Proceedings

IASC2008

Joint Meeting of
4th World Conference of the IASC and
6th Conference of the Asian Regional Section of the IASC
on Computational Statistics & Data Analysis

Pacifico Yokohama, Japan
December 5 (Fri.) – 8 (Mon.), 2008

Organized by:
IASC2008 International Organizing Committee
Japanese Society of Computational Statistics
The Institute of Statistical Mathematics

Supported by:
Commemorative Organization for the Japan World Exposition '70

Published by:
Japanese Society of Computational Statistic

Proceedings

IASC2008

Joint Meeting of
4th World Conference of the IASC and
6th Conference of the Asian Regional Section of the IASC
on Computational Statistics & Data Analysis

December 5 - 8, 2008

Pacifico Yokohama, Japan

Masahiro Mizuta
Junji Nakano
Editors

Proceedings of IASC2008

IASC2008: Joint Meeting of 4th World Conference of the IASC and 6th Conference of the Asian Regional Section of the IASC on Computational Statistics & Data Analysis
December 5 - 8, 2008 / Pacifico Yokohama, Japan

Published by: Japanese Society of Computational Statistics (JSCS)
Office c/o Statistical Information Institute for Consulting and Analysis
Daiwa building 2F, 6-3-9 Minami Aoyama, Minato-ku,
Tokyo, 107-0062, JAPAN
Phone: +81-3-5467-6946 Fax: +81-3-5467-6946

The JSCS Office will move to the address below in January 2009.
Office c/o Statistical Information Institute for Consulting and Analysis
Nohgaku Shorin Building 5F, 3-6 Kanda Jimbocho, Chiyoda-ku,
Tokyo, 101-0051, JAPAN

ISBN 978-4-9904445-1-8

Permission to host “Fortifying the analysis of Illumina data” on DSpace @Cambridge

Re: Is it possible to host a pdf of my paper [Fortifying the analysis of Illumina data] on my university library webspace?

From: Wataru SAKAMOTO
Sent: 01 April 2011 01:17
To: Mark Dunning

Dear Dr. Mark Dunning,

Thank you for your inquiry. The committees for IASC2008 have already been dissolved, so I asked the chairs of the committees at that time.

They said that you are free to use the part of your own paper in the pdf files included in the proceeding CD.

With kind regards,

Wataru Sakamoto

Scientific Secretary
International Association for Statistical Computing (IASC)

Preface

This proceedings assembles submitted and selected papers explaining presentations at IASC2008, the Joint Meeting of 4th World Conference of the IASC and 6th Conference of the Asian Regional Section of the IASC on Computational Statistics & Data Analysis, which was held on December 5 - 8, 2008 in Yokohama, Japan. We note that the separate booklet (Program & Abstracts) contains general information of IASC2008 and abstracts of all presentations which were refereed by several members of the Local Organizing Committee.

IASC2008 is an activity of International Association for Statistical Computing (IASC), a section of the International Statistical Institute (ISI) which is one of the oldest international scientific associations functioning in the modern world. IASC2008 provided a forum for researchers and practitioners from all over the world to share their knowledge about theories, methods and practices of statistical computing and computational statistics, and to discuss current important issues regarding statistical methods and data analysis in various disciplines such as medicine, business, ecology, biology, engineering, where statistical computing is indispensable.

The IASC was founded to foster world-wide interest in effective statistical computing and to exchange technical knowledge through international contact and meetings between statisticians, computing professionals, organizations, institutions, governments and the general public. In order to fulfil the objectives, it organized several conferences including the 3rd IASC World Conference held in Cyprus in 2005 and the 5th Conference of the ARS (Asian Regional Section) of IASC held in Hong Kong, China, also in 2005. IASC2008 was a successor to these two conferences and the first united conference in the history of IASC.

IASC2008 was organized by enthusiastic efforts of many colleagues in the world. We are especially grateful to the members of the international organizing committee, the scientific program committee, the local organizing committee and the executive committee of IASC2008. We would also like to thank ISS, INC. for their excellent supporting work.

Masahiro Mizuta
Chair of the Scientific Program Committee of IASC2008

Junji Nakano
Chair of the International Organizing Committee of IASC2008

Fortifying the analysis of Illumina data

Mark Dunning[1] Jonathan Cairns[1] Roslin Russell[1] Andy Lynch[1]

[1] Cancer Research UK, CRI, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK
E-mail: mark.dunning@cancer.org.uk
E-mail: jmc200@cam.ac.uk
Email: Roslin.Russell@cancer.org.uk
Email: Andy.Lynch@cancer.org.uk

Keywords: Microarray, Illumina, pre-processing

1 Introduction

Microarrays are devices that allow for the high-throughput analysis of biological samples. High-throughput meaning that, with some technologies, millions of points in the genome of interest can be investigated simultaneously. At a very basic level a microarray works by having probes that are complementary to part of the genome attached to a surface such as a glass slide. The biological sample is then treated so that the characteristic of interest will fluoresce. After allowing the sample to hybridize to the microarray, this characteristic is quantified at a particular genomic location by measuring the fluorescence of the appropriate probe. There are many opportunities for noise or bias to enter a microarray experiment and Quality Assessment (QA) and Control (QC) steps are a crucial part of any analysis.

