Affymetrix probes containing runs of contiguous guanines are not gene-specific

Graham J Upton[1], William B Langdon[1] & Andrew P Harrison[1]

[1]*Departments of Mathematical and Biological Sciences, University of Essex, Colchester, Essex, CO4 3SQ, England. These authors contributed equally to this work. Correspondence should be addressed to A. P. H. (harry@essex.ac.uk)*

High Density Oligonucleotide arrays (HDONAs), such as the Affymetrix HG-U133A GeneChip, use sets of probes chosen to match specified genes, with the expectation that if a particular gene is highly expressed then all the probes in the designated probe set will provide a consistent message signifying the gene's presence. However, we demonstrate by data mining thousands of CEL files from NCBI's GEO database that 4G-probes (defined as probes containing sequences of four or more consecutive guanine (G) bases do not react in the intended way. Rather, possibly due to the formation of G-quadruplexes, most 4G-probes are correlated, irrespective of the expression of the thousands of genes for which they were separately intended. It follows that 4G-probes should be ignored when calculating gene expression levels. Furthermore, future microarray designs should make no use of 4G-probes.

Microarrays are now commonly used to measure gene expression. One of the most popular microarray platforms is the Affymetrix GeneChip. In GeneChip arrays probe sequences of nominal length 25 bases are created by photolithography. The probes form pairs: a so-called Perfect Match (PM) probe, and a mismatch (MM) probe that is identical to the PM probe with the excep-

tion that the 13th base is the complement of that in the PM probe. Each pair of probes belongs to a probe set (typically of 11 or 16 probe pairs) with each probe set being intended to provide information concerning the prevalence of a single gene. For some genes there may be more than one dedicated probe set.

There are a number of alternative software tools for calculating a single measure of gene expression for a probe set: e.g. MAS5[1], dChip[2], RMA[3] and GCRMA[4]. To calculate the value of the expression measure, all the probes (or at least all the PM probes) in a probe set are used. It is therefore imperative that faulty probes are identified and excluded so as to minimise their impact on the biological interpretation of the data. Fortunately the existence of large datasets such as that contained in GEO[5] now provide an opportunity to identify probes that show unexpected behaviour.

Even within a probe set, subsets of probes may be measuring different exons and thus, potentially, different transcripts. Therefore, biological signals such as alternative splicing need to be taken into account[6]. In order to circumvent the biological variation in GeneChip data caused by splicing we have focused on groups of probes which map uniquely to the same exon. Indeed, by making this choice, we are identifying groups of probes whose expression should be perfectly correlated. We are therefore safe to assume that divergences away from perfect correlation are due to artifacts resulting from the technology. Naturally, measurement noise will prevent perfect correlation, and this will be most marked in cases of low expression.

Suppose probes A, B, and C refer to the same exon, whilst probe D refers to a different exon. Suppose further that probes A and B are uncorrelated, whereas probes B and C are highly

correlated, as are probes A and D. Since both probes A and B are capable of high correlations, both are responding to real signals, but the signal to which probe A is responding is not related to the exon for which it was intended. In studying the correlation between multiple probes drawn from single exons we observed some remarkably low (even negative) correlations between probes. It was this that led to our identification of a large group of misbehaving probes. These are probes that include sequences of four or more consecutive guanine (G) bases.

It has previously been reported that GeneChip probes containing a sequence of guanines have abnormal binding behaviour compared with other probes and do not covary with other probes that interrogate the same gene[7]. We confirm that these probes show abnormal behaviour with respect to the other probes in the same gene. However, we show that this is not simply due to separate isolated probes misbehaving, instead we show that it is because the probes containing runs of guanines are being affected coherently across thousands of GeneChips. We believe that the behaviour of these probes may be a consequence of the formation of G-quadruplexes[8].

We focus on the GeneChip oligonucleotide microarrays manufactured by Affymetrix. Since a major application of microarrays has been to studies of human diseases, we have concentrated our effort on the output from the most popular human GeneChips, the HG-U133A arrays, though *the results apply to all GeneChip arrays*.

We suggest that future GeneChip designs should avoid including probes containing sequences of four (or more) guanines.

