6

A novel and fast normalization method for high-density arrays

Maarten van Iterson[1], Floor A.M. Duijkers[2], Jules P.P. Meijerink[2], Pieter Admiraal[2],Gert-Jan B. van Ommen[1], Judith M. Boer[1,3], Max M. van Noesel[2] and Renée X. Menezes[2,4]

[1] *Center for Human and Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands*
[2] *Department of Pediatric Oncology/Hematology, Erasmus University Medical Center-Sophia Children's Hospital, Rotterdam, the Netherlands.*
[3] *Netherlands Bioinformatics Centre, Nijmegen, the Netherlands*
[4] *Department Epidemiology and Biostatistics, VU Medical Center, Amsterdam, the Netherlands*

***Statistical applications in genetics and molecular biology*, 11(4):1–31, 2012.**

Abstract

Background: Among the most commonly applied microarray normalization methods are intensity-dependent normalization methods such as lowess or loess algorithms. Their computational complexity makes them slow and thus less suitable for normalization of large datasets. Current implementations try to circumvent this problem by using a random subset of the data for normalization, but the impact of this modification has not been previously assessed. We developed a novel intensity-dependent normalization method for microarrays that is fast, simple and can include weighing of observations.

Results: Our normalization method is based on the P-spline scatterplot smoother using all data points for normalization. We show that using a random subset of the data for normalization should be avoided as unstable results can be produced. However, in certain cases normalization based on an invariant subset is desirable, for example, when groups

of samples before and after intervention are compared. We show in the context of DNA methylation arrays that a constant weighted P-spline normalization yields a more reliable normalization curve than the one obtained by normalization on the invariant subset only.

Conclusions: Our novel intensity-dependent normalization method is simpler and faster than current loess algorithms, and can be applied to one- and two-colour array data, similar to normalization based on loess.

Availability: An implementation of the method is currently available as an R package called *TurboNorm* from BioConductor.

6.1 Background

Intensity-dependent normalization methods are among the most commonly used methods for microarray normalization. Most of these methods are based on the lowess or loess algorithms of Cleveland [Cleveland and Devlin, 1988, Cleveland and Grosse, 1991]. A wide variety of loess-based normalization methods exists [Bolstad et al., 2003, Irizarry et al., 2008, Risso et al., 2009].

Since the density of microarrays is still expanding, increasingly large datasets are being produced. Normalization of these datasets can take a considerable amount of time and can result in memory usage problems. The widely used R `loess` [R Development Core Team] implementation is not specifically designed for the normalization of large sets of high-density microarrays. A common approach to overcome long run times involves computing the loess trend using a random subset of the data, thereby artificially reducing the size of the problem. We will show here that subset choice may have a large impact on the results, an issue that has not been assessed previously.

Normalization methods that rely on a random subset of the data assume most points are invariant between samples and, the bulk of points represent a baseline. In some cases, however, this assumption does not hold, as for example when samples are studied before and after exposure to a compound that affects most of the points. A common solution is the use of a subset of invariant features, provided it spans the entire intensity range. However, such a solution is not always possible as invariant probes are not available or they do not cover the complete intensity range.

Here, we present a new normalization method that uses invariant features without requirement to cover the complete intensity range. The method compensates for unequal coverage by using weights for all features, while the invariant features are given a higher weight. Our method is an extension of the P-spline scatterplot smoother from Eilers and Marx [1996] and can be used in general as an intensity-dependent normalization method applicable to one- and two-colour array data.

The P-spline scatterplot smoother is a special kind of penalized spline regression model for nonparametric regression. It uses equally spaced knots and a discrete penalty, similar to the approach of O'Sullivan [1986] for solving ill-

posed inverse problems using regularization and B-splines. Penalized splines have become a standard tool in scatterplot smoothing and semiparametric regression (see e.g. Ruppert et al. [2003], or the recent review by Eilers and Marx [2010].

We compare our P-spline normalization with lowess/loess-based normalization methods on a selection of the MAQC data described by Shi et al. [2006]. We then go on to show that using loess on a random subset of the data yields unstable results, and thereby potentially interesting genes could be missed. Subsequently, we apply the weighted P-spline normalization method to an array-based DNA methylation study, in which one group of samples has a baseline shift due to treatment with a demethylating agent. The advantages of the weighted normalization using invariant probes are shown in a simulation study. Run time and memory usage analysis on high-density NimbleGen tiling arrays [Heintzman et al., 2009] show that the P-spline normalization outperforms `loess`.

6.2 Methods

Introduction

Intensity-dependent normalization methods assume that the majority of the genes are unaffected between the conditions being studied. An easy way to visually verify this assumption is by using the ratio-intensity plot (RI-plot, aka MA-plot). The $\log_2$ intensity ratio (M-values) of two samples in case of single-colour microarray data, or the $\log_2$ intensity ratio of the red and green channel in case of two-colour data, are plotted against the average $\log_2$ intensities (A-values). Ideally, the majority of the data points should be centered around $M = 0$, but due to systematic errors a non-linear intensity-dependent trend is usually observed. This non-linear trend is corrected by fitting a non-linear function to the MA-plot ($M = f(A)$) and subtracting it from the observed M-values. Examples of methods that can find this non-linear trend, $f(A)$, are the lowess- and loess-algorithms [Cleveland and Devlin, 1988, Cleveland and Grosse, 1991].

These scatterplot smoothers of Cleveland are flexible, robust and versatile non-linear smoothers [Cleveland and Devlin, 1988, Cleveland and Grosse, 1991]. The older lowess is a robust iteratively locally weighted linear regression smoothing method. The more flexible loess algorithm is able to smooth regression surfaces either locally linear or quadratic as well and can include point specific weights.

