DISCLAIMER:

This document does not meet the
current format guidelines of
the Graduate School at
The University of Texas at Austin.

It has been published for
informational use only.

The Report Committee for Matthew David Bramble
Certifies that this is the approved version of the following Report:

Feature-Based Clustering of Stomach Cancer Gene Expression Data

APPROVED BY
SUPERVISING COMMITTEE:

Peter Mueller, Supervisor

Christopher S. Sullivan

Feature-Based Clustering of Stomach Cancer Gene Expression Data

by

Matthew David Bramble

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Statistics

The University of Texas at Austin

May, 2018

Dedication

Julie, thank you for your encouragement and support.

Abstract

Feature-Based Clustering of Stomach Cancer Gene Expression Data

Matthew David Bramble, M.S.Stat.

The University of Texas at Austin, 2018

Supervisor: Peter Mueller

This report presents the results of using a probabilistic clustering technique in the analysis of microRNAseq and RNAseq data from gastric cancer tumor samples deposited at TCGA (The Cancer Genome Atlas). Using the method of Hoff, who has proposed a Dirichlet process unsupervised clustering framework with feature selection, it is possible to reveal interesting structure in gastric cancer gene expression data that relates to Epstein-Barr virus (EBV) microRNA levels. This structure is not as readily identified by a typical hierarchical clustering method, and the results of this analysis contribute to an understanding of the role of EBV viral microRNAs in gastric cancer tumors.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1 BACKGROUND

The National Cancer Institute's estimate for new gastric cancer diagnoses in the U.S. for 2018 is 26,240 patients, and the estimated number of deaths from gastric cancer for 2018 is 10,800. Only 31% of gastric cancer patients will survive past 5 years.[1] The rate of occurrence of Epstein-Barr virus (EBV) association among all gastric cancers is roughly 10% worldwide, ranging from about 17% in the US to 4-5% in China.[2] EBV is a ubiquitous herpes virus that is transmitted orally, establishes a lifelong latent infection, and is thought to play a role in the oncogenesis and development of gastric cancer. EBV also is one of a small number human viruses to express microRNAs (44 identified mature microRNA transcripts) that act analogous to human host microRNAs.[3] Herpes viruses account for most known viral microRNAs, and these microRNAs aid the viruses in maintaining their hallmark latency through increasing the longevity of infected cells and allowing immune response evasion, among other effects.[4] With respect to the genesis or development of epithelial cancers such as gastric adenocarcinoma, additional effects relate to cell proliferation, transformation, and other wide ranging effects that aid in the development and persistence of tumors.[5,6,7] Potential effects also, of course, include the broad range of cancer-related effects that have been found in relation to human microRNAs.[4,8]

MicroRNAs, both human and viral, are typically short RNA sequences of roughly 22 nucleotides that bind with the RNA induced silencing complex (RISC) in the cytoplasm, thereby allowing regulation of messenger RNA levels through complementary binding, typically in the 3' untranslated region of messenger RNA transcripts. In humans, the canonical microRNA processing system involves transcription of a primary microRNA sequence of hundreds to thousands of nucleotides in length, processing by the RNase III enzyme Drosha and the dsRNA binding protein DGCR8 in the nucleus to produce a stem-loop structure of about 70 base pairs, transfer to the cytoplasm via the nuclear export factor Exportin 5, and processing in the cytoplasm by a second RNase III enzyme, Dicer, to create two 22-nucleotide mature microRNA sequences in a duplex intermediate. The resultant miRNA duplex then interacts with Argonaute 2, and one strand is incorporated into RISC, while the second RNA strand is degraded.[9] The mature microRNA guides binding to complementary sequences, typically found in the 3′ UTRs of target mRNAs, thereby repressing translation and/or degrading the target. The nucleotides at positions 2–7 on the 5'-end of the mature miRNA, referred to as the "seed" region, are important for sequence-based targeting, although other regions of the microRNA sequence can contribute to target recognition. The overall structure of the RISC-microRNA complex also is not well characterized, and this overall structure influences which regions of the sequence are available for target binding, thereby adding further complexity to the interaction.[10]

1.2 TARGET IDENTIFICATION

In terms of microRNA-target interaction, it is well accepted that a single microRNA typically targets tens, and potentially over one hundred, of different mRNA transcripts and that a single mRNA may be targeted by multiple microRNAs at a plurality of mRNA binding sites.[11] Considering the number of microRNAs that have been discovered, along with the number of potential messenger RNA targets, the combinatoric complexity between even a small set of microRNAs and their targets is extremely high. Furthermore, the interaction between each RISC-associated microRNA and its target site is complex and not fully understood. It is therefore not surprising that the sensitivity and specificity of current sequence-based prediction methods are not sufficient to provide a useful characterization of targetome interaction networks. For example, although two well-established prediction tools, TargetScan and miRanda-mirSVR, provide specificities (ability to correctly identify non-targets) in the high nineties, their sensitivities (ability to correctly identify true targets) are low, at .52 and .62 respectively.[12] In light of this insufficiency with sequence based target prediction, methods that do not rely on sequence homology to predict targets are still needed to advance research into the targetome network of microRNAs, as discussed below.

On the other hand, there are a number of laboratory techniques for definitively determining specific microRNA-target interactions, which include genome editing of predicted binding sites; reporter gene assays; gene-expression after miRNA modulation; degradome sequencing; cross-linked immuno-precipitation; and biotin-linked chromatography.[13] Although such methods are good at determining whether a microRNA binds a specific target, they are unable to give direct information about how the various target inhibition effects of a set of microRNAs influence protein levels *in vivo*. This places the question of microRNA cellular effects back within a complicated network of cellular protein interaction pathways.[14,15]

Although the number of EBV microRNAs (44) is small relative to the number of human host microRNAs (>2000), the combinatoric complexity of EBV microRNA interactions with their target mRNAs is still formidable for the reasons described above, and defining the EBV microRNA targetome is in its infancy. For this reason, various analytical techniques such as differential microRNA expression analysis that probe protein network effects of microRNAs are useful.[15,16] There are multiple recent studies that attempt to find a correlation between microRNA expression levels and mRNA expression levels in omic datasets and to thereby directly probe network-level cellular effects of microRNAs based on microRNA expression levels and gene expression levels.[17,18,19]

This reports details the use of a particular probabilistic clustering technique in support of such network-level analyses. A Dirichlet process Gaussian model was used in order to perform cluster analysis on RNAseq datasets from gastric cancer tumors that have been deposited with the Cancer Genome Atlas (TCGA) project. Interesting structure within the RNAseq data was discovered when this analysis was carried out on this data in conjunction with microRNAseq data.

2. Clustering With Feature Selection

2.1 SETUP

The clustering approach described by Hoff[20] was selected for its advantages related to unsupervised clustering based on subsets of attributes in a probabilistic model that allows identification of salient features defining each cluster. There are a variety of unsupervised clustering methods that are commonly used in the analysis of correlations in omic datasets. Many of these, such as hierarchical clustering, require that the number of end clusters be determined visually and somewhat arbitrarily by slicing a dendogram. Other algorithms such as k-means require a determination of cluster number in advance and also require the choice of a distance metric, which itself can determine the way objects cluster, particularly in high-dimensional spaces.[21] The method of Hoff, on the other hand, utilizes a Dirichlet process Gibbs sampling procedure in which the number of clusters is determined in an unsupervised manner and can be potentially infinite. Determination of the number of clusters is therefore guided by the data. In addition, the metric that is used for differentiating clusters relates to the likelihood of the data, given a probability model (in this case, a mixture of Gaussians). Therefore, if the data is generally Gaussian, then the metric of the clustering method is appropriate, and no external determination of metrics need be applied.

