Chapter

# Current State-of-the-Art of Clustering Methods for Gene Expression Data with RNA-Seq

*Ismail Jamail and Ahmed Moussa*

## Abstract

Latest developments in high-throughput cDNA sequencing (RNA-seq) have revolutionized gene expression profiling. This analysis aims to compare the expression levels of multiple genes between two or more samples, under specific circumstances or in a specific cell to give a global picture of cellular function. Thanks to these advances, gene expression data are being generated in large throughput. One of the primary data analysis tasks for gene expression studies involves data-mining techniques such as clustering and classification. Clustering, which is an unsupervised learning technique, has been widely used as a computational tool to facilitate our understanding of gene functions and regulations involved in a biological process. Cluster analysis aims to group the large number of genes present in a sample of gene expression profile data, such that similar or related genes are in same clusters, and different or unrelated genes are in distinct ones. Classification on the other hand can be used for grouping samples based on their expression profile. There are many clustering and classification algorithms that can be applied in gene expression experiments, the most widely used are hierarchical clustering, k-means clustering and model-based clustering that depend on a model to sort out the number of clusters. Depending on the data structure, a fitting clustering method must be used. In this chapter, we present a state of art of clustering algorithms and statistical approaches for grouping similar gene expression profiles that can be applied to RNA-seq data analysis and software tools dedicated to these methods. In addition, we discuss challenges in cluster analysis, and compare the performance of height commonly used clustering methods on four different public datasets from recount2.

**Keywords:** clustering, classification, RNA-seq, gene expression, adjusted Rand index, machine learning, deep learning

## 1. Introduction

In recent years, RNA-seq based on Next generation Sequencing has become an attractive alternative for conducting quantitative analysis of gene expression. This approach offers a number of advantages compared to microarray analysis such as the discovery of novel RNA species (RNA-seq is not limited by prior knowledge of the genome of the organism, it can be used for the detection of novel transcripts), the higher sensitivity for genes expressed either at low or very high level and the unbiased approach compared to microarrays that are subject to cross-hybridization

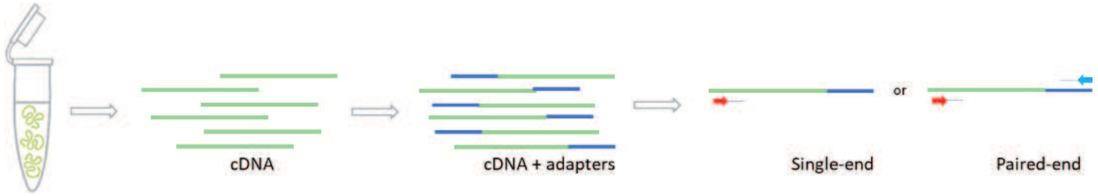

**Figure 1.**
*RNA sequencing.*

bias. Overall, RNA-seq is a better technique for many applications such as novel gene identification, differential gene expression, and splicing analysis.

The principle of RNA-seq is based on high-throughput next generation sequencing (NGS) technologies. The first step in the technique involves converting the population of RNA to be sequenced into cDNA fragments with adaptors attached to one or both ends, each molecule is then sequenced to obtain either single end short sequence reads or paired end reads [1]. These reads are stored in fastq files formats and consist of raw data for many analysis pipelines (**Figure 1**).

The primary objective of this chapter is to present algorithms for clustering gene expression data from RNA-seq. Therefore, in the first section, we will describe the different steps of the gene expression analysis workflow from preprocessing the raw reads to gene expression clustering and classification. In the second part of the chapter we will describe traditional, model-based and machine learning clustering methods for gene expression data, then we will conclude this chapter with a study for clustering samples of four public datasets from recount2, using different clustering methods and also evaluating the performance of each one using the adjusted rand index (RDI) and accuracy.

## 2. RNA-seq data analysis

RNA-seq has become a common tool for scientists to study the transcriptome complexity, and a convenient method for the analysis of differential gene expression. A typical RNA-seq data analysis workflow starts by preprocessing raw reads for contamination removal and quality control checks. The following step is to align the reads to a reference genome, or to make a de novo assembly if there is not any. Following the alignment, the quantification step aims to quantify aligned reads to produce a count matrix to use as entry data for Differential Expression (DE) analysis. Normalization and DE analysis normally go together as most of the methods have built-in normalization and accept only raw count matrix. For this study, we are more interested in the clustering step, we will perform Normalization of the raw counts separately and do the clustering without going through differential gene expression analysis. In the following section we describe with more details each step of the pipeline (**Figure 2**).

### 2.1 Preprocessing

Preprocessing raw reads consist of checking the quality of the reads, adapters trimming, removal of short reads and filtering bad quality bases. Tools like FastQC can generate a report summarizing the overall quality of the sequence information [2]. Based on this report we can determine how the quality trimming should be set up. Trimmomatic is one of many tools used to clean up the raw data. It can be used to remove adapters from the reads, trim off any low-quality bases at the ends of reads, and filter short reads that can align to multiple locations on the reference genome. Once the trimming step is done, it is a good practice to recheck the quality of the reads by rerunning FastQC.