Normalization based on P-splines

The P-spline smoother introduced by Eilers and Marx [1996] is a combination of B-splines with a difference penalty on the regression coefficients. P-splines

belong to the family of penalized splines using B-spline basis functions, where the penalization is on the curvature of the smoothed function. For the P-splines of Eilers and Marx [1996], a discrete approximation to the integrated squared second derivative of the B-splines is made. This results in an easy-to-construct penalty matrix. Additionally, by using B-splines of degree zero (piecewise constants) with equally-spaced knots allows us to construct the B-spline basis very efficiently. These properties make P-splines attractive and fast smoothers.

Here we introduce a weighted P-spline smoother. Let y represent the M-values and x the A-values, then the weighted non-linear trend, $y = f(x)$, is estimated by the weighted P-spline smoother, $y = WB\hat{\alpha}$, where B is the B-spline basis matrix constructed from x and W is a diagonal matrix of weights. Estimates of the coefficients, α, are obtained by minimizing the objective function:

$$Q = (y - B\alpha)'W(y - B\alpha) - \lambda\alpha' D'D\alpha. \tag{6.1}$$

Here D represents the matrix operator for the second-order differences. The basis for B-splines of degree zero is constructed as follows: define k equally-spaced knots on the range of x, then each row of $B_{n\times k}$ has the value 1 in only one interval between two knots and zero everywhere else. In practice $B'B$ and $B'y$ can be calculated directly without constructing the large basis B. As, $B'B$ is a diagonal matrix with each diagonal element equal to the sum of a column of B, which is just the number of x's that fall into an interval between two knots. Similarly, each element of $B'y$ is the sum of all y that correspond to x's in the interval between two knots.

We propose a weighted P-spline normalization method that uses invariant features without the requirement to cover the complete intensity range. The invariant feature set is given weight, $w = {}^{n}/_{m}$, while the remaining features have weight, $w = 1$. The rationale is that the invariant feature set, of size m, and remaining features, of size $n - m$, contribute equally to the smoothed curve. It then follows that $(n - m) = wm$ and the moderated data-dependent weight is given by $w = {}^{n}/_{m}$ assuming that the number of invariant features is much smaller than the remaining features. In this way, regions that are not covered by the invariant features still contribute to the fitted curve.

In order to show the role played by these weights, consider the estimated coefficients of the weighted P-spline smoother:

$$\hat{\alpha} = (B'WB + \lambda D'D)^{-1}B'Wy. \tag{6.2}$$

Define the B-spline basis matrix $B_{n\times k}$, with n data points and k knots as a block matrix $[B_{(n-m)\times k}|B_{m\times k}]'$ with $B_{m\times k}$ the B-spline basis matrix for the invariant features and $B_{(n-m)\times k}$ for the remaining features (reordering the B-spline basis matrix by rows has no influence on the estimated regression

${}^{(n-m)}/_{m} = {}^{n}/_{m} - 1 \approx {}^{n}/_{m}$

coefficients). Equation 2 can now be rewritten as:

$$\hat{\alpha} = \left(B'WB + \frac{\lambda}{w} D'D \right)^{-1} B'Wy, \tag{6.3}$$

where

$$W = \begin{bmatrix} w^{-1}\mathrm{I}_{n-m} & 0 \\ 0 & \mathrm{I}_m \end{bmatrix},$$

with I_{n-m} and I_m identity matrices. Observe that if the weight $w \gg 1$, $w^{-1}\mathrm{I}_{n-m}$ will vanish and equation 6.3 simplifies to equation 6.2, with $B = B_{m\times k}$, $W = I$ and $y = (y_1, \ldots, y_m)$. Note also that the penalty parameter will absorb the weight, as it is calculated *a posteriori*. Thus a *biweighted* P-spline smoother with weights $(1, w)$, where $w \gg 1$, is identical to a P-spline smoother on the invariant subset. Thus, our proposed weigthed P-spline normalization with moderated data-dependent weight is a compromise between normalization on all features and normalization based only on the invariant features.

Description of data

We make use of three experimental datasets: 1) a selection of the MAQC data [Shi et al., 2006] for comparing the P-spline normalization with lowess and to assess the effect of using a subset for normalization, 2) data from an epigenetics experiment to show the performance of the constant weighted P-spline normalization method and 3) a selection of high-density NimbleGen tiling array data described by Heintzman et al. [2009] for a run time and memory usage analysis. Simulated data is used to show the advantages of the weighted P-spline normalization.

MAQC data:

The selected data contained two different RNA sources, namely Universal Human Reference RNA from Stratagene (source A) and a Human Brain Reference RNA from Ambion (source B), each with 5 replicates. A total of 10 whole human genome oligo Agilent two-colour microarrays (G4112A) containing 43,931 probes produced by the same laboratory were selected (GEO Edgar et al. [2002] accession numbers from GSM128989 to GSM128998).

CpG island methylation array data:

The 244K CpG island methylation arrays were used to measure methylation levels in six neuro-ectodermal cell lines (NMB, SK-N-MC, CHP-100, IMR-32, AMC106c and SK-N-FI). These cell lines were treated with a demethylating agent (5-aza-2'-Deoxycytidine (DAC), Calbiochem, San Diego, USA) and a histone deacetylase inhibitor (Trichostatin A (TSA), Sigma-Aldrich, Zwijndrecht,

The Netherlands). Cells were treated for 72 hours with 30 nM DAC and 25 nM TSA was added during the last 48 hours. Culture conditions and DNA isolation procedures are described in Appendix D. Differential methylation hybridization (DMH) was used for genome-wide methylation screening and is further described in Appendix D.1, Figure D.1.

