Using the inSilicoMerging package

Jonatan Taminau*

CoMo, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

1 Merging Gene Expression Data

An increasing amount of gene expression datasets is available through public repositories like for example GEO [2] and ArrayExpress [6]. Combining such data from different studies could be beneficial for the discovery of new biological insights and could increase the statistical power of gene expression analysis. However, the use of different experimentation plans, platforms and methodologies by different research groups introduces undesired batch effects in the gene expression values. This problem hinders and complicates further analysis and can even lead to incorrect conclusions [3]. Several methods to remove this bias but at the same time to preserve the biological variance inside the data are proposed in the last years. The inSilicoMerging package combines several of the most used methods to remove this unwanted batch effects in order to actually merge different datasets. All methods are implemented in such way that they can be consistently used inside the Bioconductor framework.

2 Using the inSilicoMerging package

Using the inSilicoMerging package is straightforward since it mainly involves only a single function:

```
> merge(esets);
```

with `esets` a list of ExpressionSet objects and `method` one of the following options: `BMC`, `COMBAT`, `DWD`, `GENENORM` and `XPN`. Each of those methods is already extensively reported in literature but is nevertheless briefly explained in the following section.

In order to visually inspect a merged dataset to have some direct feedback on its effect, three different visual validation methods are provided:

*jtaminau@vub.ac.be

```
> plotMDS(eset, ...)
> plotRLE(eset,...)
> plotGeneWiseBoxPlots(eset,...)
```

`plotMDS` creates a *double-labeled* Multidimensional Scaling (MDS) plot. In this plot, all samples can be labeled by color and by symbol. This might be useful since for each sample its biological phenotype of interest and the study it originates from can be visualized simultaneously, giving an indication of the effectiveness of the used merging method. `plotRLE` creates a relative log expression (RLE) plot, which was initially proposed to measure the overall quality of a dataset but can also be used in this context. Finally, `plotGeneWiseBoxPlots` provides a local visualization by looking at the gene-wise boxplots of samples. All three methods are illustrated in the examples section.

3 Different Merging Methods

Below we list, alphabetically, the merging techniques available through this package. Note that after using any of those methods the resulting merged dataset only contains the *common* list of genes/probes between all studies.

BMC

In [8] they successfully applied a technique similar to z-score normalization for merging breast cancer datasets. They transformed the data by batch mean-centering, which means that the mean is subtracted:

$$\hat{x}_{ij}^k = x_{ij}^k - \overline{x}_i^k \tag{1}$$

This technique was proposed to eliminate multiplicative bias.

COMBAT

Empirical Bayes [4] is a method that estimates the parameters of a model for mean and variance for each gene and then adjusts the genes in each batch to meet the assumed model. The parameters are estimated by pooling information from multiple genes in each batch. It is assumed that measured gene expression values of gene i in sample j of batch k can be expressed as:

$$x_{ij}^k = \alpha_i + \mathbf{C}\beta_i + \gamma_i^k + \delta_i^k \epsilon_{ij}^k \tag{2}$$

where α_i is the overall gene expression, $\mathbf{C}$ is a design matrix for sample conditions, β_i is the vector of regression coefficients corresponding to $\mathbf{X}$, γ_i^k and δ_i^k are the additive and multiplicative batch effects for gene i in batch k respectively and ϵ_{ij}^k are error terms.

DWD

By searching for the separating hyperplane between data coming from different batches, Distance-weighted discrimination (DWD), an adaptation of Support Vector Machines (SVM), allows to remove bias by projecting the different batches on the hyperplane, calculating the batch mean $\overline{b}$ distance to the hyperplane and then subtracting the normal vector Δ of this plane multiplied by the mean [1].

$$\hat{x}_{ij}^k = x_{ij}^k - \overline{b}\Delta \tag{3}$$

GENENORM

One of the simplest mathematical transformations to make datasets more comparable is z-score normalization. In this method, for each gene expression value x_{ij} in each study separately all values are modified by subtracting the mean $\overline{x}_i$ of the gene in that dataset divided by its standard deviation σ_i:

$$\hat{x}_{ij}^k = \frac{x_{ij}^k - \overline{x}_i^k}{\sigma_{i.}^k} \tag{4}$$

No additional transformation

The most basic approach to combine two datasets is to simply *paste* them together without any transformation. This can be used as a baseline against which other techniques can be compared.

