

Analysis of Bead-level Data using beadarray

Mark Dunning

October 10, 2011

Introduction

`beadarray` is a package for the pre-processing and analysis of Illumina BeadArray. The main advantage is being able to read raw data output by Illumina's scanning software. Data presented in this form are in the same format regardless of the assay (i.e expression, genotyping, methylation) being performed. Thus, `beadarray` is able to handle all these types of data. Many functions within `beadarray` have been written to cope with this flexibility.

The BeadArray technology involves randomly arranged arrays of beads, with beads having the same probe sequence attached colloquially known as a bead-type. BeadArrays are combined in parallel on either a rectangular chip (BeadChip) or a matrix of 8 by 12 hexagonal arrays (Sentrix Array Matrix or SAM). The BeadChip is further divided into strips on the surface known as sections, with each section giving rise to a different image when scanned by BeadScan. These images, and associated text files, comprise the raw data for a `beadarray` analysis. However, for BeadChips, the number of sections assigned to each biological sample may vary from 1 on HumanHT12 chips, 2 on HumanWG6 chips or sometimes ten or more for SNP chips with large numbers of SNPs being investigated.

This vignette demonstrates the processing of bead-level data using `beadarray`. The data used are a subset of the data presented in the `BeadArrayUseCases` package and can be downloaded in the `beadarrayExampleData` package. A more comprehensive commentary on the analysis of Illumina BeadArray package is given in the vignette of `BeadArrayUseCases`, including other analysis tools that are not part of `beadarray`.

Citing beadarray

If you use `beadarray` for the analysis or pre-processing of BeadArray data please cite:

Dunning MJ, Smith ML, Ritchie ME, Tavaré S, **beadarray: R classes and methods for Illumina bead-based data**, *Bioinformatics*, **23**(16):2183-2184

1 Asking for help on beadarray

Wherever possible, questions about `beadarray` should be sent to the Bioconductor mailing list¹. This way, all problems and solutions will be kept in a searchable archive. When posting to this mailing list, please first consult the *posting guide*. In particular, state the version of `beadarray` and R that you are using², and try to provide a reproducible example of your problem. This will help us to diagnose the problem.

¹<http://www.bioconductor.org>

²This can be done by pasting the output of running the function `sessionInfo()`.

2 Reading bead-level data into beadarray

2.1 File formats

The raw images and text files required to perform a bead-level analysis are produced by Illumina's BeadScan software. Usually, it will be necessary for you to modify BeadScan's default settings to obtain bead-level data, see <http://www.compbio.group.cam.ac.uk/Resources/illumina>.

The command to read bead-level data from the current working directory is as follows. The `dir` argument may be used to specify an alternative location.

```
> BLData = readIllumina(useImages = FALSE, illuminaAnnotation = "Humanv3")
```

The `useImages` argument specifies whether beadarray will read foreground and background intensities from the TIFF images present in the directory, allowing users to experiment with strategies for image processing. In this example we set `useImages=FALSE` (often a convenient choice), and locally background corrected intensities will simply be extracted from the `txt` files. The *optical* background-correction that is referred to here is done by subtracting the *background* pixel intensities surrounding each bead. It should not be confused with another background correction further along the analysis pipeline, which may involve negative control beads to account for non-specific binding.

Note that it is not compulsory to specify which type of Illumina assay was used to generate the data. However, for expression data it is convenient to specify the name of the platform using one of the strings Humanv4, Humanv3, Humanv2, Humanv1, Mousev2, Mousev1p1, Mousev1 or Ratv1.

beadarray is able to use some of Illumina's files during analysis. These include `.locs` files, which contain the locations of *all* beads on an array (not just those that were decoded), and `.sdf` files, which contain information about the physical layout of the chip. In combination, using these files can result in significant time improvements to the detection of spatial artefacts and add additional information to some QA plots. These files are not read automatically, but if present, the path to these files is stored by beadarray for future use. If the *metrics* file generated by BeadScan is present in the directory, it will be read unaltered and stored.

3 The beadLevelData class

Once imported, the bead-level data are stored in an object of class *beadLevelData*. This class can handle raw data from both single channel and two-colour BeadArrays. Due to the random nature of the technology, each array generally has a variable number of rows of intensity data, and we use an R environment variable to store this information in a memory efficient way.

