Appl Microbiol Biotechnol (2006) 72:852–859
DOI 10.1007/s00253-006-0536-x

METHODS

Comparison of different approaches for comparative genetic analysis using microarray hybridization

Carmen Pin · Mark Reuter · Bruce Pearson · Lorna Friis · Karin Overweg · József Baranyi · Jerry Wells

Received: 24 April 2006 / Revised: 6 June 2006 / Accepted: 7 June 2006 / Published online: 25 August 2006

Abstract A robust analysis of comparative genomic microarray data is critical for meaningful genomic comparison studies. In this paper, we compare our method (implemented in a new software tool, GENCOM, freely available at http://www.ifr.ac.uk/safety/gencom) with three commonly used analysis methods: GACK (freely available at http://falkow.stanford.edu), an empirical cut-off value of twofold difference between the fluorescence intensities after LOWESS normalization or after AVERAGE normalization in which the fluorescence intensity is divided by the average fluorescence intensity of the entire data set. Each method was tested using data sets from real experiments with prior knowledge of conserved and divergent genes. GENCOM and GACK were superior when a high proportion of genes were divergent. GENCOM was the most suitable method for the data set in which the relationship between the fluorescence intensities was not linear. GENCOM has proved robust in an analysis of all the data sets tested.

C. Pin (✉) · M. Reuter · B. Pearson · K. Overweg · J. Baranyi
Institute of Food Research, Norwich Research Park,
Coloney Lane,
Norwich NR4 7UA, UK
e-mail: carmen.pin@bbsrc.ac.uk

Present address:
L. Friis
Department of Medical Microbiology and Immunology,
University of Alberta,
Edmonton, Alberta T6G 2H7, Canada

Present address:
J. Wells
Swammerdam Institute for Life Sciences,
University of Amsterdam,
Nieuwe Achtergracht 166,
Amsterdam 1018 WV, Netherlands

Keywords Microarrays · Bioinformatics · Genetic analysis · Genomotyping

Introduction

Complete genome sequences are now available from a large number of bacterial species. A comparative analysis of these sequences provides important insights into the genetic basis of bacterial diversity, evolution, and pathogenesis (Andersson and Dehio 2000; Fleischmann et al. 2002; Ivanova et al. 2003; Parkhill et al. 2003; Pearson et al. 2003). DNA microarrays have an important application in comparative genomics as they can be used to compare whole genomes of different bacterial isolates for the presence or absence of genes (Behr et al. 1999). This technique can provide information on the genetic diversity of isolates in natural populations and/or on the genome variation for environmental adaptation (Riehle et al. 2001). It can also be used to investigate strains with different virulence or physiological properties (Israel et al. 2001; Mira et al. 2002; Perrin et al. 2002; Tettelin et al. 2002). The genomic comparisons between non-pathogenic and pathogenic strains of the same species can reveal putative virulence factors for further investigation (Ivanova et al. 2003). Interspecies microarray hybridization can rapidly identify thousands of genes in uncharacterized genomes as described for *Salmonella bongori*, several serovars of *Salmonella enterica* (Chan et al. 2003; Porwollik et al. 2002), various species of *Shewanella* (Murray et al. 2001), and *Klebsiella pneumoniae* (Dong et al. 2001). The use of DNA microarrays has also been evaluated for multilocus sequence typing (Swiderek et al. 2005). In addition, mixed-plasmid microarrays have been used to evaluate the genetic

diversity in plasmids from *Escherichia coli* and *S. enterica* (Call et al. 2006).

A comparative genomic analysis can be carried out on custom-designed DNA microarrays with probes specific to all open reading frames (ORFs) of a sequenced reference genome. Fluorescently labeled genomic DNA from a reference strain is hybridized with labeled DNA from a test strain. For each gene in the reference strain, one can determine if this gene in the test strain is conserved or divergent (Shalon et al. 1996).

The approaches applied to analyze microarray data for genetic composition studies can be divided into two groups. The approaches in the first group transform the data to minimize the experimental error associated with each specific hybridization and this transformation is called data normalization. A common normalization method applied to microarray data divides the fluorescence intensity for each gene by the average fluorescence intensity calculated from the entire hybridization or a subsection of the data set (Quackenbush 2002). Another extensively applied normalization method is the so-called LOWESS normalization (Luu et al. 2001). The method applies the LOWESS non-parametric regression (Cleveland 1979) to describe the relationship between the difference (M) and the average (A) of the logarithm of the intensities, i.e., on MA plots. The LOWESS scores are used as normalization factors and they are generated locally with the values of consecutive subsections of the MA plot. In this first group of approaches, after data normalization, a constant ratio cut-off value usually combined with a statistical test is used to classify genes as conserved or divergent. The cut-off for assignment of variable genes is usually inaccurately determined by comparison of the reference strain to a similar strain with known deletions or arbitrarily based on other published cut-off values (Hatfield and Baldi 2002; Kim et al. 2002). The approaches in the second group analyze each hybridization data set according to its intrinsic experimental variability and do not necessarily normalize explicitly the data. The two techniques belonging to this group are the GACK method (freely available at http://falkow.stanford.edu) published by Kim et al. (2002) and GENCOM (GENetic COMposition analysis; freely available at http://www.ifr.ac.uk/safety/gencom) described in Pearson et al. (2003). The GACK method finds the conserved genes by fitting the normal density function to the difference of the log intensities with the highest frequencies, thereby enabling the variable genes to be identified in the tails of the density function. GENCOM implements a convergence algorithm to find the conserved genes using the properties of the relationship between the two fluorescence intensities in each hybridization. In both methods, the boundary between conserved and variable genes is dynamically determined for each microarray.

