Colon Cancer Subtypes from Gene Expression Data

Emmanuelle Dankwa Natalia Garcia Martin William Thomas

November 14, 2018

Abstract

We assess the robustness of the findings of De Sousa E Melo et al. for colon cancer subtype identification using microarray gene expression data. We reproduce the hierarchical clustering method employed in their study and discuss the use of other clustering strategies. We classify patients into the three previously defined groups using PAM as well as normal based classification and discuss the adequacy and robustness of these methods.

1 Introduction

1.1 Background

Bowel cancer, also known as colorectal cancer, occurs in the large bowel, which is made up of the colon and rectum. Colorectal cancer is the third most common malignant cancer for both males and females in the UK and the second biggest cancer killer. In 2016, 19,581 males and 15,371 females were diagnosed with colorectal cancer in England [1]. Bowel cancer is treatable and curable especially if diagnosed early. However survival drops significantly as the disease develops, suggesting the need for early diagnosis [2].

DNA (Deoxyribonucleic Acid) is referred to as the blueprint of life because it contains the instructions needed for an organism to grow, develop, survive and reproduce. One of the major functions of the DNA is to construct proteins which are responsible for carrying out most cell functions. The process by which the instructions in our DNA are converted into a functional product is known as gene expression and consists of two major steps: a transcription stage in which DNA is copied into mRNA (messenger RNA) and a translation stage where mRNA is decoded into amino acid sequences to perform cell functions [3]. The difference in DNA sequences between individuals within a population is known as genetic variation and leads to differences in the individual's phenotype, which consists of all the observable physical properties of an organism that result from the interaction of its genetic make-up with the environment, including morphology, biochemical and physiological properties, development, and behavior [4]. The most common type of genetic variation in humans are single-nucleotide polymorphisms (SNPs), which occur once in every 300 nucleotides on average, which means there are around 10 million SNPs in the human genome [5].

While most of this variability does not have any impact on health, some polymorphisms can act as biological markers for diseases such as cancer. DNA Microarrays are an important technique to identify genetic variation and can be used to classify and predict diagnostic categories based on gene expression profiles. Conventional diagnosis of cancer has been based on subjective morphological examination. On the other hand, microarray analysis aims to classify cancers objectively and accurately and to provide clinicians with information to help them allocate the most appropriate treatments to specific patient groups [6]. During this paper, we will present, test and compare several clustering and classification techniques for colon cancer subtype identification. De Sousa E Melo et al. demonstrate using an unsupervised classification strategy that three main subtypes can be recognised and state that previous to their study, two subtypes have been identified and well characterised. These are chromosomal-instable cancer and microsatellite-instable cancer. The third subtype is largely microsatellite stable and its identification is crucial as it has a very unfavorable prognosis [7].

1.2 Data

Gene expression data is highly dimensional as it contains measurements over thousands of genes and few biological samples [8]. The AMC-AJCCII-90 series (GSE33113) dataset is publicly available online and has been obtained using the Affymetrix Human Genome U133 Plus 2.0 platform. It contains information for 90 patients with stage II colon cancer who underwent curative surgery at the Academic Medical Centre (Amsterdam, Netherlands), as well as six normal colon samples. It includes gene expression data for 54,675 genes for each of these patients. Additionally, the work of De Sousa E Melo et al. made use of seven colon cancer patient datasets for classifier validation, comprising a total of 1164 unique patients.

2 Methodology

Cluster analysis is an unsupervised learning method which involves partitioning a set of objects into meaningful groups, without the use of any labelled training data. As we mentioned earlier, this is crucial in order to discover new cancer types and identify gene groups responding similarly to specific experimental conditions [8]. Cluster analysis from gene expression data involves three steps: i) pre-processing, ii) clustering, and iii) validation [9]. The first of these involves a number of data transformations including data normalization. The second step consists of selecting and applying one or several clustering methods. The resulting clusters are then validated using cluster-validation techniques, which may include evaluating the clustering structure by varying parameters in the algorithm or assessing the stability of the clusters to small variations in the data.

2.1 Data preprocessing

Before using the AMC-AJCCII-90 dataset to construct clusters, the raw data must first be processed and the dimensionality of the data reduced. This consists of several steps, which will eventually help us to identify a subset of genes which are informative for clustering individuals into groups. Here we note that our pre-processing does not mirror that of De Sousa E Melo et al., mainly due to limitations on the data available to us, but we try to replicate their procedure as best we can and are able to obtain reasonably similar results.

