Introduction of beadarraySNP

April 21, 2009

Introduction

beadarraySNP is a Bioconductor package for the analysis of Illumina genotyping BeadArray data, especially derived from the GoldenGate assay. The functionality includes importing datafiles produced by Illumina software, doing LOH analysis and performing copy number analysis.

1 Import

The package contains an artificially small example to show some of the key aspects of analysis. Normally an experiment contains 96 samples spread over four probe panels of about 1500 SNPs each. The example has data from 2 colorectal tumors and corresponding leukocyte DNA of 1 probe (or OPA) panel. This particular OPA panel contains probes on chromosomes 16-20 and X and Y. In this case all datafiles have been put in 1 directory. The import function `read.SnpSetIllumina` can be invoked to use all separate directories that are used with the Illumina software.

We start by loading in the data into a *SnpSetIllumina* object. A samplesheet is used to identify the samples in the experiment. This file can be created by the Illumina software. The next code shows the example samplesheet. The import function searches for the line starting with [Data]. The "Sample_Well", "Sample_Plate", and "Sample_Group" columns are not used, but the other columns should have meaningful values. The "Pool_ID" and "Sentrix_ID" should be identical for all samples in 1 samplesheet. Samples from multiple experiments or multiple probe panels can be combined after they are imported.

```
> datadir <- system.file("testdata", package = "beadarraySNP")
> readLines(paste(datadir, "4samples_opa4.csv", sep = "/"))

 [1] "[Header],,,,,,"
 [2] "Investigator Name,Anon,,,,,"
 [3] "Project Name,LOH,,,,,"
 [4] "Experiment Name,LOH,,,,,"
 [5] "Date,11032005,,,,,"
 [6] ""
 [7] "[Data],,,,,,"
 [8] "Sample_Name,Sample_Well,Sample_Plate,Sample_Group,Pool_ID,Sentrix_ID,Sentrix_Position"
 [9] "106NB,B12,GS0008391-HYB,training,GS0005704-OPA,1280260,R002_C012"
[10] "106TV,C12,GS0008391-HYB,training,GS0005704-OPA,1280260,R003_C012"
[11] "108NB,G10,GS0008391-HYB,training,GS0005704-OPA,1280260,R007_C010"
[12] "108TV,H10,GS0008391-HYB,training,GS0005704-OPA,1280260,R008_C010"

> SNPdata <- read.SnpSetIllumina(paste(datadir, "4samples_opa4.csv",
+     sep = "/"), datadir)
> SNPdata
```

```
SnpSetIllumina (storageMode: list)
assayData: 1453 features, 4 samples
  element names: call, callProbability, G, R
phenoData
  sampleNames: 106NB, 106TV, 108NB, 108TV
  varLabels and varMetadata description:
    Sample_Name: Sample_Name
    Sample_Well: Sample_Well
    ...: ...
    Col: Col
    (10 total)
featureData
  featureNames: rs935971, rs963598, ..., rs3093505  (1453 total)
  fvarLabels and fvarMetadata description:
    OPA: OPA
    snpid: snpid
    ...: ...
    GTS: GTS
    (8 total)
experimentData: use 'experimentData(object)'
Annotation: GS0005704-OPA
```