In this paper, we present a new software tool, GENCOM, that implements the algorithm described in Pearson et al. (2003) and we compare it with other analysis techniques using experimental hybridization data sets, particularly data sets in which the relationship between the fluorescence intensities is not linear or where variable genes predominate.

Materials and methods

Bacterial strains and growth

Campylobacter jejuni strains NCTC11168, NCTC12547, NCTC11828 (also known as 81116) (Nuijten et al. 1989), 81–176 ($pTet^+$ $pVir^+$), and DB179 ($pTet^-$ $pVir^+$) (Bacon et al. 2002) were grown at 37°C under microaerophilic conditions (10% CO_2, 5% O_2, 85% N_2; relative humidity 80%) on Skirrow agar plates or in Mueller Hinton broth using a MACS-MG-1000 controlled atmosphere workstation (DW Scientific, UK). *Streptococcus pneumoniae* strains TIGR4 (also called JNR7/87) and R6 were grown at 37°C in tryptic soy broth or on tryptic soy agar plates supplemented with 5% horse blood.

Construction of DNA microarrays

The following three DNA microarrays were used: (1) a microarray representing one replicate of all open reading frames from the *S. pneumoniae* strains TIGR4 and R6 (Dagkessamanskaia et al. 2004), (2) a sub-microarray representing four replicates of all open reading frames from the campylobacter plasmids pVir (43 genes) and pCC31 (33 genes) and 16 campylobacter chromosomal genes that are conserved in both strains, and (3) a microarray representing one replicate of all open reading frames from *C. jejuni* NCTC11168 (Pearson et al. 2003).

The microarrays were stabilized and pre-hybridized, and DNA fluorescent labeling (red and green fluorescent dye) and hybridization were carried out as previously described (Pearson et al. 2003). The microarrays were scanned using an Axon 4000A microarray scanner and data were acquired using GenePixPro 3.0 software (Axon).

Hybridization data sets analyzed in this study

Data set I consists of four independent microarray hybridization of DNA from *S. pneumoniae* strains TIGR4 and R6 to the pneumococcal microarray described in "Construction of DNA microarrays". DNA hybridization included a dye swap, i.e., TIGR4 DNA was labeled with Cy3 (green fluorescent dye) and Cy5 (red fluorescent dye) and

combined with R6 DNA labeled with Cy5 and Cy3, respectively. The dye swap had no effect on the results (not shown). This data set contains a large number of spot signals, the majority of the genes are conserved between both DNA samples, and there is a linear relationship between the fluorescence intensities of both DNA samples.

Data set II is a microarray data set of DNA from *C. jejuni* strains 81–176 ($pTet^+$ $pVir^+$) and DB179 ($pTet^-$ $pVir^+$) on the sub-microarray described in "Construction of DNA microarrays". This data set contains a small number of spot signals; an important proportion of the genes, though not the majority, is missing in one of the DNA samples, and there is a linear relationship between the fluorescence intensities of both DNA samples.

Data set III represents a microarray hybridization of DNA from *C. jejuni* strains 81–176 ($pTet^+$ $pVir^+$) and NCTC11828 ($pTet^-$ $pVir^-$) on the same sub-microarray used to generate data set II. This data set contains a small number of spot signals; the majority of the genes are missing in one of the DNA samples, and there is a linear relationship between the fluorescence intensities of both DNA samples.

Data set IV consists of a microarray hybridization of DNA from *C. jejuni* strains NCTC11168 and NCTC12547 to the *Campylobacter* microarray described in "Construction of DNA microarrays". This data set contains a large number of spot signals, the majority of the genes is conserved between both DNA samples, and there is a nonlinear relationship between both fluorescence intensities (all four data sets are provided at http://www.ifr.ac.uk/safety/gencom).