The method of Hoff also has the advantage that the model allows inference concerning which attributes of each object contribute most to defining the clusters in which objects are grouped. For example, assuming four features in a given data set, then three distinct clusters might be formed in which features 1 and 2 contribute most to defining the first cluster, feature 1 alone contributes most to the second cluster, and features 2 and 4 contribute most to the third cluster. In high-dimension spaces, such a model also is particularly advantageous, because the data model can be specified to include all features in the initial model, without resorting to a priori reductions in dimensionality. This contrasts with methods for dimensionality reduction such as principal component analysis which can collapse, as it were, important structure that is present in the data.

In our case, the objects to be clustered are 26 EBV^+ tumor samples and 100 EBV^- tumor samples, where the features of each sample comprise RNAseq counts for specific genes. The motivation for this analysis in relation to investigating microRNA-target networks is as follows. If it were possible to carry out a clustering method that could cluster EBV^+ tumors based on the expression levels of genes (which function as features in our model), and if it were therefore possible to identify the genes that contributed most to any clustering structure that is found in the data, then this would allow inference regarding the relationship between such genes and the microRNAs deriving from EBV. The resultant genes identified as most responsible for a clustering could then be inferred as having been directly or indirectly (via gene interaction networks) influenced by the EBV microRNAs.

2.2 Model

In the feature subset clustering method of Hoff, the model is produced by parameterizing cluster membership in terms of the cluster K and m-dimensional means $\mu_1, \ldots, \mu_K$ for each cluster, where $\mu_k = \mu + r_k \times \delta_k$, $r_k \in \{0 , 1\}^m$, and $\delta_k \in \Re^m$. The vector $r_k \times \delta_k$ constitutes a vector of mean shifts for group k and may be non-zero for those features that contribute to defining the cluster. As indicated above, the r vector is a binary vector belonging to each cluster that determines whether a feature will contribute to defining that cluster. Each tumor vector y_i comprises $\log_2$ RNAseq counts and is defined as: $y_i = \mu + r_i \times \delta_i + \varepsilon_i$. The distribution of (r_i, δ_i) is modeled by the Polyá urn scheme. This model and the code developed by Hoff were used with modifications in the analysis of the GC data described above.

A summary of the probability model is presented below.

$$
\begin{aligned}
&p(c, \alpha, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) \\
f \quad &\sim \quad \text{Dirichlet}(\alpha, f_0) \qquad (1) \\
\{\boldsymbol{r}_1, \boldsymbol{\delta}_1\}, \ldots, \{\boldsymbol{r}_n, \boldsymbol{\delta}_n\} \quad &\sim \quad \text{i.i.d. } f \\
\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n \quad &\sim \quad \text{i.i.d. multivariate normal}(\mathbf{0}, \text{diag}\{\sigma_1^2, \ldots, \sigma_m^2\}) \\
\boldsymbol{y}_i \quad &= \quad \boldsymbol{\mu} + \boldsymbol{r}_i \times \boldsymbol{\delta}_i + \boldsymbol{\epsilon}_i.
\end{aligned}
$$

2.3 Posterior MCMC

In accordance with the Dirichlet distribution, objects that are assigned to clusters are assigned in an exchangeable sequence. Therefore, in the Polyá urn Dirichlet process, draws from a Dirichlet distribution of cluster assignments are created for the n (i = 1:N) tumors by simply removing each tumor vector from its cluster assignment, updating the cluster parameters of its former cluster, and then re-assigning the tumor to either an existing cluster with a probability that is dependent on the number of members in the existing cluster, or a new cluster with a probability that is dependent on the updated Dirichlet parameter α, as discussed below.

$$
\Pr(c(i) = k | c(i'), i' \neq i, \alpha, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\lambda}, \boldsymbol{\tau}^2, y_1, \ldots, y_n) \propto \begin{cases} n_{k,-i} \times w_k & \text{if } k < K + 1 \\ \alpha \times w_{K+1} & \text{if } k = K + 1 \end{cases}
$$

$$
\begin{aligned}
w_k \quad &= \quad \prod_{j=1}^{m} \frac{1 + \exp\{\lambda_j + \hat{\lambda}_j^{+i}(k)\}}{1 + \exp\{\lambda_j + \hat{\lambda}_j^{-i}(k)\}} \quad \text{if } k < K + 1 \\
w_{K+1} \quad &= \quad \prod_{j=1}^{m} \frac{1 + \exp\{\lambda_j + \hat{\lambda}_j^{+i}(K)\}}{1 + \exp \lambda_j},
\end{aligned}
$$

Each weight is the ratio of the probability of the data for a cluster including the reassigned ith sample to the probability of the data for the cluster not including the reassigned ith sample. Therefore weights increase or decrease in proportion to the probabilities of the data under the new cluster assignments, and draws from that probability distribution will be successively updated, with assignments being guided by the data as the clustering space is explored.

The probability that a removed element is assigned to an existing cluster is proportional to the number of members in the cluster and the data in the cluster, parameterized in terms of λ_j which is the prior log-odds that the mean of attribute j in a cluster k differs from the mean of the attribute for the other clusters.

Updates to the other parameters of the probability model, $\{r_{(k)},\delta_{(k)}\}$, α, μ, σ^2, are made using the full conditionals of the parameters.

The following is a description concerning actual sampling of the parameters.

Sampling $\{r_{(k)},\delta_{(k)}\}$: Sampled as follows given the prior distribution:

$$f_0(\boldsymbol{r},\boldsymbol{\delta}) = \prod_{j=1}^{m} \text{binary}(r_j|\frac{e^{\lambda_j}}{1+e^{\lambda_j}}) \times \text{normal}(\delta_j|0,\tau_j^2).$$

$r_{(k)1}$, . . . $r_{(k)m}$ is sampled from {0,1} with log-odds $\lambda_j + \hat{\lambda}_{j(k)}$ after updating according to the formula below:

$$\hat{\lambda}_{j(k)} \quad = \quad \frac{1}{2}\left\{\frac{\tau_j^2}{\tau_j^2+\sigma_j^2/n_k}\frac{n_k}{\sigma_j^2}\overline{\xi_j(k)}^2 + \log\frac{\sigma_j^2/n_k}{\sigma_j^2/n_k+\tau_j^2}\right\} \ ,$$

conditional on the data, parameters τ, σ^2, and $\overline{\xi_j(k)}$, which represents the value of $y_{i,j} - \mu_j$, averaged over samples i assigned to group k. When $r_{(k),j} = 1$, then $\delta_{(k),j}$ is sampled ~N($\hat{\delta}_j$, $\hat{\tau}$j 2), with $\hat{\delta}_j = \hat{\tau}_j^2(\sum_{i:c(i)=k}(y_{i,j}-\mu_j)/\sigma_j^2)$ and $\hat{\tau}_j^2 = (n_k/\sigma_j^2 + 1/\tau_j^2)^{-1}$.

If $r_{(k),j} = 1$, then $\delta_{(k),j}$ ~N(0, τ_j^2).

Sampling α : In the sampling method of Hoff, α is reparamaterized to π, where $\pi = \alpha/(\alpha + 1) \in (0,1)$. The full conditional of α thus becomes:

$$p(\pi|K) \propto p(\pi) \times \left(\frac{\pi}{1-\pi}\right)^K \frac{\Gamma[\pi/(1-\pi)]}{\Gamma[\pi/(1-\pi)+n]}.$$

Sampling from this probability distribution is not computationally straightforward. However, because $\alpha/(\alpha + 1) \in (0,1)$ and $p(\pi)$ is a function only of K and n, if the value of the log of the function is calculated at a series of equidistant points between 0 and 1, then samples of alpha from the full conditional can be obtained by sampling from this grid of points in accordance with the assigned probabilities.

