expression dataset [1] contains 4026 genes and 96 conditions. The values in the Lymphoma dataset are integers in the range of -750 to 650. The preprocessed datasets are obtained from http://arep.med.harvard.edu/biclustering.

3.2Bicluster Plots for Yeast Dataset

Figure 1 shows eight biclusters obtained by KMeans-Binary PSO hybrid algorithm on Yeast dataset. Some of the biclusters contain all 17 conditions. All the biclusters show strikingly similar up-regulation and down regulation.

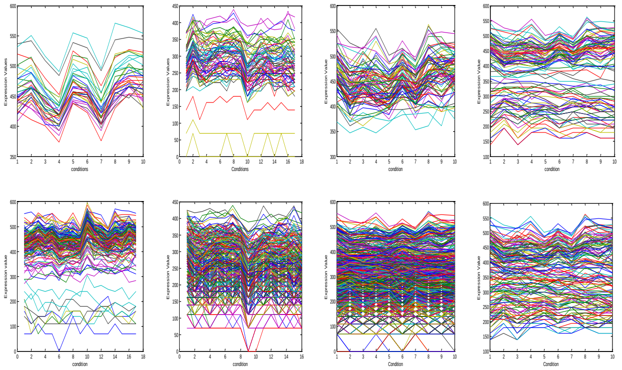


Figure 1. Eight biclusters found for the Yeast dataset.

Bicluster labels are (a), (b), (c), (d), (e), (f), (g) and (h) respectively. The details can be obtained from Table 1 using bicluster label.

Table 1. Information about biclusters from Yeast Dataset

Label	Rows	Columns	Volume	MSR
(a)	32	10	320	63.0642
(b)	75	17	1275	199.5888
(c)	80	10	800	190.3379
(d)	100	10	1000	298.6600
(e)	136	17	2312	297.9888
(f)	323	16	5168	286.1017
(g)	1030	10	1030	299.9427
(h)	150	10	1500	298.8481
(i)	882	11	9702	299.7275
(j)	1399	8	11192	299.9149
(k)	656	12	7872	299.8829
(l)	848	11	9328	299.8653
(m)	145	17	2465	299.6139
(n)	318	16	5088	281.5787

In Table 1 given above the first column reports the label of each bicluster, the second column contains the number of rows (genes), third the number of columns (conditions), fourth column contains the volume or size of the bicluster and the last column reports the mean squared residue score. Table 1 contains the details of some more biclusters which are not shown in Figure 1. The labels of these biclusters are (i), (j), (k), (l), (m) and (n).

3.3Bicluster Plots for Human Lymphoma Dataset

In Figure 2 eight biclusters obtained from Human Lymphoma dataset using KMeans-Binary PSO hybrid algorithm are shown. The algorithm is better for identifying more genes than conditions where as some other metaheuristic methods like GRASP can identify more number of conditions. The maximum number of conditions obtained here is only 27. The maximum number of genes obtained is 1180.

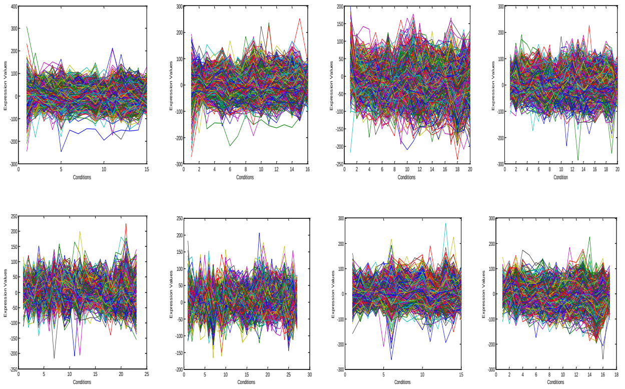


Figure 2. Eight biclusters found for the Lymphoma dataset.

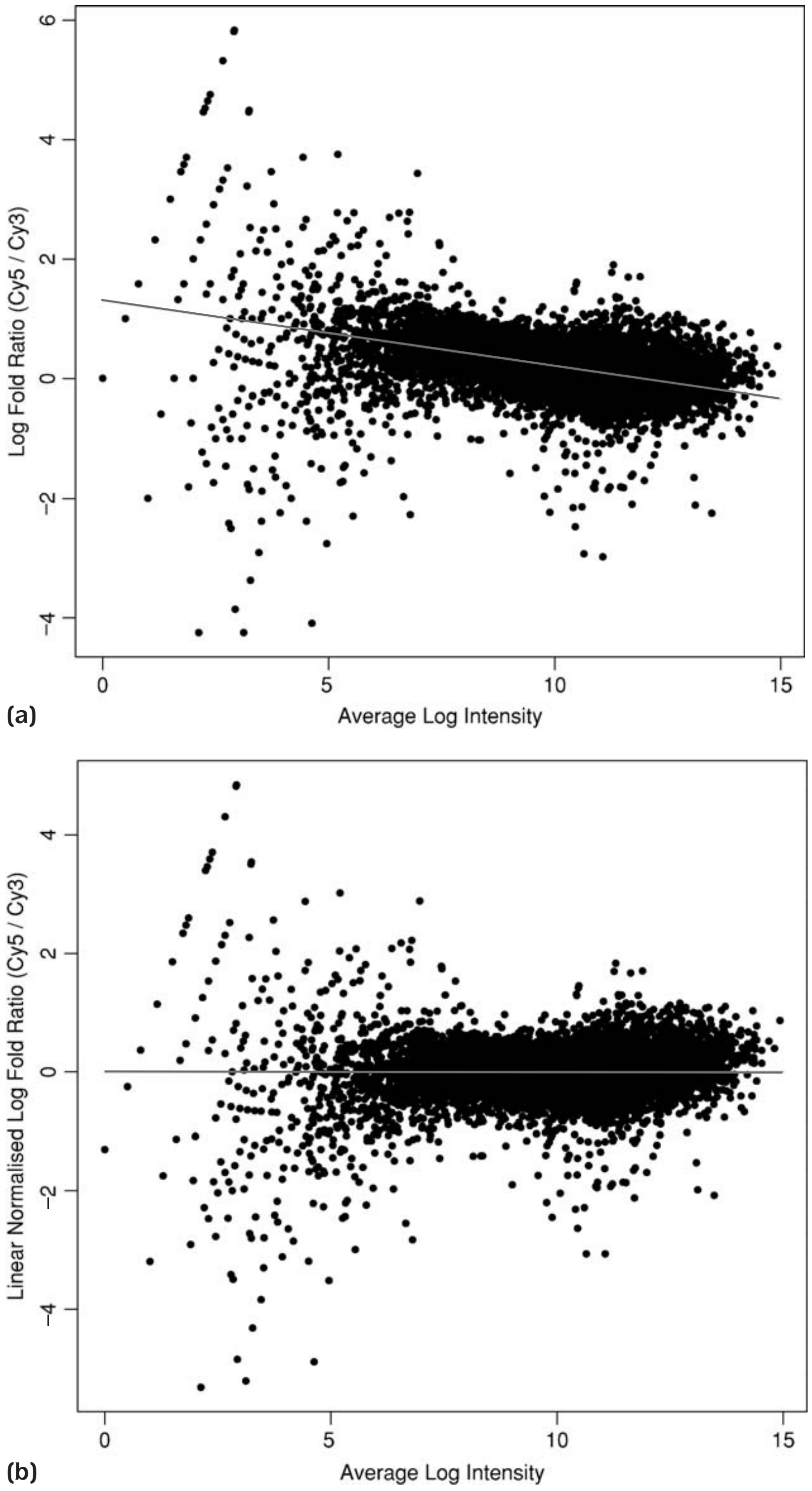
Bicluster labels are (p), (q), (r), (s), (t), (u), (v) and (w) respectively. The details can be obtained from Table 2 using bicluster label.