Microarray data analysis methods

GENCOM This method is described by Pearson et al. (2003). It assumes a linear relationship between the logarithm of the red and green fluorescence intensities of the conserved (p=present) genes, denoted as R_p and G_p, respectively, i.e., $\ln R_p=\alpha_p+\beta_p \ln G_p$ with σ_p as the standard error of the regression. The parameters α_p, β_p, and σ_p are estimated using exclusively the intensities of the conserved genes by means of a convergence algorithm. The parameter α_p takes a value different from 0 if one of the fluorescence intensities is systematically exp (α_p)-folds greater than in the other fluorescence intensity. A β_p value other than 1 means that the relationship between the green and red fluorescence intensities is not linear but should be described by a power function. Thus, both α_p and β_p are correcting factors for individual microarray data sets. The parameter σ_p is the error when fitting the regression line to the intensities of the conserved genes and is used to establish the $3\sigma_p$ boundary between the conserved and variable genes in each hybridization data set. The potentially variable genes lie outside the $3\sigma_p$ boundary. Their classification is statistically tested using an F test according to their average intensities and the variability between replicates. This method is implemented in a Visual Basic add-in for Excel and is available at http://www.ifr.ac.uk/safety/gencom. To analyze data set I, GENCOM was run in each of the four hybridization data sets. Normalization was carried out by comparing the three parameters describing the features of the conserved genes: α_p, β_p, and σ_p. The standard error of the conserved genes to the regression line, σ_p, was fairly constant in all of the hybridization data sets, but some discrepancies were noticed in α_p and β_p. For each data set, the intensities were transformed to obtain $\alpha_p=0$ and $\beta_p=1$. The transformed intensities from each array were unified in one single data set, data set I, and analyzed by GENCOM. The analysis of data set III required the initial values of −2, 1.3, and 0.005 for the parameters α_p, β_p, and σ_p, respectively. Data sets II and IV were analyzed using GENCOM without initial normalization or other requirements.

GACK This method is described by Kim et al. (2002). The normal probability density function was fitted to the highest frequency peak of the logarithm of the fluorescence intensity ratios between the average intensities by the least squares method. When the probability, according to the fitted distribution, of higher or smaller ratios than the ratio measured in a gene was smaller than 0.05, that gene was classified as absent. To unify the four data sets of *S. pneumoniae* in data set I, the intensities of each of the independent hybridization data sets were shifted so that the distributions of the logarithm of the ratios were centered in 0. The final analysis was carried out on the average intensities of the four replicates in the unified data set. In data set III, the normal density function was not fitted to the highest frequency peak but to a smaller peak, which represented the frequencies of the present genes. Data sets II and IV were analyzed by GACK without extra requirements.

LOWESS The LOWESS non-parametric regression (Cleveland 1979) was fitted on the *MA* plots [*M*=logarithm of the ratio between intensities vs *A*=average of the logarithms of the intensities; log (*R*/*G*) vs log (*RG*)]. In most cases, the fraction of data used to determine the radius of each fitting neighborhood was 0.2 as described by Luu et al. (2001). After normalization, an empirical cut-off value equivalent to twofold changes between the fluorescence intensities was used to decide whether genes were conserved or variable. Where replicates existed, a t test was carried out. The four data sets of *S. pneumoniae* were normalized independently before unifying them in data set I.

AVERAGE The fluorescence intensities were normalized by dividing them by the average intensity calculated from the entire data set as explained by Quackenbush (2002). The four data sets of *S. pneumoniae* were normalized independently before unifying them, and an empirical twofold cut-off value and a *t* test were applied to classify the genes.

Quantification of the sequence identity in the conserved genes for *S. pneumoniae*

An identity indicator (IdentIn) was calculated for the probes located in the boundary as:

$$IdentIn = \frac{alignment\ length}{amplicon\ length} \times \frac{percent\ identity\ in\ the\ alignment}{100} \quad (1)$$

Sequence identity and alignment length were determined by BLAST searching using the genome sequences as the query database. This indicator ranges from 0 to 1; if the alignment length is equal to the amplicon length and the sequence identity for that alignment/amplicon is 100%, this value will be 1. An IdentIn value of 0.95 is equivalent to 95% identity over the entire length of an amplicon or gene. This indicator is relevant for our study because it takes into account both the length of the studied sequence and the sequence identity and it can provide a good estimation of the level of identity in the boundary between present and absent genes.

Dot plot analysis

The DNA sequences of misclassified amplicons of *S. pneumoniae* were concatenated and compared with the published genome sequences of TIGR4 and R6 (window size=16, mismatch=0) using DNAMAN software (Lynnon BioSoft).

Results

We compared our method GENCOM with three commonly used analysis methods, GACK, LOWESS, and AVERAGE (see "Materials and methods"), on four distinct microarray data sets. As the sequences of all DNA samples used in this study are publicly available, we were able to identify incorrectly classified genes.

Data set I represents a typical hybridization where conserved genes predominate and hybridization and scanning are optimal. Similar results were obtained with all four analysis methods (Table 1) although GENCOM and GACK gave slightly better results; only about 5% of the genes were incorrectly classified by the latter two methods. A large proportion of the misclassified genes (111 of 120) were false positives or genes misclassified as conserved in both strains while sequence-based analysis classified them as R6 specific. Our sequence-based analysis is based on BLAST searches using the full-length genes. However, a dot plot analysis revealed substantial regions of identity between the polymerase chain reaction products printed on the microarray slide and the TIGR4 chromosome (89 out of 111). When these genes were omitted from the GENCOM or the GACK analysis, the total error was 0.8%. The incorrectly classified variable genes show fluorescence intensity ratios that are characteristic for conserved genes (Fig. 1). These genes were wrongly classified by all four methods.

