null. The Bayes factor for comparing these two models is

$$BF(G_0 : G_i) = M \frac{\prod_{j=1}^{p} |\frac{1}{2} X_j' X_j|^{\frac{n}{2}}}{\prod_{j=1}^{p} |\frac{1}{2n} X_j' X_j|^{\frac{1}{2}}} \frac{\prod_{C \in \mathcal{C}} |\frac{1}{2} X_j' X_j|^{\frac{|C|}{2}}}{\prod_{S \in \mathcal{S}} |\frac{1}{2} X_j' X_j|^{\frac{|S|}{2}}} \frac{\prod_{S \in \mathcal{S}} |\frac{1}{2} X_j' X_j|^{\frac{|S|}{2}}}{\prod_{C \in \mathcal{C}} |\frac{1}{2} X_j' X_j|^{\frac{|C|}{2}}}, \tag{14.2}$$

where $\mathcal{C}$ and $\mathcal{S}$ are the cliques and separators of $G_i$, $X_C$ refers to the columns of $X$ corresponding to the nodes in clique $C$, and

$$M = \frac{\prod_{C \in \mathcal{C}} \Gamma_{|C|} \left( \frac{n+|C|-1}{2} \right)}{\prod_{S \in \mathcal{S}} \Gamma_{|S|} \left( \frac{n+|S|-1}{2} \right)} \frac{\prod_{S \in \mathcal{S}} \Gamma_{|S|} \left( \frac{|S|}{2} \right)}{\prod_{C \in \mathcal{C}} \Gamma_{|C|} \left( \frac{|C|}{2} \right)}.$$

By applying Equation 14.2 for each triplet, we can obtain the Bayes factors for comparing models 1–7 with model 0. According to the Jeffreys (1961) scale for interpretating Bayes factors, if the Bayes factor for $BF(G_i : G_0)$ is greater than 3, there is substantial evidence to support model $i$. Therefore, we sort $BF(G_1 : G_0)$ to $BF(G_7 : G_0)$ in a decreasing order. The sorted Bayes factors are denoted as $BF_{(1)}$ to $BF_{(7)}$, and their corresponding graphs are denoted as $G_{(1)}$ to $G_{(7)}$. If $BF_{(1)}/BF_{(2)}$ is greater than 3, we conclude that this triplet supports model $G_{(1)}$. Otherwise, we conclude that there is not enough evidence to suggest which of the models this triplet supports.

Given the model that a triplet supports, we can calculate the maximum likelihood estimates for the strength of the edges (the conditional correlations). For the three-way model, causal model, inverse model, and zero effect model, if the strength of the edge between $G$ and $M$ is positive for a triplet, we filter out this triplet, because that would contradict the biological fact that microRNAs are normally negatively associated with mRNAs.

## 14.4 Application Data Example

The data sets used in our analysis include patient clinical features, gene expression profiles, and microRNA expression profiles for GBM patients. We first give an overview of the basic characteristics of these data sets and the preprocessing procedures conducted before the analysis.

### *14.4.1 Clinical Characteristics*

In the TCGA GBM study, there are 454 patients for whom clinical information (e.g., age at diagnosis and sex) is available. Column 1 in Table 14.1 gives a brief summary of these patients' clinical characteristics. The overall survival time after diagnosis with GBM is the variable in which patients are most interested.

Table 14.1 *Clinical characteristics of GBM patients in the entire patient set and the patient set for GGM analysis (i.e., reduced sample size consisting of samples with data for microRNA, mRNA, and clinical outcome)*

| Characteristics | Whole patient set | Patient set for GGM analysis |
|---|---|---|
| No. of patients | 454 | 280 |
| No. of events | 349 | 248 |
| Age (Median) | 58 | 57 |
| Age ($p$ value from Cox model) | <0.0001 | <0.0001 |
| Age (HR from Cox model) | 1.033 | 1.031 |

To apply the GGM procedure, we need to perform the following preliminary transformation:

- To obtain an estimated survival time for each patient, we first fit a Cox model using the only significant clinical feature in predicting patient overall survival, patient age, as the explanatory variable. We then compute the Breslow estimator of the baseline hazard function.
- For each censored patient $i$, we calculate the $E[\max(T_i^{est}, T_i^{obs})]$, where $T_i^{obs}$ is the observed overall survival time and $T_i^{est}$ is the estimated overall survival time. These values are imputed as the actual overall survival times as if these patients were not censored.
- Log-transformation is performed for all observed survival times and imputed values.