The raw data, in the form of 90 .CEL files, contains the results of various intensity calculations, which are measured using probes. These probes measure the expression levels for each gene and enable us to study alterations in DNA. The raw data is first normalized and summarised using *frozen robust multiarray analysis* (fRMA). In practice, fRMA is particularly useful for pre-processing multiple arrays and can help account for any batch effects which may or may not be present. In the case where we have only one batch, fRMA does not differ significantly in performance from RMA.

Pre-processing step	Number of genes retained
Initial dataset	$54,675$
fRMA	$54,675$
Barcode algorithm	$18,982$
Median absolute deviation analysis	$7,747$

Table 1: The number of genes retained for clustering after each pre-processing step.

The output of fRMA is a set of measures of gene expression for each subject, which we can then use to detect whether genes are *expressed* or *unexpressed*. Using the barcode algorithm, we can process this gene expression data and obtain a binary output telling us whether a particular gene is expressed or unexpressed. This makes use of publicly available gene expression data, which we can compare our data to. Where genes were not expressed in at least one individual, these were omitted. After this step, we have 18,982 genes remaining.

The final filtering process involved using the *mean absolute deviance* to assess which genes had the most variability, thereby identifying which genes are most informative for clustering. The mean absolute deviance for a vector x is taken to be:

$$mad(x) = k^{*}median\{|x - median(x)|\},$$

where k is a constant chosen for asymptotic normal consistency, typically $k = 1.4826$. We compute this for each gene, retaining only those genes whose median absolute deviation is greater than 0.5, before median centering the remaining gene expressions. This leaves us with 7,747 genes in total which we will use to form clusters.

2.2 Clustering Techniques

2.2.1 K-means Clustering

One clustering method we can consider is *k-means clustering.* This method attempts to partition a set of observations into k specified clusters, where each observation belongs to the cluster with the nearest mean. That is, given observations $x_1, ..., x_n$ and a partition $\mathbf{S} = \{S_1, ..., S_k\}$, we want to find

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2,$$

where $\boldsymbol{\mu_i}$ is the mean of the points in S_i.

Starting from an initialised set of k means and given a single observation, we compute the Euclidean distance between the observation and each of the means, assigning the observation to that cluster with the closest mean. This is repeated for each observation in the dataset. This gives us an initial clustering of the data. From here, we recompute the cluster means as the centroids of the observations in the new clusters, before repeating the allocation process. The procedure finishes when none of the observations are assigned to a different cluster.

While k-means clustering is efficient at clustering observations, it does suffer from several shortcomings. One of these is that k must be pre-specified. However, without prior knowledge of the number of clusters in the data, an inappropriate choice of k will lead to poor results. This is a common issue with many clustering methods, which emphasises the need to run diagnostic checks to select an appropriate number of clusters. We should also consider the possibility that the clustering algorithm converges to a local minimum, rather than a global minimum, which could again lead to misleading results.

2.2.2 K-medoids Clustering

Operating in a similar manner to k-means clustering, *k-medoids clustering* attempts to partition observations into k known clusters. However, we here use an datapoints to describe the centers of clusters, as well as an alternative metric for computing distances between cluster centers and datapoints. As before, we need to pre-specify the number of clusters k.

We define a *medoid* to be the point whose average dissimilarity to all other points in a cluster is minimal, that is, the central most point in a cluster. Rather than minimising a sum of squared Euclidean distances, we now want to minimise a sum of pairwise dissimilarities. That is, we want to find a partition such that

$$\sum_{i=1}^{n} \sum_{j=1}^{n} d(i,j) \mathbb{I}\{i \textit{ and } j \textit{ are in the same cluster}\},$$

is minimised, where $d(i,j)$ is a measure of dissimilarity between observations i and j. By taking this approach, we are able to form a clustering which is more robust to outliers.

We now assign observations to clusters according to the distance between observations and the medoids for each cluster. Where we now diverge from k-means clustering is in how we recalculate the medoids for each cluster. Within each cluster, we in turn make each datapoint the medoid and compute the overall cost (i.e. the sum of pairwise dissimilarities). If we observe a decrease in the cost, then we take that point to be the new medoid. Else, we retain the previous medoid. We then reassign observations based on the new medoids and repeat until the cost decreases no further.

