Comparison of open source (Galaxy based) and commercial pipelines for RNA-Seq data analysis

Slave Trajanoski, PhD[1];Marija Djurdjevic, Msc[1]; Andrea Groselj-Strele, PhD[1]
[1]Medical University Graz, Center for Medical Research

ABSTRACT

From the first RNA-Seq projects until today software for data analysis was constantly developed and improved. Nowadays we're finding a plethora of different solutions of tools for either single parts of RNA-Seq data analysis as well as complete pipelines that deliver gene expression results. In our work, we present a comparison (Table 1) of one commercial software from the company Partek® as two implementations: Partek® Flow® and Partek® Genomics Suite® and open source tools implemented as pipeline in a very popular web-based framework Galaxy[1] .
By using a test dataset from the NCBI Sequence Read Archive (SRA) we could evaluate performance (Table 2) and run usability tests for the two different approaches. We come to the conclusion that:
•Results from both solutions are comparable (Figure 1 to Figure 4), but one has to be very careful about parameters used in each step since they can lead to different results;
•Usability is on the side of the commercial solution, even though Galaxy developers are making good progress in this direction;
•For the sake of easy and fast analysis a lot of parameters are hidden and not changeable for the user in the commercial software, which leads to lower flexibility in comparison with the open source pipelines;
•Less management effort when using commercial software, which on the other hand is connected with license costs.

CONTACT

Slave Trajanoski
Medical University Graz
Center for Medical Research
Email:
slave.trajanoski@medunigraz.at
Phone: +43 316 385 73024
Website: zmf.medunigraz.at

INTRODUCTION

Partek® Flow®

Server based solution for NGS data analysis over client web access

Partek® Genomics Suite

Desktop standalone software for statistical analysis of microarray & Next-Generation Sequencing studies

Workflows in Galaxy web-based platform

Open source server software for integration of different tools for data processing, analysis and visualization

Partek® Flow® installed and tested on:
HP Z800 Workstation
Two 64-bit 2.66GHz 6-core CPUs
48GB RAM Memory
Linux Debian 8.5 (Jessie)

Partek® Genomic Suite® installed and tested on:
MacBook Pro
one 2.5GHz 6MB 4-core CPU
16GB 1800MHz DDR3
OS X 10.10.5 Yosemite

MUG Galaxy (https://galaxy.medunigraz.at)
256GB RAM Memory
Two 64-bit 2.4GHz 10-core CPUs
11 TB storage space for data
Linux Debian 8.5 (Jessie)

Study used for evaluation:

GEO: GSE46665, PRJNA201433, Illumina HiSeq 2000 [2]

	Partek Flow	Partek Genomics Suite	Galaxy
Import FastQ data	X	NA	X
QA/QC and trimming	X	NA	X
Alignment with STAR	X	NA	X
Quantification	X	X	X
Normalization	X	X	X
Differential gene expression analysis	X	X	X
Filter genes 0.05 + FC 2	X	X	X
Visualization	X*	X	X
Optional: enrichment	X	X	X

*Limited possibility.

Table 1. Steps in data analysis.

METHODS

Flow®

Estimation of transcript abundance:
1. Partek® E/M (Partek's optimization of the expectation-maximization algorithm)[3]
    1. Quantify to annotation model
    2. Quantify to reference (no reference genome available)
2. Cufflinks

Statistical methods for differential gene expression (DE) analysis:
1. GSA: Gene Specific Analysis is a statistical modeling approach used to test for differential expression of genes or transcripts in Partek® Flow®. GSA is capable of considering the following five response distributions: Normal, Lognormal, Lognormal with shrinkage, Negative Binomial, Poisson
2. ANOVA

Multivariate analysis (PCA for count data, hierarchical clustering)
Normalization

Genomics Suite

Estimation of transcript abundance:
1. Partek® E/M

Descriptive Statistics (Coef. Of var., mean, median, kurtosis, variance, skewness …)

Statistical methods for differential gene expression analysis:
1. ANOVA and Welch's ANOVA
2. T-test (one sample, two sample, paired)
3. Mann-Whitney, Kruskal-Walllis, Kolmogorov-Smirnov, Friedman, Quade
4. Fisher exact
5. Logistic regression
6. Multiple test
7. Power analysis

Transformation and normalization, scaling
Multivariate analysis (hierarchical clustering, SOM, MDS, PCA, CA, PLS)
Visualization (Boxplots, histograms, star plot, scatter plot, venn diagram, volcano plot, MA plot …)

DESeq2

Estimation of transcript abundance:
HTSeq (Quantify to reference genome)[4]

