

Estimation of Probe Cell Locations in High-density Synthetic-oligonucleotide DNA Microarrays

Harry Zuzan¹, Carrie Blanchette², Holly Dressman³,
Erich Huang³, Seiichi Ishida³, Jeffrey R. Marks², Joseph R. Nevins^{3,4},
Rainer Spang¹, Mike West¹, Valen E. Johnson¹

October 16, 2001

Abstract

The source of a spatial contribution to the coefficient of variation in Affymetrix GeneChip[®] probe cell summaries was determined to be caused by a misalignment of probe cell locations in the hybridisation image. Described, is an algorithm that corrects for misalignment resulting in improved allocation of pixel intensities to probe cells and a substantial reduction in the variance of pixel intensities attributed to individual probe cells.

¹Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708
Departments of ²Surgery and ³Genetics, Duke University Medical Center, Durham,
NC 27710

⁴Howard Hughes Medical Institute, Durham, NC 27710

High-density synthetic-oligonucleotide DNA microarrays (HSDMs), which are manufactured by Affymetrix under the name GeneChip[®] probe array, have miniaturised the physical area required to hybridise an RNA sample to DNA probes. On an HSDM surface, hundreds of thousands of copies of a DNA probe sequence can be fabricated in a precise location called a probe cell, and it is common for HSDMs to contain more than 400,000 probe cells. Detailed descriptions of HSDMs are found in Lockhart et al.[1] and Lipshutz et al.[2]. After a sample of fluorescent labelled RNA is hybridised to probes on an HSDM, the hybridisation data is extracted by laser confocal fluorescence scanning and recorded as a large two-dimensional array of 16 bit integers. The extracted array of integers can be presented as a grayscale image where each pixel in the scan maps to a small region on the HSDM. A low resolution image of an HSDM is shown in Figure 1. A high resolution section of Figure 1 is shown in Figure 2, where it can be seen that probe the cells are regularly spaced in a rectangular grid. In Figure 2 each probe cell occupies an area approximately 8×8 pixels and by visual inspection, most pixels can be attributed to a single probe cell. In practice, the process of attributing pixels and pixel intensities to probe cells must be performed as an automated post-processing step on the extracted two-dimensional array of pixel data. Affymetrix's software uses a proprietary algorithm to attribute pixels to probe cells. To summarise a hybridisation, Affymetrix reports three statistics for each probe cell: 1) The number of pixels attributed to the probe cell; 2) A number representative of the intensity of the probe cell's fluorescent response to hybridisation. The default choice of this number is the 75th percentile of pixel intensities attributed to the probe cell; and 3) The standard deviation of pixel intensities attributed to the probe cell.

The HSDM used to obtain the image in Figure 1 contains an array of 534×534 probe cells. This design of HSDM is manufactured under the name MU11KA and is designed for analyses of murine gene expression. The RNA hybridised to this HSDM was obtained from a murine tissue culture and Affymetrix's software was used to extract the 4733×4733 array of hybridisation data and compute the hybridisation summary. In a preliminary analysis of the hybridisation summary we investigated possible sources of error contributing to the standard deviation of intensities of pixels attributed to probe cells. Figure 3 shows a plot of standard deviation versus 75th percentile for each of the 534^2 probe cells. From this plot it is evident that pixel intensity variance increases with pixel intensity. The information in Figure 3 was used to compute the coefficient of variation with respect to the 75th percentile for each probe cell. The corresponding 534×534 array of these coefficients is presented as a grayscale image in Figure 4. The diagonal bands in Figure 4 reveal a spatial contribution to the coefficient of variation and we suspected that this pattern was due to inaccurate estimates of probe cell locations. In other words, we suspected that a misalignment problem caused pixels and their intensities to be incorrectly attributed to probe cells. This apparent misalignment motivated the development of an alignment algorithm that provides accurate estimates of probe cell locations, resulting in improved attribution of pixel intensities to probe cells.

The observation used to produce the HSDM image in Figure 1 was selected from a larger experiment that used 18 HSDMs and is typical of the remaining 17 observations. We will use the extracted HSDM data in Figure 1 as an example to illustrate

our algorithm and analyse results.

Results and discussion

Extracted HSDM pixel data is not segmented. The potential range of intensities in the extracted 4733×4733 scanned image of 16 bit hybridisation intensities had a potential range of (0,65535). In our example hybridisation image, the minimum pixel intensity was 93 and the maximum pixel intensity was 46192. The maximum appears to be an upper threshold or saturation level that could not be exceeded. All of the HSDM images in our set of 18 had similar minimum and maximum intensities. Using the top left corner of the image as the coordinate origin and letting the first coordinate index pixels from top to bottom and the second coordinate index pixels from left to right, the corners of the array of probe cells were located, by visual inspection, at the coordinates, top left (233,241), top right (226,4504), bottom left (4500, 257) and bottom right (4492, 4519). Between these corner positions, uniformly spaced probe cells would occupy areas slightly smaller than 8×8 pixels.

Each pixel in the scanned image of an HSDM represents a small region on the hybridisation surface of the physical HSDM that could be interior to a probe cell, straddle as many as four probe cells or be partly or entirely in the border area surrounding the array of probe cells. Evident in HSDM images is the effect of a blurring process. Thus, while an HSDM image is extracted, each pixel accumulates signal not only from the area of the HSDM it represents but also from a small surrounding region. By the same blurring process, each pixel loses signal to pixels nearby. Due to the discrete approximation of the HSDM surface provided by pixels and the effect of the blurring process, an HSDM image is not segmented by probe cells, even though the physical HSDM surface is segmented. The lack of image segmentation has the greatest effect on intensities of pixels representing regions on or near the perimeter of probe cells, in the sense that these pixel intensities do not represent signal accumulated from a single probe cell. As a consequence, pixels near the edges of probe cells may need to be discarded when computing hybridisation summaries. Accurate estimates of where probe cells are located with respect to the pixel grid in the HSDM image are required for the remaining pixel intensities to be representative of probe cell response to hybridisation.