RESULTS

Correlations between probes

Our results use data from 6685 HG-U133A CEL files downloaded from the NCBI Gene Expression Omnibus (GEO) repository[5]. (After purified mRNA is processed and hybridised to an array, the Affymetrix scanner stores the average fluorescence intensity of each probe in the array in a data file, known as a CEL file.) The HG-U133A array contains about 22 300 probe sets matching to about 16 000 genes. After normalising each CEL file, we examined the correlations between probes within probe sets searching for anomalies. An example is provided by the probe set 31846_at which is one of two probe sets designed to match the gene RHOD. This probe set contains 16 PM probes all drawn from the same exon. The correlation between almost any pair of these PM probes is strongly positive, with the sole exceptions being that probe pm6 (the sixth of the PM probes in this probe set) has near-zero correlations with all the other probes. This is illustrated by scatter diagrams (**Fig.1**). Although probes 5 and 16 are separated by 192 bases their log(intensities) are highly correlated ($r = 0.86$), whereas probes pm5 and pm6, though separated by just 29 bases, have log(intensities) displaying a near-zero correlation. Near-zero correlations could occur with probes whose intensities are so low that they are dominated by the background 'noise' of the chip, but that is not the case in this instance since the average normalised intensities for probes pm5, pm6 and pm16 are 225, 389 and 504, respectively.

Figure 1 near here

To test the hypothesis that probe pm6 is related to other probes, we determined its correlation with every other probe (PM and MM) in the entire array. There are 10 409 other probes (drawn from 5341 probe sets) that have correlations with probe 6 that exceed 0.8 (and no fewer than 151 that have correlations exceeding 0.95). About half these high-correlating probes are mismatch probes. An example scatter diagram (**Fig. 2**) shows the correspondence of the variation in the values of the pm6 probe with that displayed by the first PM probe in the unrelated probe set 219297_at (which was designed to measure activity of the WDR44 gene).

Figure 2 near here

Correlations between probes containing sequences of guanines

Upon listing the probes most highly correlated with probe pm6 from the 31846_at probe set (base sequence TCCTGGACTGAGAAAGGGGGTTCCT) it becomes apparent that there is a common theme: each probe contains a sequence of four or more consecutive Gs. For example, the pm1 probe in 219297_at (used in **Fig. 2**) begins with six Gs (GGGGGGATAGTCTTGTTTC-TAGCTT). By contrast, in 31846_at, probes pm5 (GAACTCCACTGCAACAGACGGGCGC) and pm16 (TTCCCACCTGTCATACTGGTAACTG) contain sequences of only 3Gs and 2Gs, respectively.

Given that the correlations are a consequence of sequences of guanines in the probe sequence, two questions that immediately arise are 'Is the *location* of the consecutive run of guanines relevant?' and 'Is the *number* of consecutive guanines relevant?'. To get clear answers to these questions we focus on probes that have only one sequence of two or more guanines. We will refer to the location of the sequence within the probe as the *G-spot* and we now examine how inter-probe correlations are affected by the location and length of the G-spot.

The effect of the location of the G-spot

We will use probes containing a single sequence of exactly four guanines to demonstrate that the location of the G-spot within the probe has a considerable bearing on the correlation. Let l denote the first base of the G-spot (so that, for these probes, $l = 1, 2, \ldots, 22$). For each value of l, **Table 1** reports the number of probes of this type and the average correlation between pairs of probes both of this type.

Table 1 near here

Table 1 shows distinct design preferences on the part of Affymetrix since probes starting GGGG are relatively common (an unfortunate choice under the circumstances) whereas cases where the GGGG sequence straddles the central probe (i.e. probes with the GGGG sequence commencing at one of locations 10 to 13) are relatively infrequent.

For all values of l the average correlation between pairs of probes with G-spot at l is significantly greater than zero, with the overall maximum at $l = 1$ and minima at $l = 5$ and $l = 21$.

Figure 3 near here

Further detail is provided by Figure 3 which shows the cumulative distribution of the $54\,615 = (331 \times 330/2)$ individual correlations between pairs of probes having $l = 1$. The figure demonstrates that fewer than 1% of these probe pairs have negative correlations, whereas 14% have correlations that exceed 0.9.