We used an identity indicator (IdentIn) to quantify the sequence identity in the genes in the boundary between conserved and variable genes (see "Materials and methods") for data set I analyzed. The IdentIn takes into account the total amplicon length and the proportion of that amplicon that is identical to a region of genomic DNA. The IdentIn had a value greater than 0.95 for all the genes classified as conserved with one exception. This is equivalent to 95% sequence identity over the entire length of a gene. The variable genes showed values between 0.05 and 0.95 again with one exception. GENCOM calculated the ratio between fluorescence intensities in the boundary between variable and conserved genes as the interval from 1.54 to 1.46; the value 1.54 was the greatest ratio detected for a conserved gene and 1.46 the smallest ratio for a variable gene.

Data sets II and III are derived from arrays with a small number of probes (92). In data set II, the conserved genes represent ca. 64% of the total number of genes. While GENCOM, GACK, and AVERAGE methods gave accurate results, LOWESS normalization resulted in 37% wrongly classified genes (Table 1; Fig. 2). In data set III, only about 17% of the genes were conserved in both DNA samples. The GACK method classified all genes correctly while GENCOM classified three genes incorrectly. The other two approaches performed rather poorly (Table 1). Figure 3 shows the results when analyzing this data set with the AVERAGE approach.

Data set IV is a genomic comparison experiment where conserved genes predominate but where the relationship

Table 1 Number of genes incorrectly classified in the data sets I to IV by the different analysis methods

Analysis	Error	Data set I (2,354 ORFs)		Data set II (92 ORFs)		Data set III (92 ORFs)		Data set IV (1,695 ORFs)
		TIGR4	R6	81–176 ($pTet^+$ $pVir^+$)	DB179 ($pTet^-$ $pVir^+$)	81–176 ($pTet^+$ $pVir^+$)	NCTC11828 ($pTet^-$ $pVir^-$)	NCTC 11168
GENCOM	False positives	7	111	0	0	0	3	–
	False negatives	0	2	0	0	0	0	1
	Total, *n* (%)	7 (0.3)	113 (4.8)	0 (0.0)	0 (0.0)	0 (0.0)	3 (3.3)	1 (0.1)
GACK	False positives	5	110	0	0	0	0	–
	False negatives	2	2	0	0	0	0	27
	Total, *n* (%)	7 (0.3)	112 (4.8)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	27 (1.6)
LOWESS	False positives	9	109	0	16	0	26	–
	False negatives	0	9	19	1	11	0	215
	Total, *n* (%)	9 (0.4)	118 (5.0)	19 (20.6)	17 (18.5)	11 (11.1)	26 (28.3)	215 (12.7)
AVERAGE	False positives	8	138	0	3	0	9	–
	False negatives	0	0	0	0	16	0	135
	Total, *n* (%)	8 (0.3)	138 (5.9)	0 (0.0)	3 (3.3)	16 (17.4)	9 (9.8)	135 (8.0)

False positives variable genes misclassified as present, *False negatives* present genes misclassified as variable

between the fluorescence intensities is not linear, i.e., the regression slope between the logarithms of the intensities is different from 1. This is an artefact from the microarray experimental procedure. GENCOM misclassified only one gene (Table 1). The algorithm found optimal values for the slope (β_p) and the intercept (α_p) of the regression line for the logarithm of the intensities of the conserved genes (α_p=1.76 and β_p=0.78). The value of the slope, 0.78, clearly shows the lack of linearity between the fluorescence intensities. The GACK method wrongly classified 27 genes as variable in the strain NCTC11168 (Fig. 4; Table 1). The LOWESS and AVERAGE methods classified 215 and 135 genes incorrectly, respectively (Table 1). The erroneous classification by the latter three methods shows the inability of these approaches to analyze data sets in which the relationship between intensities is not linear.

Discussion

We have studied the performance of four methods for comparative genetic composition analysis based on typical and atypical microarray data. As the annotated sequences of all DNA samples used in this study are publicly available,

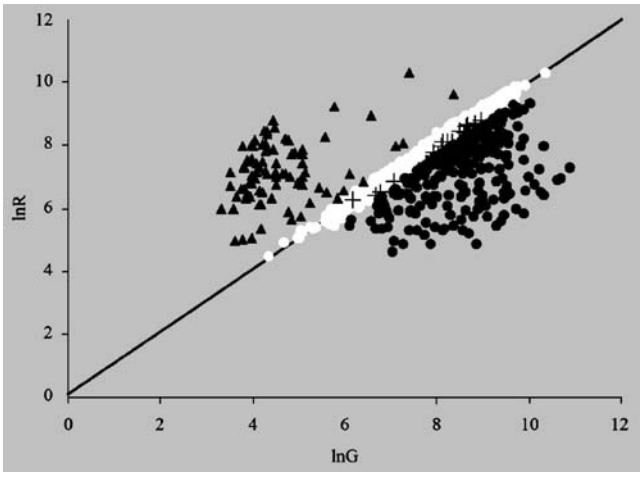

Fig. 1 Data set I: average of the natural logarithm of the fluorescence intensities from DNA microarrays hybridizing genomic DNA of *S. pneumoniae* strains TIGR4 and R6. The genes were classified with GENCOM. ○, genes present in both strains; ●, genes variable in TIGR4; ▲, genes variable in R6; +, genes variable in R6 misclassified as conserved

