Simulation Study for *vsn* Add-On Normalization and Subsample Size

Markus Schmidberger [*†] Wolgang Hartmann
Ulrich Mansmann[*]

April 16, 2015

Contents

[*]Division of Biometrics and Bioinformatics, IBE, University of Munich, 81377 Munich, Germany
[†]Email: schmidb@ibe.med.uni-muenchen.de

1 Introduction

Variance Stabilization Normalization (VSN) is a model-based method to preprocess microarray intensity data [2, 1]. The method uses a robust variant of the maximum-likelihood estimator for the stochastic model of microarray data described in the references. The model incorporates data calibration (normalization), a model for the dependence of the variance on the mean intensity, and a variance stabilizing data transformation. The method is available in the Bioconductor package *vsn* and is implemented for normalising microarray intensities, both between colors within array, and between arrays.

This simulation study analyzes the influence of different numbers of subsamples and of add-on normalization to the normalized intensities. Biased preprocessing results will be demonstrated with two public available data sets.

2 Data

For this simulation study two public available data sets from the ArrayExpress database (`http://www.ebi.ac.uk/microarray-as/ae`) were used.

E-GEOD-11121: Transcription profiling of human breast cancer cohort reveals the humoral immune system has a key prognostic impact in node-negative breast cancer.
Data: 200 HG-U133A microarray chips.
Citation: The humoral immune system has a key prognostic impact in node-negative breast cancer. Marcus Schmidt, Daniel Böhm, Christian von Törne, Eric Steiner, Alexander Puhl, Henryk Pilch, Hans-Anton Lehr, Jan G Hengstler, Heinz Kölbl, Mathias Gehrmann. Cancer Res 68(13):5405-13 (2008)

E-GEOD-12093: Transcription profiling of human breast cancer samples that were treated with tamoxifen identifier a 76-gene signature defines high-risk patients that benefit from adjuvant tamoxifen therapy.
Data: 136 HG-U133A microarray chips.
Citation: The 76-gene signature defines high-risk patients that benefit from adjuvant tamoxifen therapy. Sieuwerts Zhang, Casey McGreevy, Paradiso Cufer, Span Harbeck, Crowe Hicks, Budd Tubbs,

Sweep Lyons, Schittulli Schmitt, Talantov Golouh, Foekens Wang. Breast Cancer Res Treat – (2008)

To get the raw data (CEL files) of the experiments the *ArrayExpress* package can be used or the files can be downloaded directly from the ArrayExpress web interface.

```
> library(ArrayExpress)
> getAE(c('E-GEOD-11121','E-GEOD-12093'), extract=FALSE)
```

For quality control the *arrayQualityMetrics* package was used. This package uses six different methods for quality control and creates a qualtiy report with a summary table into a HTML output file .

```
> library(arrayQualityMetrics)
> celDir <- 'PATH'
> celf <- list.celfiles(celDir, full.names=T)
> ab <- read.affybatch(celf)
> arrayQualityMetrics(ab, outdir='QA')
```

Arrays with more than two potential problems (three stars or more in the summary table from the *arrayQualityMetrics* report) were excluded from the study. In the dataset 'E-GEOD-11121' nine arrays and in 'E-GEOD-12093' eight arrays were removed (see Table 1).

E-GEOD-11121-raw-cel-1679201568.cel	E-GEOD-12093-raw-cel-1681274912.cel
E-GEOD-11121-raw-cel-1679200783.cel	E-GEOD-12093-raw-cel-1681275553.cel
E-GEOD-11121-raw-cel-1679200049.cel	E-GEOD-12093-raw-cel-1681276139.cel
E-GEOD-11121-raw-cel-1679199817.cel	E-GEOD-12093-raw-cel-1681276420.cel
E-GEOD-11121-raw-cel-1679199644.cel	E-GEOD-12093-raw-cel-1681276775.cel
E-GEOD-11121-raw-cel-1679198997.cel	E-GEOD-12093-raw-cel-1681276985.cel
E-GEOD-11121-raw-cel-1679198126.cel	E-GEOD-12093-raw-cel-1681277333.cel
E-GEOD-11121-raw-cel-1679197959.cel	E-GEOD-12093-raw-cel-1681277423.cel
E-GEOD-11121-raw-cel-1679197894.cel	

Table 1: List of removed arrays due to quality problems identified with the *arrayQualityMetrics* package.