2.2.3 Hierarchical Clustering

Hierarchical clustering builds a cluster hierarchy displayed as a tree diagram known as a dendrogram (see Figure 1). In agglomerative hierarchical clustering, each data point is initially placed into its own singleton group. At each step, the two closest (most similar) clusters are merged, reducing the number of clusters by one. This is repeated until we obtain a single group containing all objects. Divisive clustering works the other way around, starting with one cluster and recursively splitting into the most appropriate clusters. De Sousa E Melo et al. perform hierarchical clustering with agglomerative average linkage to cluster the 90 samples with the expressed 7846 genes. Denote the data by $\{x_{ij}\},\ i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, p$, with n independent observations (patients) and p features (expressed genes), and the distance between observations i and i' by $d_{ii'}$. The squared Euclidean distance is usually chosen $\sum_j (x_{ij} - x_{i'j})^2$. Suppose we have k clusters $C_1 \dots C_k$. C_r includes observations in cluster r and define $n_r = \|C_r\|$. In average linkage hierarchical clustering, the distance between two clusters C_r, C_s is defined as the average distance between each point in one cluster and every point in the other cluster [10]. That is,

$$\text{dist}_{AL}(C_r, C_s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} d(x_{ri}, x_{sj}).$$

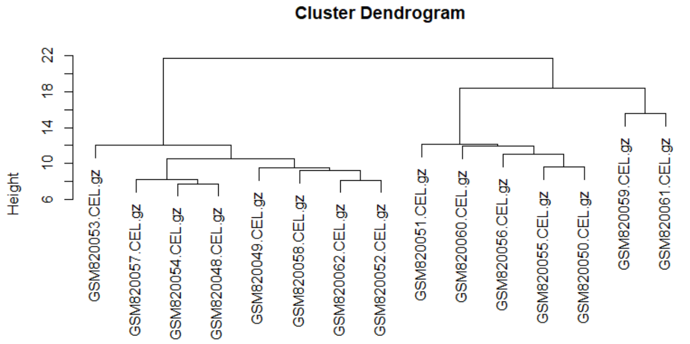

Figure 1: Example dendrogram using only 15 patients.

2.2.4 Model-based clustering

Model-based clustering (MBC) adopts a probabilistic approach to modelling and classification by viewing each data point as coming from a probability distribution. In particular, each cluster,k, is thought of as being modelled by some probability distribution f_k and the entire data is considered a mixture model of m densities, where m is the total number of clusters. We discuss and implement a particular type of the MBC later in our study (section 2.3.3).

2.2.5 Gap statistic for selecting the number of clusters

Popular methods for selecting the number of clusters include the elbow method, the average silhouette method and the gap statistic (Fig. 2B)). These approaches can be divided into global and local methods. Global methods evaluate a measure over the entire dataset and optimise it as a function of the number of clusters, while local methods test whether individual pairs of cluster should be amalgamated [11]. We will explain the gap statistic method, a global procedure, which was chosen by De Sousa E Melo et al. (2013) to support their choice of three clusters. Let $D_r = \sum_{i,i' \in C_r} d_{ii'}$ be the sum of pairwise distances for all observations in cluster r and let

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r,$$

where W_k is the pooled within-cluster sum of squares around the cluster means. The gap statistic approach standardises the graph of $\log(W_k)$ by comparing it with its expectation E_n^* under an appropriate null distribution:

$$\text{Gap}_n k = E_n^* \{\log(W_k)\} - \log(W_k).$$

The estimated optimal number of clusters is given by the value of k for which $\log(W_k)$ falls the farthest below the reference curve, i.e. the value maximising $\text{Gap}_n k$. It is worth mentioning that if smaller clusters are present within larger well-separated clusters, the function will exhibit non-monotone behavior. This means that there is value to examining the entire gap curve rather than the position of the global maximum, as local maxima may also be informative. The expectation in the formula can be estimated using Monte-Carlo methods and the R package we will use generates the reference features from a uniform distribution making use of PCA (principal component analysis) and SVD (single value decomposition) [11].

