Data mining exercise

Introduction

Your collaborator has made a CAGE experiment using 7 different tissues. As a remainder, CAGE tags are 20-21 nt long tags mapping to the genome. We often cluster CAGE tags that are close to one another on the genome to a "tag cluster". A tag cluster can then have several tags from one or several tissues.

One can view these CAGE tag clusters as being "core promoters" in the sense that they are measuring the activity and location of a core promoter.

The data file `htbinf_cage_tpms` shows the CAGE tag clusters as rows, and tissues as columns. The cell values are the TPMs from the given tissue in the cluster. There are three additional columns: the tag cluster ID, the location of the cluster in mm8 and the strand of the cluster. The collaborator now wants to know:

1. How many types of core promoters are there in terms of tissue expression patterns, and what expression patterns are these?

2. What tissues are similar to each other in terms of promoter usage? They would really like to have this as a picture and not just "values"

3. How many tissue-specific promoters are there, per tissue and just allover? It would be very helpful to calculate two "specificity scores": one for each tissue, and one summary score across all tissues for each promoter.

4. They want to have genome browser examples of the most tissue-specific promoter for each tissue

5 Lastly, they want a list with the 10 most tissue-specific promoters, taking all tissues into account.