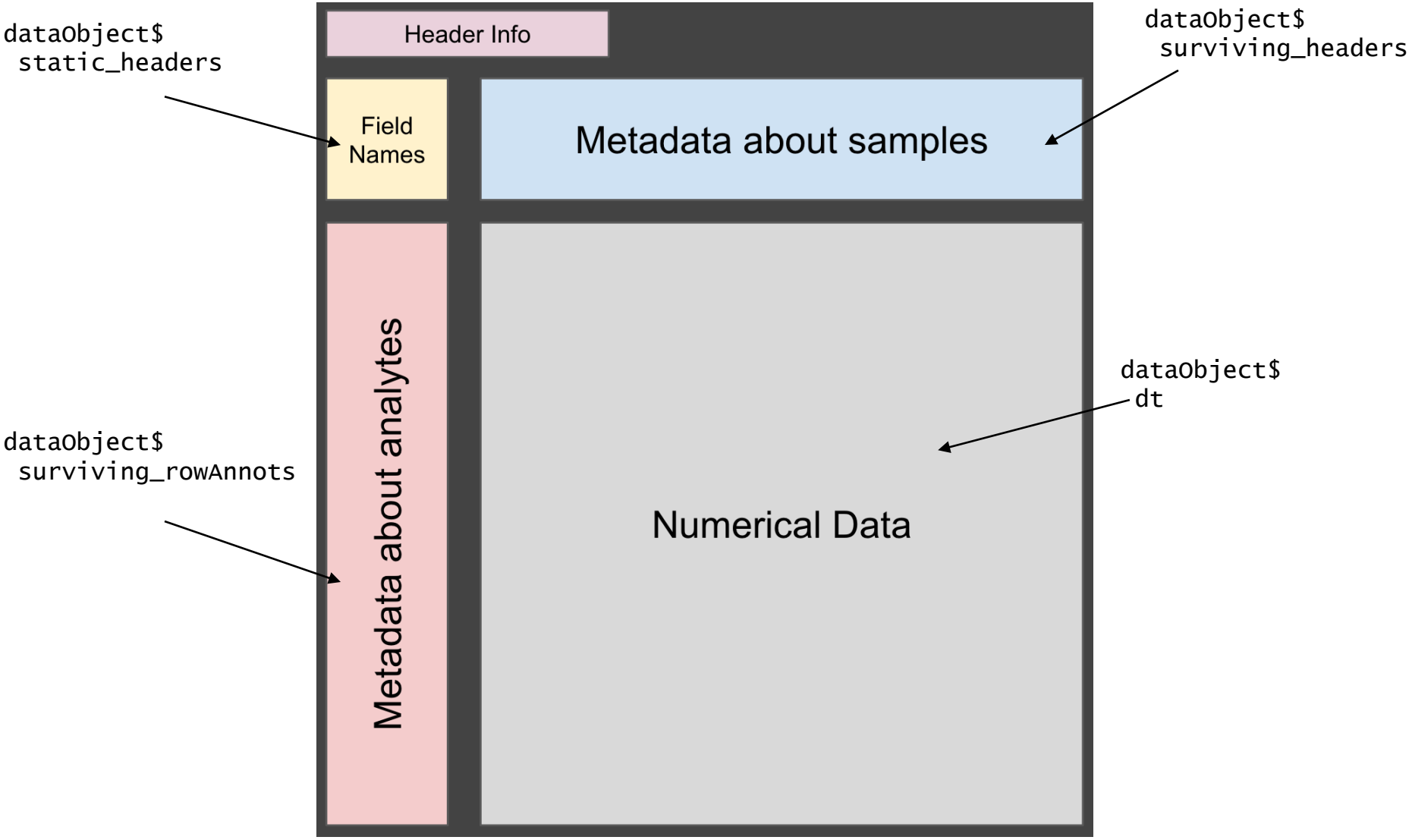
```
#download or copy the R code and source data to a directory on
your machine
#create a blank Rstudio project in this directory (or setwd to
here if just using R)

