

An overview of image-processing methods for Affymetrix GeneChips

Jose M. Arteaga-Salas^a, Harry Zuzan^a, William B. Langdon, Graham J. G. Upton and Andrew P. Harrison

Submitted: 28th August 2007; Received (in revised form): 12th October 2007

Abstract

We present an overview of image-processing methods for Affymetrix GeneChips. All GeneChips are affected to some extent by spatially coherent defects and image processing has a number of potential impacts on the downstream analysis of GeneChip data. Fortunately, there are now a number of robust and accurate algorithms, which identify the most disabling defects. One group of algorithms concentrate on the transformation from the original hybridisation DAT image to the representative CEL file. Another set uses dedicated pattern recognition routines to detect different types of hybridisation defect in replicates. A third type exploits the information provided by public repositories of GeneChips (such as GEO). The use of these algorithms improves the sensitivity of GeneChips, and should be a prerequisite for studies in which there are only few probes per relevant biological signal, such as exon arrays and SNP chips.

Keywords: *Affymetrix microarrays; image processing*

INTRODUCTION

Affymetrix GeneChip technology uses multiple independent hybridisation probes to measure the expression level for each gene investigated. Each probe is a 25-nt oligomer. Probes are designed in perfect match (PM) and corresponding mismatch (MM) pairs. Each MM is intended to act as a calibration signal for the true value given by the PM. PM and MM pairs are part of probe sets. Each probe set is designed to represent a different gene transcript. Typically probe sets consists of 11–16 PM and MM pairs. There has been tremendous success in applying GeneChip technology to illuminate the difference in gene-expression patterns for different species, tissues, phenotypes and disease states. This has led to thousands of publications, describing the results from tens of thousands of GeneChips observations, and the creation of public repositories of GeneChip data, such as NCBI's GEO [1]. Even with this successful use of GeneChips, it is clear that we can learn much more from the data. Many of the improvements in the use of GeneChips will derive from computational methods, such as being developed within the Bioconductor project ([2],

www.bioconductor.org), to remove systematic errors introduced by the technology.

Image processing of the raw data occurs prior to other calibration steps. Because of the potential for downstream inaccuracies introduced by image defects, it is important to establish reliable algorithms for calibrating GeneChips. Here we provide an overview of image processing of Affymetrix GeneChips.

FROM DAT TO CEL

The hybridisation data acquired from the scanner and scanner software supplied by the Affymetrix platform is stored in a numeric image data file with a suffix DAT. The file consists of a 512 byte header containing delimited fields of descriptive information that detail the size of the image, scanning equipment and other technical information followed by the array of pixel intensities acquired from the scanner, stored as 2-byte unsigned little-endian integers which can take values in the range (0,65535) inclusive.

^aThe first two authors contributed equally to this work.

Corresponding author: Andrew P. Harrison, Department of Mathematical Sciences, University of Essex, Wivenhoe Park, Colchester, UK, CO4 3SQ. Tel: +44 1206 872964; Fax: +44 1206 873043; E-mail: harry@essex.ac.uk

Dr Andrew Harrison is a lecturer in Mathematical Bioinformatics at the University of Essex and has worked on the analysis of Affymetrix GeneChips since 2003. His original research training was in Astrophysics.

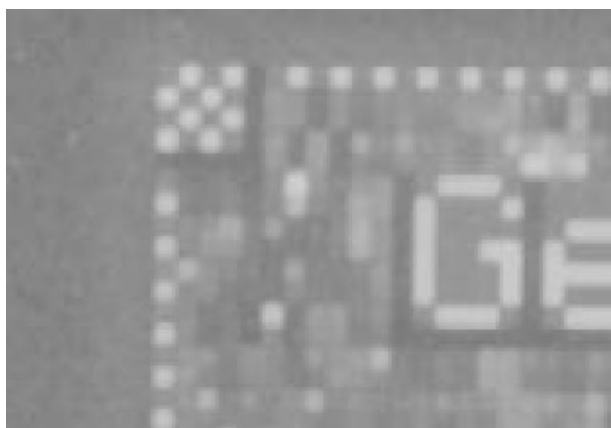


Figure 1: Fragment of DAT file, showing 4×4 checkerboard used for alignment.