Invariant subset

For normalization purposes an invariant subset of probes was required. We selected fragments without methylation-sensitive restriction sites after *Mse*I digestion. The probes (984) on these fragments are not influenced by differences in methylation (Figure D.2). Therefore, the $\log_2$ intensity ratio for these fragments is expected to be 0. The weighted normalization, especially the attribution of control probes and all other probes is further explained in the section about weighted P-spline normalization.

Simulation

In this simulation study, weighted intensity-dependent normalization is treated as a nonlinear fitting or scatterplot smoothing problem, $y = f(x)$, where y and x represent the M- and A-values respectively. Data is generated according to:

$$y_i = f(x_i) + y_0 1_{\{x_i \notin \mathcal{X}\}} + \epsilon_i \quad i = 1, \cdots, n. \tag{6.4}$$

The nonlinear trend is represented by $f(x) \propto x(1-x)^4$, where x is sample uniformly at random from $[0, 1]$. A baseline shift y_0 is added to those points that are not in the invariant set ($\mathcal{X}$) and, for simplicity, standard white noise is assumed for the error term $\epsilon \sim \mathcal{N}(0, 1)$. With this simple setup, we try to mimic the weigthed normalization process of example 2 and focus on the effect of weighing invariant probes. Here we used $n = 1000$ points, of which 10% is consider to be invariant of the treatment effect. Figure D.1 shows two simulation scenario's: a) one where the invariant set is restricted to a small range of x and b) one where the invariant set covers the whole ranges of x.

NimbleGen tiling array data:

For the run time and memory usage analysis we chose the high-resolution oligo microarray data described by Heintzman et al. [2009] containing 391K probe sets. Heintzman *et al.* developed a chromatin-immunoprecipitation-based microarray method (ChIP-chip) to locate promoters, enhancers and insulators in the human genome. From the ChIP-chip data we arbitrarily selected the NimbleGen tiling array that contained chromatin immunoprecipitated DNA from unstimulated HeLa cells using H3ac antibody (GEO accession number GSM144308).

6.3 Software

All data analyses was performed in R version 2.10.0 [R Development Core Team]. Several packages from BioConductor version 2.5 [Gentleman et al., 2004] were used. The effect of demethylating treatment on methylation levels was estimated using the *limma* package [Smyth, 2005]. Because of large differences between the cell lines, the limma-model included the factor 'cell line pair' to evaluate the treatment effect properly. Significance was defined as an FDR $< 5\%$ after multiple testing adjustment by the Benjamini-Hochberg method.

From *limma* we used the `loessFit` interface in all normalizations, as both fast lowess and loess implementations were available. The faster lowess is called by default and, when weights are supplied, loess is called.

Our implementation of the weighted P-spline method is available as an R package called *TurboNorm* from BioConductor. The package contains, besides a `loess`-like function for scatterplot smoothing, wrapper functions for both single and two-colour microarray data normalization. It also contains a function for adding the fitted curves on plots produced by the *lattice* package [Sarkar, 2009] (see for examples figures 6.1 and 6.5). See Appendix A for the vignette of the package. For profiling of the run time and memory usage, the R `Rprof` function as described by Gentleman [2009] was used. All calculations were run on an Intel®Core™2 Duo CPU 3.00GHz processor, 40GB and with an Ubuntu 9.10 64-bit architecture.

6.4 Results

Comparison with lowess

In order to show that normalization based on lowess or the P-spline method are comparable, we took a subset of samples from the MAQC data [Shi et al., 2006] and applied both normalization methods. Figure 6.1 shows the MA-plots of four technical replicates of the MAQC data, containing RNA source A. There is good agreement between lowess and the P-splines normalization as the fitted curves largely overlap.

Normalization using a random subset of probes

Here, we asses the effect of using a random subset of probes from the data for normalization. We took the same subset of samples from the MAQC data as described above. The loess-fitted curve based on all probes were compared with loess-fitted curves based on several randomly selected subsets of probes. The use of a randomly selected subset resulted in a large fluctuation of the loess-fitted curves around the $M = 0$ line, particularly at the boundaries (Figure 6.2).

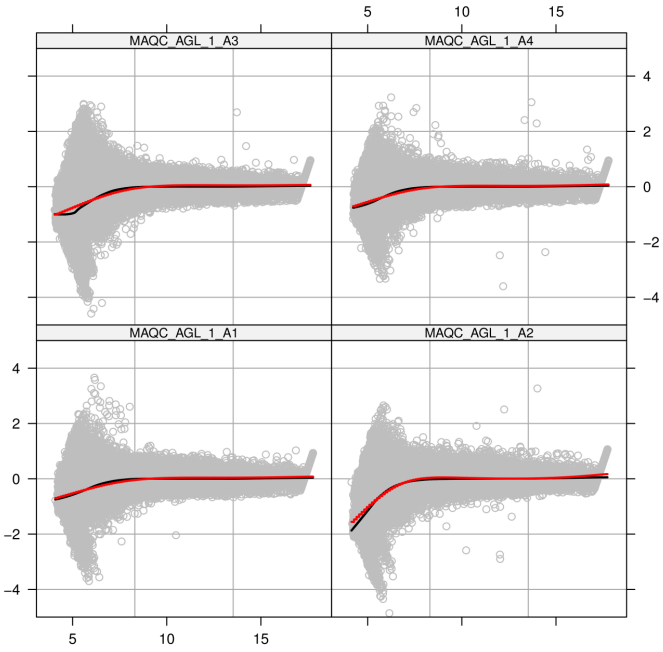

Figure 6.1: **MA-plots of four technical replicates of the MAQC data, containing RNA source A:** The black line represents the lowess curve, the red line the curve based on normalization using the P-spline method.