Table 2. Information about biclusters of Figure 2

Label	Rows	Columns	Volume	MSR
(p)	1180	15	17700	1198.8
(q)	1060	16	16960	1192.2
(r)	747	20	14940	1198.0
(s)	974	20	19480	1199.7
(t)	505	23	11615	1199.9
(u)	339	27	9153	1199.4
(v)	967	15	14505	1197.7
(w)	836	17	14212	1199.0

4.COMPARISON

The performance of KMeans-Binary PSO hybrid compared with that of SEBI [5], Cheng and Church's algorithm (CC), and the algorithm FLOC by Yang et al. [11], DBF [13] and single objective GA [3] for the Yeast dataset are given in Table 3. In Sequential Evolutionary Biclustering (SEBI) biclusters are identified using Evolutionary Computation. Cheng and Church used greedy method for finding biclusters. Yang et al. generalized the model of biclusters proposed by Cheng and Church and developed a probabilistic algorithm called FLOC to discover overlapping biclusters simultaneously. Zhang et al. [13] developed deterministic biclustering with frequent pattern mining (DBF). DBF generates biclusters in two steps. In the first step high quality seeds are generated using frequent pattern mining. These seeds are then enlarged by adding more genes and conditions. Single objective GA [3] has been used with local search to generate overlapped biclusters. In terms of average gene number, average volume and largest bicluster size KMeans-Binary


Log Fold Ratio (Cy5 / Cy3)
6
4
2
0
−2
−4
0
5
10
15
(a)
Average Log Intensity
Linear Normalised Log Fold Ratio (Cy5 / Cy3)
4
2
0
−2
−4
0
5
10
15
(b)
Average Log Intensity

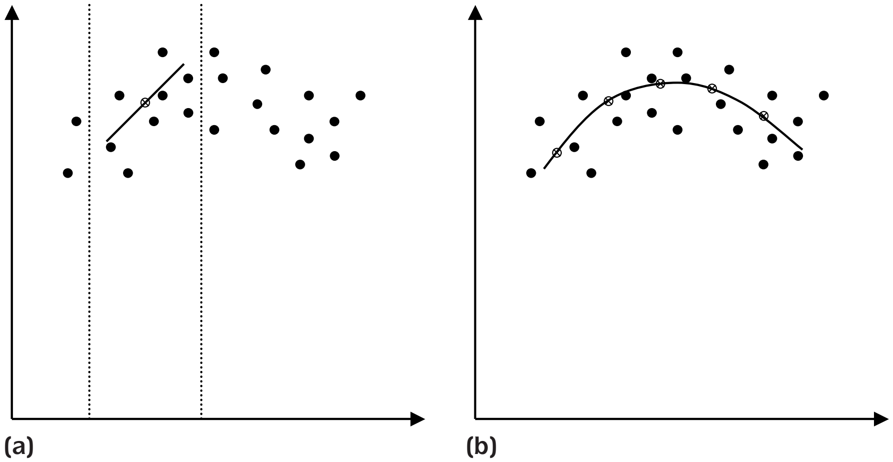


Figure 5.4: Loess normalisation. (a) Loess regression works by performing a large number of local regressions in overlapping windows across the whole range of the data set. The regression curve is usually either a straight line or a quadratic curve. (The default R implementation is a quadratic curve.) Each regression results in a central point and regression line or curve about that point. **(b)** The points and curves from the local regressions are combined to form a smooth curve across the length of the data set.

3. Apply the Loess regression to your data.
4. For each feature, calculate the normalised log ratio by subtracting the fitted value on the Loess regression from the raw log ratio.

