Widespread existence of uncorrelated probe intensities from within the same probeset on Affymetrix GeneChips

Olivia Sanchez-Graillet[1], Joanna Rowsell[1], William B. Langdon[1], Maria Stalteri[1], Jose M. Arteaga-Salas[1], Graham J. G. Upton[1] and Andrew P. Harrison[1,2]

[1] Departments of Mathematical Sciences and Biological Sciences, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, Tel: +44 (0)1206 873040

Summary

We have developed a computational pipeline to analyse large surveys of Affymetrix GeneChips, for example NCBI's Gene Expression Omnibus. GEO samples data for many organisms, tissues and phenotypes. Because of this experimental diversity, any observed correlations between probe intensities can be associated either with biology that is robust, such as common co-expression, or with systematic biases associated with the GeneChip technology.

Our bioinformatics pipeline integrates the mapping of probes to exons, quality control checks on each GeneChip which identifies flaws in hybridization quality, and the mining of correlations in intensities between groups of probes. The output from our pipeline has enabled us to identify systematic biases in GeneChip data. We are also able to use the pipeline as a discovery tool for biology.

We have discovered that in the majority of cases, Affymetrix probesets on Human GeneChips do not measure one unique block of transcription. Instead we see numerous examples of outlier probes. Our study has also identified that in a number of probesets the mismatch probes are an informative diagnostic of expression, rather than providing a measure of background contamination. We report evidence for systematic biases in GeneChip technology associated with probe-probe interactions. We also see signatures associated with post-transcriptional processing of RNA, such as alternative polyadenylation.

1 Motivation

Affymetrix GeneChip technology provides multiple measures of the expression level for each gene. Each probe is a 25-nt oligomer (25mer) and each probeset, designed to represent a different gene transcript, typically consists of eleven perfect match (PM) probes as well as corresponding mismatch (MM) probes. There has been tremendous success in applying Affymetrix GeneChip technology to illuminate the difference in gene-expression patterns for different species, tissues, phenotypes and disease states [1]. Following the publication of many experiments using GeneChips, public repositories of GeneChip data, such as GEO [2], were set up. These databases now contain the results from tens of thousands of GeneChips. The true value of having access to the combined data sources is only now becoming apparent. It is clear much more knowledge can be extracted from the combined data, than has already been published by individual studies. Many of the improvements in the use of GeneChips will derive from computational methods to either extract biological signals from the data or to remove systematic errors introduced by the technology.

[2] To whom correspondence should be addressed. E-mail: harry@essex.ac.uk

doi:10.2390/biecoll-jib-2008-98

2 Our pipeline

We have built a pipeline which analyses tens of thousands of GeneChips that are freely available in the public domain. Our pipeline brings together unique mapping of probes, quality control analysis on each GeneChip and data-mining signal intensities across many experiments. We have begun to use the products of this pipeline to make inferences about both biological systems and the biophysics of GeneChip technology.

2.1 Unique probe mappings to minimise the impact of cross-hybridization

We have previously shown that post-transcriptional processing steps, such as alternative polyadenylation and alternative splicing, leave their mark in GeneChip data [3]. In particular, probes that map to different exons may show differential regulation. In order to circumvent the biological variation in transcriptomic data caused by splicing we have focussed on groups of probes which map to the same exon. This choice identifies groups of probes whose expression should be correlated [4].

Previous work has focussed on mapping Affymetrix probes to different transcript variants in order to study alternative splicing [5, 6, 7]. However, these mappings still include probes that align exactly to more than one place in the genome. These cases are referred to as multiple targeting (MT) [8]. Probes are also susceptible to cross-hybridization (CH) where the probe sequence partially aligns to multiple genomic locations. We have attempted to avoid the issue of MT and CH by identifying probes which map uniquely to either an exon or an exon-exon junction. Ensembl [9] exon and spliced transcript sequences (release 48) and probe sequences for 12 Affymetrix Human GeneChip chip types [10] were used. The probe sequences were aligned against the spliced transcripts and exons using MegaBLAST [11]. Care was taken to ensure that the probe did not align to the antisense version of the gene (i.e. alignments with start position greater than end position). Ensembl exons with the same length, sequence and genomic coordinates have different identifiers if they are different sequence types (coding or non-coding). In our analysis we considered such exons as being synonymous because they will produce the same RNA sequence and will be equally able to hybridize to a probe which aligns to the exon.

In order to account for both multiple-targeting and cross-hybridization we calculated the alignment "value" for each probe by multiplying the alignment length and the percentage sequence identity. For example a probe that aligns to a sequence with 25 bases and percentage sequence identity of 80% would have an alignment value of 20 (25*0.8).

Figures 1 and 2 demonstrate which probes are considered to be mapping uniquely to an exon. We consider a probe to be mapping uniquely to an exon if it:

- aligns exactly (25 bases, 100% identity) to only one exon and to any of its synonyms (i.e. the probe maps to the same genomic region)
- maps to only one place on the same exon
- does not map to any exon-exon junctions

- does not map partially to any other exon (i.e. does not have alignment value between 20 and 25 with any other exon)

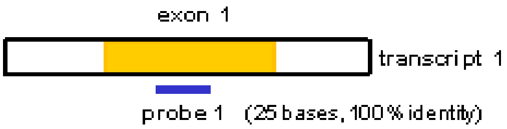

Figure 1: An example of a probe considered to be mapping uniquely to an exon. Probe 1 aligns to only one place in the exon with 25 bases (100% identity) and nowhere else.