The *beadLevelData* class contains a number of slots useful for describing Illumina data. The data that have been extracted from the text files are found in the *beadData* slot. This can be thought of as a list, which can be indexed by name or a numeric value representing a particular array-section. A data frame holds the data for that array-section, with the number of rows being the number of beads on the section. For convenience, the function `getBeadData` is used to access data held in the *beadData* slot. The function `insertBeadData` can be used to assign new data to this slot.

Data types with one value per array-section can be stored in the *sectionData* slot. For instance, any metrics information present in the directory used by `readIllumina` will be stored here. This is also a convenient place to store any QC information derived during the pre-processing of the data, as we will see.

The numeric identifiers for the bead-types in the *beadLevelData* are known as ArrayAddress IDs in Illumina's annotation files. For downstream analysis it is convenient to convert these into the form `ILMN_...` used in most annotation packages. Mapping objects to convert these IDs are supplied with beadarray in the *extdata* directory, but this conversion may be performed automatically if the annotation of the *beadLevelData* object is known. For this example dataset, beads that could not be decoded

are assigned a special ArrayAddress ID of 0. For two-channel data, the intensities from the Red channel and associated coordinates are also stored in the object.

```
> class(exampleBLdata)
> slotNames(exampleBLdata)
> exampleBLdata[[1]][1:10, ]
> getBeadData(exampleBLdata, array = 1, what = "Grn")[1:10]
> uIDs = unique(getBeadData(exampleBLdata, array = 1, what = "ProbeID"))
> uIDs[1:10]
```

4 Transformation Functions

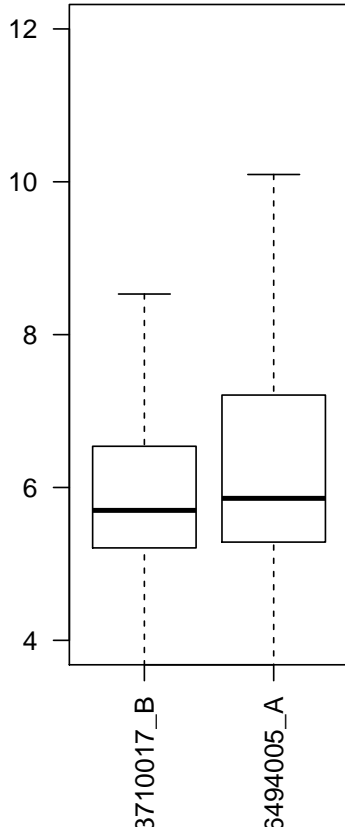
A more flexible way to obtain per-bead data from a beadLevelData object is to define a transformation function that takes as arguments the beadLevelData object and an array index. The function then manipulates the data in the desired manner and returns a vector the same length as the number of beads on the array. The `logGreenChannelTransform` is the default transformation in many plotting / QA functions within `beadarray`. Users with two-channel data may also wish to experiment with the similarly defined `logRedChannelTransform` or `logRatioTransform` when plotting.

```
> log2(exampleBLdata[[1]][1:10, 2])
> logGreenChannelTransform
> logGreenChannelTransform(exampleBLdata, array = 1)[1:10]
> logRedChannelTransform
```

5 Boxplots and imageplots

Two standard quality assessment plots supported by `beadarray` are the imageplot and boxplot. Boxplots can be used to compare foreground and background intensities between arrays. Image plots can be used to identify spatial artefacts on the array surface that can occur from mis-handling or scanning problems. With the raw bead-level data, we can plot false images of each array. This kind of visualisation is not possible when using the summarised BeadStudio output, as the summary values are averaged over spatial positions. Image plots in R are also more convenient than scrutinising the original tiffs, as multiple arrays can be visualised on the one page. By default, the array surface is plotted with the longest edge going horizontally. Both the `boxplot` and `imageplot` functions take a transformation function as an argument, with the default to do a \log_2 transformation on the green channel.