### *14.4.2 microRNA Data Set*

The microRNA data were obtained by the University of North Carolina H-miRNA 8 × 15K array. We downloaded the level 3 microRNA data, which had been through normalization, directly from TCGA. The expression levels of 534 microRNAs were recorded for each patient in the microRNA data set. All 534 microRNAs were considered in our model.

### *14.4.3 mRNA Data Set*

The gene expression data were obtained by using the Affymetrix Human Genome U133A chip. We downloaded level 2 mRNA expression data, which had been through initial processing. There are 280 patient samples with all three types of information available. We only used these samples in the following

GGM analysis (see Table 14.1). We first normalized the mRNA data globally using the BrainArray CDF and RMA normalization method. After normalization, there are 12,126 measurements for each patient, and each measurement only corresponds to one gene. Because low-expressing genes are subject to much greater random measurement error, genes with uniformly low expression are discarded according to the following steps.

- Divide the 280 patients into a short survival group and a long survival group using 2 years as the cutoff point, which is clinically meaningful according to Stupp et al. (2005). The mean expression level for each gene is calculated for the two groups, respectively.
- A gene is considered underexpressed if the mean normalized expressions for both groups are less than 5.

There are 7,785 genes left after this screening step. Next, we select the top 1,000 genes most relevant to patient survival time after adjustment for patient age. This is done by fitting 7785 Cox models with age and each gene expression as predictors.

In summary, the data we used in the GGM analyses include 280 patients with their overall survival time, 534 microRNA expressions, and 1,000 gene expressions, resulting in $534 \times 1{,}000$ triplets. We are interested in which of the eight graphical models in Figure 14.2 best represent the relationships among microRNA expression, gene expression, and patient survival time in each triplet.

### *14.4.4 Analysis Results*

We categorize all the triplets into eight groups according to which of the eight models their Bayes factors indicate. The gray bars in Figure 14.3 summarize the number of triplets in each group.

The maximum likelihood estimates are computed for the strength of the edges (the conditional correlations) for each triplet given the graphical model that its Bayes factor indicates. If the strength of the edge between $G$ and $M$ is positive for a triplet, we filter out this triplet because it contradicts with the biological fact that microRNAs normally repress the expression of their target genes. Triplets that pass this filter for each graphical model are summarized in the black bars in Figure 14.3.

We calculate how many different microRNAs are involved in the biologically meaningful triplets in each group, and for each microRNA, we count the number of triplets with this microRNA. The results are shown in Figures 14.4 and 14.5 for the three-way model, the independent model, the causal model, and the microRNA model, respectively. In these figures, the x-axes show the different

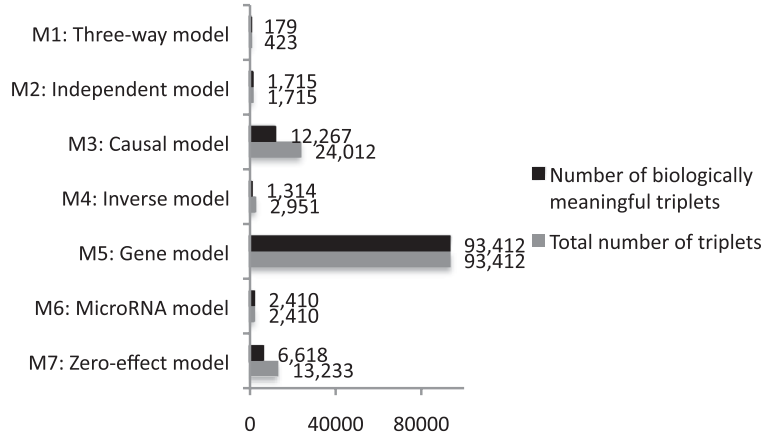


Figure 14.3 Number of triplets in each group categorized by which of the models their Bayes factors indicate.

microRNAs sorted according to the number of triplets they are involved in. The y-axes indicate the exact number of triplets in which each microRNA appears.