Figure 6.3 shows the influence of the randomly chosen subsets of probes on the significance of differentially expressed genes using a volcano plot. An empirical Bayes linear regression model [Smyth, 2005] was fitted to compare RNA sources A and B. The largest changes, as indicated by the length of the arrow, occur at the 'main vent of the volcano', as genes with fold-changes in this area are mostly affected by changes of the fitted curve. This phenomenon was observed with other probe subsets but affecting specific genes for each subset considered (Figure D.2).

The overlap of significant differentially expressed up- and down-regulated genes between normalization using subsets with that of normalization using all probes is shown in table 6.1. If 1% of the data is used, there is an overlap of around 90% for both up- and down-regulated genes with the normalized data based on all probes. In other words, there may be an increase of false positives and false negatives when a subset is used. When 10% or more of all the probes is used, the overlap is at least 98%. We should point out, however, that these

results are data-dependent and, in particular, the MAQC data is likely to yield higher overlaps than in most studies, as it only involves technical variability.

Table 6.1: The percentage overlap of significant differentially expressed up- and down-regulated genes between normalization using subsets of different sizes with that of normalization using all probes. A 5% FDR threshold was used on the Benjamini-Hochberg corrected p-values.

	subset (%)	0.1	0.5	1	5	10	25
up		62	84	89	96	98	99
down		71	87	93	94	98	99

Normalization of CpG island methylation array data using an invariant subset

Normalization based on a subset of the data is sometimes necessary, to see this we will use the CpG island dataset and demonstrate the use of the weighted P-spline normalization method. Treatment of cell lines with demethylating agents resulted in a general decrease of methylation levels between untreated and treated cell line pairs (see raw signals, Figure 6.4, upper panel). Conventional array normalization methods such as lowess and variance stabilization normalization [Huber et al., 2002] or the P-spline method without using weights did not preserve the demethylating treatment effect (Figure 6.4 or D.3 middle panel, Figure 6.5 left panel). Indeed, after lowess normalization we found a significant decrease of methylation for only 1.7% of the probes and an increase of methylation for 2.2% of the probes. Results were similar for unweighted P-spline normalization. Clearly this does not reflect the decrease of overall methylation level after treatment. It instead results from a violation of the assumption, made by these as most normalization methods, that the bulk of the data is invariant between samples.

To correct this, we used a subset of invariant probes, that could be deduced from the arrays used (see Description of datasets: *Invariant subset* and Figure D.1). Normalization based on this subset resulted in a significant increase of methylation for 0.7% of the probes and a decrease for 69.9%, after weighted P-spline normalization. Similar results can be observed for loess in the right panel of Figure 6.5 (see also Figure D.3). The invariant probes' intensity distribution was optimal in the middle intensity range, while only few invariant probesets represented the highest or lowest intensity ranges. This diminishes the estimates's reliability on the boundaries (Appendix, Figures D.4 and D.5) and led us to apply a weighted version of the P-spline as well as the loess normalization method. Weights were assigned in such a way that invariant probes received

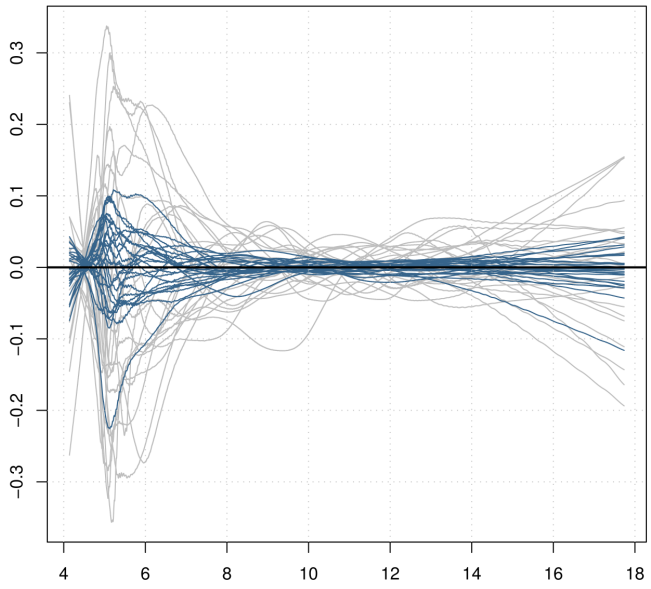

Figure 6.2: **Differences between normalized M-values using loess on all points and loess on a random subset:** The curves represent differences between the loess-fitted curve on all data points with that of loess curves for 25 randomly selected subsets of the data with sizes of 10% (dark blue) or of 1% (grey) of the original number of probes for a single Agilent array of the MAQC data. On the y-axis the difference between normalized M-values are displayed, and on the x-axis the intensity (A-values).

a higher weight than the other probes, such that the total influence was kept equal to that of the remaining probes (see Methods). This weighted normalization had a superior coverage in the boundary regions than normalization on only the invariant probes (Figure 6.6). In Figures 6.4 and 6.5, we show that the weighted P-spline normalization retains the biological difference. This resulted in decreased methylation for 36.4% of the probes and increased methylation for only 1.1%. In comparison, a weighted loess normalization showed similar results with a significant decrease of methylation for 32.6% of the probes and an increase for 1.2% of the probes after the combined epigenetic treatment.