2.2.6 Consensus clustering

In gene expression data, the problem of a relatively small sample size is compounded by very high dimensionality, making the clusters sensitive to noise and susceptible to over-fitting [12]. Agglomerative hierarchical clustering (HC) is intuitive and visually appealing, but it does not provide an objective criterion to establish the number of clusters and the clusters' boundaries. Methods such as k-means provide defined clusters and boundaries but are not easy to visualise and requires the choice of the number of clusters. Model-based models are often based on asymptotic approximations whose accuracy decreases for small sample sizes as is the case with gene expression data. Monti et al. introduce a more robust method to overcome some of these difficulties by representing the consensus across multiple runs of a clustering algorithm of choice using bootstrapping. It involves recording a number of pairwise consensus values, which is the number of times a pair of items appeared within the same cluster as a proportion of times they appeared in the same bootstrap sample. This model-independent resampling-based method allows us to assess stability and sensitivity to initial conditions and to visualise the cluster boundaries as well as the appropriate number of clusters. The study of De Sousa E Melo et al. employed the consensus procedure with hierarchical agglomerative clustering.

In order to quantify the agreement among the clustering runs over the perturbed datasets, we define a symmetrical $N \times N$ consensus matrix $\mathcal{M}$, where N is the number of items we want to cluster. Let $D^{(1)}, D^{(2)}, \ldots, D^{(H)}$ be the list of perturbed datasets obtained by resampling the original data D, where H is the number of resampling iterations. Define $M^{(h)}$ to be the $N \times N$ connectivity matrix corresponding to $D^{(h)}$ such that $M^{(h)}(i, j) = 1$ if entries i and j are in the same cluster and 0 otherwise and let $I^{(h)}$ be the $N \times N$ indicator matrix such that the $I^{(h)}(i, j) = 1$ if both items i and j are present in the dataset $D^{(h)}$ and 0 otherwise. Then, the consensus matrix is given by

$$\mathcal{M}(i, j) = \frac{\sum_h M^{(h)}(i, j)}{\sum_h I^{(h)}(i, j)}.$$

$\mathcal{M}$ has entries between 0 and 1. We denote the case where all the entries are either 0 or 1 as perfect consensus. If the matrix entries were arranged such that adjacent entries correspond to items belonging to the same cluster, perfect consensus would give us a block-diagonal matrix with non-overlapping blocks of $1's$ along the diagonal (each block corresponding to a cluster) surrounded by $0's$[12]. We can associate a colour gradient to the range of real numbers between 0 and 1 so that white correspond to zero and a dark colour corresponds to 1, so that higher consensus will lead to better defined blocks. Consensus clustering can then be nicely visualised as a heat map.

2.2.7 The Adjusted Rand Index (ARI)

While we can use consensus clustering to obtain a more robust clustering, we might like to be sure that an alternative clustering method will not cluster the data in a completely different way. Given two different clusterings, it is very difficult to determine which clustering is more reflective of the true grouping structure. However, we are able to compare these clusterings directly and assess the proportion of observations that are assigned to the same cluster in both procedures. That is, we can compute a measure of agreement between two clusterings. One such measure is the *Rand index.*

Given n observations $x_1, ..., x_n$ and two different clusterings of the observations, we define the Rand index to be

$$R = \frac{a + b}{\binom{n}{2}},$$

where a denotes the number of pairs of observations which are clustered together by both methods and b denotes the number of pairs of observations which are not clustered together by both methods. That is, $a + b$ represents the number of agreements between the two clustering algorithms. The Rand index takes on values between 0 and 1, with 0 corresponding to total disagreement and 1 corresponding to the clusterings being identical.

Alternatively, we can correct the Rand index for any agreements between two clustering methods that occur by chance, which accounts for the fact that the expected value of the Rand index is not constant. This correction yields the *adjusted Rand index.* To demonstrate how we can compute the adjusted Rand index, we need to first summarise the similarity between two clusterings in a contingency table. Let $\mathbf{C} = \{C_1, ..., C_r\}$ and $\mathbf{K} = \{K_1, ..., K_s\}$ denote two clusterings of $x_1, ..., x_n$, with n_{ij} being the number of points in common between clusters C_i and K_j. We can then generate the contingency table shown in Table 2.

K **C**	K_1	K_2	...	K_s	Sums
C_1	n_{11}	n_{12}	...	n_{1s}	a_1
C_2	n_{21}	n_{22}	...	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C_r	n_{r1}	n_{r2}	...	n_{rs}	a_r
Sums	b_1	b_2	...	b_s	

Table 2: Contingency table summarising the relationship between two clusterings of n observations.

The adjusted Rand index can then be defined as

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}},$$

which enables us to assess the level of agreement between two clustering methods.