When the GeneChip hybridisation surface is scanned, it is not known a priori which pixels will correspond to which probe cells. Attributing pixel elements to probe cells is an image segmentation problem. Since the probe cells are square and laid out in a rectangular lattice, the process of segmentation is often called grid alignment or registration. To aid grid alignment, there are 4×4 checkerboard patterns of alternating bright and dim probe cells in each of the four corners of the hybridisation surface (Figure 1). The alternation of intensity is effected through a combination of probe sequence and the addition of exogenous target sequences to reagent kits used in sample processing. Affymetrix's image-processing software first locates the four checkerboard sequences then establishes a preliminary grid through bi-linear interpolation between the four corners. Recent chip designs incorporate a lattice of 2×2 checkerboard patterns on the hybridisation surface to aid or validate the interpolation. After the preliminary interpolation, the grid alignment is locally refined; however, we are not aware of manuscripts from Affymetrix that adequately describe the current method of refinement.

Once local refinement of the grid alignment is complete, a set of pixels is assigned to each probe cell. In early versions of Affymetrix chips and software, pixels could comprise a rectangular array. In current versions, all the arrays of pixels corresponding to probe cells are square. Each array of pixels mapped to a probe cell has its perimeter pixels deemed the least reliable and discarded. One reason for the lack of reliability is that even if the grid alignment is optimal, some pixels will still carry signal from more than one probe cell. Given some

misalignment of the grid, it is the pixels around the perimeter of a probe cell that will be most affected by misalignment. A second reason for discarding the pixels is that the signal of a probe cell tends to be weakest around its edges. As a consequence, if there were a 5×5 array of pixels attributed to a probe cell, then this would be reduced to a 3×3 array of central pixels. From this reduced set of pixels, the 75th percentile is reported as the estimate of the probe cell intensity, and the standard deviation and number of pixels used to compute the 75th percentile are also reported. This information is stored in the cell intensity file with a suffix CEL. A mapping of probe cells to aggregates of probe cells, called probe sets that are each related to either a gene or quality control purpose can be found in the cell description file with a suffix CDF. A description of all relevant files can be found on the Affymetrix Developers Network at <http://www.affymetrix.com>.

The CEL file provides a summary of the DAT file. Any inaccurate information in this summary can only be recovered by reprocessing the DAT file. Such errors may be important, but may not be evident. Inadequate segmentation of the hybridisation stored in the DAT file is one source of error, and there have been several proposed solutions alternative to the vendor supplied grid alignment [3–5].

At this point, we have discussed segmentation of the image as if the information contained in a single pixel was perfectly accurate. In fact, another important source of error is blur, meaning that each pixel records signal not just from the region of the hybridisation surface that it covers. It also records signals from its surrounding region and it also loses some of its signal to its surrounding region. Pixels covering high-intensity regions will tend to lose more signals to their surrounding pixels than they gain and pixels covering low intensity regions will tend to accumulate extra signal from their neighbours. In successive versions of the Affymetrix chips and scanners the physical size of probe cells has decreased from approximately 24×24 microns to 5×5 microns. This means that the ratio of the perimeter to area of a probe cell is increasing and hence so is the proportion of unreliable information in a hybridisation image. Increasing the number of pixels can be of little help unless there is a corresponding reduction in blur. There appears to have been little analysis, independent of the vendor, in this area.

Finally the recorded pixel values will be truncated in the range (0,65535) for 16-bit unsigned integers. Truncation of signal at the high end is called fluorescent saturation. Such saturation can be controlled by amplification of signal in the scanner photomultiplier tube and circuit board [Dan Bartell, personal communication]. Naturally reducing the amplification also reduces the resolution of the signal at the low end.

Saturation of fluorescent signal should not be confused with chemical saturation [6], which is a function of inhibition of binding caused by pre-existing hybridisation duplexes. Since probe affinity is highly probe-sequence dependent [7], chemical saturation can be restricted to some extent by selecting against probes with extremely high affinity in the design of a chip.

Typically, the 75th percentile of pixel intensities is the only part of the summary used in downstream analysis methods. The pixels near the centre of a probe cell with hybridisation signal above background noise tend to be much brighter than those around the edges. Discarding the pixels at the cell edges means, the (remaining) pixels attributed to a probe cell represent a heterogeneous population and hence the SD is typically of little use in downstream estimates of gene-expression levels, other than aiding the detection of outliers, grid misalignment and defects in the hybridisation surface [3–5]. Since the photon count for each pixel is amplified, distributional assumptions such as Poisson counts for probe cell intensities are not close to being valid and cannot be usefully applied to downstream estimates of gene expression. Therefore the 75th percentile of pixel intensities attributed to a probe cell, given in the CEL file, is effectively the hybridisation summary of gene expression. If the grid alignment is accurate, image blur is not an important factor and the hybridisation image is free from large defects and recorded within the range of 16-bit pixels with at most a few pixel intensities truncated, then the CEL file should adequately summarise the contents of the DAT file for the purpose of estimating gene expression.