Sampling μ: The mean μ of parameter j μ_j is sampled as follows:

$$\mu_j \sim \text{normal}(\hat{\mu}_j, \hat{\sigma}_j^2) \ , \text{ where } \hat{\sigma}_j^2 = (n/\sigma_j^2 + 1/v)^{-1}, \ \hat{\mu}_j = \hat{\sigma}_j^2(\textstyle\sum_{i=1}^{n} \varepsilon_{i,j}/\sigma_j^2 + m/v)$$

and υ is a parameter of the prior for μ_j.

Sampling σ^2_j: The precision $1/\sigma^2_j$ is sampled from the full conditional as follows:

$$1/\sigma_j^2 \sim \mathrm{gamma}[\nu_1 + n/2, \nu_2 + \textstyle\sum_{i=1}^{n}(\varepsilon_{i,j} - \mu_j)^2)/2]$$

where ν_1, and ν_2 are parameters of the prior for the precision.

Sequential updating of the parameters as described above provides convergence on estimates for the parameters c, r, α, μ, σ^2, which can be obtained by averaging samples taken over a suitable range of iterations of the algorithm. In our case, the algorithm was run for 10,000 iterations, and the final 1000 iterations were used as samples for cluster assignment mode determination.

3. Data Analysis

3.1 DATA

MicroRNAseq data and RNAseq data were obtained from TCGA through the GDC data portal.[22] the complete microRNAseq raw counts dataset consisting of roughly 400 gastric cancer tumor samples was downloaded, and alignment of the microRNAseq reads to EBV stem-loop sequences registered at miRBase was carried out using Bowtie2 (Version 2.3.4). The count data was represented as counts per million of the transcripts mapped to human stem-loop and EBV stem-loop sequences for microRNA data. In the RNAseq expression datasets obtained from TCGA, the data was expressed in counts per kilobase (transcript length) per million transcripts mapped to human RNA sequences (TPKM). The latter datasets are preprocessed data made available by TCGA, and no further pre-processing was carried out on the data.

Alignment of microRNAseq reads to EBV microRNA stem-loop sequences yielded a range of counts for each of the 44 EBV microRNAs across all of the tumors, and a histogram of count values (normalized with respect to total human microRNA level) was constructed in order to identify which of the roughly 400 tumors was EBV$^+$. As can be seen from the histogram data in Fig. 1 and in Fig. 2 which shows an expanded count region from Fig. 1, the number of tumors having normalized total EBV microRNA counts of 0.00 to < 0.02 portrays a discernible stochastic tail attributable to random mappings of transcripts to the EBV stem-loop sequences, with the tail falling off rapidly to zero at around 0.001. It can be inferred that random mappings are responsible for the counts in this tail region. Based on these results, an EBV microRNA/human microRNA ratio of 0.02 was taken as the cutoff in the determination of whether significant EBV microRNA load exists. This yielded 26 tumor samples with significant EBV load (referred to below as EBV$^+$ tumors). This EBV$^+$ tumor count of 26 is supported by the conclusions of others in the field who have analyzed the same TCGA gastric cancer data.[23] All of these 26 EBV$^+$ tumors, along with 100 of the EBV$^-$ tumors selected at random from the remainder of the tumors, were used in subsequent analysis, giving 126 objects to be clustered. These objects were clustered based on features comprising 95 experimentally-verified EBV microRNA targets, as will be described below.

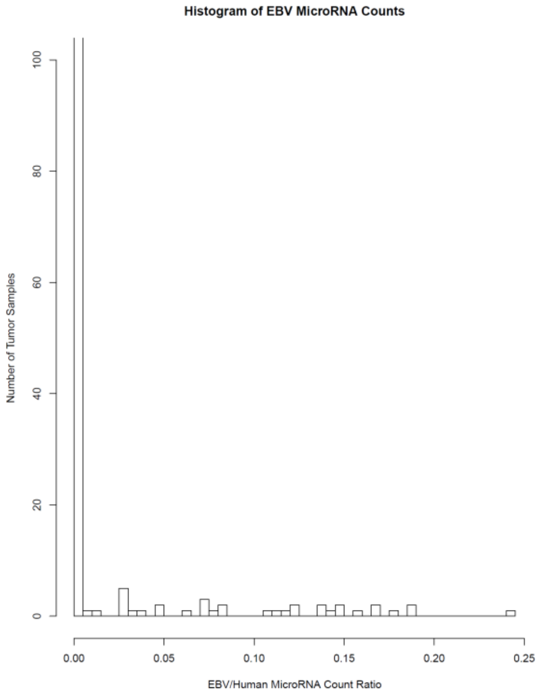

Fig. 1. Distribution of EBV microRNA Loads

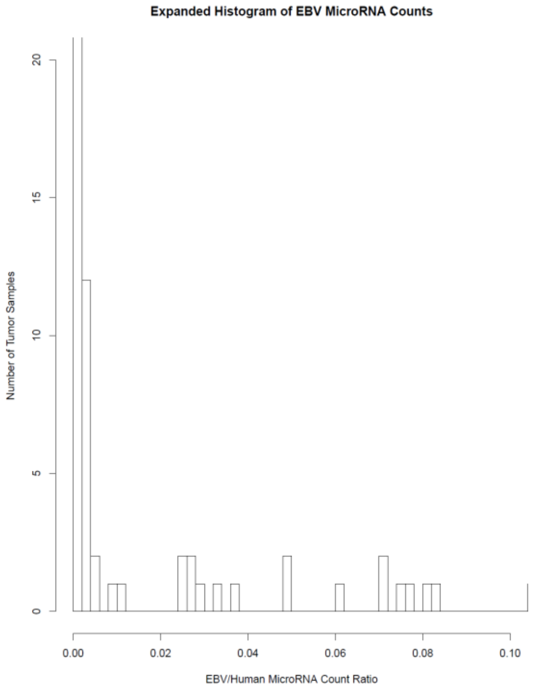

Fig. 2. Expanded Region from Fig. 1

Regarding the RNAseq gene expression data obtained from TCGA, ENSG gene names were first converted to standard gene names using the BIOMART R package. In order to run a limited dataset to test the practicality of the clustering algorithm, clustering was carried out on a set of features (genes) know to be targeted by EBV microRNA. The curated genes were found through a literature search of EBV microRNA gene targets that have been experimentally confirmed by various methods, including HITS-CLIP, PAR-CLIP, and luciferase reporter knockdown.[2,3,7,24,25] In addition, because there were roughly 400 EBV$^-$ tumors, making clustering with this algorithm intractable due to running time, 100 tumors were selected at random among the EBV$^-$ tumors. Therefore, the final data included a total of 126 tumors, each having 96 features (gene expression counts). No gene feature had more than 10% missing values, and missing values were replaced using the mean value of the feature determined for those tumors for which the feature was not missing (using the respective means within the EBV+ and EBV- groups for missing values within those groups).

The amenability of the data to clustering using the proposed Dirichlet Gaussian mixture model depends on the features being normally distributed. A $\log_2$ transformation of the data resulted in surprisingly strong normality, as can be seen in the histograms for the features shown in Fig. 3. The y-axis represents the number of tumor samples, and the x-axis represents the $\log_2$ values of the RNAseq count data for the feature (FPKM; frequency per kilobase of transcript per million mapped reads).

