Institut für Informatik

Praktische Informatik und Bioinformatik
Prof. Dr. Ralf Zimmer
Dr. Gergely Csaba
Dr. Florian Erhard

Wintersemester 2015/2016

GoBi: Excercise 4 (Bonus)

Differential Expression

Deadline: Thursday, 21.01.2016, 14:00

Save your solution to /home/proj/biocluster/genprakt/${stud_account}/Solution4. Also provide an executable jar file (containing the sources) in this directory that allows to reproduce your results. The jar should print a usage info if invoked without parameters.

Bonus Task 1 (In-depth evaluation of simulated data):

We applied the mapper methods star, tophat2, gsnap and contextmap on the simulated data, and extracted the gene counts using the HTseq library:

`http://www-huber.embl.de/HTSeq/doc/overview.html`

using all possible combinations of intersection-nonempty, intersection-strict, union and gene, transcript, exon, CDS see: `http://www-huber.embl.de/HTSeq/doc/count.html#count`.

You find the htseq-count results under:

```
data/diff_simulation/(f1,f2)/(r1,r2,r3)/(star,contextmap,gsnap,tophat2).counts/*.counts
```

Combine the htseq-count results and create input for EnrichmentBrowser for all combinations, evaluate the results, and create a html report for the evaluation.

The report should include:

- an overall ranking over all combinations (mapper, count extracting method, DE-method) along with the corresponding statistics from Task 1 and links to detail views (the plots from Task 1 and a table of not called differential genes with their properties (p-value, foldchange)).
- a comparison of mappers: barplot + table. Select for each mapper the overall best extraction mode,DE method combination, and compare these (and provide the properties of this selection).
- a comparison of extracting methods: barplot + table. Select for each extracting method the

overall best mapper,DE method combination, and compare these (and provide the properties of this selection).

- a comparison of DE-methods: barplot + table. Select for each DE-method the overall best mapper,extracor combination, and compare these (and provide the properties of this selection).

Bonus Task 2 (In-depth evaluation of your feature count program):

Compute gene counts with your program from the previus assignment with the following settings

- NRP per gene
- gene-unique NRP per gene (i.e. RP-s not overlapping multiple genes)
- transcriptome mapping NRP per gene
- gene-unique transcriptome mapping NRP per gene

Combine the extracted counts to EnrichmentBrowser's input and apply all three DE-methods, and update the evaluation from Bonus Task 1!

Bonus Task 3 (Derive EM algorithm):

Derive the expected value of the log likelihood function $Q(\mu, \sigma^2|\mu_k, \sigma_k^2)$ from the likelihood function in Task 4. Prove that the weighted sample mean and variance indeed maximize Q.