Textual description of affydata

September 10, 2009

This is a simple data package. It contains an example data set containing part of the data from a Dilution experiment. The full data is publicly available from Gene Logic `http://qolotus02.genelogic.com/datasets.nsf/`. The help file obtained via `?Dilution` describes the data set.

1 Normalization of Dilution Data

We start by loading the library and the data.

```
> library(affydata)
> data(Dilution)
```

This will create the `Dilution` object of class `AffyBatch`. As described in the `affy` vignette, `print` (or `show`) will display summary information. These objects represent data from one experiment. The `AffyBatch` class combines the information of various *CEL* files with a common *CDF* file. This class is designed to keep information of one experiment. The probe level data is contained in this object.

```
> Dilution
```

```
AffyBatch object
size of arrays=640x640 features (35221 kb)
cdf=HG_U95Av2 (12625 affyids)
number of samples=4
number of genes=12625
annotation=hgu95av2
notes=
```

The data in `Dilution` is a small sample of probe sets from 2 sets of duplicate arrays hybridized with different concentrations of the same RNA. This information is part of the `AffyBatch` and can be accessed with the `phenoData` and `pData` methods:

```
> phenoData(Dilution)

An object of class "AnnotatedDataFrame"
  sampleNames: 20A, 20B, 10A, 10B
  varLabels and varMetadata description:
    liver: amount of liver RNA hybridized to array in micrograms
    sn19: amount of central nervous system RNA hybridized to array in micrograms

    scanner: ID number of scanner used

> pData(Dilution)

    liver sn19 scanner
20A    20    0       1
20B    20    0       2
10A    10    0       1
10B    10    0       2
```

Various researchers have pointed out the need for normalization of Affymetrix arrays. Let's look at an example. The first two arrays in `Dilution` are technical replicates (same RNA), so the intensities obtained from these should be about the same. The second 2 are also replicates. The second arrays are hybridized to twice as much RNA so the intensities should be in general bigger. However, notice that the scanner effect is stronger than the RNA concentration effect.

```
> pData(Dilution)

    liver sn19 scanner
20A    20    0       1
20B    20    0       2
10A    10    0       1
10B    10    0       2
```

The `boxplot` method for the `AffyBatch` class, shown in Figure 1, shows this is the case.

The method `boxplot` can be used to show PM, MM or both intensities.

As discussed in the next section this plot shows that we need to normalize these arrays.

Figure 1 shows the need for normalization. For example arrays scanned using scanner 1 are globally larger than those scanned with 2.

Another way to see that normalization is needed is by looking at log ratio versus average log intensity (MVA) plots. The method `mva.pairs` will show all MVA plots of each pairwise comparison on the top right half and the interquartile range (IQR) of the

```
> par(mfrow = c(1, 1))
> boxplot(Dilution, col = c(2, 2, 3, 3))
```

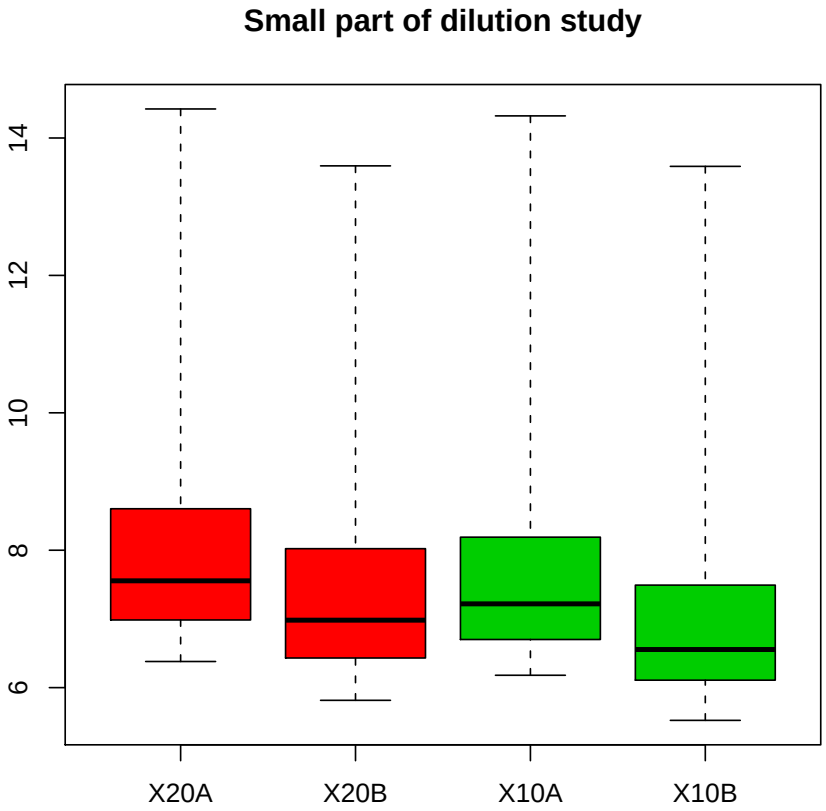

Figure 1: Boxplot of arrays in dilution data.

```
> gn <- sample(geneNames(Dilution), 100)
> pms <- pm(Dilution[, 3:4], gn)
> mva.pairs(pms)
```