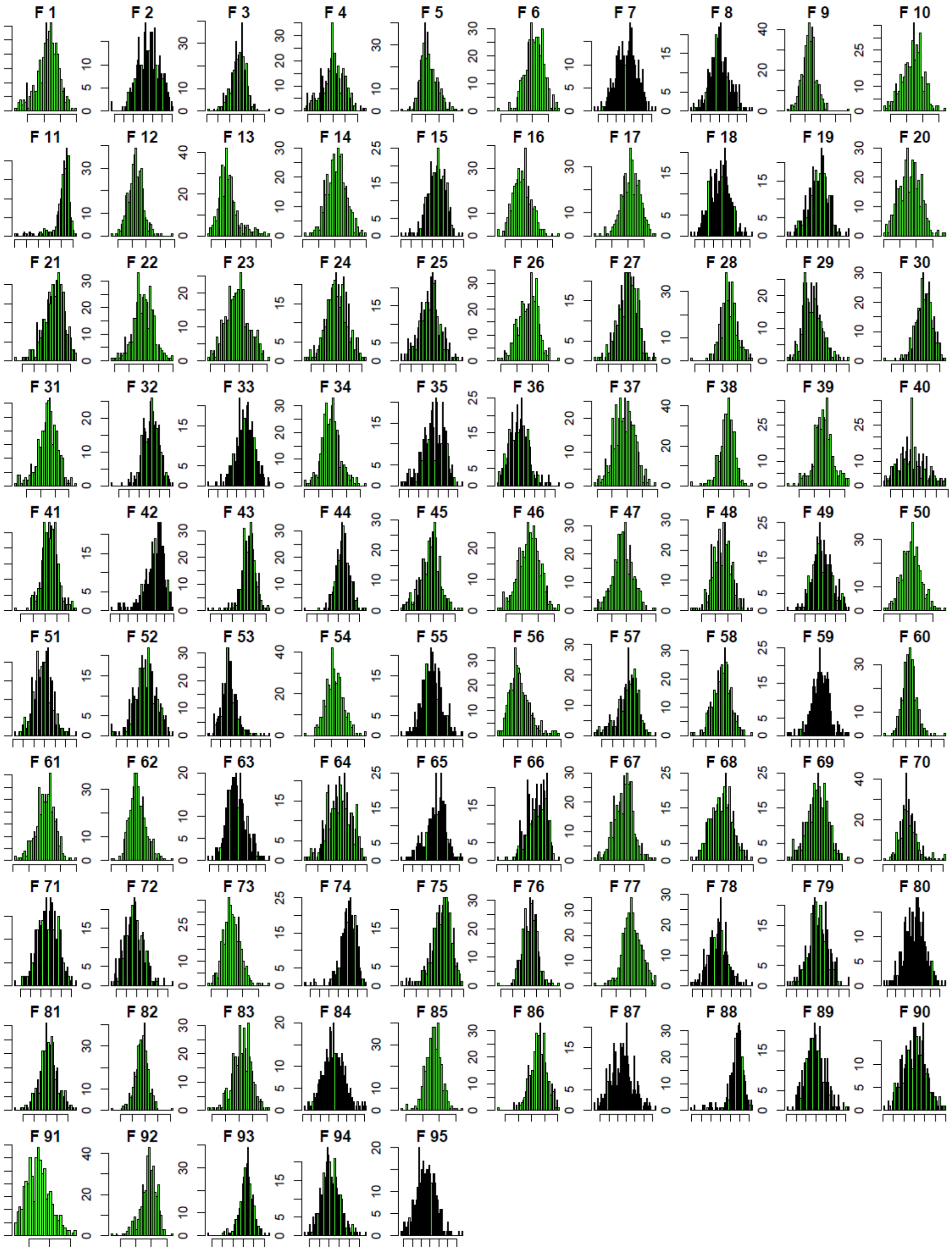

Fig. 2 Distributions for Curated Gene Features

3.2 Clustering With Feature Selection

Clustering of the data matrix was carried out using code adapted from the R and C code produced by Hoff.[20] To aid in visual inspection of the resulting cluster assignment vector, the tumor vector supplied to the clustering algorithm was arranged with the EBV$^+$ tumors placed as the first 26 elements, arranged from increasing to decreasing total EBV microRNA load, with the remaining 100 EBV$^-$ tumors arranged randomly. The vector of cluster assignments shown in Fig. 3 was obtained, indicating cluster assignments for each of the 126 tumors in the above arrangement. The upper row of each section of the figure denotes the index of each tumor, and the lower row indicates the cluster assignment. As can be seen from the figure, among the EBV$^+$ tumors (orange or green indices), there is a striking clustering in lower EBV microRNA loads (indices 11-26), where 12 of the 16 tumor samples are assigned to cluster 2.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
5	5	16	10	5	10	1	5	10	5	2	2	2	10	5	10	2	2	2	2

21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
2	2	2	2	2	4	3	3	11	1	3	3	4	1	4	12	6	2	8	4

41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
12	1	13	14	6	14	1	11	2	7	1	5	8	3	7	9	6	7	3	9

61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
9	14	1	3	17	5	3	7	1	9	3	1	13	7	1	12	15	13	8	6

81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
4	8	7	7	6	4	11	16	1	6	6	15	9	8	13	4	11	1	1	1

100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115
7	15	9	5	3	3	3	4	4	6	1	11	12	4	12	8

116	117	118	119	120	121	122	123	124	125	126
7	4	3	13	2	1	18	5	6	8	8

Fig. 3. Complete Cluster Assignments

The ordering of the EBV$^+$ tumors (indices 1-26) was set based on the ratio of total mapped EBV microRNA load to total mapped microRNA load (total mapped EBV microRNA load + total mapped human microRNA load). Given that this ordering revealed an interesting grouping within the lower EBV microRNA loads, the clusterings were investigated under different tumor orderings. The two other parameters available for reordering the tumors were mapped EBV microRNA load/library size ratio and mapped human microRNA load/library size ratio. The cluster assignments of the EBV$^+$ tumors corresponding to indices 1-26 are shown below under these alternate orderings.

EBV/Human	Human/Library	EBV/Library
5	4	2
5	2	10
16	2	1
10	2	5
5	2	5
10	2	5
1	2	10
5	2	10
10	2	16
5	2	5
2	2	5
2	2	2
2	10	2
10	5	2
5	5	2
10	10	10
2	1	10
2	10	2
2	2	5
2	10	2
2	5	2
2	10	2
2	5	2
2	5	2
2	16	4
4	5	2

Table 1. Comparison of Clusters Under Different Orderings

In Table 1, when the EBV$^+$ tumors were first ordered by the ratio of EBV microRNA load to human microRNA load *(EBV/Human ratio*; left column), a strong clustering was initially obtained. This ratio was used in order to normalize the EBV microRNA counts with respect to variation in sample preparation, thus allowing comparison across microRNAseq libraries. Fig. 4 below shows the correlation between EBV microRNA levels and human microRNA levels. It is clear that human microRNA levels increase with increasing EBV microRNA levels, as would be expected because the two together make up the total number of fragments in each microRNA library. Indeed, when the tumors are ordered by human microRNA load (center column), the detected cluster association shifts to the higher human microRNA loads. The simple correlation referred to above explains the differing clustering orientations in the three columns in Table 4. However, it could still be the case that human microRNA levels are the main cause, or a partial cause, of the structure found in the data. An additional test that could be carried out would be to normalize with respect to tumor purity estimated based on RNAseq data, or even to simply normalize based on RNAseq library size. It is clear that the normalization method that is used here is not transparent and could potentially influence the results.

