BRIEFINGS IN FUNCTIONAL GENOMICS AND PROTEOMICS. VOL 8. NO 3. 199–212 doi:10.1093/bfgp/elp027

On the causes of outliers in Affymetrix GeneChip data

Graham J. G. Upton, Olivia Sanchez-Graillet, Joanna Rowsell, Jose M. Arteaga-Salas, Neil S. Graham, Maria A. Stalteri, Farhat N. Memon, Sean T. May and Andrew P. Harrison

Abstract

We describe various types of outliers seen in Affymetrix GeneChip data. We have been able to utilise the data in the Gene Expression Omnibus to screen GeneChips across a range of scales, from single probes, to spatially adjacent fractions of arrays, to whole arrays, to whole experiments. In this review we describe a number of causes for why some reported intensities might be misleading on GeneChips.

Keywords: *Affymetrix Genechips; outliers; blur; degradation; scanners; probe correlations*

INTRODUCTION

Affymetrix GeneChip technology has proven to be an effective way to measure the co-expression of tens of thousands of genes. This has resulted in many thousands of publications that have detailed expression changes across many tissues, developmental stages, phenotypes and diseases, for most of the major model organisms. Much of the data is now deposited in public resources, such as the Gene Expression Omnibus [1]. A number of groups are now exploring the best way to mine that data.

Although GeneChips measure many genes simultaneously, most published studies have only utilised a relatively small number of conditions. This has led to many analysts referring to the 'curse of dimensionality', since it is not clear how best to extract statistically significant changes, which are also detailing interesting biology from comparisons of just a few GeneChips. As a result many users of GeneChips report thousands of genes to be differentially expressed when comparing two or more conditions, but then choose, for pragmatic reasons, to focus on unravelling the biology of only a few tens of genes. However, the availability of large collections of GeneChip experiments enables us to overcome the problem since we now have more conditions studied than there are genes. However, in order to use these surveys to extract biological signals, it is imperative that reliable quality control checks are developed to assess the impact of any biases in the data.

Corresponding author. Andrew P. Harrison, Department of Mathematical Sciences and Department of Biological Sciences, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, UK. Tel: +44 (0)1206 872964; Fax: +44 (0)1206 873043; E-mail: harry@essex.ac.uk

Graham J. G. Upton is Professor of Applied Statistics at the University of Essex. With more than 100 publications, he is the co-author of *The Oxford Dictionary of Statistics*.

Olivia Sanchez-Graillet is a Research Fellow in Integrative Bioinformatics at the University of Essex, focussing on developing tools for mining large surveys of GeneChip data.

Joanna Rowsell is a PhD student at the University of Essex, working on the analysis of splicing patterns within GeneChip data.

Jose Arteaga-Salas is a PhD student at the University of Essex, working on detecting spatial defects in GeneChip data.

Neil Graham is a Research Fellow in Bioinformatics at the Nottingham Arabidopsis Stock Centre, working on improving cross-species studies of the transcriptome.

Maria Stalteri is a Research Fellow in Bioinformatics at the University of Essex, working on integrating biological annotation into the analysis of GeneChip experiments.

Farhat Memon is a PhD student at the University of Essex, working on developing cloud computing as a resource for mining ultra-large surveys of GeneChip data.

Sean May is the Director of the Nottingham Arabidopsis Stock Centre, holds a Faculty post in the School of Biosciences at the University of Nottingham, and his focussed on improving the interpretation of post-genomic experiments.

Andrew P. Harrison is a Senior Lecturer in Mathematical Bioinformatics at the University of Essex. He has worked on the analysis of Affymetrix GeneChips since 2003. His original research training was in Astrophysics.

An Affymetrix GeneChip consists of a high-density array of oligonucleotides, and each transcript is detected by a particular group of 25mer probe sequences, known as a probe-set. The collation of the intensities in multiple probes, into a single measure of expression for a gene, has a significant impact upon what can be inferred from GeneChip data [2, 3]. Although many of the expression measure calculations try to correct for outliers, it is important to remove known outliers prior to the calculation of the expression measure. Otherwise, they will act to alter the detection of other potential outliers within the expression measure calculations. Known outliers might appear to tentatively support the values seen by others in their probe-set, even though this is coincidental. We therefore advise that known outliers should not be included in the calculations of expression.

Some probes with likely problems can be identified even without looking at the output from an experiment. Clearly probes that do not map to the same transcript as others within their set, or those that map to more than one transcript, will, potentially, produce discrepant results. As a result there have been a number of efforts to generate up-to-date mappings between the probe sequences and the information in genomic databases [4, 5]. Also, chains of nucleic acids may undergo folding. If either a probe sequence, or a fragment of a target complementary to a probe, shows regions of self-complementarity, then this affects hybridisation efficiency [6]. Estimates of folding rates suggest that ~10% of probes may be affected by folding [7].

In this review we focus on newly discovered biases that are invisible within a single experiment, but become apparent when the data from several experiments are compared. We have been able to utilise the data in GEO to screen GeneChips across a range of scales, from whole experiments, to whole arrays, to spatially adjacent fractions of arrays, to single probes. This information can then be utilised by analysts who have only a relatively small number of GeneChips available, in order to sharpen their results, potentially leading to novel insights into their biology of interest. We will describe a number of potential causes for why some reported intensities might be misleading on GeneChips. Unless otherwise stated, the results have been derived from a study of 2756 HG-U133_Plus_2 GeneChips downloaded from GEO in 2007.

METHODOLOGY

The results reported here are the outcomes of a continuing and ongoing examination of GeneChips. A natural next step was to question whether the probes in probe-sets are equally responsive to changes in expression level. An effective procedure was provided by the use of heatmaps [9], with heatmaps for a range of Human GeneChips being available at http://bioinformatics.essex.ac.uk/users/harry/. The study of heatmaps reveals many instances of individual probes that are not well-correlated with the remaining members of an otherwise well-correlated probe-set. Subsequent research has shown that there can be many different causes of the poor correlations: those that we have identified are described in the next sections. However, the list is not exhaustive: we have identified other groups of probes that appear to behave unusually. These are not listed here since our research has not yet identified the underlying causes.

Heatmaps for entire probe-sets may mislead, however, since alternative splicing and polyadenylation can cause probes for the same gene to show differential expression with respect to each other [8]. It is also entirely possible for one exon to be up-regulated at the same time as another exon from the same gene is down-regulated. Similarly, groupings of probes either side of a polyadenylation signal may show differential expression. Because of these possibilities we have focussed on studying groupings of probes that map to the same exon: all such probes should certainly be well correlated with one another. We study unique probes, those that only target one place on a single exon [9] according to the exon definitions from Ensembl. This means that we only utilise a fraction of the probes available on the GeneChip, but the fraction we use will have been chosen to provide reliable measurements.

CAUSES OF OUTLIERS

We subdivide the causes into effects that are directly related to the base sequences in the probes and effects that are unrelated to the base sequences. Our results are based on the CEL files that we have downloaded from GEO. These CEL files are identified by 'GSM' numbers, with each separate experiment being identified by a 'GSE' number.

Outlier probes related to probe sequences

Runs of guanines

Probes containing runs of guanine show abnormal binding affinities, with respect to models of hybridisation, and are typically outliers with respect to the rest of the probe-set to which they are assigned [10]. We have recently confirmed that probes containing G-spots, i.e. runs of four, or more, contiguous guanines, usually show uncorrelated behaviour with the rest of the probe-set to which they are assigned. But we have gone further in discovering that probes containing G-spots report intensities that are correlated with the other probes containing runs of guanine [11, 12] (Figure 1).