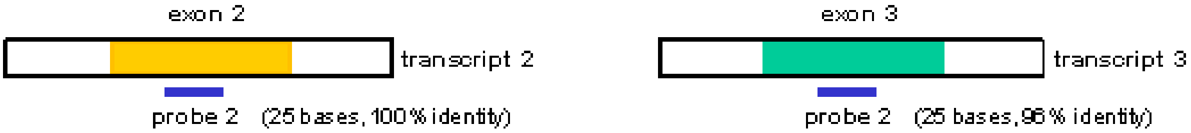
exon 2
transcript 2
probe 2 (25 bases, 100% identity)
exon 3
transcript 3
probe 2 (25 bases, 96% identity)

Figure 2: An example of a probe which is not considered to be mapping uniquely to an exon. Probe 2 aligns to one exon with 25 bases (100% identity) but also partially (25 bases and 96% identity) to another exon with an alignment value of 24 (25*0.96).

We have also identified those probes that map uniquely to an exon-exon junction. We consider a probe to be mapping uniquely to an exon-exon junction if it:

- aligns exactly (all 25 bases, 100% identity) to only one position on a spliced transcript

- does not map partially to any other spliced transcript (i.e. does not have an alignment value between 20 and 25 to any other spliced transcript)

- does not map to any exon with 25 bases, 100% identity

The unique junction probe may map to the same exon pair in different transcripts of the same gene. Figures 3, 4 and 5 demonstrate which probes are considered to be mapping uniquely to an exon-exon junction.

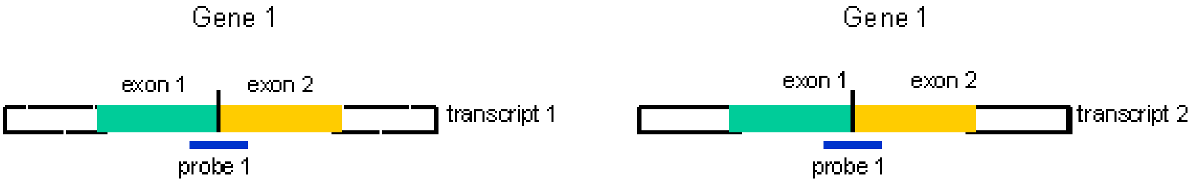

Figure 3: An example of a probe considered to be mapping uniquely to an exon-exon junction. Probe 1 maps to two exon-exon junctions that contain the same exons from different transcripts of the same gene (Gene 1). The exons of both transcripts are the same in terms of length and genomic position, i.e. the probe is mapping to the same genomic region.

We have created a database containing information about the probes, exons and transcripts. The computational system that we have developed implements the definition of unique mappings and uses the database information. By using this system, all probes which map uniquely to the same exon, as well as probes mapping uniquely to an exon-exon junction have been identified.

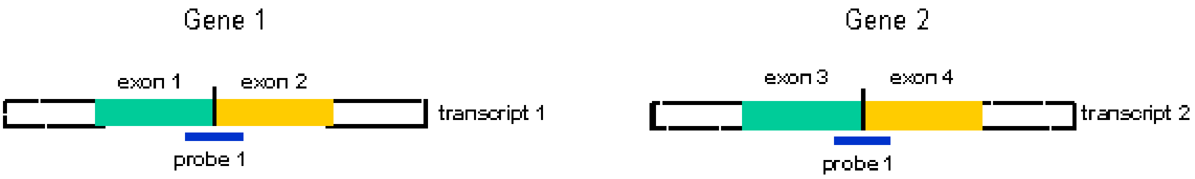

Figure 4: An example of a probe not considered to be mapping uniquely to an exon-exon junction. Probe 1 maps to two junctions from different genes (Gene 1 and Gene 2).

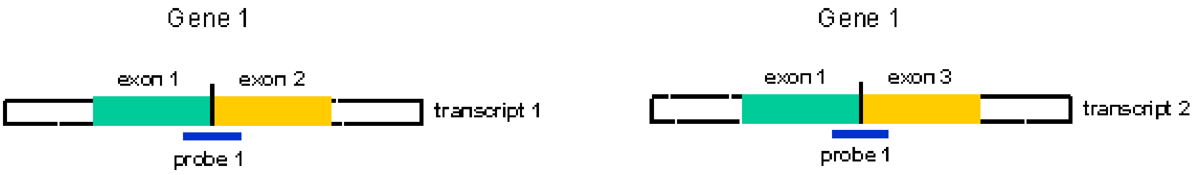

Figure 5: A different example of a probe not considered to be mapping uniquely to an exon-exon junction. Probe 1 maps to two exon-exon junctions which are from the same gene but contain different exons (i.e. probe 1 maps to different genomic regions) and therefore the mapping is not unique.

2.2 Correlation matrices for probesets

The preprocessing of GeneChips is an active research field with a number of different algorithms being developed [12]. The calibration of microarrays requires a correction for background signals as well as normalization of the data and the calculation of an expression measure. We have identified the calculation of the expression measure as being the dominant cause of variance in the lists of genes reporting to be differentially expressed in experiments [13].