Alignment must accomodate a spatial deformation. Because probe cells are laid out in a rectangular array, the first step in estimating probe cell locations is to identify the coordinates of the probe cells at the four corners of the array. Once the corner locations are established, locations of the remaining probe cells can be estimated by linear interpolation. Past experience has provided us with evidence that probe cells are not equally spaced and linear interpolation can be inaccurate by as many as three pixels in both the vertical and horizontal directions. But deviations from an interpolated lattice were in all cases gradual, and can be modelled as a continuous deformation of the HSDM as has been noted by Schadt et al.[3]. To accomodate this deformation we used interpolated locations as initial estimates of probe cell locations and employed an iterative algorithm that gradually translated the estimated locations of individual probe cells while maintaining strong local lattice relationships among neighbouring probe cells.

To motivate the alignment algorithm, which is similar to the “facet model” described in Laading et al.[4], we refer to Figure 2 where probe cells boundaries are evident by abrupt changes in pixel-to-pixel intensities near neighbouring probe cells. A 6×6 array of pixels superimposed over a probe cell, will tend to have smallest variance if it is aligned to the central region of the probe cell and misalignment will in most cases cause an increase in variance. Our alignment algorithm takes into account the tendency of misalignment to increase this variance. Our alignment algorithm also takes into consideration the local lattice structure of the gradually deformed HSDM. Given an estimate of the location of a probe cell, our algorithm samples several additional nearby locations as prospective revised estimates. The choice of location to retain as the revised estimate is the location that best reduces variance and maintains the local lattice structure with respect to the current estimated locations of neighbouring probe cells. Since it is advantageous to choose both the prospective location with the smallest variance and the prospective location that best maintains local lattice structure, we employ a penalty for each of these two criteria and base our decision on which prospective location to choose as that which minimises a weighted average of the two penalties. Once the estimate of a probe cell’s location has been revised, it may be revisited and is subject to future revision after the estimated locations of nearby probe cells have been revised.

The alignment algorithm. We now describe the implementation of our alignment algorithm in detail. Let j be a variable that indexes the array of probe cells and let c_j be the current estimate of the coordinates of the centre of probe cell j . Although pixels occupy discrete locations, we consider the elements of c_j to be continuous. Let \mathcal{N}_j be the set of indices of the eight neighbours of probe cell j and let \bar{c}_j be the centroid of $\{c_k\}_{k\in\mathcal{N}_j}$. Based on the criterion of retaining local lattice structure, the optimal revised estimate of c_j is \bar{c}_j and we use \bar{c}_j as the centre of a 3×3 rectangular grid of evenly spaced locations which we sample in order to propose revised estimates of c_j . Call these nine locations a_{uv} , with $u, v \in \{-1, 0, 1\}$, $a_{00} = \bar{c}_j$ and let $\delta \in (0, 1]$ be the distance separating adjacent locations. At each a_{uv} , the penalty for locating the probe cell out of alignment with its neighbours is $t_{uv} = \sqrt{u^2 + v^2}$ and the variance, s_{uv}^2 , of pixel intensities is computed within a 6×6 pixel boundary centred at each a_{uv} . This 6×6 pixel boundary can partially include pixels it intersects and the contribution of pixel volumes to variance is weighted according to their partial areas within the boundary. The decision of which a_{uv} to choose as the revised estimate of c_j is based on minimising a weighted average of the penalties t_{uv} and s_{uv}^2 . After evaluating alignments over trial runs, we found it best to first log transform the pixel data and an effective weighted penalty to be $\delta t_{uv} + 5s_{uv}^2/S^2$, where S^2 is the mean of s_{00}^2 over all probe cells.

Prior to the first iteration of the alignment algorithm, initial probe cell locations were estimated by interpolation, S^2 was computed and δ was set to 0.5. Call this initialisation the completion of iteration 0. For each subsequent iteration, the probe cell locations, c_j , $j = 1, \dots, 534^2$ were updated sequentially. In each case, the revised c_j immediately replaced the estimate from the previous iteration as did the contribution of s_{00}^2 to S^2 . Thus, each computation of location \bar{c}_j was based on the most recently revised members in $\{c_k\}_{k\in\mathcal{N}_j}$. After each iteration, δ was decremented by

0.05 and the iterations ceased when δ was no longer greater than 0.

Aligning to log transformed data. Not all probe cells in an HSDM image provide the same amount of information regarding their locations. Many probe cells record little or no RNA hybridisation, and hence, little or no information about their boundaries. In these cases it is expected that any s_{uv}^2 would constitute a similar penalty for all combinations of u and v and the revised c_j would tend to be \bar{c}_j . The alignment algorithm weights the penalties so that in early iterations, probe cells carrying substantial information about their boundaries drive the deformation of the array of estimated probe cell locations. Decrementing the value of δ reinforces the smoothness of the deformation in later iterations. Use of the logarithm of pixel values encourages probe cells that provide substantial information about their boundaries to drive the deformation uniformly.

Evidence of improved alignment. In the hybridisation summary of our example HSDM provided by Affymetrix’s software, the 75th percentile and standard deviation of almost all probe cells were computed from 36 pixels. In order to compare our hybridisation summaries to those provided by Affymetrix, the mean of the 6×6 pixel regions used to compute variance when estimating probe cell locations was used represent probe cell intensities. We prefer to use the mean to represent probe cell intensity. In practice, we could have computed the mean of a smaller region, say, 5×5 pixels in size, if we thought those means would more accurately estimate probe cell intensities.

The scatter plot in Figure 5 shows the standard deviation versus mean of pixel intensity allocated to each probe cell using the initial probe cell locations obtained by interpolation. The most notable difference between Figures 5 and 3 is how means and percentiles differ in behavior as saturation is approached. The mean squared error in our interpolated probe cell locations was 2595497.24 compared to 1369705.03 in Affymetrix’s. Figure 6 is the equivalent of Figure 5 but after our alignment algorithm was applied. In Figure 6 it appears that the coefficient of variation has improved due to better estimates of probe cell locations. The mean squared error in the aligned probe cell locations was reduced to 616329.64. This is a large decrease, but a decrease in pixel variance within probe cells is an expected outcome of the application of our alignment algorithm and does not indicate improved alignment without corroboration. Figures 5 and 6 are both fan shaped plots and further comparisons are difficult because the points are too numerous to be labelled. A comparison of means prior to and after alignment is shown in Figure 7. This plot indicates that nothing drastic has happened to probe cell means during alignment. There is a consistent transition from more conservative estimates of probe cell means at the low end to larger estimates of probe cell means at the high end indicating that estimates of probe cell means may be less affected by neighbouring probe cells using the aligned probe cell coordinates. The average translation of estimated probe cell centres during alignment was 1.031 pixels and the maximum was 2.685 pixels. Finally, to corroborate that we have improved estimates of probe cell mean intensities from our alignment algorithm, in Figure 8 we plotted the coefficients of variation for the revised estimated locations of probe cells as an image. The only areas of this image that indicate spatial pat-

terns in the coefficient of variation are areas where pixels are saturated and a strip at the bottom where hybridisation was weak or non-existent, as can be verified in Figure 1. The spatial contribution to the coefficient of variation seen in Figure 4 has been removed and we conclude that this is because our alignment algorithm provides improved estimates of probe cell locations.