A microarray surface is typically scanned by laser to produce an image of the fluorescence, usually to be processed by the manufacturers' software. All of the probes in the image are located, and then intensities for each probe are calculated using the local pixels. However, the pixel intensities measured on the image may be influenced by factors other than hybridization such as optical noise from the scanner, foreign items deposited on the array or flaws on the array surface. Commonly, these effects are known as spatial artefacts and must be dealt with as part of the QC process (Reimers and Weinstein (2005)).

One manufacturer, Illumina, produces microarrays that work differently to the traditional technology. Rather than attaching probes directly to a surface, they are attached to beads that in turn are randomly positioned on an array (Kuhn et al (2004)). For each genomic location being investigated, there will typically be thirty to forty beads on the array with the same probe attached (known collectively as a bead-type). It is generally believed that this high degree of replication, with the spatial randomization, makes Illumina data robust and the platform is widely used. However the majority of analysts use data summarized by bead-type which restricts the ability to perform QA. Researchers are now realizing the benefits of analyzing data at the bead-level rather than using the bead-type summaries provided by Illumina's BeadStudio software (Dunning et al. (2008), Stokes et al. (2007)) and researchers have begun depositing bead-level data in microarray data repositories (Acevedo et al. (2008), Wilson et al. (2008)).

2 Illumina QA/QC

There are currently two types of QA measurement provided by Illumina. A text file (usually called *Metrics.txt*) is automatically generated, giving scores from 0 to 1 for how well each image was 'registered' and 'focused'. However, interpretation of these values is difficult. The file also records the 5th and 95th percentile of the foreground, and potentially useful data such as the time at which the arrays were scanned. Additionally, BeadStudio gives the opportunity to view quality control reports based on the summarized data from the control beads that Illumina put on each array. These controls generally report presence or absence of hybridization and are described in a technical note produced by Illumina (www.illumina.com/downloads/GX_QualityControl_TechNote.pdf).

The Human6 V3 chip used in this paper has 664 negative control bead-types. These have probe sequences that should not target the genome being investigated, and no signal should be observed. Whether other bead-types are showing expression can be assessed by comparison to the negative controls. There are also sample-independent controls that have probe sequences complementary to oligonucleotides spiked into the hybridization solution. These should always have detectable signal. For six such bead types the concentration is controlled at three levels ("medium", "low", "high"), and there should be

a perceptible gradient. There is also one housekeeping bead-type targeting the gene ACTB, which is generally believed to be ubiquitously expressed, and two bead-types with sequences complementary to the biotin used for staining the arrays. Other controls use of pairs of "perfect match" and "mismatch" bead-types where the "mismatch" probe sequence differs at one location from the "perfect match" sequence and should result in a lower signal.

The summarized values presented by BeadStudio have undergone two processes that might manage spatial artefacts. First there is a local background correction step that, while not detrimental, has negligible impact. Second there is an outlier removal stage that, within each bead-type, discards observations that are more than three median absolute deviations (MADs) from the bead-type median. However the values are not logged before outlier calculation, with the result that low outliers are difficult, or sometimes impossible, to detect. No account is made for the number of beads in a bead-type in performing this calculation, nor is spatial information used, thus rendering the process suboptimal for detecting outliers driven by a spatial artefact. A final pitfall is that with only one iteration of the outlier detection process, a slide with multiple spatial artefacts may see the beads from one artefact being generally discarded, but find that the beads from the remaining artefacts are kept in the analysis as they were undetectable until the first artefact was removed.

The current usage of the control summary reports provided by Illumina is the comparison of average values of particular controls across the entire experiment after completion. Ideally though, QA issues would be addressed at the earliest opportunity. Moreover, BeadStudio does not indicate the number of outliers that have been removed, and so there is no warning that there may have been defects on the array; a matter that is of concern if the outlier process has not adequately addressed the matter. Even if such defects have been satisfactorily addressed, it would be useful to know of them in order that the causes can be investigated and eradicated from future experiments.

In this paper, we describe a more rigorous QA procedure for Illumina expression data, which can identify arrays that have completely failed, whose measurements cannot be trusted, or which require some adjustment. This consists of html report pages, quality statistics, and an Illumina-specific implementation of the Harshlight (Suarez-Farinas et al. (2005)) approach known as BASH (**b**ead**a**rray **S**ubversion of **H**arshlight). BASH, which is described elsewhere (Cairns et al. (2008)), has been tailor-made for Illumina arrays and available within the *beadarray* R package (Dunning et al. (2007)). The benefits of using bead-level QA scores and processes will be demonstrated using Illumina HumanV3 gene expression chips.

3 Proposed QA scores for Illumina expression arrays

We have previously noted that spatial artefacts can be found on Illumina arrays by using the imageplot function in *beadarray* (Dunning et al. (2007)). With the amount of replication on arrays, and random placement of beads, it should be possible to prevent artefacts from spoiling an analysis. Illumina recommend that outliers are removed prior to creating summarized data, and a simple modification to this outlier removal is to log-transform the data prior to applying the 3 MAD rule. As part of outlier QA process, we record the number of outliers on an array as calculated by Illumina, and after applying a log-transformation. Furthermore, these numbers are partitioned by the 2 strips that comprise an array, and 9 segments that comprise a strip, to highlight patterns. It should be noted that Illumina remove outliers on an array-wide basis. However, there is often a gradient along a BeadChip that leads to systematic differences in values between the two strips, for which reason we remove outliers on a strip-wide basis. In this paper where we refer to "Illumina outlier removal", we apply Illumina's rule but to a single strip only.