The base assumption behind all the expression measures is that multiple probes from within the same probeset measure the same thing. Each probe should, ideally, provide an accurate and linear response to increasing amounts of the gene target. Although many probes perform as desired, there are many probes which are noticeably less responsive to target concentration. Some probes are either unresponsive (no hybridization signal) or invariant (same hybridization signal) across many observations. We have begun to examine the correlations between probes from within the same probeset. Our analysis is starting to produce results that indicate that many, if not all, of the existing expression measure calculations may be missing interesting biophysical effects. These include clear evidence for a significant fraction of mismatch probes which are measuring the effects of signal rather than background. We are also seeking to identify spurious probes in order to remove them from downstream analysis.

We use large numbers of GeneChip experiments obtained from GEO in order to identify those probes whose expression changes are corrupted in a systematic way. We have begun to identify the sources of these systematic errors through studying relationships apparent across tens of thousands of microarray experiments. In February 2007 we downloaded almost 40,000 Affymetrix GeneChip CEL files from GEO. Human, mouse and rat samples make up the majority of the CEL files.

Many of the GeneChips we have downloaded contain spatial flaws in their hybridization. We have developed methods to correct for these defects [14, 15], building upon our earlier work [16]. For all the chips of a given design we first identify and then remove defects from downstream analysis and then row quantile normalise the chip intensities so as to transform the intensities into a standard distribution. Row quantile normalization is quantile normalization in that the distribution of intensities is forced to take a standard form, but it also takes into account the way Affymetrix GeneChips are generated. Probes in a row have similar sequences. This is due to the way the probes are laid out by Affymetrix. This means all probes in a row tend to have similar GC content which in turn leads to individual rows showing biases in intensities different to their immediate neighbours. Quantile Normalizing each row independently acts to remove these biases.

We transform all intensities onto a log2 scale and then correlate the probe signals across all examples of a chip with a given design, e.g. HG-U133A, against the signals from all the probes taken from the same probeset. Figure 6 shows an example for two perfect match probes, PM9 and PM11 from probeset 208772_at, on the HG-U133A array. We collate all the correlations for a given probeset, including the correlations between perfect match probes, mismatch probes and perfect match and mismatch probes, into a matrix which we colour-code according to the correlation value. The data in Figure 6 is transformed into one number, the correlation. This is one of the squares in the matrix shown in Figure 7.

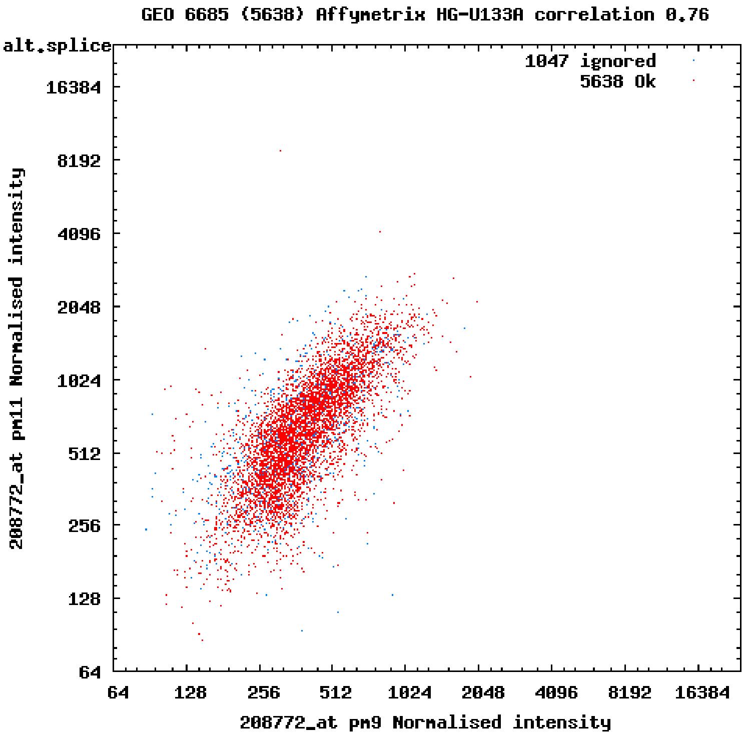

Figure 6: A scatter plot of the log2 intensities between perfect match probes 9 and probe 11 from probeset 208772_at, obtained from 5638 HG-U133A GeneChips. For this particular probe pair, 1047 of the original set of 6685 chips were flagged as possibly providing spurious signals or being close to dubious values.

The correlation matrices are proving to be an effective tool for establishing systematic biases in probes and probesets. Figure 8 shows the correlation matrix for probeset 200750_s_at on the HG-U133A chip. Figure 8 indicates that all the PM probes are correlated closely with each other, which means they can be reliably condensed into a single measure of expression. However, Figure 8 also shows that the MM probes are also closely correlated with the PM probes, and thus they are measuring the same thing as the PM, i.e. they are not measuring a background signal. Figure 9 shows that there is little or no correlation between any of the probes in probeset 204921_at on the HG-U133A chip. It is unlikely that this set can be used for diagnostic purposes, and care should be taken in interpretation if this probeset shows up as being differentially expressed in any experiment. Figure 10 shows that PM 5 and its mismatch in probeset 201131_s_at on HG-U133A are outliers, and should not be included in the calculation of any expression measure.

We have generated correlation matrices for all of the GeneChip types of the Human data in GEO (up to February 2007). The matrices can be obtained from `http://bioinformatics.essex.ac.uk/users/wlangdon`. We intend to produce similar correlation matrices for all the other organisms for which GeneChips exist, and which contain sufficient data to enable robust correlation values to be calculated.