DETECTING DEFECTS

It is clear that a variety of defects affect GeneChip images, e.g. Figure 2: Suárez-Fariña *et al.* [8] discuss ‘dirt’, ‘dark and bright spots’, ‘dark clouds’ and ‘shadowy circles’; Upton and Lloyd [9] refer to

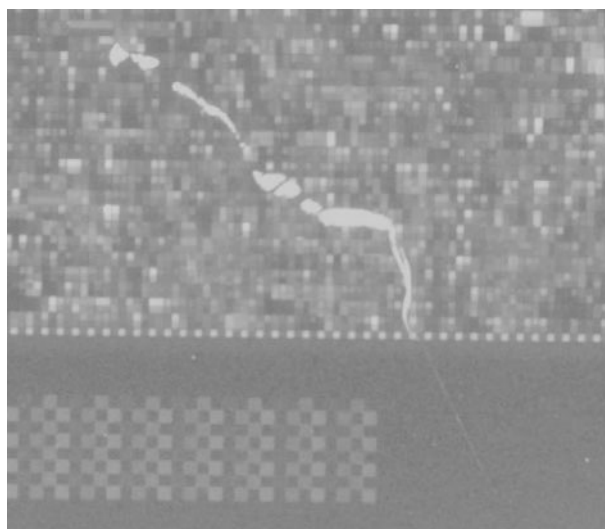


Figure 2: Fragment of HG-UI33A DAT file, showing linear defect from 10 to 40 μm wide at edge of GeneChip.

‘blob’, ‘lines’, ‘rectangular enhancement’ and ‘coffee rings’. It is difficult to see these unwanted spatial patterns by eye directly on a single Affymetrix GeneChip because of the large dynamic range of probe intensities, but progress is possible either by comparing each probe against some reference value, e.g. Figure 3, or by modelling the values in sets of arrays [10, 11].

The choice of reference values

In order to be convinced that a given cell in an array is an outlier, we must first have some notion of what value would be expected in that cell (a *reference value*). Whatever procedure is used the aim is the same: to identify an outcome that would occur only rarely by chance, but would occur relatively frequently in the presence of positive spatial autocorrelation. Notice that it is not useful to look for the occurrence of an event simply because it should rarely occur in the absence of flaws—the event must be one that regularly occurs in the presence of spatial flaws. We are looking for a test that has both high *sensitivity* (high conditional probability of discovering a spatial flaw given that one is present) and high *specificity* (high conditional probability of judging a cell is flawless, given that it is flawless). We can almost always improve a procedure’s specificity by introducing a second pass in which cells initially deemed to be flawed are discarded because they have insufficient neighbours judged to be flawed. This relies on the assumption that the external factors causing a flaw will rarely affect single probes, but will usually

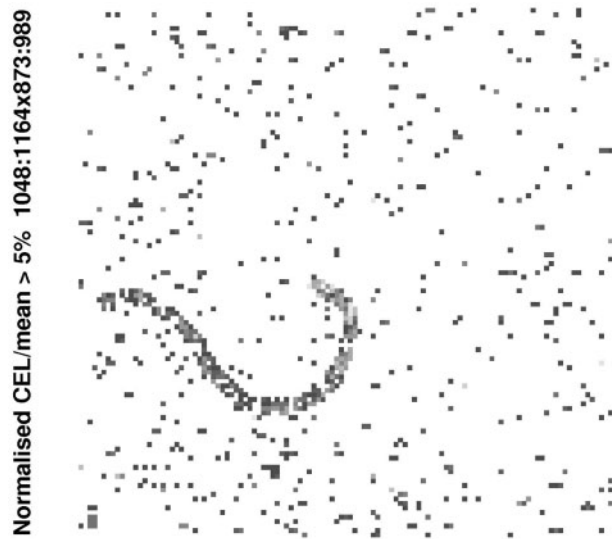


Figure 3: Example of linear defect in GEO CEL file (inverted question mark). The fragment is 0.1×0.1 of an Affymetrix GeneChip Human Genome UI33 Plus 2.0 Array from a female colon cancer. The log image has been enhanced by plotting only quantile normalised probe values which are $>5\%$ bigger than average.

leave their footprint over a considerable region of the chip. There are a number of candidates for the choice of reference value.

