Quality Control of NGS data: tutorial in Galaxy

In this tutorial you will compare a high-quality fastq file to a low-quality fastq file, using the FastQC program in Galaxy. Then you will try to clean up the low-quality file and assess the improvements.

Use one of these Galaxy servers:

- http://galaxy.nbic.nl/ → log in using your (novel) galaxy account
- http://www.usegalaxy.org

The good and bad fastq file are located at:

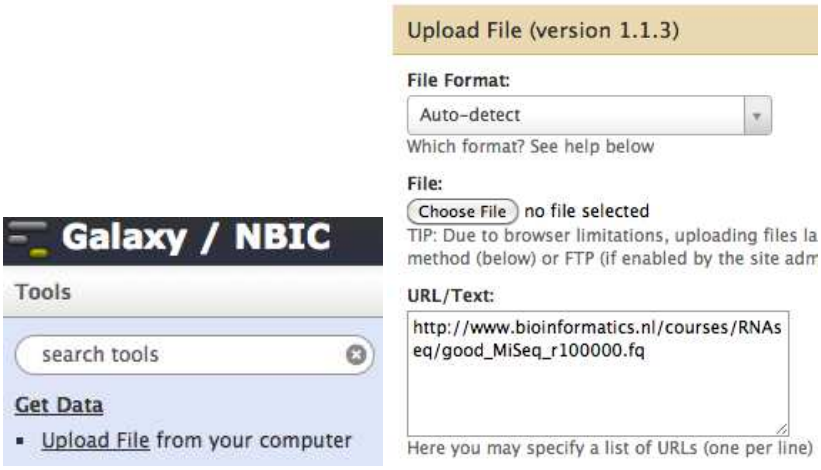
- http://www.bioinformatics.nl/courses/RNAseq/good_MiSeq_r100000.fq
- http://www.bioinformatics.nl/courses/RNAseq/bad_MiSeq_r100000.fq

Use the FastQC manual to interpret the QC reports:

- http://www.bioinformatics.nl/courses/RNAseq/FastQC_Manual.pdf

Load both data files into Galaxy:

- Use the "Get Data" tool from the menu and paste the URL in the textbox



- How many sequences are in each file?

Generate a FastQC report for both files:

Look at both FastQC reports and assess the differences:

- Use the "eye" button in your History

- Which encoding is used for the quality values?
- What is the GC content in the data sets?
- Which graphs give errors or warnings? What seems to be the problem?
- What are the main differences between the reports?

Clean up the bad fastq file by connecting various tools from the section "NGS: QC and manipulation". You may experiment with different strategies to see what works best. For example:

- FASTQ Groomer (Input FASTQ quality scores type = Sanger), followed by FASTQ Quality Trimmer (think about which settings you want to use)
- "Filter by quality" followed by "FASTQ Quality Trimmer"

Generate a new FastQC report for the cleaned file(s):

- What result did you expect?
- Did you get the expected result?
- How many sequences does the cleaned file contain?
- What is the sequence length in the cleaned file? Do all sequences have the same length?
- Can you think of other or further cleaning steps? Try them if there's any time left.