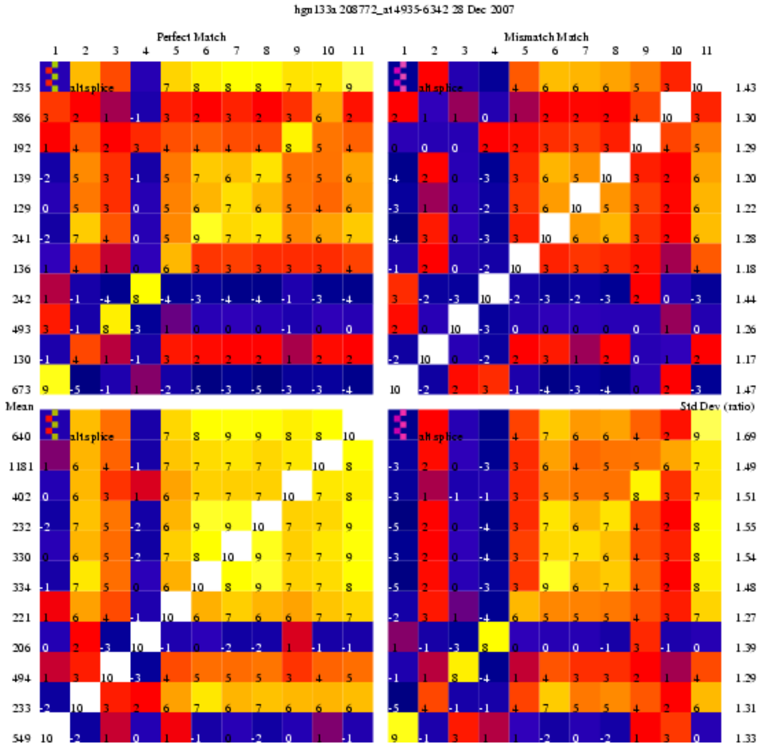

Figure 7: The correlations in intensities (log2) between probes in probeset 208772_at on the HG-U133A array. The numbers to the left indicate the mean linear intensity of each probe across all the data in GEO. The numbers to the right indicate the standard deviation of the normalised log2 intensities for all the data in GEO. The lower left quadrant details the correlations between perfect match probes and all the other perfect match probes in the probeset. The upper right quadrant details the correlations between all the mismatch probes. The upper left quadrant details the correlations between the perfect match and mismatch probes. The number in each square is the correlation ×10. The matrix is diagonally symmetric, and the diagonal corresponds to comparing a probe with itself – perfect correlation, and hence is scored 10. The correlation calculated for PM probes 9 and 11 , the data in Figure 3, is reported as 8 (0.76 multiplied by 10 and rounded).

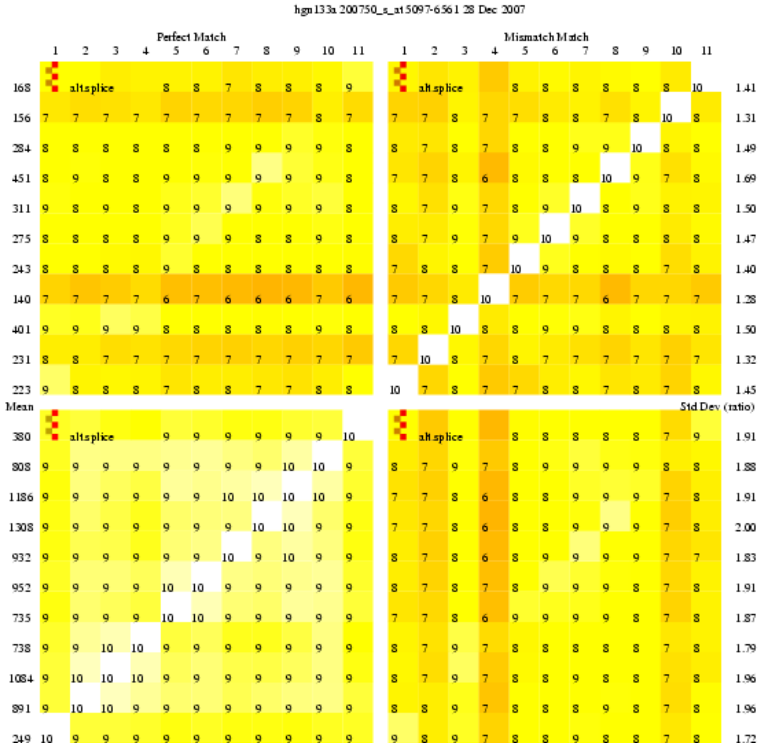

Figure 8: The correlations in intensities (log2) between probes in probeset 200750_s_at on HG-U133A. There is considerable correlation between all the probes, including mismatch probes. This suggests that in this probeset mismatch probes contain information about the expression of the transcript, rather than as a measure of the background signal.