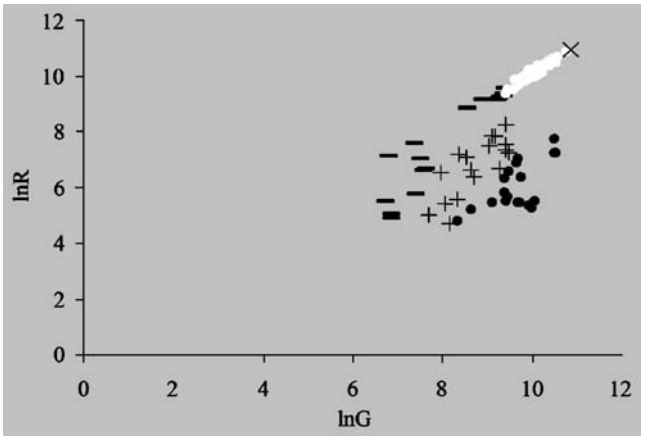

Fig. 2 Data set II: a microarray hybridization of *C. jejuni* strains 81–176 ($pTet^+$ $pVir^+$) and DB179 ($pTet^-$ $pVir^+$) on the slide constructed with the two plasmids, pTet (33 genes) and pVir (43 genes), plus 16 extra chromosomal genes present in both strains. The genes were classified with the LOWESS approach. ○, genes present in both strains; ●, genes variable in DB179; +, genes variable in DB179 misclassified as conserved; – genes present in 81–176 misclassified as variable; ×, genes present in DB179 misclassified as variable

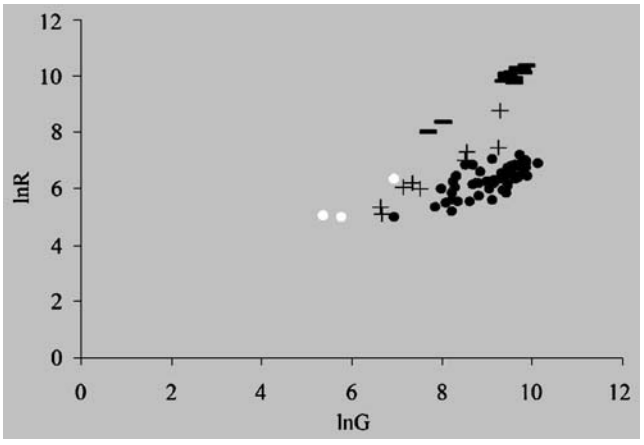

Fig. 3 Data set III: a microarray hybridization of *C. jejuni* strains 81–176 ($pTet^+$ $pVir^+$) and NCTC11828 ($pTet^-$ $pVir^-$) on the slide constructed with the two plasmids, pTet (33 genes) and pVir (43 genes), plus 16 extra chromosomal genes present in both strains. The genes were classified with AVERAGE approach. ○, genes present in both strains; ●, genes variable in NCTC11828; +, genes variable in NCTC11828 misclassified as conserved; –, genes present in 81–176 misclassified as variable

the incorrectly microarray-based classified genes could be identified. Data set I represents a typical comparative genomic microarray; the number of conserved genes is much greater than the number of divergent genes and the relationship between the fluorescence intensities is linear. This represents an optimal data set which should result in robust results with any appropriate statistical analysis method. All four methods indeed gave good results (Table 1) although GENCOM and GACK performed slightly better. This is because these methods base the establishment of the boundary between conserved and variable genes on the variability of the conserved genes only. The derived cut-off value between divergent and conserved genes was ca. 1.5-fold changes for data set I and

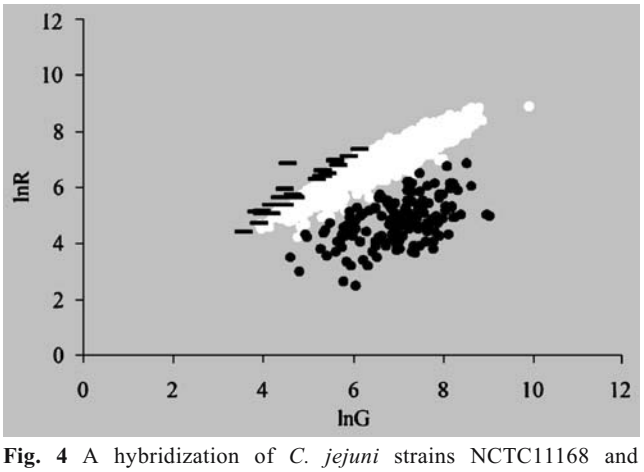

Fig. 4 A hybridization of *C. jejuni* strains NCTC11168 and NCTC12547 on the microarray constructed with the whole genome of the strain NCTC11168. The genes were classified with GACK. ○, genes present in both strains; ● genes variable in NCTC12547; – genes present in NCTC11168 misclassified as variable

this represents a 95% nucleotide sequence identity over the entire length of the probe. The correspondence between fluorescence intensity ratio and sequence identity depends on the analysis approach and the method for quantifying sequence identity. Taboada et al. (2005) indeed reported a percentage identity of ca. 90% between cDNA probe and gene for a logarithm ratio equal to 1, while for the same ratio other authors estimated a 70% sequence identity (Gaudriault et al. 2006).