The effect of the length of the G-spot

We next examine how varying the number of consecutive guanine bases affects the correlation (**Table 2)**. For simplicity and since the correlation is greatest when the G-spot is at the 5' end of the probe, in this section all the probes start with the G-spot. Table 2 demonstrates that, while the correlation between probes beginning with exactly three guanines is appreciably greater than zero, it is pairs of probes beginning with four or more guanines for which the correlation is remarkable. As further confirmation we looked also at probes beginning with the sequence GGXGG (where X is any base other than G). The average correlation amongst pairs of these probes was 0.06, confirming that for high correlations *consecutive* guanines are required.

Table 2 near here

Correlation between probes having different locations for their G-spots

Returning to probes containing a single sequence of four guanines, **Table 3** displays the average correlations between two such probes, when one has its G-spot at the start of the probe ($l = 1$) and the other does not. Whilst all the average correlations considerably exceed zero, the peak correlation when both probes have $l = 1$ is accentuated.

Table 3 near here

For probes containing a single sequence of four guanines, Tables 1 and 3 provide information concerning 43 of the 253 average correlations corresponding to pairs of values of l. Figure 4 provides a contour diagram that gives an overview of the entire correlation surface. Denoting the two values of l by l_1 and l_2 there is a sharp peak at $l_1 = l_2 = 1$, a ridge along $l_1 = l_2$, with a secondary peak near $l_1 = l_2 = 14$ and a general decrease as $|l_1 - l_2|$ increases.

Figure 4 near here

Other types of array

It seemed unlikely that the effect was related in any way to the organism under investigation. To confirm this we analysed data from a set of ATH-121501 GeneChips (for *Arabidopsis thaliania*): the average correlation between probes starting with four Gs was 0.86.

Discussion

The previous section has demonstrated that probes containing a G-spot of four or more bases are very likely to be highly correlated with many other probes not in their own probe set. The phenomenon is evidently not related to genetics, so that it is clear that the pragmatic solution is simply to eliminate G-spot probes from future array designs. However, we cannot resist making some suggestions concerning the possible causes of the G-spot effect. In particular, we believe the G-spot effect results from probe-probe interactions occurring on GeneChips.

The potential for the formation of G-quadruplexes

The high density of synthesis sites on the surface of Affymetrix GeneChips leads to crowded conditions on the array surface[9]. Assuming a stepwise synthesis yield for probes of 95% per base and that the density of initiation sites for probe synthesis is 5×10^{17} molecules/m^2, the average distance between full-length 25mer probes is about 3 nm. As the lengths of the probes may be up to 22 nm, it is thus likely that probes can come into contact[10].

The high density of probes results in considerable differences between the rates and efficiencies of hybridisation for probes in solution and for probes tethered to a surface[11]. These differences may be due to electrostatic repulsion of the high charge density on arrays resulting from the phosphate backbones of the probes[12]. The electrostatic effects act to reduce the stability of a probe-target duplex[12] and it has been suggested[13] that probe-probe associations involving only a few residues will be able to compete with the formation of probe-target duplexes. There have been initial attempts to model probe-probe duplexes[10]. However, a full model is not computationally tractable[10] and there are presently no theoretical results which describe under what conditions probe-probe interactions occur. We believe the co-ordinated behaviour of G-spot probes results not from a probe-probe dimer but from a higher-order binding of four DNA strands.

The Hoogsteen hydrogen-bonded guanine (G)-tetrad is a four-stranded DNA spiral stack held together by eight hydrogen bonds per level[8]. Even G-quadruplexes formed by quite short runs of Gs along the 4 DNA strands can be thermally stable up to 90°C [14]. G-quadruplexes are stabilised by positive sodium or potassium cations centrally placed between adjacent (G)-tetrads. The cations are thus close to four electronegative oxygens in the (G)-tetrad above and four more in the (G)-tetrad below and act to reduce the repulsion of the oxygen atoms via the formation of cation-dipole interactions. We suggest that probes in close proximity which contain a run of four or more contiguous guanines, may sometimes interact to form a G-quadruplex.