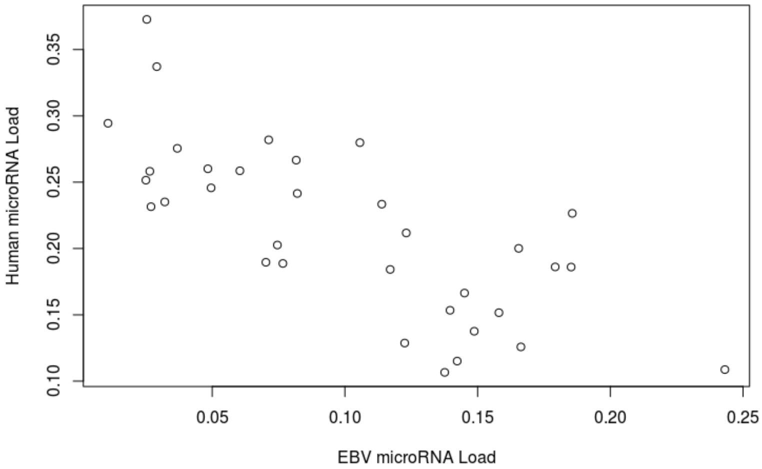

Fig. 4 Correlation Between Human and EBV microRNA Levels

Clustering was also carried out using the standard HCLUST R package in order to determine whether a typical hierarchical clustering method could also find this structure in the data. Given that the data expresses the degree of normality shown above, the standard squared Euclidean distance and centroid linkage method of HCLUST was used. The results are shown in Fig 5. As shown in the figure, clustering of the low-EBV-load tumors is portrayed in the dendogram and the circularized dendogram. Although a relative clear clustering can be seen in the circularized dendogram, the clustering is not nearly as pronounced as that seen with the probabilistic clustering method. Furthermore, if the traditional horizontal dendogram is inspected in the attempt to fix a cut point for the dendogram to obtain the best grouping of all of the more clustered EBV$^+$ objects (e.g., those with the lower EBV loads), it is not at all clear where to place the vertical cut line. Moreover, the method is not able to provide any information regarding which of the features are more significant than others in defining the clusters.

