Uncovering the expression patterns of chimeric transcripts using surveys of Affymetrix GeneChips

Joanna Rowsell[1*], Renata da Silva Camargo[1], William B. Langdon[1***] Maria A. Stalteri[1] and Andrew P. Harrison[1****]**

[1]Departments of Biological and Mathematical Sciences, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, UK

Summary

Background: A chimeric transcript is a single RNA sequence which results from the transcription of two adjacent genes. Recent studies estimate that at least 4% of tandem human gene pairs may form chimeric transcripts. Affymetrix GeneChip data are used to study the expression patterns of tens of thousands of genes and the probe sequences used in these microarrays can potentially map to exotic RNA sequences such as chimeras.
Results: We have studied human chimeras and investigated their expression patterns using large surveys of Affymetrix microarray data obtained from the Gene Expression Omnibus. We show that for six probe sets, a unique probe mapping to a transcript produced by one of the adjacent genes can be used to identify the expression patterns of readthrough transcripts. Furthermore, unique probes mapping to an intergenic exon present only in the MASK-BP3 chimera can be used directly to study the expression levels of this transcript.
Conclusions: We have attempted to implement a new method for identifying tandem chimerism. In this analysis unambiguous probes are needed to measure run-off transcription and probes that map to intergenic exons are particularly valuable for identifying the expression of chimeras.

1 Background

The complexity of RNA and its modification during and post transcription have been extensively studied. Alternative splicing [1] and alternative polyadenylation [2] are two key processes which can greatly affect RNA, modifying the coding sequence and untranslated regions of a transcript. With single genes producing variant transcripts, the huge discrepancy between the approximately 25,000 human protein-coding genes and the approximate estimate of 84,000 proteins could be explained.

In addition to transcript variants, new types of RNA sequences have been discovered. For instance, instead of the normal independent transcription of a gene, a single RNA sequence can be formed from two adjacent genes Figure 1. The resulting fused transcripts are known as transcription-induced chimeras (TICs) [3]. Typically exons from both genes are present in the chimeric RNA sequence with the intergenic region removed during splicing [4]. Some chimeric transcripts are translated into bifunctional proteins with properties from the proteins

[*]Currently at the Institute of Genetics, University of Nottingham
[**]Currently at Lonza Biologics plc, Cambridge, UK
[***]Currently at the Department of Computer Science, King's College London
[****]To whom correspondence should be addressed. E-mail: harry@essex.ac.uk

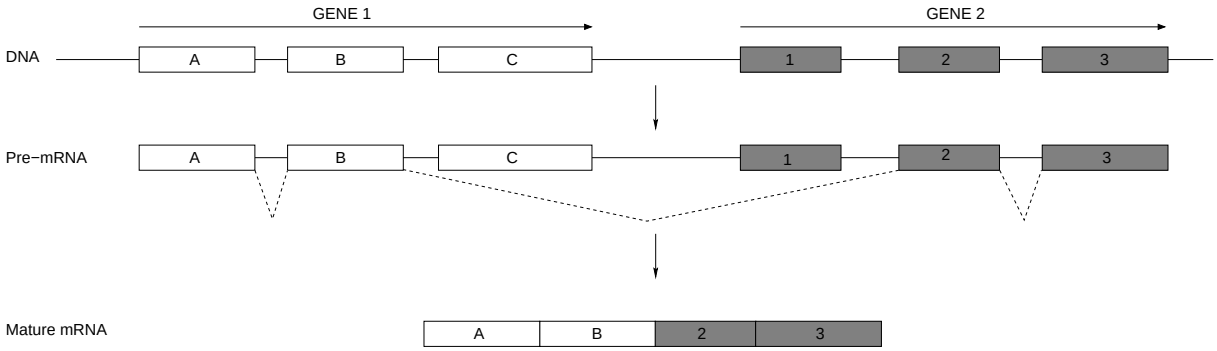


Figure 1: An illustration of transcription-induced chimerism adapted from [4]. One transcript is produced from two adjacent genes and the intergenic region is spliced out. The resulting mRNA sequence is known as a chimeric transcript or transcription-induced chimera. The boxes are exons and the dotted lines show where the pre-mRNA is spliced. The arrows above the genes indicate the direction of transcription.

of both original genes (e.g. [5, 6]). In the last decade, a few cases of readthrough transcription have been reported (e.g. [7]). Recently different techniques have been implemented to estimate the number of adjacent human gene pairs that are transcribed into a single RNA strand [3, 4]. For example, analysis in the ENCODE regions which form 1% of the human genome suggests that at least 4% of tandem gene pairs can be involved in readthrough transcription [3].