It has been argued[7] that probes do not form G-quadruplexes on GeneChips because the probes are immobilised and so it must be the targets that form quadruplexes which cause G-spot

probes to show abnormal binding. However, since the probes are sufficiently close to each other, and attached via linkers, they have enough flexibility to interact closely. Moreover, because the probes run in parallel and contain identical sequences, we believe that this provides an ideal opportunity for G-quadruplexes to form where there are runs of contiguous guanines. The coherence between all G-spot probes leads us to suggest that the problem lies with the probes and the GeneChip technology rather than the incoherently randomly segmented targets themselves.

Brightness and chip-to-chip variability of the G-spot probes

The formation of a G-quadruplex will result in four probes having their guanines facing inwards towards the quadruplex. Thus these bases will not be available to hybridise with targets. Yet probes starting with GGGG are on average about twice as bright as other strongly correlated probes whilst containing only an average number of Cs and Gs. We suggest the fact that G-spot probes tend to be bright may be due to the nature of the hybridisation on the surface of GeneChips resulting from the high packing density of probes. Models of the hybridisation dynamics of surface-immobilised DNA[15] show that as probes interact more strongly so the nucleation sites available are modified with resulting changes in the hybridisation affinity related to the packing density of probes. When further apart the affinity between probe and target increases rapidly. The effective association rate is proportional to $(\text{probe density})^{-1.8}$. We suggest that, on the surface of a chip, in a G-spot region, there will be a number of probes that form G-quadruplexes. The G-quadruplex acts to bind four probes together and these probes do not hybridise to the target. This means that the

remaining probes have more space and will have increased target affinity due to a lower probe density. Indeed the run of Gs on the remaining probes is available to act as an efficient nucleation site for hybridisation. This could encourage non-specific binding of labelled targets.

Implications for the use of existing GeneChips

Our findings have several implications. The extent to which a particular 25 base sequence will form probe-probe interactions may depend upon a range of factors which vary from experiment to experiment. Thus probe-probe interactions need to be taken into account when modelling the affinity of the probe.

We have detected the G-spot effect from looking at the correlations between probes. Thousands of probes behave coherently from sample to sample. We suggest that there is one or more aspect to the preparation of each GeneChip and/or sample which affects the extent of the formation of G-quadruplexes across the whole GeneChip. There are many things which effect the stability of quadruplexes. These include monovalent cations. Potassium has a larger affinity for a quadruplex than sodium. (However sodium is likely to be the dominant cation during hybridisation). Conversely lithium acts to destabilise G-quadruplexes. Molecular crowding also helps to induce quadruplex formation[16]. (However we suggest this should be constant from chip to chip). Ethanol has recently been shown to be a better inducer of quadruplexes than even potassium cations[17] (ethanol is used in the preparation of nucleic acids). Even the life-history of the chip, such as whether it has been stored at low/high temperatures, or preheating the Chip prior to hybridisation,

may all alter the population of quadruplexes on the surface of the chip.

The existence of correlations between probes that only have a relatively short sequence in common suggests that hybridisation on GeneChips may be dominated by a few hot spots for some probes. In such probes the other bases have less influence on the binding between them and labelled target molecules. To correct for the effects of cross-hybridisation, greater weight needs to be attached to these hot spots rather than only studying the overlap between probe and target across the whole probe.

Our results also imply that designers of future high density oligonucleotide arrays need to avoid runs of contiguous guanines and other such sequences that act to stabilise probe-probe interactions.

Methods

During 2007 we downloaded CEL files from the NCBI Gene Expression Omnibus (GEO) repository. By the end of that year we had tens of thousands of CEL files, including 6685 examples of the most popular GeneChip produced by Affymetrix for the human genome: the HG-U133A array. These CEL files, from 162 separate GEO series (GSE) of experiments, were created between January 2002 and February 2006 and subsequently uploaded to GEO by many independent experimenters.

The next step was to create "heatmaps" illustrating the correlations (in the log space) between all probes within each probe set. These were created using information from all the 6685 CEL files.