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`

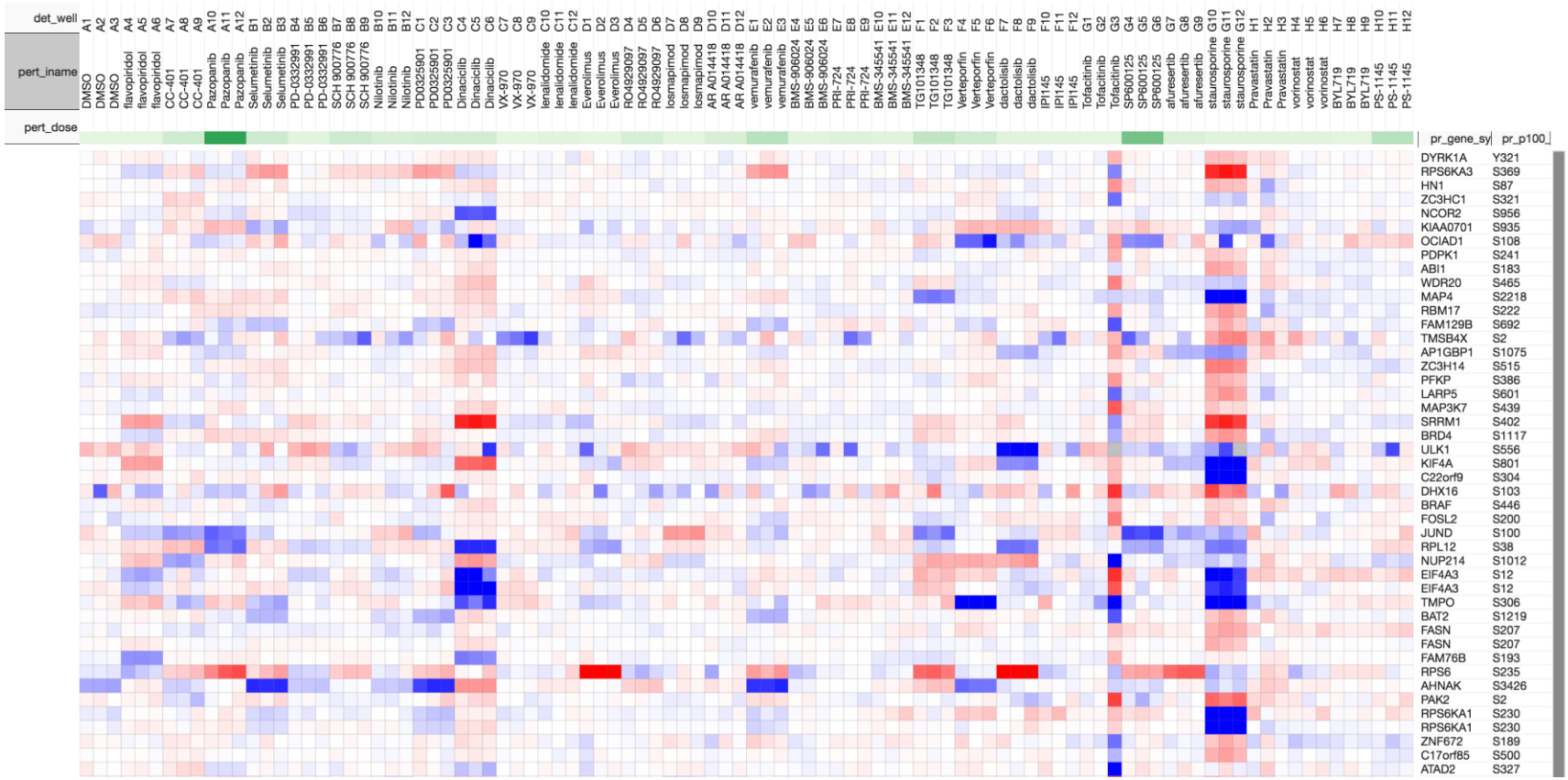
  (meaning RPS6 phosphorylated on Ser123).

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