XPN

The basic idea behind the cross-platform normalization [7] approach is to identify homogeneous blocks (clusters) of gene and samples in both studies that have similar expression characteristics. In XPN, a gene measurement can be considered as a scaled and shifted block mean. For a platform k, gene i and sample j, the recorded gene expression is given by:

$$x_{ij}^k = A_{\alpha^*(i),\beta_k^*(j)}^k b_i^k + c_i^k + \sigma_i^k \epsilon_{ij}^k \tag{5}$$

where A_{α^*,β^*}^k is a block mean and b_i^k and c_i^k represent gene and platform specific sensitivity and offset parameters respectively. The functions $\alpha^*()$ and $\beta^*()$ map a specific gene measurement in a sample to their corresponding multi-platform cluster. The noise variables ϵ_{ij}^k are assumed independent standard gaussians. `XPN` uses an iterative scheme to update the parameters in Equation 5 until convergence to a local minimum, giving:

$$\hat{x}_{ij}^k = \hat{A}_{\alpha^*(i),\beta_k^*(j)}^k \hat{b}_i^k + \hat{c}_i^k + \hat{\sigma}_i^k \hat{\epsilon}_{ij}^k \tag{6}$$

More details can be found in [7].

3.1 Merging two-by-two

Some merging technique are only reported and implemented to merge exactly two studies (e.g. `XPN` [7] and `DWD` [1]). In order to be able to merge any number of studies, this package added an additional step. This step combines all studies two-by-two and is called recursively on the intermediate results until only one, merged, dataset remains. Its behavior is illustrated in the following example:

```
list of studies = [ A ; B ; C ; D ; E ]
m(X,Y) = applying merging technique 'm' on dataset 'X' and 'Y'
combineByTwo:
  iteration 1 : [ E ; m(A,B) ; m(C,D) ]   => [ E ; AB ; CD ]
  iteration 2 : [ CD ; m(E,AB) ]          => [ CD ; EAB ]
  iteration 3 : [ m(CD,EAB) ]             => [ CDEAB ]
```

4 Example

For this example we retrieve two Lung Cancer datasets using the inSilicoDb package [9]. Both datasets were assayed on a different platform (Affymetrix Human Genome U133A Array versus Affymetrix Human Genome U133 Plus 2.0 Array) and where preprocessed using fRMA [5].

```
> library(inSilicoDb)
> eset1 = getDataset("GSE19804", "GPL570", norm="FRMA", features = "gene", curation = 17470)
> eset2 = getDataset("GSE10072", "GPL96",  norm="FRMA", features = "gene", curation = 17469)
> esets = list(eset1,eset2);
```

Both studies contain normal and tumor samples and are already consistently annotated with a common `Disease` feature:

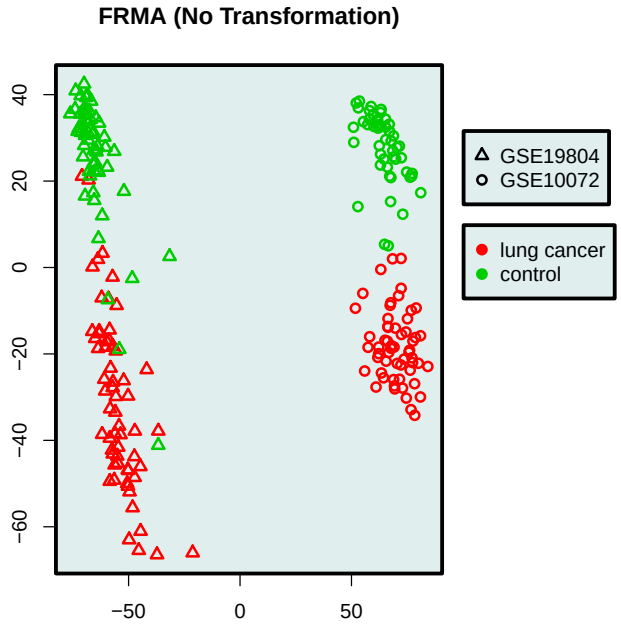
```
> table(pData(eset1)[,"Disease"]);

    control lung cancer
         60          60

> table(pData(eset2)[,"Disease"]);

    control lung cancer
         49          58
```