For data set I, the LOWESS scores used for normalization were inaccurate for genes with low average intensities (see Fig. 5). This would normally result in a noticeable increase of wrongly classified genes. However, the disturbance in the data after LOWESS normalization was compensated by the use of a broad twofold arbitrary cut-off. This is why the LOWESS method resulted in only a very slight increase of the number of wrongly classified genes. For the AVERAGE method, the overestimation of the number of conserved genes in data set I was due to the use of the broad twofold cut-off.

When normalizing a microarray data set, the fluorescence intensities of the two DNA samples are assumed to be functionally related,

$$R = f(G) \tag{2}$$

where R is the fluorescence intensity from DNA sample 1 (labeled with a red fluorescent dye) and G is from DNA sample 2 (labeled with a green fluorescent dye). Distinct methods identify f in different ways. For example, AVERAGE (Quackenbush 2002) assumes that the majority of the genes are conserved and that the relationship between the fluorescence intensities is linear, i.e., $f(G)=KG$ where K is the constant of proportionality and estimated as the ratio between the averages of the fluorescence intensities over the entire data set. When the assumptions are not met, as in data sets III and IV, AVERAGE results in a high number of incorrectly classified genes. In data set III, only 17% of the genes are conserved and K is calculated with the fluorescence intensities from mainly variable genes. In data set IV, the relationship between the two fluorescent intensities is not linear.

LOWESS normalization assumes that the fluorescence intensity ratios of the variable genes are homogenously distributed on the *MA* plot, i.e., $f(G)=KG$, where the logarithm of K represents the prediction calculated after applying LOWESS regression to the MA plot. The LOWESS method failed to analyze data sets II, III, and IV correctly for the following reasons. The main reason is that, in DNA–DNA hybridization, the logarithms of the ratios of the absent genes on the *MA* plots are not distributed homogeneously, namely, they are biased to the lower average intensities. The disturbance effect of the missing genes on the LOWESS scores affects mainly the genes with low

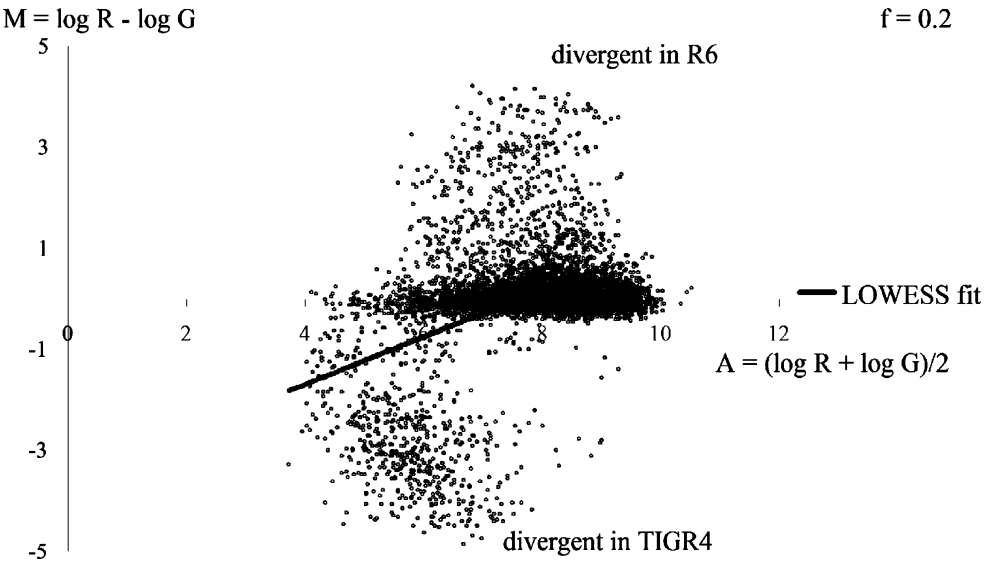

Fig. 5 LOWESS fit on the *MA* plot derived from data set I

fluorescence intensities. For this reason, in data set IV, mainly the genes with low A values were wrongly classified. The greater the number of absent genes, the less homogeneous is the distribution of the ratios on the MA plots and the LOWESS normalization becomes more erroneous. Thus, the LOWESS normalization method is more sensitive to the proportion of absent genes than the AVERAGE method.

Normalization can be considered an intrinsic feature in both the GENCOM and GACK methods. When analyzing a data set derived from a unique microarray slide, explicit normalization is not needed. Normalization must be carried out explicitly only to unify data from different hybridization.