This data was validated by LUMA analysis for global methylation levels and methylation specific PCR for gene-specific methylation (Appendix, D). In five

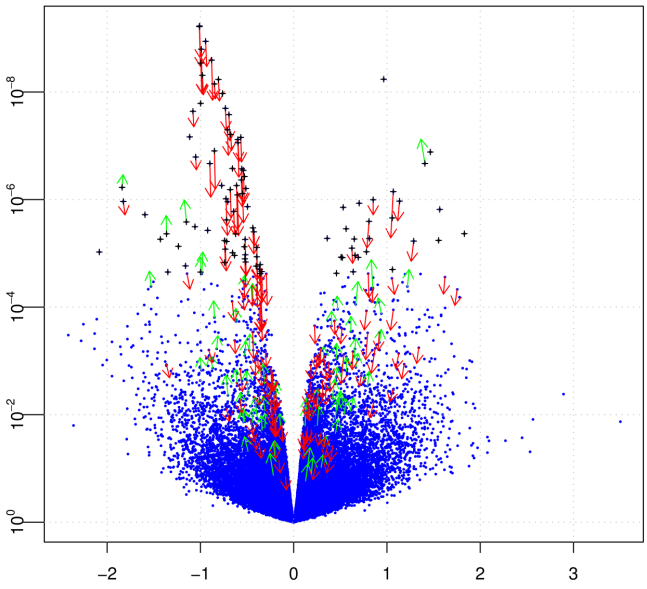

Figure 6.3: **Volcano plot for the selection of the MAQC data:** The volcano plot shows the difference between loess normalization on a subset (1% of the probes), with a loess normalization using all probes. The volcano plot shows $-\log_{10}$(P-value) on the y-axis and the fold-change on the x-axis. The 0.5% of the largest changes (euclidean distance) are indicated by arrows, with increasing significance from subset to the full data indicated by green arrows and decreasing significance in red. The top 100 of most significant probes are indicated by + sign.

cell lines, the average demethylation was 9.3%, which ranged between no change in methylation (AMC106c and NMB) to more than 20% demethylation (CHP-100) (Figure 6.7). Additionally, gene specific demethylation was validated by an unmethylated specific PCR (USP) of two known hypermethylated genes. Figure D.5 (of the Appendix) shows an increase of unmethylated product after demethylating treatment.

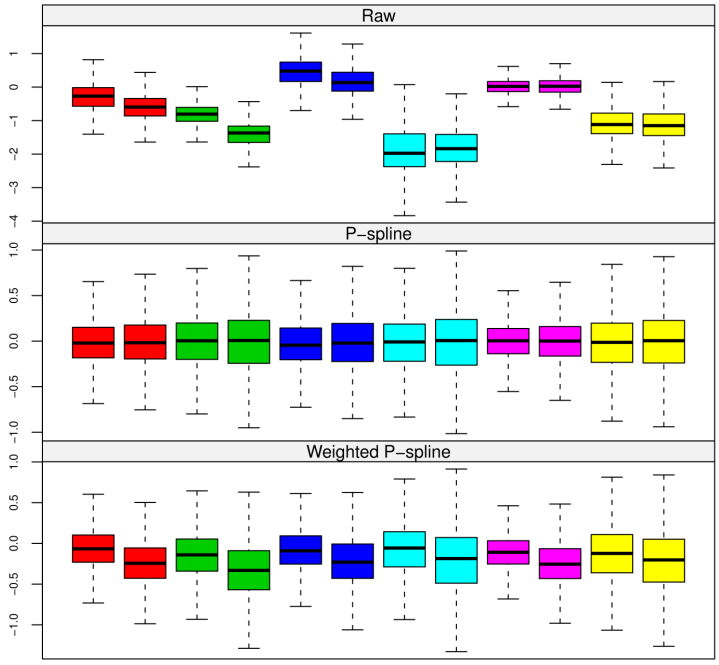

Figure 6.4: **Box and whisker plots of the raw, P-spline and weighted P-spline normalized methylation data:** On the x-axis six cell line pairs NMB, SK-N-MC, CHP, IMR, AMC and FISK (from left to right) are displayed. Each pair consist of the untreated and treated cell line where the treatment was with Decitabine and Trichostatin A. The y-axis displayed the methylation ratio. Note the different scale of the y-axis for the highest panel.

Simulation

Both `loess` and the P-spline smoother were used to obtain smoothed fits, with and without weights. Moderated data-dependent weights were used as described in Section 6.2. As expected, the unweighted fits follow the noninvariant trend (solid black line) as they are much more abundant. However, the weighted fits capture much better the invariant trend (dashed black line). In scenario a, where the invariant probes are restricted to a certain range, the fitted curves outside this range follow the noninvariant trend. Both `loess` and P-spline perform well.