We believe that G-quadruplexes are forming on GeneChips and that their formation acts to confuse the interpretation of probes containing runs of guanine. Neighbouring probes can readily come into physical contact on Affymetrix GeneChips and each of the probes consists of the same sequence. Such identical sequences are not able to Watson–Crick hydrogen bond to each other. But, rather than Watson–Crick hydrogen bonding to their associated transcripts, probes containing runs of guanine will form Hoogsteen hydrogen bonds with adjacent neighbours (which also contain runs of guanine). Under this scenario, a group of four probes will form a tetrad of interactions, and a stack of tetrads resulting from the run of guanines will stabilise the grouping. These probes will all have their guanines pointing into the quadruplex, and so will not be available to hybridise to target. But, if groups of four probes form quadruplexes, this results in extra space outside the quadruplexes enabling remaining unbonded probes to hybridise strongly to targets. Further details on the biophysics of GeneChips, particularly related to the impact of G-quadruplexes, can be found in [12].

Motifs related to chip preparation

A family of sequence motifs, centred on GCCTCCC and related to the preparation of target, confuses the interpretation of data from GeneChip experiments [13]. Amplification of mRNA occurs during the preparation of the RNA target with the most popular amplification method being the Eberwine technique [14]. This method is based on the incorporation of a T7-binding site with an oligo-dT primer in the first strand of cDNA. The variants of the T7 primer have the core T7-binding domain in common. The T7-binding domain is essential for the first interaction with the RNA polymerase and the start of transcription. The bases flanking the core T7-site are called spacer sequences [13]. The 5′ spacer is introduced to avoid having the T7 binding site on the end of the primer. The 3′ spacer is the reflection of the viral T7 sequence at the relevant position in the polymerase binding site [15]. However, a more common belief is that the 3′ spacer is just a separator between the T7 site and the oligo-d(T) stretch [13]. The final step of the amplification procedure is the incorporation of

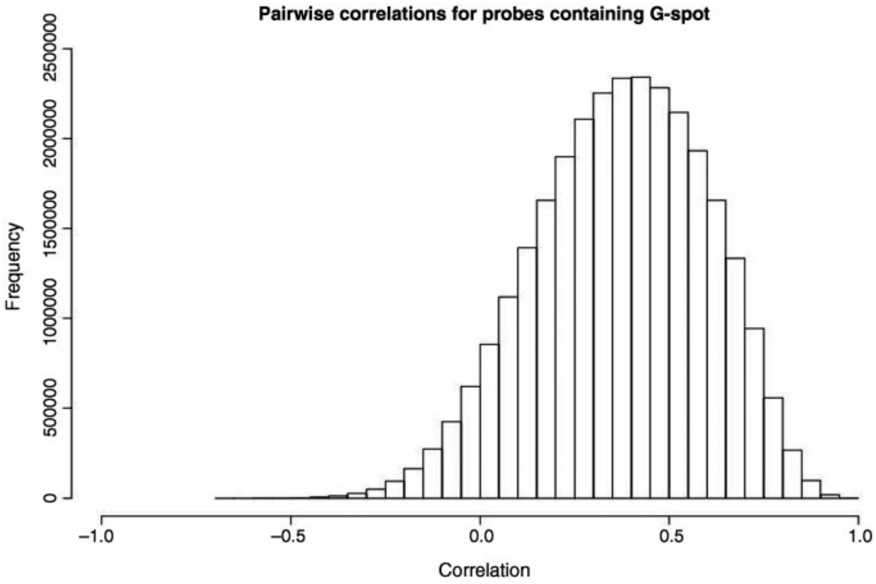
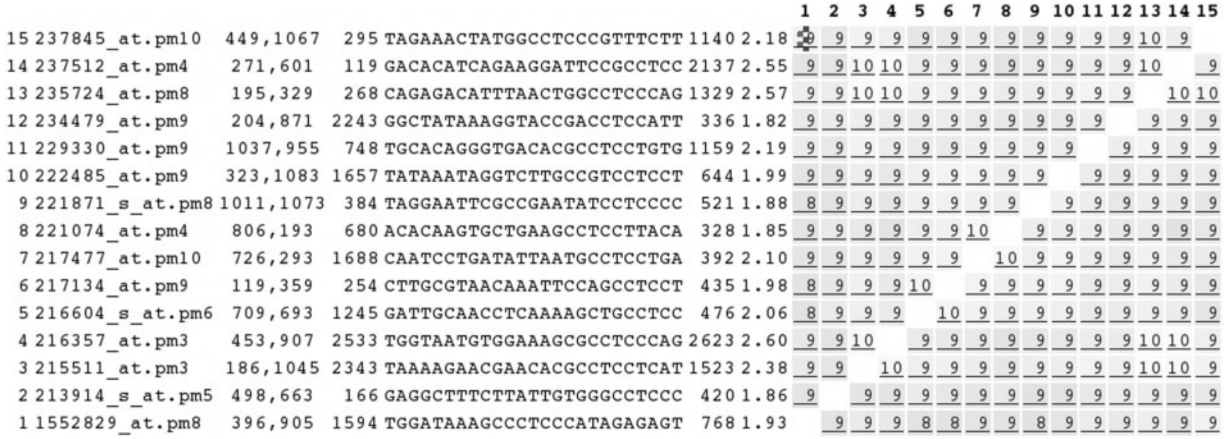


Figure 1: Distributions of pair-wise correlations (based on 2756 downloaded HG-U133.Plus.2 CEL files) between each of the probes that contain G-spots and map uniquely to single exons.



								1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
15	237845_at.pm10	449,1067	295	TAGAAACTATGGCCTCCCGTTTCTT	1140	2.18	9	9	9	9	9	9	9	9	9	9	9	9	10	9	
14	237512_at.pm4	271,601	119	GACACATCAGAAGGATTCCGCCTCC	2137	2.55	9	9	10	10	9	9	9	9	9	9	9	9	10		9
13	235724_at.pm8	195,329	268	CAGAGACATTTAACTGGCCTCCCAG	1329	2.57	9	9	10	10	9	9	9	9	9	9	9	9		10	10
12	234479_at.pm9	204,871	2243	GGCTATAAAGGTACCGACCTCCATT	336	1.82	9	9	9	9	9	9	9	9	9	9	9		9	9	9
11	229330_at.pm9	1037,955	748	TGCACAGGGTGACACGCCTCCTGTG	1159	2.19	9	9	9	9	9	9	9	9	9	9		9	9	9	9
10	222485_at.pm9	323,1083	1657	TATAAATAGGTCTTGCCGTCCTCCT	644	1.99	9	9	9	9	9	9	9	9	9		9	9	9	9	9
9	221871_s_at.pm8	1011,1073	384	TAGGAATTCGCCGAATATCCTCCCC	521	1.88	8	9	9	9	9	9	9	9		9	9	9	9	9	9
8	221074_at.pm4	806,193	680	ACACAAGTGCTGAAGCCTCCTTACA	328	1.85	9	9	9	9	9	9	10		9	9	9	9	9	9	9
7	217477_at.pm10	726,293	1688	CAATCCTGATATTAATGCCTCCTGA	392	2.10	9	9	9	9	9	9		10	9	9	9	9	9	9	9
6	217134_at.pm9	119,359	254	CTTGCGTAACAAATTCCAGCCTCCT	435	1.98	8	9	9	9	10		9	9	9	9	9	9	9	9	9
5	216604_s_at.pm6	709,693	1245	GATTGCAACCTCAAAAGCTGCCTCC	476	2.06	8	9	9	9		10	9	9	9	9	9	9	9	9	9
4	216357_at.pm3	453,907	2533	TGGTAATGTGGAAAGCGCCTCCCAG	2623	2.60	9	9	10		9	9	9	9	9	9	9	9	10	10	9
3	215511_at.pm3	186,1045	2343	TAAAAGAACGAACACGCCTCCTCAT	1523	2.38	9	9		10	9	9	9	9	9	9	9	9	10	10	9
2	213914_s_at.pm5	498,663	166	GAGGCTTTCTTATTGTGGGCCTCCC	420	1.86	9		9	9	9	9	9	9	9	9	9	9	9	9	9
1	1552829_at.pm8	396,905	1594	TGGATAAAGCCCTCCCATAGAGAGT	768	1.93		9	9	9	8	8	9	9	8	9	9	9	9	9	9