The GACK normalization assumes proportionality between intensities, $f(G)=KG$, and K represents the parameter known as the mean of the normal distribution fitted to the log intensity ratios of the conserved genes only. The method fails if the assumption of normality for the log intensity ratios is not met, as it does in data set IV. The relationship between the fluorescence intensities in data set IV is not linear, and for this reason the distribution of the log intensity ratios is far from normal and not even symmetric. As a consequence, when applying the GACK method, genes were wrongly classified.

However, GENCOM can correctly analyze data sets in which the relationship between the fluorescence intensities is not linear. GENCOM assumes that the relationship between intensities is:

$$f(G) = KG^C \tag{3}$$

By taking logarithms, Eq. (3) can be transformed as follows:

$$ln\, R = \alpha + \beta\, ln\, G \tag{4}$$

where α and β are estimated by linear regression on the intensities of the conserved genes. The relationship between the parameters of Eqs. (3) and (4) is α=ln (K) and β =C.

GENCOM can satisfactorily analyze and/or normalize data set IV, where the relationship between intensities is nonlinear.

Both methods, GACK and GENCOM, give good results independently of the proportion of variable genes. However, the analysis of data set III, although satisfactory, was not straightforward either for GENCOM or for GACK method. The difficulty lies in the small number of conserved genes, only 16. In the case of the GACK approach, the main difficulty was to construct a histogram with this small number of conserved genes to fit the normal distribution. To analyze this data set with GENCOM, it was necessary to search for optimal initial values for the parameters of the conserved genes. One of the requirements of the algorithm is that the initial values must ensure the selection of some truly conserved genes in the first iteration.

We have shown the importance of normalizing comparative genomic microarray hybridization data sets according to conserved genes only. The normalization techniques that use the entire hybridization data set may be successful only if the proportion of variable genes is small. The analysis approach must be able to accurately detect and analyze atypical data, such as data sets in which absent genes predominate or the relationship between the intensities is not linear. We showed that the GACK and GENCOM methods accurately analyzed the data sets in which the majority of the genes were absent. However, only GENCOM was able to correctly analyze the data set in which the relationship between the raw intensities is not linear. So, GENCOM performs best for comparative genetic composition analysis based on microarray hybridization.

Acknowledgements We gratefully acknowledge support from the BBSRC core strategic grants 434.1213A and 4311209A. JMW also wishes to acknowledge support from EC project TCS-TARGETS QLK2-CT-2000-00543.

References

Andersson SG, Dehio C (2000) *Rickettsia prowazekii* and *Bartonella henselae*: differences in the intracellular life styles revisited. Int J Med Microbiol 290:135–141

Bacon DJ, Alm RA, Hu L, Hickey TE, Ewing CP, Batchelor RA, Trust TJ, Guerry P (2002) DNA sequence and mutational analyses of the pVir plasmid of *Campylobacter jejuni* 81–176. Infect Immun 70:6242–6250

Behr MA, Wilson MA, Gill WP, Salamon H, Schoolnik GK, Rane S, Small PM (1999) Comparative genomics of BCG vaccines by whole-genome DNA microarray. Science 284:1520–1523

Call DR, Kang MS, Daniels J, Besser TE (2006) Assessing genetic diversity in plasmids from *Escherichia coli* and *Salmonella enterica* using a mixed-plasmid microarray. J Appl Microbiol 100:15–28

Chan K, Baker S, Kim CC, Detweiler CS, Dougan G, Falkow S (2003) Genomic comparison of *Salmonella enterica* serovars and *Salmonella bongori* by use of an *S. enterica* serovar typhimurium DNA microarray. J Bacteriol 185:553–563

Cleveland W (1979) Robust locally weighted regression and smoothing scatterplots. J Am Stat Assoc 74:829–836

Dagkessamanskaia A, Moscoso M, Henard V, Guiral S, Overweg K, Reuter M, Martin B, Wells JM, Claverys J (2004) Interconnection of competence, stress and CiaR regulons in *Streptococcus pneumoniae*: competence triggers stationary phase autolysis of ciaR mutant cells. Mol Microbiol 51:1071–1086

Dong Y, Glasner JD, Blattner FR, Triplett EW (2001) Genomic interspecies microarray hybridization: rapid discovery of three thousand genes in the maize endophyte, *Klebsiella pneumoniae* 342, by microarray hybridization with *Escherichia coli* K-12 open reading frames. Appl Environ Microbiol 67:1911–1921

Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, Peterson J, DeBoy R, Dodson R, Gwinn M, Haft D, Hickey E, Kolonay JF, Nelson WC, Umayam LA, Ermolaeva M, Salzberg SL, Delcher A, Utterback T, Weidman J, Khouri H, Gill J, Mikula A, Bishai W, Jacobs WR Jr, Venter JC, Fraser CM (2002) Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. J Bacteriol 184:5479–5490

Gaudriault S, Duchaud E, Lanois A, Canoy AS, Bourot S, Derose R, Kunst F, Boemare N, Givaudan A (2006) Whole-genome comparison between *Photorhabdus* strains to identify genomic regions involved in the specificity of nematode interaction. J Bacteriol 188:809–814

