multiClust: An R-package for Identifying Biologically Relevant Clusters in Cancer Transcriptome Profiles

Nathan Lawlor[1,*], Alec Fabbri[2,*], Peiyong Guan[3,4], Joshy George[5] and R. Krishna Murthy Karuturi[5]

[1]Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA. [2]Department of Biomedical Engineering, University of Connecticut, Storrs, CT, USA. [3]Genome Institute of Singapore, A*STAR (Agency for Science, Technology and Research), Singapore. [4]School of Computer Science and Engineering, Nanyang Technological University, Singapore. [5]The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA. *These authors contributed equally to this work.

ABSTRACT: Clustering is carried out to identify patterns in transcriptomics profiles to determine clinically relevant subgroups of patients. Feature (gene) selection is a critical and an integral part of the process. Currently, there are many feature selection and clustering methods to identify the relevant genes and perform clustering of samples. However, choosing an appropriate methodology is difficult. In addition, extensive feature selection methods have not been supported by the available packages. Hence, we developed an integrative R-package called *multiClust* that allows researchers to experiment with the choice of combination of methods for gene selection and clustering with ease. Using *multiClust*, we identified the best performing clustering methodology in the context of clinical outcome. Our observations demonstrate that simple methods such as variance-based ranking perform well on the majority of data sets, provided that the appropriate number of genes is selected. However, different gene ranking and selection methods remain relevant as no methodology works for all studies.

KEYWORDS: R software, gene selection, data clustering, clinical outcome

CITATION: Lawlor et al. *multiClust*: An R-package for Identifying Biologically Relevant Clusters in Cancer Transcriptome Profiles. *Cancer Informatics* 2016:15 103–114 doi: 10.4137/CIN.S38000.

TYPE: Software or Database Review

RECEIVED: November 27, 2015. **RESUBMITTED:** March 28, 2016. **ACCEPTED FOR PUBLICATION:** April 03, 2016.

ACADEMIC EDITOR: J. T. Efird, Editor in Chief

PEER REVIEW: Six peer reviewers contributed to the peer review report. Reviewers' reports totaled 2,103 words, excluding any confidential comments to the academic editor.

FUNDING: The authors thank The Jackson Laboratory for Genomic Medicine and the University of Connecticut for funding this study. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: Krishna.karuturi@jax.org

Paper subject to independent expert blind peer review. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Published by Libertas Academica. Learn more about this journal.

Background

Cancer and many other diseases are caused by the aberrant alterations of genes responsible for proper cell growth, differentiation, and division at different levels of regulation, for example, genomic mutations, transcription, and translation. There is considerable heterogeneity in every cancer type, and improvement in the identification of homogeneous subgroups of patients would significantly improve patient care and decrease morbidity and mortality due to disease.[1–4] Examining the gene expression patterns of cancer has proven to be effective in identifying novel clinically relevant subgroups and genetic signatures.[4–6] Common clinical outcome variables of interest are distance metastasis, tumor recurrence, and patient survival. These variables further help in understanding the underlying fundamental biology of regulating and selecting appropriate patient care options for different cancer subgroups. For example, in the study by Calon et al, gene expression profiling analysis of colorectal cancer samples identified elevated expression of transforming growth factor (TGFB1) and TGFB3 in poor-prognosis cases. The transcriptomic profiling results of these experiments lead to the identification of a therapeutic TGF-β inhibitor, LY2157299, capable of reducing disease progression in an immunodeficient mouse model.[6]

Clustering algorithms are used to identify patterns in gene expression and discover the clinically relevant subgroups. These algorithms use the expression of selected genes and choice of distance/similarity metrics to discover samples that exhibit similar transcriptomic profiles, and hence the sample subgroups are identified. A variety of clustering algorithms have been applied for such purposes in cancer studies. Hierarchical clustering is the most common form of clustering used in the literature[7–10] followed by *k*-means (partitional) clustering.[7,8,10,11] Biclustering[12] and consensus clustering[11] have also been applied to discover the patient subgroups with distinctive transcriptional profiles from each other, although these methods are not most commonly used. Additional clustering methods include but are not limited to density-based, grid-based, correlation, spectral, gravitational, and herd clustering.[13–15]

Hierarchical clustering is one of the most popularly employed methods of clustering due to its amenability for visualization of the relationships and because there is no need to set the number of clusters a priori.[7] Hierarchical clustering divides observations into clusters and creates a tree diagram or a dendrogram where each node of the tree represents a cluster. This algorithm is used to visualize each sample's/gene's relationship with other samples/genes within the cluster and among all clusters.[7,11] This method of clustering is dependent, in addition to the right choice of genes, on the choice of distance metric or how the distance between data points is defined for clustering. A common distance metric used in hierarchical clustering is Euclidean distance, or the straight line distance between two points in geometric space. In addition, Pearson's correlation coefficient and unsigned Pearson's correlation coefficient have also been popularly used. Ultimately, using different distance metrics can influence the shape and organization of the clusters.[16,17] The linkage method is another important component of hierarchical clustering that determines the distance between two clusters. Many linkage criteria are available, which include complete, single, average linkage, and Ward's minimum variance method.[16,18] One of the most common linkage methods is average linkage, that is, the distance between two clusters is calculated as the average distance of all pairs of points in the clusters. However, there is no clear evidence for the superiority of the average linkage over the other linkage methods.[16,19–21]

Another important class of data clustering is partitioning. Partitioning methods such as *k*-means clustering divide the samples into a predetermined number of clusters. The number of clusters is generally chosen at random and samples are assigned to the cluster with the closest centroid, or average geometric position of all the samples in the space. After all the samples are assigned to a cluster, the centroids are readjusted to the average position of all the observations in the cluster. This process is repeated until the cluster centroids do not change.[22,23]

An important step toward applying the above clustering algorithms on transcriptomic data is feature (gene or probe) selection, or selecting a subset of variables or features relevant to clustering.[24,25] Feature selection is an integral part of this discovery process because the quality and relevance of clustering are sensitive to the selected features or genes. Feature selection is needed in the context of whole transcriptomic profile analysis because there are many irrelevant genes that may dilute signal required for clustering tumor samples.[26] A typical feature selection method involves ranking genes based on a criterion and selecting a certain number of genes at the top of the ranked list (Fig. 1). Some frequently used gene ranking methods include variance-based ranking, Gaussian Mixture Modeling, coefficient of variation, and mean-variance methods.[25,27–29] However, these methods remain relevant as none of them have been shown to be superior to the others and are often inconsistent in identifying biologically relevant clusters across different types of data.

A challenge in feature selection is determining which method is the most effective at identifying the subset of relevant genes.[8,9,24] Another challenge in most feature selection methods is determining how many features, or genes in the dataset, are deemed relevant. Current approaches have been arbitrary with no clear attention paid to the effect of the number of genes on the quality of the resultant clustering algorithm. Choosing the appropriate method is difficult, as the recent work[9] demonstrates that no method is a clear winner based on generic clustering metrics such as F-score and Entropy.

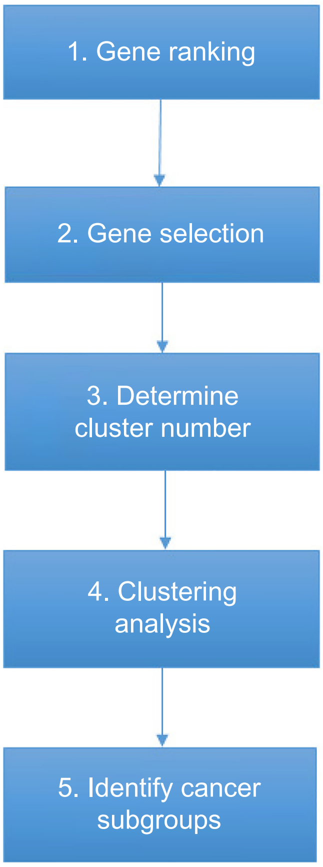

Figure 1. Flowchart depicting the commonly used steps in a typical cluster analysis of cancer gene expression data. First, genes are ranked based on a chosen statistical metric expected to capture the relevance of genes. Second, the number of relevant genes is selected. Third, the appropriate number of clusters is chosen to discretize genes and samples. Fourth, clustering is performed using either a hierarchical or a *k*-means clustering algorithm. Lastly, Cox Proportional Hazard Model tests are used to correlate sample clustering with clinical outcome and identify cancer subgroups.