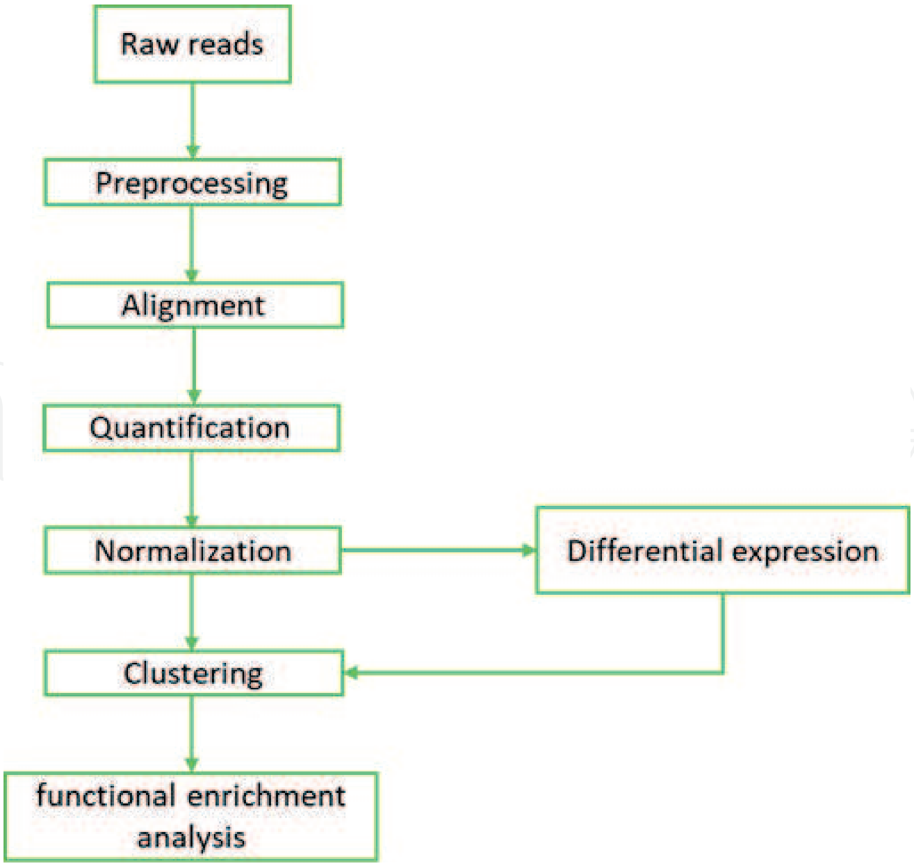

**Figure 2.**
*RNA-seq data analysis workflow.*

### 2.2 Alignment

Now that we have explored the quality of our raw reads, we can move on to read alignment. Read alignment is one of the first steps required for many different types of analysis. It aims to map the huge number of short RNA sequences generated by NGS instruments (reads) to a reference genome in order to identify the correct genomic loci from which the read originated. In RNA-seq, alignment is a major step for the calculation of transcript or gene expression levels; several splice-aware alignment methods have been developed for RNA-seq experiments such as STAR, HISAT2 or TopHat. These aligners are designed to specifically address many of the challenges of RNA-seq data mapping using a strategy to account for spliced alignments [3–5].

### 2.3 Quantification

Quantification of gene expression is to count the number of reads that map to each gene using methods such as HTSeq-count, FeatureCounts or kallisto [6–8]. This step is crucial if we want to do a gene differential expression analysis, which means to identify genes (or transcripts), if any, that have a statistically significant difference in abundance across the experimental groups or conditions.

### 2.4 Normalization

The read counts generated in the quantification step need to be normalized to make accurate comparisons of gene expression between samples or when doing an exploratory data analysis. Several normalization methods are used for this purpose such as CPM (counts per million), TPM (transcripts per kilobase million), RPKM/FPKM

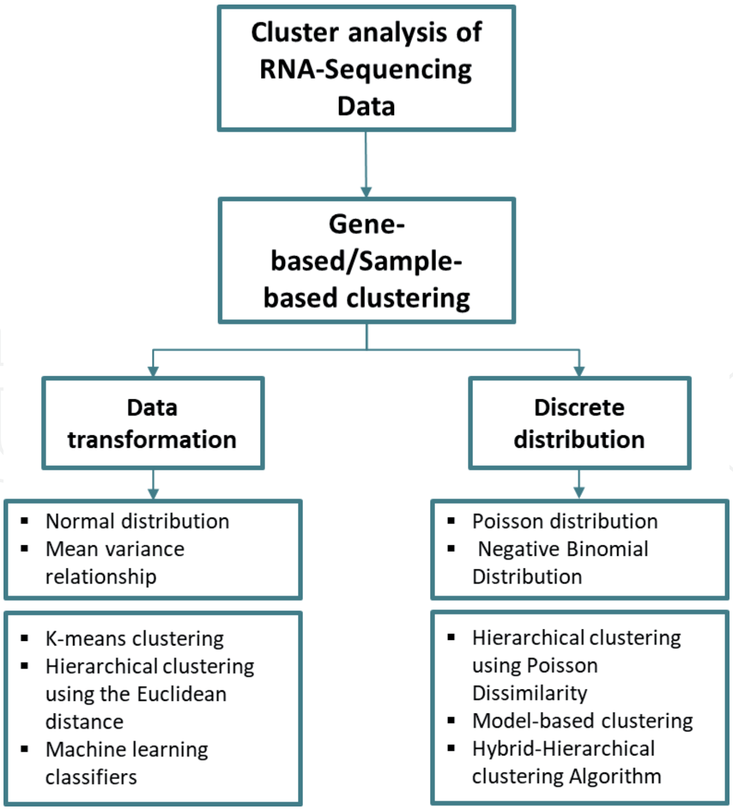
Cluster analysis of RNA-Sequencing Data
Gene-based/Sample-based clustering
Data transformation
Discrete distribution
Normal distribution
Mean variance relationship
Poisson distribution
Negative Binomial Distribution
K-means clustering
Hierarchical clustering using the Euclidean distance
Machine learning classifiers
Hierarchical clustering using Poisson Dissimilarity
Model-based clustering
Hybrid-Hierarchical clustering Algorithm

**Figure 3.**
*Cluster analysis of RNA-sequencing data.*

(reads/fragments per kilobase of exon per million reads/fragments mapped), DESeq2's median of ratios and EdgeR's trimmed mean of M values (TMM) [9].

### 2.5 Clustering

Cluster analysis techniques have proven to be helpful to understand gene expression data by uncovering unknown relationships among genes and unveiling different subtypes of diseases when it comes to clustering biological samples [10]. In the following section, we present methods for sample-based and gene-based clustering, starting with traditional methods used after data transformation then model-based clustering for data generated using a combination of probability distributions (**Figure 3**).