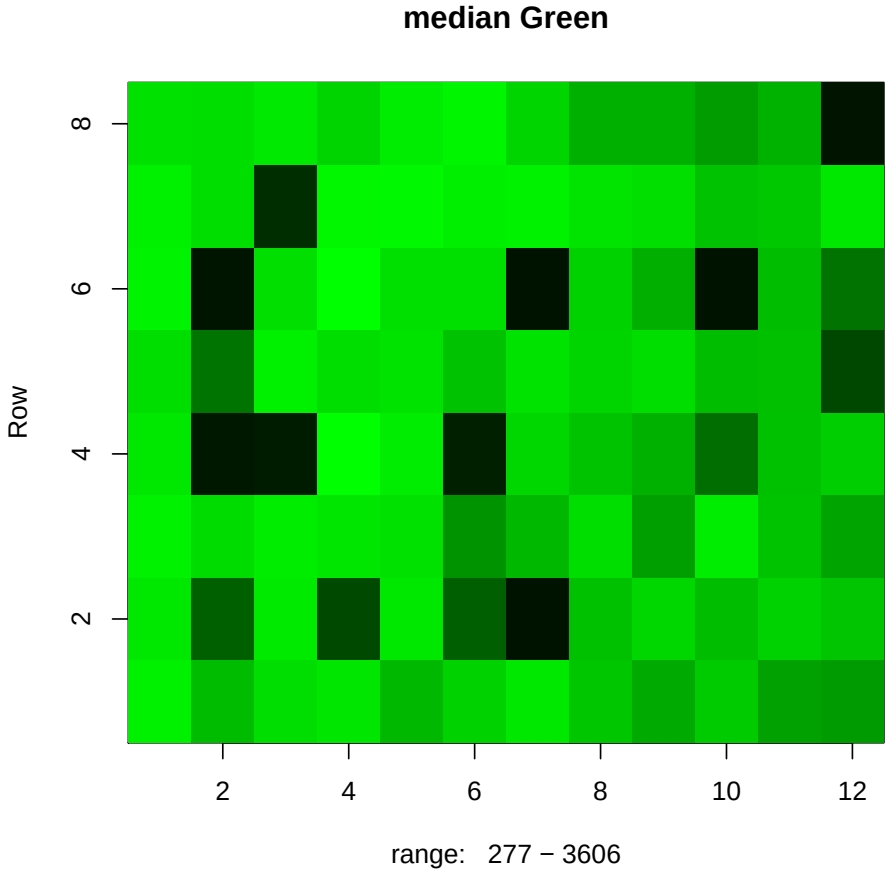
The targets file contains extra information on the samples which are important during normalization. The `NorTum` column indicates normal and tumor samples. The tumor samples are not used for the invariant set in between sample normalization. The `Gender` column is used to properly normalize the sex chromosomes. The following lines show a possible way to add these columns to the `phenoData` slot of the object. If a `NorTum` and/or `Gender` column is in the `phenoData` slot they will be automatically used later on.

```
> pd <- read.AnnotatedDataFrame(paste(datadir, "targets.txt", sep = "/"),
+     sep = "\t")
> pData(SNPdata) <- cbind(pData(SNPdata), pData(pd))
```

2 Quality control

Illumina Sentrix arrays contain 96 wells to process up to 96 samples. Each well can be used with a separate set of probes or OPA panel. The example shows the QC of the full experiment form which the example files were taken. This experiment consisted of 24 samples with 4 probe panels each

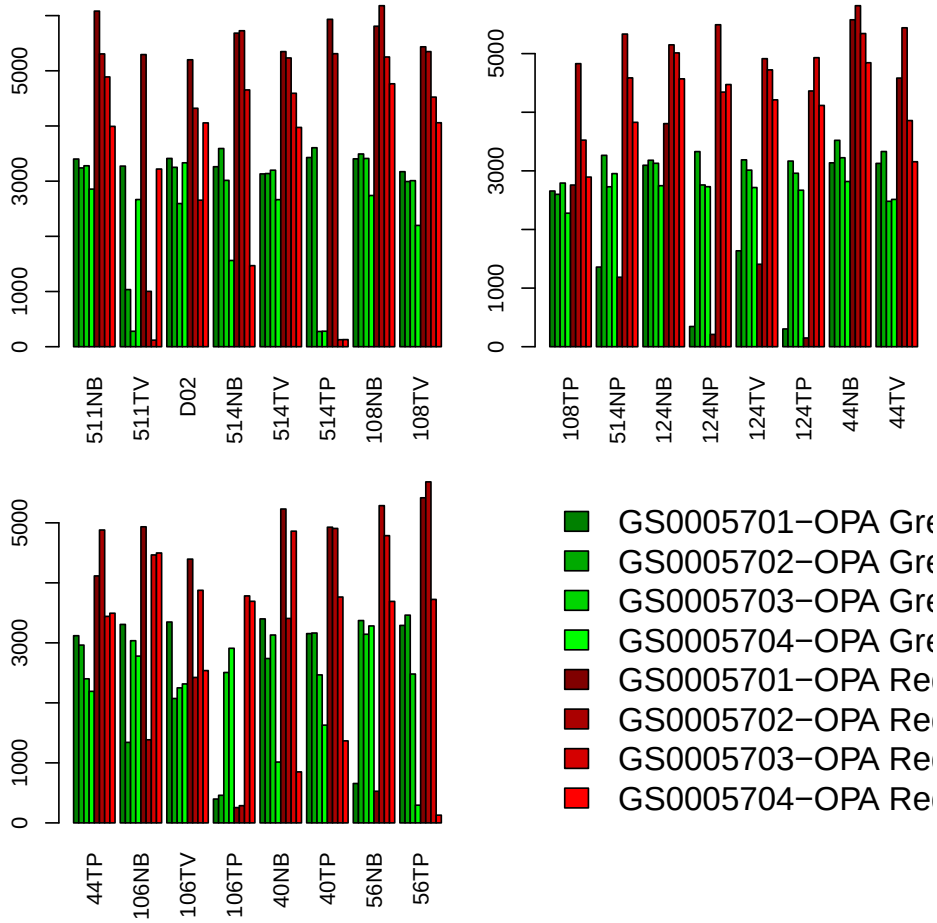
```
> qc <- calculateQCarray(SNPdata)
> data(QC.260)
> plotQC(QC.260, "greenMed")
```



