Digital Expression Explorer: Data Processing Method

July 2015

- SRA archives are downloaded from NCBI SRA by ASCP transfer
- MD5 sums are generated
- Fastq dump from the SRA toolkit (v2.4.4) to generate a sample of 40,000 sequences
- FastQC is used to perform quality checks on the data sample
- The format (SE vs PE) and read length is diagnosed
- Fastq data are dumped. Read 2 is discarded for PE data.
- Quality trimming with fastq_quality_trimmer from FastX-Toolkit 0.0.13 with parameters -t 20 -l 19 -Q33.
- STAR v2.4.0g1 is used to map reads to the respective genome specifying the corresponding GTF file "--sjdbGTFfile". Otherwise default parameters.
- Samtools v0.1.19-44428 is used for bam conversion and sorting.
- Featurecounts v1.4.2 is used to summarise gene-wise tag counts with the following parameters "-Q 10 -s 0".

For color-space data:

- abi-dump (SRA toolkit) is used to extract the sra archive.
- SolidTrimmer.py is used to quality trim sequences with minimum quality of 20 of and minimum read length of 19.
- SubJunc aligner v1.4.2 is used to map colorspace reads to the genome with the default settings. **(Liao et al 2013)**.