## 3. Clustering methods for gene expression data

### 3.1 Data transformation methods

Traditional clustering algorithms like hierarchical clustering and k-means cannot be directly applied to RNA-seq count data, to apply these methods for cluster analysis of RNA-seq data, that tend to follow an over-dispersed Poisson or negative binomial distribution, we need to transform the data in order to have a distribution closer to the normal distribution. In the following section, we present popular methods for data transformation:

- Logarithmic, widely used method to deal with skewed data in many research domains, often used to reduce the variability of the data and make the data conform more closely to the normal distribution. However, it was demonstrated in [11], that in most circumstances the log transformation does not help make data less variable or more normal and may, in some circumstances, make data more variable and more skewed.

- Variance stabilizing transformation: This method was used to transform microarray data to stabilize the asymptotic variance over the full range of the data [12].

- Eight data transformations (r, r2, rv, rv2, l, l2, lv, and lv2) for RNA-seq data analysis were proposed in [13], these methods deal with the two common properties when it come to the count matrix generated in the quantification step, Sparsity and Skewness; Sparsity means that many counts in the count matrix are zero. Skewness means that the histogram of all counts in the count matrix is usually skewed.

### 3.2 Clustering methods based on normal distribution

#### *3.2.1 Hierarchical methods*

Hierarchical clustering method is the most popular method for gene expression data analysis. In hierarchical clustering, genes with similar expression patterns are grouped together and are connected by a series of branches (clustering tree or dendrogram). Experiments with similar expression profiles can also be grouped together using the same method. This clustering technique is divided into two types: agglomerative and divisive. In an agglomerative or bottom-up clustering method each observation is assigned to its own cluster. In a comparative study on Cancer data [14], three variants of Hierarchical Clustering Algorithms (HCAs): Single-Linkage (SL), Average-Linkage (AL) and Complete-Linkage (CL) with 12 distance measure have been used to cluster RNA-seq Samples. The same methods will be used in our study along with hierarchical clustering with Poisson distribution [15].

#### *3.2.2 k-medoids*

K-medoids is a partitional clustering algorithm proposed in 1987 by Kaufman and Rousseeuw. It is a variant of the K-means algorithm that is less sensitive to noise and outliers because it uses medoids as cluster centers instead of means that are easily influenced by extreme values. Medoids are the most centrally located objects of the clusters, with a minimum sum of distances to other points. After searching for k representative objects in a data set, the algorithm which is called Partitioning Around Medoids (PAM) assigns each object to the closest medoid in order to create clusters. Like in k-means the number of classes to be generated needs to be specified.

### 3.3 Model-based clustering

Yaqing Si et al. described a number of Model based clustering methods for RNA-seq data in their paper [16], these methods assume that data are generated by a mixture of probability distributions: Poisson distribution when only technical replicates are used and Negative binomial distribution when working with biological replicates. The first method they proposed is a model-based clustering method with the expectation-maximization algorithm (MB-EM) for clustering RNA-seq gene expression profile. The expectation-maximization algorithm is widely used in many computational biology applications, the authors in [17] explain how this algorithm works and when it is used. The second method is an initialization algorithm for cluster centers, the idea behind this method is to randomly choose one cluster center and then gradually add centers by selecting genes based on the distance between each gene and each of the selected centers. Two other stochastic algorithms have been proposed in this paper, a stochastic version of the expectation-maximization algorithm and a classification expectation maximization algorithm with simulated annealing. The last method in this paper is a model-Based Hybrid-Hierarchical Clustering Algorithm, it does not require to pre-specify the number of clusters to be generated as it is required by the previous methods. The authors propose to use agglomerative clustering starting with k0 clusters to speed up the calculation, then, it repeatedly identifies the two clusters that are closest together and merges the two most similar clusters. This method was called hybrid because it combines two steps: Obtaining the initial K0 clusters using one of the previous described algorithms then agglomerative clustering to build the hierarchical tree.

### 3.4 Classification and clustering algorithms of machine learning for RNA-seq data

Classification in machine learning is a supervised learning approach in which the algorithm learns from the data given to it and makes new observations, then applies the conclusions to new data. Clustering on the other hand is an unsupervised learning problem for grouping unlabeled features. The learning algorithm that learns the model from the training data and maps the input data to a specific class is called classifier, in the following section, we briefly present three widely used classifiers for grouping RNA-seq data.

- Random forests (RF): an ensemble method that trains a large number of individual decision trees, each tree gives a class prediction, the category that wins the majority votes is used as the final decision of the random forest model. The algorithm can perform both classification and regression tasks and has better accuracy among current algorithms.

- Support Vector Machine (SVM): one of the most popular supervised learning models, used for both classification and regression, the data points are separated using an optimal hyperplane or a set of hyperplanes in a multidimensional space with the maximum possible margin between support vectors.

- Poisson linear discriminant analysis: an approach used for the classification and clustering of RNA-seq data using a Poisson log linear model [15].

To test these algorithms, we used MLSeq (Machine learning interface for RNA-sequencing data) which is an R package including more than 80 machine learning algorithms and a pipeline to classify RNA-seq data including normalization, filtering and transformation steps [18].

### 3.5 Clustering with deep learning

Deep learning is also a technique that can be used to learn better data representation of high-dimensional data. The two recently published surveys [19, 20] present a taxonomy of existing deep clustering algorithms, by describing the different Neural Network Architecture that exists for feature representation, clustering loss function and Performance Evaluation Metrics for Deep Clustering. In [20], the authors categorize current deep clustering models into following three categories:

- Auto-Encoders Based Deep Clustering

- CDNN-Based Deep Clustering (feed-forward networks trained only by specific clustering)

- Generative Adversarial Network (GAN)

These approaches are already used in the analysis of RNA-seq data, for example, an unsupervised deep embedding algorithm that clusters single cell (scRNA-seq) data was proposed in [21], another paper use a Lasso model and a multilayer feed-forward artificial neural network to analyze RNA-Seq gene expression data [22]. In [23], the authors used a Deep Neural Network model from the R package h2o for cancer data classification and in [24], ladder networks were used for gene expression classification.