Other types of plots (”intensityMed”,”greenMed”,”redMed”,”validn”,”annotation”, ”samples”) show the median intensity, median red intensity or identifying information.

```
> par(mfrow = c(2, 2), mar = c(4, 2, 1, 1))
> reportSamplePanelQC(QC.260, by = 8)
> SNPdata <- removeLowQualitySamples(SNPdata, 1500, 100, "OPA")

GS0005704-OPA 4 subsamples, 0 removed
```



Based on these findings, low quality samples can be removed from the experiment

3 Normalization

Normalization to calculate copy number is a multi-step process. In Oosting(2007) we have determined that the following procedure provides the optimal strategy:

1. Perform quantile normalization between both colors of a sample. This is allowed because the frequencies of both alleles throughout a sample are nearly identical in practice. This action also neutralizes any dye bias.

2. Scale each sample using the median of the high quality heterozygous SNPs as the normalization factor. Genomic regions that show copy number alterations are likely to show LOH(loss of heterozygosity), or are harder to genotype leading to a decrease quality score of the call.

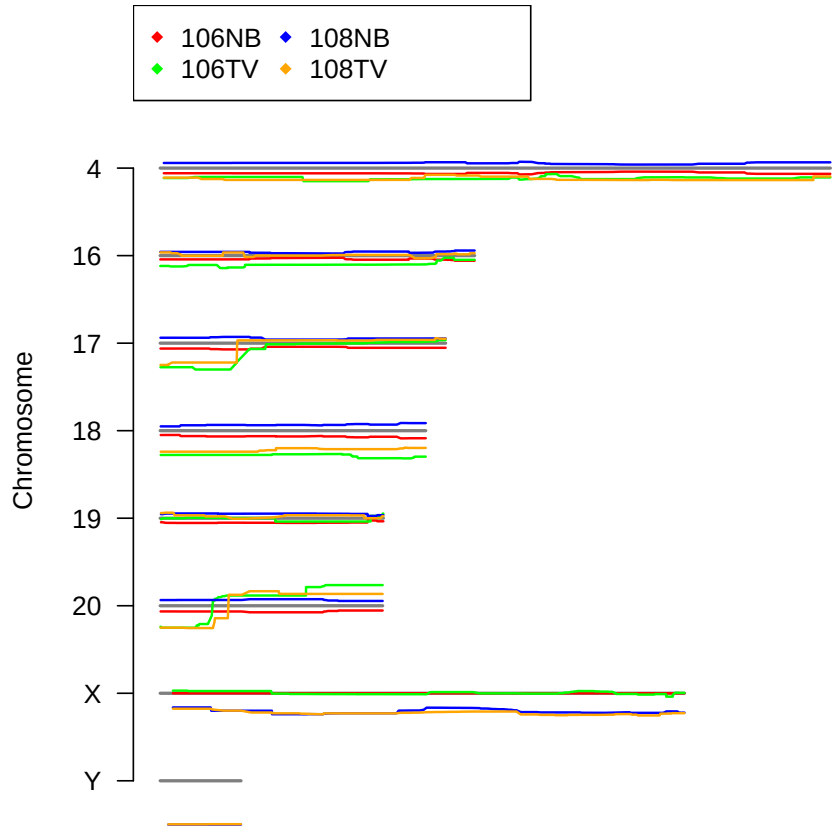
3. Scale each probe using the normal samples in the experiment. Assume that these samples are diploid, and have a copy number of 2.

```
> SNPnrm <- normalizeBetweenAlleles.SNP(SNPdata)
> SNPnrm <- normalizeWithinArrays.SNP(SNPnrm, callscore = 0.8,
+     relative = TRUE, fixed = FALSE, quantilepersample = TRUE)
> SNPnrm <- normalizeLoci.SNP(SNPnrm, normalizeTo = 2)
```