EXAMPLE 5.5 NON-LINEAR REGRESSION APPLIED TO DATA SET 5A

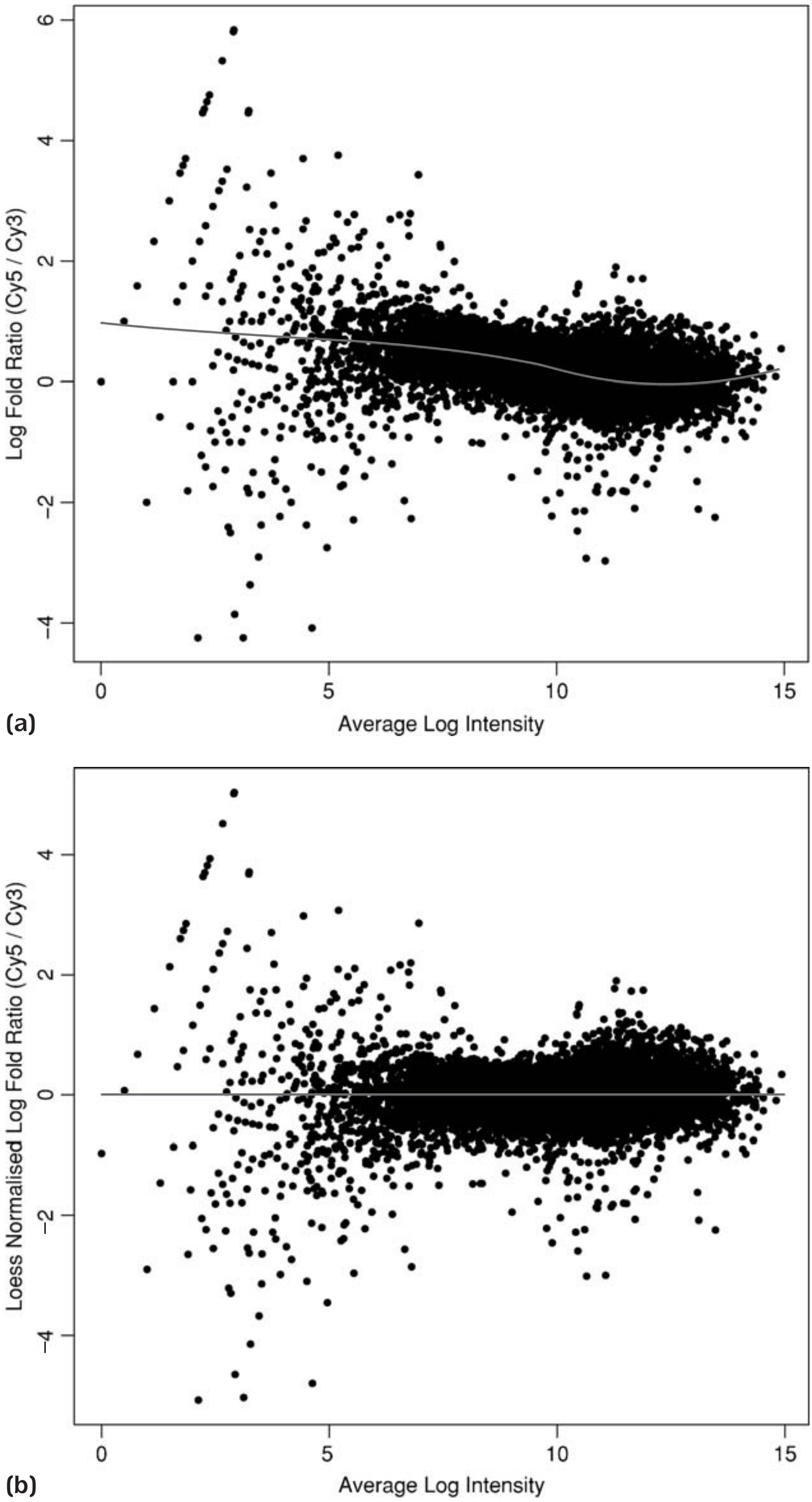
Loess regression is applied to the human fibroblast data set 5A (Figure 5.5a).[7] The curve fits the data very well. The normalised data (Figure 5.5b) are balanced about zero and are ready for analysis for differentially expressed genes.

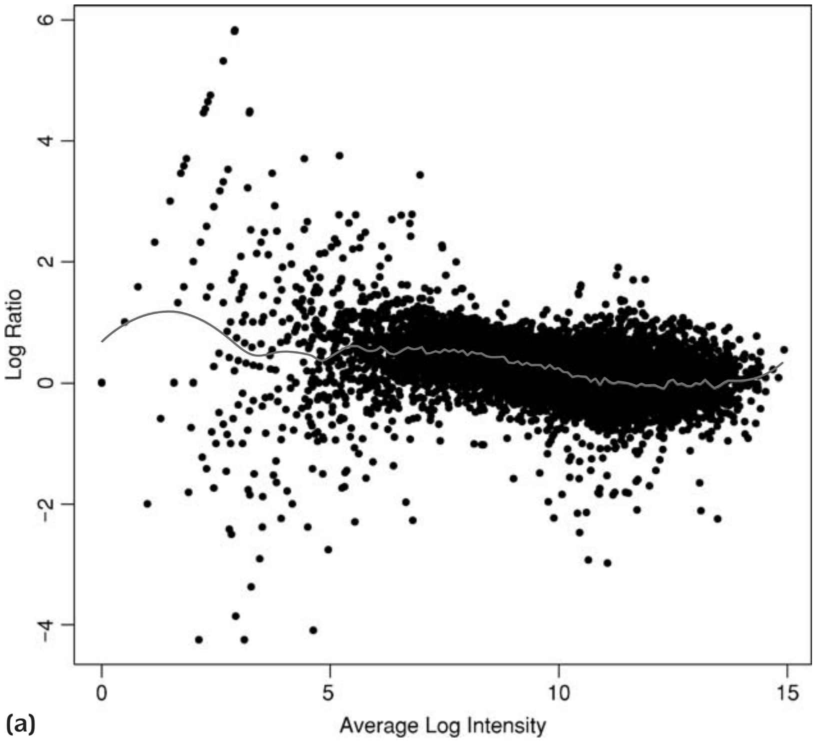
Although Loess is an advanced statistical technique, it is important to remember that it is no more than a computational method for drawing a best-fit curve through a cloud of points. There is no conceptual or theoretical underpinning to the curve produced by Loess; it is only a scaling of the data.

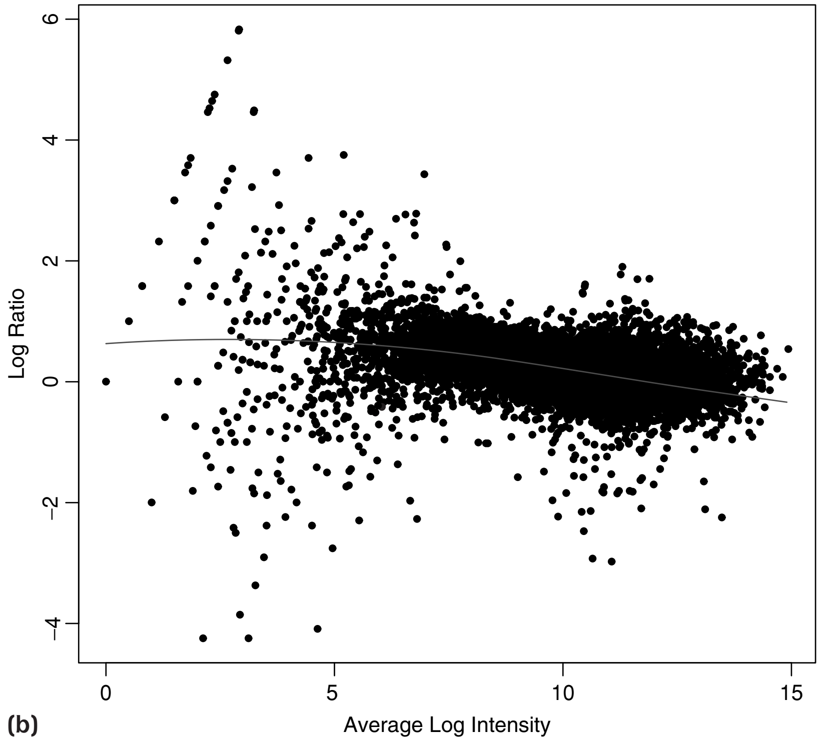
Loess regression has a number of parameters, which can be set by the user, whose values will have an impact on the way in which the curve fits the data. The most important of these is the size of the window, which determines the smoothness of the regression. If the window is too small, the curve will be too sensitive to local ups and