Figure 2: A collection of probes from different probe-sets which map to antisense transcripts and which are highly correlated with each other across the 2756 HG-U133.Plus.2 CEL files downloaded from GEO. The columns indicate the probe order in the heatmap, probe identifier (pm means perfect match followed by a number that indicates the position of the probe in the probeset), *x* coordinate in the CEL file, *y* coordinate in the CEL file, interrogation position on the consensus/exemplar sequence defined by Affymetrix, sequence, geometric mean of the intensities across GEO, and standard deviation (of the logs of intensities respectively). The numbers in each of the cells represent the rounded correlation × 10.

the 3′ spacer sequence and its transcription by the T7 RNA polymerase. In this way, the 3′ spacer sequence becomes the leader sequence for all the copies of the amplified RNA. The 3′ spacer sequence of the Affymetrix primer is transcribed as the 9-mer leader sequence 5′GGGAGGCGG3′ and its complementary motif is 5′CCGCCTCCC3′. Kerkhoven *et al*. [13] noted that the T7 spacer sequences caused hybridisation artefacts on those probes containing complementary sequences of the spacer sequence (T7 3′ spacer motifs). Because there are no rules for eliminating these complementary sequences in the probe design of Affymetrix, probes containing fractions of these complementary sequences, such as GCCTCCC or CCTCC, may hybridise to the leader sequence of every amplified RNA molecule.

We find that the motif identified by Kerkhoven *et al*. [13] has some impact on the correlations observed from GeneChips. An example is shown in Figure 2 using data from 2756 CEL files. All of the probes here belong to different probe-sets and map to different genes, and have low or negative correlations with the rest of the probes in their own probe-sets. But all of these probes have the subsequence ‘CCTCC’ in common and show high correlation with each other. It is our belief that these correlations are not resulting from biology, but instead from cross-hybridisation of the probes to the primer-spacer appended to transcripts, as suggested by Kerkhoven *et al*. [13]. Pairs of probes which both contain CCTCCC show correlations typically between 0.5 and 0.8, whereas probes containing GCCTCCC show higher correlations ≥0.8 (Figure 3).

Other motifs

There is a wide variation in general expression levels that is unlikely to be entirely a reflection of the variations in the transcriptome being studied. Given that the effects of the G-quadruplexes (G-spots) and the primer spacers (CCTCC) may be related to chip preparation, it is natural to look for further examples. As a step in that direction, we have identified those probes that contain specified motifs. We have accomplished this for all motifs of lengths three to seven. For the longest of these motifs there are $4^7 = 16\,384$ possible sequences ranging from AAAAAAA to TTTTTTT. The most frequently occurring on the HG-U133_Plus_2 is the sequence TTTCTCT that appears in 5265 probes. However, the sequence CCCCCCG appears only in 79 probes, and for that reason we have not studied longer motifs. For each motif we have calculated its average value for each CEL file (multiplicatively scaled to have the same mean): thus for TTTCTCT, we have averaged the expression levels for the 5265 probes in which that motif occurs.

With this large number of probes, any variation from one CEL file to another is unlikely to be due to biological variation. Despite the overall scaling of the

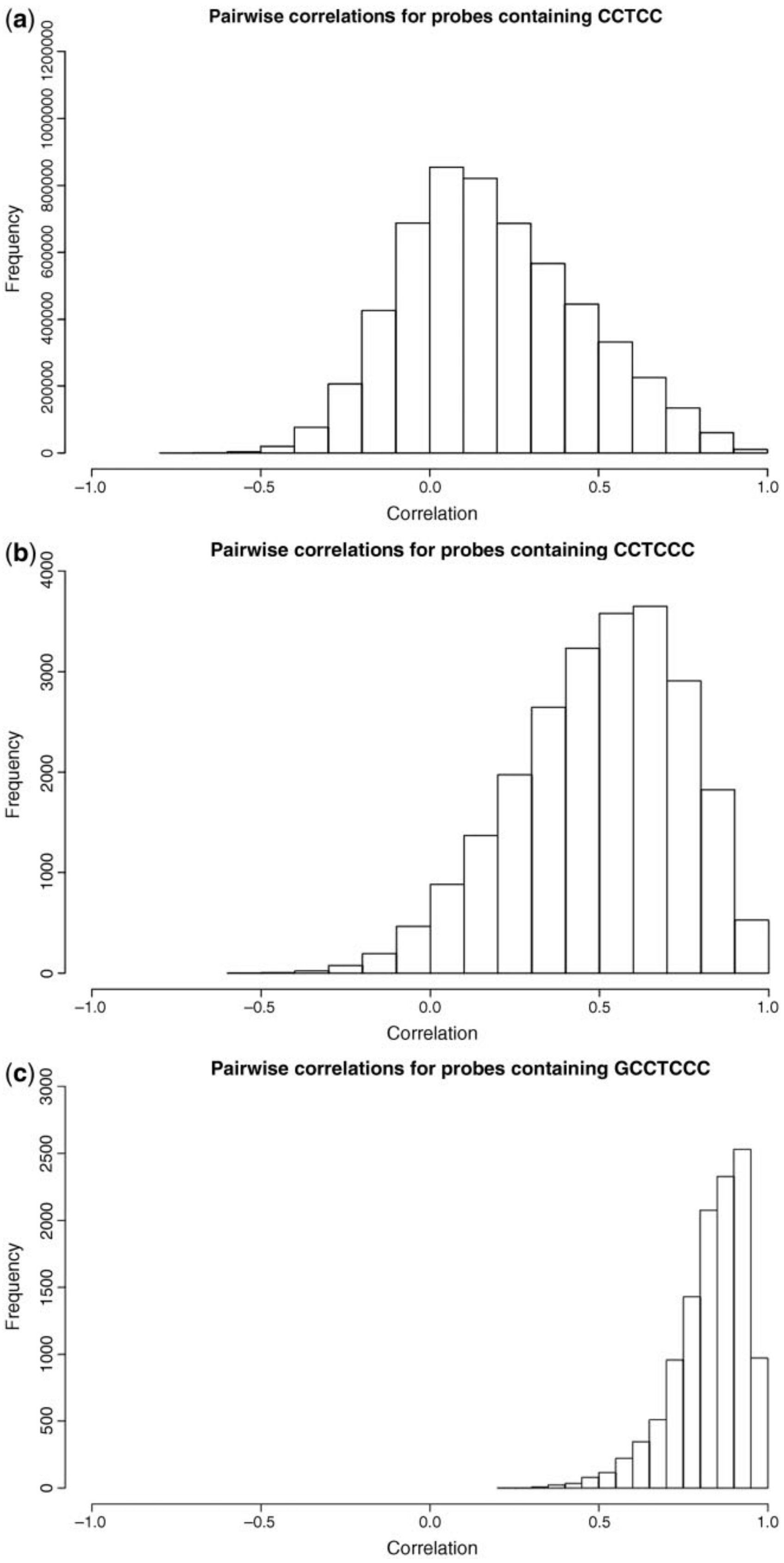


Figure 3: Distributions of pair-wise correlations between each of the probes that map uniquely to single exons and that: (**a**) contain the motif CCTCC; (**b**) contain the motif CCTCCC; (**c**) contain the motif GCCTCCC.

CEL files there will, of course, be variability in the averages for a given motif. We now calculate '*z*-scores' for each motif average (by comparing each CEL average with the mean for the remaining 2755 CEL averages and dividing by their standard deviation). There will be many *z*-scores that exceed the usual values judged significant (we are calculating ~45 million such values) so we concentrated on those with an absolute value >6 (corresponding to a chance probability of ~1 in 5 million): we found >24 000 such values. These were not randomly spread amongst the CEL files, nor amongst the motifs.