There is currently no package that facilitates the choice of gene selection methods from a variety of options (ranking and choice of number of genes) in combination with commonly used clustering algorithms. Hence, in this review, we describe our R-package that offers gene ranking and selection methods from a wide variety of methods used in the literature and choice of hierarchical or *k*-means clustering algorithm. The quality of gene selection and clustering is studied in the context of clinical relevance of clustering, specifically patient clinical outcome, in cancer studies rather than generic score of quality of clustering. Our package and methodologies are focused on clinical relevance in contrast to the abstract clustering metrics used in the recent work.[9] Ultimately, our software combines aspects of gene selection, ranking, clustering, and clinical outcome analysis conveniently into a single package. The suite of tools available in our package will help researchers to identify biologically homogenous sample groups based on transcriptomic profiles and determine the association of these groupings with clinical outcome.

Methods and Package Workflow

In this section, we describe the components and workflow of our R-package *multiClust*. Our integrative R-package contains:

1. An in-depth vignette explaining how to use our package and obtain publicly available gene expression datasets with clinical outcome information
2. Four gene ranking options that order genes based on different statistical criteria:
   a. CV_Rank
   b. CV_Guided (novel method)
   c. SD_Rank
   d. Poly
3. Four different ways to select the number of genes:
   a. Fixed
   b. Percent
   c. Poly (novel method)
   d. GMM
4. Two ways to determine the cluster number:
   a. Fixed
   b. Gap statistic
5. Two clustering algorithms:
   a. Hierarchical clustering
   b. *k*-means clustering
6. A function to calculate the average gene expression in each sample cluster
7. A function to correlate sample clusters with clinical outcome

A summary of the necessary steps to use our package is depicted in Figure 2.

If users intend to obtain gene expression and clinical data from public databases such as Gene Expression Omnibus (GEO),[30] they should refer to our package vignette for more guidelines (Supplementary Package Vignette File). Our package is compatible for use with both microarray and RNA-seq gene expression data. This vignette includes instructions about how to obtain gene expression and clinical data from GEO using R as well as gene expression normalization procedures for both data types. Common methods of normalization used for GEO microarray datasets include robust multi-array average (RMA) normalization, or quantile normalization and log2 scaling.[31,32] When using RNA-seq data of fragments per kilobase of transcript per million mapped reads (FPKM) values, a common method of normalization is to log2 transform the data; however, other systematic methods are available.[33] After downloading and preprocessing the gene expression dataset (Fig. 2), users have the choice of using our package function or other available R functions to load the gene expression matrix into an R object. Users who have their own gene expression datasets may also use our package function to load their data into R. However, if the expression data are already loaded as an R object, the user may proceed to the gene ranking and selection steps in the package.

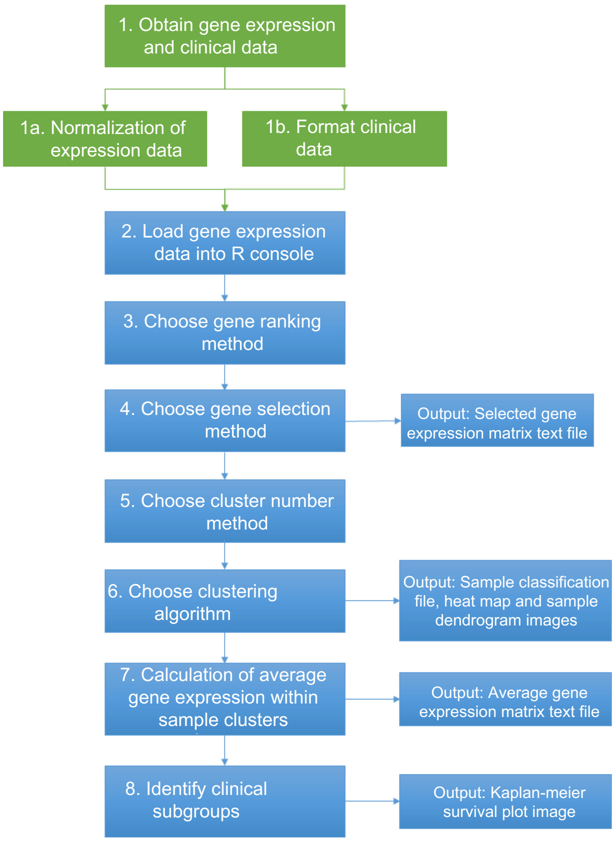
1. Obtain gene expression and clinical data
1a. Normalization of expression data
1b. Format clinical data
2. Load gene expression data into R console
3. Choose gene ranking method
4. Choose gene selection method
Output: Selected gene expression matrix text file
5. Choose cluster number method
6. Choose clustering algorithm
Output: Sample classification file, heat map and sample dendrogram images
7. Calculation of average gene expression within sample clusters
Output: Average gene expression matrix text file
8. Identify clinical subgroups
Output: Kaplan-meier survival plot image

Figure 2. Flowchart of the *multiClust* workflow and the output files produced at each step. The boxes in green refer to the steps to be performed before using *multiClust*. These steps include obtaining the gene expression and clinical data, normalizing the expression data, and formatting the clinical data. Boxes in blue are the steps that are performed while using the *multiClust* package.

After obtaining the gene expression and clinical data and loading them into R, the next step is to specify the gene ranking option. As summarized in Table 1, there are four different gene ranking methods to choose from within *multiClust* including a novel method called “CV_Guided”. Following gene ranking, the user chooses the desired gene selection method. In Table 2, there are four different options to choose from. The “Fixed” and “Percent” methods require the user to input a positive integer corresponding to a specific amount of genes to be selected from the dataset. The “Poly” and “GMM” methods are adaptive gene selection methods in the sense that they select a different number of genes for each dataset based on a statistical criterion. These two adaptive methods fit mathematical functions to the gene expression dataset to identify an appropriate number of genes to be selected. Therefore, users do not need to specify the number of genes to be selected.

Table 1. Choice of gene ranking methods offered in *multiClust.*

GENE RANKING METHOD	DESCRIPTION	REFERENCES
CV_Rank	Genes are ranked from highest to lowest coefficient of variation values.	Li et al.[56] and Hall et al.[57]
CV_Guided	Every gene within the set is then plotted on a mean vs. standard deviation graph. A line is plotted starting from the origin with a slope of the CV of the entire dataset. The mean and standard deviation cutoffs move along this line in a positive direction away from the origin until an equal or less than the number of desired genes is above the cutoffs.	Modification to "CV_Rank" method
SD_Rank	Genes are ranked from highest to lowest standard deviation values.	Miller et al 2005,[27] Tothill et al.[46], Gibbons and Roth 2002,[58] Eisen 2002[59]
Poly	Sequentially fits three second-degree polynomial functions of mean and standard deviation to the dataset to determine the most variable genes.	Kharchenko 2014[29]

The Poly algorithm is a novel iterative method akin to a robust regression method that uses a series of three second-order polynomial regressions to filter out genes in each dataset. Each of the three linear regressions maps the expected standard deviation of a gene given its mean. After each regression is calculated, the genes with a standard deviation below the expected value will be filtered out. When the next regression is being calculated, the genes that are filtered out are not included in the calculations. This model is more selective than the constant standard deviation cutoff filter, as each standard deviation cutoff function is dependent on the mean of the remaining genes. The filtering process was carried out specifically three times, as it removes more than 90% of genes with low variance.[29]