Statistical methods for differential gene expression analysis:
The package DESeq2 provides methods to test for differential expression by use of **negative binomial generalized linear models**; the estimates of dispersion and logarithmic fold changes incorporate data-driven prior distributions. For significance testing, DESeq2 uses a **Wald test**.[5]

Multivariate analysis (PCA, hierarchical clustering, dispersion estimates, MA−plot)
Geometric normalization

Cufflinks

Estimation of transcript abundance:
Cufflinks (de novo assembling and abundance estimation) – FPKM (Fragments Per Kilobase Of Exon Per Million Fragments Mapped)[6]

Statistical methods for DE gene expression analysis:
Cuffdiff uses the test statistics T = E[log(y)]/Var[log(y)], where y is the ratio of the normalized counts between two conditions, and this ratio approximately follows a normal distribution; hence, a t-test is used to calculate the P value for DE.[7]

Multivariate analysis (PCA, MDS, Scatter Matrix. Dendrogram, Volcano …)
Normalization (geometric normalisation, classic FPKM, quartile)

RESULTS

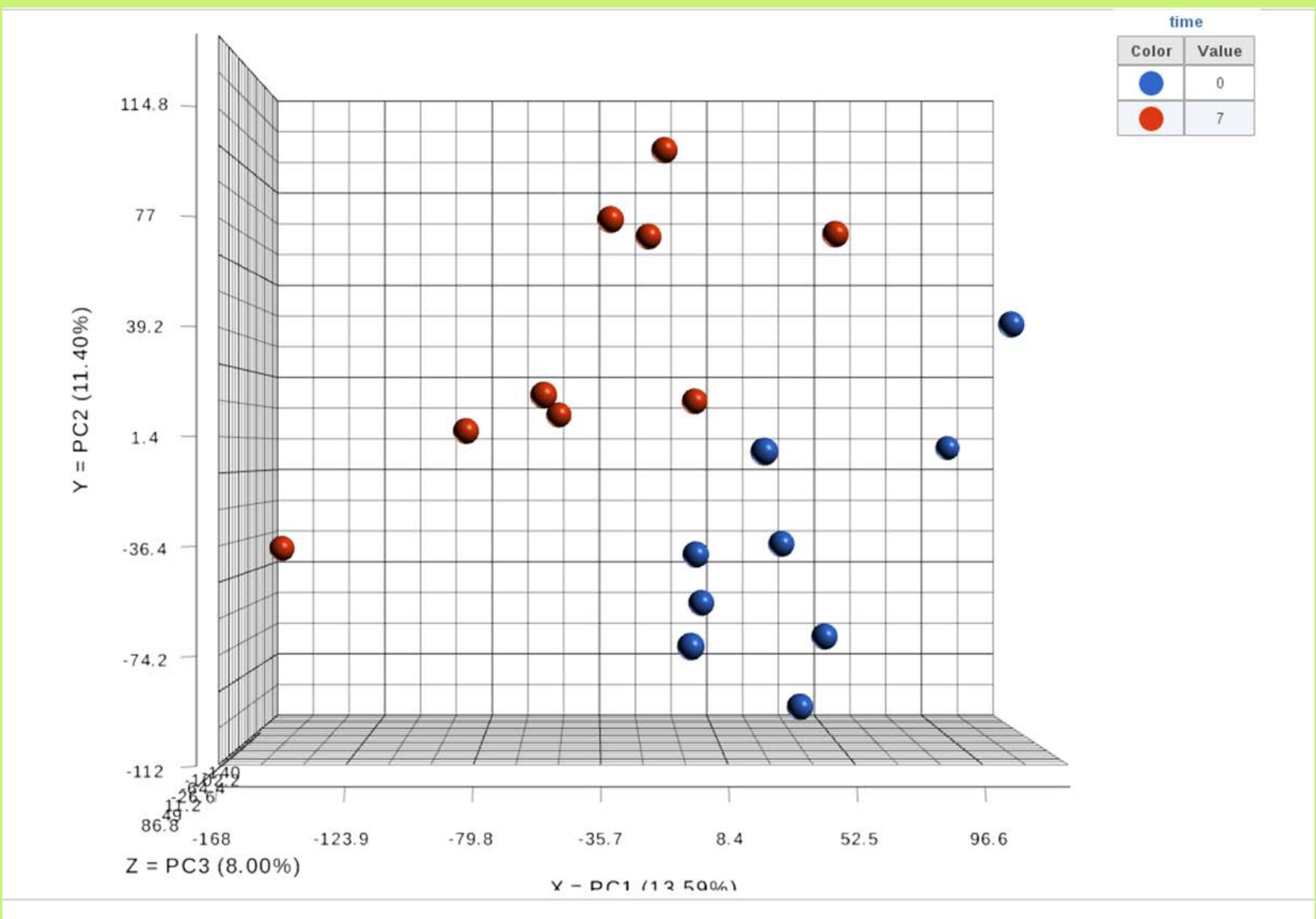

Figure 1. Partek® Flow® PCA diagram.