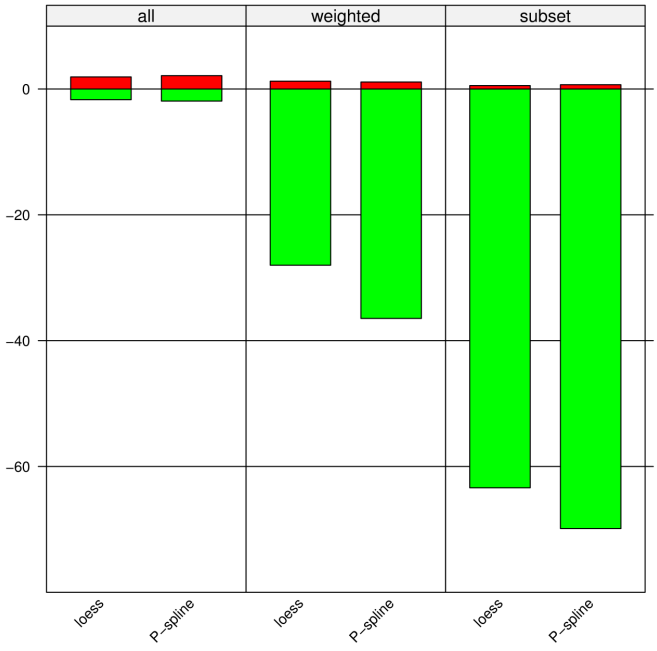

Figure 6.5: **Methylation changes with different normalization methods:** The bar plots show the percentages of probes that significantly decrease or increase in methylation after demethylating treatment using different normalization methods. These were loess or the P-spline method using all probes (left panel) weighting the invariant probes (middle panel) and using the invariant probes only. Decreased methylation is indicated in green and increased methylation in red.

Run time and memory usage analysis

Run time and memory usage analysis were performed on normalization of the NimbleGen tiling array data involving 391K probe sets, described by Heintzman et al. [2009]. They used previously loess-based intensity-dependent normalization on these high-resolution oligo microarray data [Heintzman et al., 2007].

Random subsets of increasing size were drawn, on which a normalization was performed and run time and memory use was profiled. Both lowess and P-spline outperformed loess in terms of speed, with the P-spline algorithm the fastest. The normalization of one complete array took a fraction of a second using lowess and the P-spline algorithm, but more than 4s for loess (Figure 6.8).

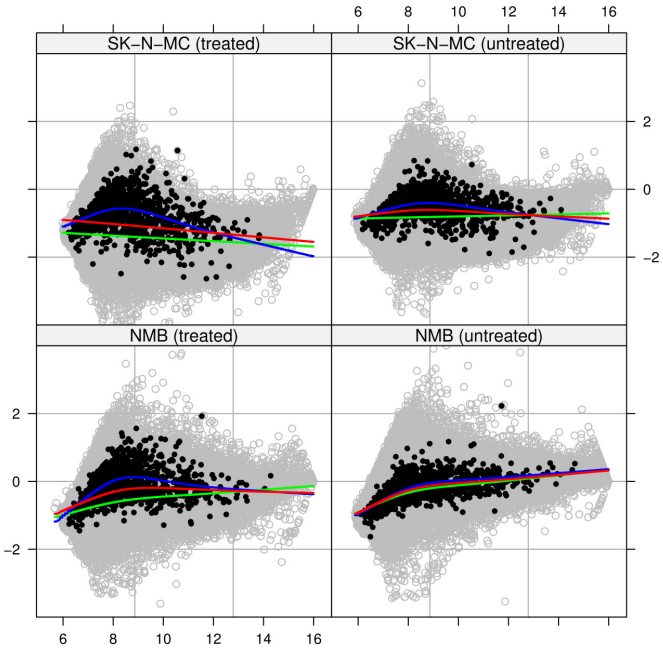

Figure 6.6: **MA-plots for two cell lines, NMB and SK-N-MC:** Each cell line is hybridized with and without treatment of Decitabine and Trichostatin A. The green curve is computed using all points, the blue curve uses only the invariant probes, and the red curve uses the invariant probes with larger weights (≈ 250) than the remaining probes.

On memory requirements, lowess and the P-spline agorithm are comparable in memory usage, using around 10Mb in memory for normalization of the whole array, whereas the loess memory usage were 5 times higher. This difference can be explained by the computational complexity of the algorithms used.

The run time complexity of the P-spline algorithm is approximately $\mathrm{O}(N + M^3)$, since construction of the piecewise constant basis takes $\mathrm{O}(N)$ and solving the reduced linear system of equations $\mathrm{O}(M^3)$. Here M is the number of knots (by default 100 in our implementation) and N is the number of data points. On the other hand, lowess and loess are at least of $\mathrm{O}(MN^3)$ because on M carefully chosen locations a least-square fit of $\mathrm{O}(N^3)$ is performed, with M the number of equidistant intervals in case of lowess and the number of bins in the case of loess. Lowess uses as default equidistant intervals of size $1/100$ of the range of the data and thus $M = 100$. Loess uses a data-dependent number of

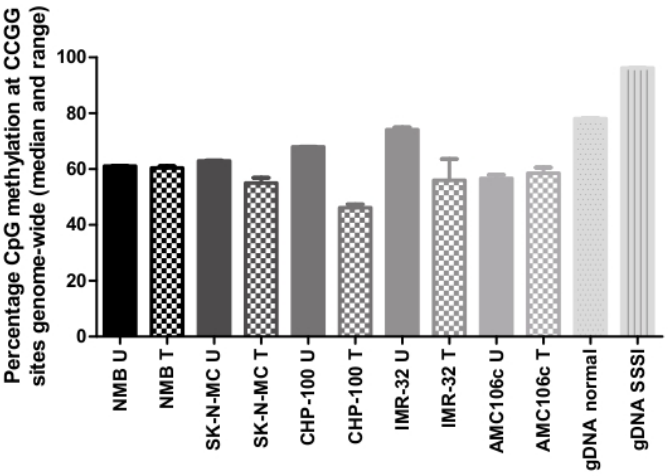

Figure 6.7: **LUMA for measurement of global DNA methylation levels:** This barplot shows global methylation levels at HpaII restriction sites for normal and SSSI treated controls and five neuro-ectodermal cell lines untreated (U) and treated (T) with Decitabine and Trichostatin A. The digestions and measurements were performed in duplicate.

bins, for which it is difficult to give an estimate of the number of bins used.