3 Number of Subsamples

As default for `AffyBatch` objects the `vsn2()` function uses 30.000 subsamples (rows=probes) for the calculation of the model parameters using the maximum-likelihood estimator. However on current chip types there are more than 500.000 probes and therefore in default less than 5% will be used for the estimation of the model. Long computation times and the main memory requirements are the main reasons for the sample size reduction.

```
> library(vsn)
> fit <- vsn2(ab, subsample=100000)
> x <- predict(fit, newdata=ab)
```

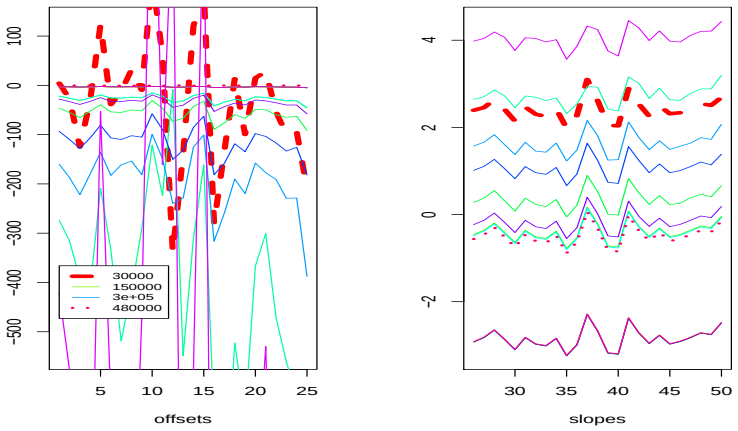
Figure 1 plots the values of the offset and slope parameters for 25 and 50 arrays and different numbers of subsamples. The dashed red line is the default subsample with 30.000 rows and the dotted pink line indicates the parameters using all probes. Depending on the number of subsamples the parameters are changing very strong. The default subsample is a good approximation for several other subsamples, but there are strong outliers too. This indicates that in a bad case the model parameters will be estimated on an unsuitable subsample.

Figure 2 shows the boxplots of the intensities for every array. Blue boxes are calculated with 30.000 subsamples (default) and red boxes with 420.000 subsamples. The boxplots reflect a good variance stabilization for all arrays. But due to the differences in the estimated parameters (see Figure 1) there are strong differences in the location of the means and in the outliers.

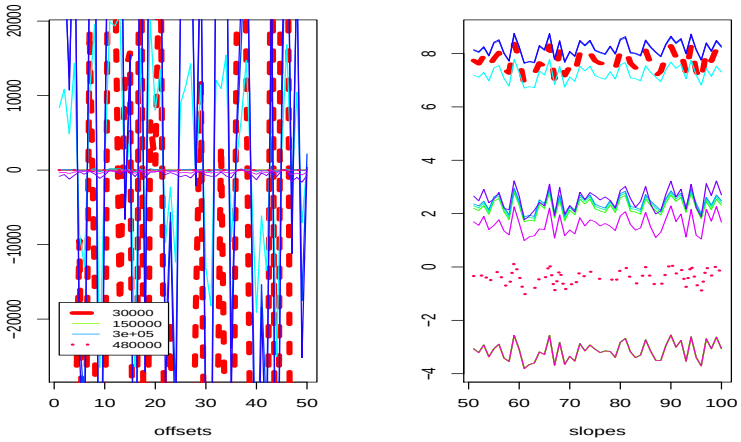
4 Add-On Normalization

Add-On normalization allows to normalize a new microarray with respect to a normalization performed for a set off arrays. The *vsn* package offers this technique which is needed to apply gene expression profiles for classification or prognosis to a new patient.

In the simulation study 20 arrays were normalized with `vsn2()`. After this two times 20 arrays were added using the vsn add-on normalization. Additionally the same 60 arrays were normalized with vsn together (using the same subsample) and than compared to the results of the add-on normalization.