## 4. Clustering algorithms and software packages/tools corresponding to the algorithms

Clustering algorithms and software packages corresponding to the algorithms are shown in **Table 1**.

| Methods | Implementation in R |
|---|---|
| Hierarchical clustering | hclust() function in “stats” |
| k-means | “cluster”, “factoextra” |
| k-medoids | “cluster”, “factoextra” |
| SOM | “kohonen” |
| Model-based clustering with the expectation-maximization algorithm (MB-EM).<br>Stochastic version of the expectation-maximization algorithms (Deterministic annealing (DA) algorithm).<br>Classification expectation maximization (CEM) algorithm with simulated annealing (SA). | “MBCluster.Seq” |
| Machine learning algorithms | “MLSeq” |

**Table 1.**
*Clustering algorithms and software packages corresponding to the algorithms.*

## 5. Clustering of public RNA-seq data from recount2

Recount2 is a multi-experiment resource of analysis-ready RNA-seq gene and exon count datasets. It contains 2041 different studies and over 70,000 human RNA-seq [25]. We selected for our study four different datasets based on the number of samples and the number of classes. We then performed sample-based clustering on each dataset and compared the results to the classes in the phenotype table in recount2 to evaluate the performance of each method. The methods used to classify the data are 3 subtypes of the hierarchical clustering with the Euclidean distance, hierarchical clustering with a Poisson model and k-medoids.

### 5.1 Datasets

Description of the four datasets from recount2 is shown in **Table 2**.

### 5.2 Adjusted Rand Index

There are several similarity measures for cluster evaluation, we chose to work with the adjusted Rand index which is the corrected-for-chance version of the Rand index. It is a measure used in data clustering to evaluate the performance of a clustering method, by comparing the results of a clustering algorithm against known classes from external criteria [26]. In our study, we performed different sample-based classification method on four different datasets, after that, we compared the results to the class labels we associated to each sample based on the field "characterization of the samples" in the phenotype table in recount2, then we used the ARI for cluster validation.

### 5.3 Standard deviation

**Figures 4–6** are examples to show the standard deviation of the transformed data, across samples, against the mean, using the shifted logarithm transformation, the regularized log transformation and the variance stabilizing transformation.

| Dataset (accession) | Number of samples | Number of classes | Classes |
|---|---|---|---|
| SRP032789 | 20 | 4 | 17 breast tumor samples of three different subtypes:<br>• TNBC.<br>• Non-TNBC.<br>• HER2-positive. |
| SRP049097 | 54 | 4 | 3 subtypes of Leiomyosarcoma:<br>• 8 LMS cases from subtype I<br>• 6 cases from subtype II<br>• 3 cases from subtype III<br>• 7 cases of normal tissues |
| SRP042620 | 168 | 6 | • 28 breast cancer cell lines.<br>• 42 Triple Negative Breast Cancer (TNBC) primary tumors.<br>• 42 Estrogen Receptor Positive (ER+) and HER2 Negative Breast Cancer primary tumors.<br>• 30 uninvolved breast tissue samples that were adjacent to ER+ primary tumors.<br>• 5 breast tissue samples from reduction mammoplasty procedures performed on patients with no known cancer.<br>• 21 uninvolved breast tissue samples that were adjacent to TNBC primary tumors. |
| SRP044668 | 94 | 3 | • 39 contrast-enhancing glioma core samples.<br>• 36 non-enhancing FLAIR glioma margin samples.<br>• 17 non-neoplastic brain tissue samples. |