There are 179 clinically meaningful triplets supporting the three-way model. In the 179 triplets, there are 42 different microRNAs. Of the 42 microRNAs, 3 mircoRNAs – hsa-mir-148a, hsa-mir-221, and hsa-mir-222 – are from the 10-microRNA list derived by Srinivasan et al. (2011) that can predict survival in GBM.

We also plot the graphical model estimation for the eight triplets with the greatest Bayes factor for the three-way model, independent model, causal model, and microRNA model compared with the null model, respectively (see Figures 14.6 and 14.7). The solid edges in the figures indicate negative associations, and dashed edges in the figures indicate positive associations. The width of the edges shows the strength of the associations. Many microRNAs from the 10-microRNA list are shown in these graphs. For example, the top panel of Figure 14.6 shows that hsa-mir-148a negatively affects patient survival by modulating different genes such as ZEB1 and WAC. Furthermore, it also affects patient survival negatively, which means that a higher expression of hsa-mir-148a suggests a shorter progression-free survival time. This result is the same as the conclusion drawn by Srinivasan et al. (2011), who identified hsa-mir-148a as a risky microRNA with a hazard ratio greater than 1. Similarly, we found that, as in Srinivasan et al. (2011), the hsa-mir-221 and hsa-mir-222 are also risky microRNAs. On one hand, these microRNAs negatively influence patient survival by modulating the expression of their target genes, and on the other hand, they also directly affect patient survival negatively. In the bottom

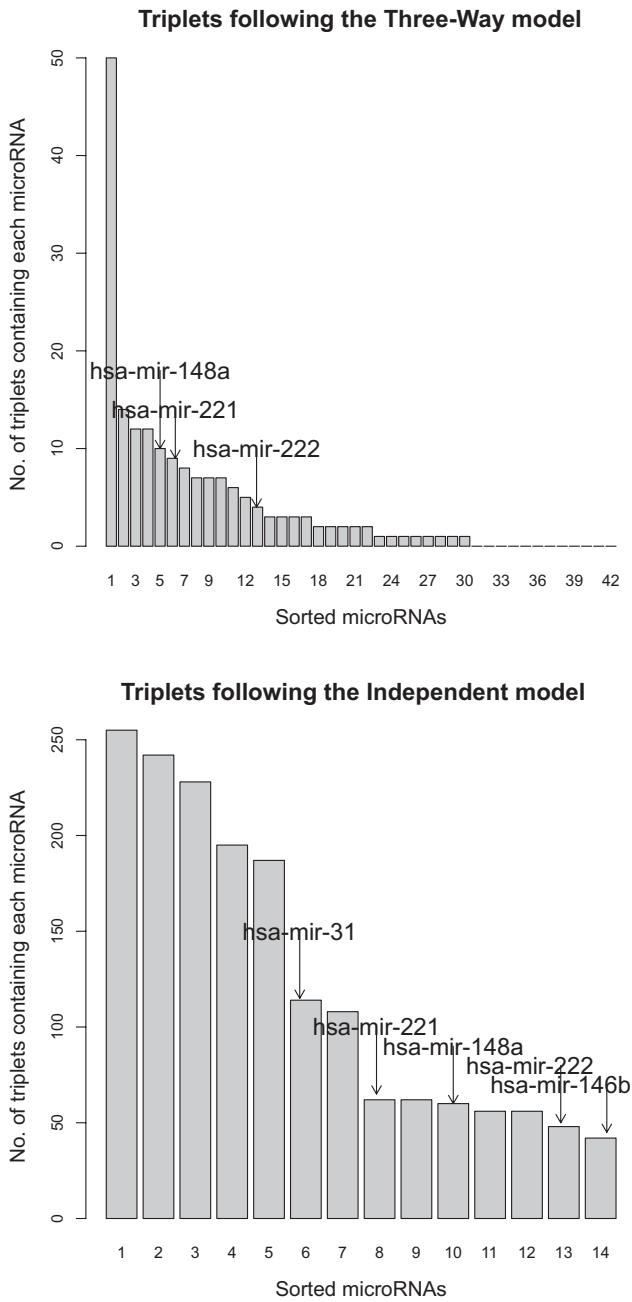


Figure 14.4 Top panel: triplets supporting the three-way model (42 different microRNAs are sorted according to number of different triplets); bottom panel: triplets supporting the independent model (14 different microRNAs are sorted according to number of different triplets).

