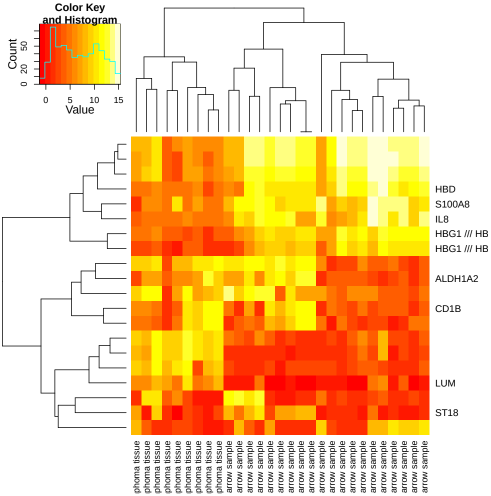
Introduction to the ExpressionSet

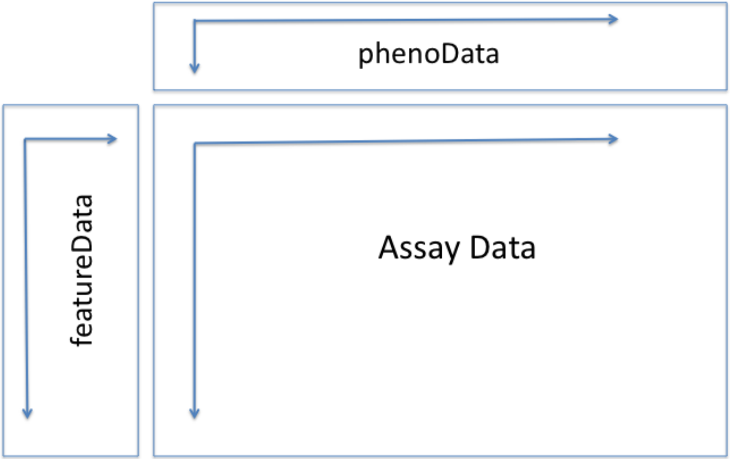
Sean Davis

February 3, 2014

Our Goal


Color Key
and Histogram
Count
0
20
50
0
5
10
15
Value
HBD
S100A8
IL8
HBG1 /// HB
HBG1 /// HB
ALDH1A2
CD1B
LUM
ST18
phoma tissue
phoma tissue
phoma tissue
phoma tissue
phoma tissue
phoma tissue
phoma tissue
phoma tissue
phoma tissue
arrow sample
arrow sample
arrow sample
arrow sample
arrow sample
arrow sample
arrow sample
arrow sample
arrow sample
arrow sample
arrow sample
arrow sample
arrow sample
arrow sample
arrow sample
arrow sample
arrow sample
arrow sample
arrow sample
arrow sample

The Genomics Table Triumvirate



Gene information–featureData

```
head(fData(eset))
```

```
##           Gene Symbol ENTREZ_GENE_ID
## 1007_s_at        DDR1            780
## 1053_at          RFC2           5982
## 117_at          HSPA6           3310
## 121_at           PAX8           7849
## 1255_g_at      GUCA1A           2978
## 1294_at          UBA7           7318
##                                           RefSeq Transcript ID
## 1007_s_at              NM_001954 /// NM_013993 /// NM_013994
## 1053_at                                NM_002914 /// NM_181471
## 117_at                                               NM_002155
## 121_at    NM_003466 /// NM_013952 /// NM_013953 /// NM_013992
## 1255_g_at                                            NM_000409
## 1294_at                                              NM_003335
##                                             Gene Title
## 1007_s_at discoidin domain receptor tyrosine kinase 1
## 1053_at   replication factor C (activator 1) 2, 40kDa
## 117_at           heat shock 70kDa protein 6 (HSP70B')
## 121_at                                   paired box 8
```

Sample Information–phenoData

```
head(pData(eset))
```

```
##              title geo_accession submission_date type channel_count
## GSM27065  T-LL-1      GSM27065     Jul 20 2004   RNA             1
## GSM27066  T-LL-2      GSM27066     Jul 20 2004   RNA             1
## GSM27068  T-LL-3      GSM27068     Jul 20 2004   RNA             1
## GSM27079 T-ALL-1      GSM27079     Jul 20 2004   RNA             1
## GSM27082 T-ALL-2      GSM27082     Jul 20 2004   RNA             1
## GSM27083 T-ALL-3      GSM27083     Jul 20 2004   RNA             1
##              source_name_ch1 molecule_ch1 taxid_ch1
## GSM27065    Lymphoma tissue    total RNA      9606
## GSM27066    Lymphoma tissue    total RNA      9606
## GSM27068    Lymphoma tissue    total RNA      9606
## GSM27079 Bone marrow sample    total RNA      9606
## GSM27082 Bone marrow sample    total RNA      9606
## GSM27083 Bone marrow sample    total RNA      9606
```