In the GMM method, Gaussian Mixture Models are used to calculate the number of Gaussian distributions (K) for each gene within a dataset using the "mclust" package.[34] For genes with multiple levels of expression, the statistical outputs are used to compute a modified t-score (T_{Rank}). These scores along with simulated data scores are used to identify the number of genes at a 1% false discovery rate (FDR); however, users have the ability to adjust this FDR cutoff. The T_{Rank} score for a gene (g^*_i) is computed as shown below:

$$T_{Rank}(g_i^*) = \sum_{\substack{a,b\in\{1,2\ldots K_i\}\\ a\neq b}} \frac{|\mu_{ia}-\mu_{ib}|}{\sqrt{\sigma_{ia}^2+\sigma_{ib}^2+s_0^2}} \frac{\pi_{ia}\pi_{ib}}{(\pi_{ia}+\pi_{ib})^2}$$

The first term $\frac{|\mu_{ia}-\mu_{ib}|}{\sqrt{\sigma_{ia}^2+\sigma_{ib}^2+s_0^2}}$ of T_{Rank} (g^*_i) is the standardized difference between the mean of the *a*th and *b*th Gaussians; it measures the differentiability of the two Gaussians, whereas the second term $\frac{\pi_{ia}\pi_{ib}}{(\pi_{ia}+\pi_{ib})^2}$ of T_{Rank} (g^*_i) measures the relative distribution of the samples into the two groups, which reaches its maximum value when the samples are evenly distributed since $\sum_{k=1}^{K_i} \pi_{ik} = 1$ for each gene. In the second term, π_{ia} and π_{ib} represent the mixing proportions of the *a*th and *b*th Gaussians for gene "i".

Figure 3 portrays scatterplot images that illustrate the top 1000 genes selected by the four ranking methods. In all of the scatterplots, the selected genes are plotted against the total genes based on their mean and standard deviation values across each sample. Selected genes are depicted in red, while filtered genes are represented in black. Once the

Table 2. Choice of four gene selection methods available in *multiClust*.

GENE SELECTION METHOD	DESCRIPTION	REFERENCES
Fixed	Users define a positive integer of genes they want to select from their gene expression dataset.	Tothill et al 2008[46] and Jorissen et al 2009[35]
Percent	Users define a positive integer between 1 and 100 to select a percentage of total genes of the dataset.	Simon and Lam 2002[60]
Poly (Adaptive)	This method fits three second-degree polynomial functions of mean and standard deviation to the dataset. Afterward, it returns a positive integer indicating the number of genes in the dataset with mean and standard deviation values higher than the fit polynomial functions.	Kharchenko 2014[29]
GMM (Adaptive)	Uses Gaussian mixture modeling (GMM) and a score based on each genes' mean, variance, mixing proportion, and Gaussian assignment to select genes. GMM is used to model the individual gene expression across the different samples and categorize them into different levels of expression. After determining the best fitting GMM's for each gene, relevant genes are selected using a metric that measures the standardized difference between the mean of the genes in the dataset.	Fraley et al 2012[28] and Fraley et al 2002[61]

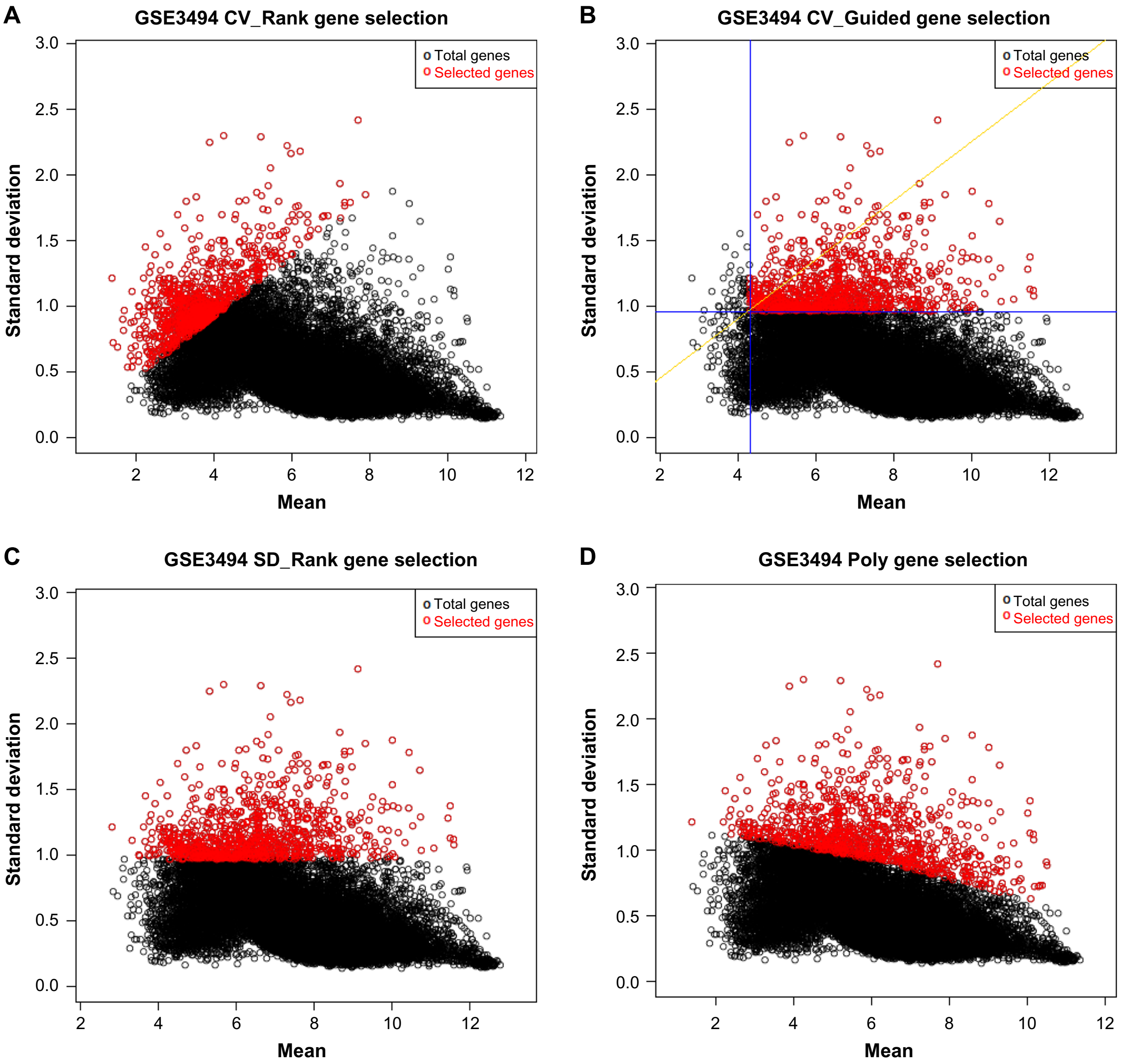

Figure 3. Illustration of choice of relevant genes by various gene ranking methods on a mean-variance plot using the GSE3494 breast cancer dataset.[27] (**A**) CV_Rank method. (**B**) CV_Guided, the yellow line is the coefficient of variation for the dataset, while the blue lines represent the mean and standard deviation cutoffs for a given number of genes to be selected. (**C**) SD_Rank. (**D**) Poly method.

gene ranking and selection process is completed, *multiClust* produces a text file for the user that contains a matrix of the samples and selected genes. The rows of this matrix represent the selected genes and the columns represent the tumor samples.

After selecting genes, users have the choice of determining the number of clusters to group samples into two different methods as summarized in Table 3. The first method "Fixed" requires the user to input the number of clusters they would like to partition samples into. This number is

Table 3. Summary of the methods available to identify the number of clusters in *multiClust.*

CLUSTER NUMBER METHOD	DESCRIPTION	REFERENCES
Fixed	User defines a positive integer (>1) to specify the number of clusters to divide samples into.	Marisa et al 2013,[4] Calon et al 2015,[6] Jorissen et al 2009[35]
Gap Statistic	This method (from the package "cluster") determines the number of clusters to split the samples into by calculating a goodness of clustering measure by using Gap Statistic.	Maechler 2015,[36] Tibshirani et al 2001[37]

generally arbitrary; however, in many studies, the number of clusters is often chosen as the number of known molecular subtypes of that cancer.[4,6,35] In contrast, the second method, "Gap Statistic" uses a function from the R package "cluster" to calculate a goodness of clustering measure for the tumor samples.[36] Following this calculation, the cluster number with the highest goodness of clustering score is chosen to discretize samples.[37]

Next, users have the option to cluster samples via hierarchical clustering or *k*-means clustering. For the hierarchical clustering option, our package outputs a PDF image of the sample dendrogram and Java TreeView heat map files of clustered genes and samples. The Java TreeView ATR, GTR, and CDT files can be loaded into Java TreeView[38] to view the clustering of genes and samples. Examples of these sample dendrograms and heat map are given in Figures 4 and 5. For both the clustering options, *multiClust* outputs a spreadsheet specifying the cluster assignment of each sample.