We now can simply merge both studies without applying any transformation:

```
> library(inSilicoMerging);
> eset_FRMA = merge(esets);
```

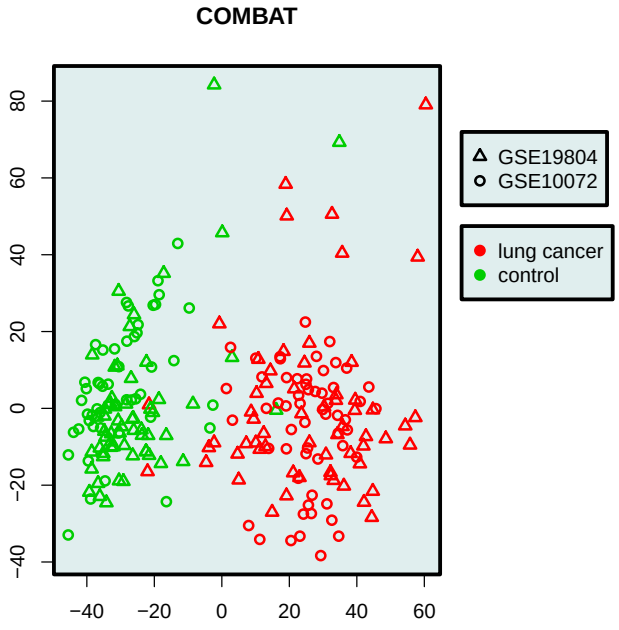
To further investigate the combined data we can use the `plotMDS` function to have a first visual inspection.

```
> plotMDS(eset_FRMA,
+         colLabel="Disease",
+         symLabel="Study",
+         main="FRMA (No Transformation)");
```



From this plot we can immediately notice a very strong dataset-bias (probably due to the difference in platform) while we would expect that all control samples from both studies would cluster together. Let us try another method to see if we can solve this issue:

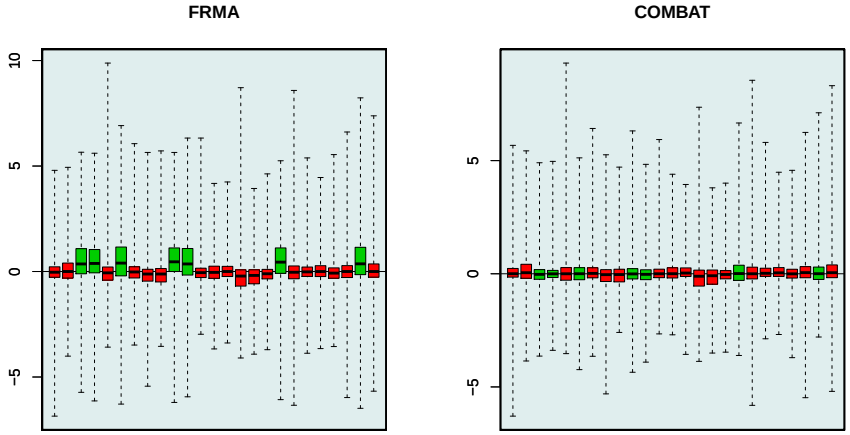
```
> eset_COMBAT = merge(esets, method="COMBAT");
> plotMDS(eset_COMBAT,
+         colLabel="Disease",
+         symLabel="Study",
+         main="COMBAT");
```



This clearly looks better. Both studies are mixed together and the biological phenotype of interest (tumor versus normal) is preserved in the merged dataset.