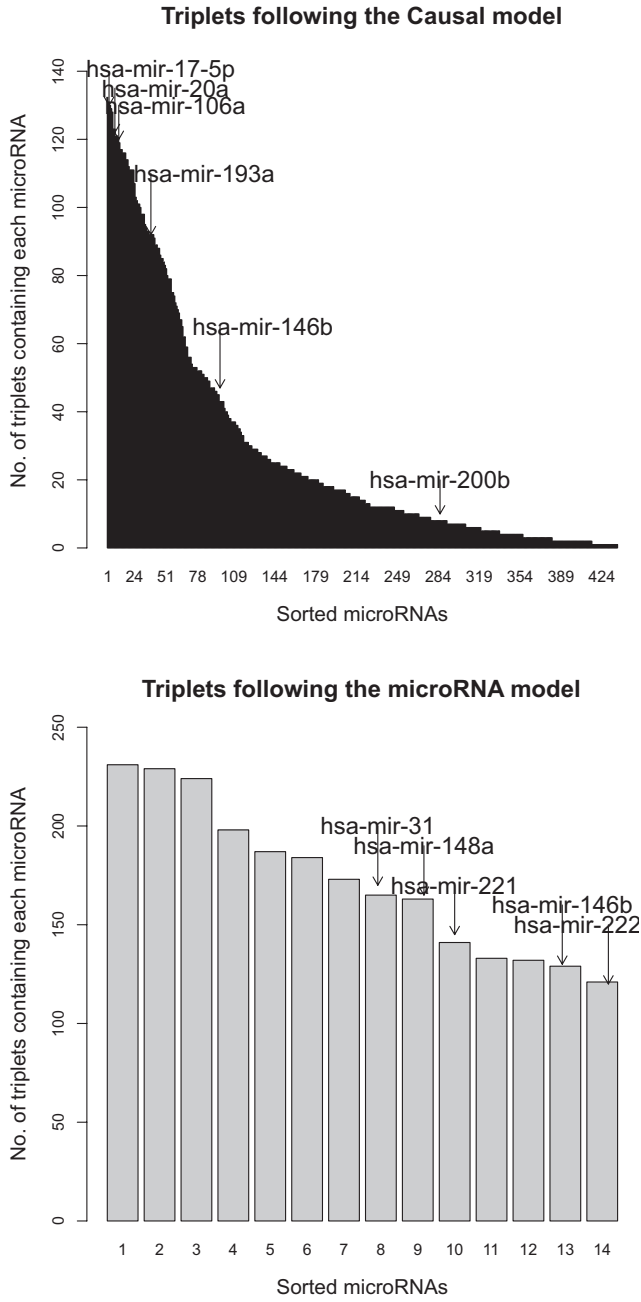


Figure 14.5 Top panel: triplets supporting the causal model (458 different microRNAs are sorted according to number of different triplets); bottom panel: triplets supporting the microRNA model (14 different microRNAs are sorted according to number of different triplets).