The GSE2125 group gave many unusually high *z*-values for motifs including CTTC and TTCTC and unusually low values for motifs including the complementary AAGAG sequence. However, possibly the clearest evidence of a 'motif bias' links the members of GSE3678 and GSE4219 with the TTTTTTG and TTTTTTT motifs. For example, GSE4219 comprises 24 CEL files (12 in each of the subseries GSE4217 and GSE4218) of which 20 had *z*-values exceeding 6 for the TTTTTTG motif and 22 for the TTTTTTT motif. The remaining 10 CEL files having high *z*-values for the TTTTTTT motif are all members of the GSE3678 set, which comprises 14 CEL files.

Outlier probes associated with where and when the chip was run

The effect of the scanner

In the case of both G-spots and the CCTCC motif, the probes affected were poorly correlated with others in their probe-set, but were well correlated with others in their group. There are many other probes that also display markedly low correlations with the members of an otherwise highly intercorrelated probe-set [9]. A first step is to ask whether these probes are strongly correlated with any others on the array.

For one subgroup of probes, the strong correlations were with the lower-magnitude probes on the edge of the array which are probes having no biological significance. These probes are intended to provide a reference pattern of roughly alternating bright and very bright signals that enable the subsequent analysis of the scanned GeneChip to correctly allocate signals to probes. At first sight, therefore, the presence of high correlations between biologically relevant probes and these control probes is surprising.

Table 1: The correlation between the values of a probe and the values at (0, 0), based on a random sample of 300 CEL files

Size of nearest neighbour	50–	150–	400–	1000–	3000–	8000–
Median correlation with (0, 0)	0.07	0.13	0.25	0.50	0.70	0.85

The probes are subdivided by the size of their largest neighbour as reported in the CEL file. The groups are non-overlapping.

The answer lies in the magnitudes of neighbouring probes: all the lower-magnitude edge probes are adjacent to very bright edge probes, while all the affected non-control probes also lie next to high-magnitude probes.

Table 1 demonstrates how the size of a neighbour matters. For every perfect-match and mismatch probe we calculated the correlation (over the 2756 CEL files) of the values of that probe and the values of an arbitrarily chosen control probe—we chose the probe at (0, 0). We assigned each perfect match or mismatch probe to a class determined by the magnitude of the probe's nearest neighbour. The table reports the median correlations for each class, with the results very clearly showing that the magnitudes of these unwanted correlations are indeed determined by the magnitude of the neighbouring large probe.

Each probe value in a CEL file is derived from a set of (typically 25) individual measurements made by a standard Affymetrix scanner [16]. The standard algorithm takes the 75th percentile of these values to calculate the intensity recorded in the CEL file. The CEL file also provides the standard deviation and number of measurements as part of its output. Moreover, the identity number of the scanner and the date and time of the scan are given in the CEL file's header. If a scanner is not well focussed then what should be a sharp image of a bright spot will be blurred across its neighbours (in the same way that stars appear to twinkle). For a poorly focussed scanner the effect on neighbours will be most noticeable in the case of the brightest probes, as is suggested by Table 1. The potential outcome is illustrated in Figure 4 which is a scatter diagram of the values of two overlapping members of the 233481_at probe-set.

In Figure 4, the line of equality suggests that probe 5 usually has values that are only slightly larger than those for probe 4, and the expected linear relation between the values of the two

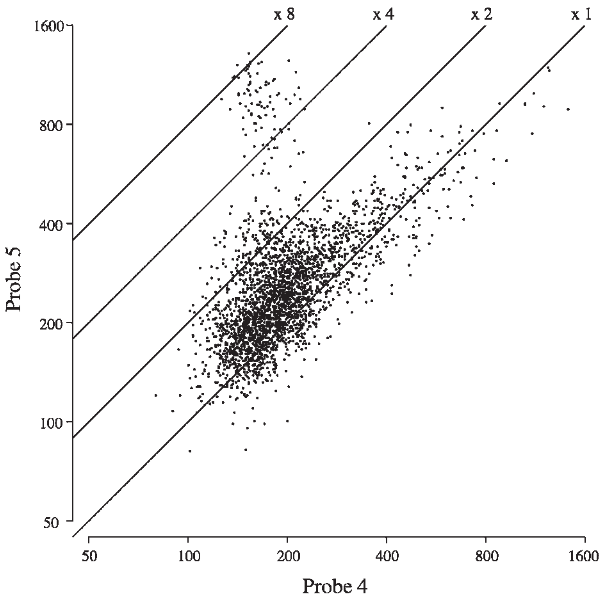


Figure 4: A scatter diagram showing the intensities for two members of the probe-set 233481_at from the 2756 HG-U133_Plus_2 CEL files. The two probes have a 19-base sequence in common, so that a high correlation would be anticipated. However, probe 5 lies adjacent to a high-value probe (mean > 20 000), whereas probe 4 does not (highest neighbouring mean 70).

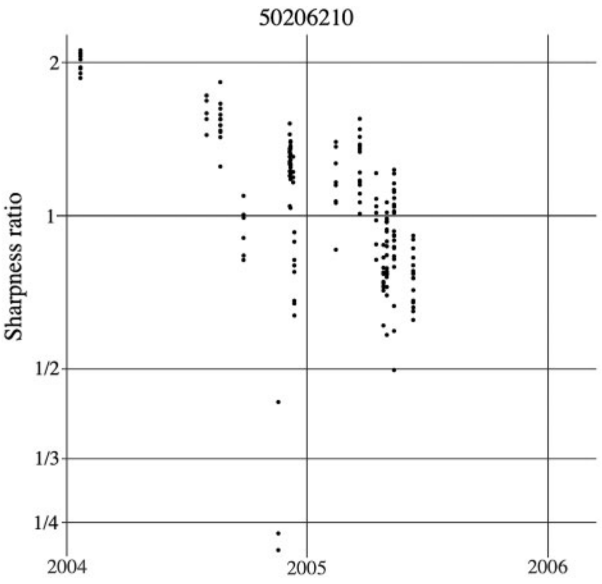


Figure 5: An example of the apparent change in behaviour of a scanner with time. Each point represents one of the 310 downloaded HG-U133_Plus_2 CEL files that were scanned by scanner 50206210. In early 2004 the scanner produced very sharp images, but the quality steadily deteriorated through that year. An improvement late in the year carried through to 2005, but by mid 2005, the images were, at best, only of average quality.

probes is visible to an extent. However, there is an obvious cloud of points at the top left corresponding to cases where the value of probe 5 is as much as eight times the value of probe 4. There are also many other cases in which the value for probe 5 is more than double that of probe 4, suggesting that a smaller amount of blur is inflating these values. The CEL files in the top left of Figure 4 are not a random selection: they represent a series of files all scanned by the same scanner during a period of about four weeks when 'there were numerous breakdowns, and we had an Affy field engineer working on both the scanners and the chip washers nearly every week' (Michael Bittner, personal communication).

Time

Scanners are used to scan many types of array, and we have not attempted to follow the usage of a scanner through these different types across time. For the HG-U133_Plus_2 CEL files, however, we can report evidence that (i) some scanners appear to give more blurred images than others, and (ii) in general scanners appear to become less accurate with time. Plotting a measure of blur (the 'Sharpness ratio' statistic used in Figure 5 compares the value of S for a given CEL file against the average value of S, where S is the sum of the squared differences between a probe value and the average of its four immediate neighbours, summed over all probes in the file) against time shows discontinuities suggesting that scanner performance can be capable of considerable improvement (presumably by servicing or renewal of a component). The results for all 310 downloaded files for scanner 50206210 are shown.