Rather than use the summarized output from BeadStudio, we derive QA measures from the bead-level data for the control probes. We have adapted Illumina's approach for assessing whether a given bead-type is expressed or not. Within an array, Illumina discard negative control bead-types whose summary values are more than three MADs from the median for the negative controls. Illumina then rank the summarized intensity for each other bead-type against the summarized values for the remaining negative control bead-types and calculate a detection p-value $1 - R/N$, where R is the relative rank of the bead intensity when compared to the N remaining negative controls. Thus, if a particular bead has higher intensity than all the negative controls it will be assigned a value of 0.

With access to bead-level data we can test each individual observation for a particular bead-type, and can include all observations that were reported on the array, unlike Illumina's implementation which takes place after outliers have been removed. Low expression of a few beads of a positive control bead-

type is likely to be due to a defect on the array surface. Thus, our QA scores include the percentage of positive control beads that are detected for housekeeping, biotin, and hybridization controls. Additionally, rather than always making comparison to the negative controls we calculate detection scores for medium hybridization control beads against the low hybridization controls, and for the high hybridization controls against the medium hybridization controls. There are some merits to also looking at the detection scores of the individual negative control beads, but for this paper the set of QA measures derived from the control beads, that we propose, and the abbreviations that are used in this paper, are as follows:

HkpDet - %age of housekeeping control beads that are detected compared to the negative controls.

BioDet - %age of biotin labelling control beads that are detected compared to the negative controls.

LowDet - %age of "low" control beads that are detected compared to the negative controls.

MedDet - %age of "medium" control beads that are detected compared to the negative controls.

HighDet - %age of "high" control beads that are detected compared to the negative controls.

MvsL - %age of "medium" control beads that are detected compared to the "low" controls.

HvsM - %age of "high" control beads that are detected compared to the "medium" controls.

The QA scores described above can be computed and stored for a BeadLevelList object read in by *beadarray*. If requested by the user, per-strip quality summary plots can be generated and compiled into an HTML page. An HTML page can also be generated that includes a summary of the QA measures for the entire chip.

4 Overview of BASH

Harshlight is an algorithm for masking physical defects on microarray chips. The algorithm first identifies *Extended defects*, where an array has substantial low-frequency variation across the surface, by assessing the variance of a "smoothed" image (relative to the variance of the actual image). After smoothing out local bead-specific variation, there should be low variance on a good array. Harshlight also seeks to find more localized artefacts on arrays by identifying clusters of outliers. Two separate algorithms search for areas with unusual numbers of outliers (*Diffuse Defects*) and larger than expected numbers of connected outliers (*Compact defects*). We have subverted Harshlight from its Affymetrix origins to produce the Illumina-specific BASH. BASH utilizes Illumina's within-array replication rather than requiring multiple arrays, and addresses a number of issues such as the hexagonal grid employed by Illumina while using an iterative procedure to pick up secondary spatial artefacts on an array. A more detailed description of BASH has been submitted elsewhere, so here we will focus on the benefits of using BASH.

5 Data

We used a dataset of 96 tumours hybridized onto Illumina gene expression Human6 V3 chips. For each sample, around 48,000 bead types can be interrogated, with each chip capable of analyzing six samples. As has been noted, each sample is hybridized to two physically separate strips within an array. An image is produced for both strips, and while BeadStudio combines the two strips, we will take advantage of the fact that we effectively have two technical replicate arrays for the same sample. Since the two strips are adjacent on the chip, and hybridized at the same time, it is reasonable to assume that, in the absence of technical artefacts, we should observe the same levels of expression on the two strips.

Using the *beadarray* software, the bead-level data were read for each strip on each of the chips separately (not possible using BeadStudio). Our set of QA scores was computed on these bead-level data before creating summarized data according to the 3 MAD rule with and without applying a $\log_2$ transformation. Summarized data were also created without removing outliers. This process was then repeated after using BASH.

6 Results

On the whole, the positive control scores for the tumour dataset were encouraging. The number of strips (out of 192) that had HkpDet, BioDet, LowDet, MedDet, HighDet, MvsL and HvsM scores of 100% were 184, 169, 169, 176, 179, 162 and 120 respectively. Notable examples of positive control failure were strips 4250605024_A_1 and 4250605024_A_2 where only 64% of housekeeping beads were detected, despite nearly 100% of all other positive control beads being detected. Thus, we can conclude that the hybridization was successful, but that the sample quality may be poor. A further indicator of the quality of this sample is given in the metrics file produced by Illumina, with the 95th percentile of bead intensities being 47 or 5.55 on the $\log_2$ scale, which is a value we would associate with the background level. Only two other strips (4250605058_A_2 and 4250605058_B_1) saw fewer than 96% of the housekeeping beads detected and these occur on the same chip. We give the QA scores for all strips on this particular chip in Table 1.

Table 1: Table of QA measures (see Section 3) and BASH extended defect scores for chip 4250605058 of the tumour dataset.