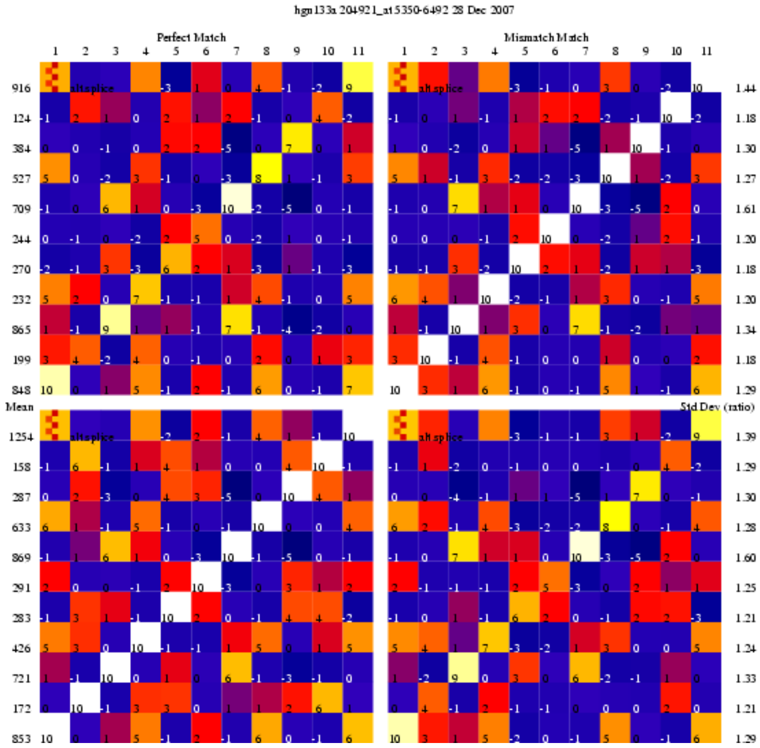

Figure 9: The correlation in intensities (log2) for probes in probeset 204921_at on HG-U133A. For this probeset there is little correlation between any of the probes – this indicates that this probeset will be uninformative.

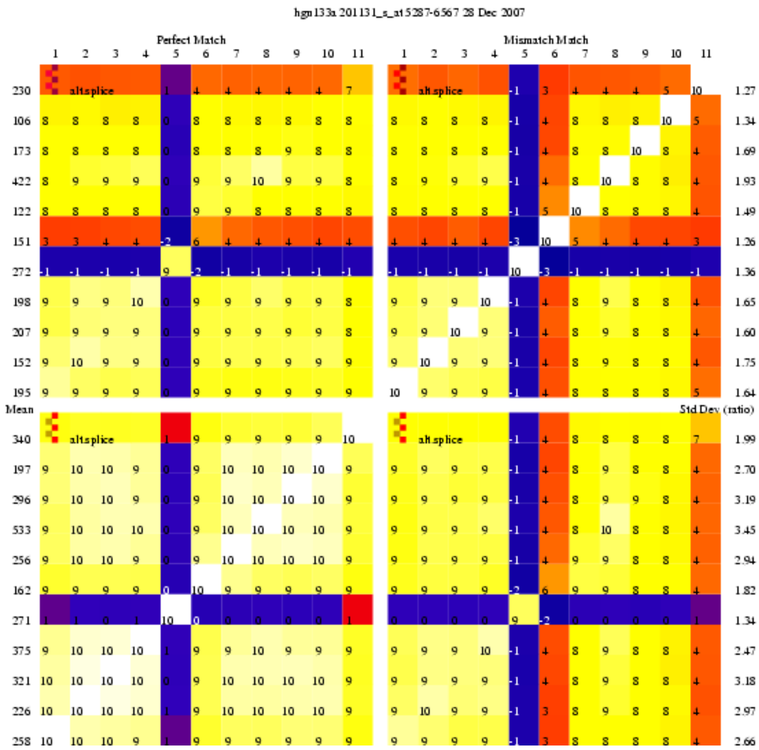

Figure 10: The correlations in intensities (log2) between probes in probeset 201131_s_at on HG-U133A. Probe 5 for both the PM and MM appears poorly correlated with the other probes. Outliers should be removed before the calculation of expression.

2.3 Using correlation matrices to identify signatures associated with post-transcriptional processing of RNA

Correlation matrices were created for every Ensembl exon (release 48) with unique Affymetrix probes aligning in the sense direction of the exon. These provide good controls because the groups of probes that align to the same exon are expected to show concordance in their expression. However, the effects of polyadenylation may result in a block of correlation within the matrix of an exon. Figures 11 and 12 both show blocks of high correlation in the matrix at positions of polyadenylation (polyA) sites, which we have obtained from the ASTD database [17]. We have started to analyse these steps as well as similar step structures due to competing splice sites (data not shown).

2.4 Why probes are not behaving as expected

In studying the correlation between probes drawn from a single exon we observe some remarkably low (even large negative) correlations between probes. Figure 13 shows a group of probes from within a probeset that all map to the same exon. Probes 6, 7 and 8 are all correlated indicating that they are measuring something. But these are not correlated with the rest of the probes in the probeset. Figure 14 shows that probes 6, 7 and 8 are closely correlated with tens of thousands of other probes. Indeed, they are correlated with the outlier probe seen in Figure 10, PM5 in probeset 201131_s_at. The probes within the family have a common run of four or more contiguous guanines. Probes containing runs of contiguous guanine have been shown to have abnormal binding affinities and be typically outliers within a probeset [18]. We con-