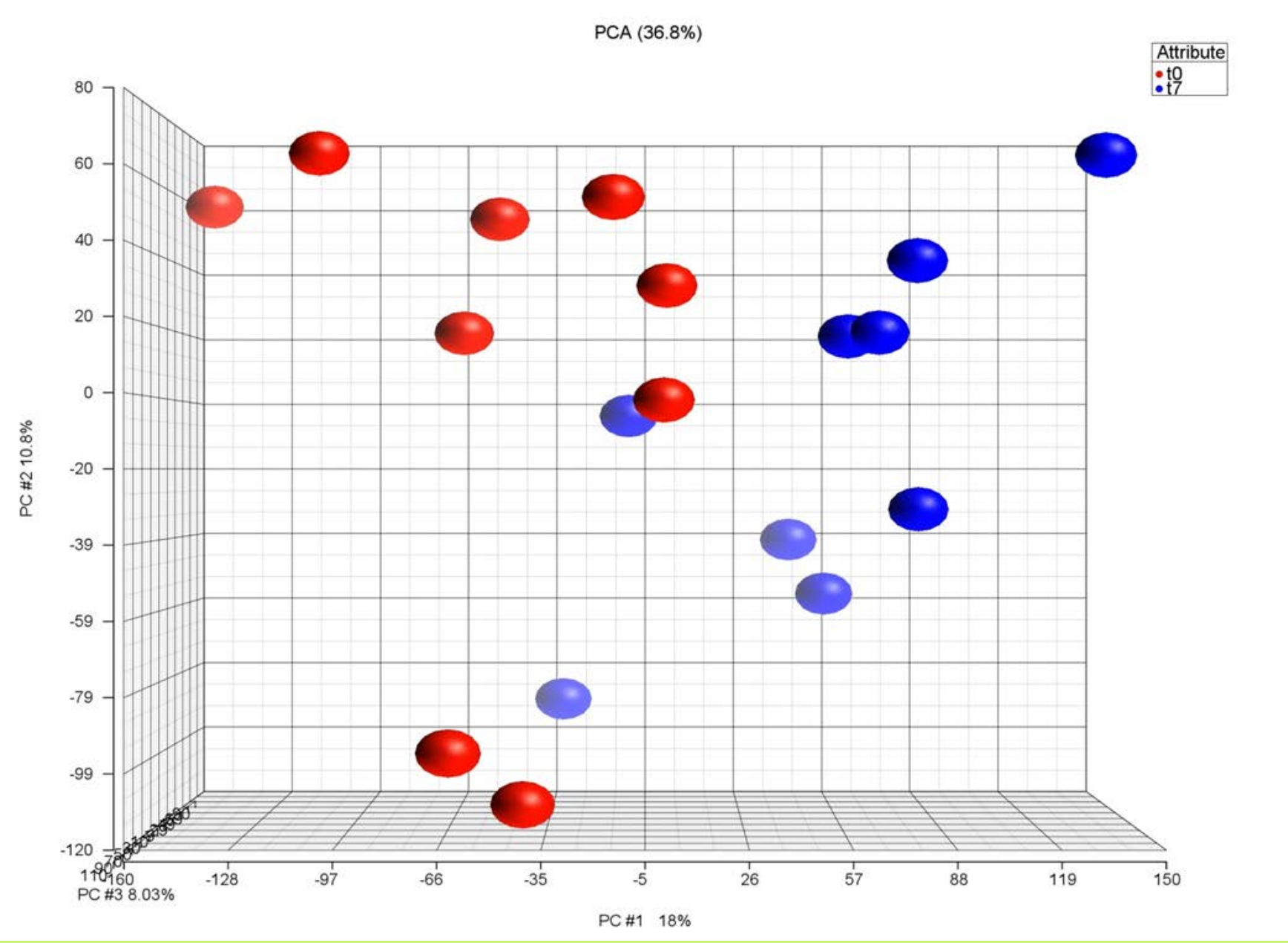

Figure 2. Partek® Genomix Suite® PCA diagram.