Each CEL file was separately log normalised and potential spatial flaws identified[18,19]. To avoid problems with results being dominated by a few outliers we excluded data for each probe if they were more than three standard deviations from the probe's mean. Secondly we excluded not only data flagged as potentially in a spatial flaw but also data within 60μm of a spatial flaw. Even after this ultra-cautious treatment, we had many thousands of data for each of approximately half a million probes. The resulting 22 299 visualisations are at http://bioinformatics.essex.ac.uk/users/wlangdon/HG-U133A. Inspection of the heatmaps provided an efficient method for identifying probes that, despite having reasonable average magnitudes, had low correlations with the other probes included in their subset.

When calculating the correlation between probes containing runs of Gs, firstly only PM and MM probes with a single sequence of 2 or more Gs were selected. These were then divided into subgroups according to the length of the run of Gs and the location of the first G in the sequence. To avoid inflating the average correlation by including probes that would have been expected to be correlated in the absence of the G-spot effect, within each subgroup only the first probe in any probe set was used. Similarly where probes have identical sequences, only one was include in the averages. In Tables 1, 2 and 3 all possible correlations between pairs of probes were calculated and averaged.

1. Affymetrix. Microarray suite users guide, 5th edition (2001).

2. Li, C. & Wong, W. H. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences* **98**, 31–

36 (2001).

3. Irizarry, R. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research* **31**, e15 (2003).

4. Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F. & Spencer, F. A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association* **99**, 909–917 (2004).

5. Barrett, T. *et al.* NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Research* **35**, D760–D765 (2007).

6. Stalteri, M. & Harrison, A. Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips. *BMC Bioinformatics* **8**, 13 (2007).

7. Wu., C., Zhao, H., Baggerly, K., Carta, R. & Zhang, L. Short oligonucleotide probes containing G-stacks display abnormal binding affinity on affymetrix microarrays. *Bioinformatics* **23**, 2566–2572 (2007).

8. Burge, S., Parkinson, G., Hazel, P., Todd, A. & Neidle, S. Quadruplex DNA: sequence topology and structure. *Nucleic Acids Research* **34**, 5402–5415 (2006).

9. Glazer, M. *et al.* Kinetics of oligonucleotide hybridization to photolithographically patterned DNA arrays. *Analytical Biochemistry* **358**, 225–238 (2006).

10. Burden, C., Pittlekow, Y. & Wilson, S. Adsorption models of hybridisation and post-hybridisation behaviour on oligonucleotide microarrays. *Journal of Physics: Condensed Matter* **18**, 5545–5565 (2006).

11. Peterson, A., Heaton, R. & Georgiadis, R. The effect of surface probe density on DNA hybridization. *Nucleic Acids Research* **29**, 5163–5168 (2001).

12. Vainrub, A. & Pettitt, B. M. Coulomb blockage of hybridization in two-dimensional DNA arrays. *Physical Review E* **66**, 041905 (2002).

13. Forman, J., Walton, I., Stern, D., Rava, R. & Trulson, O. Thermodynamics of duplex formation and mismatch discrimination on photolithographically synthesised oligonucleotide arrays. *in N. Leontis and J. SantaLucia, editors of Molecular Modeling of Nucleic Acids, ACS Symposium Series, Am. Chem. Soc.* **682**, 206–228 (1998).

14. Blume, S., Guarcello, V., Zacharias, W. & Miller, D. Divalent transition metal cations counteract potassium-induced quadruplex assembly of oligo(dG) sequences. *Nucleic Acids Research* **25**, 617–625 (1997).

15. Hagan, M. & Chakraborty, A. Hybridization dynamics of surface immobilised DNA. *Chemical Physics* **120**, 4958–4968 (2004).

16. Kan, Z. *et al.* Molecular crowding induces telomere G-quadruplex formation under salt-deficient conditions and enhances its competition with duplex formation. *Angew. Chem. Int. Ed.* **45**, 1629 (2006).

17. Vorlickova, M., Bednarova, K. & Kypr, J. Ethanol is a better inducer of DNA guanine tetraplexes than potassium cations. *Biopolymers* **82**, 253–260 (2006).