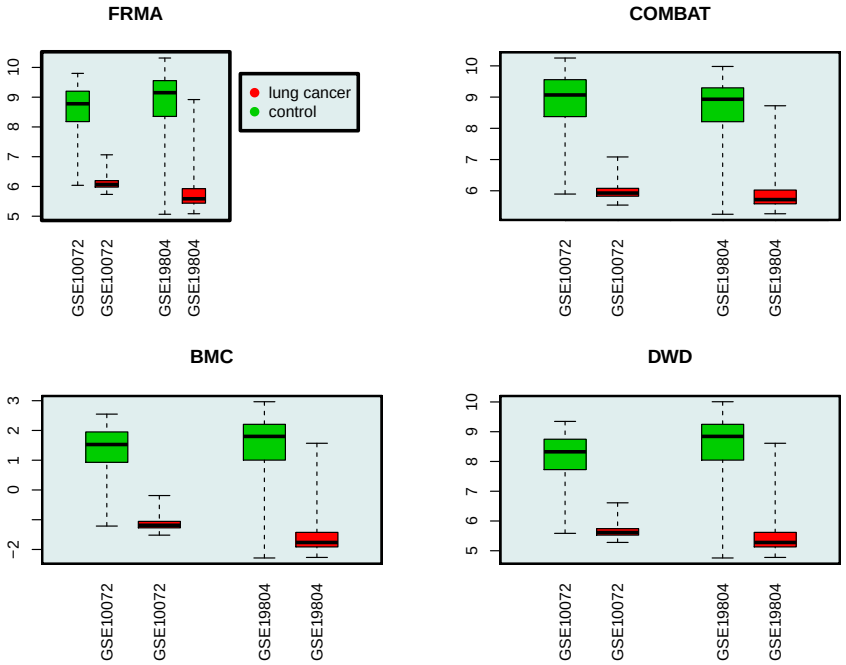
In a similar way we can use the other visualization methods too. To illustrate the RLE plots we only select 25 (random) samples for clarity purposes. We can compare the merging without transformation on the left and after using the `COMBAT` method on the right. In this plot we color the samples based on the study they originate from.

```
> par(mfrow=c(1,2))
> select = sample(1:ncol(eset_FRMA),25);
> plotRLE(eset_FRMA[,select], colLabel="Study", legend=FALSE, main="FRMA");
> plotRLE(eset_COMBAT[,select], colLabel="Study", legend=FALSE, main="COMBAT");
```



Finally, in the last visualization method the local effect of each method on the gene level can be illustrated with a gene-wise boxplot. We arbitrary select the `CA4` genes to investigate:

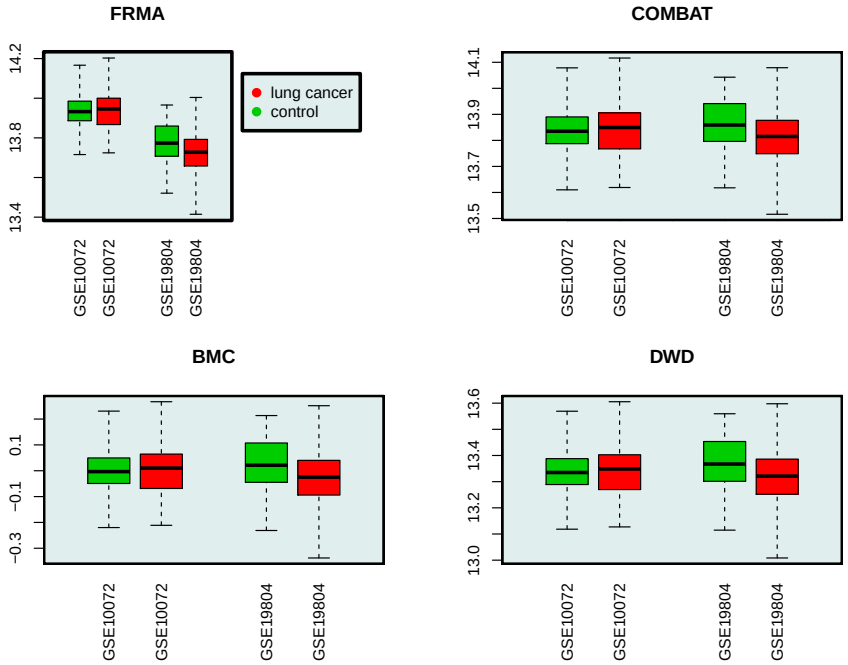
```
> gene = "CA4";
> eset_BMC = merge(esets, method="BMC");
> eset_DWD = merge(esets, method="DWD");
> par(mfrow=c(2,2));
> plotGeneWiseBoxPlot(eset_FRMA, colLabel="Disease", batchLabel="Study",
+                     gene=gene, legend=TRUE, main="FRMA");
> plotGeneWiseBoxPlot(eset_COMBAT, colLabel="Disease", batchLabel="Study",
+                     gene=gene, legend=FALSE, main="COMBAT");
> plotGeneWiseBoxPlot(eset_BMC, colLabel="Disease", batchLabel="Study",
+                     gene=gene, legend=FALSE, main="BMC");
> plotGeneWiseBoxPlot(eset_DWD, colLabel="Disease", batchLabel="Study",
+                     gene=gene, legend=FALSE, main="DWD");
```



In contrast to the two previous methods which illustrated the global bias between the two datasets we have a very local view this time. This gene is clearly differentially expressed (ok, maybe it was not that arbitrary after all :-)) in both studies and without transformation the dataset-bias is not problematic. All merging methods take this into account and only small modification are performed.