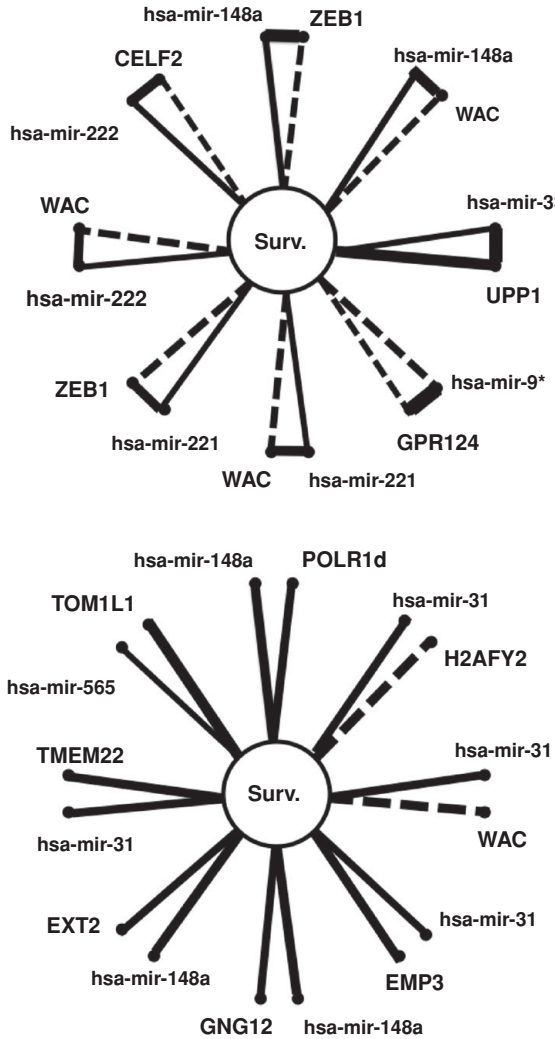


Figure 14.6 Triplets with greatest BFs supporting the three-way model (top panel) and independent model (bottom panel) compared with the null model (solid edge: negative association; dashed edge: positive association; the width of an edge: the strength of the association).

panel of Figure 14.6, genes and microRNAs affect patient survival independently. As shown in Srinivasan et al. (2011), hsa-mir-148a and hsa-mir-31 are risky microRNAs.

The top panel of Figure 14.7 shows the microRNAs and genes that are consistent with fundamental biological mechanisms. Because the microRNAs

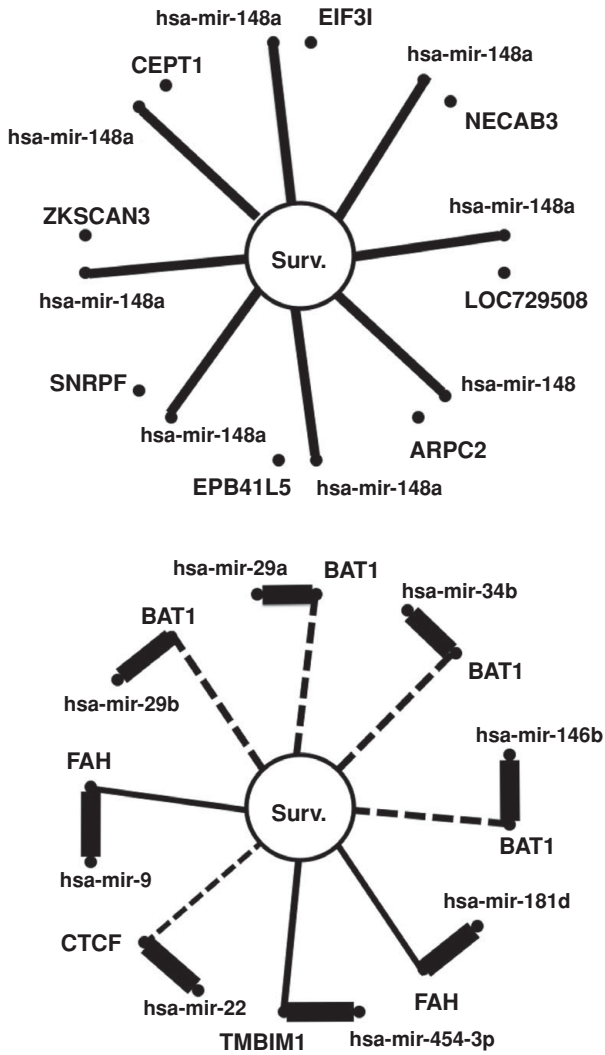


Figure 14.7 Triplets with greatest BFs supporting the causal model (top panel) and microRNA model (bottom panel) compared with the null model (solid edge: negative association; dashed edge: positive association; the width of an edge: the strength of the association).

in this figure are not directly related to survival, many of them can be missed if the analysis is done by using only microRNA data (Srinivasan et al., 2011). In these microRNAs, hsa-mir-29a, hsa-mir-34b, hsa-mir-146b, hsa-mir-22, and hsa-mir-29b are risky microRNAs, and hsa-mir-181d, hsa-mir-454-3p, and hsa-mir-9 are protective microRNAs. The bottom panel of Figure 14.7 shows that