Filtering of Gene Expression Values

1. All data are first downloaded from: http://www-genome.wi.mit.edu/cancer/pub/glioma .
2. “Brain_Classics.res” and “Brain_NonClassics.res” files are combined to get the gene expression levels of all samples (=50).
3. Then the values are preprocessed by using a threshold of 20 and a ceiling of 16000. If a value is less than 20, it is replaced by 20; similarly, if a value is greater than 16000, it is replaced by 16000.
4. Then those genes are excluded which violate *max(g)-min(g)>100* and *max(g)/min(g)>3*, leaving a total of 4434 genes.

Normalization Method

1. We linearly scale all gene expression values in the range [0, 1].
2. Suppose *x* is a gene expression value of a gene *g*, the scaled value would be: $\frac{x-\min(g)}{\max(g)-\min(g)}$ where *min(g)* and *max(g)* are the minimum and maximum value of gene expressions of *g* among different samples.
3. If you want to linearly scale *x* in the range [a, b], use the following formula: $(b-a)\frac{x-\min(g)}{\max(g)-\min(g)}+a$
4. If you want to transform *x* to *N(0,1)*, use the following formula:

$\frac{x-\mu}{\sigma}$ where μ and σ are the mean and standard deviation of *g* across all samples.