After the clustering of genes and samples, *multiClust* calculates the average expression of each gene within each sample cluster number. Following this step, a text file containing a matrix of the average expression for each selected gene in each cluster is produced. This will be useful in defining the gene expression signature for each sample cluster identified.

Lastly, our package uses Cox Proportional Hazard Models to test the clinical relevance of the sample clusters and thereby identify novel cancer subgroups. After cluster analysis of selected genes and tumor samples, a Kaplan–Meier plot is produced (Fig. 6) portraying patient survival probability over time. A Cox proportional hazard model test from the R package "survival" is performed for each clustered dataset and clinical outcome measure provided.[39,40] Several of the gene expression datasets used in our study had multiple clinical outcome measures available as shown in Table 1. As a result, we performed multiple survival analysis tests on these datasets and only one survival test on those datasets with one clinical outcome measure. Following each survival test, a *P*-value is outputted to specify if there was a significant correlation between the patient sample grouping and clinical outcome. A *P* value ≤ 0.05 is regarded as significant.

Results

In this study, we used *multiClust* to identify the best performing clustering methodology (gene ranking and selection methods in conjunction with both clustering algorithms) in context to clinically relevant clusters in 14 cancer datasets with 21 clinical outcome measures in total. Detailed results

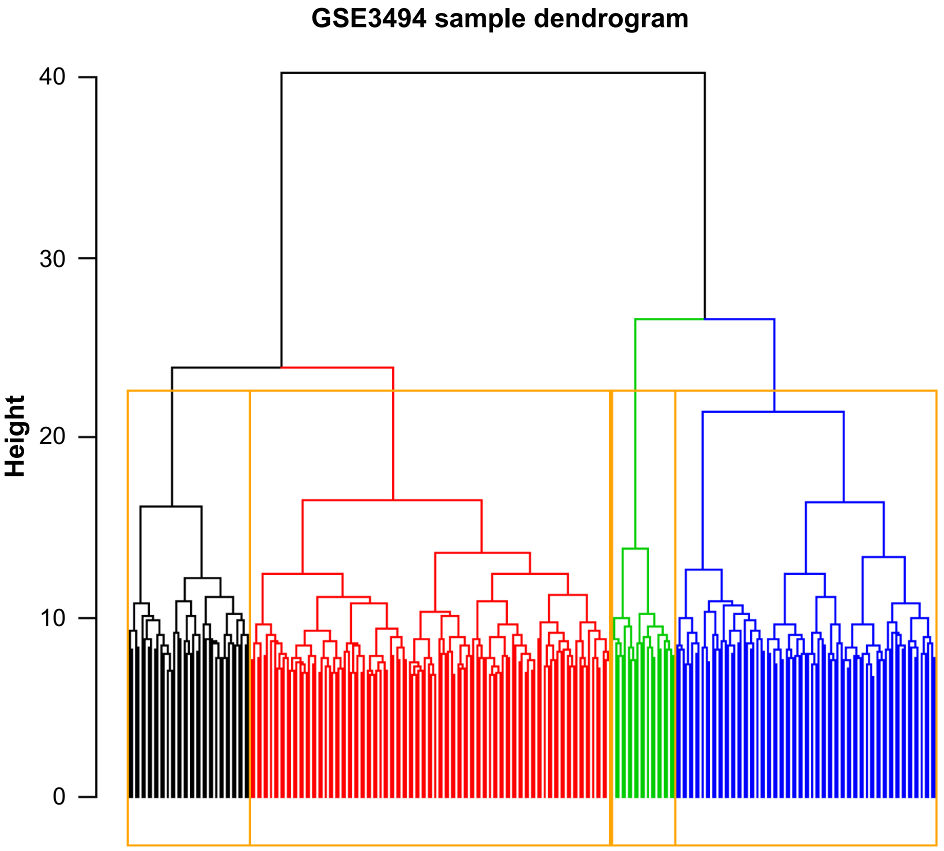

Figure 4. The dendrogram of tumor samples from the GSE3494[27] breast cancer dataset generated using hierarchical clustering. The top 1000 genes were selected via the SD_Rank method. Genes and samples were clustered using Euclidean distance and Ward.D2 linkage. Samples were split into four clusters.

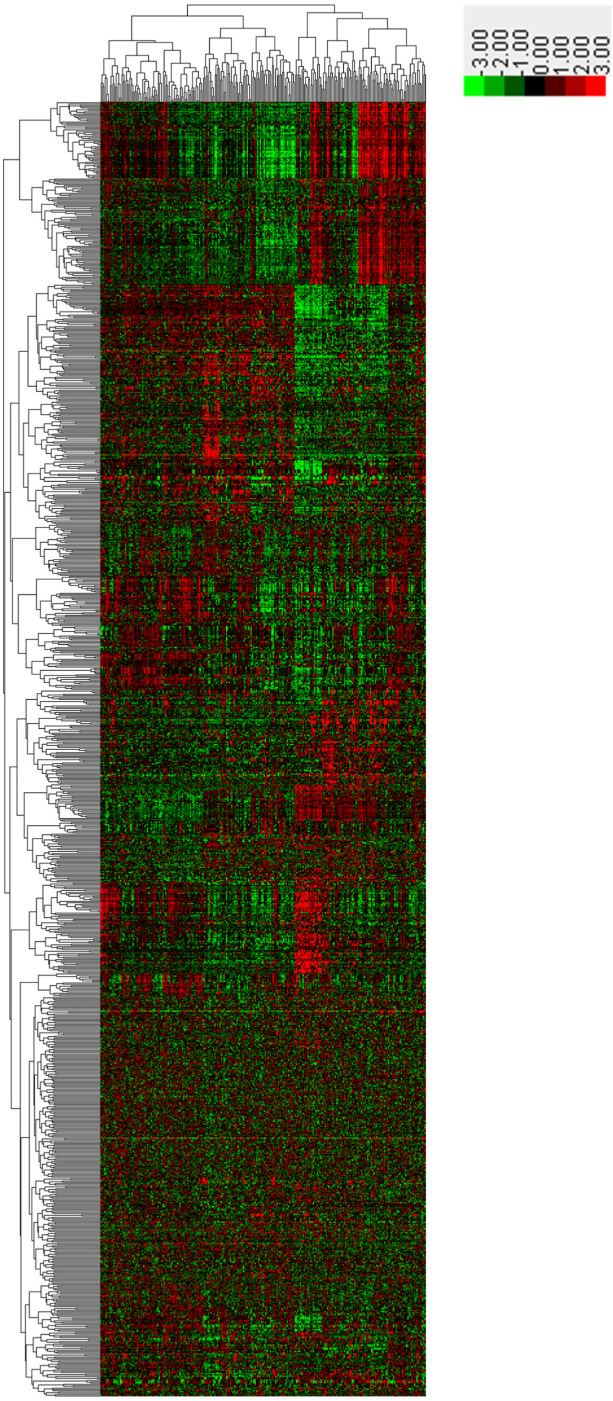

Figure 5. A Java TreeView heat map of clustering of breast cancer dataset, GSE3494.[27] A total of 1000 genes were selected using SD_Rank and were clustered using centered Pearson's correlation. Samples were clustered using Euclidean distance metrics with Ward.D2 linkage. In this heat map, rows are genes and columns are tumor samples.

