Comparing R-based methods and Cuffdiff2 for analysis of RNA-seq data in Galaxy

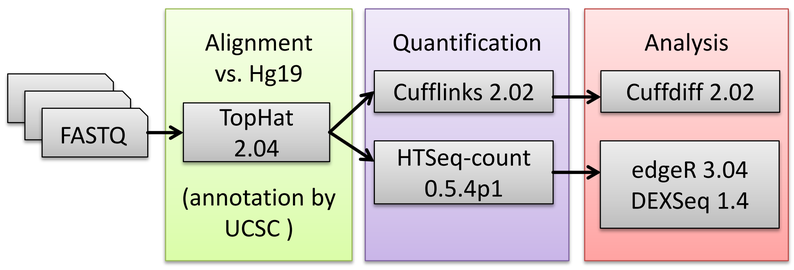
René Böttcher (1,4), Saskia Hiltemann (1,2), Bram Stoker (2), A. Marije Hoogland (3), Leon Mei (5), G.J.L.H. van Leenders (3), Peter Beyerlein (4), Andrew Stubbs (2), Guido Jenster (1)

1 Dept. of Urology; 2 Dept. of Bioinformatics; 3 Dept. of Pathology, Erasmus MC, Rotterdam, The Netherlands;
4 Dept. of Bioinformatics, Technical University of Applied Sciences Wildau, Wildau, Germany; 5 Bioassist, Netherlands Bioinformatics Center (NBIC), Nijmegen, The Netherlands

Approaches for analyzing RNA-seq data

Differential expression (DE) and differential exon usage (DEU) in RNA-seq data are commonly investigated by Cufflinks and Cuffdiff at the moment. However, previous work demonstrated that Cuffdiff, prior to version 2, does not capture the biological variation between groups containing many replicates. Therefore, we set out to implement two R-based methods (edgeR and DEXSeq) in Galaxy and to compare their performance with a recent release of Cuffdiff2.

Methodology

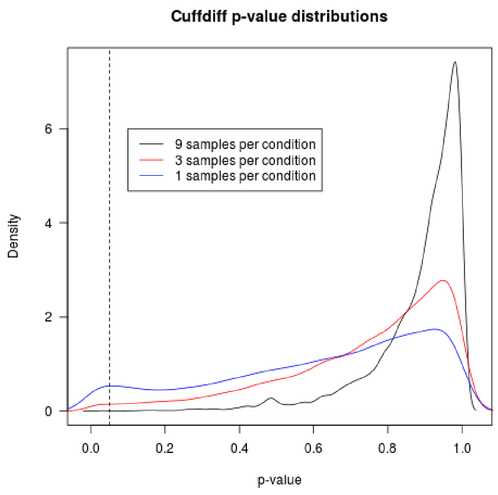


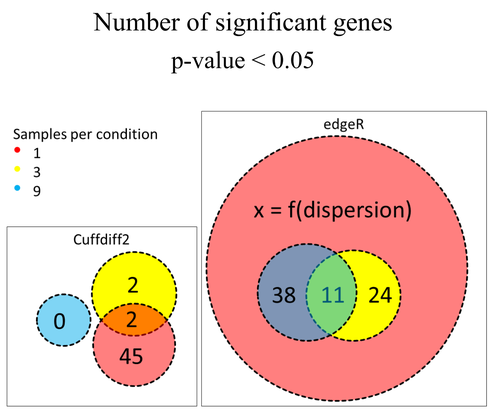
Our test case consisted of 27 prostate cancer samples grouped in three distinct conditions (nine samples per condition).

The groups are based on assigned Gleason scores, indicating progressing tumor stage with increasing number and 4+3 being worse than 3+4. Furthermore, two of these conditions represent tissue samples from the same patient and were therefore treated as paired samples. In addition, we used publicly available data to confirm our findings (Kannan et al., 2011), which consisted of 10 tumor – normal pairs from prostate cancer patients.

Results

The distribution of p-values derived from the Cuffdiff 2.0.2 depends on the number of samples per condition. Starting with a one vs. one comparison between unpaired samples (3+4 vs. 3+3), Cuffdiff reports 47 genes as significant considering a p-value threshold of 0.05. This number drops to four when three samples are used per condition. Lastly, when using nine biological replicates per condition, Cuffdiff does not report any significant genes. This is also the case when higher stage tumors or paired samples are considered.
In contrast, edgeR and DEXSeq both are able to model increased variance and provide significant results for all investigated contrasts (see Table 1, FDR < 0.05 and 3 genes with DEU, adj. p-value < 0.1). Furthermore, the independent result lists of Cuffdiff and edgeR share only one gene. Notably, both edgeR and DEXSeq do not produce usable results in case of no replicates. This is logical, since the estimated dispersion has to be inserted manually and therefore, any results obtained are depending on this estimate. Thus, the authors do not advise to use their software without replicates.
To also take into account that cancer samples have an intrisically high variance, we also compared 10 tumor and 10 normal samples in a paired fashion. Again, Cuffdiff2 was not able to find any genes as statistically significant, whilst edgeR returns a list of 2986 genes (p-value < 0.05) and 1570 genes (p-value < 0.01).