Accuracy and reproducibility of probe cell intensities. Figures 9 and 10 plot Affymetrix’s 75th percentiles versus our pre- and post-alignment means. The average absolute difference between Affymetrix’s 75th percentiles and our pre- and post-alignment means were 539.3 and 606.54 respectively. Although it is tempting to conclude that a linear relationship exists in both of Figures 9 and 10, the fact that almost all of these probe cells contain different probe sequences must be considered. Each probe cell binds target RNA with an affinity that is unique to its probe sequence and the apparent strength of the linear relationship may be due to the broad range of probe-target affinities present on an HSDM. The disparity encountered by choosing the mean in place of the 75th percentile combined with improvements in alignment may prove to be large in relation to a given probe cell and its probe sequence. When deciding how individual probe cell intensities should be represented, accuracy is at stake. This is in contrast to comparisons of entire hybridisations, where reproducibility is at stake [5]. The need for accuracy suggests that a measure of central tendency such as the mean or median should be used to represent probe cell intensity. In future experiments, where suitable replications of hybridisations are available, we will study the extent of improvement in reproducibility given improvements in accuracy.

References

- [1] Lockhart, D.J. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14**, 1675–1680. (1996)
- [2] Lipshutz, R.J., Fodor, S.P.A., Gingeras, T.R. & Lockhart D.J. High density synthetic oligonucleotide arrays. *Nat. Genet.* **21**, 20–24. (1999)
- [3] Schadt, E.E., Li, C., Su, C. & Wong, W.H. Analyzing High-Density Oligonucleotide Gene Expression Array Data *J. Cell. Biochem.* **80**, 192–202. (2000)
- [4] Laading, J.L., McCulloch, C., Johnson, V.E., Gilland, D. & Jaszczak, R.J. A hierarchical feature based deformation model applied to 4D cardiac SPECT data. *Lecture Notes in Computer Science: Information Processing in Medical Imaging*. 266-279 (Springer-Verlag, Berlin; 1999).
- [5] Li, C. & Wong, W.H. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* **98**, 31–36. (2001).

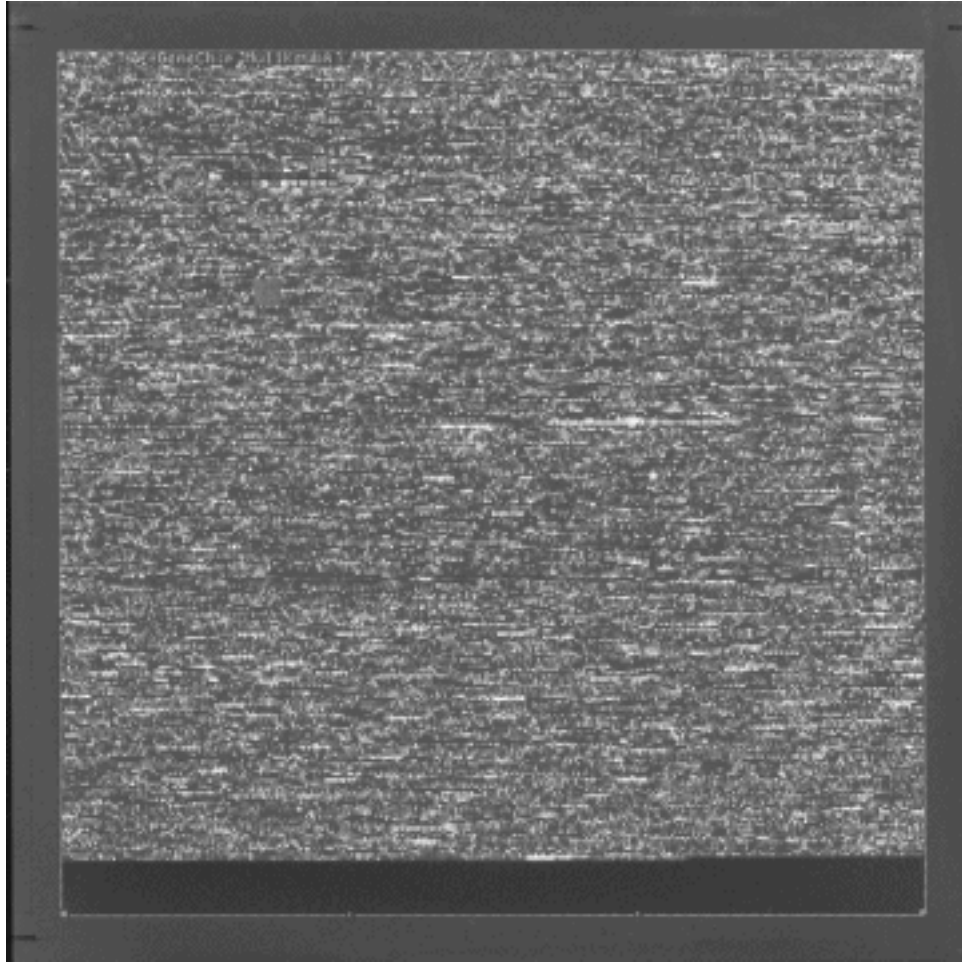


Figure 1: A low resolution log-transformed HSDM image of an MU11KA HSDM containing 534×534 probe cells. Actual size is 4733×4733 pixels . Surrounding the array of probe cells is a border region. The dark band in the lower region of probe cells exhibit little or no hybridisation.