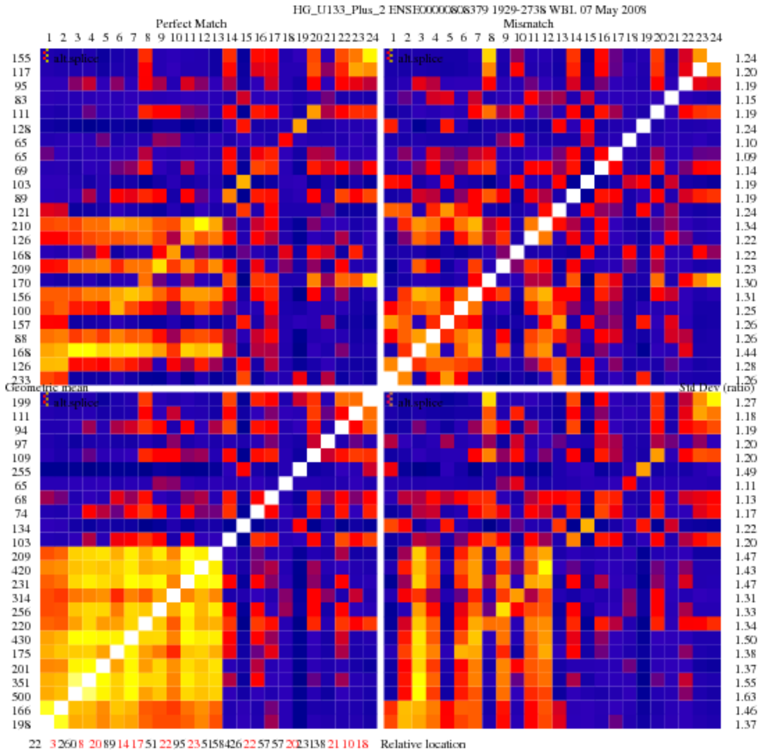

Figure 11: A correlation matrix for exon ENSE00000808379 with probes from HG-U133_Plus_2. Probes 1-13 lie before a polyA site and probes 14-24 are after. The matrix suggests that the GeneChips are detecting expression from transcripts which frequently use this site, as there is little correlated expression beyond the site. The numbers under the lower left quadrant (PMvsPM) detail the position of probes on the exon. The 1st base of probe 1 is at position 22 with respect to the start of the ENSEMBL exon, and the 1st base of probe 2 is at position 25 (22 + 3).

firm this observation, and go further by identifying that the probes containing runs of guanine are closely correlated. We expect this signature results from four adjacent probes containing a run of guanines interacting to form a G-quadruplex [19]. Determining the existence of probe-probe interactions on GeneChips may have widespread implications for the users of Affymetrix GeneChip 3' gene expression, tiling, exon and Genotyping arrays.

We are presently seeking other explanations for the causes of outliers. We expect that in a number of cases, the probe sequence will overlap the location of a Single Nucleotide Polymorphism (SNP). SNPs have been shown to cause divergent values for the differences in PM and MM intensities [20]. SNPs have also been found to cause the false identification of expressed quantitative trait loci using GeneChip data [21].

A fraction of the probes show little or no expression across all of the experiments in GEO. These probes are not useful measures of expression, and should therefore be excluded from any expression measure calculation. We are identifying such probes and exploring the reasons behind why probes are unresponsive.

Another issue we are considering is how to deal with probes which share a significant overlap in sequence. Indeed there are a number of examples of probe sequences which have multiple copies, i.e. the sequences of the probes are identical but they are assigned to separate probe-sets. There are also a large number of overlapping probes in which a significant fraction of sequence is shared. These overlapping probes are often strongly correlated, as expected, but these are not truly independent measurements. However, many of the summary expression

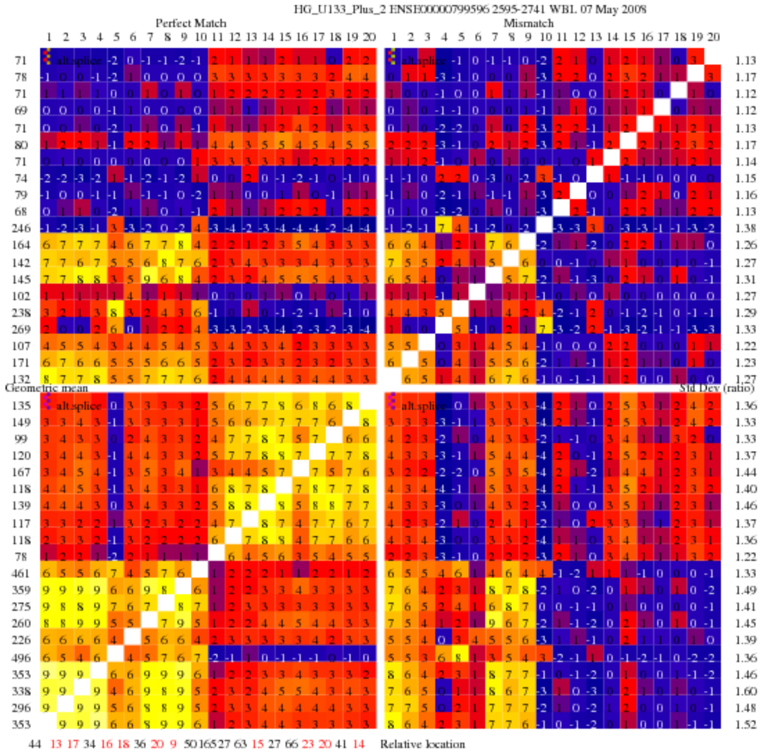

Figure 12: A correlation matrix for exon ENSE00000799596 with probes from HG-U133_Plus_2. Probes 1-10 align to the exon before a polyA site and probes 11-20 align after. The matrix suggests that there is a considerable population of GeneChips measuring transcripts that did not terminate at this site, because there is correlated expression beyond the site.