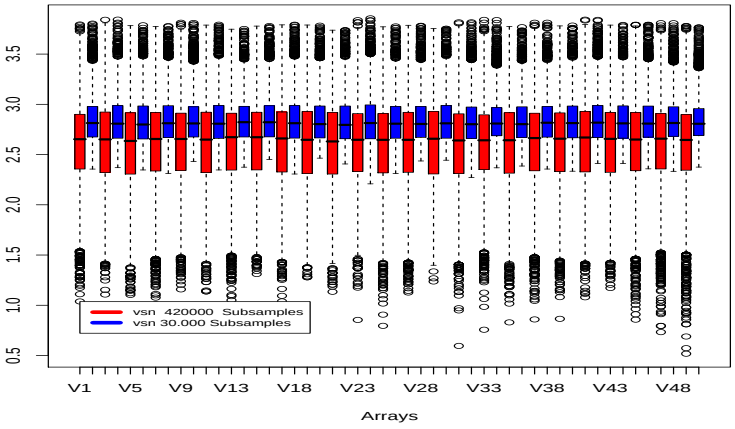


(a) E-GEOD-11121

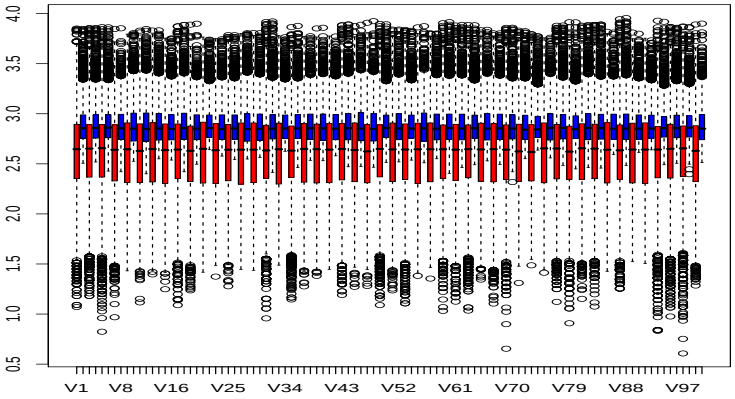


(b) E-GEOD-12093

Figure 1: Values of the vsn coefficients for different numbers of subsamples.



(a) E-GEOD-11121



(b) E-GEOD-12093

Figure 2: Boxplots of intensities after vsn normalization with two different numbers of subsamples.

```
> set.seed(1234)
> fit_ref <- vsn2(ab[,1:20])
> fit_add1 <- vsn2(ab[,21:40], reference=fit_ref)
> fit_add2 <- vsn2(ab[,41:60], reference=fit_ref)
> set.seed(1234)
> fit_comp <- vsn2(ab)
```

Figure 3 shows the influence of the add-on normalization to the model parameters. Blue lines are the parameters for the complete vsn normalization and red lines for the reference and add-on normalization. For both data sets there are strong differences between the parameters, but they have the same structure.

Figure 4 plots the boxplots of the intensity values. The boxplots reflect a good variance stabilization for all arrays. But due to the differences in the estimated parameters (see Figure 3) there are strong differences in the location of the means and in the outliers.

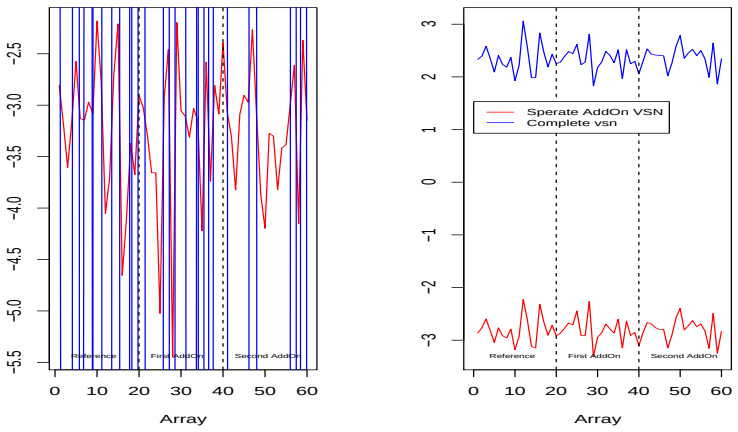
There is no influence from the number of add-on batches, because for every add-on array the same reference data (parameters) will be used. The more add-on data used the bigger the difference between add-on (+ reference) and complete vsn noramlization gets.