Assay Data (expression)–assayData

```
head(assayDataElement(eset, "exprs"))
# OR
head(exprs(eset))
```

```
##             GSM27065 GSM27066 GSM27068 GSM27069 GSM27071 GSM27072 GSM27074
## 1007_s_at      408.1    141.1    283.1     62.3    194.1    173.9    190.2
## 1053_at         28.0     70.8     50.6     68.8     48.5     13.6     41.8
## 117_at          12.0     31.8     36.3     61.5     81.8     55.1     26.9
## 121_at         273.6    255.8    218.9    322.6    199.1    456.0    129.3
## 1255_g_at       23.9      4.8     12.5     22.5     21.3     14.8     15.9
## 1294_at        335.7    112.4    452.3    245.2    524.6    444.1    169.1
##             GSM27075 GSM27077 GSM27079 GSM27082 GSM27083 GSM27085 GSM27087
## 1007_s_at      285.8    536.8    237.5    159.8    212.1    113.2     59.9
## 1053_at         50.5     39.9     33.0    129.9    114.9    183.8     87.4
## 117_at          27.8     37.2     37.1     25.6     54.4     42.8     38.8
## 121_at         274.9    265.6    380.4    341.2    268.8    273.6    233.5
## 1255_g_at        3.0     25.2     21.3     19.5     26.0     17.7      7.7
## 1294_at        220.3    240.0    186.8    263.9    145.0    107.9    142.2
##             GSM27088 GSM27091 GSM27093 GSM27094 GSM27095 GSM27097 GSM27098
## 1007_s_at       59.9     68.4    121.8    195.9     73.8    400.1    171.2
```

The Parts of an ExpressionSet

The Data
typically a *matrix*-like object (or multiple matrices)

The Parts of an ExpressionSet

The Data
typically a *matrix*-like object (or multiple matrices)

The Sample Annotations
typically a dataframe-like object

The Parts of an ExpressionSet

The Data
: typically a *matrix*-like object (or multiple matrices)

The Sample Annotations
: typically a dataframe-like object

The Feature Annotation (Gene information)
: typically a dataframe-like object

An Example ExpressionSet

```
browseURL("http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE
library(GEOquery)
eset = getGEO("GSE1577")[[1]]

## Found 1 file(s)
## GSE1577_series_matrix.txt.gz
## Using locally cached version:
/var/folders/21/8t47kwys6vqb8606kdfn71780000gn/T//RtmpXcLYOS/GSE
## Using locally cached version of GPL96 found here:
##
/var/folders/21/8t47kwys6vqb8606kdfn71780000gn/T//RtmpXcLYOS/GPL
```

An Example ExpressionSet

```
class(eset)

## [1] "ExpressionSet"
## attr(,"package")
## [1] "Biobase"
```

```
help(ExpressionSet)
help("ExpressionSet-class")
```

An Example ExpressionSet

```
show(eset)

## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 22283 features, 29 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: GSM27065 GSM27066 ... GSM27109 (29 total)
##   varLabels: title geo_accession ... data_row_count (25 total)
##   varMetadata: labelDescription
## featureData
##   featureNames: 1007_s_at 1053_at ... NA.13868 (22283 total)
##   fvarLabels: ID GB_ACC ... Gene Ontology Molecular Function (16
##     total)
##   fvarMetadata: Column Description labelDescription
## experimentData: use 'experimentData(object)'
## Annotation: GPL96
```

An Example ExpressionSet

```
# dimensions of the ExpressionSet
dim(eset)

## Features  Samples
##    22283       29

# Number of features
nrow(eset)

## Features
##    22283

# number of samples
ncol(eset)

## Samples
##      29
```

An Example ExpressionSet

```
# 'names' of samples--must be unique
sampleNames(eset)[1:8]

## [1] "GSM27065" "GSM27066" "GSM27068" "GSM27069" "GSM27071" "GSM27072"
## [7] "GSM27074" "GSM27075"

# 'names' of features--must be unique
featureNames(eset)[1:8]

## [1] "1007_s_at" "1053_at"   "117_at"    "121_at"    "1255_g_at" "1294_at"
## [7] "1316_at"   "1320_at"
```

Sample Information–phenoData

```
head(pData(eset))
class(pData(eset))
colnames(pData(eset))
all.equal(sampleNames(eset), rownames(pData(eset)))
summary(pData(eset))
```

Gene information–featureData

```
head(fData(eset))
class(fData(eset))
colnames(fData(eset))
all.equal(featureNames(eset), rownames(fData(eset)))
summary(fData(eset))
```

Assay Data (expression)–assayData

```
head(assayDataElement(eset, "exprs"))
# OR
head(exprs(eset))
class(exprs(eset))
summary(exprs(eset))
all.equal(colnames(exprs(eset)), sampleNames(eset))
all.equal(rownames(exprs(eset)), featureNames(eset))
```

Subsetting ExpressionSets

- Subsetting works similarly to data.frames or matrices
- Columns represent samples
- Rows represent genes (or features)

```
eset[1:10, ]
eset[, 1:10]
```

subsetting ExpressionSets

How can we subset our ExpressionSet to include only the “Bone Marrow” samples?