[7] The R statistical software is very well equipped to perform Loess regression. It can be found in the *modreg* package. Suppose the fibroblast data set were in a data frame called *fibroblast*, with variables *average* and *lratio* containing the average log intensity and log ratio. Then the R commands to apply Loess normalisation would be:

```
attach(fibroblast)
lmodel <- loess(lratio∼average)
fibroblast$normlratio <- lratio - predict.loess(lmodel,
       average)
```


6
4
2
0
−2
−4
Log Fold Ratio (Cy5 / Cy3)
0
5
10
15
Average Log Intensity
(a)
4
2
0
−2
−4
Loess Normalised Log Fold Ratio (Cy5 / Cy3)
0
5
10
15
Average Log Intensity
(b)


Log Ratio
6
4
2
0
−2
−4
0
5
10
15
(a)
Average Log Intensity


Log Ratio
6
4
2
0
−2
−4
0
5
10
15
(b)
Average Log Intensity

Second, it is common for spatial bias to arise from the array not being horizontal in the scanner, which may have no relationship to the variabilities between different pins.

SECTION 5.4 BETWEEN-ARRAY NORMALISATION

Section 5.3 described normalisation methods that can be used to compare the Cy3 and Cy5 channels of a single array. This section looks at normalisation methods that allow you to make comparisons between samples hybridised to different arrays, which could be either two-colour arrays or Affymetrix arrays. In such experiments, each hybridisation reaction may be slightly different, and so the overall intensities of

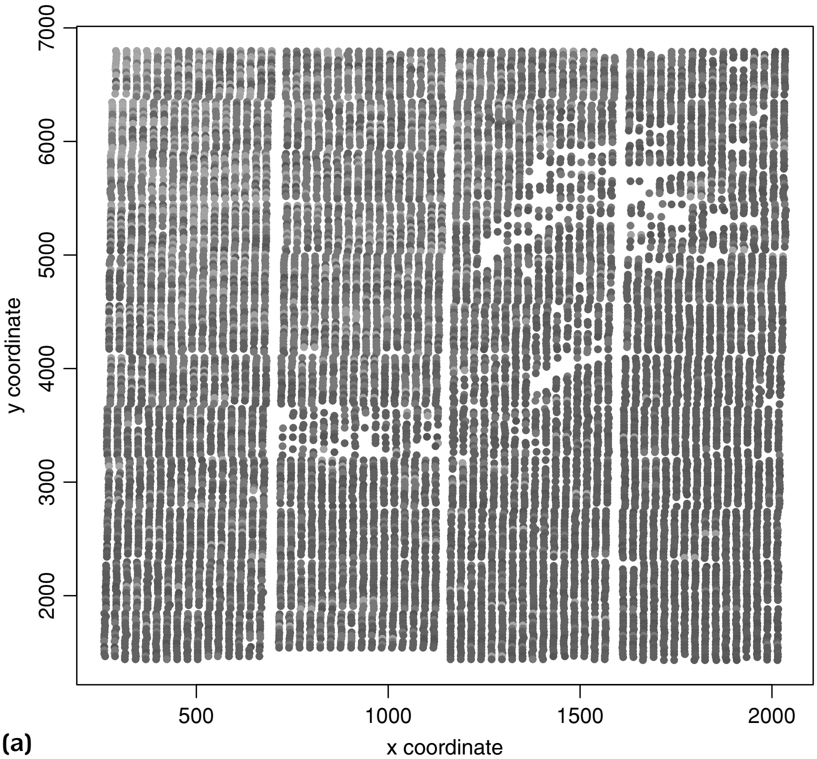


Figure 5.7: Spatial bias on a microarray and two-dimensional Loess regression. (a) False-colour representation of the log ratios of a microarray, with mouse kidney in Cy3 and liver from the same mouse in Cy5 (data set 5B). Each spot represents a feature. The *x* and *y* coordinates of each spot correspond to the *x* and *y* coordinates of the feature on the array. The colour of the spot represents the log ratio (Cy5/Cy3) of the feature, with red spots having a positive log ratio and green spots having a negative log ratio. There is a strong spatial bias on the array, with green spots in the top-left-hand corner and red spots in the bottom-right-hand corner. The areas of the array with missing spots represent features that have been flagged by the image-processing software, or features with a higher background than signal that have been removed from the data set. **(b)** The same data, but with the fit of a two-dimensional Loess surface to the log ratios superimposed as contours. The contours follow the colour trend, going from negative at top left to positive at bottom right. **(c)** False-colour plot of the normalised log ratio values of the features. These are calculated by subtracting the fitted values of the Loess surface from the raw log ratios. There is no spatial bias on the normalised data. (*Please see also the color section at the middle of the book.*)

(*continued*)