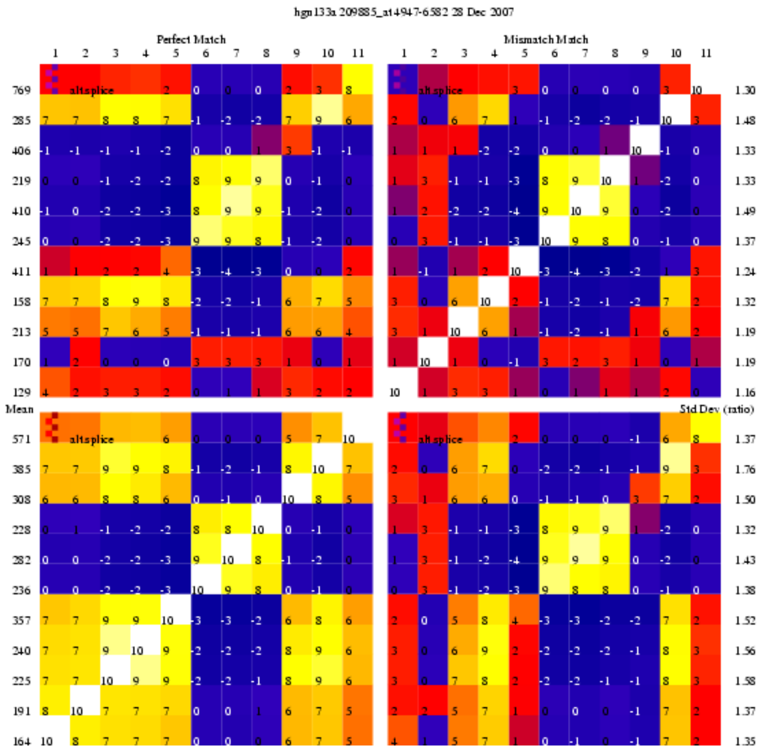

Figure 13: The correlation matrix for probeset 209885_at on the HG-U133A array. All the PM probes map uniquely to one exon, ENSE00001219272, and yet they show two different correlation blocks (PM 1,2,3,4,5,9,10 and 11 in one block, and PM 6,7 and 8 in the other block).

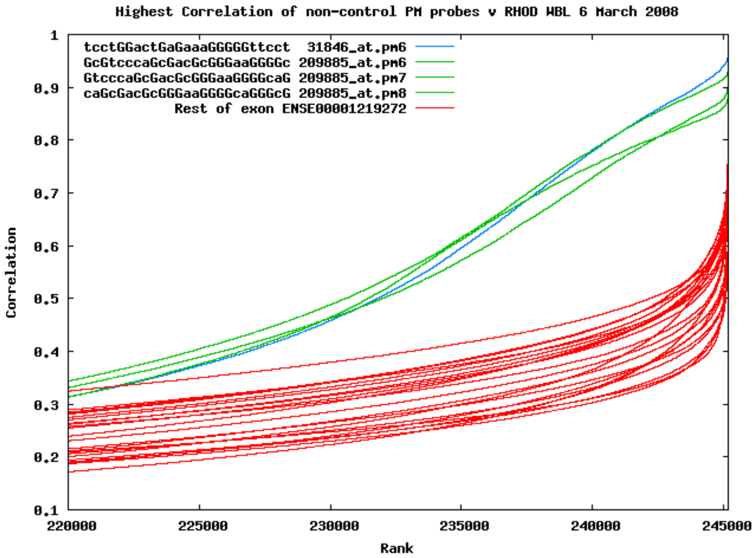

Figure 14: Rank of correlations between each of the probes mapping to exon ENSE00001219272 and the rest of the probes on the GeneChip HG-U133Av2. The several thousand probes that show close correlation have runs of contiguous guanines. We suggest this signal comes from probe-probe interactions resulting in G-quadruplexes.

level quantifiers make no distinction between independent and dependent probes. Moreover, we have found that relatively small overlaps in sequence can result in significant correlation due to cross-hybridization. It is crucial to separate the biological cause of the correlation from the biophysical cause of the correlation, and we are exploring methods to do this.

3 Summary

We have developed a computational pipeline to analyse tens of thousands of Affymetrix Human GeneChips, freely available from the Gene Expression Omnibus. The data samples many tissues and phenotypes. Because of this experimental diversity, any observed correlations between probe intensities can be associated either with biology that is robust, such as common co-expression, or with systematic biases associated with the GeneChip technology.

We generate matrices of correlations of the intensities of all the probes within each of the Affymetrix-defined probesets as well as for groups of probes which map uniquely to individual exons. The unique mappings reduce the effects of multiple targeting and cross-hybridization. The focus on exons also minimises the impact of alternative splicing which sometimes causes probes within a probeset to behave incoherently.

In the majority of cases, probesets do not measure one solid block of transcription. Instead there are numerous examples of outlier probes. In a number of probesets the mismatch probes indicate expression rather than providing a measure of background signal. Probes containing runs of four or more contiguous guanines are correlated with other similar probes and so do not measure gene expression. We suggest this systematic bias in GeneChip data results from probe-probe interactions. Furthermore, post-transcriptional processing events such as alternative polyadenylation leave a clear mark in a number of correlation matrices.