Outliers concentrated in one region of the chip

A large number of microarray experiments available in GEO contain replicate data in their hybridisations. Replication, although often regarded as expensive and time consuming, is an important tool to measure random noise in microarray experiments. When biological replicate arrays are produced, at each location of the array we would expect to observe little variation between the intensity values. Naturally, this variation will be higher in some locations than in others. However, if locations with unusually high variation are concentrated in compact regions of

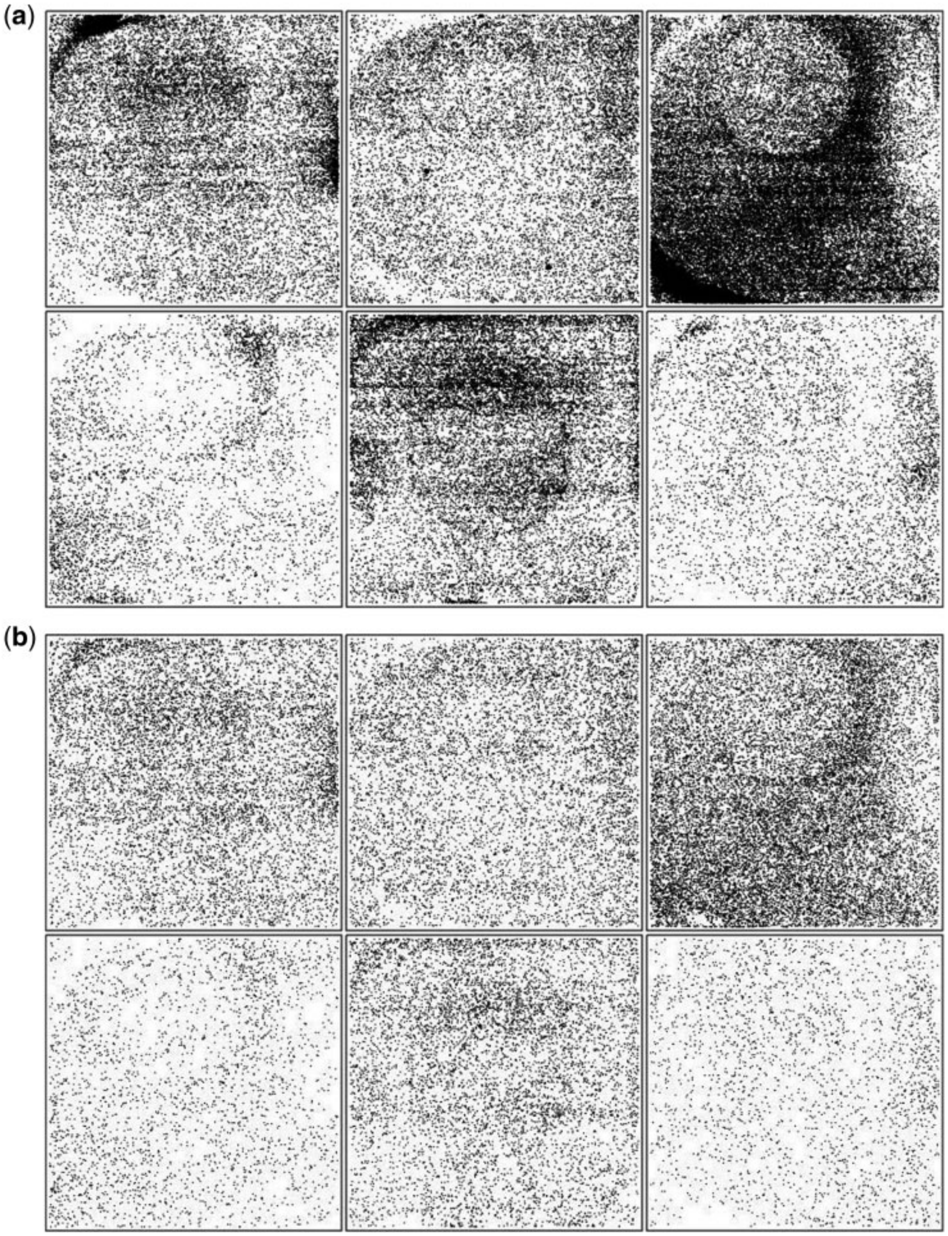


Figure 6: (**a**) Spatial defects in three replicates of the HG-U133.Plus.2 array; (**b**) Remaining flaws in the same three replicates after applying the LPE and CPP adjustments in sequence. The upper rows show the locations of unusually large values and the lower rows the locations of unusually low values.

the array, then a spatial defect might be contaminating the data.

Spatial defects in Affymetrix GeneChip data are typically manifested as blobs, scratches, rings or arcs concentrated, particularly towards the edges of the array [16–19]. We have developed methods to visualise spatial defects by comparing each location in all the available replicates with a reference set of values [19]. By comparison with a reference set, it is possible to identify locations with unusually low or unusually large values in all the replicates. If these locations are concentrated in compact regions of the array, then spatial defects are visible. Figure 6a shows an example of circular spatial defects visible in three replicates of the HG-U133_Plus_2 array, obtained from GEO.

Minimising spatial defects is an important, though often neglected, step in microarray data analysis. Such defects are likely to affect the intensity values in a subset of the probes within a probe-set, and thus impact upon the gene expression measurements assigned to that probe-set [16]. Thus, it is important to utilise appropriate tools to minimise the effects of these defects. We have developed two methods to assist with flaw reduction: the local probe effect adjustment (LPE) and the complementary probe pair adjustment (CPP) [19]. The methods can be utilised independently or sequentially. Both methods use the spatial compactness of the defects to minimise the magnitude of the defects and hence the impact of the defects on the affected probe-sets. Standard normalisation procedures, such as quantile normalisation, make no such adjustments. Figure 6b shows the spatial defects that remain after normalising the data sequentially first with LPE and then with CPP. Although some circular patterns are still visible, it is clear that the methods are effective in spatial defect reduction.

For experiments with replicate arrays, the set of reference values to visualise spatial defects is typically obtained with the median of all replicates [19]. However, in the absence of replicate arrays an alternative reference set is required. A reliable set of comparison values can be obtained by calculating the geometric mean of the intensity values (on a probe by probe basis) for arrays available in GEO [17]. For robustness it is recommended that the upper and lower 5% of the observed values are discarded so as to avoid the potential effects of outliers.

Outlier arrays

Identifying outlier arrays

With a few exceptions such as the large-scale cancer studies, experimenters typically analyse just a few arrays. They may compare values across arrays, but are unlikely to compare whole arrays and will certainly be unaware if all their arrays (and hence all their results) are atypical of those obtained by the wider community. Figure 7 is a scatter diagram of the median value plotted against the ratio of the quartiles for some 2756 arrays of the same type (HG-U133_Plus_2). The medians and quartiles have been determined for each array from the complete set of 1 354 896 probe values (therefore including perfect matches, mismatches and controls).

Figure 7 shows that the typical HG-U133_Plus_2 CEL file has a median expression level of about 100, with the value of the upper quartile being about double that of the lower quartile. Some extreme results are indicated. One CEL file (GSM46839 at bottom left) shows barely any expression and is evidently uninformative. In contrast, another (GSM53147 at extreme right) has expression levels of typical variability, but at around 20 times the standard level. Both these data sets, and others with unusually high or low medians, or with low quartile ratios, should probably not be used for interpretation of genetic effects.