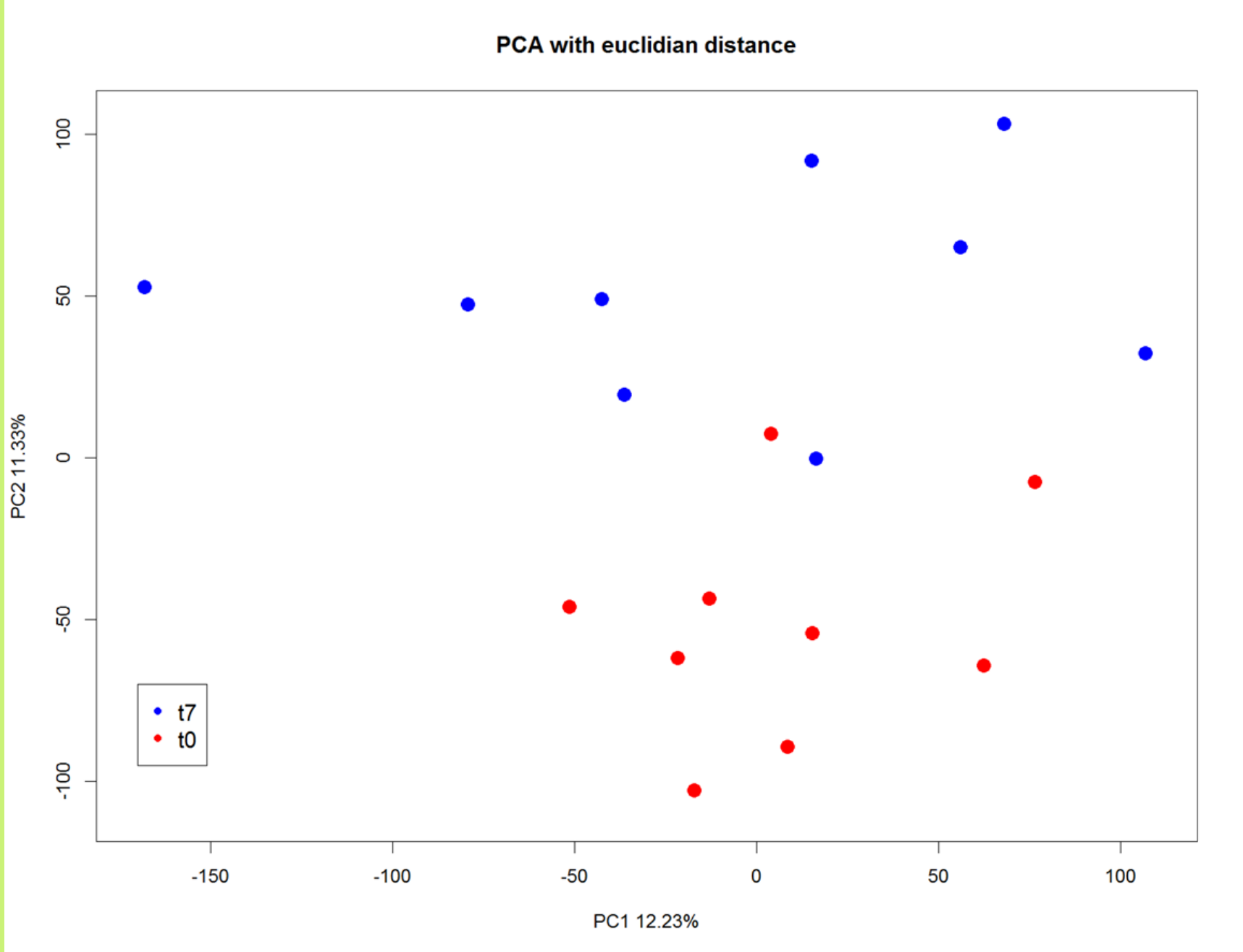

Figure 3. PCA diagram from Galaxy based analysis.

RESULTS

	Partek Flow		Partek Genomic Suite		Galaxy HTSeq	
time in hh:mm:ss	4 Samples	25 Samples	4 Samples2	25 Samples3	4 Samples4	18 Samples
Import Time	00:04:44	00:30:52				
PreAlignment QA/QC	00:15:38	02:02:39			00:06:00	00:49:00
Trim bases	00:06:23	00:57:51				
Alignment STAR-2.4.1d	00:37:11	03:22:18			00:13:00	00:55:00
Post-Alignment QA/QC	00:19:13	01:02:48	00:17:00	01:25:00		
Filter Alignment	00:18:37	01:07:47				
Coverage report	00:51:20	01:36:30				
Quantification to annotation model (Partek E/M)	00:11:16	00:26:05	02:39:00	07:00:00		
Differential Gene expression (GSA)	00:00:14	00:00:15				
Quantifiy to transcriptome (CuffLinks)	07:34:09					
Transcriptome expression analysis (CuffDiff)	00:35:43					
Qauntification and assembly (HTSeq)					00:54:00	01:23:00
Differential Gene expression (DeSeq)					00:01:00	00:04:00
Total 1 complete pipeline (E/M)	**02:44:36**	**11:07:05**				
Total 2 complete pipeline (Cuff...)	**10:42:58**					
Total 3 quantification and diff. Exprs.			**02:56:00**	**08:25:00**		
Total 4 complete pipeline (HTSeq)					**01:14:00**	**03:11:00**

Table 2. Times needed for calculation.