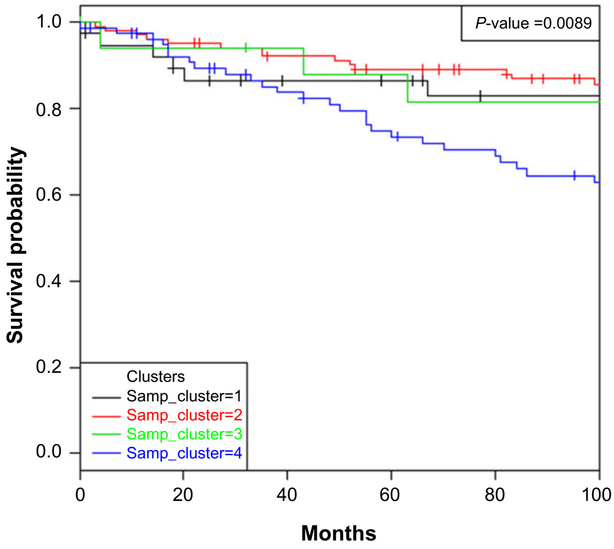

Figure 6. Kaplan–Meier plot of disease specific survival (DSS) of patients from the GSE3494 dataset.[27] Patients were clustered using the top 1000 genes on an SD_Rank ranking of relevance, and hierarchical clustering coupled with Euclidean distance and Ward.D2 linkage.

of these tests are reported in the Supplementary Comparative Analysis File. Information regarding the dataset names, gene probe numbers, patient numbers, cancer type, and more can be viewed in Table 4.

We downloaded publicly available solid tumor gene expression datasets with clinical data from NCBI GEO and The Cancer Genome Atlas (TCGA). These gene expression and clinical outcome files can be viewed in the Supplementary Gene Expression File and Supplementary Clinical Outcome File.

Gene expression datasets were downloaded from GEO using the "GEOquery" package[30,41] and then normalized using the robust multichip average (RMA) procedure. Specifically, the Bioconductor package "affy"[31] was used in R to perform the normalization. Datasets consisting of multiple gene expression datasets such as the GSE26712-14764 cohort were prepared using an alternative method. First, the raw expression data were merged into a collective dataset and then quantile normalized using the Bioconductor package "preprocessCore"[32] and then log2 scaled. Ultimately, both these types of normalization methods resulted in gene expression values being in a range of approximately 3–16. The TCGA Glioblastoma dataset was downloaded from the TCGA Data Portal and was composed of median expression data from three different expression centers: The Broad Institute, University of North Carolina, and the Lawrence Berkley National Laboratory.[42] These data were normalized prior to obtaining it from the TCGA database. Prior to clustering, all gene expression data were further normalized between a range of 0 and 1 via feature scaling.

For each dataset, a fixed number of clusters was determined in order to divide samples. This number was kept the same during each *k*-means and hierarchical clustering experiment performed on the dataset. The number of clusters for each dataset is shown in Table 1; the choice was as specified in their respective literature.

Hierarchical and *k*-means clustering were examined in this study using the same distance metric of Euclidean distance. Hierarchical clustering using Euclidean distance and average linkage was the least effective at identifying clinically relevant subgroups in the datasets we examined

Table 4. Summary of gene expression datasets used in this study.

DATASET	CANCER TYPE	PATIENT NUMBER	GENE PROBE NUMBER	CLUSTER NUMBER	CLINICAL OUTCOME	REFERENCE(S)
GSE2034	Breast	286	22,283	6	RFS	Wang et al 2005[62]
GSE3494	Breast	251	22,283	4	DSS	Miller et al 2005[27]
GSE17705	Breast	298	22,283	3	RFS	Symmans et al 2010[63]
GSE25066	Breast	508	22,283	3	DFS	Hatzis et al 2011[64]
GSE9899	Ovarian	285	54,669	6	PFS, OS	Tothill et al 2008[46]
GSE26712-14764	Ovarian	265	22,283	5	OS	Bonome et al 2008,[65] Denkert et al 2009[66]
GSE14333	Colon	290	54,675	3	DFS	Jorissen et al 2009[35]
GSE17536	Colon	177	54,675	3	OS, DFS, DSS	Smith et al 2010[67]
GSE17538	Colon	238	54,675	3	OS, DFS, DSS	Smith et al 2010[67]
GSE39582	Colon	585	54,675	6	OS, RFS	Marisa et al 2013[4]
GSE30219	Lung	307	54,675	3	DFS	Rousseaux et al 2013[68]
GSE50081	Lung	181	54,675	3	OS, DFS	Der et al 2014[69]
GSE68465	Lung	443	22,283	3	OS	Director's Challenge Consortium… et al 2008[70]
TCGA-GBM	Glioblastoma	446	17,814	5	OS	The Cancer Genome Atlas et al 2013[71]

Abbreviations: RFS, relapse-free survival; DSS, disease-specific survival; DFS, disease-free survival; PFS, progression-free survival; OS, overall survival.

for all combinations of gene ranking and selection methods. For this choice of clustering and linkage method, the most notable results were the top 1000 genes in the CV_Guided ranking and the top 1% genes in Poly ranking of genes. Both gene selection methods identified clinically relevant clusters in 5 out of the 14 datasets (Fig. 7B). In contrast, hierarchical clustering using Euclidean distance in conjunction with Ward.D2 linkage proved to be more effective at identifying clinically relevant clusters. All combinations of the gene ranking and gene selection methods used identified clinically relevant clusters in at least 6 out of the 14 datasets to have a significant correlation between sample clustering and patient clinical outcome. Specifically, the choice of top 1000 genes in the "SD_Rank" method identified clinically relevant clusters in 10 datasets (Fig. 7A). Lastly, all gene ranking and selection methods followed by *k*-means clustering using Euclidean distance identified clinically relevant clusters in the greatest number of datasets, though it varied substantially based on the choice of gene ranking and selection methods. Each combination of gene ranking and selection method with *k*-means clustering could identify clinically relevant clusters in at least 9 datasets (Fig. 7C). The most efficient combination of gene selection and ranking methods for *k*-means clustering was selecting the "Adaptive-GMM" determined number of genes with the SD_Rank method, which identified clinically relevant clusters in 13 of the 14 datasets we tested.

As mentioned earlier, some gene expression datasets provided multiple clinical outcome measures (RFS, DFS, etc.) and survival tests were conducted for these respective datasets. A panel of graphs comparing the number of survival tests with a significant association with sample clustering and clinical outcome across the different gene ranking, gene selection, and clustering methods can be seen in Figure 8. Similarly, the method of choosing top 1000 genes in CV_Guided ranking and the method of choosing top 1% genes in Poly ranking were the most effective for hierarchical clustering using Euclidean distance and average linkage. These combinations of gene ranking and selection yielded 6 survival tests out of 21 that had significant correlation (Fig. 8B). All other gene ranking and selection methods for this clustering algorithm produced less than six survival tests with statistical significance. For the hierarchical clustering, using Euclidean distance and Ward.D2 linkage, the combination of gene selection and ranking that produced the highest number of significant results was the top 1000 genes in CV_Guided ranking. This combination of methods yielded 14 survival tests, whereas all other groups of methods produced less than 14 survival tests with significant relationships between sample clustering and clinical outcome (Fig. 8A). Overall, hierarchical clustering with Ward.D2 linkage produced more survival tests with significant associations when compared to the average linkage clustering. Lastly, the *k*-means clustering algorithm yielded 19 significant survival tests using the SD_Rank and GMM gene estimation method (Fig. 8C). Similar to the previous findings in Figure 7 above, the *k*-means clustering algorithm proves to be the most effective at identifying significant associations between the tumor sample clusters and clinical outcome than that of hierarchical clustering algorithms.