18. Langdon, W. B., Upton, G. J. G., da Silva Camargo, R. & Harrison, A. P. A survey of spatial defects in Homo Sapiens Affymetrix GeneChips. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2007). Submitted.

19. Langdon, W. B. & Harrison, A. P. Row quantile normalisation of microarrays. *Genomics* Submited to CAMDA 2007 Special Issue.

Acknowledgements WBL was funded through research grant BBE0017421 from the Biotechnology and Biological Sciences Research Council. We acknowledge the vital contribution of Dr Renata da Silva Camargo, who created the Essex archive and identified the single-exon subsets.

Competing Interests The authors declare that they have no competing financial interests .

Correspondence Correspondence and requests for materials should be addressed to Dr. A. P. Harrison. (email: harry@essex.ac.uk).

Table 1: Dependence of correlations between 4G-probes on the location of the G-spot ($l = 1$ means the probe's 5' end starts with GGGG)

Location of G-spot, l	1	2	3	4	5	6	7	8	9	10	11
Number of probes	331	173	220	229	265	203	225	224	218	340	187
Average correlation	0.68	0.38	0.36	0.38	0.30	0.34	0.40	0.41	0.47	0.59	0.49
Location of G-spot, l	12	13	14	15	16	17	18	19	20	21	22
Number of probes	185	207	181	194	234	235	244	251	284	224	250
Average correlation	0.55	0.60	0.60	0.59	0.58	0.56	0.56	0.51	0.46	0.44	0.45

Table 2: Correlations between probes having their single sequence of k Gs starting with the first base

Length of starting sequence, k	2	3	4	5	6	7
Number of probes	5189	1279	331	67	11	5
Average correlation	0.03	0.15	0.68	0.83	0.91	0.94

Table 3: Dependence of correlation between 4G-probes with G-spots in locations 1 and l

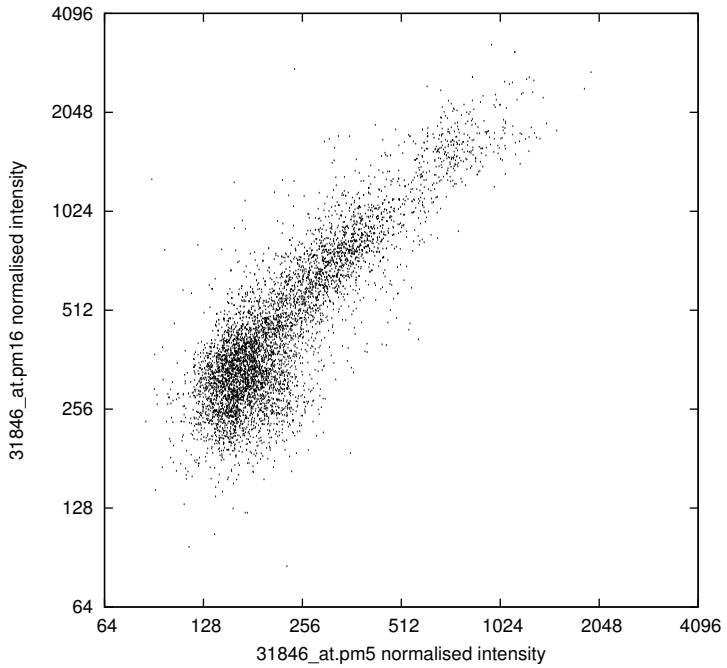
G-spot in 2$^{\mathrm{nd}}$ probe, l	1	2	3	4	5	6	7	8	9	10	11
Average correlation	0.68	0.30	0.29	0.28	0.25	0.28	0.29	0.28	0.29	0.33	0.28
G-spot in 2$^{\mathrm{nd}}$ probe, l	12	13	14	15	16	17	18	19	20	21	22
Average correlation	0.31	0.32	0.31	0.29	0.30	0.28	0.26	0.23	0.22	0.19	0.19