**Table 2.**
*Description of the four datasets from recount2.*

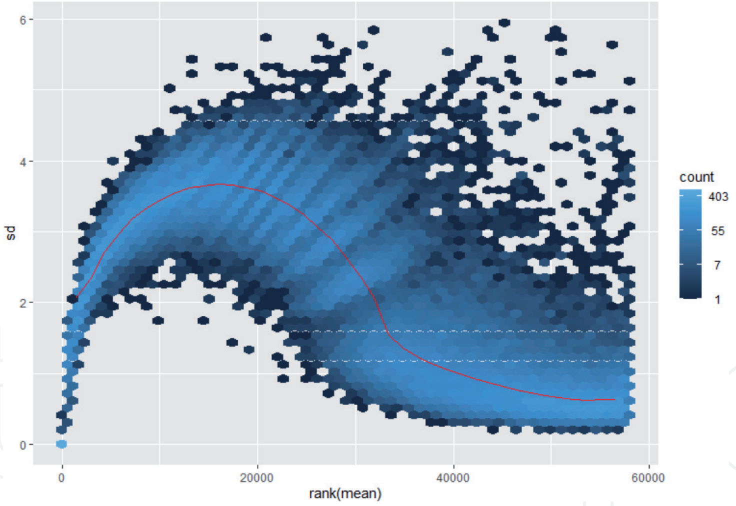

**Figure 4.**
*Standard deviation of the transformed data using the shifted logarithm transformation.*

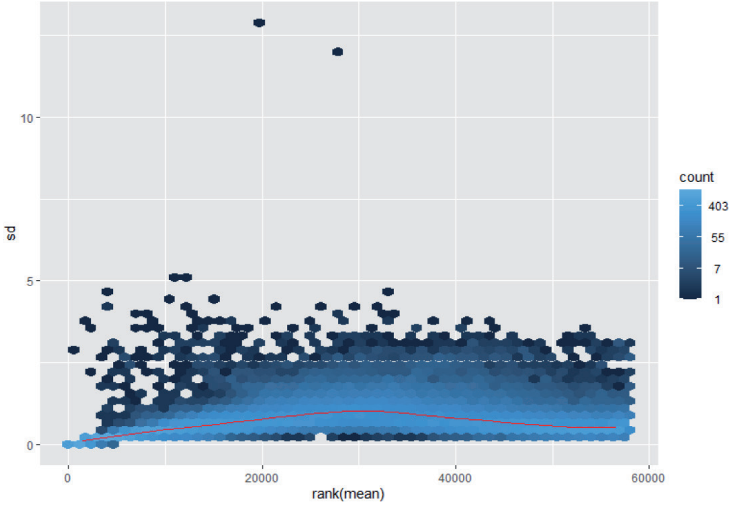

**Figure 5.**
*Standard deviation of the transformed data using the regularized log transformation.*

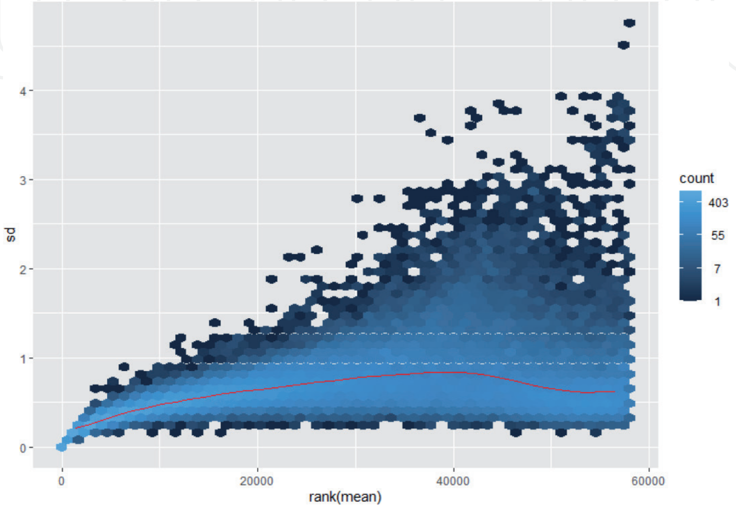

**Figure 6.**
*Standard deviation of the transformed data using the variance stabilizing transformation.*

| hclus (complete) | hclust (single) | hclust (average) | hclust (complete) | k-medoids |
|---|---|---|---|---|
| Euclidean | Euclidean | Euclidean | Poisson distance | Euclidean |
| 0.4146015 | 0.3818763 | 0.4146015 | 0.4146015 | 0.6798897 |

**Table 3.**
*Performance of clustering methods (SRP032789).*

| hclus (complete) | hclust (single) | hclust (average) | hclust (complete) | k-medoids |
|---|---|---|---|---|
| Euclidean | Euclidean | Euclidean | Poisson distance | Euclidean |
| 0.02880412 | −0.003409256 | 0.0005777741 | 0.1874828 | 0.2791547 |

**Table 4.**
*Performance of clustering methods (SRP049097).*

| hclus (complete) | hclust (single) | hclust (average) | hclust (complete) | k-medoids |
|---|---|---|---|---|
| Euclidean | Euclidean | Euclidean | Poisson distance | Euclidean |
| 0.1944569 | 0.005551586 | 0.1285448 | 0.1468464 | 0.2579758 |

**Table 5.**
*Performance of clustering methods (SRP042620).*

| hclus (complete) | hclust (single) | hclust (average) | hclust (complete) | k-medoids |
|---|---|---|---|---|
| Euclidean | Euclidean | Euclidean | Poisson distance | Euclidean |
| 0.2379903 | −0.007755123 | 0.399417 | 0.2657942 | 0.3771837 |

**Table 6.**
*Performance of clustering methods (SRP044668).*

## 5.4 Machine learning classification

Three widely used machine learning algorithms were used for the classification of the four datasets, Random forests, support vector machine and Poisson linear discriminant analysis. To perform this analysis, we first split the data into two parts as training and test sets, with 70% of samples for the training dataset, and the remaining 30% samples for the testing dataset, the training set is used to fit the parameters of the model, that is used thereafter to predict the responses for the observations in the test dataset. Normalization was applied with Deseq median ratio method and the variance stabilizing transformation was applied for the normalization of the dataset. The model was trained using 5-fold cross validation repeated 2 times. The number of levels for tuning parameters is set to 10.

## 5.5 Results

| Classifier | Accuracy | Kappa |
|---|---|---|
| rf | 1 | 1 |
| SVM | 0.6667 | 0.5 |
| PLDA | 1 | 1 |

**Table 7.**
*Classification results for SRP032789 data.*

| Classifier | Accuracy | Kappa |
|---|---|---|
| rf | 0.8235 | 0.765 |
| SVM | 0.7647 | 0.6909 |
| PLDA | 0.7647 | 0.6866 |

**Table 8.**
*Classification results for SRP049097 data.*

| Classifier | Accuracy | Kappa |
|---|---|---|
| rf | 0.9412 | 0.9249 |
| SVM | 0.5882 | 0.4685 |
| PLDA | 0.7843 | 0.7267 |

**Table 9.**
*Classification results for SRP042620 data.*

| Classifier | Accuracy | Kappa |
|---|---|---|
| rf | 0.8214 | 0.7271 |
| SVM | 0.6786 | 0.5218 |
| PLDA | 0.7143 | 0.5573 |

**Table 10.**
*Classification results for SRP044668 data.*

### 5.6 Computational time

All experiments are performed on a machine with 16 GB RAM, 1024 GB hard disk running with a windows operating system and MLSeq R package.

| | SRP032789 | SRP049097 | SRP042620 | SRP044668 |
|---|---|---|---|---|
| rf | 176.67 | 781.31 | 4234.89 | 1412.19 |
| SVM | 1080.92 | 2333.52 | 6645.21 | 1597.89 |
| PLDA | 31.45 | 60.93 | 234.98 | 72.66 |