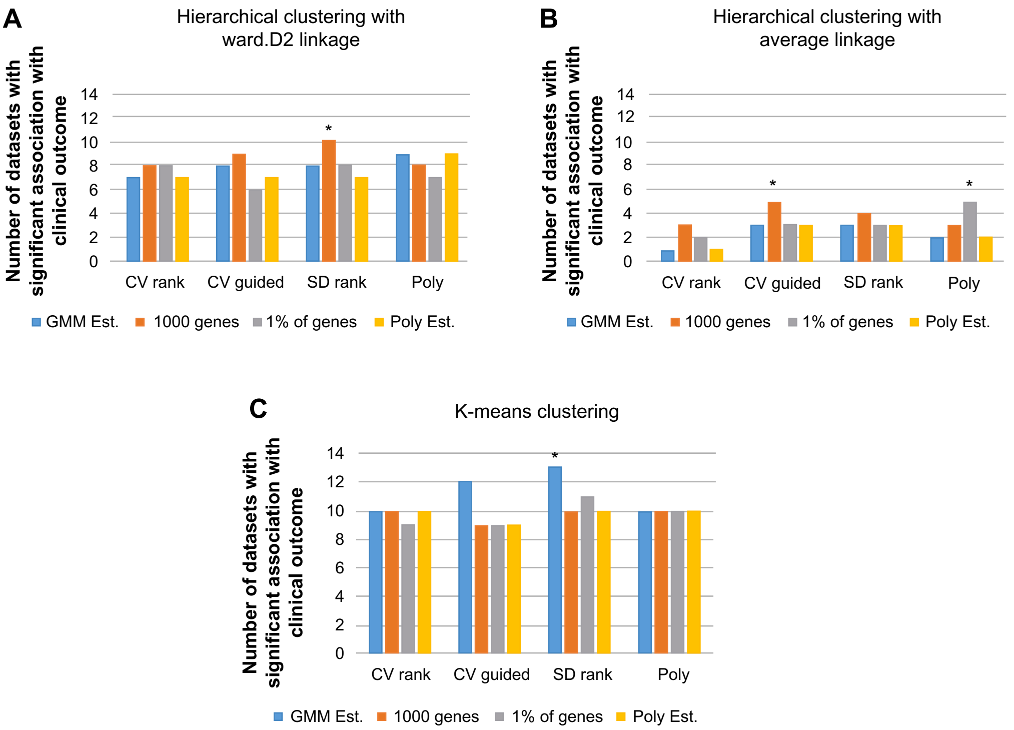

Figure 7. Histograms of performance on clinical relevance of clusters identified for different combinations of gene ranking and gene selection.
Notes: Hierarchical clustering using (**A**) Ward.D2 linkage, (**B**) average linkage. (**C**) *k*-means clustering. All methods used a Euclidean distance metric. Each bar shows the number of datasets that demonstrated a significant association with tumor sample clustering and respective patient clinical outcome. Bars marked with "*" are the gene ranking and selection methods that identified clinical outcome relevant clusters in the highest number of datasets for each type of clustering.

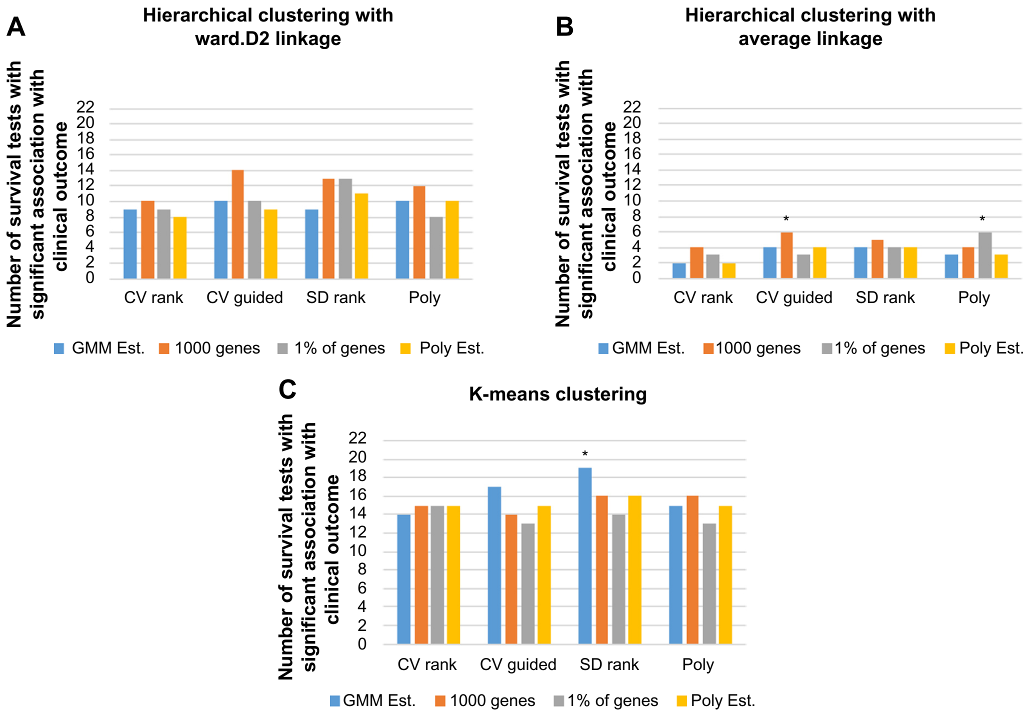

Figure 8. Histograms of performance on clinical relevance of clusters identified for different combinations of gene ranking and gene selection.
Notes: Hierarchical clustering using (**A**) Ward.D2 linkage, (**B**) average linkage. (**C**) *k*-means clustering. All methods used a Euclidean distance metric. Each bar shows the number of survival tests that demonstrated a significant association with tumor sample clustering and respective patient clinical outcome. Bars marked with "*" are the gene ranking and selection methods that identified clinical outcome relevant clusters in the highest number of survival tests for each type of clustering.

Discussion

Tuning the overall methodology for clustering including gene ranking, gene selection, and clustering method with appropriate choice of parameters is essential to elicit relevant sample groups using gene expression data. There are several studies that have evaluated the efficacy of various gene expression clustering methods.[7,9,17] In addition, there are several Bioconductor packages that have been made available for studying gene expression clustering. Examples include "iBBiG"[43] and "rqubic"[44] for biclustering of gene expression data. Similarly, the "EBarrays"[45] software package offers tools for analyzing microarray data such as the expectation maximization algorithm for gene expression mixture models. However, there are currently no known software packages that permit the tuning of overall gene selection and clustering methodologies to examine the efficacy of these steps.

To meet such a need, we presented an R-package called *multiClust* that accommodates a selection of methods at each step of overall methodology from a wide variety of options used in the literature. As clinical outcome is one of the most important criteria in consideration in biomedical research, the package allows evaluation of clusters based on their clinical relevance.

As a use case, we used *multiClust* to evaluate various combinations of options in the clustering methodology to identify the best performing clustering methodology in context to clinically relevant clusters in large cancer studies. In contrast to using generic quality metrics, clinical outcome was used as an objective function to help reveal what matters to the biomedical research community. After testing the various gene ranking and selection methods offered in *multiClust*, we determined that the SD_Rank gene ranking method with the GMM gene selection method was the most effective combination when using a *k*-means clustering framework. These methods yielded 13 out of 14 datasets and 19 out of 21 survival tests with significant association between sample clustering and clinical outcome (Fig. 7C and 8C). The SD_Rank gene ranking method orders genes by their standard deviation values. From a hypothetical stand point, variation among gene expression could correlate to variation among clinical outcome. It is possible that varying gene expression of biologically important genes can directly result in better or worse patient clinical outcomes.[1,35,46]

The SD_Rank method has not been highlighted in the literature, as it is sensitive to the choice of number of features. However, this method was identified to perform the best when used in conjunction with our GMM-based TRank method of identifying the number of genes. The GMM gene selection method fits Gaussian mixture models to the data using an Expectation-Maximization (EM) algorithm[28] in combination with a *t*-test like scoring function, which factors in variance and separation between Gaussians when discretizing gene expression. GMM is often more flexible at grouping samples into clusters of varying sizes, rather than tending to produce clusters of equal size. This method determines an estimated number of variable features that can be used to effectively assign samples into clearly distinct groups, which is key to identifying clinically relevant subgroups of cancer.