Strip	HkpDet	BioDet	LowDet	MedDet	HighDet	MvsL	HvsM	BASHext
A_1	100.0	100.0	100.0	100.0	100.0	100.0	100	0.06
A_2	93.5	94.1	89.1	98.1	90.99	79.3	84.7	0.11
B_1	95.0	93.6	85.6	93.9	89.93	85.1	75.5	0.10
B_2	100.0	100.0	100.0	100.0	100.0	100.0	99.3	0.11
C_1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.09
C_2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.11
D_1	100.0	100.0	100.0	100.0	100.0	100.0	98.1	0.10
D_2	98.0	100.0	100.0	100.0	100.0	100.0	98.3	0.13
E_1	100.0	100.0	100.0	100.0	100.0	100.0	97.4	0.16
E_2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.16
F_1	100.0	100.0	100.0	100.0	100.0	100.0	98.7	0.12
F_2	100.0	99.1	100.0	100.0	100.0	100.0	97.6	0.14

Although the percentages are still high in our two aberrant strips, the design of the positive control beads is such that their intensities should always be detected above background and we should see 100% of the observations on each strip being detected. The two strips in question have consistently low values, while their replicates perform well suggesting a technical rather than biological cause. Any unusual behaviour of this kind warrants investigation. The fact that we are able to detect failure in some of the control beads on these strips is a departure from the method of quality control used by Illumina, which would throw away the defective beads before presenting the summarized data to the user. Thus, the strips appear normal in the summarized data given by BeadStudio (Table 2) and these errors are hidden from the user. Arrays A and B show no obvious signs of reduced signal, while the detection score tells us that the mean values exceed all of the negative control intensities, but doesn't highlight that up to 10% of the beads contributing to that average are failing.

Table 2: Table of Illumina QA measures (naturally by array rather than by strip) for the same chip. The first three columns provide the average intensities for biotin, hybridization and housekeeping controls, while the final three columns provide the detection p-values.

	Ave.Biotin	Ave.Hyb	Ave.Hkpg	DetBio	DetHyb	DetHkpg
A	12141	12717	14187	0	0	0
B	8561	9456	9340	0	0	0
C	9131	10036	12144	0	0	0
D	8847	9632	9767	0	0	0
E	8978	10042	3406	0	0	0
F	9085	9585	12692	0	0	0

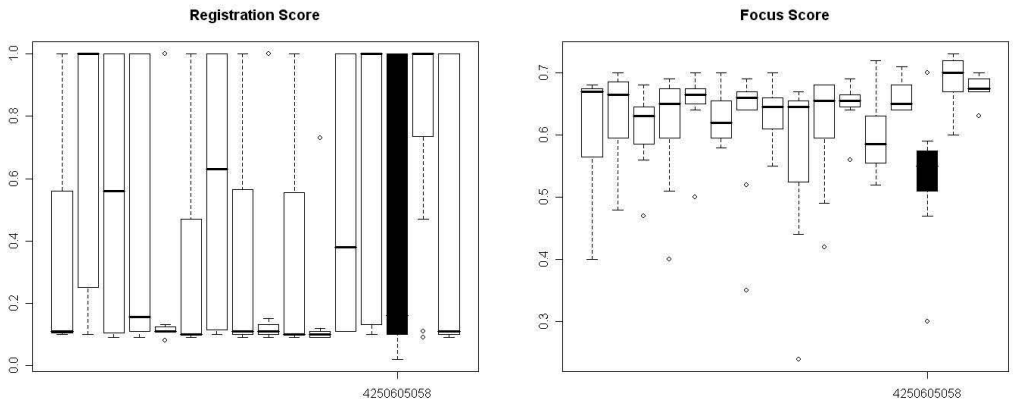

Figure 1: Depicting the registration and focus quality metrics for the strips by the 16 chips in this data set. Our example chip is labelled.

Examination of the metrics files at least raises some questions about these strips (Figure 1), with the Registration values particularly low for the two strips that we have highlighted, and the Focus values being generally low across the chip. However many of the registration values are lower that would be anticipated, and these values don't reveal that the problems with the strips may be localized (and thus removed). The low focus measures are difficult to interpret, and there are other arrays that we do not suspect are problematic but which have similar values.

Most of the strips on this array have Extended Defect scores that are higher than we would be comfortable with. On the other chips, the IQR for extended defect scores is 0.062 to 0.075. The values that we are seeing on chip 4250605058 suggest that there is a low-frequency spatial artefact (such as a gradient along the arrays) that may impede both outlier and spatial artefact identification. The focus metric scores of illumina may reflect this same phenomenon. It may be possible to correct for this using a surface-normalization method (perhaps a loess based method). Certainly, the observations from this chip will be less precise and it may be desirable to down-weight them in any analysis.

The quality control reports we produce are in HTML format, but we present some of the key elements in Figure 2 where a poor quality array is contrasted with a better quality array.