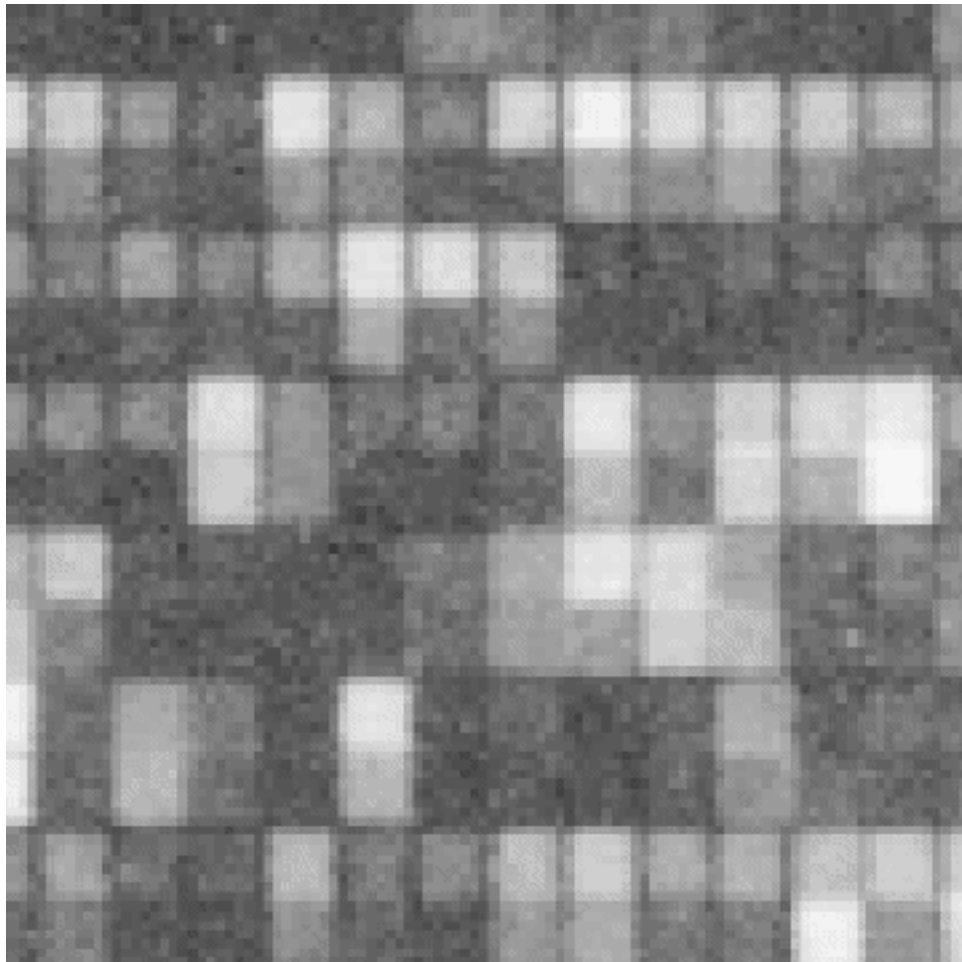


Figure 2: A 100×100 pixel region of the log-transformed HSDM image in Figure 1. Probe cells are evident as squares of 8×8 pixels organised in a lattice. Bright probe cells indicate numerous copies of fluorescent labelled RNA bound to probe DNA sequences.

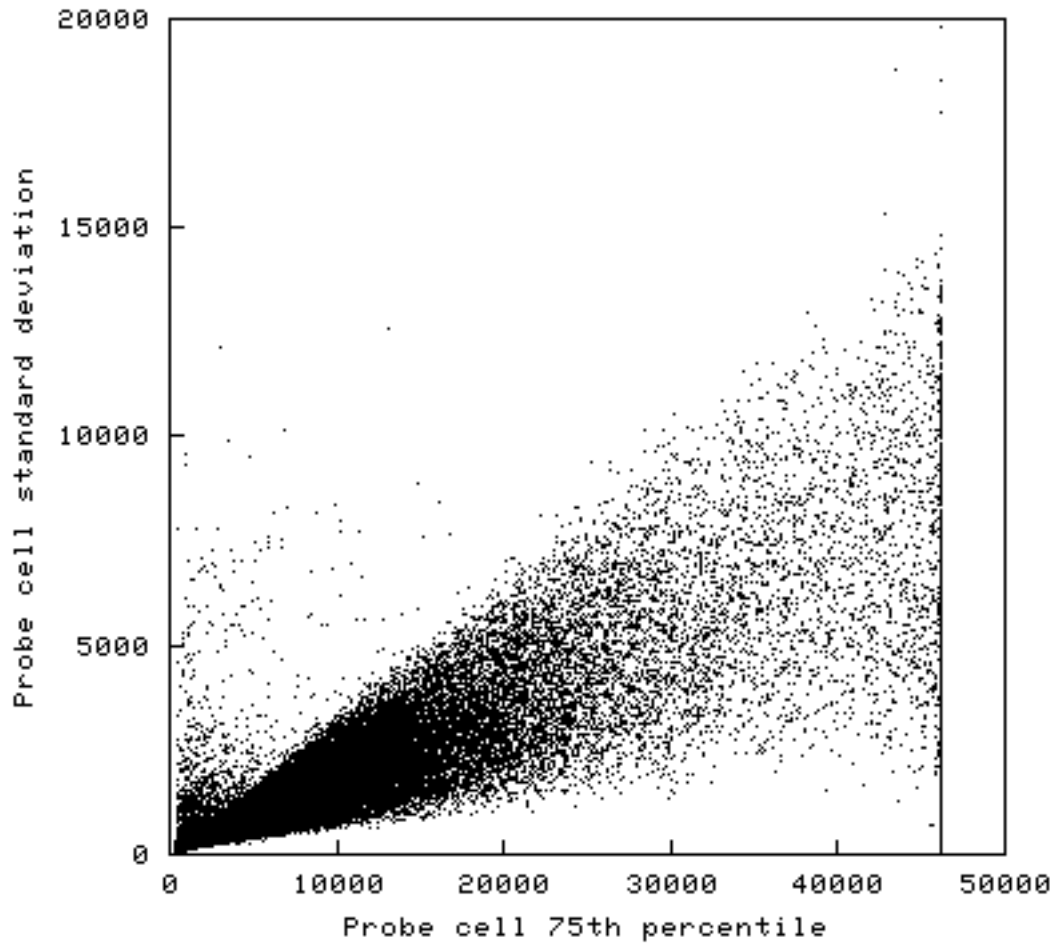


Figure 3: A plot of standard deviations of pixels allocated to probe cells versus their 75 percentiles in the hybridisation summary provided by Affymetrix. Saturation of pixel intensity at 46192 is evident as a vertical line on the right side.