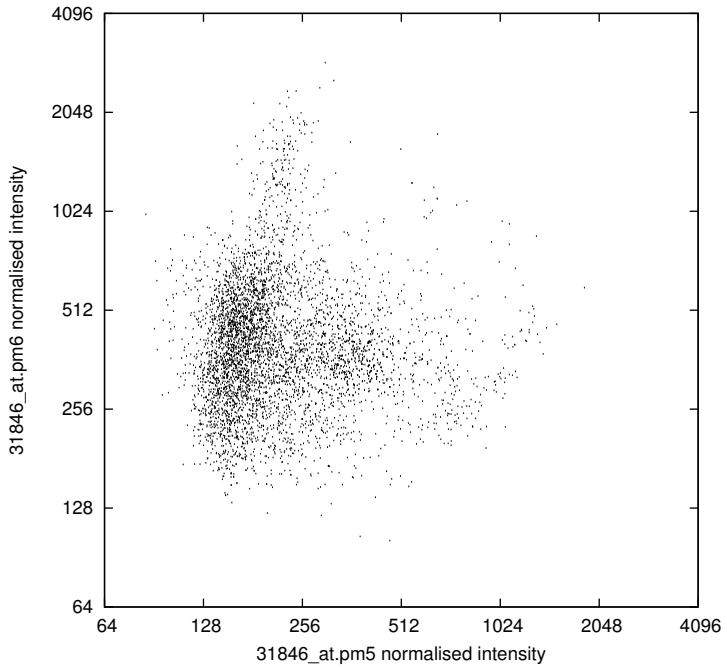
Figure 1 Scatter diagrams for probes from probe set 31846_at which matches the gene RHOD. (i) Probes PM 5 and PM 16 ($r = 0.86$); (ii) Probes PM 5 and PM 6 ($r = -0.01$).

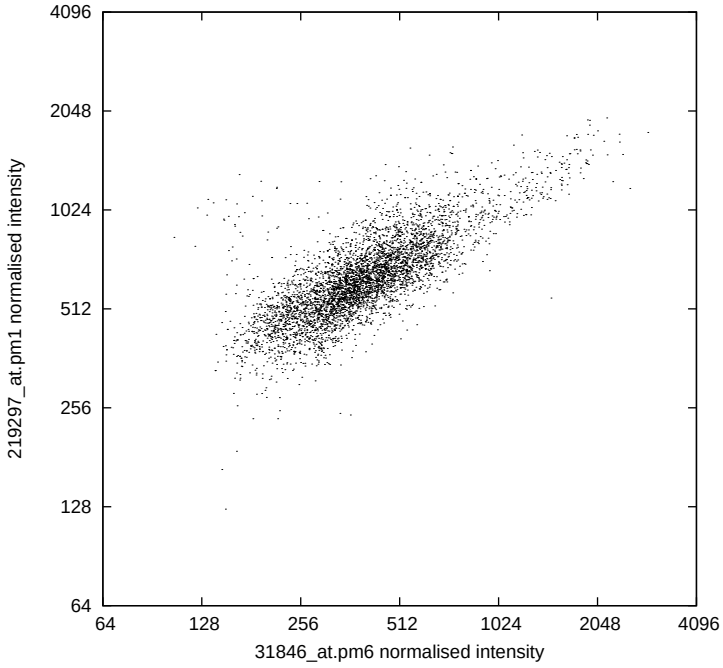
Figure 2 Scatter diagram comparing probe pm6 from probe set 31846_at with probe pm1 from probe set 219297_at ($r = 0.78$).