For other genes this situation can vary, for example for a relatively stable gene:

```
> gene = "RPL37A";
> par(mfrow=c(2,2));
> plotGeneWiseBoxPlot(eset_FRMA, colLabel="Disease", batchLabel="Study",
+                     gene=gene, legend=TRUE, main="FRMA");
> plotGeneWiseBoxPlot(eset_COMBAT, colLabel="Disease", batchLabel="Study",
+                     gene=gene, legend=FALSE, main="COMBAT");
> plotGeneWiseBoxPlot(eset_BMC, colLabel="Disease", batchLabel="Study",
+                     gene=gene, legend=FALSE, main="BMC");
> plotGeneWiseBoxPlot(eset_DWD, colLabel="Disease", batchLabel="Study",
+                     gene=gene, legend=FALSE, main="DWD");
```



As this example illustrates, it is now straightforward to merge a number of gene expression studies by applying different existing methods. A number of simple visualization tools are provided for a first inspection of the merged dataset(s).

5 Session Info

```
> sessionInfo()

R version 3.1.1 (2014-07-10)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets  methods
[8] base
```

```
other attached packages:
[1] inSilicoMerging_1.8.7 DWD_0.11              Matrix_1.1-4
[4] inSilicoDb_2.0.1      RCurl_1.95-4.3        bitops_1.0-6
[7] Biobase_2.24.0        BiocGenerics_0.10.0   rjson_0.2.14

loaded via a namespace (and not attached):
[1] grid_3.1.1      lattice_0.20-29 tools_3.1.1
```

References

[1] Monica Benito, Joel Parker, Quan Du, Junyuan Wu, Dong Xiang, Charles M. Perou, and J. S. Marron. Adjustment of systematic microarray data biases. *Bioinformatics*, 20(1):105–114, 2004.

[2] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207–10, Jan 2002.

[3] Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*, 11(10):733–9, 2010.

[4] Cheng Li and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[5] Matthew N McCall, Benjamin M Bolstad, and Rafael A Irizarry. Frozen robust multiarray analysis (fRMA). *Biostatistics*, 11(2):242–53, 2010.

[6] Helen Parkinson, Ugis Sarkans, Nikolay Kolesnikov, Niran Abeygunawardena, Tony Burdett, Miroslaw Dylag, Ibrahim Emam, Anna Farne, Emma Hastings, Ele Holloway, Natalja Kurbatova, Margus Lukk, James Malone, Roby Mani, Ekaterina Pilicheva, Gabriella Rustici, Anjan Sharma, Eleanor Williams, Tomasz Adamusiak, Marco Brandizi, Nataliya Sklyar, and Alvis Brazma. Arrayexpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res*, 39(Database issue):D1002–4, Jan 2011.

[7] Andrey A. Shabalin, Håkon Tjelmeland, Cheng Fan, Charles M. Perou, and Andrew B. Nobel. Merging two gene-expression studies via cross-platform normalization. *Bioinformatics*, 24(9):1154–1160, 2008.

[8] Andrew Sims, Graeme Smethurst, Yvonne Hey, Michal Okoniewski, Stuart Pepper, Anthony Howell, Crispin Miller, and Robert Clarke. The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets - improving meta-analysis and prediction of prognosis. *BMC Medical Genomics*, 1(1):42, 2008.

[9] Jonatan Taminau, David Steenhoff, Alain Coletta, Stijn Meganck, Cosmin Lazar, Virginie de Schaetzen, Robin Duque, Colin Molter, Hugues Bersini, Ann Nowé, and David Y. Weiss Solís. inSilicoDb: an R/Bioconductor package for accessing human Affymetrix expert-curated datasets from GEO. *Bioinformatics*, 27(22):3204–3205, 2011.