5 Conclusion

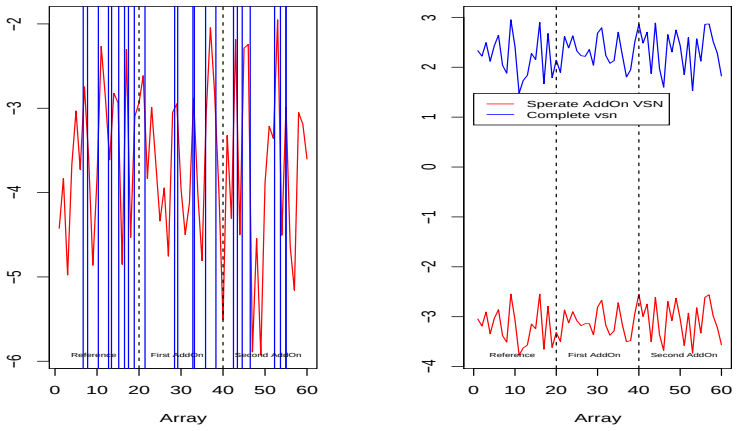
Same results can be found using more arrays and other data sets. The results are not shown due to overflow in the graphics using more than 150 arrays.

Using more than 30.000 subsamples increases the computation time but in theorey increases the accuracy of the statistical model too. A misuse of the add-on technique to normalize many patients to a base set of a few patients results in a biased preprocessing result. The number of add-on arrays should be not bigger than the number of reference arrays.

Therefore the power of parallel computing or the *affyPara* [3] is required to use vsn normalization with more than 30.000 subsamples and to normalize all microarray data together.

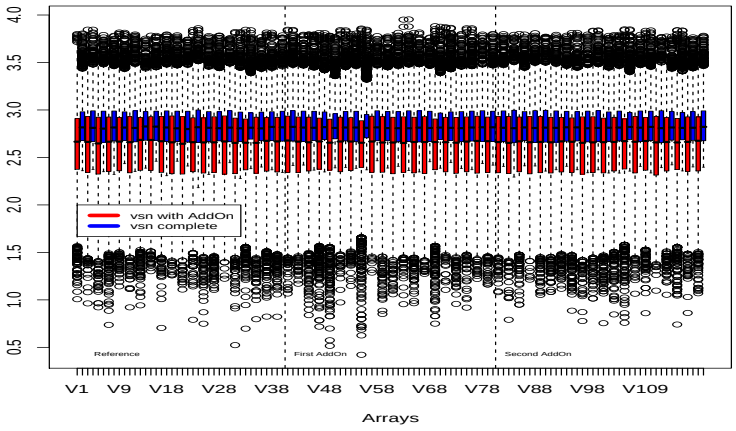


(a) E-GEOD-11121

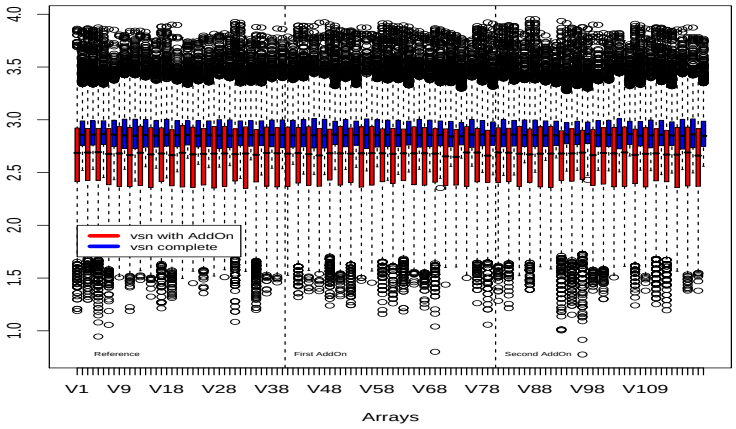


(b) E-GEOD-12093

Figure 3: Values of the vsn coefficients compared for complete vsn normalization and add-on normalization with 20 reference arrays and two times 20 arrays add-on.



(a) E-GEOD-11121



(b) E-GEOD-12093

Figure 4: Boxplots of intensities for complete vsn normalization and vsn Add-On normalization.

References

[1] Wolfgang Huber, Anja von Heydebreck, Holger Sueltmann, Annemarie Poustka, and Martin Vingron. Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical Applications in Genetics and Molecular Biology*, 2(1):3, 2007.

[2] Wolfgang Huber, Anja von Heydebreck, Holger Sültmann, Annemarie Poustka, and Martin Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:S96–104, 2002.

[3] Markus Schmidberger and Ulrich Mansmann. Parallelized preprocessing algorithms for high-density oligonucleotide array data. In *22th International Parallel and Distributed Processing Symposium (IPDPS 2008)*, 2008.