2.3 Patient classification

2.3.1 PAM: Prediction Analysis for Microarrays

In their study, De Sousa E Melo et al. (2013) used the prediction analysis for microarrays (PAM) [6] in developing a classifier to predict cancer sub-type for samples from other arrays. In predicting the cluster of a sample, PAM makes use of only a proportion of the total number of genes in the array : these genes are those that have stable expression in all samples within a particular cluster. This is achieved by the "nearest shrunken centroid" method described below:

Let x_{ij} represent the gene expression for genes $i = 1, 2, \ldots p$ from $j = 1, 2, \ldots n$ samples and let C_k be the indices of the n_k samples in cluster k. Suppose also that there exists K clusters for the data. Define the i^{th} component of the centroid for class k as: $\bar{x}_{ik} = \sum_{j \epsilon C_k} x_{ij}/n_k$ and the i^{th} component of the overall centroid as $\bar{x}_i = \sum_{j=1}^{n} x_{ij}/n$. The class centroids are shrunk toward the overall centroid by 'soft thresholding' and the use of a shrinkage parameter (threshold), Δ, which is determined by M-fold cross-validation and chosen to have minimal cross-validation and test errors. As the value of Δ increases, more genes are dropped out of the set of genes to be used for cluster prediction, leaving the most representative. To classify a test sample, PAM computes the squared distance from the gene expression profile of this sample to each of the shrunken cluster centroids. The cluster whose centroid has the minimum squared distances to the gene expression profile of the test sample is its predicted cluster.The main advantage of PAM lies in its ability to eliminate noisy genes thus producing a more accurate classifier.

2.3.2 Multivariate normal-based classification: Finite Mixture of Gaussians (FMG)

In our analysis, we propose the Finite Mixture of Gaussians (FMG) method for patient classification and compare that to the PAM method used by De Sousa E Melo et al. (2003). FMG adopts a model-based clustering approach as discussed in section 2.2.3. Let $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_i, ..., \mathbf{x}_n\}$ denote a sample of n independent and identically distributed observations. A probability function defines the distribution of each observation through a finite mixture model of G components:

$$f(\boldsymbol{x}_i; \boldsymbol{\Psi}) = \sum_{k=1}^{G} \pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k)$$

where $\boldsymbol{\Psi}$ denotes the set of parameters for the mixture model, π_k represents a mixing weight for the k^{th} component where $\sum_{k=1}^{G} \pi_k = 1, f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k)$ is the k^{th} component density for the $x_i^t h$ observation which has parameter vector θ_k and G is the number of mixture components. FGM assumes a multivariate Gaussian distribution for each component; $f_k(\boldsymbol{x}; \boldsymbol{\theta}_k) \sim N(\boldsymbol{\mu}_k), \sum_k$ where each cluster k has center $\boldsymbol{\mu}_k$ and covariance matrix $\sum_k$. A key consideration to be made in finite mixture modelling is the number of components to be included in the mixture and the kind of covariance parameterisation to use. We would make use of the Bayesian Information Criterion (BIC) which is based on penalising the log likelihood with the addition of more components.

3 Results

3.1 Clustering

Previous work suggests that we should have at least two clusters corresponding to chromosomal-instable and microsatellite-instable cancers, while De Sousa E Melo et al. provide biological support for the existence of a third cancer sub-type, largely microsatellite stable. The use of 10+ clusters may be preferable in terms of the gap statistic (Fig. 2B), but we have no biological evidence for such a large number of clusters. Indeed, given that there are only 90 subjects in our dataset, using 10 or more clusters may simply lead to overfitting of the data based on some confounding factors. We must also consider the difficulties that having 10 different subgroups would introduce at the treatment level, as tailoring treatments to 10 different groups of subjects may be neither practical nor cost effective. Taking this into account, the gap statistic would suggest that we should divide the subjects into only 3 clusters.

The choice of 3 clusters is further supported by the output obtained from consensus clustering (Fig. 2A and 2C). In Fig. 2A, the consensus values shown in the matrices range from 0 (patients never clustered together) to 1 (always clustered together), marked by white to dark blue. The consensus matrices are ordered by the consensus clustering which is depicted as a dendrogram at the top of the heatmap. An appropriate value of K would correspond to clearly defined blocks with values close to “perfect”, that is, shown as white or dark blue, while lighter blue regions correspond to more ambiguous values. Hence, $K = 2$ and $K = 3$ seem appropriate. In Fig. 2C, the tracking plot indicates the stability of the clusters. Patients that change cluster often, represented by changing colours within a column, indicate unstable membership and clusters with an abundance of unstable members suggest an unstable cluster [13]. This plot suggests that $K = 3$ is a reasonable choice.

