Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Detecting artifacts on SNP chips

Maurizio Pellegrino*, Mayte Suárez-Fariñas*, Marcelo O. Magnasco, and Knut M. Wittkowski

***Equal contribution**

Abstract

Microscopists are familiar with many blemishes that fluorescence images can have due to dust and debris, glass flaws, uneven distribution of fluids or surface coatings, etc. Microarray scans do show similar artifacts, which might affect subsequent analysis. We developed a tool, *Harshlight*, for the detection and masking of blemishes in HDONA microarray chips. *Harshlight* uses a combination of statistic and image processing methods to identify defects. We demonstrate that *Harshlight* can be widely used for chips with different technologies thanks to its user-tunable parameters. Here we report its application to SNP microarrays.

Background

Analysis of hybridized microarrays starts with scanning the fluorescent image. The quality of data scanned from a microarray is affected by a plethora of potential confounders, which may act during printing/manufacturing, hybridization, washing, and reading. For high-density oligonucleotide arrays (HDONAs) such as Affymetrix GeneChip® oligonucleotide (Affy) arrays, each chip contains a number of probes specifically designed to assess the overall quality of the biochemistry. Affymetrix software and packages from Bioconductor project for R [1] provide for a number of criteria and tools to assess overall chip quality, such as percent present calls, scaling factor, background intensity, raw Q, and degradation plots. However, these criteria and tools have little sensitivity to detect *spatially localized* artifacts which can substantially affect the sensitivity of detecting physiological (i.e., small) differences. In [2] we presented a simple method to "harshlight" these blemishes in HDONAs chips to render them evident.

The method produces an Error Image (**E**) for each chip, which indicates the deviation of this chip's log-intensities from the other chips in the experiment. Formally, **E** is calculated as $\mathbf{E}^{(i)} = \mathbf{L}^{(i)} - median_j\ \mathbf{L}^{(j)}$ where $\mathbf{L}^{(j)}$ is the log-intensity matrix of chip *i*. Given that the intensity of each cell is highly determined by the sequence of the probe [3], this deviation should be near zero except for the probes belonging to the probe sets related to genes that are differentially expressed. The algorithm detects outliers based on **E**.

Subsequently [4], we developed an R-package, *Harshlight*, which automatically spots blemishes on Affymetrix HUG95 and HUG133 gene expression arrays based on suspicious patterns in the error image (**E**) using diagnostic tests based on both image processing and statistical approaches [5]. For analytical purposes, the defects were divided in three sets (extended, diffuse, and compact), due to the size and probably the nature of the defects (Figure 1).

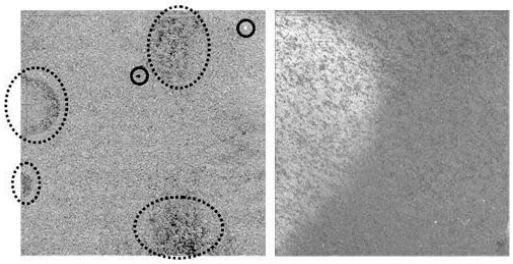

Figure 1 Three types of defects. Left. Solid circles mark compact defects and dashed circles outline areas with diffuse defects. Right. A chip with a large defect that invalidates its further use.

The need for such an algorithm was validated by statistical analysis of large collections of gene expression chips: all chip collections we examined have substantial numbers of small localized defects, and the damaging effect of such localized defects was assessed by simulating localized damage to curated chip collections. The effectiveness of the algorithm in detecting such defects and correcting the damage to gene expression values was demonstrated by

applying it to the Affymetrix callibration datasets using the Affycomp suite. In this manuscript, we shall describe how the tunable parameters of *Harshlight* permit its use for different chip technologies and families; specifically, we present analysis of the SNP chip GeneChip® Human Mapping 100K Set [6].

Results

Harshlight has been extensively and successfully tested on gene expression HDONAs [4]. Due to the wide variety of existing chip technologies and applications, we wanted to extend our analysis to other chips and validate our method. The GeneChip® Human Mapping 100K Set (single nucleotide polymorphism or SNP chip) is comprised of a set of two arrays and covers 100,000 SNPs present in the human genome, making it a valuable tool for genome-wide analysis. For each position tiled within each SNP, all *four* possible nucleotides are represented and the probe quadruples are randomly distributed across the chip. For gene expression HDONAs, each probe pair in the probeset representing each gene contained only two probes, a perfect match (PM) and a single MM probe. Due to this and other differences in technology, the parameters used to detect defects in HDONAs were modified to analyze SNP chips.

We found that the analysis of diffuse defects can greatly benefit from fine-tuning parameters to distinct chip families. Diffuse blemishes are characterized by areas with a high density of blemished probes (outliers), most likely due to defects in the hybridization stage [4]. In the case of diffuse defects, the outliers are defined as pixels whose intensity values are higher (bright outliers) or lower (dark outliers) than a certain percentage compared to the expected intensity.

