Institut für Informatik Wintersemester 2015/2016

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

GoBi: Excercise 3 (Bonus)

Differential Expression

Deadline: Thursday, 17.12.2015, 14:00
alternative Deadline: Thursday, 05.01.2016, 14:00

Bonus Task 1: Quality check

A widely used tool for assessing the quality of RNA-seq results is FastQC. You find in `/home/proj/biosoft/praktikum/genprakt/assignments/a3/data/fastqc` the fastqc results for the 8 paired-end RNA-seq run.

Read the manual for FastQC; `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/`

and interpret the results.

Bonus Task 2 (Detect potential PCR duplicates):

Adapt your program from Task 1 to annotate each read pair by the PCR copy index. The PCR copy index is an increasing number for all read pairs mapping to the same genomic region.

Hint:

An efficient way to compute PCR copy index is to collect genomic region vectors in an appropriate data structure. If you encounter a genomic region vector, look into this data structure whether it is already present and how many times. For memory efficiency, always remove from this data structure genomic region vectors that will not be encountered anymore.

Bonus Task 3 (Extended feature statistics):

Compute further statistics for each of the BAM files from Task 2 and create additional plots based on your extracted information:

- **quality**: barplot (y-axis percentage of mapped NRP)
  - multimapped NRP
  - no mismatch NRP
  - mismatch <= 2 NRP
  - mismatch <= 3 NRP
  - no clipping NRP
  - clipping <= 5 NRP
- **intergenic**: barplot (y-axis percentage of intergenic NRP)
  - gene-proximal NRP (distance ≥ 500)
  - antisense NRP
  - intergenic-spliced NRP
- **uniqueness**: barplot (y-axis percentage of gene mapping NRP)
  - gene-unique NRP
  - multi-gene $x <= 3$ NRP
  - intergenic-antisense NRP
  - gene-tr-unique NRP
  - gene-tr-unique $x <= 3$
  - gene-merged (any y)
  - gene-merged unique
  - intronic
- **biotype**: barplot (x-axis percentage of transcript-unique NRP) showing the 10 most frequent biotypes (left to right descending)

Bonus Task 4 (Gene counts without PCR duplicates):

To check for the effect of PCR duplicates, repeat Task 3 for the number of unique regions (NURs) instead of NRPs (NURs can be easily derived e.g. by computing the NRPs for read pairs with PCR copy index 0)

Bonus Task 5 (Visualization):

One of the standard desktop visualization tool for mapped RNA-seq data is IGV. `https://www.broadinstitute.org/igv/`.

Use the installed version in `/home/proj/biosoft/software/IGV_2.3.60/igv.sh`

along with the prepared genome `/home/proj/biosoft/praktikum/genprakt/igv/ens.hg.19.genome`

and annotation: `/home/proj/biosoft/praktikum/genprakt/igv/Homo_sapiens.GRCh37.75.gtf`

to visualize some selected regions of interests, by loading the contextmap bamfiles as tracks.

Use your extracted information to reate a IGV region file corresponding to the following regions:

- regions corresponding to the 5 genes from `liver/ENCSR000AFB/repl1` where the proportion of the NRP-s for most abundant transcript to all transcriptomic NRP is the smallest. Use also the Sashimi-plot visaulization (see IGV website for how and why).
- the regions corresponding to the top 10 genes ordered by max(transcript-RPKM)-min(transcript-RPKM) over all runs.
- the regions corresponding to the top 10 genes ordered by max(intron-RPKM)-min(intron-RPKM) over all runs.

Create for each point above a snapshot from the visualization and discuss your findings shortly.