Use of replicates

When an array is one of a set of technical replicates, some average of the values in the same location in the remaining replicates can be calculated. The Harshlight package [8] uses the median value of a group of replicates as the reference value.

Use of non-replicates

Information is available for the same type of array for data collected from many different sources. It is therefore possible to assign a representative value to each location in the arrays. Indeed, for many array types there are publicly available sets of data generally consisting of arrays that have been analysed by previous researchers and then deposited in a database (such as the NCBI's GEO database) in case they prove of more general use. It is natural that as technology progresses the meaning of 'large scale' should change. Until recently a large-scale study might use dozens or even hundreds of arrays. However, with the advent of large repositories (GEO contains tens of thousands of arrays and is roughly doubling in size every year) studies of data generated from several experiments, or laboratories,

are becoming common. Whilst analysing several hundred GeneChips from over a dozen different studies, Reimers and Weinstein [12] used the 20% trimmed mean of each probe across chips (for the same tissue) as the reference value.

In assessing the characteristics of a particular chip, Reimers and Weinstein [12] used a 'heat map' in which each value was assigned its percentile ranking based on the historical information. Where there are many previous examples of the same type of chip, a logistically simpler alternative to retaining the full array information would be to retain only the median, M and the lower and upper quartiles (L and U). The 'heat' for a value ν would then be provided by calculating $z = (\nu - M)/(U - L)$. Higher resolution can be obtained by obtaining deciles, or centiles, for each probe position. These are alternatives currently being studied by the Essex Bioinformatics group using the thousands of arrays publicly available within GEO, e.g. Figure 4.

Using a single array

Detection using only the values from the current array will be most useful when working with rare or expensive arrays with no replicates and little previously collected data available for comparison. Another important use of arrays without replicates is in the context of proof of principle experiments where experimenters need to show that they can produce sufficient quantities of RNA for future microarray studies. Stem cell differentiation experiments, primary cell cultures that are difficult to maintain and RNA extractions from small biopsies are examples. Even when we have only one array, major flaws defects can still be discovered, using an adaptation of the z -statistic given in the Use of Non-replicates section.

Detecting spatial features

Regardless of whether the chip being studied is one of a group of replicates, or is a singleton, the general approach is the same. A 'code' (whose nature is discussed below) is attached to each cell of the chip. The chip is then scanned to find aggregations of cells that have been assigned the same code—these are then candidates for interpretation as cells displaying spatial flaws. The procedure works best when the number of codes is not large (for example, 5–15). If there are too few codes then huge areas of cells having the same codes would appear by chance. If there are too many codes then it becomes difficult