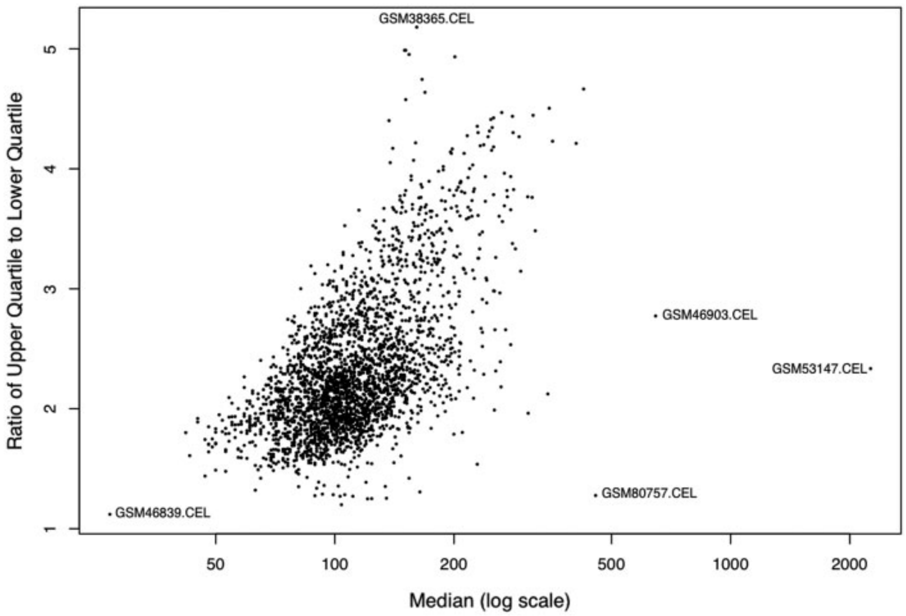


Figure 7: A scatter diagram showing the variations in general expression level and within-file variability in a sample of 2756 HG-U133_Plus_2 CEL files. For each data set the ratio of the quartiles is plotted against (on a log scale) the median expression value.

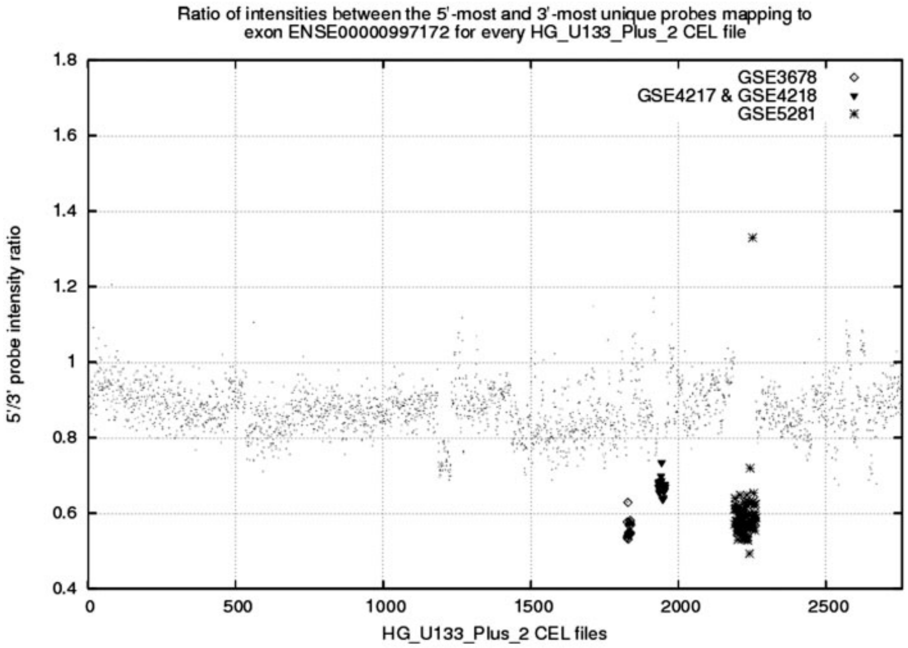


Figure 8: Ratio of the expression levels of the probes at either end of exon ENSE00000997172. The 2756 CEL files are in lexicographic order. Unusual GSE groups are indicated.

Some CEL files have unremarkable median values, but show extreme variation between their high and low values. This suggests that these are arrays without blur and are therefore observing values with unusual clarity. One such is GSM38365 (at the top of Figure 7). It is noteworthy that the other points in this region of Figure 7 correspond to other CEL files collected as part of the same experiment (GSE2125).

Degradation

We have calculated the intensity ratio of the 5′-most probe to the 3′-most probe for 10 378 exons in every HG-U133_Plus_2 CEL file. Each probe pair measures the expression of a single exon and they have constant binding affinities so that the intensity ratio provides a consistent measure of RNA fall-off. Figure 8 shows the results, for a single exon probe pair, for the 2756 CEL files. Experiments GSE3678 and GSE4219 (GSE4217 and GSE4218) again stand out in some exons as does experiment GSE5281. Similar results are obtained for each exon probe pair.

PREVALENCE AND IMPACT

Runs of guanines

The presence of G-spot probes was first identified by analysing human arrays, but has subsequently been identified on the arrays for a wide range of species including mouse, plants and yeast. The effect of removing the G-spot probes on the analysis of data will depend on the GeneChip used in the experiment. This is because the number of probes that contain G-spots differs widely for different GeneChips (Table 2).

Table 2: Numbers of G-spot probes in selected GeneChip arrays

Array	**HG-U133. Plus.2**	**Arabidopsis ATH1**	**Drosophila 2**	**Yeast genome 2.0**
Number of G-spot probes	32 547	6680	64	36

The popular Arabidopsis ATH1 and Human HG-U133_Plus_2 arrays contain tens of thousands of probe-sets include G-spot probes. The standard algorithms such as GCRMA are designed to cope with outliers and may well be able to cope with a single G-spot probe (assuming that the probe-set is not affected by other types of outliers), but will have difficulty when there are two or more G-spot probes, since they will have correlated values. Table 3 indicates the size of the problem for these arrays: about 5% of ATH1 Probe-sets and >15% of HG-U133_Plus_2 probe-sets include at least two G-spot probes.

Table 3: Numbers of G-spot probes in probe-sets, and numbers of probe-sets, for two popular GeneChip arrays

	Number of G-spot probes				**Number of probe-sets**	
	1	**2**	**3–5**	**6–11**	**Total**	**With G-spot probe**
Arabidopsis ATH1	4031	914	224	11	22810	5180
HG-U133.Plus.2	10824	5251	2920	174	54660	19169

The effects can be demonstrated by analysing the Arabidopsis AtGenExpress developmental series data set [20] with and without G-spot probes included. The number of probe-sets identified as differentially expressed (ANOVA test, $P < 0.01$, Benjamini and Hochberg false discovery rate multiple testing correction) across tissues, changes from 22 410 with G-spot included to 22 393 without G-spots, with 22 298 probe-sets common to both and 95 specific to the analysis with G-spots excluded.Similar results are obtained when the same analysis is performed to identify genes differentially expressed by time in the same data set. The number changes from 18 620 including G-spots to 18 636 excluding G-spots, with 18 355 common to both analysis and 281 unique to G-spot excluded analysis.

The removal of G-spot probes also has an effect on the correlation of probe-sets across data sets. The two probe-sets containing eleven G-spot probes (257331_at and 261946_at) are very highly correlated (correlation score = 0.917) across the developmental series data. Probe-sets with fewer G-spots are also highly correlated, for example the probe-sets 265369_s_at and 263560_s_at (both containing seven G-spot probes) have a correlation score of 0.936 across the leaf samples of the developmental series. Removal of the G-spot probes reduces this correlation to 0.247, indicating that these probe-sets are probably not correlated and the high correlation score is due to the G-spot probes. Similar results are found with probe-sets containing fewer G-spots.

These results demonstrate that the inclusion of the G-spot probes could produce misleading results when analysing data to identify differentially expressed genes or producing gene interaction networks from correlation scores.