Our results have widespread implications because of the pervasive use of GeneChips in modern biological research.

Acknowledgments

OSG, WBL and MS are supported by a grant from the BBSRC (BB/E001742/1). JR is supported by a Strategic Studentship from the BBSRC (BBS/S/H/2005/11996A). JMAS is supported by a CONACYT scholarship (178596). We are grateful to Dr Renata da Silva Camargo who started the development of unique mappings to exons. We are also grateful to Giulietta Spudich and Dr Gautier Koscielny for help and advice about mappings.

References

[1] Dennise D. Dalma-Weiszhausz, Janet Warrington, Eugene Y. Tanimoto, and C. Garrett Miyada. The Affymetrix GeneChip platform: an overview. *Methods Enzymol.*, 410:3–28, 2006.

[2] T. Barrett, T. Suzek, D. Troup, Wilhite S., W.-C. Ngau, P. Ledoux, D. Rudnev, A. Lash, W. Fujibuchi, and R. Edgar. NCBI GEO: mining millions of expression profiles – database and tools. *Nucleic Acids Research*, 33(Database issue):D562–D566, 2005.

[3] Maria A. Stalteri and Andrew P. Harrison. Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips. *BMC Bioinformatics*, 8(8):13, 2007.

[4] Homin K. Lee, Amy K. Hsu, Jon Sajdak, Jie Qin, and Paul Pavlidis. Coexpression analysis of human genes across many microarray data sets. *Genome Research*, 14(6):1085–1094, 2004.

[5] Manhong Dai, Pinglang Wang, Andrew D. Boyd, Georgi Kostov, Brian Athey, Edward G. Jones, William E. Bunney, Richard M. Myers, Terry P. Speed, Huda Akil, Stanley J. Watson, and Fan Meng. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Research*, 33(20):e175, 2005.

[6] Jun Lu, Joseph C. Lee, Marc L. Salit, and Margaret C. Cam. Transcript-based redefinition of grouped oligonucleotide probe sets using AceView: High-resolution annotation for microarrays. *BMC Bioinformatics*, 8:108, 2007.

[7] Davide Rambaldi, Barbara Felice, Viviane Praz, Philip Bucher, Davide Cittaro, and Alessandro Guffanti. Splicy: a web-based tool for the prediction of possible alternative splicing events from Affymetrix probeset data. *BMC Bioinformatics*, 8(Suppl 1):S17, 2007.

[8] Michal J. Okoniewski and Crispin J. Miller. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics*, 7:276, 2006.

[9] T. J. P. Hubbard, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, J. Herrero, R. Holland, K. Howe, K. Howe, N. Johnson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, C. Melsopp, K. Megy, P. Meidl, B. Ouverdin, A. Parker, A. Prlic, S. Rice, D. Rios,

M. Schuster, I. Sealy, J. Severin, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, M. Wood, T. Cox, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, P. Flicek, A. Kasprzyk, G. Proctor, S. Searle, J. Smith, A. Ureta-Vidal, and E. Birney. Ensembl 2007. *Nucleic Acids Research*, 35(suppl_1):D610–617, 2007.

[10] Affymetrix. `www.affymetrix.com`.

[11] Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb Miller. A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, 7(1–2):203–214, 2000.

[12] Rafael A. Irizarry, Zhijin Wu, and Harris A. Jaffee. Comparison of Affymetrix GeneChip expression measures. *Bioinformatics*, 22(7):789–794, 2006.

[13] Andrew P. Harrison, Caroline E. Johnston, and Christine A. Orengo. Establishing a major cause of discrepancy in the calibration of Affymetrix GeneChips. *BMC Bioinformatics*, 8:195, 2007.

[14] Jose M. Arteaga-Salas, Harry Zuzan, William B. Langdon, Graham J. G. Upton, and Andrew P. Harrison. An overview of image-processing methods for Affymetrix GeneChips. *Briefings in Bioinformatics*, 9(1):25–33, 2008.

[15] W. B. Langdon, G. J. G. Upton, R. Camargo, and A. Harrison. A survey of spatial defects in Homo sapiens Affymetrix GeneChips. Transactions on Computational Biology and Bioinformatics (submitted), 2008.

[16] Graham J. G. Upton and Julie C. Lloyd. Oligonucleotide arrays: information from replication and spatial structure. *Bioinformatics*, 21(22):4162–4168, 2005.

[17] The alternative splicing and transcript diversity (ASTD) database. `http://www.ebi.ac.uk/astd`.

[18] Chunlei Wu, Haitao Zhao, Keith Baggerly, Roberto Carta, and Li Zhang. Short oligonucleotide probes containing G-stacks display abnormal binding affinity on Affymetrix microarrays. *Bioinformatics*, 23(19):2566–2572, 2007.

[19] G. J. G. Upton, W. B. Langdon, and A. Harrison. Incorrect measurement of gene expression by microarrays. Genome Biology (submitted), 2008.

[20] Sunita Kumari, Lalit K. Verma, and Jennifer W. Weller. AffyMAPSDetector: a software tool to characterize Affymetrix GeneChip expression arrays with respect to SNPs. *BMC Bioinformatics*, 8:276, 2007.

[21] R. Alberts, P. Terpstra, Y. Li, R. Breitling, J.-P. Nap, and R. Jansen. Sequence polymorphisms cause many false cis eQTLs. *PLoS ONE*, 7:e622, 2007.