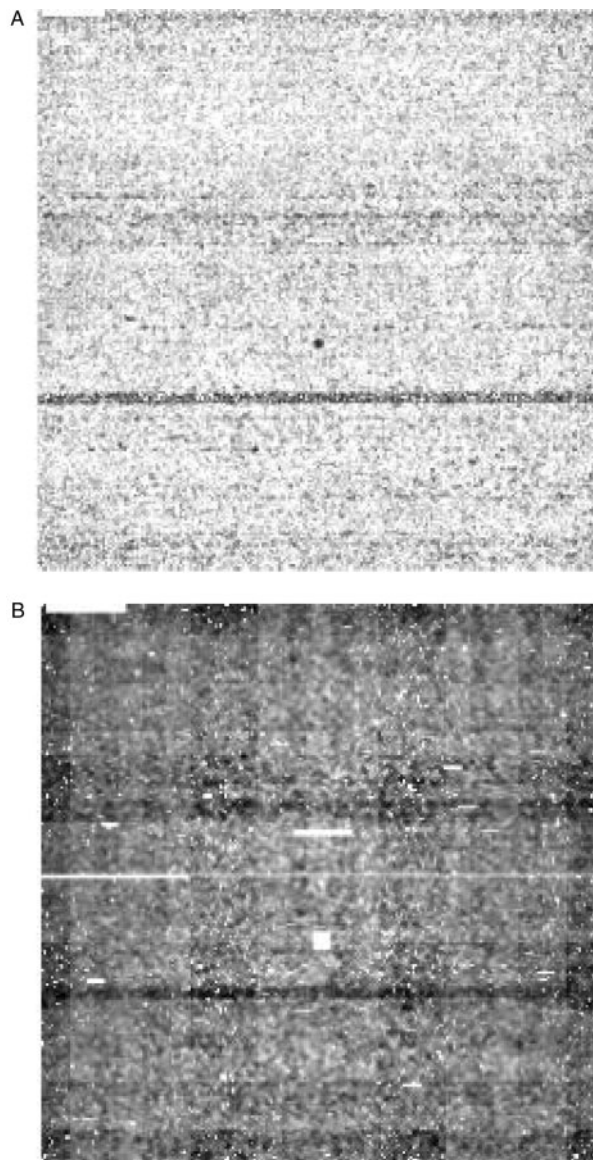


Figure 4: (A) Distribution of suspiciously low probe values in 2754 HG-UI33.Plus2 GEO CEL files. Log of mean. Darker is higher. Black indicates >10% of probes at this location were suspect. Note black square near centre is an artefact. (B) Distribution of suspiciously low probe values in 6685 HG-UI33 GEO CEL files.

to find regions having the same code even in the presence of spatial flaws.

When there are replicates, the code at a location might be the identifier of the replicate having the largest value at that location (after some form of standardisation using quantiles or multiplicative scaling). Alternatively the code might identify the replicate having the smallest value at that location. Perhaps more usefully, the code might have a sign (\pm) identifying not only the replicate having the most

extreme (in some sense) value at that location but also its nature (extremely large, or extremely small).

In the case of a single array being compared with some array of standard values, each cell might be accorded a code indicating the extent of its departure from the standard (for example, the log-fold change expressed to the nearest integer, or the integer part of the z -statistic defined in the Use of Non-replicates section). If there is no standard array available for comparison, then the code might be the log of the ratio of the cell value to the array's median value, again expressed as an integer.

Assessing significance

Even in a good chip, it is possible to find small clusters of identically coded cells by chance. Standard statistical significance levels (e.g. 1% or 5%) make no sense in the context of the hundreds of thousands of tests that might be associated with microarrays, since the experimenter will be overwhelmed by false positives. It is therefore necessary to devise tests for which 'significant' events are events occurring by chance on the order of 0.0001% of occasions, or less frequently.

It is easy to calculate the probabilities of occurrence for an unusual event centred on a specific cell (for example the probability that six or more of the eight cells directly bordering a given cell have the same code). This helps in identifying what might be constituted as an unlikely event. However, calculating the probability that an array contains a compact region of cells having the same codes somewhere within the array is much more difficult, and the simple alternative is to simulate random patterns and observe the proportions of occasions on which events of interest occur [8].

The types of defects

Many authors have attempted classifications of defects by their type or extent. Here we focus primarily on that presented by the Harshlight package [8].

Classification by type (Harshlight)

The package reports the defects that it finds under three headings, using specific pattern-recognition methods for each type:

- (i) *Extended defects* are defined as 'covering a large area' with the result that there is 'substantial variation in the overall intensity from one region of the chip to another'. The package authors

recommend, that an array be discarded if the proportion of variations in the error image explained by the background exceeds $\sim 30\%$ then the chip should be discarded. Suárez-Fariña *et al.* [8] find this type of defect to be rare ($\sim 2\%$ of chips), but report that several of the chips in the SpikeIn measurements [13] are affected by such defects.

- (ii) *Compact defects* are described as ‘small connected clusters of outliers’, such as might be created by specks of dirt [14]. Because of their size, the percentage of a microarray that is affected by these is small. Harshlight detects these defects by first creating an ‘outlier’ image of spurious probe values and then uses the FloodFill algorithm [15] to detect clusters of connected outliers. For every flagged pixel in the image, the algorithm recursively looks for other flagged pixels in its neighbourhood. If any are found, these pixels are assigned to the same cluster number, and the process stops when no more connected pixels can be found. Reimers and Weinstein [12] described these as ‘distinctive artefacts’.
- (iii) *Diffuse defects* are described as ‘areas with densely distributed, although not necessarily connected outliers’. Diffuse defects appear as regions of the image in which there are a large number of outliers, when compared to other regions of the image. To outline the area of blemishes, Hubbell *et al.* [8] introduce a closing procedure, a technique used in image processing to close breaks in features of an image [16]. A circular kernel is centred on each pixel and flagged if any of the pixels within the kernel are flagged. This ‘dilation’ pass causes the borders of the defects to grow, and so to accurately determine the boundaries of the defect, an ‘erosion’ pass has the kernel flagged only if all the pixels inside the window are flagged.