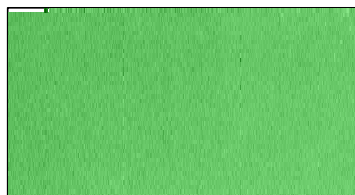
```
> par(mfrow = c(1, 2))
> boxplot(exampleBLdata, las = 2, outline = FALSE, ylim = c(4,
+ 12))
```



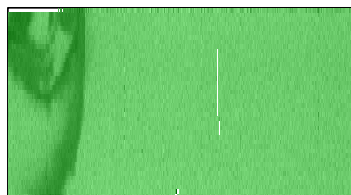
The imageplot can be configured in many ways (see manual page for more details). Sections from a BeadChip often have one edge that is much longer than the other, and it is important to recognise this when producing the plots. By default, `beadarray` makes imageplots with the longest edge on the x-axis (suitable for widescreen monitors). However, with `horizontal = FALSE`, the imageplot will be displayed in the same orientation as the original TIFF image from the directory. With the `squareSize` we can control how many pixels from the original image make up the pixels in the resulting imageplot. The following code produces imageplots for the first three array-sections in the example dataset. Note that we also change the colour scheme to represent low and high intensities by light and dark green respectively.

If `.locs` information is available to `beadarray`, it will be able to determine the optimal `squareSize` parameter. If not (as with our example dataset), the user may have to experiment with different values for `squareSize`.

```
> par(mfrow = c(1, 2), mar = c(2, 2, 2, 2))
> imageplot(exampleBLdata, array = 1, low = "lightgreen", high = "darkgreen")
> imageplot(exampleBLdata, array = 2, low = "lightgreen", high = "darkgreen")
```



z-range 4.8 to 9.3 (saturation 4.8, 9.3)



z-range 5 to 9.6 (saturation 5, 9.6)

6 BASH

BASH is a method for managing the spatial artefacts that may be found on an array as described in Cairns et al (2008). BASH uses the methodology developed for the Harshlight package, but altered to exploit the availability of replicated observations on the same array. The algorithm first identifies Extended defects, where an array has gradual but significant shifts across the surface. BASH also seeks to find more localized artifacts on arrays by classifying features that have unusual intensities as outliers and then finding outliers close to each other on the array. Two separate algorithms then search for areas with a larger numbers of outliers than would be expected by chance (Diffuse Defects) and large connected clusters of outliers (Compact defects). The random nature (both in position and numbers of each feature type) of Illumina arrays mean that the Harshlight algorithm must proceed in a different way to the original Harshlight implementation. Whereas Affymetrix probes have replicates on other arrays, Illumina beads are replicated on the same array. We can therefore generate an error image based on how much each bead differs from the median of its replicates' intensities, instead of replicates on other arrays. Having performed manipulations to the error image, we can then find outliers on this image by bead type, determining which beads are more than 3 Median Absolute Deviations, or MADs, from the median.

Finally, since Illumina arrays are randomly arranged and use a hexagonal grid rather than rectangular, BASH has it's own method for creating networks of beads on the array. However, if `.locs` files are available to `beadarray` the time taken for this step will be improved considerably.

```
> bsh = BASH(exampleBLdata, array = 1:2)
```

The result of `bash` includes quality control scores; the number of beads masked in total and the extended score.

```
> bsh$QC
> par(mfrow = c(1, 2))
> barplot(bsh$QC[, 1], main = "Number of beads masked")
> barplot(bsh$QC[, 2], main = "Extended Score")
```

The weights themselves can be stored using `setWeights`. These will be taken into when summarizing the bead-level data. The QC tables themselves can be appended to the `sectionData` slot of `BLData`.

```
> for (i in 1:10) {
+   BLData <- setWeights(exampleBLdata, wts = bsh$wts[[i]], array = i)
+ }
> BLData <- insertSectionData(exampleBLdata, what = "BASHQC", data = bsh$QC)
```