Tab. 1: Results of DE / DEU analysis, comparing 9 samples per condition.

Method	7 (3+4) vs. 6 (3+3)	7 (4+3) vs. 6 (3+3)	7 vs. 7 (paired samples)
Cufflinks	0	0	0
edgeR	49	230	111
DEXSeq	3	6	8

Moreover, the runtime and computational requirements of the Cufflinks2/Cuffdiff2 pipeline exceed edgeR by far and do not scale well with an increasing number of samples.
These findings suggest that Cuffdiff2 in its current release (2.02) is not suited for analysis of larger cohorts and thus alternative approaches are required.

The TraIT Project

TraIT aims to develop a long-lasting IT infrastructure for molecular medicine that will facilitate the collection, storage, analysis, archiving, sharing and securing of data generated in operational CTMM research projects.
The project will build on existing expertise to create an IT infrastructure that will help to accelerate the molecular research thoughout the Dutch Life Sciences and its Health sector.

The Galaxy server can be accessed via the NBIC website:

galaxy.nbic.nl

Galaxy HPC CLOUD Architecture

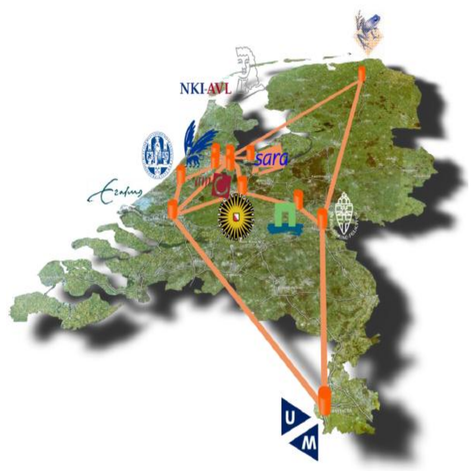
The rapid evolution of NGS technologies together with decreasing costs create a challenge of storing and analyzing the vast amount of sequencing data generated by experimental biologists. Configuring suitable data analysis software and access to readily available computation and storage are the two major bottlenecks faced by many research groups.

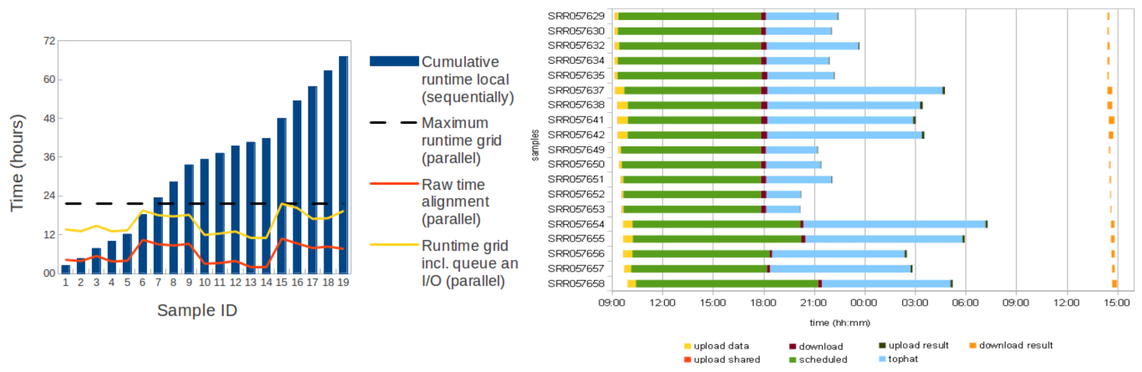
* In collaboration with NBIC, BiGGrid, Mattias de Hollander (NIOO/NBIC)

RNA-seq analysis on the Life Science Grid

To demonstrate the advantages of grid /cloud usage for the analysis of RNA-seq data, we performed alignments of the 10 cancer - normal samples pairs on both the Dutch Life Science Grid (LSG) and a local workstation. Since alignments are one of the bottleneck operations in sequencing analysis, a major speed-up can leverage research efficiency data throughput.

The machine environment of our workstation consisted of an 8-core processor and 24 GB RAM. In comparison, we made use of the 'long' queue of the LSG comprising a total of 2332 cores, whereas 8 cores were assigned per TopHat instance. As can be seen below, runtime of a single process on the LSG far exceeds the one on a workstation due to a slower processing speed and time required for up-/download as well as queuing. This behaviour is reversed when more samples are being processed owing to parallelization of the actual alignment processes. Therefore, grid computation requires as much time as the slowest process, whereas runtime increases sequentially for local computation.





Conclusion

Our Galaxy implementations of edgeR and DEXSeq workflows provide an accurate high-throughput analysis and performance comparisons of different RNA-seq tools in Galaxy. Both tools seem to be able to deal with an increased biological variation due to replicates and allow for paired samples. Since Cuffdiff2 is under active development, we expect an improved release targeting the issues described above. Until then, we recommend to adapt the RNA-seq workflow and provide alternative tools depending on the number of biological replicates per condition. Other alternatives besides edgeR include DESeq / DESeq2, BitSeq, EBSeq or the combination of Voom and Limma. We hope that our experiences will help developers and users to save time and effort when analyzing RNA-seq data.