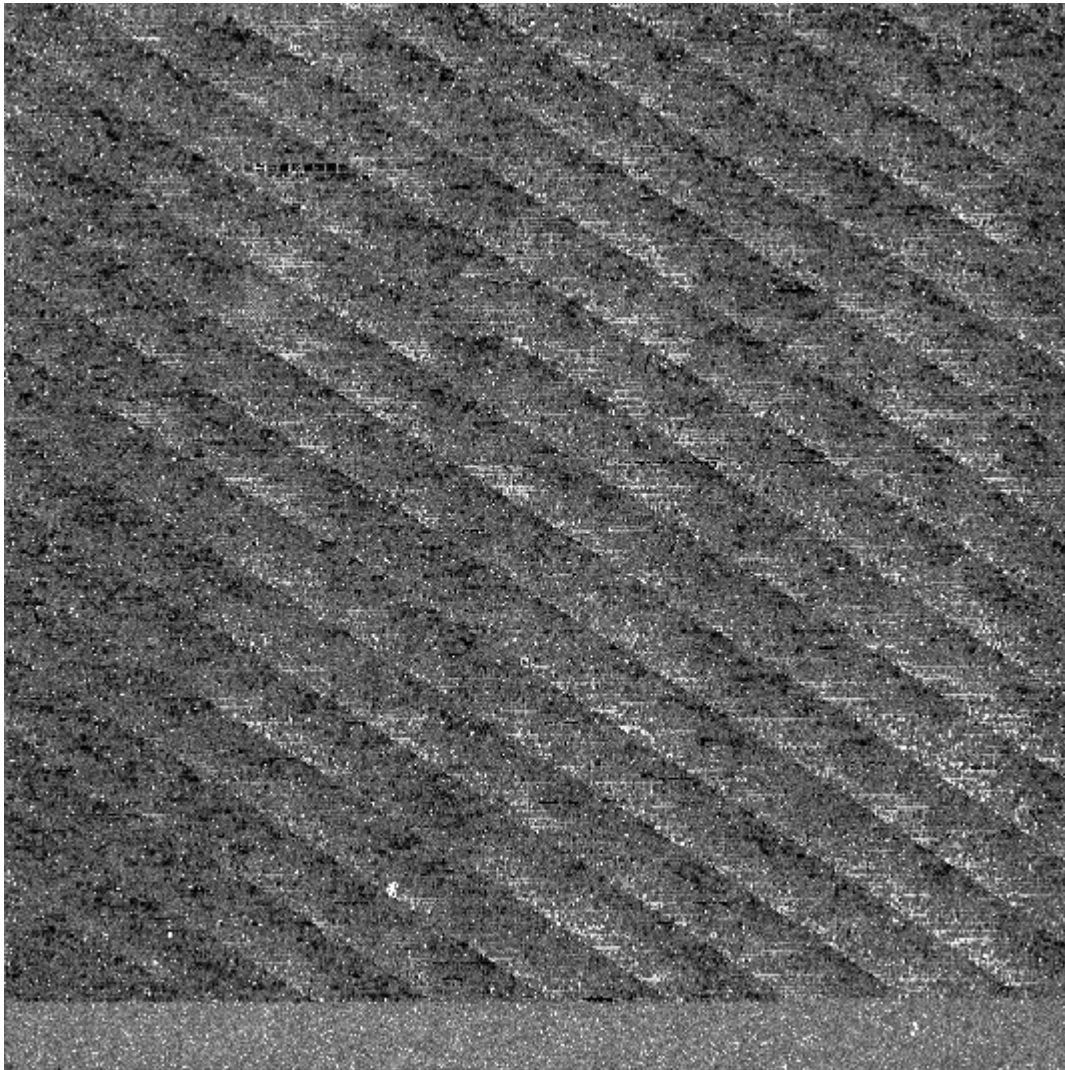


Figure 4: An image displaying the coefficient of variation in each probe cell computed from the points in Figure 3. The banding pattern provides evidence that a spatial source of variation contributes to the magnitude of variance of pixel intensities attributed to probe cells.