Furthermore, the SD_Rank method proved to be the most effective when using a *k*-means clustering framework. The *k*-means algorithm uses Euclidean distance and variance to cluster samples.[47] This clustering algorithm may be the most effective because it groups samples in the same way in which genes were ranked and ordered. Several studies have demonstrated that *k*-means clustering is more effective at analyzing the similarity of observations in a cluster relative to all other observations within the cluster than that of hierarchical clustering.[48,49] Furthermore, in many instances of the literature, *k*-means clustering has been shown to outperform hierarchical clustering in the context of gene expression profiling.[48–50] While hierarchical clustering remains to be a popular method for visualization of data, methods such as *k*-means clustering tend to be more meaningful at identifying biologically relevant patterns and clusters in gene expression data.[50]

Hierarchical clustering methods often have problems with incorrectly merging neighboring clusters.[48] Observations can be mistakenly assigned to the same cluster at early stages of the clustering process when they rightfully belong to different clusters. In hierarchical clustering, the assignment of the observation to the cluster cannot be changed. However, in *k*-means clustering, observations can be assigned to new clusters after the centroids have been modified. Similar to hierarchical clustering, the *k*-means algorithm has its own limitations. This algorithm is often limited by the starting positions of the cluster centers. A common solution to this problem is to use multiple restarts and place the clusters at random initial positions as an attempt to obtain a near optimal minimum.[51] Such an option that has been included in our package, however, was not examined because determining the optimality of starting points was beyond the scope of this study.

Furthermore, our results show that hierarchical clustering using the Ward.D2 linkage criteria was more effective than that of the average linkage method, which is commonly used in gene expression studies. Average linkage takes the mean distance between all pairs of data points of two clusters and therefore does not give special weight to sample outliers. Therefore, if these outliers are merged with larger clusters, the interesting local cluster structure can be lost.[16] It is possible that hierarchical clustering with average linkage criteria is not effective at assigning samples with variable features into the appropriate clusters. In other words, samples with completely different gene expression profiles may be mistakenly grouped into a sample cluster, therefore preventing us from identifying cancer patient groups with distinct clinical outcome. Ward.D2 linkage implements the Ward's minimum variance method,[18] which uses an error sum of squares calculation to cluster samples. This linkage criteria measure the sum of squared differences between each data observation and the group's mean.[18] Such a method groups observations while minimizing the sum of squares, ensuring that similar observations end up in the same clusters. This linkage method is more effective at identifying sample groups with similar expression profiles because it groups samples based on the variation of their expression profiles rather than an average of their profiles. Even though the methodology we suggested above yielded the most datasets and survival tests, there are cases for which the alternative methodologies also worked. Hence, we also need to study other types of diseases and gene expression datasets to identify the respective best-performing methodologies. Our *multiClust* package serves such a purpose in addition to helping to identify dataset-specific methodology.

Another future objective will be to incorporate more complex model-based clustering algorithms into our *multiClust* package. *k*-means and hierarchical clustering were chosen as the primary methodologies to examine in our package and this study because they are regarded as some of the most popular and well-understood clustering algorithms.[48–51] Furthermore, these two algorithms were chosen for comparative analysis because of their low computational intensiveness in comparison to that of model-based algorithms.[49,50] This allowed for rapid testing of multiple gene ranking and selection options when using our 14 gene expression datasets. However, model-based clustering via parsimonious Gaussian mixture models and linear mixed-effects models are also beneficial options for clustering of tissue samples and gene profiles even when dataset size is small.[28,52] Such methods have been demonstrated to perform well with high-dimensional gene expression data and produce better clustering when compared to conventional clustering methods.[28] In addition, extensions of the density-based spatial clustering of applications with noise algorithm have been proposed to obtain biologically useful patterns from large gene expression datasets.[53–55]

In conclusion, our R package *multiClust* offers numerous commonly used and new options for each step in the clustering methodology to test a wide variety of gene ranking, gene selection, and clustering algorithms for the identification of clinically relevant cancer subgroups in large gene expression datasets. Our analysis using this package shows that no single gene ranking or gene selection method is consistently best at identifying clinically relevant subgroups of cancer across all the clustering frameworks tested. However, our results suggest that the SD_Rank gene ranking method coupled with the GMM gene selection method was most effective at identifying clinically relevant subgroups of cancer in a *k*-means clustering context. Nonetheless, *multiClust* provides multiple options of gene selection and ranking for users to study clustering in the context of patient clinical outcome.

Acknowledgments

The authors appreciate the assistance of the Jackson Laboratory Internship Program Directors Walter Nakonechny and Melanie Sinche, as well as the guidance of Dr. Joerg Graf and Elaine Mirkin from the University of Connecticut. Additionally, the authors thank the staff of the Computational Sciences Department of The Jackson Laboratory for their feedback and support.

The release version of our multiClust package is now available on Bioconductor: (https://www.bioconductor.org/packages/3.3/bioc/html/multiClust.html).

Author Contributions

Authored the *multiClust* R-package and conducted experiments: NL, AF. Contributed to functions related to GMM, TRank in the package, experiments and contributed to the development of methods: PG. Obtained and processed the necessary data: NL, AF, PG. Conceived the project: RK. Led the project: RKMK, JG. All the authors contributed to the preparation of the article.

Supplementary Material

Supplementary Gene Expression File. This folder contains the gene expression datasets used in our study. It can be downloaded here (approx. 750MB): http://la-press.com/cr_data/multi_clust_supplementary_gene_expression_file.zip

Supplementary Clinical Outcome File. This folder contains the patient clinical outcome information used in our study.

Supplementary Package Vignette File. This PDF file contains information about the *multiClust* package and how to use the software.

Supplementary Comparative Analysis File. This folder contains spreadsheets showing the number of survival tests and datasets determined to have a significant association with clinical outcome for each gene ranking, gene selection, and clustering method used in our study.

REFERENCES

1. van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415:530–6.
2. Rapin N, Bagger FO, Jendholm J, et al. Comparing cancer vs normal gene expression profiles identifies new disease entities and common transcriptional programs in AML patients. *Blood*. 2014;123:894–904.
3. Nevins JR, Potti A. Mining gene expression profiles: expression signatures as cancer phenotypes. *Nat Rev Genet*. 2007;8:601–9.
4. Marisa L, de Reyniès A, Duval A, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med*. 2013;10:e1001453.
5. Oh SC, Park YY, Park ES, et al. Prognostic gene expression signature associated with two molecularly distinct subtypes of colorectal cancer. *Gut*. 2012;61:1291–8.
6. Calon A, Lonardo E, Berenguer-Llergo A, et al. Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nat Genet*. 2015;47:320–9.
7. D'haeseleer P. How does gene expression clustering work? *Nat Biotechnol*. 2005;23:1499–501.
8. Chandrasekhar T, Thangavel K, Elayaraja E. *Effective Clustering Algorithms for Gene Expression Data*. 2012. *ArXiv12014914 Cs Q-Bio*. Available at: http://arxiv.org/abs/1201.4914.
9. Wiwie C, Baumbach J, Röttger R. Comparing the performance of biomedical clustering methods. *Nat Methods*. 2015;12(11):1033–8.
10. Jain AK, Dubes RC. *Algorithms for Clustering Data*. Upper Saddle River, NJ: Prentice-Hall, Inc; 1988.
11. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach Learn*. 2003;52:91–118.
12. Oghabian A, Kilpinen S, Hautaniemi S, Czeizler E. Biclustering methods: biological relevance and application in gene expression analysis. *PLoS One*. 2014;9:e90801.