For every pixel *i*, the proportion of outliers in the surrounding window _$_i$ is computed as:

$$p_i = \frac{1}{n_S} \sum_{j \in \Theta_i} O_j$$

A binomial test is then used to look for areas in which the number of outliers p_i is higher than the overall proportion of outliers p_o in the whole **E**.

$$\mathbf{D}_i = \begin{cases} 1 & \text{if } p_i > b_{1-\alpha}(p_o, n_S) \\ 0 & \text{otherwise} \end{cases}$$

Where 1 indicates an outlier pixel.

Given the technology behind SNP chips (i.e. a higher number of probes for each sample) we would expect more diffuse defects present on the surface of the chips. Therefore, we adjusted the parameters accordingly, as shown in **Table 1**.

Table 1 Parameters used in diffuse defects detection

Parameters	HDONAs	SNP
Threshold bright defects	40% more than original value	75% more than original value
Threshold dark defects	35% less than original value	67% less than original value
binomial test: p-value	0.001	0.0001

The definition of outliers plays a critical role in this analysis. To account for a higher probability of finding diffuse defects, we increased the difference from expected intensity for the definition of bright and dark outliers. In addition, we decreased the p-value for the binomial test, to account for a change in sensitivity needed.

From our analysis, we could successfully identify diffuse defects on SNP chips. *Harshlight* can reliably detect both evident defects (Figure 2), and blemishes not apparent to the naked eye, across a wide variety of chip designs.

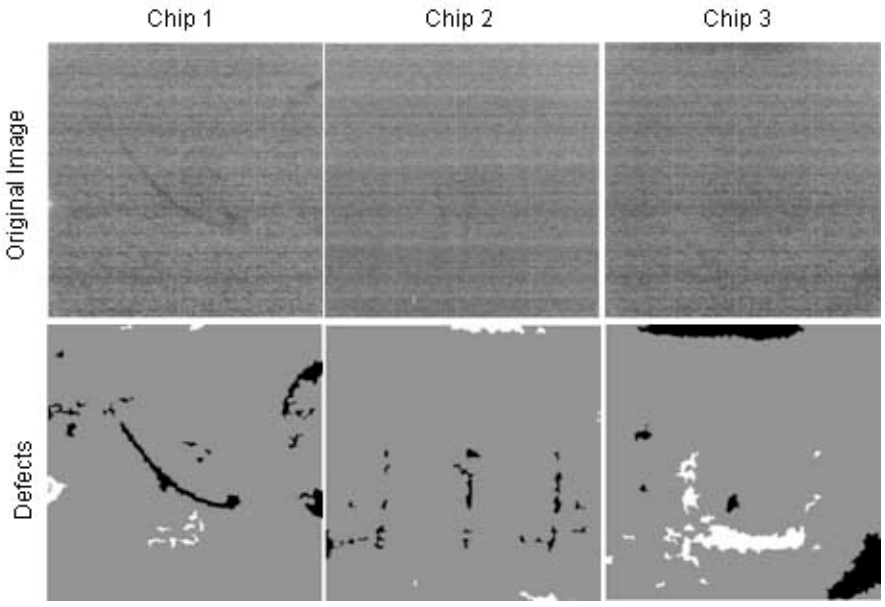

Figure 2: Examples of SNP chip analysis. Chip 1: the original chip image shows defects easily detectable by naked eye. These blemishes are recognized effectively as diffuse defects. Chip 2 and 3: the original chip image doesn't show any apparent defect. *Harshlight* can efficiently detect and mask diffuse defects. The three chips were analyzed with the same parameter set.

Conclusions

Localized blemishes on microarray chips can impact further data analysis on any experimental design where subtle changes are to be measured. *Harshlight* is a useful tool to automatically detect and mask blemishes on microarray chips. Even though tested on HDONAs chips, *Harshlight* is able to reliably identify defects on chips built with different technologies, thanks to its user-tunable parameters.

References

1. Ihaka R, Gentleman R: **R: A language for data analysis and graphics.**
Journal of Computational and Graphical Statistics 1996, **5**(3)**:**299-314.

2. Suárez-Fariñas M, Haider A, Wittkowski KM: **"Harshlighting" small blemishes on microarrays.**
BMC BIOINFORMATICS 2005, **6:65.**

3. Naef F, Magnasco MO: **Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays.**
Phys Rev E Stat Nonlin Soft Matter Phys 2003, **68**(1 Pt 1)**:**011906. Epub 2003 Jul 16.

4. Suárez-Fariñas M*, Pellegrino M*, Wittkowski KM, Magnasco MO: **Harshlight: a "corrective make-up" program for microarray chips.**
*Equal contribution
BMC BIOINFORMATICS 2005, ***6:294.***

5. AffyWebdata: **Affymetrix website.**
[http://www.affymetrix.com/support/technical/sample_data/datasets.affx]

6. **Affymetrix website.**
[http://www.affymetrix.com/products/arrays/specific/100k.affx]