The study of the effects of single nucleotide polymorphisms (SNPs) on hybridisation patterns seen in GeneChips has been undertaken by several groups. The study by Kumari *et al.* [21] noted that analysts of GeneChips should consider the location of SNPs in order to not miss valuable information, whilst Alberts *et al.* [22] observed that genetic variation affects hybridisation patterns. Our own analysis indicates that there is not a strong relationship between the location of a SNP within a probe and the corresponding impact on the probe's behaviour with respect to the other probes within a probe-set. A cross-tabulation of probes (SNP: yes/no; outlier: yes/no) shows no evidence that SNPs lead to outlier probes. Moreover, we find that almost half of the probes which contain SNPs and are outliers also contain runs of guanine (G-spots) or CCTCC (the primer spacer). The existence of multiple sources of outliers means that care needs to be taken when interpreting why a particular probe appears to be slightly different to its peers from the same probe-set in any single experiment.

We also find that groups of probes which map uniquely to exons in the antisense direction are frequently not correlated with each other. But the intensities from such antisense probes, which also contain either G-spots or CCTCC, may show enhanced correlations with their respective families of outliers, e.g. Figure 2. The existence of apparent expression from a subset of probes may act to confuse the interpretation of searches for antisense expression on Affymetrix GeneChips.

Motifs related to chip preparation

We have studied the relative impact of the G-spot and the primer spacer upon the correlations seen in GeneChip surveys. We have checked whether the high correlations between pairs of antisense probes were provoked either by the occurrences of G-spots or by the occurrences of CCTCC. Langdon *et al.* [12] showed that modifying CDFs, by removing probes containing G-spots, led to improvements in the ability of RMA to identify true positives in the spike-in experiments, using the Affycomp tool of [2]. We have repeated this experiment, by removing probes containing CCTCC, but find that in this case there is little change to the results. In conjunction with the correlation analysis, this indicates that the G-spot effect is stronger than the CCTCC motif effect in causing outlier probes. This is not surprising since the 5-base CCTCC motif will be naturally less prevalent than the 4-base GGGG motif.

Table 4 gives information concerning the prevalence of these probes.

Although there are fewer CCTCC probes, they can nevertheless be devastating in their effect on the interpretation of gene prevalence, as Figure 9 indicates. This heatmap shows the correlations between the 11 probes in probe-set 396_f_at, designed for erythropoietin (EPOR) and the 16 probes in probe-set 47571_at, designed for the zinc finger gene 236 (ZNF236). All 11 probes in the first probe-set contain CCTCC as do 9 of the probes in the second probe-set. A glance at the figure quickly reveals which probes these are. It seems very unlikely that either probe-set is fit for purpose.

Table 4: Prevalence of the CCTCC motif for two popular GeneChip arrays

	Number of probes containing the CCTCC motif			
Number of G-spot probes	**1**	**2**	**3–5**	**6–11**
Arabidopsis ATH1	3273	870	167	5
HG-U133Plus2	7959	2410	517	17

Other motifs

The earlier discussion revealed that entire groups of CEL files showed biased results for probes with particular motifs. The most extreme case was the inflated values in some groups for probes containing the TTTTTT sequence, but these were not the only affected groups. We particularly noted strong biases involving probes containing the TAT motif in GSE3678, involving all of AAAA, ACTGC, TGGGA and GCCA in GSE6021 and GSE6022, and involving AGA in GSE2125. We suspect that these biases are due to variations in preparation of the material, though the nature of the tissue being analysed may be relevant. It would appear that about 5% of GSEs are associated with some bias of this type.

The effect of the scanner

In general it is rather difficult to be sure of the relative contributions of the scanner, the washers and the laboratory itself, on the clarity with which data are visualised, though the extreme case represented by the cloud of data points at the top of Figure 4 is certainly associated with the scanner. That particular scanner was known to be faulty and was being

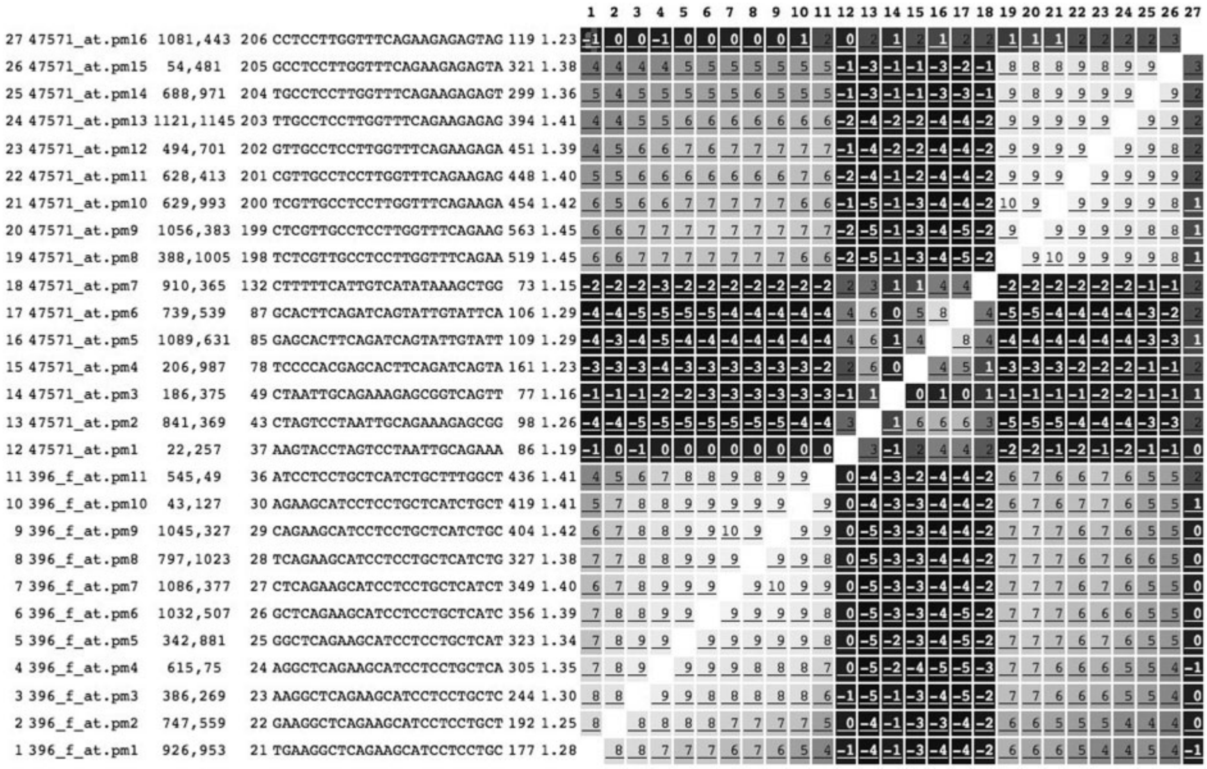


Figure 9: Heatmap showing unwanted correlations between two unrelated probe-sets. The correlations are between probes containing the CCTCC motif. Key as for Figure 2.

adjusted by Affymetrix technicians during the period in which high blurring occurred. We have recently downloaded a further 10 000 CEL files from GEO and can reassure users that the scanner is no longer associated with highly blurred data. Nevertheless, it appears that the vast majority of CEL files are noticeably blurred (as Figure 4 suggests). This will affect all probes. Results will not be biased with respect to genes, but high expression levels will be under-stated and low expression levels will be over-stated.

Trapped bubbles, scratches, etc

Most CEL files appear to contain minor imperfections. Although the numbers of affected probes are small (relative to the number of probes on an array), about one CEL file in three contains at least one 5 × 5 region in which a majority of the probes have an extreme value (a value that, after standardisation, is >3 SDs from the normal value for that probe). About 1% of CEL files have as many as 500 (overlapping) flawed regions of this type.

Outlier CEL files

As Figure 7 suggests, only about 0.2% of CEL files have means that are >4 SDs from the median. Such files should most probably be discarded, as should the 0.1% of CEL files that have probes with values having a variance <4 SDs below that usually observed.

