Figure 10.1. Four different visualizations of the same data, the mean-difference plot of the unprocessed probe intensities from a pair of microarrays. Panel a) shows the usual scatterplot. Because of the large number of points, it is a rather featureless black blot. Panel b) shows the result of a hexagon binning procedure. The color code at each hexagon represents the number of data points that it contains. In panel c) we see a color representation of a smooth density on the \((x, y)\)-plane calculated from the data using a kernel density estimator. In the sparse regions of the density, the plot is augmented by black dots that represent individual data points. In the denser regions, these are omitted. Note that at the boundary between sparse and dense region (as assigned by the algorithm), a visual artifact is created. Panel d) shows the usual scatterplot, however with points colored according to the local density.
10.4 Heatmaps

Heatmaps, or false color images have a reasonably long history, as has the notion of rearranging the columns and rows to show structure in the data. They were applied to microarray data by Eisen et al. (1998) and have become a standard visualization method for this type of data.

A heatmap is a two-dimensional, rectangular, colored grid. It displays data that themselves come in the form of a rectangular matrix. The color of each rectangle is determined by the value of the corresponding entry in the matrix. The rows and columns of the matrix can be rearranged independently. Usually they are reordered so that similar rows are placed next to each other, and the same for columns. Among the orderings that are widely used are those derived from a hierarchical clustering, but many other orderings are possible. If hierarchical clustering is used, then it is customary that the dendrograms are provided as well. In many cases the resulting image has rectangular regions that are relatively homogeneous and hence the graphic can aid in determining which rows (generally the genes) have similar expression values within which subgroups of samples (generally the columns).

The function `heatmap` is an implementation with many options. In particular, users can control the ordering of rows and columns independently from each other. They can use row and column labels of their own choosing or select their own color scheme. They can also add a colored bar to annotate either the row or the column data (e.g., to show the association with some phenotype). And perhaps most importantly they can take the standard R implementation and extend it in any way they would like.

We return to the ALL example to provide a demonstration of a heatmap. We select two small subgroups of patients for this examination, the ALL1/AF4 group and the E2A/PBX1 group.

```r
> library("ALL")
> data("ALL")
> selSamples <- ALL$mol.biol %in% c("ALL1/AF4",
+ "E2A/PBX1")
> ALLs <- ALL[, selSamples]
> ALLs$mol.biol <- factor(ALLs$mol.biol)
> colnames(exprs(ALLs)) <- paste(ALLs$mol.biol,
+ colnames(exprs(ALLs)))
```

There are 15 samples and they are stored in the `exprSet` object `ALLs`. Among the total of 12625 probes in this data set, we select those with mean expression larger than 100 in at least one of the two groups, and a p value of the two-sample t test of less than 0.0002.

```r
> library("genefilter")
> meanThr <- log2(100)
> g <- ALLs$mol.biol
```