Harshlight also aims to minimise the impact that the calculation of compact defects has on the detection of diffuse defects. It achieves this by detecting the areas of compact regions reported within areas assigned to diffuse region. If the ratio of areas is $< 50\%$ then the compact defects are not removed prior to the calculation of diffuse defects.

A saturated blob finder

Blobs of ‘large, spatially contiguous clusters of signal from high-intensity distributions’, have been

observed to frequently impact (10–20%) upon the new generation of SNP, exon and tiling arrays [17]. Song *et al.* [17] developed the Microarray Blob Remover (MBR) interactive tool to counteract the effects of the blobs. Initially, MBR scans the image with a 100×100 square, sliding in steps of 50 probes in both directions, counting the number of probes whose values are above the k th quantile of intensities. If the probes above the threshold occupy $> 50\%$ of the square, MBR rescans the square with a circle of radius 20, sliding in steps of two probes. If more than $p\%$ of the probes in the circle have intensities above the $(k - 5)$ th quantile, then all the probes in the circle are flagged as being within a blob and can then be ignored by subsequent downstream analysis. We expect the method outlined by Song *et al.* [17] can be readily adapted to detecting contiguous regions of very low signal.

Classification by context

Reimers and Weinstein [12], study the variations in the magnitudes of the 20th and 80th percentiles across an array. They observe that there may be:

- (i) Shifts in the background signal level (i.e. of the 20th percentile)
- (ii) Shifts of signal scale (i.e. of the differences between the 80th and 20th percentiles)

They report that they find many such irregularities in otherwise flawless chips.

Flaws detected as model residuals

Expression levels vary both by probe set and from one array to another. The Bioconductor package affyPLM [10], permits the user a flexible choice of model for use with collections of related or unrelated arrays. As with any model, residuals should be examined for patterns and examples of the often remarkable patterns that may be found (reflecting a variety of flaws) are presented [11].

THE IMPACT OF THE DEFECTS

Image defects will likely affect a subset of the probes within a probe set, and thus impact upon the expression measure assigned to the probe set. Suárez-Fariña *et al.* [8] used Harshlight prior to calculating expression measures through the MAS5 [18] and GCRMA [19] algorithms and discovered that MAS5 is strongly affected by blemishes, whereas GCRMA

is affected to a smaller, but still relevant, level. This was accounted for by noting the robust averaging employed by GCRMA. Reimers and Weinstein [12] used MAS5 and RMA [20] in the affy package of Bioconductor and reported that RMA is notably more robust than MAS5 to the negative effects of blemishes typically seen in GeneChips. However, RMA does worse than MAS5 for larger blemish perturbations. Reimers and Weinstein [12] also observe that the MAS5 PM correction reduces regional biases in scale factor and background, whereas RMA does not.

Suárez-Fariña *et al.* [8] also benchmarked the impact of Harshlight on the Affycomp suite, an ensemble of tests on spike-in measurements that it is used to benchmark expression measure calculations [21]. Pre-processing the SpikeIn133 dataset with Harshlight results in a significant improvement, with ROC curves showing an improved ability to detect true positives for a given false-positive rate using both MAS5 and GCRMA.

Interestingly, Suárez-Fariña *et al.* [8] also discuss the possibility that ‘diffuse defects’ are probe-sequence dependent. This may have important ramifications because widely used expression measures such as GCRMA, which build statistical models to describe cross-hybridisation, may have inadvertently used probes whose signals have an extra noise component caused by diffuse defects. We are unsure of the magnitude of such an effect but its possibility suggests that developers should use a package such as Harshlight to clean up spike-in data prior to using it for calibration purposes.

Song *et al.* [17] focus on detecting blobs of saturated emission. After removing the blobs, Song *et al.* [17] report a significant improvement in the sensitivity and false-discovery rate of a tiling array analysis.