Hatfield GW, Baldi P (2002) DNA microarrays and gene expression: from experiments to data analysis and modeling. Cambridge University Press, Cambridge, pp 53–72

Israel DA, Salama N, Krishna U, Rieger UM, Atherton JC, Falkow S, Peek RM Jr (2001) *Helicobacter pylori* genetic diversity within the gastric niche of a single human host. Proc Natl Acad Sci U S A 98:14625–14630

Ivanova N, Sorokin A, Anderson I, Galleron N, Candelon B, Kapatral V, Bhattacharyya A, Reznik G, Mikhailova N, Lapidus A, Chu L, Mazur M, Goltsman E, Larsen N, D'Souza M, Walunas T, Grechkin Y, Pusch G, Haselkorn R, Fonstein M, Ehrlich SD, Overbeek R, Kyrpides N (2003) Genome sequence of *Bacillus cereus* and comparative analysis with *Bacillus anthracis*. Nature 423:87–91

Kim CC, Joyce EA, Chan K, Falkow S (2002) Improved analytical methods for microarray-based genome-composition analysis. Genome Biol 3:RESEARCH0065

Luu P, Yang YH, Dudoit S, Speed TP (2001) Normalization of cDNA microarray data. SPIE BIOS

Mira A, Klasson L, Andersson SG (2002) Microbial genome evolution: sources of variability. Curr Opin Microbiol 5:506–512

Murray AE, Lies D, Li G, Nealson K, Zhou J, Tiedje JM (2001) DNA/DNA hybridization to microarrays reveals gene-specific differences between closely related microbial genomes. Proc Natl Acad Sci U S A 98:9853–9858

Nuijten PJ, Bleumink-Pluym NM, Gaastra W, van der Zeijst BA (1989) Flagellin expression in *Campylobacter jejuni* is regulated at the transcriptional level. Infect Immun 57:1084–1088

Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, Harris DE, Holden MT, Churcher CM, Bentley SD, Mungall KL, Cerdeno-Tarraga AM, Temple L, James K, Harris B, Quail MA, Achtman M, Atkin R, Baker S, Basham D, Bason N, Cherevach I, Chillingworth T, Collins M, Cronin A, Davis P, Doggett J, Feltwell T, Goble A, Hamlin N, Hauser H, Holroyd S, Jagels K, Leather S, Moule S, Norberczak H, O'Neil S, Ormond D, Price C, Rabbinowitsch E, Rutter S, Sanders M, Saunders D, Seeger K, Sharp S, Simmonds M, Skelton J, Squares R, Squares S, Stevens K, Unwin L, Whitehead S, Barrell BG, Maskell DJ (2003) Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. Nat Genet 35:32–40

Pearson BM, Pin C, Wright J, I'Anson K, Humphrey T, Wells JM (2003) Comparative genome analysis of *Campylobacter jejuni* using whole genome DNA microarrays. FEBS Lett 554:224–230

Perrin A, Bonacorsi S, Carbonnelle E, Talibi D, Dessen P, Nassif X, Tinsley C (2002) Comparative genomics identifies the genetic islands that distinguish *Neisseria meningitidis*, the agent of cerebrospinal meningitis, from other *Neisseria* species. Infect Immun 70:7063–7072

Porwollik S, Wong RM, McClelland M (2002) Evolutionary genomics of *Salmonella*: gene acquisitions revealed by microarray analysis. Proc Natl Acad Sci U S A 99:8956–8961

Quackenbush J (2002) Microarray data normalization and transformation. Nat Genet 32 Suppl:496–501

Riehle MM, Bennett AF, Long AD (2001) Genetic architecture of thermal adaptation in *Escherichia coli*. Proc Natl Acad Sci USA 98:525–530

Shalon D, Smith SJ, Brown PO (1996) A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res 6:639–645

Swiderek H, Claus H, Frosch M, Vogel U (2005) Evaluation of custom-made DNA microarrays for multilocus sequence typing of *Neisseria meningitidis*. Int J Med Microbiol 295: 39–45

Taboada EN, Acedillo RR, Luebbert CC, Findlay WA, Nash JH (2005) A new approach for the analysis of bacterial microarray-based comparative genomic hybridization: insights from an empirical study. BMC Genomics 6:78

Tettelin H, Masignani V, Cieslewicz MJ, Eisen JA, Peterson S, Wessels MR, Paulsen IT, Nelson KE, Margarit I, Read TD, Madoff LC, Wolf AM, Beanan MJ, Brinkac LM, Daugherty SC, DeBoy RT, Durkin AS, Kolonay JF, Madupu R, Lewis MR, Radune D, Fedorova NB, Scanlan D, Khouri H, Mulligan S, Carty HA, Cline RT, Van Aken SE, Gill J, Scarselli M, Mora M, Iacobini ET, Brettoni C, Galli G, Mariani M, Vegni F, Maione D, Rinaudo D, Rappuoli R, Telford JL, Kasper DL, Grandi G, Fraser CM (2002) Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. Proc Natl Acad Sci U S A 99:12391–12396