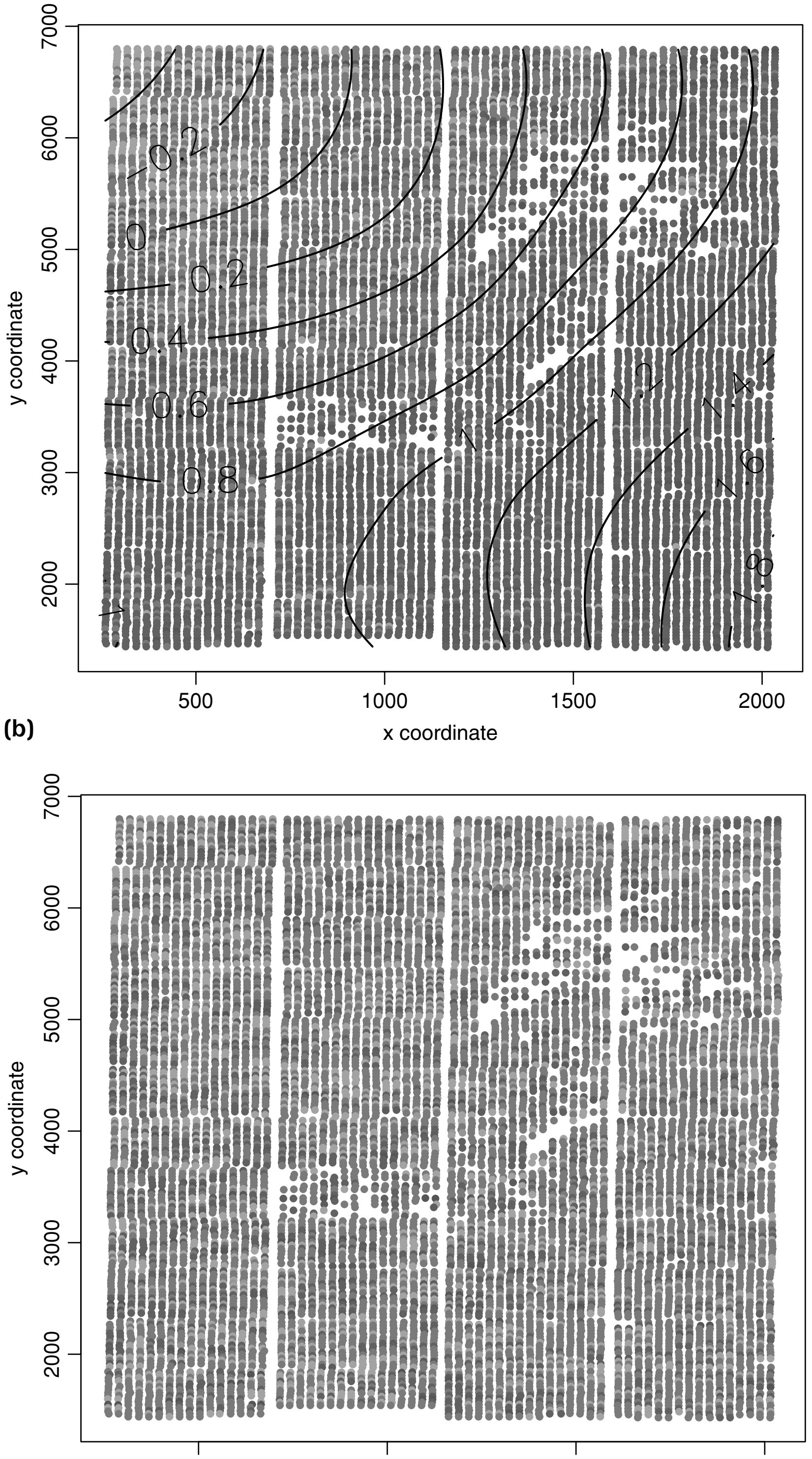
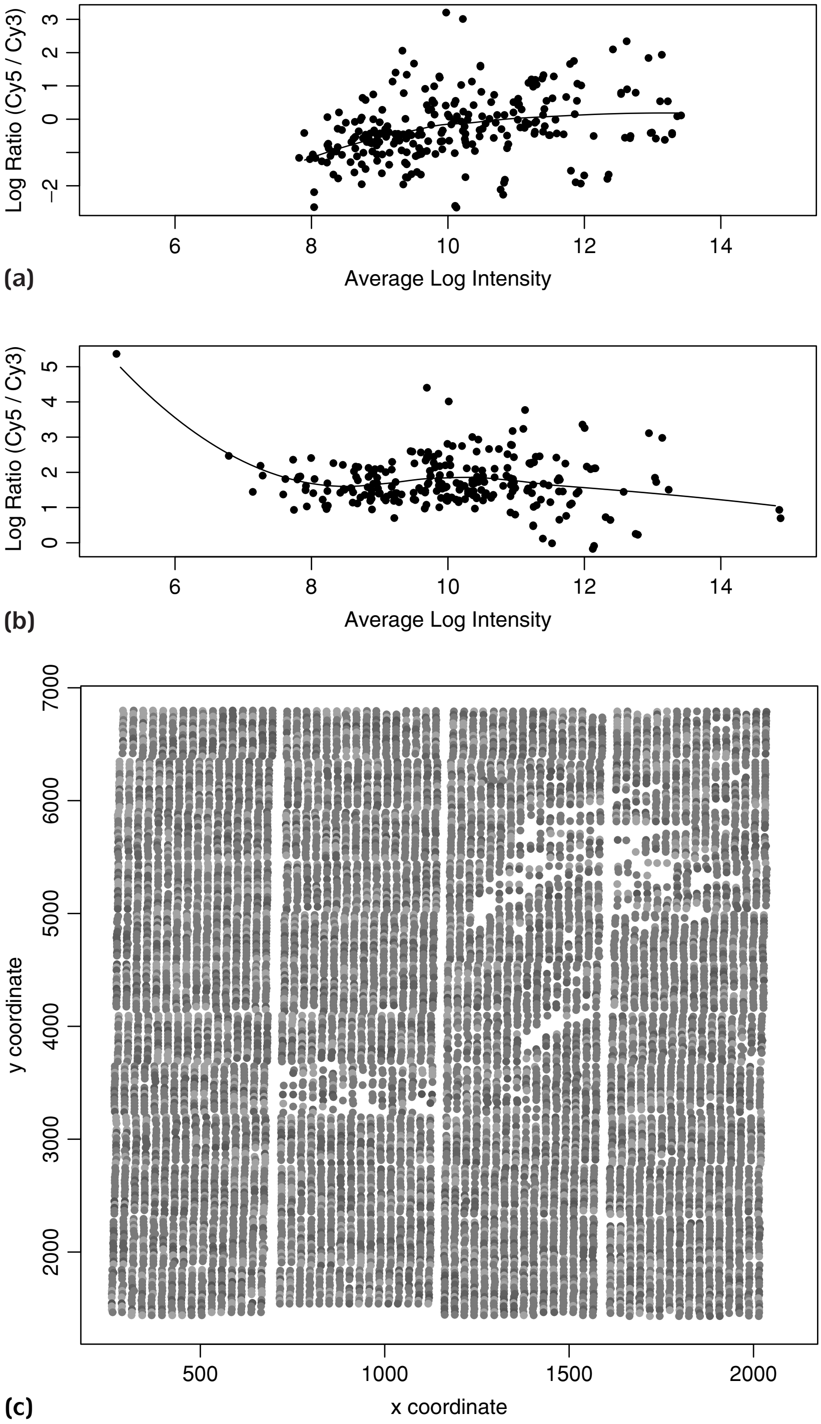


Figure 5.7: (*continued*)


Log Ratio (Cy5 / Cy3)
3
2
1
0
−2
6
8
10
12
14
(a)
Average Log Intensity
Log Ratio (Cy5 / Cy3)
5
4
3
2
1
0
6
8
10
12
14
(b)
Average Log Intensity
y coordinate
7000
6000
5000
4000
3000
2000
500
1000
1500
2000
(c)
x coordinate

