# Data mining exercise

# An emulation of real life as a data scientist in biology

- In absalon, there is now a data mining challenge, to do in groups, under 'assignments'
- It gives a data set and a vague problem statement from some experimentalists who don't know data mining, statistics or bioinformatics
- You have an experimentalist who knows everything there is to know about the experiment around (me). But he knows no bioinformatics or R whatsoever.
- You have this whole lecture plus at least 1h on the next (Tuesday). The idea is that these 4h will ve enough. I will select some groups to present (make an rmd file, knit to html)
- All groups will submit their solution - I have made an 'assigment' for this. I will make all solutions available to all

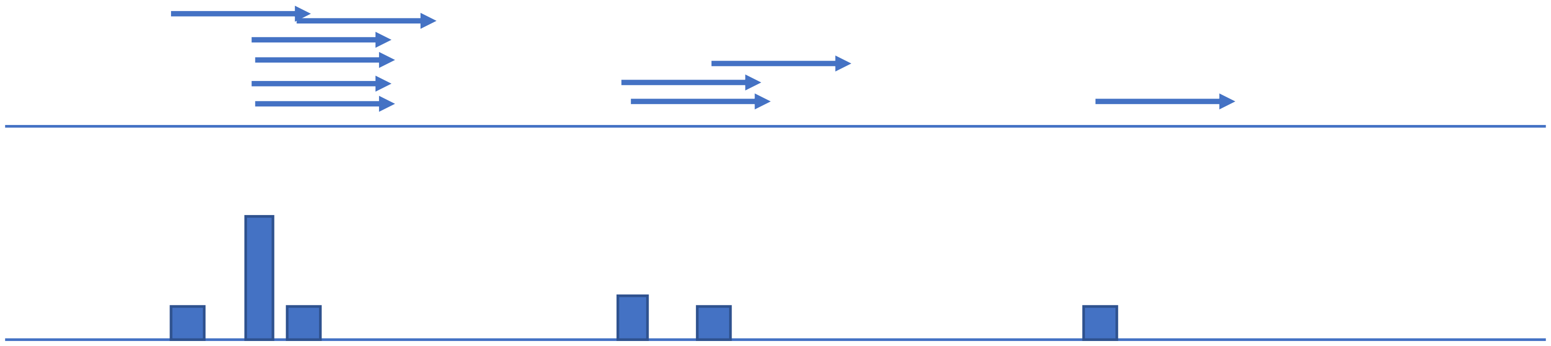
# The data

- CAGE libraries are made for a set of tissues in mouse
- Each row in your file is a CAGE tag cluster (next slide)
- Each row has TPM-normalized expression values for each tissue, and and ID and mm8 locations (and strand)

# CAGE tag cluster

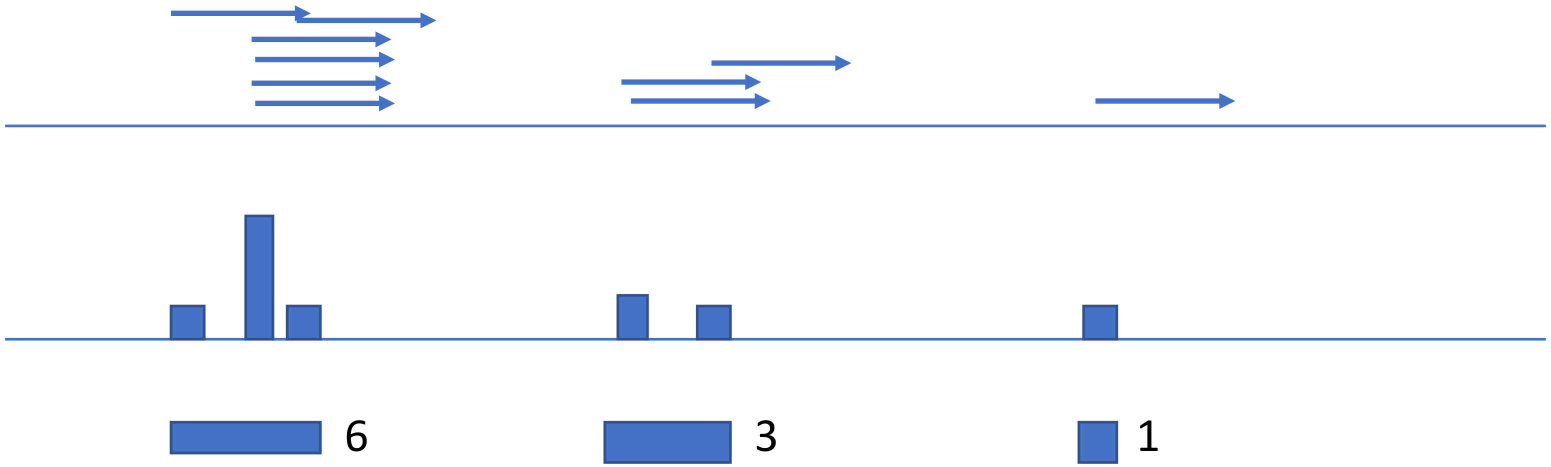
Reminder: CAGE tags are, effectivley, the first 30 bp of RNAs from the 5' end