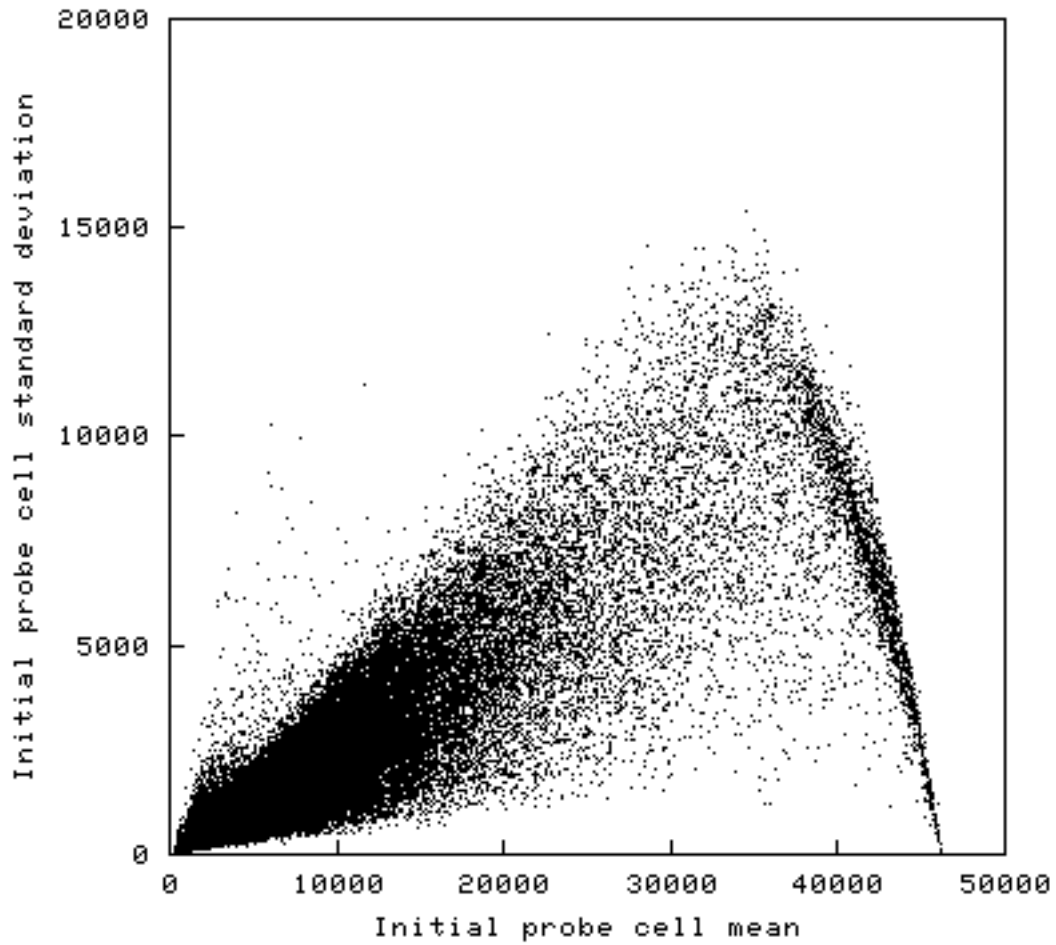


Figure 5: A plot of standard deviations of pixels allocated to probe cells versus their mean using initial interpolated probe cell locations.

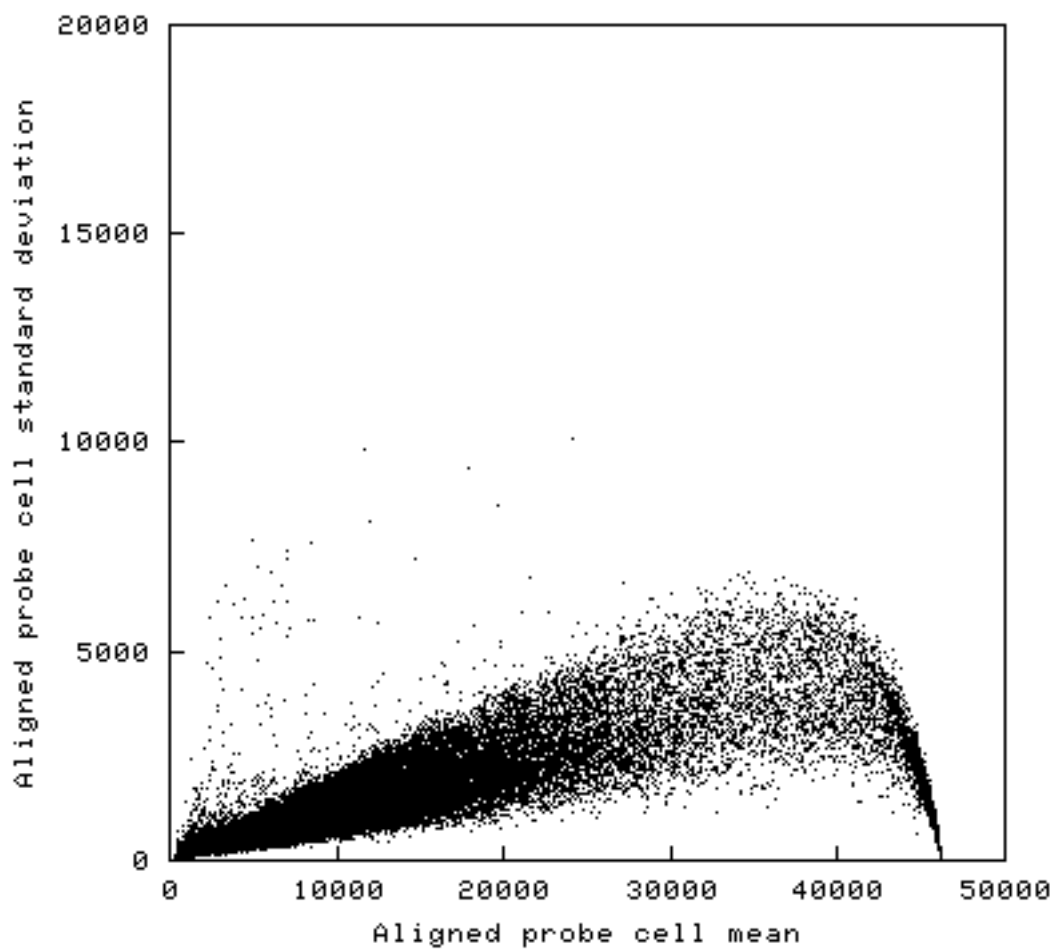


Figure 6: A plot of standard deviations of pixels allocated to probe cells versus their mean using aligned probe cell locations.