**Table 11.**
*Computational time in seconds.*

## 6. Discussion and conclusion

Clustering the samples of the three datasets to the sub-classes defined in the phenotype table of recounts2 was not easy. We first tried to visualize the separation between the subtypes using principal component analysis (**Figures 7** and **8**–**10**), then using 4 variants of the hierarchical clustering and k-medoids we classified the samples of each dataset (**Figures 11** and **12** show the hierarchical clustering plots of the dataset SRP032789). The performance of the 5 methods was different depending on the dataset (**Tables 3**–**5**), making it impossible to make a general system of recommendation. However, we can see that the k-medoid method has relatively

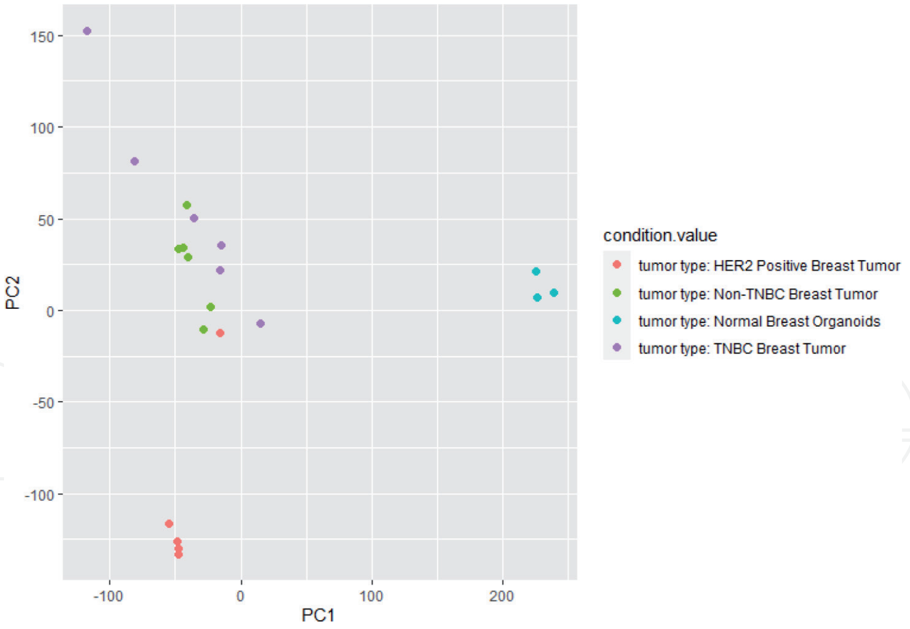

**Figure 7.**
*PCA of data from the study SRP032789.*

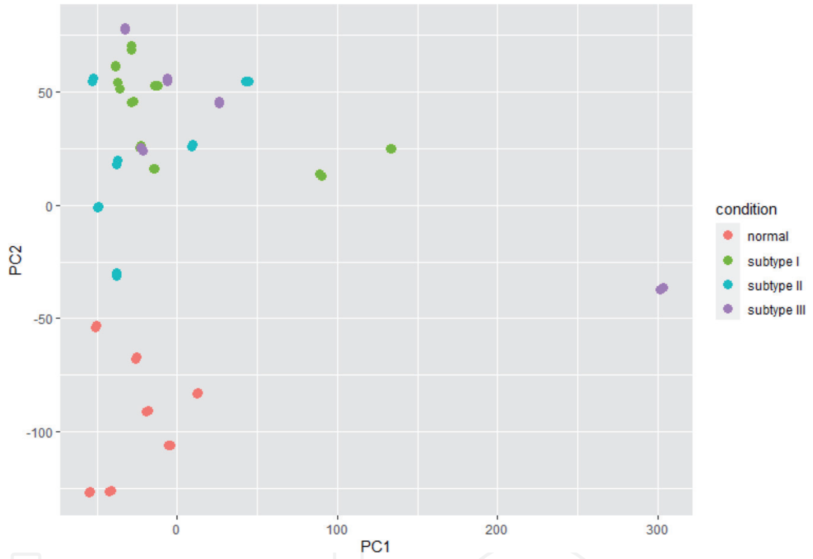

**Figure 8.**
*PCA of the data from the study SRP049097.*

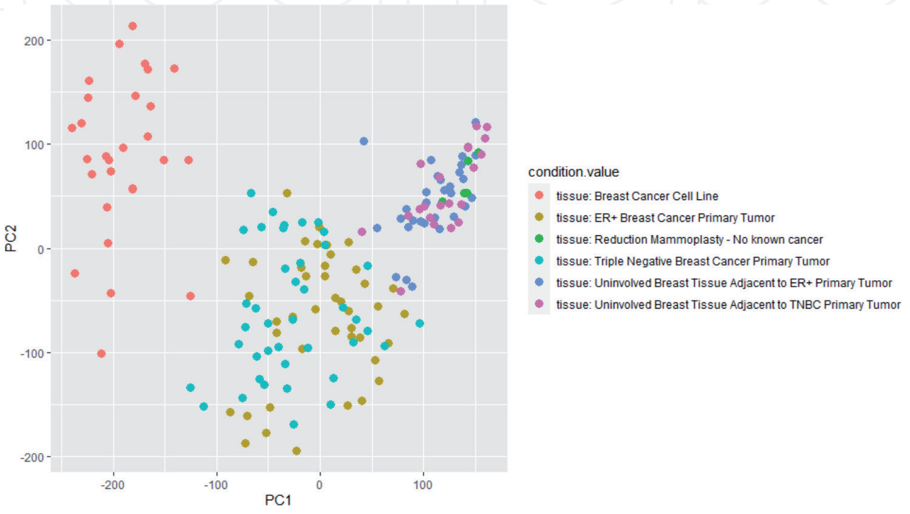

**Figure 9.**
*PCA of the data from the study SRP042620.*

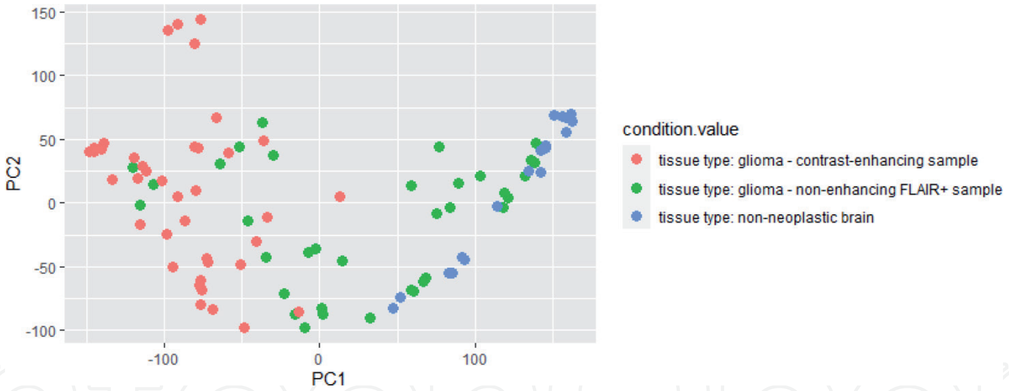

**Figure 10.**
*PCA of the data from the study SRP044668.*