Each CAGE tag is mapped to the genome – if you would visualize this in the genome browser, it would look like a 'barplot' on the genome where we count the occurence s of 5' ends of tags, like this:



# CAGE tag cluster

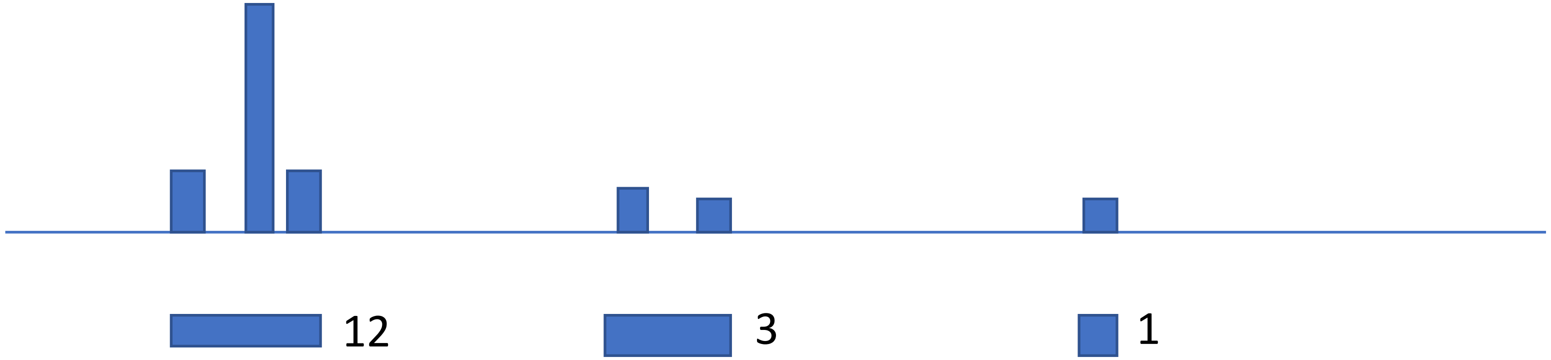
It is helpful to ‘clump’ together tags that map close to one another, like this



These CAGE tag clusters have a start, end and a strand, and also a count of CAGE tags

In our data, clusters can have CAGE tags from several tissues

In our data, clusters can have CAGE tags from several tissues



10 liver
2 lung

0 liver
1 lung

1 brain

So, aside from location and strand, each cluster will have a row in an expression matrix that says how many tags from each tissue

| | liver | lung | brain |
|---|---|---|---|
| Cluster1 | 10 | 2 | 0 |
| Cluster2 | 0 | 1 | 0 |

As a last step, such counts are normalized, becasue the library sizes are different (by tags-per-million)

# So, again: the data

- CAGE libraries are made for a set of tissues in mouse
- Each row in your file is a CAGE tag cluster
- Each row has TPM-normalized expression values for each tissue, and and ID and mm8 locations (and strand)

# Tips

- This is means to emulate a real-life setting
- Make use of the experimentalist that knows everything about the experimt (=me)
- This person also sort knows what he wants, but cannot explain it in bioinformatics terms.
- In most real data science projects – including this one, unlikel home works etc the challenge is not *how* to do analyses but figuring out *what questions to ask.* Which is really what the core of science is.

# Schedule

- We will work in the home work groups the rest of the lecture (we will not return to this room until next lecture).
- The experimentalist will walk between groups randomly, but you can send him a text if you need assistance (22456668)
- At Tuesday, we will start in the home work groups and you will have one hour to finish, and then we meet in the big lecture room at 1400.