4 Reporting

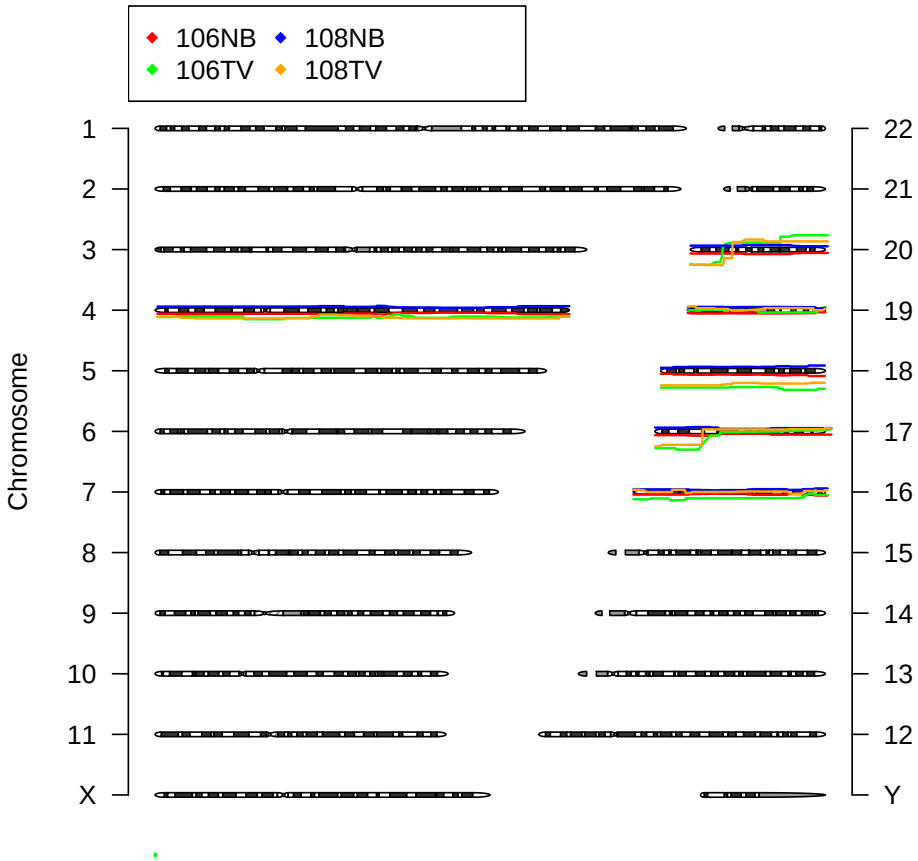
Although the OPA panels contain distinct chromosomes, also a few spurious SNPs on other chromosomes are in it. We first select the probes that are located on the chromosomes for this OPA panel.

```
> SNPnrm <- SNPnrm[featureData(SNPnrm)$CHR %in% c("4", "16", "17",
+     "18", "19", "20", "X", "Y"), ]
> reportSamplesSmoothCopyNumber(SNPnrm, normalizedTo = 2, smooth.lambda = 4)
```



A figure is created of all 4 samples in the dataset with copynumber along the chromosomes.

```
> reportSamplesSmoothCopyNumber(SNPnrm, normalizedTo = 2, paintCytobands = TRUE,
+     smooth.lambda = 4, organism = "hsa", sexChromosomes = TRUE)
```



By using the `organism` a figure is created that shows all chromosomes in 2 columns. Experimental data is plotted wherever it is available in the dataset.

5 References

Oosting J, Lips EH, van Eijk R, Eilers PH, Szuhai K, Wijmenga C, Morreau H, van Wezel T. High-resolution copy number analysis of paraffin-embedded archival tissue using SNP BeadArrays. Genome Res. 2007 Mar;17(3):368-76. Epub 2007 Jan 31.

6 Session Information

The version number of R and packages loaded for generating the vignette were:

- R version 2.9.0 (2009-04-17), `x86_64-unknown-linux-gnu`
- Locale: `LC_CTYPE=en_US;LC_NUMERIC=C;LC_TIME=en_US;LC_COLLATE=en_US;LC_MONETARY=C;LC_MESSAGES=en_US;`
- Base packages: base, datasets, graphics, grDevices, grid, methods, stats, tools, utils
- Other packages: beadarraySNP 1.10.0, Biobase 2.4.0, limma 2.18.0, lodplot 1.1, quantreg 4.27, quantsmooth 1.10.0, SparseM 0.79