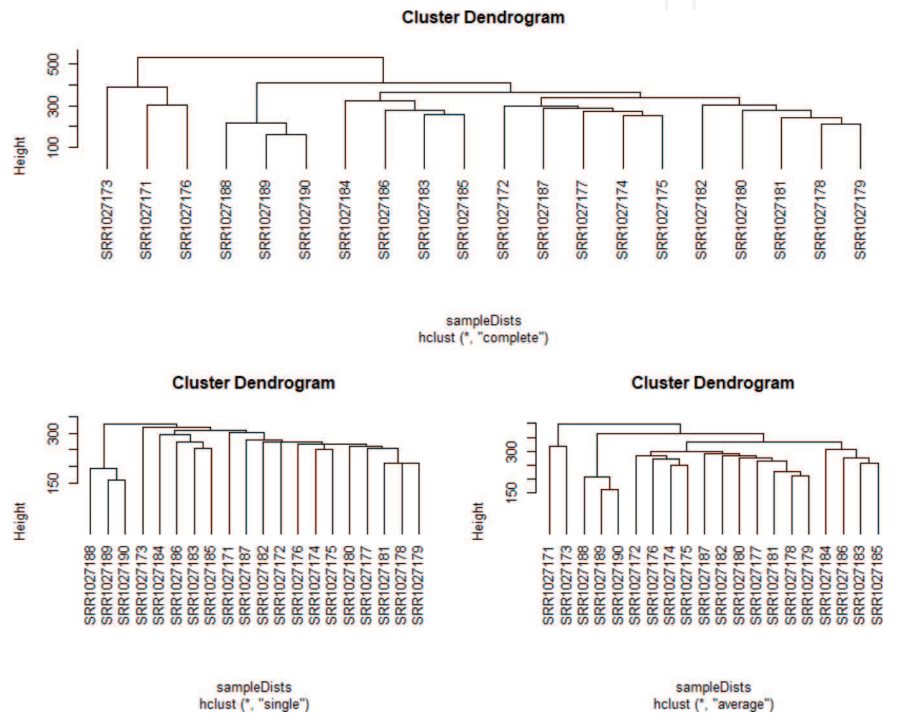

**Figure 11.**
*Dendrograms obtained for the dataset from SRP032789 study using three variants of the hierarchical clustering method with the Euclidean distance.*

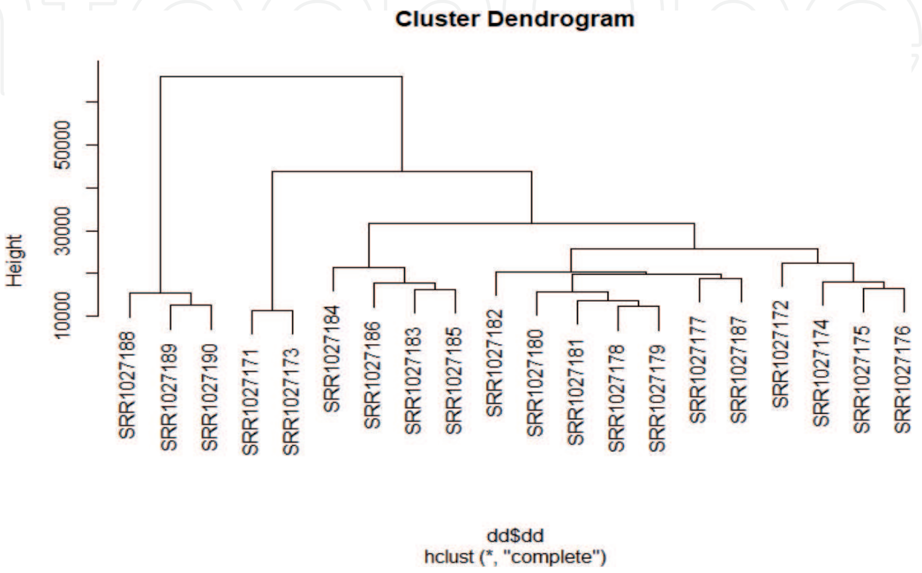

**Figure 12.**
*Dendrograms obtained for the dataset from the study SRP032789 using the hierarchical clustering method with the Poisson distance.*

better performance than the other methods for all the datasets. In the second part of the study, we compared a few machine learning methods used for the classification of RNA-seq data. The performance of the models surpasses the classical methods used before, also RF and PLDA performed better than SVM which does not perform very well when the data set is large and has noise. Note that the model accuracies given in this study should not be considered as a generalization. The results can depend on several criteria: normalization and transformation methods, gene-wise overdispersions, outliers, number of classes etc. (**Tables 6–11**).

## Conflict of interest

The authors declare no conflict of interest.