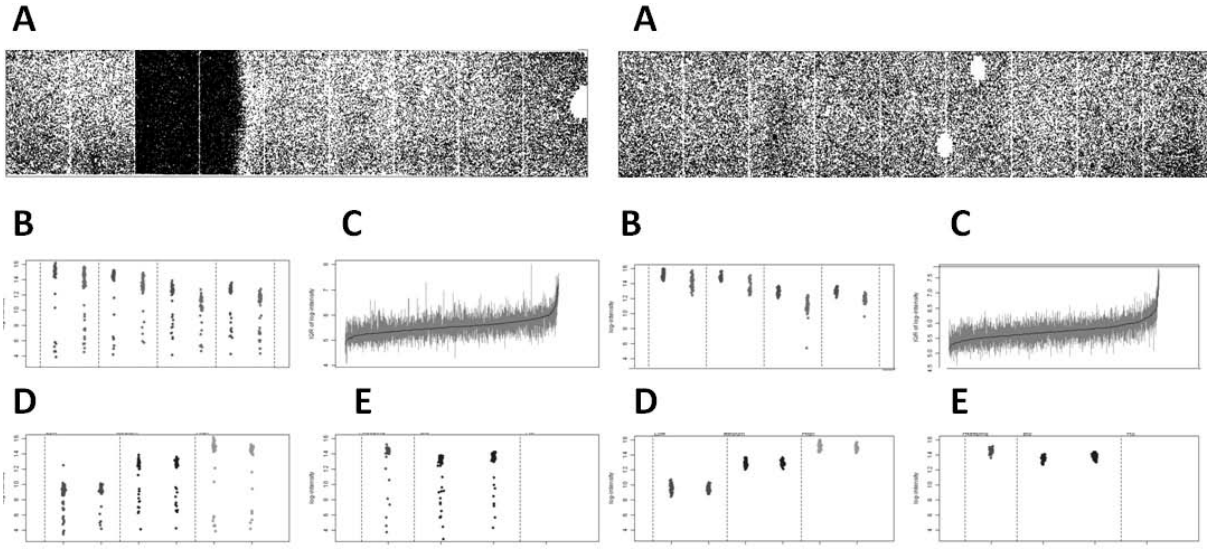

Figure 2: Comparison of two quality control reports for contrasting strips (4250605058_A_1 and 4250605058_A_2).

The most striking element (labelled A) is a plot of outlier locations. Note that the outlier plots show

an obvious spatial artefact on the poor quality strip, with substantial numbers of outliers in segments three and four, whereas no obvious defects can be seen on the good quality strip (the regions of white are areas where no beads could be successfully identified by Illumina). This artefact is associated with low intensities, making this a prime example of an artefact that Illumina outlier identification would struggle to identify. In the HTML report numbers of beads and numbers of outliers per segment are also given.

The remaining elements within the reports are plots of each bead for the perfect and mismatch paired bead-types (B) of which there are four pairs for this chip, plots of the 664 negative control bead-types (C), where each bead-type is represented by a vertical line covering the inter-quartile range, plots of each bead for the low-medium-high control probes (D) with two bead-types for each level, and plots of Housekeeping and biotin controls (E) that should be high in all cases. Whilst the positive controls on the poor quality strip have high average intensity, for each control there is a trail of lower intensity observations. Again we note that these problems will not be apparent from Illumina's summary data. It is also worth noting that the amalgamation of perfect match controls for comparison with the amalgamated mismatch controls, as can be graphed in BeadStudio, is not advisable due to the variation between perfect matches (and that between mismatches) as seen in Figure 2.

Table 3 gives the percentages of outliers found for all the strips of our example chip, both in total and broken down by the nine segments that constitute a strip. Generally, each segment has around 5% to 6% of beads as outliers, although the strips identified as particularly problematic in Table 1 are found to have many more outliers. Note that strip A_2 is the poor quality strip depicted in Figure 2, and while after viewing that figure we would reasonably wish to exclude the entirety of segment 3, in fact three quarters of those beads would be retained in the analysis using Illumina outlier rules. This motivates the implementation of the BASH methodology, especially since detecting spatial artefacts by-eye would be time-consuming and difficult to reproduce exactly. The number of outliers on a given strip in the tumour dataset was generally between 5% and 6% (25th and 75th quantiles of 5.2 and 6.0), so the strips A_2 and B_1 discussed here are quite exceptional with 8.0% and 7.8% outliers respectively. Thus, along with the control QA scores, the number of outliers can be used as a guideline to identify defective strips.

Table 3: Table of the percentage of Illumina-called outliers seen in the nine segments (S1 to S9), and in total, of the twelve strips on a chip.

Strip	S1	S2	S3	S4	S5	S6	S7	S8	S9	Total
A_1	5.0	4.6	4.8	4.8	4.7	4.9	4.0	4.5	5.7	4.8
A_2	4.4	4.0	25.6	15.2	3.6	3.4	3.9	4.4	6.1	8.0
B_1	5.0	4.4	4.3	18.1	11.0	11.8	4.6	4.5	6.6	7.8
B_2	5.7	4.3	4.4	4.3	4.6	5.2	5.5	6.0	9.2	5.5
C_1	6.8	6.0	6.3	6.2	5.7	6.0	5.1	6.2	8.7	6.4
C_2	6.5	5.9	5.8	5.5	5.6	5.0	4.9	6.4	8.7	6.0
D_1	6.1	5.5	5.3	5.1	5.2	5.7	5.2	6.2	8.1	5.8
D_2	5.4	4.8	4.7	4.5	4.7	4.9	5.1	6.4	9.4	5.5
E_1	3.1	2.8	2.9	2.9	3.4	3.9	4.4	5.8	8.2	4.2
E_2	3.1	2.8	2.7	2.8	3.2	3.6	4.4	5.6	8.8	4.1
F_1	7.7	7.1	6.2	5.8	5.8	5.2	5.0	7.1	12.2	6.9
F_2	6.9	5.5	5.5	5.0	4.8	4.6	5.2	7.2	13.3	6.5

In Table 4, the percentages of beads excluded by BASH are given in the same format as for Table 3. In contrast to regular outlier removal, BASH excludes all of the observations in the severely affected segments. Figure 3 illustrates the performance of BASH for the A_2 strip compared to the Illumina outliers. It can be seen that BASH not only masks all beads inside the obvious spatial artefact, but does not remove as many beads on the rest of the array. It is generally true that BASH removes few beads when there is no obvious local spatial artefacts, but note that for many of the strips on this chip BASH removes more beads than were initially identifies as outliers. In particular BASH is removing beads primarily from segments 1 and 2, while the Illumina-called outliers were more common in segments 8 and 9 (save for the strips with spatial artefacts). This can be explained by a gradient hinted at by the extended defect scores, coupled with the different scales upon which the analyses are performed. If there is a gradient going from "low" in segment 1 to "high" in segment 9, then the Illumina method that more

Table 4: Table of the beads removed by BASH for the strips of our example chip.