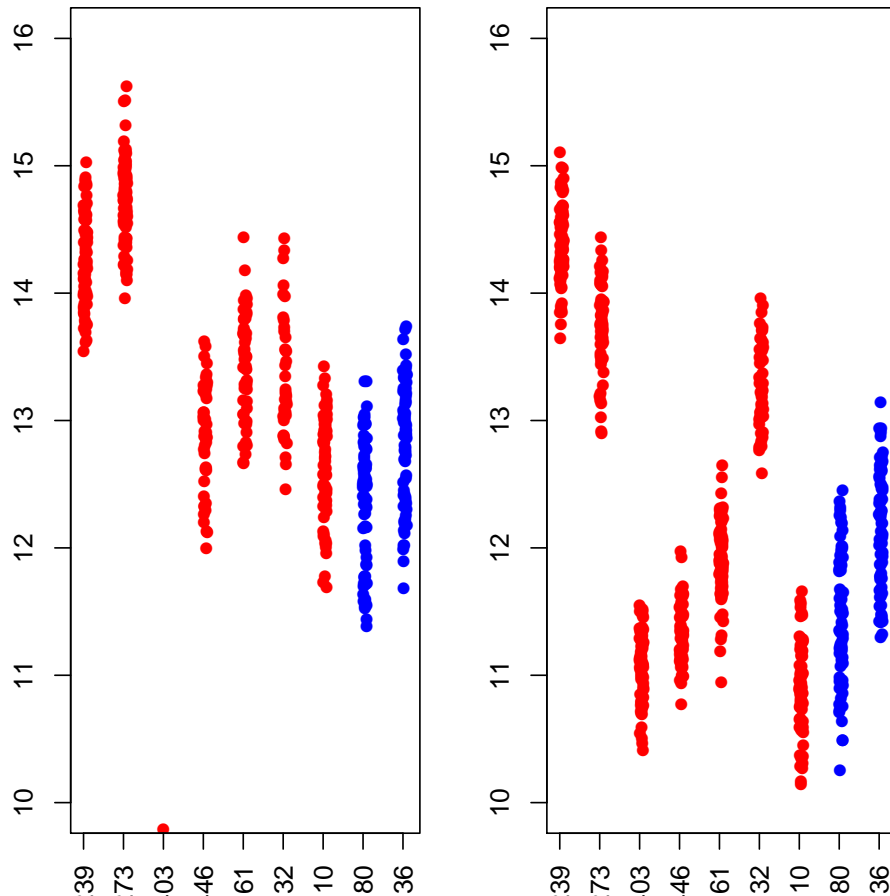
7 Using control information

Illumina have designed a number of control probes for each expression platform. For expression arrays, we store the ArrayAddressIDs of the control probes for in the ExpressionControlData object. Otherwise a data frame may be used to define these ids. As the example data described in this vignette were derived using an obsolete technology, we have stored the control information with the package in the `controlProfile` object. ArrayAddressIDs are listed in the first column, and the type of control in the second column. Objects of this form can be used in various quality assessment functions in `beadarray`.

```
> controlProfile <- makeControlProfile("Humanv3")
> head(controlProfile)
> table(controlProfile[, 2])
```

Two particular controls on expression arrays are housekeeping and biotin controls. With the `poscontPlot` function, we can plot the intensities of any ArrayAddressIDs that are annotated as belonging to the Housekeeping or Biotin group in the `ExpressionControlData` object or `controlProfile`. The `poscontPlot` is flexible in allowing other "tags" in the `controlProfile`, in the example below we configure the plot to show both housekeeping and negative controls in the same plot.

```
> par(mfrow = c(1, 2), mar = c(2, 2, 2, 2))
> for (i in 1:2) {
+   poscontPlot(exampleBLdata, array = i, ylim = c(10, 16))
+ }
```



With knowledge of which `ArrayAddressIDs` match control types, we can easily provide summaries of these control types on each array. In `quickSummary` the mean and standard deviation of all control types is taken for a specified array, using intensities of all beads that correspond to the control type. Note that these summaries may not correspond to similar quantities reported in Illumina's BeadStudio software, as the BeadStudio summaries are produced after removing outliers (see later).

The `makeQCTable` function extends this functionality to produce a table of summaries for all sections in the `beadLevelData` object. These data can be stored in the `sectionData` slot for future reference.