Some authors do not attempt to correct for defects; for example, Reimers and Weinstein [12] suggest their method is most useful in the quality assessment step rather than as a resource aiding the correction of arrays. However, the Harshlight package [8] does try to correct the spatial biases seen in GeneChips. Harshlight attempts a repair of locations identified as flawed by replacing the reported value by the reported value of all the replicates at that location. Arteaga-Salas and Upton [22] have recently presented a different alternative, the Local Probe Effect followed by a Complementary Probe Pair adjustment (LPECPP), that makes use of the

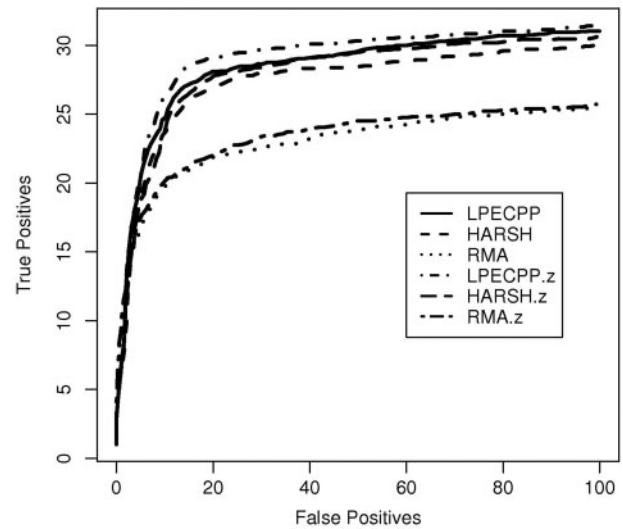


Figure 5: A receiver–operator curve measuring the impact of image processing on the ability of the RMA algorithm to detect genes having a fold change of 2. The ROC curves are generated using affycomp [21] on the HGU 133A spike-in measurements [13]. Harsh shows the results for RMA following pre-processing with Harshlight [8] and LCECPP shows the results for RMA following pre-processing with LPECPP [22]. The z measures have used the Zuzan [4] algorithm to transform the DAT files into CEL files, whereas the others use the standard Affymetrix transformation.

information found in perfect match probes to adjust their corresponding mismatch probe, and vice versa. We have found that after both types of probes have been adjusted, the majority of the spatial blemishes will be removed.

We report a preliminary benchmarking of image processing using the Affycomp package [21], following the suggestion of Suárez-Fariña *et al.* [8]. We find that the grid alignment transformation in moving from a DAT file to a CEL file, described by Zuzan *et al.* [4], acts to improve the sensitivity of RMA, Figure 5, as well as for GCRMA and MAS (data not shown). By either removing [8] or correcting-for [22], image defects in the CEL file, we see a further improvement. We are presently exploring how best to optimise the grid alignment and to correct for the effects of defects.

Spatial defects are already recognised by Affymetrix and the user community. Once defects have been detected, there is the choice of whether to discard the data having the defect, or find some way to determine the true signal after removing the noise caused by the defect. There are a huge number of bits of information on a chip so that a flaw amounting to 5% of the chip appears of

little account. However, there is comparatively little information on any particular gene so losing even one observation may prove a serious loss. If one is looking for subtle changes in the data, when there are only a small number of probes per transcript unit (and there are only 4 probes per exon on the new Exon arrays), then one should not be dealing with corrupted data. In the future, with increasing systems biology use of array data, a far larger portion of the data will be used. In those circumstances, where most of the data are used in a model, and noting all GeneChips contain errors to a greater or lesser degree, the discrepancies caused by image defects can end up making a disproportionate contribution to error. This will make the modelling more difficult.

SUMMARY AND CONCLUSIONS

All GeneChips are affected to some extent by spatially coherent defects. Fortunately image processing has a number of potential beneficial impacts on the downstream analysis of GeneChip data. There are now a number of robust and accurate algorithms that identify the most disabling of defects.

There are a number of algorithms that aim to reliably transform a DAT file into a CEL file. However, several of these methods are yet to be published and we are not aware of work that has benchmarked these algorithms against each other.

There are a number of packages such as Harshlight [8], which now enable biologists to scan their chips for defects. Moreover, because Harshlight is integrated within bioconductor, scientists can now integrate image-processing analysis with downstream analysis routines such as the calculation of expression measures and the determination of differentially expressed genes.

At present many algorithms identify spatially adjoining regions of defects, and treat the strength of the defect as a binary value, all or nothing. The resulting image is black and white, whereas in reality a heat map of defect contours, particularly towards the boundaries, may be more realistic. This will be particularly relevant if the defects result from inhomogeneous mixing of transcripts. Using large surveys, such as stored in GEO, will enable reliable estimates of the typical level of spatial defects.