13. Jain AK, Murty MN, Flynn PJ. Data clustering: a review. *ACM Comput Surv*. 1999;31:264–323.
14. Wong KC. *A Short Survey on Data Clustering Algorithms*. 2015. *ArXiv151109123 Cs Stat*. Available at: http://arxiv.org/abs/1511.09123.
15. Berkhin P. In: Kogan J, Nicholas C, Teboulle M, eds. *Grouping Multidimensional Data*. Berlin, Heidelberg: Springer; 2006:25–71. Available at: http://link.springer.com/chapter/10.1007/3-540-28349-8_2.
16. Andreopoulos B, An A, Wang X, Schroeder M. A roadmap of clustering algorithms: finding a match for a biomedical application. *Brief Bioinform*. 2009;10:297–314.
17. Xu R, Wunsch DC. Clustering algorithms in biomedical research: a review. *IEEE Rev Biomed Eng*. 2010;3:120–54.
18. Murtagh F, Legendre P. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *J Classif*. 2014;31:274–95.
19. Pal SK, Bandyopadhyay S, Biswas S. Pattern Recognition and Machine Intelligence: First International Conference, PReMI 2005, Kolkata, India, December 20–22, 2005, Proceedings. Berlin: Springer Science & Business Media; 2005.
20. Chesnevar CI, Rivaherrera EO, de la Ossowski S, Vouros G. Agreement Technologies: Second International Conference, AT 2013, Beijing, China, August 1–2, 2013. Proceedings. Berlin, Heidelberg: Springer; 2013.
21. Oakes MP, Ji M. *Quantitative Methods in Corpus-based Translation Studies: A Practical Guide to Descriptive Translation Research*. Amsterdam: John Benjamins Publishing; 2012.
22. Moftah HM, Azar AT, Al-Shammari ET, Ghali NI, Hassanien AE, Shoman M. Adaptive k-means clustering algorithm for MR breast image segmentation. *Neural Comput Appl*. 2013;24:1917–28.
23. Zheng B, Yoon SW, Lam SS. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Syst Appl*. 2014;41:1476–82.
24. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. Vol 103. New York: Springer; 2013.
25. Lee IH, Lushington GH, Visvanathan M. A filter-based feature selection approach for identifying potential biomarkers for lung cancer. *J Clin Bioinforma*. 2011;1:11.
26. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res*. 2003;3:1157–82.
27. Miller LD, Smeds J, George J, et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci U S A*. 2005;102:13550–5.
28. McNicholas PD, Murphy TB. Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics*. 2010;26:2705–12.
29. Kharchenko P. *Identifying Highly Variable Genes*. 2014. Available at: http://pklab.med.harvard.edu/scw2014/subpop_tutorial.html.
30. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets – update. *Nucleic Acids Res*. 2013;41:D991–5.
31. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy – analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*. 2004;20:307–15.
32. Bolstad B. *A Collection of Pre-Processing Functions*. R package version 1.30.0; 2015.
33. Zwiener I, Frisch B, Binder H. Transforming RNA-Seq data to improve the performance of prognostic gene signatures. *PLoS One*. 2014;9:e85150.
34. Fraley C, Raftery AE, Murphy TB, Scrucca L. *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*. 2012. Available at: http://my.ilstu.edu/~mxu2/mat456/mcluster.pdf.
35. Jorissen RN, Gibbs P, Christie M, et al. Metastasis-associated gene expression changes predict poor outcomes in patients with dukes stage B and C colorectal cancer. *Clin Cancer Res*. 2009;15:7642–51.
36. Maechler M, Rousseeuw P, Struyf A, et al. *Package 'Cluster'*. 2015. Available at: https://cran.r-project.org/web/packages/cluster/cluster.pdf.
37. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. *J R Stat Soc Ser B Stat Methodol*. 2001;63:411–23.
38. Saldanha AJ. Java Treeview – extensible visualization of microarray data. *Bioinformatics*. 2004;20:3246–8.
39. Therneau T. *A Package for Survival Analysis in S*. 1999. Available at: http://www.mayo.edu/research/documents/tr53pdf/doc-10027379.
40. Therneau T. *Package 'Survival'*. 2015. Available at: https://cran.r-project.org/web/packages/survival/survival.pdf.
41. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*. 2007;23:1846–7.
42. The Cancer Genome Atlas Research Network. Corrigendum: comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*. 2013;494:506–6.
43. Gusenleitner D, Culhane A. *iBBiG: Iterative Binary Biclustering of Genesets*. R package version 1.14.0; 2011. Available at: http://bcb.dfci.harvard.edu/~aedin/publications/
44. Li G, Ma Q, Tang H, Paterson A, Xu Y. QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Res*. 2009; 37(15):e101.
45. Yuan M, Newton M, Sarkar D, Kendziorski C. *EBarrays: Unified Approach for Simultaneous Gene Clustering and Differential Expression Identification*. R Package Version 2340; 2007.
46. Tothill RW, Tinker AV, George J, et al. Australian Ovarian Cancer Study Group. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin Cancer Res*. 2008;14:5198–208.
47. Kanungo T, Mount D, Netanyahu NS, Piatko C, Silverman R, Wu A. An efficient k-means clustering algorithm: analysis and implementation. *IEEE Trans Pattern Anal Mach Intell*. 2002;24:881–92.
48. Steinbach M, Karypis G, Kumar V. A comparison of document clustering techniques. *KDD Workshop on Text Mining*. Minneapolis, Minnesota; 2000.
49. Kaur M, Kaur U. Comparison between K-mean and hierarchical algorithm using query redirection. *Int J Adv Res Comput Sci Softw Eng*. 2013;3:1454–9.
50. Thalamuthu A, Mukhopadhyay I, Zheng X, Tseng GC. Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*. 2006;22:2405–12.
51. Tzortzis GF, Likas CL. The global kernel -means algorithm for clustering in feature space. *IEEE Trans Neural Netw*. 2009;20:1181–94.
52. McLachlan GJ, Flack LK, Ng SK, Wang K. Clustering of gene expression data via normal mixture models. *Methods Mol Biol*. 2013;972:103–19.
53. Jiang D, Pei J, Zhang A. DHC: a density-based hierarchical clustering method for time series gene expression data. *Third IEEE Symposium on Bioinformatics and Bioengineering, 2003. Proceedings*. Buffalo, NY: IEEE; 2003:393–400.
54. Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: a survey. *IEEE Trans Knowl Data Eng*. 2004;16:1370–86.
55. Edla DR, Jana PK. A prototype-based modified DBSCAN for gene clustering. *Procedia Technol*. 2012;6:485–92.
56. Li W, Fan M, Xiong M. SamCluster: an integrated scheme for automatic discovery of sample classes using gene expression profile. *Bioinformatics*. 2003;19:811–7.
57. Hall JS, Iype R, Senra J, et al. Investigation of radiosensitivity gene signatures in cancer cell lines. *PLoS One*. 2014;9(1):e86329.
58. Gibbons FD, Roth FP. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res*. 2002;12:1574–81.
59. Eisen M. *Cluster 3.0 Manual*. 2002. Available at: http://bonsai.hgc.jp/~mdehoon/software/cluster/cluster3.pdf.
60. Simon R, Lam A. *BRB-ArrayTools Version 3.0 User's Manual*. 2002. Available at: http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CB0QFjAAahUKEwiN1-nSv8zIAhUEPT4KHc2VAr8&url=http%3A%2F%2Flinus.nci.nih.gov%2Fpilot%2FManual.doc&usg=AFQjCNGV1GfYzEgK2H4yh0IPi2PdPzuNHw&sig2=MLr4Djq7 WSppY8wb2nIaeg.
61. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc*. 2002;97:611–31.
62. Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*. 2005;365:671–9.
63. Symmans WF, Hatzis C, Sotiriou C, et al. Genomic index of sensitivity to endocrine therapy for breast cancer. *J Clin Oncol*. 2010;28:4111–9.
64. Hatzis C, Pusztai L, Valero V, et al. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA*. 2011;305:1873–81.
65. Bonome T, Levine DA, Shih J, et al. A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. *Cancer Res*. 2008;68:5478–86.
66. Denkert C, Budczies J, Darb-Esfahani S, et al. A prognostic gene expression index in ovarian cancer – validation across different independent data sets. *J Pathol*. 2009;218:273–80.
67. Smith JJ, Deane NG, Wu F, et al. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology*. 2010;138:958–68.
68. Rousseaux S, Debernardi A, Jacquiau B, et al. Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. *Sci Transl Med*. 2013;5:186ra66.
69. Der SD, Sykes J, Pintilie M, et al. Validation of a histology-independent prognostic gene signature for early-stage, non-small-cell lung cancer including stage IA patients. *J Thorac Oncol*. 2014;9:59–64.
70. Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma, Shedden K, Taylor JM, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med*. 2008;14:822–7.
71. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*. 2013;45:1113–20.