It is also informative to compare the expression level of various control types to the background level of the array. This is done by the `controlProbeDetection` function that returns the percentage of each control type that are significantly expressed above background level. Obviously for positive controls we would prefer this percentage to be near 100 on a good quality array.

```
> quickSummary(exampleBLdata, array = 1)
> qcReport <- makeQCTable(exampleBLdata)
> head(qcReport)[, 1:5]
> BLdata <- insertSectionData(exampleBLdata, what = "BeadLevelQC",
+   data = qcReport)
> names(exampleBLdata)
> for (i in 1:2) {
+   print(controlProbeDetection(exampleBLdata, array = i, negativeTag = "negative"))
+ }
```

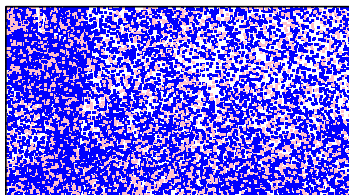
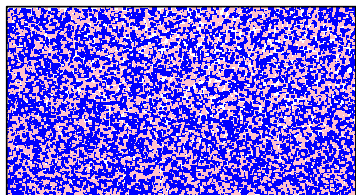
The generation of QA plots for all sections in the `beadLevelData` object is provided by the `expressionQCPipeline` function. Results are generated in a directory of the users choosing. This report may be generated at any point of the analysis. If the `overwrite` parameter is set to `FALSE`, then any existing plots in the directory will not be re-generated. Furthermore, QC tables that have been stored in the `beadLevelData` object already can be used.

8 Outlier removal and plotting

Before combining the observations for each bead-type on an array, Illumina remove any observations with outlying intensity (more than 3 median absolute deviations from the median). This step can be repeated in `beadarray` and it is sometimes useful to see where these outliers are located on the array surface. Often, they will coincide with beads masked by BASH or with any spatial artefacts that may be seen.

Users are able to define their own functions to identify outliers. Such functions must take a list of intensities and corresponding `ArrayAddressIDs` and return indices of which observations are found to be outliers.

```
> par(mfrow = c(1, 2), mar = c(2, 2, 2, 2))
> outlierplot(exampleBLdata, array = 1)
> outlierplot(exampleBLdata, array = 2)
```



9 Summarization

The summarization procedure takes the `BLData` object, where each bead-type is represented by differing numbers of observations on each array, and produces a summarized object to make comparisons between arrays. For each array section represented in the `BLData` object, all observations are extracted, transformed, and then grouped together according to their `ArrayAddressID`. Outliers are removed and the mean and standard deviation of the remaining beads are calculated.

The `illuminaChannel` class is used to define how summarization proceeds with specification of a transformation function, a function to remove outliers and function to calculate the means and standard deviation. The code below creates two different summarized objects; one which uses mean and standard deviations, and one which uses median and standard errors.

```
> BSDData <- summarize(exampleBLdata)

> myMedian <- function(x) median(x, na.rm = TRUE)
> myMad <- function(x) mad(x, na.rm = TRUE)
> greenChannel2 <- new("illuminaChannel", logGreenChannelTransform,
+   illuminaOutlierMethod, myMedian, myMad, "G")
> BSDData2 <- summarize(exampleBLdata, list(greenChannel2))
```



```
> BSData
```

The `BSData` object is very similar to the *ExpressionSet* class in Biobase. However, to accomodate the unique features of Illumina data we have added an `nObservations` slot, which gives the number of beads that we used to create the summary values for each bead-type on each array after outlier removal.

It is possible to have multiple channels, each of which is summarized in a different manner, in the same `ExpressionSetIllumina` object. This is achieved by passing a list of `illuminaChannel` objects to `summarize`. This would be especially useful for two-channel data, where the Red and Green channels, and some combination of the two would be of interest in the analysis. In the example code below we summarize both the non background-corrected and background-corrected intensities in the same object. The `channel` function is used to select the data for one of these channels, which returns an object of type *ExpressionSetIllumina*.

The detection score is a standard measure for Illumina expression experiments, and can be viewed as an empirical estimate of the p-value for the null hypothesis that there is no expression. These can be calculated for summarized data provided that the identity of the negative controls on the array is known. For further analysis of the summarized object, see the separate beadsummary vignette `beadsummary.pdf`.

```
> det = calculateDetection(BSData)
> head(det)
> Detection(BSData) <- det

> sessionInfo()
```