A desirable resource is something similar to the excellent affycomp [21], which benchmarks different image-processing algorithms on a SpikeIn dataset. There is no ‘true’ knowledge of defects, but it is

important to establish where algorithms agree and disagree about the identity and strength of the defects upon a single dataset. Since image processing can be used to clean-up data, it will improve the accuracy of some gene-expression measurements and the implications biologists draw from them. Moreover, there is a clear link between modelling chemical saturation [6], with the observation of significant spatial variations in the upper values seen for PM and MM [12]. There is also a link between modelling cross-hybridisation events [19] with the observation of significant spatial variations in non-specific background [12]. It is our opinion that image-processing, statistical models of the physics of GeneChips, and the calculation of expression measures, should be treated together.

Key Points

- We demonstrate that correcting for defects makes a noticeable improvement in the sensitivity of Affymetrix GeneChips.
- We expect that image-processing methods will be most beneficial when used to analyse the new generation of arrays such as SNP chips and exon arrays.
- We believe this is the first overview of analysing Affymetrix GeneChips using image-processing methods.

Acknowledgements

J.M.A.S. is supported by a CONACYT scholarship (178596) and W.B.L. is supported by a grant from the BBSRC (BB/E001742/1).

Conflict of interest statement: H. Zuzan reports a financial interest in a patent for estimating probe cell locations.

References

1. Barrett T, Suzek T, Troup D, *et al.* NCBI GEO: mining millions of expression profiles-database and tools. *Nucleic Acids Res* 2005;**33**:D562–6.
2. Gentleman R, Carey V, Huber W, *et al.* *Quality assessment of Affymetrix GeneChip Data, in Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Springer, 2005.
3. Schadt EE, Li C, Ellis B, Wong WH. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression data. *J Cell Bio Supplement* 2001;**37**:120–5.
4. Zuzan H., Blanchette C., Dressman H. *et al.* [<http://ftp.stat.duke.edu/WorkingPapers/02-05.pdf>]
5. Baggerly K.A. [http://www.mdanderson.org/pdf/biostats_utmdabtr00503.pdf]
6. Hekstra D, Taussig AR, Magnasco M, Naef F. Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Res* 2003;**31**:1962–8.
7. Naef F, Magnasco M. Solving the riddle of the bright mismatches. *Phys Rev E* 2003;**68**:011906.
8. Suárez-Fariña M, Pellegrino M, Wittkowski K, Magnasco M. Harshlight: a corrective make-up program for microarray chips. *BMC Bioinformatics* 2005;**6**:294.

9. Upton GJG, Lloyd JC. Oligonucleotide arrays: Information from replication and spatial structure. *Bioinformatics* 2005;**21**: 4162–8.
10. Bolstad BM, Collin F, Brettschneider J, *et al.* Quality assessment of Affymetrix GeneChip Data. In: Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S, (eds). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Springer, 2005;33–47.
11. Bolstad B. Website: <http://plmimagegallery.bmbolstad.com/index.html>, (7 November 2007, date last accessed)
12. Reimers M, Weinstein JN. Quality assessment of microarrays: Visualization of spatial artefacts and quantitation of regional biases. *BMC Bioinformatics* 2005;**6**:166.
13. AffyWebdata: Affymetrix Website. http://www.affymetrix.com/support/technical/sample_data/datasets.affx, (7 November 2007, date last accessed)
14. Suárez-Fariña M, Haider A, Wittkowski K. Harshlighting small blemishes on microarrays. *BMC Bioinformatics* 2005;**6**:65.
15. Cormen TH, Leiserson CE, Rivest RL, Stein C. *Introduction to algorithms: 2nd edition*. Cambridge, MA: The MIT Press, 2001
16. Russ JC. *The Image Processing Handbook: 4th Edition*, Boca Raton, FL: CRC Press, 2002.
17. Song J, Maghsoudi K, Li W, *et al.* Microarray blob-defect removal improves array analysis. *Bioinformatics* 2007;**23**: 966–71.
18. Hubbell E, Liu W, Mei R. Robust estimators for expression analysis. *Bioinformatics* 2002;**18**:1585–92.
19. Wu Z, Irizarry RA, Gentleman R, *et al.* A model based background adjustment for oligonucleotide expression arrays. *J Am Stat Assoc* 2004;**99**:909–17.
20. Irizarry RA, Bolstad BM, Collin F, *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 2003;**31**:15.
21. Cope LM, Irizarry RA, Jaffee HA, *et al.* A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics* 2004;**20**:323–31.
22. Arteaga-Salas J.M. and Upton GJG. A normalisation method for oligonucleotide arrays motivated by spatial biases. *Proceedings of the 26th Leeds Annual Statistical Research Workshop*, University of Leeds, 2007; 26, pp. 95–8.