## Author details

Ismail Jamail* and Ahmed Moussa
System and Data Engineering Team - SDET, Tangier, Morocco

*Address all correspondence to: jamail@ensat.ac.ma

## References

[1] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet [Internet]. 2009 Jan; 10(1):57-63. Available from: http://dx.doi.org/10.1038/nrg2484

[2] Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

[3] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics [Internet]. 2012 Oct 25;29(1):15-21. Available from: http://dx.doi.org/10.1093/bioinformatics/bts635

[4] Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods [Internet]. 2015 Mar 9;12(4):357-60. Available from: http://dx.doi.org/10.1038/nmeth.3317

[5] Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics [Internet]. 2009 Mar 16;25(9):1105-11. Available from: http://dx.doi.org/10.1093/bioinformatics/btp120

[6] Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics [Internet]. 2014 Sep 25;31(2):166-9. Available from: http://dx.doi.org/10.1093/bioinformatics/btu638

[7] Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics [Internet]. 2013 Nov 13;30(7):923-30. Available from: http://dx.doi.org/10.1093/bioinformatics/btt656

[8] Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol [Internet]. 2016 Apr 4;34(5):525-7. Available from: http://dx.doi.org/10.1038/nbt.3519

[9] Li P, Piao Y, Shon HS, Ryu KH. Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. BMC Bioinformatics [Internet]. 2015 Oct 28;16(1). Available from: http://dx.doi.org/10.1186/s12859-015-0778-7

[10] Vidman L, Källberg D, Rydén P. Cluster analysis on high dimensional RNA-seq data with applications to cancer research - An evaluation study. Nazarov PV, editor. PLoS ONE [Internet]. 2019 Dec 5;14(12):e0219102. Available from: http://dx.doi.org/10.1371/journal.pone.0219102

[11] Feng, Changyong et al. “Log-transformation and its implications for data analysis.” Shanghai archives of psychiatry vol. 26,2 (2014): 105-9. doi:10.3969/j.issn.1002-0829.2014.02.009

[12] Durbin BP, Hardin JS, Hawkins DM, Rocke DM. A variance-stabilizing transformation for gene-expression microarray data. Bioinformatics [Internet]. 2002 Jul 1;18(Suppl 1):S105-10. Available from: http://dx.doi.org/10.1093/bioinformatics/18.suppl_1.S105

[13] Zhang Z, Yu D, Seo M, Hersh CP, Weiss ST, Qiu W. Novel Data Transformations for RNA-seq Differential Expression Analysis. Sci Rep [Internet]. 2019 Mar 18;9(1). Available from: http://dx.doi.org/10.1038/s41598-019-41315-w

[14] Jaskowiak PA, Costa IG, Campello RJGB. Clustering of RNA-Seq samples: Comparison study on

cancer data. Methods [Internet]. 2018 Jan;132:42-9. Available from: http://dx.doi.org/10.1016/j.ymeth.2017.07.023

[15] Witten DM. Classification and clustering of sequencing data using a Poisson model. Ann Appl Stat [Internet]. 2011 Dec;5(4):2493-518. Available from: http://dx.doi.org/10.1214/11-AOAS493

[16] Si Y, Liu P, Li P, Brutnell TP. Model-based clustering for RNA-seq data. Bioinformatics [Internet]. 2013 Nov 4;30(2):197-205. Available from: http://dx.doi.org/10.1093/bioinformatics/btt632

[17] Do CB, Batzoglou S. What is the expectation maximization algorithm? Nat Biotechnol [Internet]. 2008 Aug;26(8):897-9. Available from: http://dx.doi.org/10.1038/nbt1406

[18] Goksuluk D, Zararsiz G, Korkmaz S, Eldem V, Zararsiz GE, Ozcetin E, et al. MLSeq: Machine learning interface for RNA-sequencing data. Computer Methods and Programs in Biomedicine [Internet]. 2019 Jul;175:223-31. Available from: http://dx.doi.org/10.1016/j.cmpb.2019.04.007

[19] Aljalbout, Elie et al. Clustering with Deep Learning: Taxonomy and New Methods. *ArXiv* abs/1801.07648. 2018

[20] Min E, Guo X, Liu Q, Zhang G, Cui J, Long J. A Survey of Clustering With Deep Learning: From the Perspective of Network Architecture. IEEE Access [Internet]. 2018;6:39501-14. Available from: http://dx.doi.org/10.1109/ACCESS.2018.2855437

[21] Li X, Wang K, Lyu Y, Pan H, Zhang J, Stambolian D, et al. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. Nat Commun [Internet]. 2020 May 11;11(1). Available from: http://dx.doi.org/10.1038/s41467-020-15851-3

[22] Urda D, Montes-Torres J, Moreno F, Franco L, Jerez JM. Deep Learning to Analyze RNA-Seq Gene Expression Data. In: Advances in Computational Intelligence [Internet]. Springer International Publishing; 2017. p. 50-9. Available from: http://dx.doi.org/10.1007/978-3-319-59147-6_5

[23] Sharma A, Rani R. An Optimized Framework for Cancer Classification Using Deep Learning and Genetic Algorithm. j med imaging hlth inform [Internet]. 2017 Dec 1;7(8):1851-6. Available from: http://dx.doi.org/10.1166/jmihi.2017.2266

[24] Golcuk G, Tuncel MA, Canakoglu A. Exploiting Ladder Networks for Gene Expression Classification. In: Bioinformatics and Biomedical Engineering [Internet]. Springer International Publishing; 2018. p. 270-8. Available from: http://dx.doi.org/10.1007/978-3-319-78723-7_23

[25] Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, et al. Reproducible RNA-seq analysis using recount2. Nat Biotechnol [Internet]. 2017 Apr;35(4):319-21. Available from: http://dx.doi.org/10.1038/nbt.3838

[26] Santos JM, Embrechts M. On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification. In: Artificial Neural Networks – ICANN 2009 [Internet]. Springer Berlin Heidelberg; 2009. p. 175-84. Available from: http://dx.doi.org/10.1007/978-3-642-04277-5_18