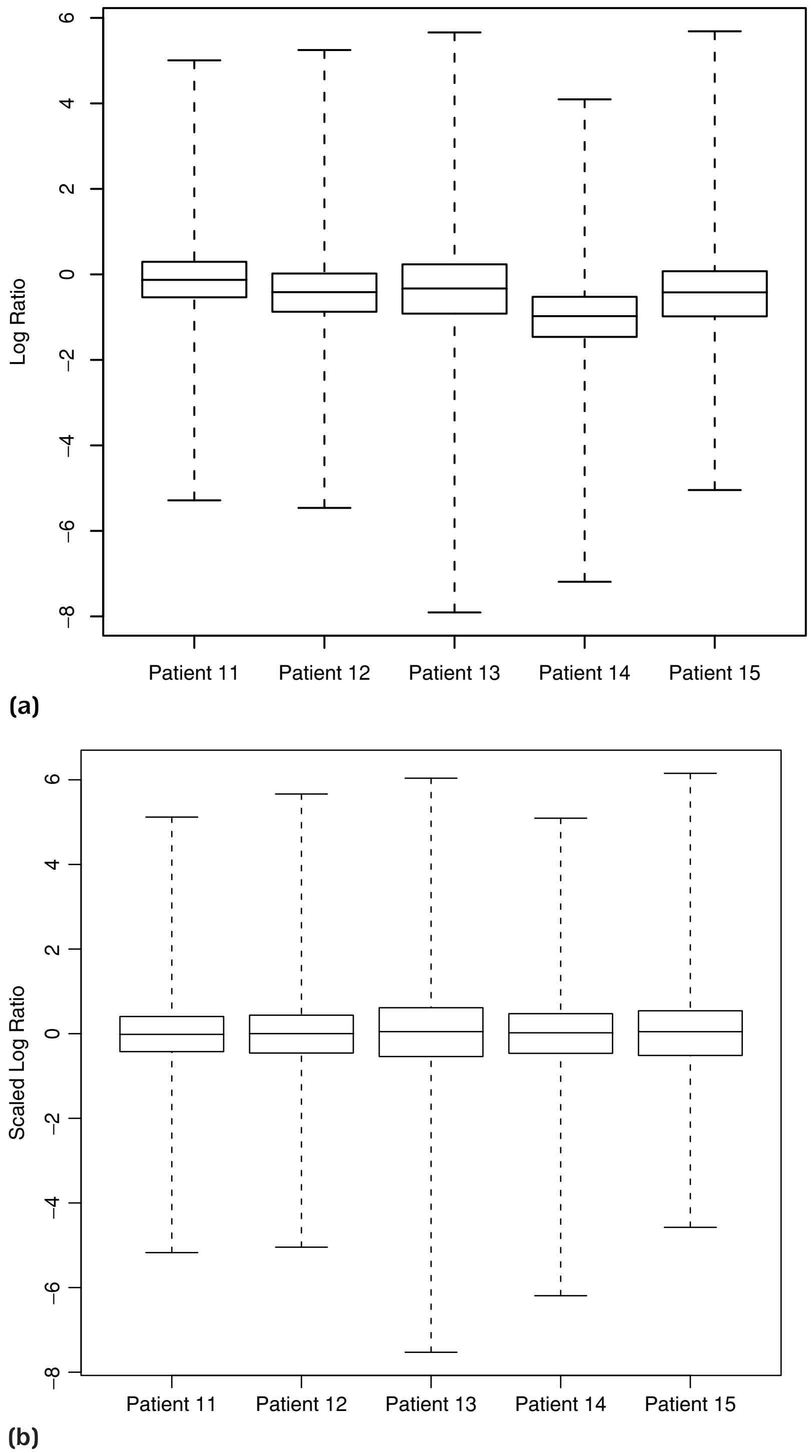


Figure 5.9: Scaling, centering and distribution normalisation. Different methods that allow the comparison of samples on many arrays for data analyses such as cluster analyses (Chapter 7) and classification analyses (Chapter 8). The data in this figure are five patients suffering from diffuse large B-cell lymphomas (data set 5C). **(a)** Box plot of the raw log ratios of the five patients. The distribution of log ratios for all patients is shown on one plot so they can be easily compared. The line at the center of each box represents the mean (or median) value of the distribution; the size of the box represents the standard deviations (or median absolute deviation from the median) of the distribution; the two horizontal lines bracketing the box (sometimes called whiskers) represent the extreme values of the distribution.