Strip	S1	S2	S3	S4	S5	S6	S7	S8	S9	Total
A_1	0.7	0.1	0.6	0.6	0.0	0.4	0.5	0.6	0.6	0.4
A_2	1.0	0.1	100.0	80.7	0.0	1.4	0.0	0.0	0.0	21.2
B_1	1.6	2.4	0.1	100.0	64.1	68.0	0.0	0.0	0.0	26.1
B_2	44.1	6.6	3.3	0.8	0.1	0.0	0.0	0.0	0.0	6.3
C_1	49.8	19.2	6.6	0.6	0.4	0.0	0.0	0.0	0.2	8.7
C_2	38.7	20.9	5.8	0.6	0.0	0.0	0.0	0.0	0.0	7.5
D_1	43.6	21.2	5.8	1.1	0.1	0.0	0.0	0.1	0.0	8.1
D_2	41.1	10.6	3.8	0.7	0.3	0.0	0.0	0.0	0.0	6.5
E_1	17.4	4.9	2.7	0.7	0.2	0.1	0.0	0.0	0.0	2.9
E_2	18.6	5.4	2.2	0.7	0.4	0.2	0.0	0.0	0.0	3.1
F_1	52.6	23.6	7.0	3.4	0.4	0.0	0.0	0.0	0.0	9.8
F_2	50.7	13.0	4.1	1.6	1.1	0.7	0.0	0.0	0.1	8.0

easily calls high outliers will call them in segment 9. After taking logs, these outlying values are drawn in to the main group and the lower outliers become visible which BASH removes from segment 1.

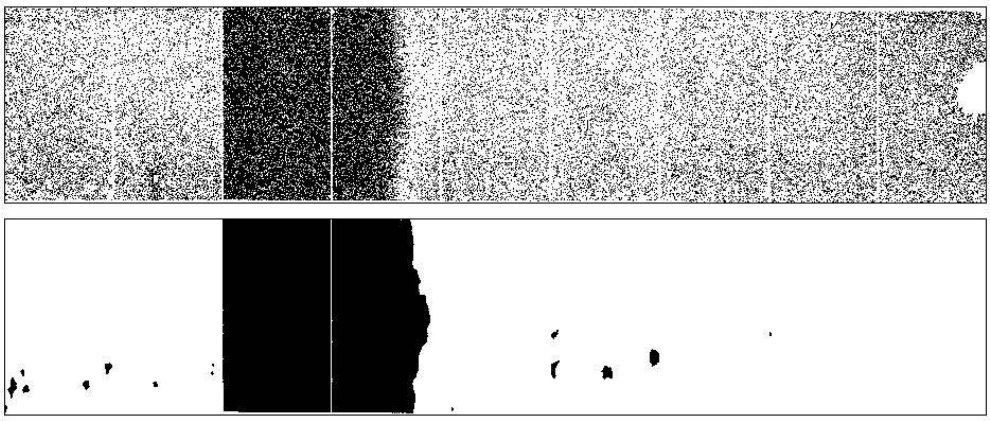

Figure 3: Locations of beads that found to be outliers on 4250605058_A_2 using Illumina's default method of a 3 MAD cut-off from the unlogged median (as seen in Figure 1), compared to beads masked using BASH. It can be seen that BASH not only masks all beads inside the obvious spatial artefact, but does not remove as many beads on the rest of the array.

If this is perceived to be problematic then there are a number of points to consider. Firstly BASH provides some basic gradient-correction options that can be applied before BASHing, and others could be applied before even calling BASH. Secondly, BASH allows for a certain amount of tuning of parameter values, and may be adjusted for a particular laboratory or experiment. Thirdly, *beadarray* provides tools for manually editing the mask that BASH returns, and finally it should be noted that the removal of some good beads should be less detrimental than the retaining of affected beads due to the redundancy built in to the Illumina BeadArray platform. Out of the box, on a single strip, the median number of bead-types (out of $\sim 48,000$ bead-types) with fewer than five observations was 9 (IQR 6 to 11). After outlier removal, there was a median of 15 bead-types (IQR 10 to 18). After additionally applying BASH, this rises to a median of 17 bead-types (IQR 14 to 22). Only 18 strips (out of 192) see the number rise by more than 10 after applying BASH and seven of these are on our already-identified poor quality chip.

It should be noted that the definition of Illumina outliers prevents any bead-type from disappearing entirely. This is perfectly possible though if one applies BASH. While this may be of concern, consider that if for a bead-type with just three observations, all three fell in the same spatial artefact then none would be removed by Illumina. While losing all of the observations for a particular bead-type may be distressing, to have no ability to discard seriously biased estimates would be a greater concern. These issues are amplified once one considers the Illumina HumanWG-12 chip.

