SQA_1v2

RNA-seq analysis using Galaxy platform

(Modified by Lucia Martin Caballero after Ramon Ramos Barrales, as of August 2016)

Get started with Galaxy

1. Enter the galaxy server from your computer, within the LRZ or through a VPN connection (to be downloaded from the LRZ webpage)
    a. Copy into your browser http://blum-galaxy.genzentrum.lmu.de/galaxy/
    b. Go into user->login
    user: **********
    pwd: ********
2. Get your data from Shared data->Data libraries->Ladurner
3. In history, click  ->create new and import the fastq files from the Ladurner middle panel. The heavier file is the actual data while the lighter one is the index file.
4. History->  ->copy datasets. Select source history->Pombe Genome and destination history->New history. Copy the datasets **genes.gtf** (with the gene annotations) and **Schizosaccharomyces_pombe.ASM294v2.27.dna.genome.fa** (with the pombe genome)

Do the data analysis

1. If your data is decompressed->NGS: Demultiplex.->Illumina Demultiplex.
    - paired end run? -> "Single-end".
    - Insert Single-end -> add the heaviest file here.
    - Use Illumina index read for demultiplexing -> "YES"
    - Add the lighter file.
    - Filter reads without barcode -> "No"
    - Mismatch in Index -> "1"
    - How many barcodes? -> introduce the number of barcodes used and the sequence of them.

    If your data is compressed into gzip for spacesaving reasons-> Je-Demultiplex-Illu
    -Is this single or paired library->Single-end
    -FASTQ file-> choose the two heavier files
    -First index file-> choose the lightest file
    -Barcode list input type-> Paste the barcodes list in a text field within the form-> add the barcodes used in the following format:

*Please note the barcodes must be 8 characters long!! In my case, used the short barcode (6 characters) plus the first two on the right of the barcode (AT)
-Do your reads contain UMIs?->No
-Add matched barcode at the end of the read header->yes
-MM->1
-MMD->1
-Q->10 (if it doesn't work because it is too stringent, try to go down to 5)
-V-> detect automatically
-extra number of bases to be trimmed right after the barcode-> 0
-extra number of bases to be trimmed from the read end-> 0
-Replace white space in the read name/header with specified symbols-> :
-Keep unassigned reads-> yes
Execute

2. 2. NGS: QC and manipulation. FastQC:Read QC.

- Do this for each Demultiplexed file. Usually you can get bad quality for: Per base sequence content, Per base GD content and Sequence Duplication Levels. Kmer Content use to don't be perfect.

3. NGS: Mapping. Map with Bowtie for Illumina.

- Are files FASTA Format? -> No
- Select a reference genome from the history (**Schizosaccharomyces_pombe.ASM294v2.27.dna.genome.fa)**.
- Use default options for building indices.
- In general we use Single-end libraries.
- Use commonly used settings for Bowtie analysis at least something special have to be done (like increase the numbers of repeats allowed).
- Suppress the header in the output SAM file -> NO

4. NGS: SAM Tools. SAM to BAM.

- You can leave the reference list with the locally cached option or select the pombe genome.
- Minimum mapping quality -> 0
- Add the database to the bam files generated if they are not linked already. Click in the file, if in database there is a question mark, then add the genome by clicking in edit attributes (pencil icon).

5. Convert Formats. BAM to BigWig or BedGraph (for keeping strand information)

- At the moment should be a bug in the server and this conversion give an error. Do the conversion directly using the edit attribute (pencil icon) of the

bam file and then selecting “convert format” option. Lucia: I tried this and gave an error.

-For BedGraph: NGS: Bedtools-> Create a BedGraph of genome coverage (select the strand that you want).

6. Download the BigWig files to your computer and check them in a genome browser (**IGV and IGB** works fine).

7. NGS: Gene expression. HTSeq-count.

- Select the bam file and the genes annotation file (genes.gtf).
- Feature type -> exon
- Attribute to be used as feature ID -> “gene_id” is by default. If you prefere gene names type in **“gene_name”**
- Min alignment quality -> 0
- Select mode? -> Union (intersection_nonempty could be interesting to check).
- Is paired-end? -> Usually no
- Select strand specifity –> Reverse

8. NGS: Gene expression. DESeq2.

- FDR -> 0.05
- Add HTSeq files for different groups (e.g wt and mutant).
- Use beta prior? -> Yes
- Fitting of dispersions to the mean intensity -> Parametric.