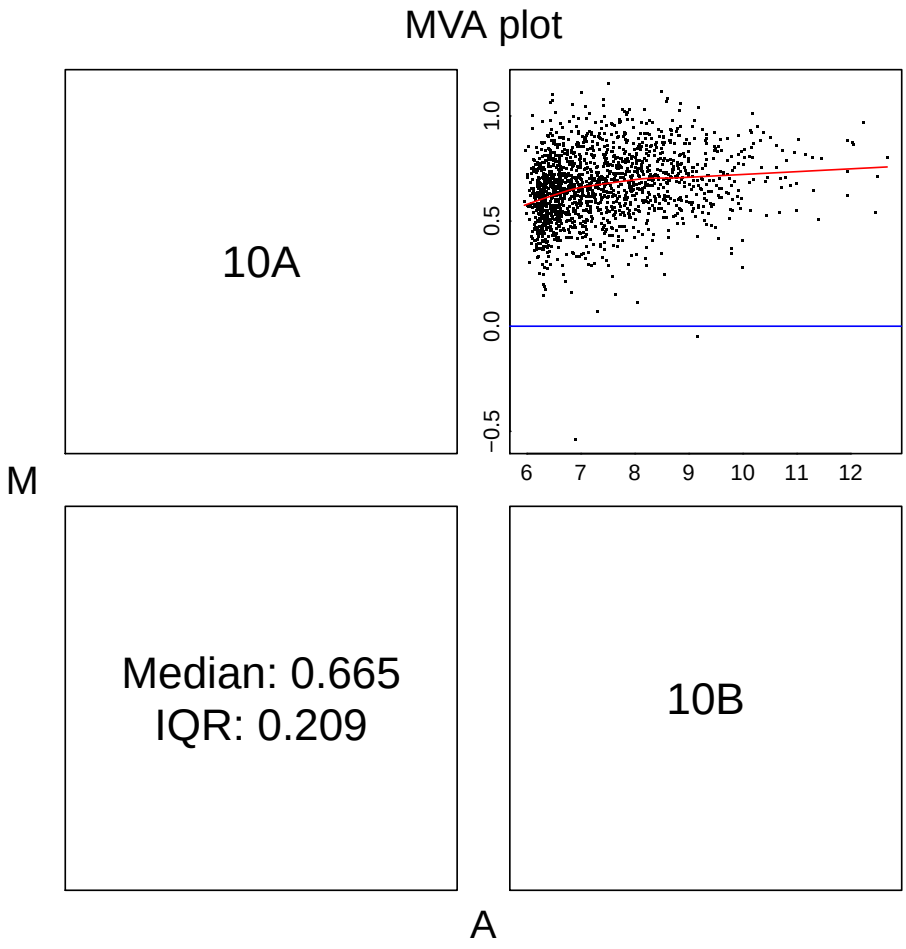

Figure 2: MVA pairs for first two arrays in dilution data

log ratios on the bottom left half. For replicates and cases where most genes are not differentially expressed, we want the cloud of points to be around 0 and the IQR to be small.

The method `normalize` lets one normalize the data.

```
> normalized.Dilution <- Biobase::combine(normalize(Dilution[,
+     1:2]), normalize(Dilution[, 3:4]))
```

We normalize the two concentration groups separately. Notice the function `merge` permits us to put together two `AffyBatch` objects.

Various methods are available for normalization (see the help file). The default is quantile normalization. All the available methods are obtained using this function:

```
> normalize.methods(Dilution)
```

```
[1] "constant"         "contrasts"        "invariantset"     "loess"
[5] "methods"          "qspline"          "quantiles"        "quantiles.robust"
```

and can be called using the `method` argument of the normalize function.

Figures 1 and 1 show the boxplot and mva pairs plot after normalization. The normalization routine seems to correct the boxplots and mva plots.

```
> boxplot(normalized.Dilution, col = c(2, 2, 3, 3), main = "Normalized Arrays")
```