The effect of having a poor quality strip is illustrated in Figure 4. A series of MA-plots are shown comparing the strips 4250605058_B_1 and 4250605058_B_2 after different methods were used to summarize the bead-level data from each strip separately. An MA-plot compares the log-ratio (M) of the measures from the two strips (y-axis) to the average log-intensity (A). Since the two strips are technical replicates, the MA plot should consist of a horizontal line (at M=0 in the absence of a gradient along the chip). Note that with no outlier removal, and Illumina-outlier removal, the spatial artefact is influential and shifts the value of M to below 0. With outlier removal on the log-scale, in the absence of BASH, only low outliers are liable to be removed from the affected chip and so the value of M shifts to above 0. Note that at the lower end of the A range ($A < 7$) we are measuring only noise and no signal, so the level of agreement that we can anticipate is limited and there is a resultant increase in the magnitude of $M-$values. This is a limitation of the log-transformation.

The effect of BASH in Figure 4 is striking, with far fewer values of M greater than 0.5 in magnitude. The effects are clearer in Figure 5, where the change in M after using BASH is depicted on the y-axis. Here we see that for two arrays with spatial artefacts, BASH improves agreement between technical replicates, and for an array with no spatial artefact it is not detrimental. Additionally, for the arrays with very poor housekeeping control scores (4250605024_A_1 and 4250605024_A_2) no improvement was seen by using BASH (not shown), suggesting that an appropriate course of action of these arrays would be to remove them from the analysis.

7 Discussion

Illumina arrays have seen a rapid increase in use and there is a perceived confidence in the results they produce. However, it is still worth remembering the principles learnt from other technologies and that careful QA is required at all stages of the analysis. For instance, spatial artefacts are rarely investigated for Illumina arrays and contrary to popular belief, Illumina's method for removing outliers does not always cope well with such artefacts. As the size of experiments that are being conceived to use Illumina arrays grows (experiments involving hundreds or even thousands of arrays are feasible) it is essential that an automated process is in place to screen for poor quality data at the time it is produced, rather than waiting until the entire experiment has been completed. To this end, we present a QA framework based on the bead-level data produced at the time of scanning, rather than the results obtained using BeadStudio. As the calculation of our proposed QA metrics are implemented in the `R` open-source language, there is naturally the opportunity to incorporate them into existing analysis pipelines. Although we would still recommend some scrutiny of the HTML reports generated by beadarry to guide decisions over whether particular arrays should be excluded rather than automating the whole process. A natural extension to our QA measures to import them into existing MSPC (Multivariate Statistical Process Control) software and see the relative contributions of each score to overall quality. A large collection of previously collected data could then be used to inform QA of newly generated data.

Although spatial artefacts and hybridisation failure were uncommon in this dataset, the combination of using QA measures based on bead-level data for the positive controls and the number of outliers on an array were able to flag potentially unreliable arrays. The BASH method performed well in being able to detect spatial artefacts on arrays by removing more beads on arrays with substantial spatial artefacts whilst retaining more beads on good quality arrays. In both these scenarios we would expect more reliable bead type averages to be produced.

At present, our QA measures are only applicable to Illumina expression although they should be easy to translate to other Illumina assays such as genotyping. The notion of QA on the bead-level data is especially important to convey and we hope that recent trends in making bead-level data available continue. As part of our analysis pipeline we also re-annotate the contents of each chip to make sure the probe sequences are well-designed (`http://www.compbio.group.cam.ac.uk/`). For the HumanV3 chip we confirm that the housekeeping control is a perfect match for the gene ACTB and that negative controls hybridisation controls have low probability of matching the human genome. Therefore we can discount the possibility of non-specific or cross-hybridisation affecting the results from the controls on this chip. However, this is not the case for an older chip (the HumanV2) where two out of the 14 housekeeping controls that were used had no perfect match to any region of the genome. Therefore the generally lower hybridisation seen for these control should be attributed to poor probe design and not to array defects.

BASH makes no assumptions about the content of the array and thus can run on any type of Illumina data. It would be especially interesting to see how BASH performs on high-density genotyping arrays,

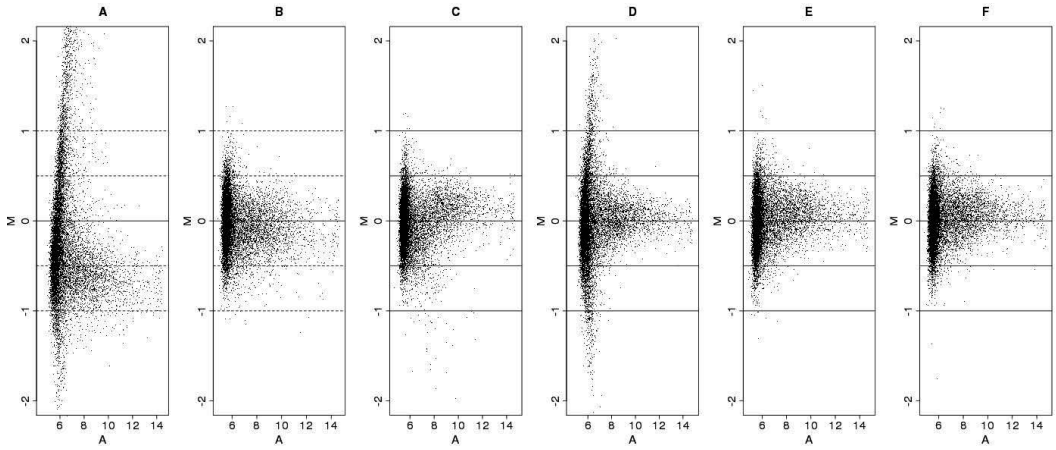

Figure 4: MA-plots comparing the two strips from an array containing a spatial artefact. The y-axis shows the log-ratio of the estimates from the two strips, which should be zero as the strips are technical replicates. Depicted are plots after no outlier removal (A), Illumina outlier removal (B) log outlier removal (C) and the same again but applying BASH first (D-F)