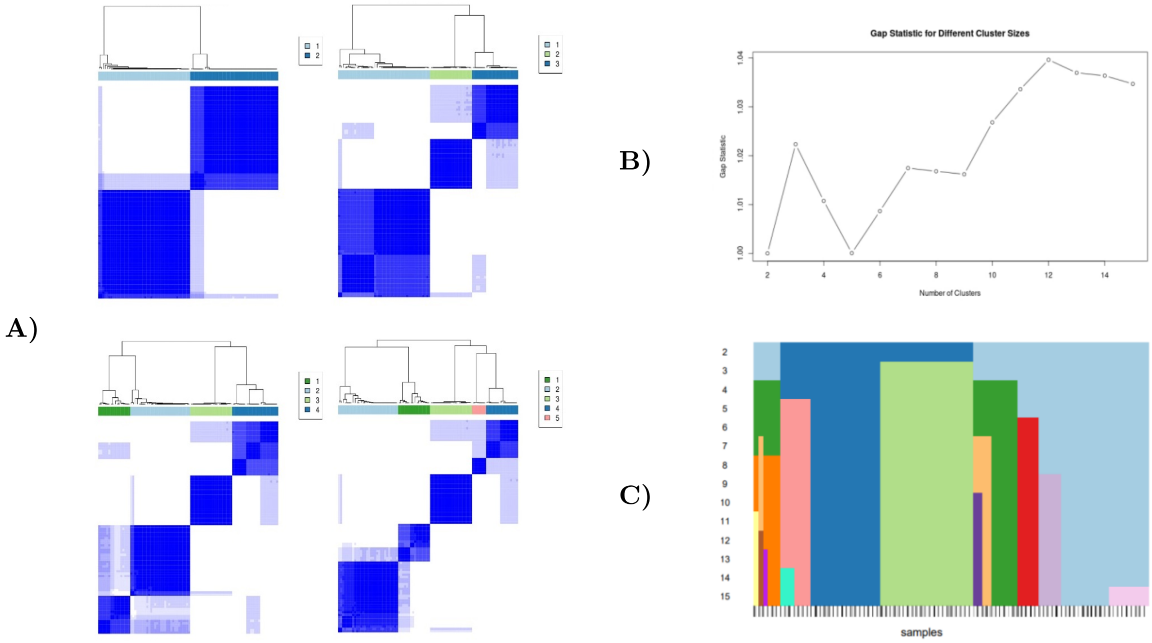

Figure 2: A) Consensus matrices for $K = 2$ to $K = 6$ produced with hierarchical clustering with 0.98 subsampling ratio and 1000 iterations using the selected 7,747 genes. B) Values of the gap statistic for different numbers of clusters with 100 Bootstrap iterations. C) Tracking plot for $K_{max} = 15$.

Following the approach of De Sousa E Melo et al., we employed hierarchical consensus clustering to obtain a clustering for the 90 patients. This utilised 7,747 genes which were selected during data pre-processing. If we filter the genes further to the 146 genes chosen for classifier construction by De Sousa E Malo et al., we are able to obtain the heatmap shown in Fig. 3, which displays the gene expression of the classified 90 patients in 3 sub-types.