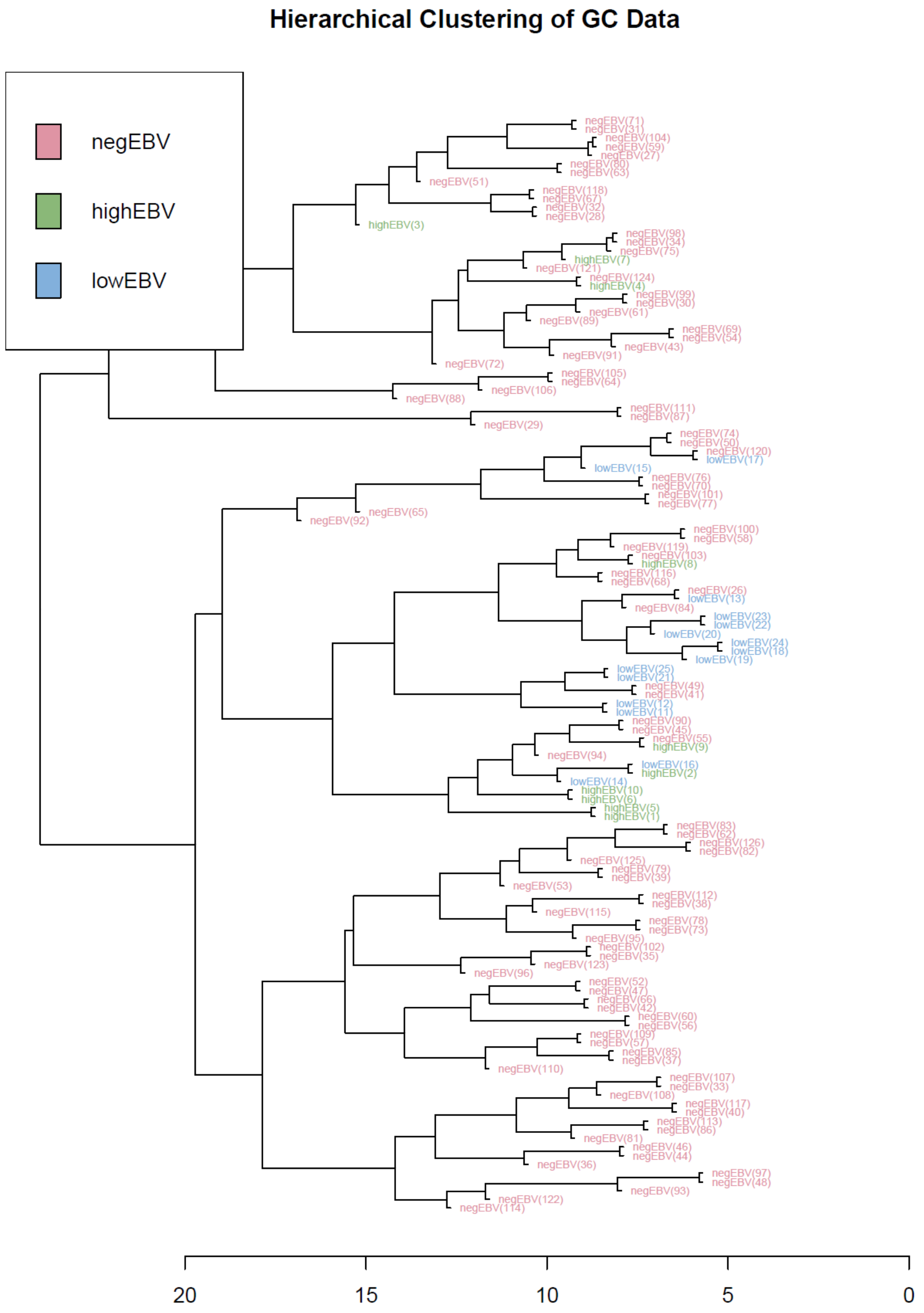

Fig. 5 Hierarchical Clustering of Gastric Cancer Data

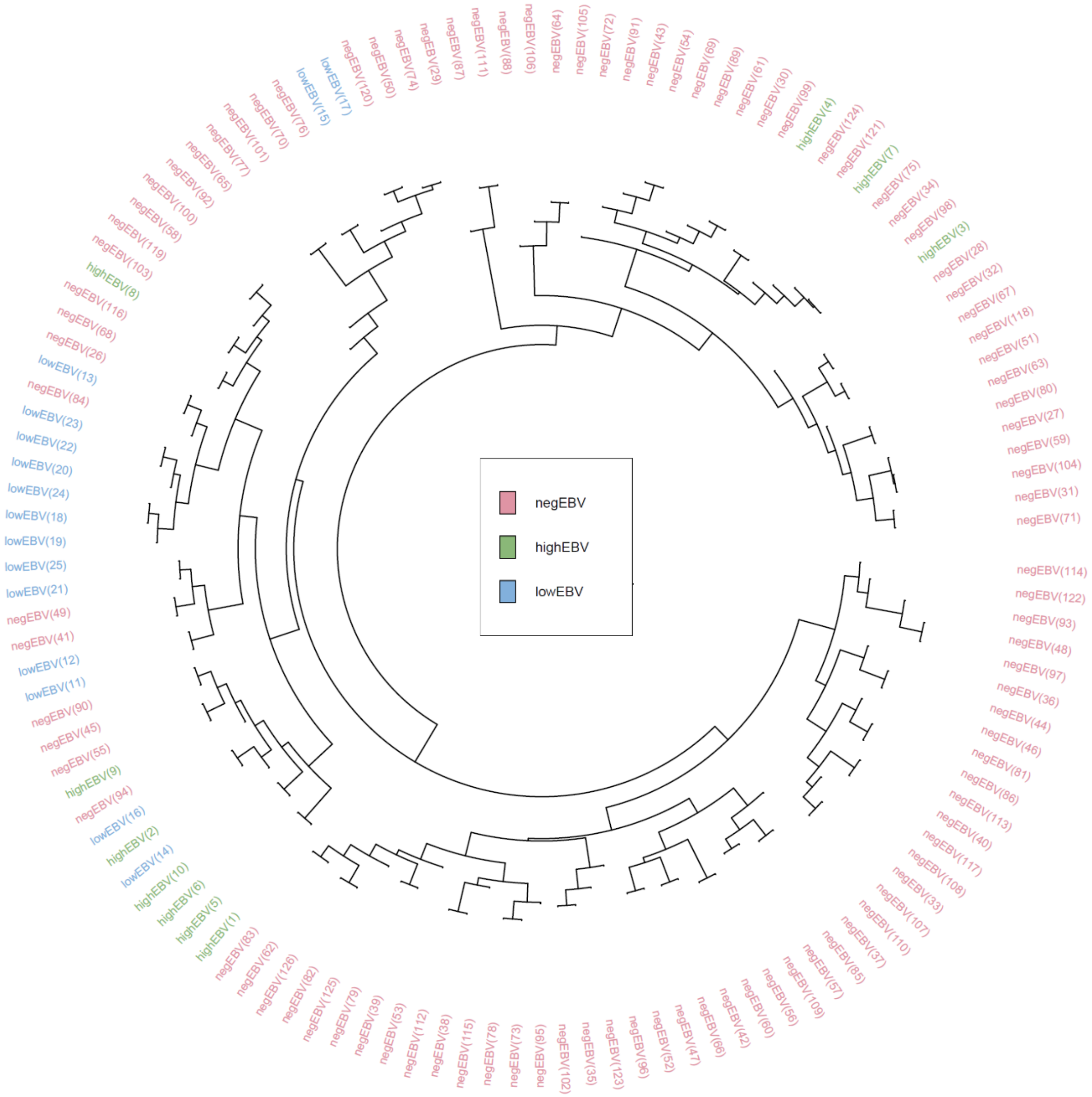

Fig. 5 Hierarchical Clustering of Gastric Cancer Data

3.3 ANALYSIS OF RELEVANCE

In determining the relevance of particular features, the parameters of each of the clusters were sampled in the same manner as described above using the clustering algorithm, but with the cluster assignment fixed for each object. A relevance matrix was thereby obtained by summing the r_k vector at each MCMC iteration and dividing by the total number of samples. In the above model of f_0, a beta(a_λ,b_λ) prior is placed on the reparamaterized form of λ: $e^\lambda/(1 + e^\lambda)$:

$$f_0(\boldsymbol{r}, \boldsymbol{\delta}) = \prod_{j=1}^{m} \text{binary}(r_j | \frac{e^\lambda}{1+e^\lambda}) \times \text{normal}(\delta_j | 0, \tau_j^2 = \eta \times \sigma_j^2)$$

A beta(1,1) prior was used by default in the analysis, and a scatter plot of relevance and feature number is shown for the beta(1,1) prior in upper part of Fig. 6 below. As can be seen, there are roughly 45 features that have a relevance of greater than .95, and at the point when relevance begins to drop off appreciably, the number of features increases to over 60. This does not help much in identifying a small set of features that contribute to the clustering, although it does provide a clue about which features do not contribute significantly to clustering. With the aim of gaining a more restrictive test for relevant features, the parameters of the beta prior were switched to values that provide sampling close to zero and close to one. With a beta(.01,100) prior shown in the bottom right plot, a more restrictive relevance test was obtained at the cutoff of .95 insofar as the number of relevant features was decreased to 25. However as is seen in the plot, this success is not exactly functional, as there are still many features with very similar relevance values in the neighborhood of the cutoff, and the critical point is therefore somewhat arbitrary. Given the seeming arbitrariness of determining a cutoff, one might simply select the top n features, provided they achieve a certain degree of relevance (e.g., 0.95). The features with the top ten relevance values are therefore listed below.

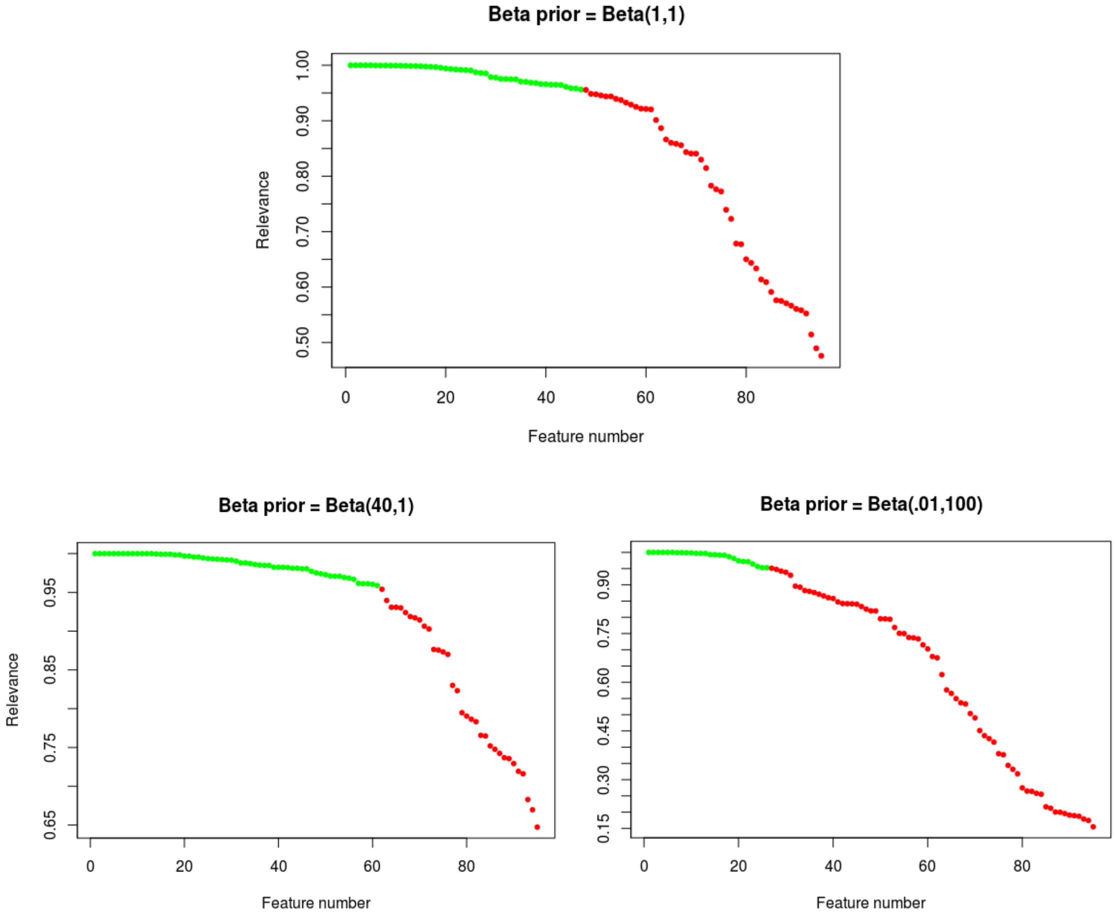

Fig. 6. Results of Testing Different Beta Prior Parameters

The following is a list and brief description of the top ten features identified by feature selection obtained from the UniProt database.[26]

CLEC7A

Lectin that functions as pattern receptor specific for beta-1,3-linked and beta-1,6-linked glucans, such as cell wall constituents from pathogenic bacteria and fungi. Necessary for the TLR2-mediated inflammatory response and for TLR2-mediated activation of NF-kappa-B. Enhances cytokine production in macrophages and dendritic cells. Mediates production of reactive oxygen species in the cell. Mediates phagocytosis of *C.albicans* conidia. Binds T-cells in a way that does not involve their surface glycans and plays a role in T-cell activation. Stimulates T-cell proliferation (By similarity).

