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

Wintersemester 2015/2016

GoBi: Excercise 3

Differential Expression

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

Save your solution to /home/proj/biocluster/genprakt/${stud_account}/Solution3. 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.

You find the input files in the directory

`/home/proj/biosoft/praktikum/genprakt/assignments/a2/data/`

We will analyze RNA-seq data from the ENCODE project
(see `https://www.encodeproject.org`) from two tissues (liver and melanocyte of skin) with 2 x 2 replicates each. The experiments were carried out with a strand-specific paired-end RNA-seq protocol.

Task 1 (Mapped read annotation):

You find four small bam files containing the mappings of several thousand read pairs in
`/home/proj/biosoft/praktikum/genprakt/assignments/a3/data/debug_bams/*_annot.bam`

The goal of this task is to get prepared for feature statistics and feature counting (the following tasks) by annotating each read pair in the bam file with the categories used later and by comparing these annotations to reference annotations given by us.

The read pair features you have to annotate are:

- Mismatch count
- Clipping size
- Split count or annotate split-inconsistent:true (skip all further annotation in this case)
- Gene count: number of genes that are matched by the read pair (for the *match* definitions, see below)

- For each matched gene, annotate matched transcripts: If it matches one or more transcripts, write Gene-id,Gene-biotype:transcript-id1,transcript-id2,... If it does not match any transcript but it matches the merged transcript, write Gene-id,Gene-biotype:MERGED. If it does neither of those, write Gene-id,Gene-biotype:INTRON; separate genes by the pipe symbol
- Gene distance: distance to the next sense gene (only for intergenic read pairs)
- Antisense: would it match a gene if the read pair was on the other strand (true/false; only if no gene is matched)

A read pair does not match any gene if reads are mapped to different chromosomes. A read pair is split inconsistent (annotate skip-inconsistent: true) if there is at least one read base in one of the reads that is within a split of the other read (skip all other annotations in these cases).

A **gene** is matched, if the read pair is fully contained in the gene body disregarding introns. A read pair matches a **transcript**, if it is contained in the transcript and intron boundaries are consistent. It matches the **merged transcript**, if each read base is part of some exon.

A read pair is intergenic if no gene is contained between its first base and its last base.

Use your GTF parser (with the GTF file h.ens.75) and the htsjdk library to annotated read pairs and process the bamfiles. Write the annotations as tab-separated list into the XX attribute of the read encountered last. You can compare your annotations to the XX attribute in the files in the debug folder. Apply your annotation on the four bamfiles in

`/home/proj/biosoft/praktikum/genprakt/assignments/a3/data/test_bams`

and save the results (as tab separated files (`readid<tab>annotations`) into your solution with the following names:
`contextmap.annot, gsnap.annot, star.annot, tophat2.annot`.

Hints:

- To keep the bamfiles as small as possible we removed the sequence and quality strings. You have to set the validation strigency to `SAMFileReader.ValidationStringency.SILENT` in the your `SAMFileReader` instance before processing.
- Ignore reads if: they are secondary or supplementary, they are unmapped, their mate is unmapped or mate pairs are mapped to the same strand of the same chromosome.
- process SAMRecords for the paired-end data together, (i.e. while iterating over the records, keep open records in an appropriate structure, and wait for the mate pair to come - use for this the method `getReadName()`, `getMateUnmappedFlag`, `getMateAlignmentStart()`). Note that some mapper can also report chimeric read-pairs (produced be fusion-transcripts from different genes sometimes even from different chromosomes. (see `getReferenceName()`, `getMateReferenceName()`).

- you can retrieve read splits, i.e. introns from the aligned blocks per read `getAlignedBlocks()`.
- you can derive the number of mismatches by checking the attribute NM or nM or XM (the attribute depends on the method, but you can check any of these attributes is set (e.g. `null != getAttribute("NM")`). If it is set it is an `Integer`. Important: $STAR$ reports the number of mismatches per read-pair, not per read.
- you can derive clippings by comparing `getUnclippedStart(),getUnclippedEnd()` to `getAlignmentStart(),getAlignmentEnd()`
- RNA-seq experiments can be strand specific or not strand specific. (One can derive this information from a mapping, how?). If the experiment is strand specific and paired-end the first of the read pair can be on the sense or the antisense strand (the mate have to be mapped correspondingly to the other strand). If the experiment is strand specific, use the information in the processing, for example a read pair might be unique for a gene if considering strandness, or be categorized as intergenic (antisense) while mapped to the wrong strand. **The used ENCODE data are strand-specific and the first read maps to the antisense strand.**

Implementation hints:

- Write a class that contains all annotations along with the given read pair. It may also provide methods to compute the annotations.
- Let this class also implement a method to parse annotations from a SAMRecord (our reference); implement *equals* to check your computed annotations against ours.
- Such a class would greatly facilitate the subsequent tasks.

Task 2 (Feature statistics):

You find mappings with different mapper methods (star, tophat2, gsnap, contextmap) for all runs under `/home/proj/biosoft/praktikum/genprakt/assignments/a3/data/ENCODE_DATA`

as defined in the tab separated file `mappinginfos.tsv` in the same directory along with the condition, experiment, replicate, number of read-pairs, mapper method.

Use your annotation methods (or the methods of the SAMRecord class, when appropriate) to compute statistics for each mapping. Create a barplot for each bam-file (x-axis: the following categories, y-axis: mio. reads):

- all NRP
- mapped NRP
- multimapped NRP
- transcriptomic NRP
- merged-tr NRP
- intronic NRP
- antisense NRP
- intergenic NRP

Provide for all labels the percentage of reads on all NRP in parenthesis.

Implementation hints:

- Use your annotation class from Task 1 together with several *Counter* objects to compute all the statistics.
- Do **NOT** write annotated bam-files / tab separated annotation files as in Task 2.

Task 3 (Gene counts):

Extend your program from Task 2 to also write a table of NRPs per gene. It should be clear by now that there are many ways to assign read pairs to genes. Report NRPs for the following categories:

- NRP mapping within the gene's genomic region
- NRP within the gene's genomic region but not within any other gene's genomic region
- intronic NRP
- transcriptomic NRP
- merged-transcriptomic NRP
- NRP for the transcript with highest NRP among the gene's annotated transcripts

Transform computed NRPs to FPKM values and plot:

- **gene-tr-FPKM plot**: cumulative plot showing the transcript-FPKM of the genes, whereas $transcript{-}FPKM(x) = (10^9 * \text{transcript mapping NRP in gene } x)/(\text{all NRP} * \text{merged transcript length})$
- **gene-intron-FPKM plot**: cumulative plot showing the intron-FPKM of the genes, whereas $intron{-}FPKM(x) = (10^9 * \text{intron mapping NRP in gene } x)/(\text{all NRP} * \text{merged transcript length})$
- **main-tr plot**: cumulative plot coverage percentage (covP) of the most abundant transcript to the transcriptomic abundance of the gene $covP(x) = (100 * \text{NRP in most abundant transcript in gene } x/(\text{all t}$

Use these plots and create html files comparing different mapper pro RNA-seq run (8 html files) and comparing the replicates from the two conditions for each mapper (2 x 4 html files).

Implementation hints:

- Again, use your annotation class from Task 2 together with a *Counter* object for each gene to compute the gene counts.
- Do **NOT** write annotated bam-files / tab separated annotation files as in Task 2.