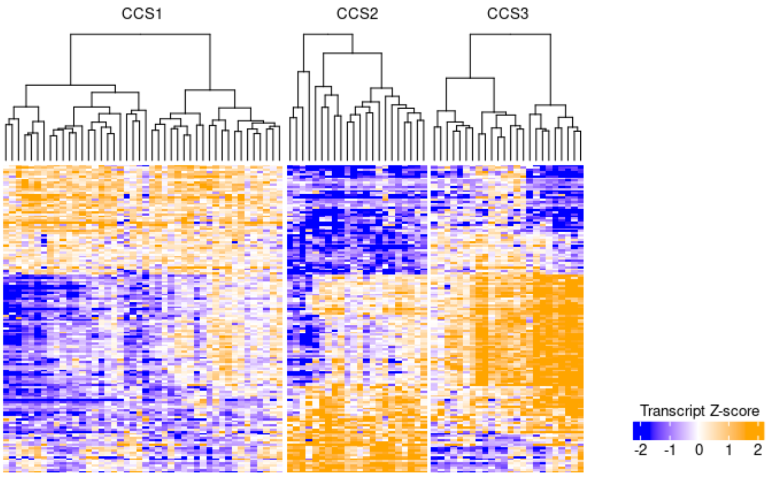

Figure 3: Heatmap of hierarchical clustering of the 90 patients (columns) using the 146 classifying genes (rows). Points correspond to median-centered $\log_2$ gene expression values (orange, high expression; blue, low expression).

We also used a Finite Gaussian mixture (FGM) model to cluster the 7,747 genes retained from the AMC-AJCCII-90 dataset. First, we fitted models over different number of components and chose the model with the highest Bayesian Information Criterion (BIC) value as the optimal. The optimal model suggested 4 components for the dataset, which is a different result from what was obtained using the gap statistic.

Comparison of Clustering Results

Classification methods	RI (Rand index)	ARI (adjusted RI)
HC with and without consensus	0.829	0.646
K means with and without consensus	1	1
K-medoids clustering with and without consensus	1	1
Model-based clustering and HC with consensus	0.7583021	0.4546874

Table 3: Measures of agreement between different clustering methods.

The adjusted Rand index (ARI) was used as the main measure for assessing the level of agreement between cluster assignment on the same data set using different approaches. Table 4 displays results of the ARI and the Rand index (RI) for pairs of methods used. Consensus clustering had no impact on cluster assignments using the k-means and k-medoids methods. There was a fair level of agreement (ARI = 0.646) between the hierarchical clustering method with and without consensus clustering. A high level of discrepancy existed between the assignments by the model-based clustering and the results from hierarchical clustering with consensus (ARI =0.4546874).

3.2 Patient Classification

We adopted the Prediction Analysis for Microarrays (PAM) method employed by De Sousa E Melo et al. to classify patients into clusters. The PAM classifier was trained using the 7,747 genes from pre-processing and 10-fold cross-validation was performed to obtain the optimal shrinkage/threshold parameter (Δ) for minimal errors. The minimum Δ value which additionally gives a smaller set of genes to be used in classification was estimated to be 4.625 (Fig. 4). At this value, we obtain only 164 significant genes to be used in classification. The trained classifier at $\Delta = 4.625$ was then used in cluster prediction for the 90 patients and results compared to predictions from other approaches such as the model-based approach.

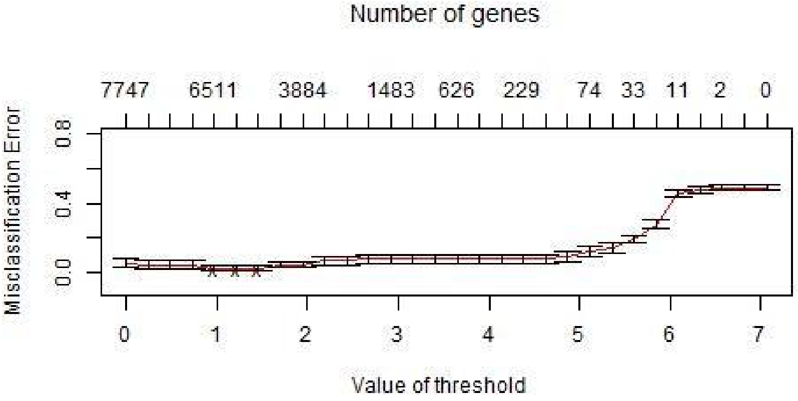

Figure 4: Plot showing cross-validation errors for different threshold levels (Δ).

4 Discussion

Within this report, we have both attempted to replicate the analysis of De Dousa E Melo et al. and assess the robustness of their findings. Because of the unavailability of some data, which in turn restricted our ability to account for batch effects, we were not able to pre-process the data in the same manner. This means that our gene filtering procedures led to somewhat different gene selections. We also question the choice of a median absolute deviation of 0.5 as a threshold for identifying those genes with the most variability. In our analysis, we found this choice fairly arbitrary, and it is possible that the clustering results may have changed if a different threshold had been chosen.