Degradation

Figure 8 suggests that, as one would expect, there is noticeable degradation in the majority of CEL files. We are developing a procedure to quantify that degradation (in terms of proportion per base) so that in future it may be possible to compensate for the degradation and thereby provide more coherent probesets. Figure 8 demonstrates that about 2% of the files that we examined were so severely degraded that their worth was questionable.

CONCLUSIONS

The curse of dimensionality applies to post-genomic experiments which report the expression of tens of thousands of genes in only a few conditions. Many of the analysis problems previously faced by users of microarray experiments will likely be faced by analysts of ultra-high-throughput sequencing experiments in the future. It is therefore probable that many of the ideas developed to analyse microarray data can be recycled. Although we have focussed here on the mining of GeneChips, we expect that once large amounts of new sequence data are deposited in public databases, similar biases will be found. Moreover, it is possible that once this data has been produced, the community will start to move onto next–next-generation technology, and be, in all likelihood, facing the curse of dimensionality once more. Since this cycle could repeat indefinitely, it is important to establish how best to mine large surveys of expression data, of whichever flavour, in order to establish what biology can be inferred from such a wealth of untapped data.

A central part of this endeavour will be to develop a detailed understanding of the causes of imperfections in the data. The widespread uptake of Affymetrix technology, using standardised probe and chip designs, hybridisation protocols and scanners, means that the GeneChip data now in GEO provides a wonderful opportunity to develop this new area of science. We hope that this review helps to highlight the rapid progress being made in this direction. The authors of many papers utilising Affymetrix GeneChips will have only scratched the surface of understanding what their expression values are telling them about their biological system of interest, and analysis improvements can be widely applied. Moreover, the data has resulted from large investments of time and money by scientists, and their funding sources. Importantly, in the current economic climate, this novel area of science is incredibly cost-effective as it allows us to mine hundreds of thousands of GeneChips only a few years after some of the community balked at the cost of generating a handful of such experiments.

Mining large surveys of GeneChip data is starting to identify probes that show unusual behaviour. The existence of correlations between outliers may result from different processes affecting particular subsequences within the 25mer sequence. One such family has been associated with how the RNA is prepared for hybridisation and another family has been associated with how probes in close proximity interact with each other. But in both of these cases a combination of identifying individual cases of outliers [10, 13], and a search for enhanced correlations, such as described here and in [11], provide effective ways of identifying which probes are not reliable target RNA measurements.

Moreover, our strategy of searching for correlations in probes which should not be correlated has enabled us to identify a further group of correlated outliers, resulting from how the raw image is captured. Light from bright probes leaks over into the neighbouring cells. All neighbours of bright probes will show correlated changes in the amount of leaked light, and the fraction of light will depend upon the optical properties of the scanners. We hypothesise that there could be other biases affecting subsets of the probes on GeneChips. We believe our strategy of studying outlier correlations will be effective at identifying the non-biological causes of other families of outliers. We are developing methods to perform this analysis [23].

SUPPLEMENTARY MATERIAL

Supplementary material is available at *BFGP* online.

Key Points

- Our results are based on meta-analysis of thousands of CEL files.
- Individual probes with particular motifs, e.g. G-spots and CCTCC, may mislead.
- Some CEL files are compromised by blur.
- The blur from scanners is time-dependent.
- RNA degradation may seriously affect some probe values.

Acknowledgements

The authors are grateful to Dr William Langdon for producing a number of software tools used in our research.

FUNDING

BBSRC grant (No. BBE0017421) (to O.S.G. and M.S.); Strategic Studentship from the BBSRC (No. BBSSH2005A11996A) (to J.R.); CONACYT scholarship (No. 178596) (to J.M.A.S.); BBSRC grant (No. BBE01772X1) (to N.G.); scholarship from the University of Sindh (NO.SU/PLAN/F.SCH/611) (to F.M.).

References

1. Barrett T, Troup DB, Wilhite SE, *et al.* NCBI GEO: mining tens of millions of expression profiles – database and tools update. *Nucleic Acids Res* 2007;**35**:D760–5.
2. Cope LM, Irizarry RA, Jaffee H, *et al.* A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics* 2004;**20**:323–31.
3. Harrison AP, Johnston CE, Orengo CA. Establishing a major cause of discrepancy in the calibration of Affymetrix GeneChips. *BMC Bioinformatics* 2007;**8**:195.
4. Dai M, Wang P, Boyd AD, *et al.* Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* 2005;**33**:e175.
5. Okoniewski MJ, Miller CJ. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics* 2006;**7**:276.
6. Gao Y, Wolf LK, Georgiadis RM. Secondary structure effects on DNA hybridization kinetics: a solution versus surface comparison. *Nucleic Acids Res* 2006;**34**:3370–7.
7. Heim T, Tranchevent LC, Carlon E, *et al.* Physics-based analysis of Affymetrix microarray data. *J Phys Chem B* 2006;**110**:22786–95.
8. Stalteri MA, Harrison AP. Interpretation of multiple probesets mapping to the same gene in Affymetrix GeneChips. *BMC Bioinformatics* 2007;**8**:13.
9. Sanchez-Graillet O, Rowsell J, Langdon WB, *et al.* Widespread existence of uncorrelated probe intensities from within the same probeset on Affymetrix GeneChips. *J Integrat Bioinform* 2008;**5**:98.
10. Wu C, Zhao H, Baggerly K, *et al.* Short oligonucleotide probes containing G-stacks display abnormal binding affinity on Affymetrix microarrays. *Bioinformatics* 2007;**23**:2566–72.
11. Upton GJ, Langdon WB, Harrison AP. G-spots cause incorrect expression measurement in Affymetrix microarrays. *BMC Genomics* 2008;**9**:613.
12. Langdon WB, Upton GJ, Harrison AP. Probes containing runs of guanines provide insights into the biophysics and bioinformatics of Affymetrix GeneChips. *Brief Bioinform* 2009;**10**:259–77.
13. Kerkhoven RM, Sie D, Nieuwland M, *et al.* The T7-primer is a source of experimental bias and introduces variability between microarray platforms. *PLoS ONE* 2008;**3**:4.
14. Van Gelder RN, von Zastrow ME, Yool A, *et al.* Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proc Natl Acad Sci USA* 1990;**87**:1663–7.
15. Baugh LR, Hill AA, Brown EL, *et al.* Quantitative analysis of mRNA amplification by in vitro ranscription. *Nucleic Acids Res* 2001;**29**:E29.
16. Arteaga-Salas JM, Zuzan H, Langdon WB, *et al.* An overview of image processing methods for Affymetrix GeneChips. *Brief Bioinform* 2008;**9**:25.
17. Langdon WB, Upton GJ, Camargo R, *et al.* A survey of spatial defects in Homo sapiens Affymetrix GeneChips. *Trans Comp Biology Bioinform* 2008, doi:10.1109/TCBB.2008.108.
18. Upton GJ, Lloyd CJ. Oligonucleotide arrays: information from replication and spatial structure. *Bioinformatics* 2005;**21**:4162.
19. Arteaga-Salas JM, Harrison AP, Upton GJG. Reducing spatial flaws in oligonucleotide arrays by using neighbourhood information. *Stat Appl Genet Mol Biol* 2008;**7**:29.
20. Schmid M, Davison TS, Henz SR, *et al.* A gene expression map of Arabidopsis development. *Nat Genet* 2005;**37**:501–6.
21. Kumari S, Verma LK, Weller JW. AffyMAPSDetector: a software tool to characterize Affymetrix GeneChip expression arrays with respect to SNPS. *BMC Bioinformatics* 2007;**8**:276.
22. Alberts R, Terpstra P, Li Y, *et al.* Sequence polymorphisms cause many false *cis* eQTLs. *PLoS One* 2007;**2**:e622.
23. Langdon WB, Harrison AP. Evolving DNA motifs to predict GeneChip probe performance. *Algorithms Mol Biol* 2009;**4**:6.