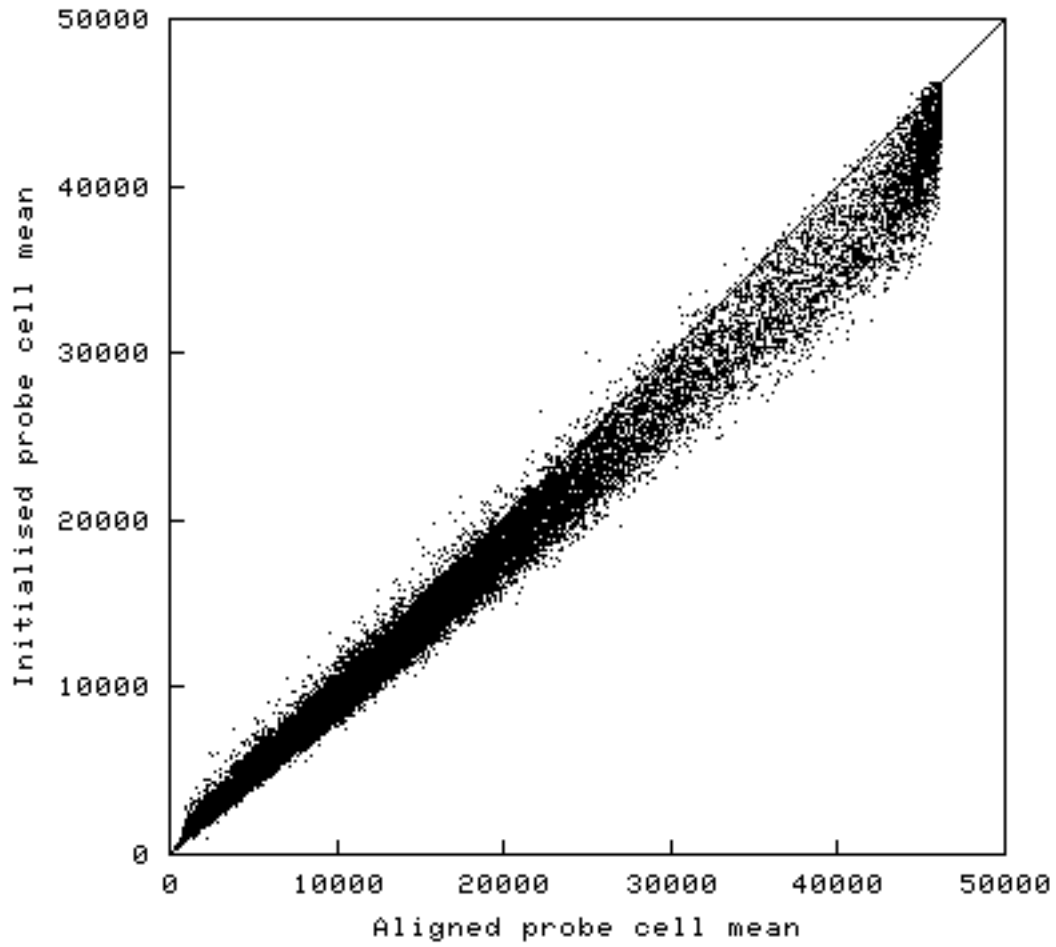


Figure 7: A plot of probe cell means from their initial interpolated locations versus means from their final probe cell locations.

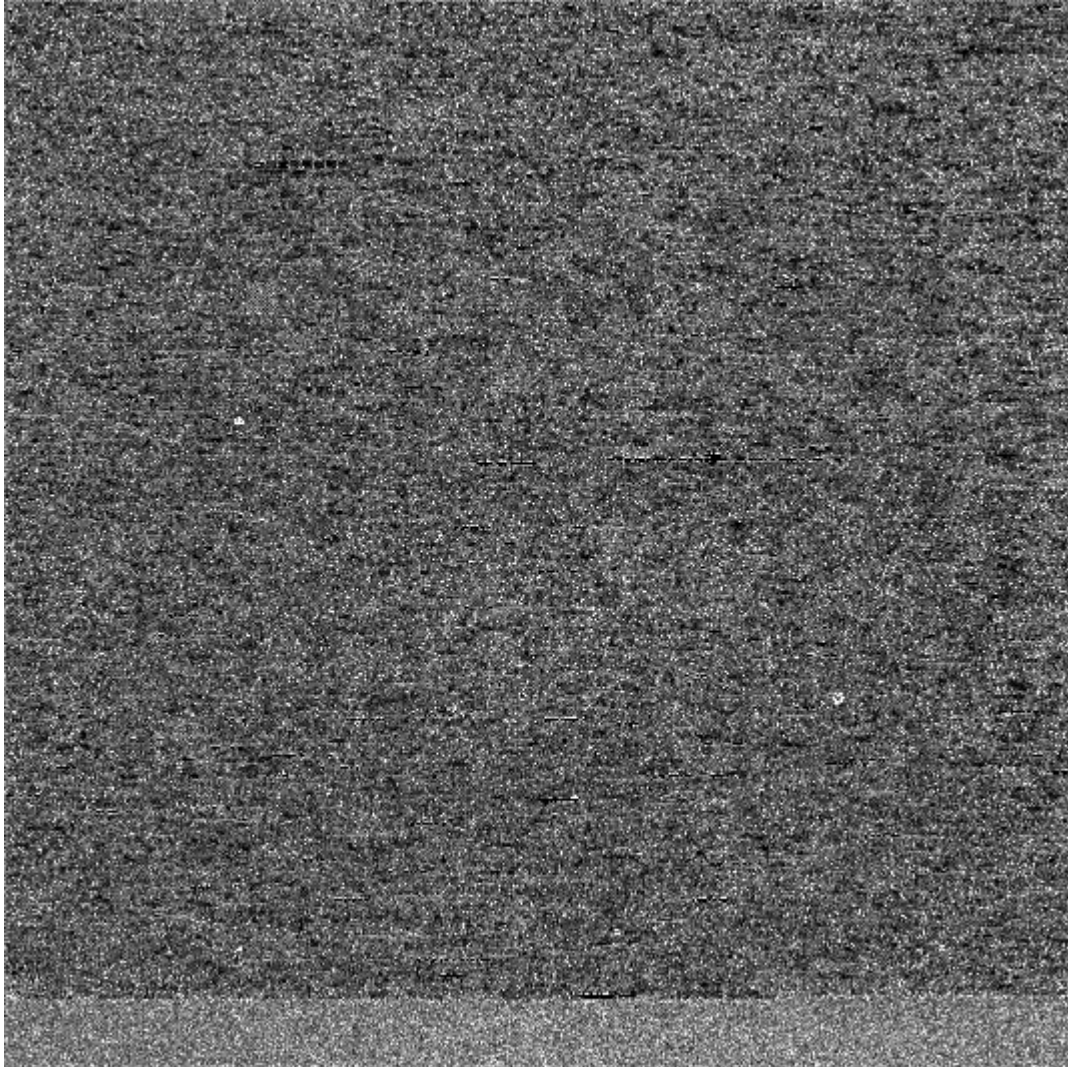


Figure 8: An image displaying the coefficient of variation for each probe cell after alignment of probe cell locations. With the exception of saturated regions and a band where little or no hybridisation took place, a spatial contribution to the variance within probe cells is no longer evident.