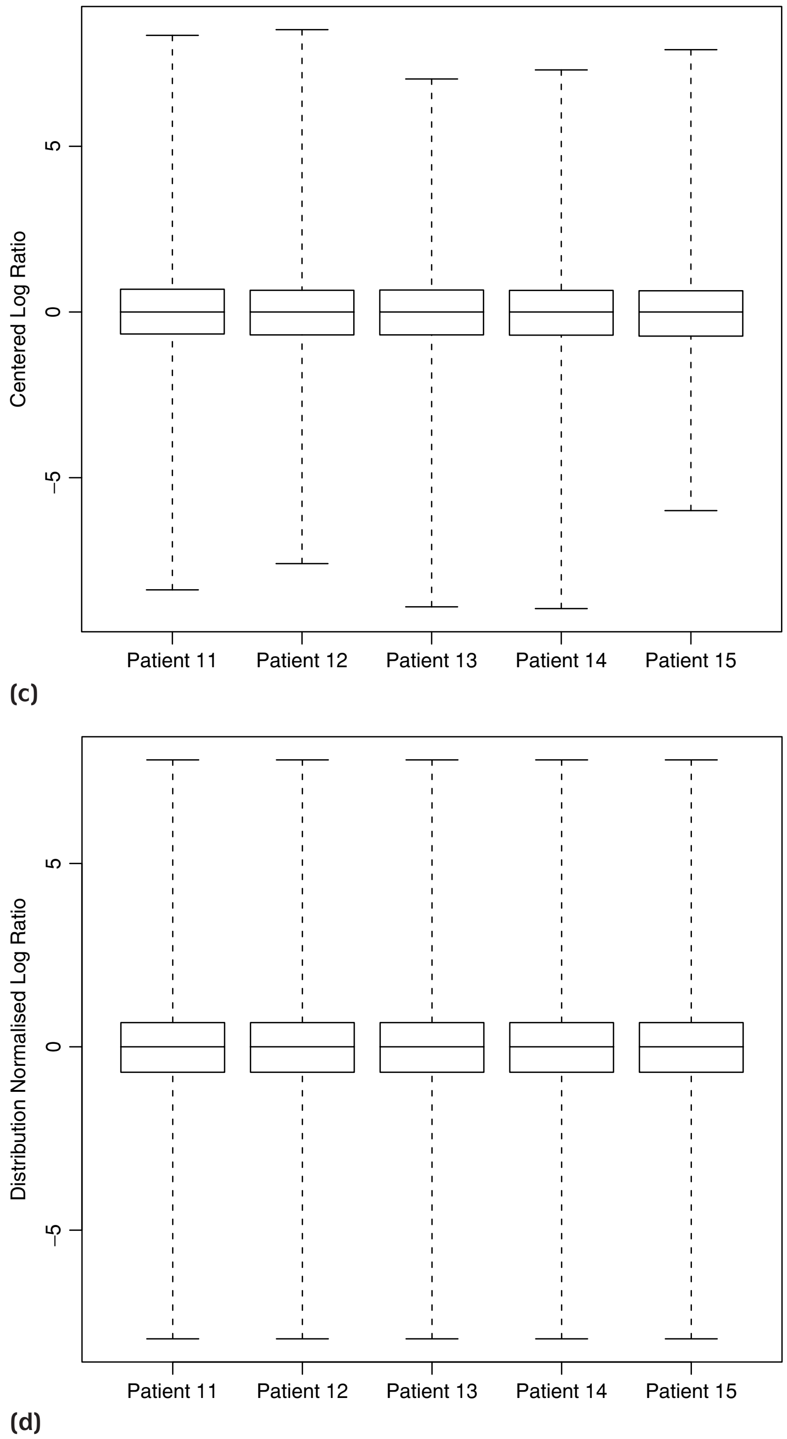
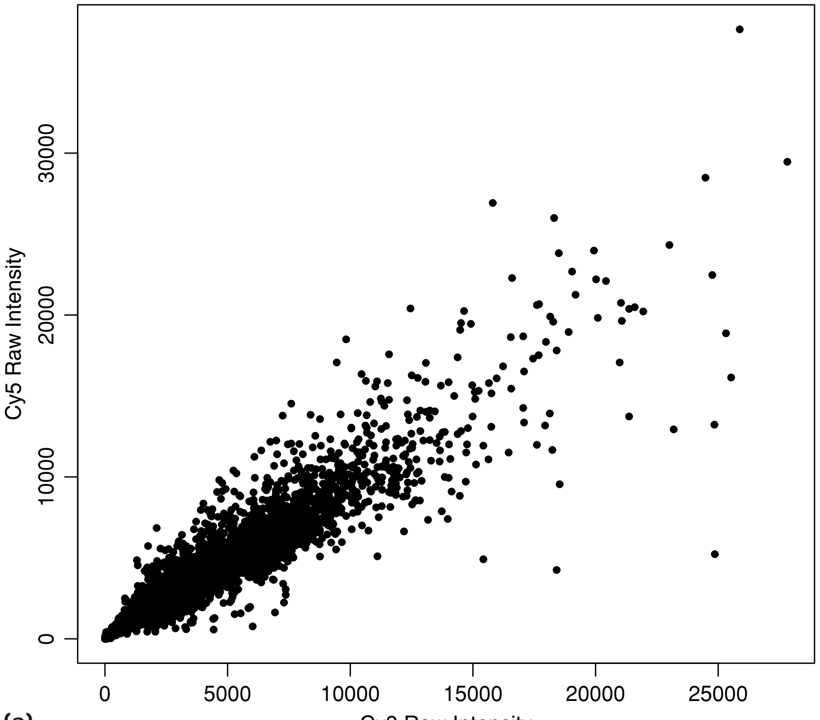


Figure 5.9: (*continued*) In this plot, the five patients all have different means, standard deviations and distributions. **(b)** The data has been scaled by subtracting the mean of the distribution from each of the log ratio values of each patient. The means of the distributions are all equal to zero. **(c)** The data has been centered by subtracting the mean of the distribution and dividing by the standard deviation. The centered distributions for each patient have mean 0 and standard deviation 1. Centering is useful when using correlation as a distance measure (Chapter 7). **(d)** The data has been distribution normalised so that each patient has the same set of measurement values; the distributions for all five patients are identical.



(a)

Figure 5.1: Plots of Cy3 vs. Cy5 for data set 5A. Human foreskin fibroblasts have been infected with *Toxoplasma gondii* for a period of 1 hour. A sample has been prepared, labelled with Cy5 (red), and hybridised to a microarray with approximately 23,000 features. The Cy3 (green) channel is a sample prepared from uninfected fibroblasts. Because the infectious period is short, most genes in this experiment are not differentially expressed. **(a)** Scatterplot of the (background-subtracted) raw intensities; each point on the graph represents a feature on the array, with the *x* coordinate representing the Cy3 intensity, and the *y* coordinate representing the Cy5 intensity. The graph shows two weaknesses of the raw data that would have a negative impact on further data analysis:

1. Most of the data is bunched in the bottom-left-hand corner, with very little data in the majority of the plot.
2. The variability of the data increases with intensity, so that it is very small when the intensity is small and very large when the intensity is large.

(b) Scatterplot of the log (to base 2) intensities. This plot is better than (a). The data is spread evenly across the intensity range, and the variability of the data is the same at most intensities. The genes with log intensity less than 5 have slightly higher variability, but these genes are very low expressed and are below the detection level of microarray technology.

The straight line is a linear regression through the data. The linear regression is not perfect (the data appears to bend upwards away from the line at high intensities), but is approximately right. The intercept is 1.4, and the gradient is 0.88. If the two channels were behaving identically, the intercept would be 0 and the gradient would be 1. We conclude that the two Cy dyes behave differently at different intensities; this could result from differential dye incorporation or different responses of the dyes to the lasers.