CXCL11

Chemotactic for interleukin-activated T-cells but not unstimulated T-cells, neutrophils or monocytes. Induces calcium release in activated T-cells. Binds to CXCR3. May play an important role in CNS diseases which involve T-cell recruitment. May play a role in skin immune responses.

PDCD1LG2

Involved in the costimulatory signal, essential for T-cell proliferation and IFNG production in a PDCD1-independent manner. Interaction with PDCD1 inhibits T-cell proliferation by blocking cell cycle progression and cytokine production (By similarity).

CMKLR1

Receptor for the chemoattractant adipokine chemerin/RARRES2 and for the omega-3 fatty acid derived molecule resolvin E1. Interaction with RARRES2 induces activation of intracellular signaling molecules, such as SKY, MAPK1/3 (ERK1/2), MAPK14/P38MAPK and PI3K leading to multifunctional effects, like, reduction of immune responses, enhancing of adipogenesis and angionesis. Resolvin E1 down-regulates cytokine production in macrophages by reducing the activation of MAPK1/3 (ERK1/2) and NF-kappa-B. Positively regulates adipogenesis and adipocyte metabolism. Acts as a coreceptor for several SIV strains (SIVMAC316, SIVMAC239, SIVMACL7E-FR and SIVSM62A), as well as a primary HIV-1 strain (92UG024-2).

CCR2

Receptor for the CCL2, CCL7 and CCL13 chemokines (PubMed:23408426). Receptor for the beta-defensin DEFB106A/DEFB106B (PubMed:23938203). Transduces a signal by increasing intracellular calcium ion levels (By similarity). Upon CCL2 ligation, mediates chemotaxis and migration induction through the activation of the PI3K cascade, the small G protein Rac and lamellipodium protrusion (Probable).

CLEC2D

Receptor for KLRB1 that protects target cells against natural killer cell-mediated lysis (PubMed:20843815, PubMed:16339513). Inhibits osteoclast formation (PubMed:14753741, PubMed:15123656). Inhibits bone resorption (PubMed:14753741). Modulates the release of interferon-gamma (PubMed:15104121). Binds high molecular weight sulfated glycosaminoglycans (PubMed:15123656).

CD200R1

Inhibitory receptor for the CD200/OX2 cell surface glycoprotein. Limits inflammation by inhibiting the expression of proinflammatory molecules including TNF-alpha, interferons, and inducible nitric oxide synthase (iNOS) in response to selected stimuli. Also binds to HHV-8 K14 viral CD200 homolog with identical affinity and kinetics as the host CD200

CARD8

Inhibits NF-kappa-B activation. May participate in a regulatory mechanism that coordinates cellular responses controlled by NF-kappa-B transcription factor. May be a component of the inflammasome, a protein complex which also includes PYCARD, NALP2 and CASP1 and whose function would be the activation of proinflammatory caspases.

CXCL16

Acts as a scavenger receptor on macrophages, which specifically binds to OxLDL (oxidized low density lipoprotein), suggesting that it may be involved in pathophysiology such as atherogenesis (By similarity). Induces a strong chemotactic response. Induces calcium mobilization. Binds to CXCR6/Bonzo.

SH2B3

Links T-cell receptor activation signal to phospholipase C-gamma-1, GRB2 and phosphatidylinositol 3-kinase.

Considering that many of these features seem to be involved in immune response, and that it is likely that a major function of EBV microRNAs relates to immune system avoidance and inhibition of T-cell recognition, the analytical tool CIBERSORT was used in order to determine whether there is any difference in immune cell ratios both between EBV$^+$ and EBV$^-$ tumors, and between the low-load EBV$^+$ tumors and high-load EBV$^+$ tumors. CIBERSORT uses RNAseq expression levels for a subset of roughly 500 genes in order to estimate relative and absolute proportions of infiltrating immune cells in a cell sample. An RNAseq expression matrix for the TCGA gastric cancer tumor samples used above was subsetted using the 500 genes, and the data matrix was submitted online to Cibersort. The result shown in Fig. 7 was obtained. The same ordering of tumor samples was used as in the previous analysis (upper 26 rows: EBV$^+$, ordered from higher to lower EBV microRNA loads). The data obtained for the EBV$^+$ and EBV$^-$ tumor samples was roughly normal, and so the results were compared using simple T-tests. With regard to differences between EBV$^+$ and EBV$^-$ tumors, significant differences between the two groups were obtained for T cells CD8$^+$, T cells CD4 memory resting, T cells CD4 memory activated, Macrophages M0, and Macrophages M1, as shown in Table

7. The same test was carried out for the high-microRNA load and low-microRNA load EBV$^+$ groups, but significant results were not obtained.

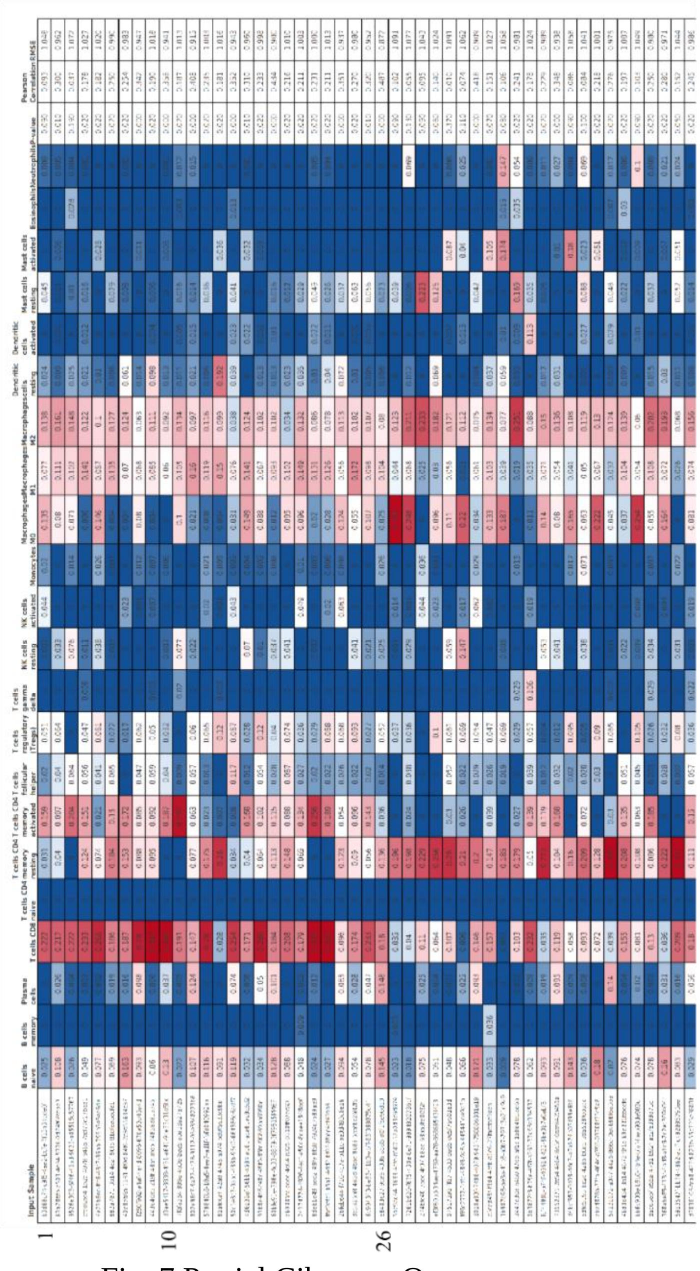
Input Sample
B cells naive
B cells memory
Plasma cells
T cells CD8
T cells CD4 naive
T cells CD4 memory resting
T cells CD4 memory activated
T cells follicular helper
T cells regulatory (Tregs)
T cells gamma delta
NK cells resting
NK cells activated
Monocytes
Macrophages M0
Macrophages M1
Macrophages M2
Dendritic cells resting
Dendritic cells activated
Mast cells resting
Mast cells activated
Eosinophils
Neutrophils
P-value
Pearson Correlation
RMSE
1
10
26