6.5 Discussion

We showed that our weighted P-spline normalization method is an efficient approach for intensity-dependent microarray normalization. It is also as versatile as the loess algorithm of Cleveland and Grosse [1991], being applicable to one- and two-colour array data, but with lower computational complexity. The P-spline normalization method yields considerably faster implementations, avoiding the need for subset-based normalization and allowing for normalization using a set of invariant features via weights.

Normalization methods that are based on loess frequently use a subset of the data of which, to our knowledge, the effect has not yet been assessed. The cyclic loess normalization method [Bolstad et al., 2003] as implemented in the *affy* package calls the R `loess`-function, by default using 5000 randomly chosen probes, although this number can be altered by the user as it is soft-coded. The loess normalization method as implemented in the *marray* package, on the other hand, calls the R `loess`-function with a fixed, hard-coded, number of 2000 randomly chosen probes as subset. Here we have shown that the results of such a normalization are unstable. When subsets are randomly chosen they

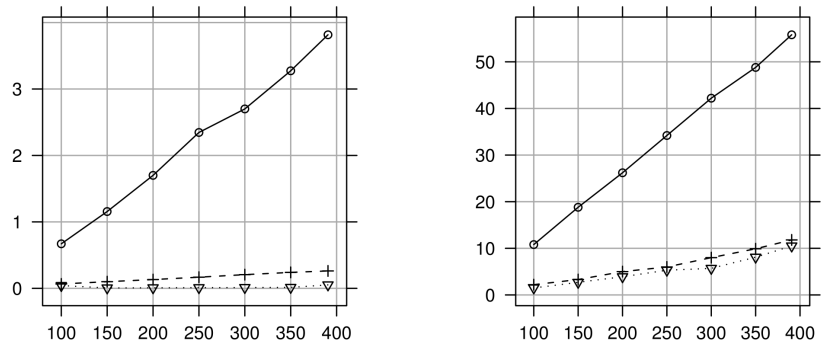

Figure 6.8: **Performance of different implementations of lowess/loess and the P-spline method:** Run time was measured in seconds (left-panel) and memory usages in Mb (right panel) as function of the number of probes in the data (x-axis $\times 1000$) for P-spline ($\bigtriangledown$), lowess ($+$) and loess ($\circ$).

may not represent the entire dataset well, either because they do not span the entire data range or because they do not represent the trend. In particular, the less dense boundary regions of the MA-plot tend to be under-sampled. In an example using the MAQC data, 10% of the probes yielded already acceptable overlap with normalization based on all probes. In practice, however, most studies involve smaller signal-to-noise ratios than the MAQC data, so a stable subset-based normalization is unlikely to be achieved using a percentage as low as 10%. In addition, even if the lower bound of 10% is considered to suffice, for a 100K array this would involve using 10.000 probes, which is already a large number of probes for current loess implementations.

In some studies, the typical normalization assumption that the majority of the probes remains unchanged across samples does not hold, as was true for our methylation dataset containing samples before and after treatment. The same effect is observed for low-density custom arrays [Oshlack et al., 2007]. Oshlack *et al.* proposed a modification of the lowess algorithm by using probe-specific quantitative weights. Arbitrarily, we chose the weight for the invariant probes versus the remaining probes as the weight that gives both groups of probes an equal attribution in the normalization curve fit. This implies that the estimated trend will be somewhere in between the trends of invariant and noninvariant probes. If desired, larger weights could be used to force the estimated trend closer to the invariant probes trend.

The results of the methylation array experiment after using our new nor-

malization are generally comparable to the demethylating effect found in the validation experiments. The difference in demethylation between the array analysis (36.4% demethylation) and the demethylation measured by a LUMA validation experiment (10%) can be explained by the methodological difference. LUMA is a quantitative measurement of methylation throughout the genome, whereas the array data are relative measurements of methylation in selected gene promoter areas. These differences render an exact comparison of percentages impossible.

Specifically in the context of methylation arrays, others [Ibrahim et al., 2006, Irizarry et al., 2008, Laird, 2010] also recognized that current microarray normalization methods fail to adequately normalize the data. Irizarry et al. [2008] proposed a method, CHARM, to detect methylation difference between samples, which also corrects for CG-content. As their method requires control probes to be present on the array, it is hard to be generalized to other types of data. Our method is much more versatile, as it can be used in any study where lowess/loess is typically used.

The lowess/loess algorithms make use of robust estimation techniques to guard against outliers that can be seen on custom arrays. It is also possible to achieve this type of robustness by adapting the P-spline algorithm, for example by using the L_1 norm, instead if L_2, when writing equation 6.2. Another robustifying modification is to apply iteratively weighted regression using the residuals as weights. These extensions are beyond the scope of this work and will appear elsewhere.

6.6 Conclusions

We propose a versatile and fast intensity-dependent normalization method applicable to one- and two-colour array data. The method has comparable properties to lowess and loess, but its lower computational complexity allows it to be faster and more memory efficient.

6.7 Bibliography

W.S. Cleveland and S.J. Devlin. Locally-weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83:596–610, 1988.

W.S. Cleveland and E. Grosse. Computational methods for local regression. *Statistics and Computing*, 1:47–62, 1991.

B.M. Bolstad, R.A. Irizarry, M. Astrand, and T.P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

R.A. Irizarry, C. Ladd-Acosta, B. Carvalho, H. Wu, S.A. Brandenburg, J.A. Jeddeloh, B. Wen, and A.P. Feinberg. Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Research*, 18(5):780–90, 2008.