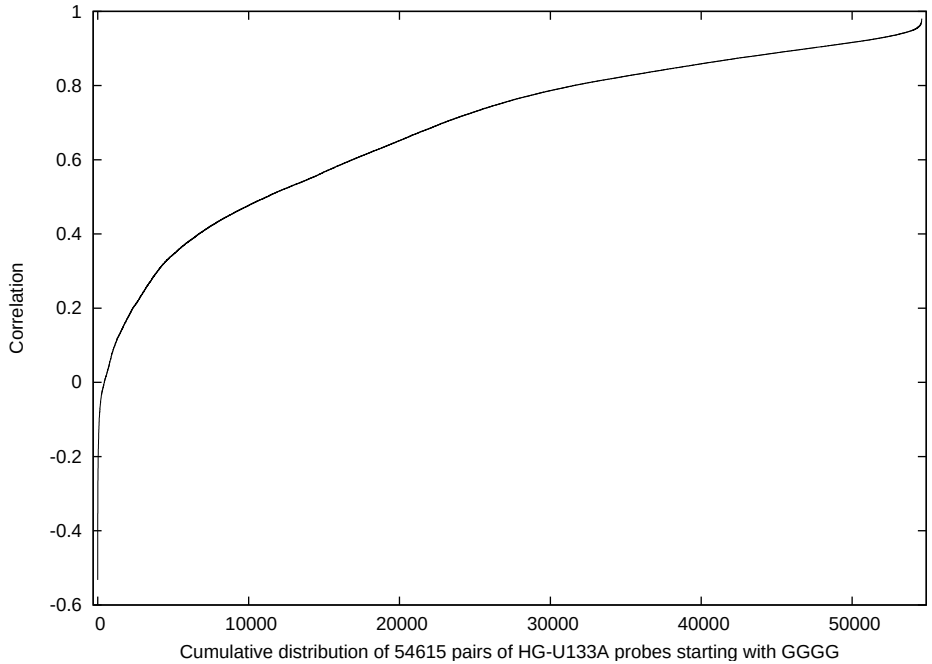
Figure 3 The distribution of the correlation between pairs of probes that begin with the sequence GGGG and contain no other runs of Gs.

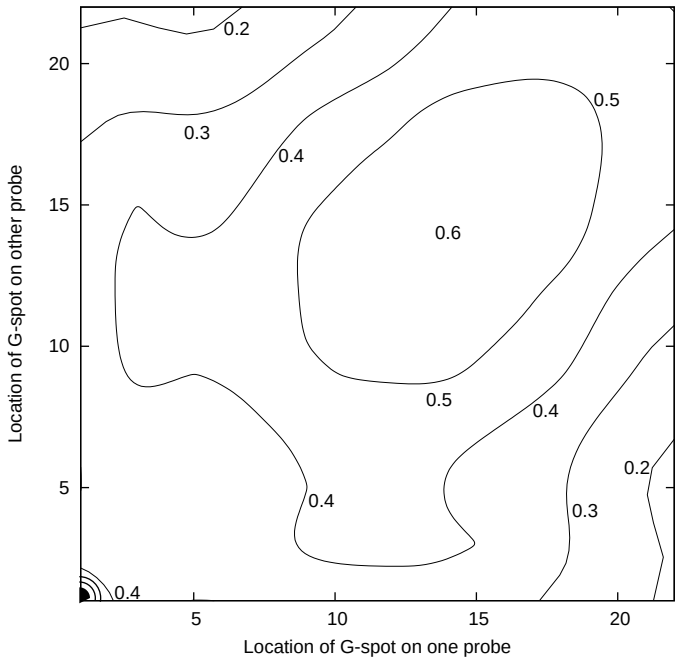
Figure 4 Contour diagram showing how average correlation varies with location of G-spot for pairs of probes (each with a single sequence of four guanines). The values in Table 1 correspond to the main diagonal and those in Table 3 to two edges. The maximum is at bottom left.


4096
2048
1024
512
256
128
64
31846_at.pm16 normalised intensity
64
128
256
512
1024
2048
4096
31846_at.pm5 normalised intensity


4096
2048
1024
512
256
128
64
31846_at.pm6 normalised intensity
64
128
256
512
1024
2048
4096
31846_at.pm5 normalised intensity


4096
2048
1024
512
256
128
64
219297_at.pm1 normalised intensity
64
128
256
512
1024
2048
4096
31846_at.pm6 normalised intensity


1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
Correlation
0
10000
20000
30000
40000
50000
Cumulative distribution of 54615 pairs of HG-U133A probes starting with GGGG


0.2
0.3
0.4
0.5
0.6
0.5
0.4
0.2
0.3
0.4
0.4
20
15
10
5
5
10
15
20
Location of G-spot on other probe
Location of G-spot on one probe