Fig. 7 Partial Cibersort Output

Immune Cell Type	P Value EBV+ vs. EBV^{-}	P value, EBV^{+} low-load vs. EBV^{+} high-load
B cells memory	0.004167	-0.01796
T cells CD8	1.07×10^{-7}	0.109585
T cells CD4 memory resting	5.58×10^{-08}	-0.09971
T cells CD4 memory activated	2.49×10^{-05}	0.075827
Macrophages M0	3.01×10^{-05}	-0.0637
Macrophages M1	1.22×10^{-07}	0.047537

Table 2. T-Test Results of EBV^{+} vs. EBV^{-} Tumors and EBV^{+} High-Load and Low-Load Tumors

4. Conclusion

The feature subset clustering method of Hoff was used in order to analyze gastric cancer data in the hope of revealing structure among tumors in relation to EBV status. Insofar as the clustering method of Hoff provides information regarding relevant features responsible for clustering, while also allowing these relevant features to vary among clusters, the method was considered as a potential method for revealing targets of EBV microRNAs that does not rely on sequence-based prediction. The results show that the method allowed easier interpretation of clustering data in comparison to a standard hierarchical clustering method, which ultimately did not afford an obvious way to identify relevant clusters. The probabilistic clustering method provided more obvious cluster assignments, and indeed a striking clustering was observed that corresponded to the level of EBV microRNA load within the EBV$^+$ tumors. With regard to determination of feature relevance, the method was less fruitful than had been hoped. While the method provided relevance probabilities, a large number of features (two-thirds) had probabilities in the 90% or greater range, with probabilities falling off dramatically in only roughly one-third of the features. From a practical standpoint, this still leaves a great many features to investigate as potentially significant drivers of the clustering. Even when placing an extremely stringent prior on the λ parameter in the model, the number of features having high significance could not be reduced from a practical standpoint. That being said, it is probably the case that selection of confirmed EBV microRNA target genes likely contributed to this high proportion of relevant genes. Also, considering current understanding of microRNA biology, it may be the case that the results are an accurate reflection of the diffuse and complex manner in which microRNAs cooperate to influence expression of a large number of genes at once. Future analysis would benefit from a more detailed investigation of the specific EBV microRNAs that correlate with the cluster that was found. Also, using a more random selection of genes that are implicated in aspects of cancer development, not just those that are confirmed microRNA targets, may lead to new discovery of genes or sets of genes that are influenced by EBV microRNAs.

References

1. National Cancer Institute, Surveillance, Epidemiology, and End Results Program. https://seer.cancer.gov/statfacts/html/stomach.html (Accessed April 19, 2018).

2. Kanda, T. *et al.* Clustered MicroRNAs of the Epstein-Barr Virus Cooperatively Downregulate an Epithelial Cell-Specific Metastasis Suppressor. *J. Virol.* **89,** 2684–2697 (2015).

3. Skalsky, R. L. *et al.* The Viral and Cellular MicroRNA Targetome in Lymphoblastoid Cell Lines. *PLoS Pathog* **8,** (2012).

4. Kincaid, R. P., Sullivan, C. S. Virus-Encoded microRNAs: An Overview and a Look To the Future. *PLoS Pathog* **8,** (2012).

5. Marquitz, A. R., Mathur, A., Nam, C. S. & Raab-Traub, N. The Epstein-Barr Virus BART microRNAs target the pro-apoptotic protein Bim. *Virology* **412,** 392–400 (2011).

6. Zhang, J. *et al.* The oncogenic role of Epstein–Barr virus-encoded microRNAs in Epstein–Barr virus-associated gastric carcinoma. *Journal of Cellular and Molecular Medicine* **22,** 38 (2018).

7. Albanese, M., Tagawa, T., Buschle, A. & Hammerschmidt, W. MicroRNAs of Epstein-Barr Virus Control Innate and Adaptive Antiviral Immunity. *J. Virol.* **91,** e01667-16 (2017).

8. Lujambio, A. & Lowe, S. W. The microcosmos of cancer. *Nature* **482,** 347–355 (2012).

9. Burke, J. M., Kelenis, D. P., Kincaid, R. P. & Sullivan, C. S. A central role for the primary microRNA stem in guiding the position and efficiency of Drosha processing of a viral pri-miRNA. *RNA* **20,** 1068–1077 (2014).

10. Cullen, B. R. MicroRNAs as Mediators of Viral Immune Evasion. *Nat Immunol* **14,** 205–210 (2013).

11. Hashimoto, Y., Akiyama, Y. & Yuasa, Y. Multiple-to-Multiple Relationships between MicroRNAs and Target Genes in Gastric Cancer. *PLoS One* **8,** (2013).

12. Oliveira, A. C. *et al.* Combining Results from Distinct MicroRNA Target Prediction Tools Enhances the Performance of Analyses. *Front Genet* **8,** (2017).

13. Cloonan, N. Re-thinking miRNA-mRNA interactions: Intertwining issues confound target discovery. *Bioessays* **37,** 379–388 (2015).

14. Haecker, I. & Renne, R. HITS-CLIP and PAR-CLIP advance viral miRNA targetome analysis. *Crit Rev Eukaryot Gene Expr* **24,** 101–116 (2014).

15. Bracken, C. P., Scott, H. S. & Goodall, G. J. A network-biology perspective of microRNA function and dysfunction in cancer. *Nature Reviews. Genetics; London* **17,** 719–732 (2016).

16. Wang, W. *et al.* MicroRNA profiling of CD3+CD56+ cytokine-induced killer cells. *Scientific Reports* **5**, 9571 (2015).

17. Liu, B. *et al.* Identifying functional miRNA–mRNA regulatory modules with correspondence latent dirichlet allocation. *Bioinformatics* **26,** 3105–3111 (2010).

18. Petralia, F. *et al.* A new method to study the change of miRNA–mRNA interactions due to environmental exposures. *Bioinformatics* **33,** i199–i207 (2017).

19. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE* **5,** (2010).

20. Hoff, P. D. Model-based subspace clustering. *Bayesian Anal.* **1,** 321–344 (2006).

21. Sirinukunwattana, K., Savage, R. S., Bari, M. F., Snead, D. R. J. & Rajpoot, N. M. Bayesian Hierarchical Clustering for Studying Cancer Gene Expression Data with Unknown Statistics. *PLOS ONE* **8,** e75748 (2013).

22. NCI Genomic Data Commons. Available at: https://portal.gdc.cancer.gov/. (Accessed: 1st April 2018).

23. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513,** 202–209 (2014).

24. Haecker, I. & Renne, R. HITS-CLIP and PAR-CLIP advance viral miRNA targetome analysis. *Crit Rev Eukaryot Gene Expr* **24,** 101–116 (2014).

25. miRTarBase 7.0: the experimentally validated microRNA-target interactions database. Available at: http://mirtarbase.mbc.nctu.edu.tw/php/index.php. (Accessed: April 2018).

26. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. **45**, D158-D169 (2017). Available at: http://www.uniprot.org (Accessed: April, 2018).