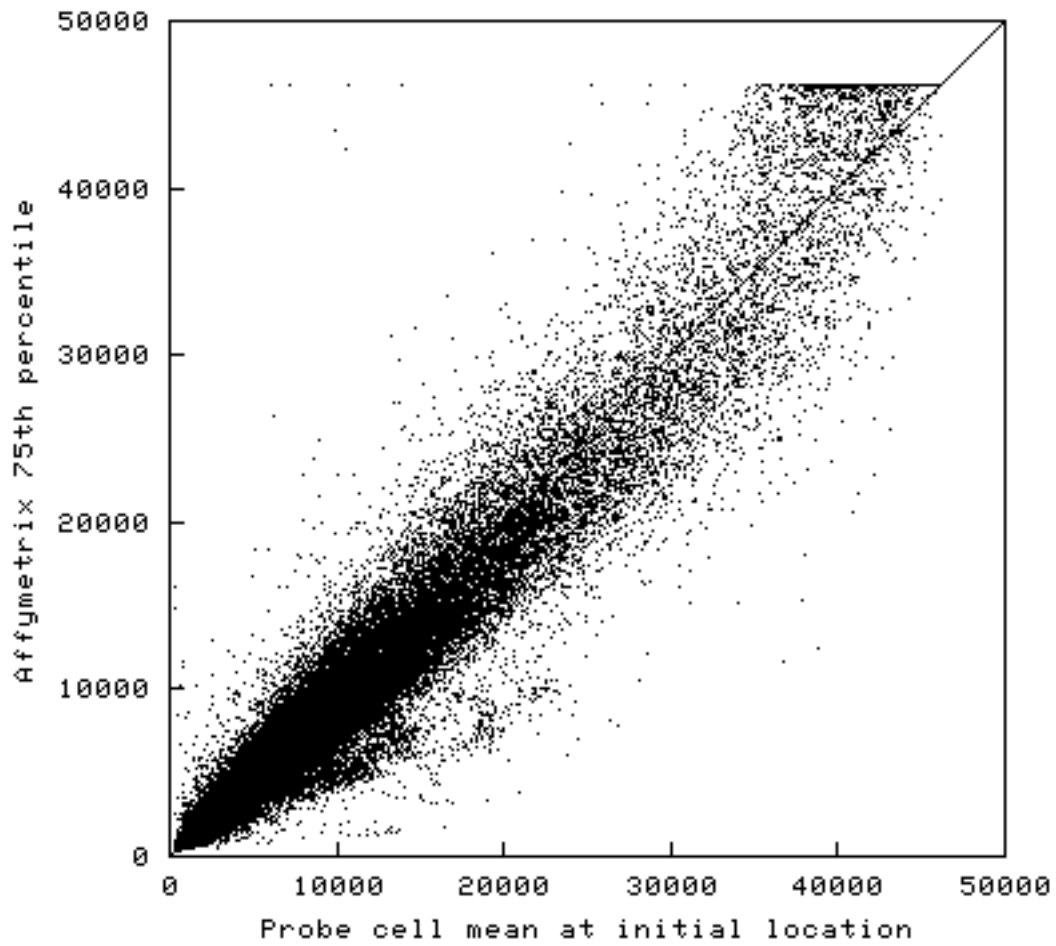


Figure 9: A plot of Affymetrix's 75th percentiles versus initial probe cell means calculated at interpolated probe cell locations prior to application of the alignment algorithm. The mean absolute pairwise difference between the 75th percentiles and the means was 539.3

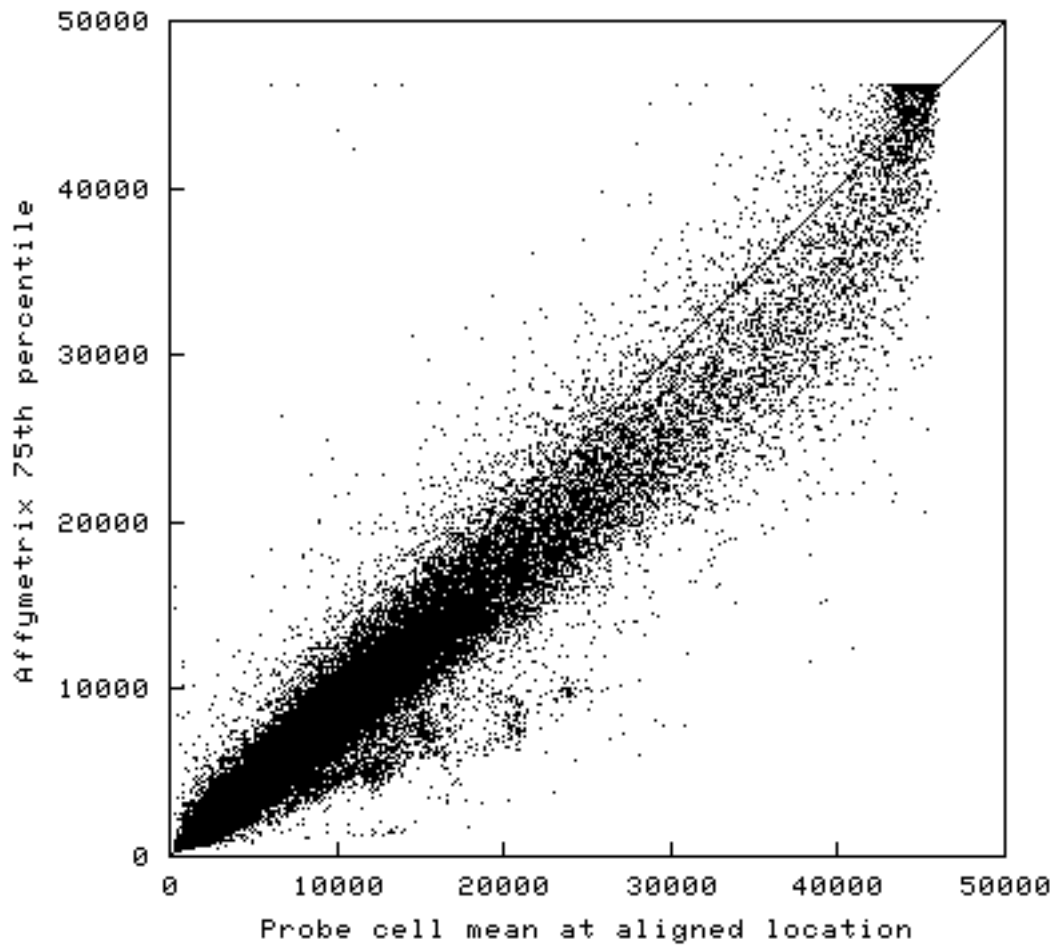


Figure 10: A plot of Affymetrix's 75th percentiles versus probe cell means calculated at aligned locations after application of the alignment algorithm. The mean absolute pairwise difference between the 75th percentiles and the means was 606.54