(*continued*)

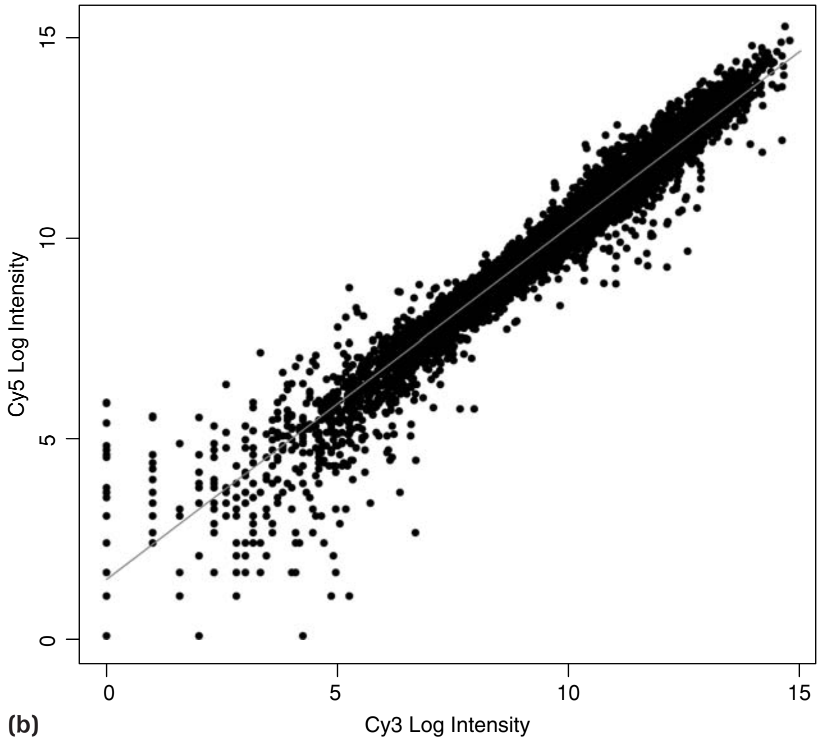


Figure 5.1: (*continued*)

EXAMPLE 5.1 TAKING THE LOG OF *TOXOPLASMA GONDII* DATA (DATA SET 5A)

Fibroblasts taken from human foreskin have been infected with *Toxoplasma gondii*. Samples from uninfected cells and cells treated with *T. gondii* for 1 hour are hybridised to two channels of a microarray with approximately 23,000 features.[5] The researchers want to identify genes that have been differentially expressed.

The raw data (Figure 5.1a) do not satisfy the requirements for effective analysis. Most of the features are in the bottom-left part of the graph; the variability increases with intensity, and the distribution of the intensities is not bell-shaped but very heavily right-skewed (Figures 5.2a and 5.2b).

The logged data (Figure 5.1b), however, do satisfy the requirements. The data are well spread across the range of log intensity values the variability is approximately constant at all intensities and would appear to be normally distributed (with the exception of very low expressed genes, whose intensities are likely to be unreliable); and the distribution of intensities (Figures 5.2c and 5.2d) are closer to being bell-shaped (although these distributions are also slightly right-skewed).

[5] The paper from which this data has been derived is given at the end of the chapter. The data is available from the Stanford Microarray Database.

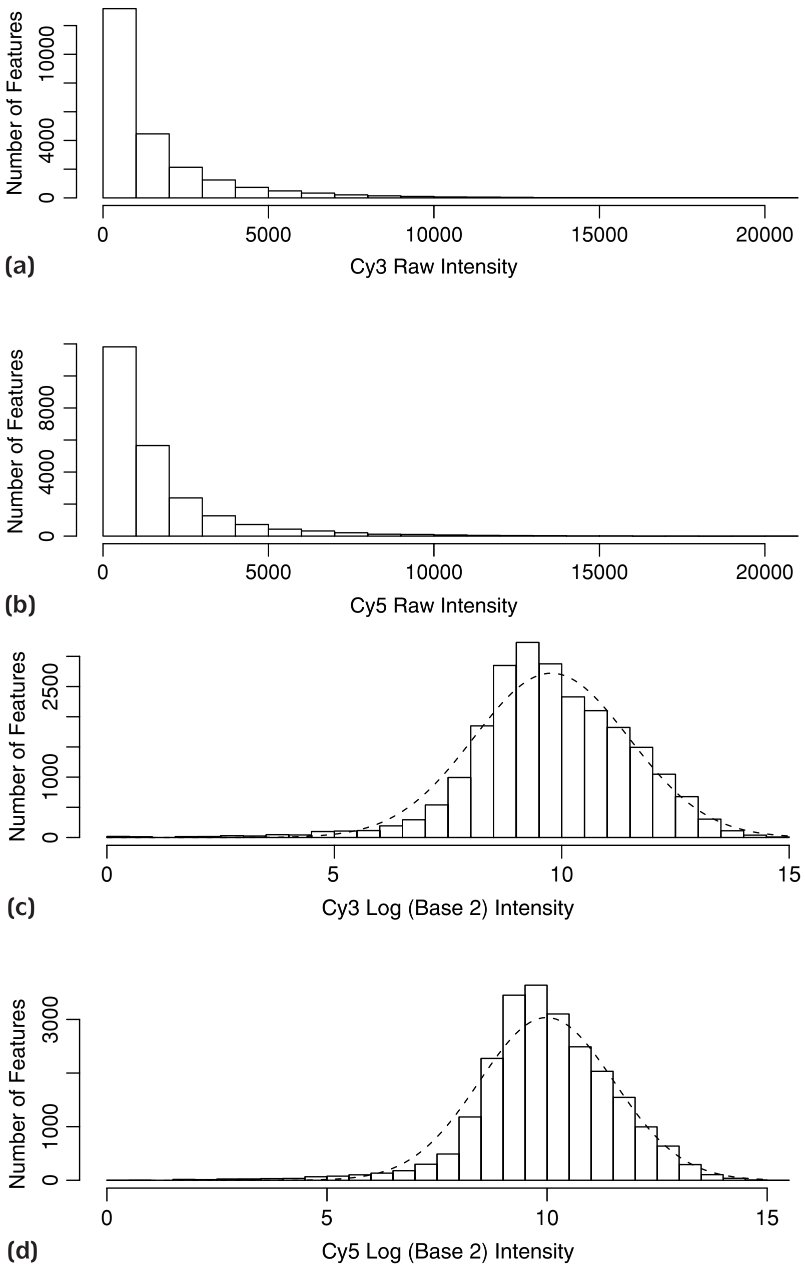


Figure 5.2: Histograms of the raw and log Cy3 and Cy5 intensities. Histograms of the intensities of the features for the human fibroblast data. **(a)** The raw intensities for the Cy3 channel; the data is right-skewed, with the majority of features having low intensity and decreasing numbers of features having higher intensity. **(b)** The raw intensities for the Cy5 channel; the pattern is the same as (a). **(c)** The log intensities for the Cy3 channel; the intensities are closer to a bell-shaped normal curve (shown as a dashed line). There is still a slight right skew, but the logged data is better for data analysis than the raw data. **(d)** The log intensities for the Cy5 channel, along with a normal curve (dashed line). As with (c), the intensities are approximately normal, with a slight right skew.