subsetting ExpressionSets

How can we subset our ExpressionSet to include only the “Bone Marrow” samples?

```
levels(pData(eset)$source_name_ch1)
```

```
## [1] "Bone marrow sample" "Lymphoma tissue"
```

subsetting ExpressionSets

How can we subset our ExpressionSet to include only the “Bone Marrow” samples?

```
levels(pData(eset)$source_name_ch1)

## [1] "Bone marrow sample" "Lymphoma tissue"
```

```
eset[, pData(eset)$source_name_ch1 == "Bone marrow sample"]

## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 22283 features, 20 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: GSM27079 GSM27082 ... GSM27109 (20 total)
##   varLabels: title geo_accession ... data_row_count (25 total)
##   varMetadata: labelDescription
## featureData
##   featureNames: 1007_s_at 1053_at ... NA.13868 (22283 total)
##   fvarLabels: ID GB_ACC ... Gene Ontology Molecular Function (16
##     total)
##   fvarMetadata: Column Description labelDescription
```

Heatmap Preliminaries

- Load the gplots library and look at the heatmap.2 help

```
library(gplots)
`?`(heatmap.2)
```

- The heatmap.2 function takes as a minimal input a matrix of values.

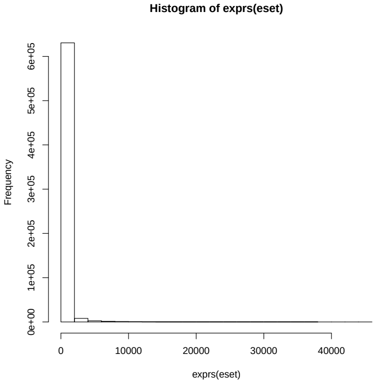
Prepare for heatmap

Are our values in the log space?

Prepare for heatmap

Are our values in the log space?

```r
summary(exprs(eset))
hist(exprs(eset))
```



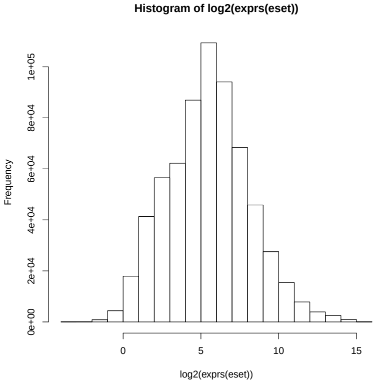
Prepare for heatmap

How can we put the expression values in log2 space?

Prepare for heatmap

How can we put the expression values in log2 space?

```
summary(log2(exprs(eset)))
hist(log2(exprs(eset)))
```



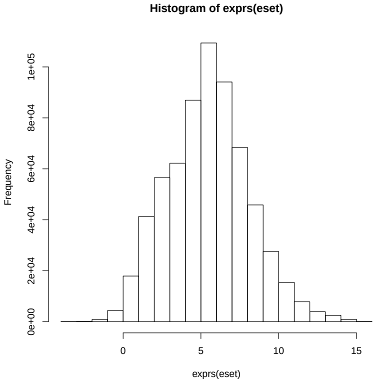
Prepare for heatmap

How can we replace the expression values in our ExpressionSet with the log2-transformed values?

Prepare for heatmap

How can we replace the expression values in our ExpressionSet with the log2-transformed values?

```
exprs(eset) = log2(exprs(eset))
hist(exprs(eset))
```



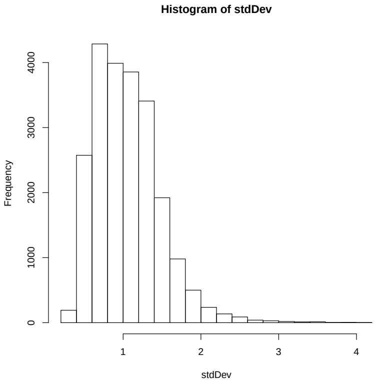
Prepare for heatmap

How can we compute the expression standard deviation for each feature in the ExpressionSet?

Prepare for heatmap

How can we compute the expression standard deviation for each feature in the ExpressionSet?

```
stdDev = apply(exprs(eset), 1, sd)
hist(stdDev)
```



Prepare for heatmap

How can we subset the expression data to include the top 20 most variable (most informative) genes?

Prepare for heatmap

How can we subset the expression data to include the top 20 most variable (most informative) genes?

```
eset2 = eset[order(stdDev, decreasing = TRUE)[1:20], ]
```

Heatmap

```
heatmap.2(exprs(eset2), trace = "none")
```

Challenge Exercises

- Use the pheatmap package to make a more interesting and informative heatmap that includes the sample types as a color bar.
- Make a multidimensional scaling (MDS) plot of the samples using the top 200 most variable genes.
- Use the featureData in the ExpressionSet and the grep function to construct a new ExpressionSet subset that contains transcription factor genes.
- Make a histogram of the Pearson correlation of the correlation between the first gene and all the other genes.