Example 2 A typical day at the computer

Marc Corrales Berjano

December 7, 2016

Contents

- I will make an exaple for someone doing data analysis since I dont know which

are the advantages to 'real developping'.

Purpose :: We will take a look at the expression profiles of 25 Drosophila cell lines.

1 DONE Start a new Docker container

2 TODO Start to track the project in git

- Magit

```
git init
git remote add origin https://github.com/histonemark/retreteando.git
git add *
git commit -m 'Retreteando'
```

3 TODO Get the data

- Lets download the data from GEO.

```
wget https://dl.dropboxusercontent.com/u/3975383/Drosophila_25_cell_lines.txt
ls | grep '^Drosophila'
```

4 TODO Clean and prepare the Data

- How does the data look like

```
head Drosophila_25_cell_lines.txt
```

- Let's pretend that it was not so beautifully R ready and we needed to clean

it a bit and we have to remove 3 'evil' genes 'FBgn0000022','FBgn0000253' and 'FBgn0036608'.

```
import sys

toremove = ['FBgn0000022','FBgn0000253','FBgn0036608']

with open('Drosophila_25_cell_lines.txt') as f:
    sys.stdout.write('%s' % f.readline()) # Header
    for line in f:
        items = line.split()
        gname = items[0]
        if gname in toremove: continue
        sys.stdout.write('%s' % line)
```

M-x toggle-truncate-lines

5 TODO Explore the Data, Plot , statistics

- Lets load the clean data.

```
cells <- read.delim('./Dmel_25Cl_clean.tsv')
dim(cells)
```

- Lets look at it

```
head(cells)
```

- Lets see how the genes cluster.

```
#cells$color <- 'black'
#cells$color[cells$]
PCA <- prcomp(cells[7:29], scale=T) # Dont take RPKMs
plot(PCA$x)
```

How pretty much all the genes seem to have a similar expression level except for a group of 50-100 that seem to drive all the expression in the variance. Blah blah blah.

6 TODO Send a report, publish, export

How do you want to save or show your code/results:

- As a pdf