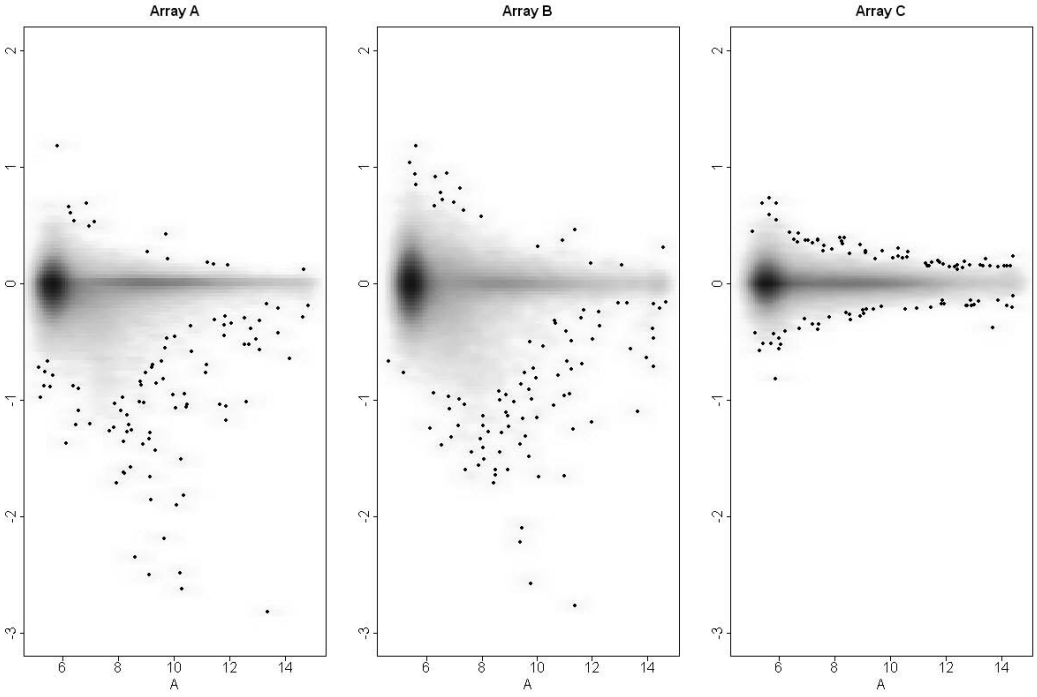

Figure 5: Depicting the change in M following BASH (y-axis) against A. Values below 0 show greater agreement following BASH. Figures are provided for three arrays, two containing a spatial artefact and a third containing no artefact.

where due to the sheer number of SNPs investigated, each sample is spread across multiple strips. However, the question of how one removes outliers on two-colour data remains to be addressed. Another interesting application of BASH is to the new Human HT-12 chips created by Illumina. These chips have 12 samples (rather than 6) with 48,000 bead types and only one strip for each sample. Thus, there are fewer observations for each bead-type and BeadStudio's outlier calling may be detrimentally affected. BASH, since it performs well for the HumanWG-6 arrays we presume will work well for the Human HT-12s, but will lack the technical replicates to validate this.

References

Acevedo L.G., Bieda M., Green R., Farnham P.J. (2008) Analysis of the mechanisms mediating tumor-specific changes in gene expression in human liver tumors. *Cancer Res 2008* 68(8):2641-51

Cairns J.M., Dunning M.J., Ritchie M.E., Russell R and Lynch A.G. (2008) BASH: A tool for managing BeadArray spatial artefacts. *Bioinformatics* Submitted

Dunning, M.J., Smith, M. L., Ritchie, M. E. Tavaré, S. (2007). beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics*, 23(16): 2183-2184

Dunning, M.J, Barbosa-Morais, N.L., Lynch A.G., Tavaré, S. and Ritchie M.E., (2008) Statistical issues in the analysis of Illumina data. *BMC Bioinformatics*, 9(85)

Kuhn, K., S.C. Baker, E. Chudin ,M.H. Lieu, S.Oeser, Bennett, H., Rigault,. P,. Barker, D, McDaniel., T.K. and Chee., M. (2004) A novel, high-performance random array platform for quantitative gene expression profiling *Genome Research*, 14(11) : 2347–2356.

Lin, S.M., Du, P., Huber, W., and Kibbe, W.A. (2008). Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Research*, 36(2)

Reimers M. and Weinstein J.N., (2005) Quality assessment of microarrays: Visualization of spatial artifacts and quantitation of regional biases *BMC Bioinformatics*, 6(166)

Suarez-Farinas, M., and Haider, A., and Wittkowski, K.M., (2005). Harshlighting small blemishes on microarrays. *BMC Bioinformatics*, 6(65)

Stokes, T.H., Han, X., Moffitt, R.A. and Wang M.D., (2007) Extending microarray quality control and analysis algorithms to illumina chip platform *Proceedings of the 29th annual international conference of the IEEE (EMBS)*

Wilson M.D., Barbosa-Morais, N.L., Schmidt, D., Conboy, C.M., Vanes, L., Tyybulewicz, V.L.J., Fisher, E.M.C., Tavaré. S, and D.T. Odom. (2008). Species-specific transcription in mice carrying human chromosome 21 *Science Express* 11 September 2008