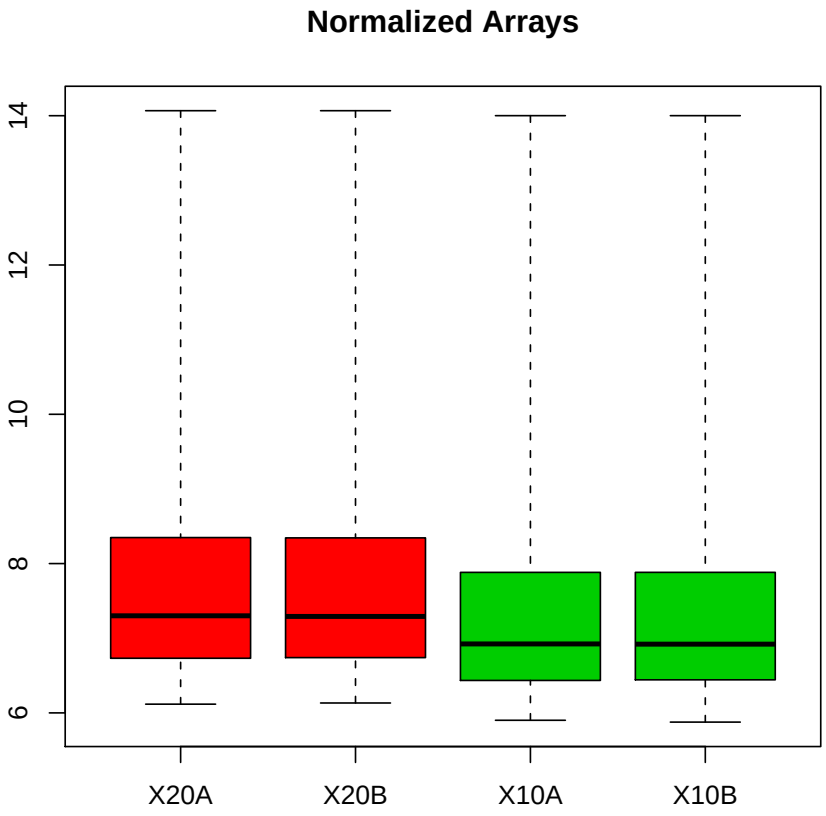

Figure 3: Boxplot of first arrays in normalized dilution data.

```
> pms <- pm(normalized.Dilution[, 3:4], gn)
> mva.pairs(pms)
```

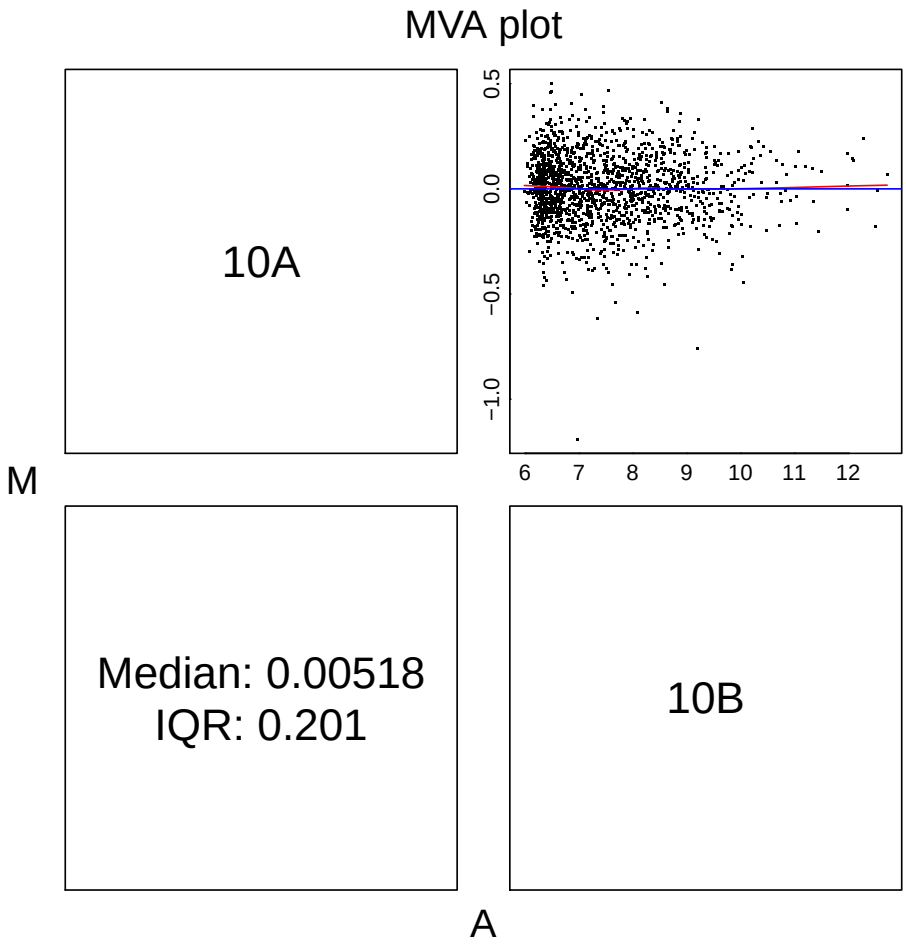

Figure 4: MVA pairs for first two replicate arrays in normalized dilution data