Within their paper, De Dousa E Melo et al. employ consensus clustering to assess the robustness of their clustering procedure, further utilising the gap statistic to verify that the choice of 3 clusters were appropriate. The choice of consensus clustering felt particularly appropriate given the relatively small size of the dataset, as bootstrapping enabled us to account for both random noise and outliers in the data. However, adjusting the subsampling proportion for the bootstrapping led to notably different results, calling into question how such a proportion should be selected. While using slightly different genes in our clustering, we obtained similar results and feel that De Sousa E Melo et al. tackled the issue of robustness fairly well. While consensus clustering provides a robust solution to sub-type identification, other ensemble techniques are available and would be worth exploring. For instance, heterogeneous ensemble clustering allows us to combine results from multiple clustering algorithms as well as different choices of parameters such as alternative measures of distance.

Giving a clustering of subjects, it is then important to determine whether the subjects truly differ from one another in terms of their diagnosis so that an appropriate treatment can be identified. De Sousa E Melo et al. explored the differences between the clusters by producing Kaplan-Meier plots for the individuals within each cluster. This allowed them to assess the differences in survival rate between clusters. In particular, they found that while two of the clusters had similar survival rates, the third cluster had noticeably lower survival. This suggests that for some individuals, alternative treatments may need to be considered, which further emphasises the need to identify individuals with this particular subtype of colon cancer.

References

[1] ONS, "Cancer registration statistics, England: 2016." https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/cancerregistrationstatisticsengland/previousReleases/, 2018. [Online; accessed 10-Nov-2018].

[2] BowelCancerUK, "Bowel cancer." https://www.bowelcanceruk.org.uk/about-bowel-cancer/bowel-cancer/, 2018. [Online; accessed 10-Nov-2018].

[3] S. Tarek, R. A. E., and M. Shoman, "Gene expression based cancer classification," *Egyptian Informatics Journal*, vol. 18, no. 3, pp. 151–159, 2017.

[4] EBI, "Human genetic variation (I): an introduction." https://www.ebi.ac.uk/training/online/course/human-genetic-variation-introduction/, 2018. [Online; accessed 10-Nov-2018].

[5] NIH, "What are single nucleotide polymorphisms (SNPs)." https://ghr.nlm.nih.gov/primer/genomic/research/snp/, 2018. [Online; accessed 11-Nov-2018].

[6] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proceedings of the National Academy of Sciences*, vol. 99, no. 10, pp. 6567–6572, 2002.

[7] F. De Sousa E Melo *et al.*, "Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions," *Nature medicine*, vol. 19, no. 5, p. 614, 2013.

[8] P. Jaskowiak, R. Campello, and I. Costa, "On the selection of appropriate distances for gene expression data clustering," in *BMC bioinformatics*, vol. 15, p. S2, BioMed Central, 2014.

[9] J. Handl, J. Knowles, and D. Kell, "Computational cluster validation in post-genomic data analysis," *Bioinformatics*, vol. 21, no. 15, pp. 3201–3212, 2005.

[10] R. Donamukkala, D. Huber, A. Kapuria, and M. Hebert, "Automatic class selection and prototyping for 3-d object databases," in *Proc. Int'l Conf. 3-D Digital Imaging and Modeling*, 2003.

[11] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, 2001.

[12] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," *Machine learning*, vol. 52, no. 1-2, pp. 91–118, 2003.

[13] M. Wilkerson, "ConsensusClusterPlus (Tutorial)," 2016.

[14] C. Fraley, A. Raftery, *et al.*, "Model-based methods of classification: using the mclust software in chemometrics," *Journal of Statistical Software*, vol. 18, no. 6, pp. 1–13, 2007.

[15] R. Govindarajan, J. Duraiyan, K. Kaliyappan, and M. Palanisamy, "Microarray and its applications," *Journal of pharmacy & bioallied sciences*, vol. 4, no. Suppl 2, p. S310, 2012.

[16] D. Wang *et al.*, "Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome," *Science*, vol. 280, no. 5366, pp. 1077–1082, 1998.

[17] S. Markowitz and M. Bertagnolli, "Molecular basis of colorectal cancer," *New England Journal of Medicine*, vol. 361, no. 25, pp. 2449–2460, 2009.

[18] M. de Souto, I. Costa, D. de Araujo, T. Ludermir, and A. Schliep, "Clustering cancer gene expression data: a comparative study," *BMC bioinformatics*, vol. 9, no. 1, p. 497, 2008.