D. Risso, M.S. Massa, M. Chiogna, and C. Romualdi. A modified LOESS normalization applied to microRNA arrays: a comparative evaluation. *Bioinformatics*, 25(20):2685–2691, 2009.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org. ISBN 3-900051-07-0.

P.H.C. Eilers and B.D. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121, 1996.

F. O'Sullivan. A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1 (4):502–518, 1986.

D Ruppert, M.P. Wand, and R.J. Carroll. *Semiparametric regression*. New York: Cambridge University Press., 2003.

P.H.C. Eilers and B.D. Marx. Splines, knots and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, n/a, doi:10.1002/wics.125:1, 2010.

L. Shi, L.H. Reid, W.D. Jones, R. Shippy, J.A. Warrington, S.C. Baker, P.J. Collins, F. de Longueville, E.S. Kawasaki, K.Y. Lee, Y. Luo, Y.A. Sun, J.C. Willey, R.A. Setterquist, G.M. Fischer, W.Tong, Y.P. Dragan, D.J. Dix, F.W. Frueh, F.M. Goodsaid, D.Herman, R.V. Jensen, C.D. Johnson, E.K. Lobenhofer, R.K. Puri, U. Schrf, J. Thierry-Mieg, C. Wang, M. Wilson, P.K. Wolber, L. Zhang, S. Amur, W. Bao, C.C. Barbacioru, A.B. Lucas, V. Bertholet, C. Boysen, B. Bromley, D. Brown, A. Brunner, R. Canales, X.M. Cao, T.A. Cebula, J.J. Chen, J. Cheng, T.M. Chu, E. Chudin, J. Corson, J.C. Corton, L.J. Croner, C. Davies, T.S. Davison, G. Delenstarr, X. Deng, D. Doris, A.C. Eklund, X.H. Fan, H. Fang, S. Fulmer-Smentek, J.C. Fuscoe, K. Gallagher, W. Ge, L. Guo, X. Guo, J. Hager, P.K. Haje, J. Han, T. Han, H.C. Harbottle, S.C. Harris, E. Hatchwell, C.A. Hauser, S. Hester, H. Hong, P. Hurban, S.A. Jackson, H. Ji, C.R. Knight, W.P. Kuo, J.E. LeClerc, S. Levy, Q.Z. Li, C. Liu, Y. Liu, M.J. Lombardi, Y. Ma, S.R. Magnuson, B. Maqsodi, T. McDaniel, N. Mei, O. Myklebost, B. Ning, N. Novoradovskaya, M.S. Orr, T.W. Osborn, A. Papallo, T.A. Patterson, R.G. Perkins, E.H. Peters, R. Peterson, K.L. Philips, P.S. Pine, L. Pusztai, F. Qian, H. Ren, M. Rosen, B.A. Rosenzweig, R.R. Samaha, M. Schena, G.P. Schroth, S. Shchegrova, D.D. Smith, F. Staedtler, Z. Su, H. Sun, Z. Szallasi, Z. Tezak, D. Thierry-Mieg, K.L. Thompson, I. Tikhonova, Y. Turpaz, B. Vallanat, C. Van, S.J. Walker, S.J. Wang, Y. Wang, R. Wolfinger, A. Wong, J. Wu, C. Xiao, Q. Xie, J. Xu, W. Yang, L. Zhang, S. Zhong, Y. Zong, and W.Jr. Slikker. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24(9):1151–61, 2006.

N.D. Heintzman, G.C. Hon, R.D. Hawkins, P. Kheradpour, A. Stark, L.F. Harp, Z. Ye, L.K. Lee, R.K. Stuart, C.W. Ching, K.A. Ching, J.E. Antosiewicz-Bourget, H. Liu, X. Zhang, R.D. Green, V.V. Lobanenkov, R. Stewart, J.A. Thomson, G.E. Crawford, M. Kellis, and B. Ren. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112, 2009.

R. Edgar, M. Domrachev, and A.E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A.J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J.Y.H. Yang, and J. Zhang. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004. URL http://genomebiology.com/2004/5/10/R80.

Gordon K. Smyth. Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, and W. Huber R. Irizarry, editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer, New York, 2005.

D. Sarkar. *lattice: Lattice Graphics*, 2009. URL http://CRAN.R-project.org/package=lattice. R package version 0.17-26.

R. Gentleman. *R programming for Bioinformatics*. Chapman & Hall/CRC, 2009.

W. Huber, A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18:S96–104, 2002.

N.D. Heintzman, R.K. Stuart, G. Hon, Y. Fu, C.W. Ching, R.D. Hawkins, L.O. Barrera, S. Van Calcar, C. Qu, K.A. Ching, W. Wang, Z. Weng, R.D. Green, G.E. Crawford, and B. Ren. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, 39(3):311–8, 2007.

A. Oshlack, D. Emslie, L.M. Corcoran, and G.K. Smyth. Normalization of boutique two-color microarrays with a high proportion of differentially expressed probes. *Genome Biology*, 8 (1):R2, 2007.

A.E.K. Ibrahim, N.P. Thorne, K. Baird, N.L. Barbosa-Morais, S. Tavare, V.P. Collins, A.H. Wyllie, M.J. Arends, and J.D. Brenton. MMASS: an optimized array-based method for assessing CpG island methylation. *Nucleic Acids Research*, 34(20):e136, 2006.

P.W. Laird. Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics*, 11(3):191–203, 2010.