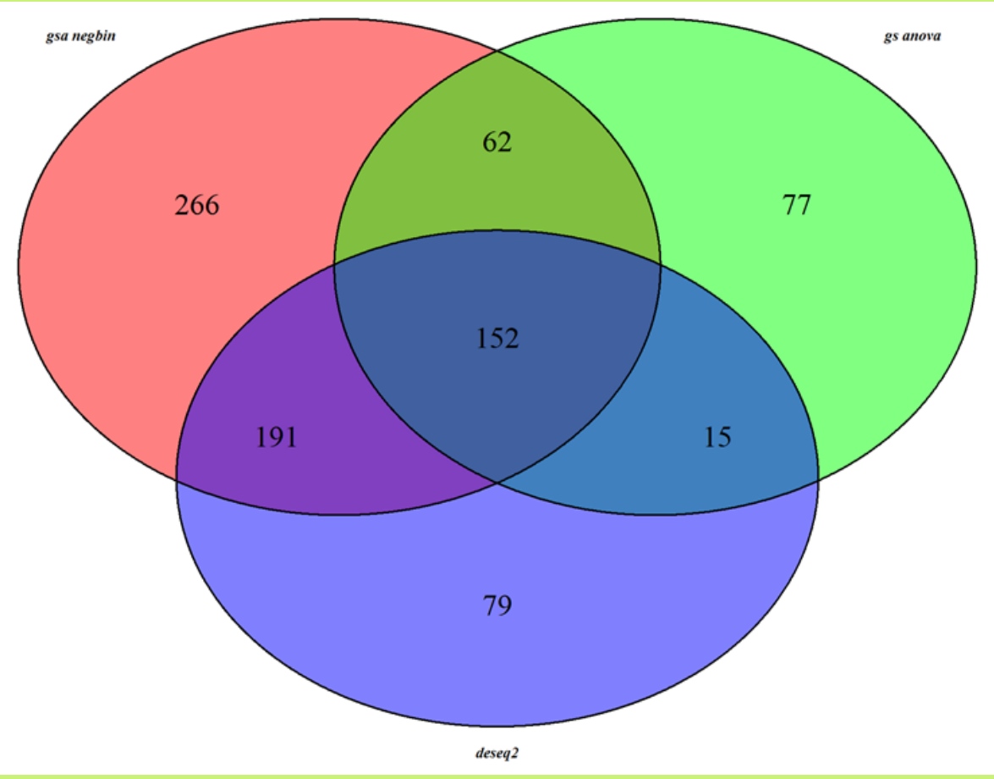

Figure 4. Number of Differentially expressed genes and their overlaps.

CONCLUSIONS

RNA-Seq is gaining on popularity between the scientists and a lot of tools and pipelines are already available for data analysis. By testing commercial and open source applications for data analysis we conclude that generated results are comparable, but main issue remains choosing proper parameters in each step in the analysis for the RNA-Seq experiment. Of course administration and management requires more effort when using open source solution, which on the other hand provides more flexibility and access to latest methods. In contrast usage of commercial software requires licensing fees, has less administration and is not so flexible in concern of changing parameters or implementation of new methods.

REFERENCES

1. Afgan E, et al. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. 2016;44(W1):W3..
2. Mitchell A et al., Identification of differentially expressed transcripts and pathways in blood one week and six months following implant of left ventricular assist devices., PLoS One, 2013 Oct 21;8(10):e77951
3. Xing Y, Yu T, Wu YN, Roy M, Kim J, Lee C. An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. Nucleic Acids Res. 2006;34:3150-3160.
4. Simon Anders, Paul Theodor Pyl, Wolfgang Huber, HTSeq — A Python framework to work with high-throughput sequencing data, Bioinformatics (2014)
5. Michael I Love, Wolfgang Huber, Simon Anders - Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biology (2014)
6. Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rin, Lior Pachter - Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, Nature Protocols (2012)
7. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D. - Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data, Genome Biology (2013)