Here we investigate whether chimeric transcripts can be detected using surveys of Affymetrix GeneChips which contain hundreds of thousands of probes for measuring the expression patterns of tens of thousands of genes. The Affymetrix arrays contain millions of 25-base single DNA strands known as probes which measure the expression of genes. For every PM probe which is perfectly complementary to the RNA sequence, there is a mismatch (MM) probe which is exactly the same as the PM probe except that the middle base is different. Expressed sequence tags (ESTs) were used in the design of the Affymetrix arrays [8] and due to the extensive use of such sequences it was predicted that some probes may be mapping to exotic RNA sequences such as chimeric transcripts [9].

It is often a condition of publication that GeneChip datasets are made publicly available and therefore large amounts of GeneChip data are deposited into databases such as the NCBI Gene Expression Omnibus (GEO) database [10]. As a result there are many freely available Affymetrix data files that can be used collectively as a large data set enabling many different conditions to be studied. This wealth of data provides a potentially cost-effective way of analysing chimeric transcripts. In this paper we report our exploration of GEO as a powerful resource for identifying the expression of chimeric transcripts.

2 Methods

2.1 Data

Probe sequences for 15 Affymetrix human arrays (Table 1) were downloaded from the support section of the Affymetrix website in September 2007 (http://www.affymetrix.com/index.affx). The order of the probes in each of the probe sets is the one defined in the NetAffx section of the Affymetrix website and is based on transcript positioning. The NCBI reference sequences

Human Genome Array	Mean intensity	Human Genome Array	Mean intensity
HC_G110	-	HG_U133B	-
HG_Focus	-	HG_U133A_2	187
HT_HG_U133A	-	HG_U95A	720
HT_HG_U133B	-	HG_U95Av2	339
HT_HG_U133_Plus_B	-	HG_U95B	111
HG_U133_Plus_2	135	HG_U95C	-
HG_U133A	251	HG_U95D	-
		HG_U95E	132

Table 1: Affymetrix Human GeneChips used in the probe mapping. The mean probe intensity of all probes over all cel files is shown for the arrays with unique probe mappings.

(RefSeq) for the individual gene transcripts and chimeric transcripts were obtained from the NCBI Entrez CoreNucleotide database in December 2007 (http://www.ncbi.nlm.nih.gov/sites/entrez?db=nuccore). Ensembl exon sequences were downloaded in January 2008 (Release 48; http://www.ensembl.org/index.html). The cel files were downloaded from the Gene Expression Omnibus (GEO) in February 2007 the only exception being GSE5949 which was downloaded in November 2007 (http://www.ncbi.nlm.nih.gov/geo/).

2.2 Alignments, normalization and background elimination

The EMBOSS Smith-Waterman tool [11] was used for sequence alignments. Standalone Blast version 2.2.17 for Linux (ia32) was obtained from the NCBI in September 2007 (http://www.ncbi.nlm.nih.gov/BLAST/download.shtml). Probe sequences were aligned against the RefSeq transcripts and exon sequences using MegaBLAST [12] with word size and minimal hit score of 25.

The probe intensities were normalized to a log scale against a reference chip for each chip design (an average across all of GEO, February 2007 and including GSE5949) [13]. To identify possible spatial defects the normalized data were studied and suspiciously similar areas on the chip were highlighted [13].

To reduce the impact of background noise on the true probe intensities we attempted to eliminate the background intensities by subtracting the logarithm (base 2) of the mismatch intensities from the perfect match intensities.

3 Results

3.1 Probes mapping to chimeric transcripts

The literature was searched to find human chimeric transcripts caused by readthrough transcription. Cases where a mRNA NCBI reference sequence (RefSeq) could not be found for the chimeric transcripts were excluded from the analysis. The probe sequences of fifteen Affymetrix GeneChips (Table 1) were aligned against the reference sequence transcripts (Supplementary Table 1) of both genes (and variants) and the chimeric RefSeq transcript (and variants) using MegaBLAST [12]. The results of the probe mappings are described in the sections

below. Although many probes map to each of the RefSeq transcripts only the unique probe mappings are discussed since these are the most useful for uncovering expression patterns. Here a unique mapping indicates a probe which aligns to only one of the transcripts in the set (both individual gene transcripts, chimeric transcript and any variants). The individual gene transcript refers to the independent transcript of either the upstream or downstream gene when the adjacent genes are not involved in transcription-induced chimerism. The only cases with unique probe mappings were MASK-BP3 [7], HCC-2/HCC-1 [14] and SSF1-P2Y$_{11}$ [15] (Supplementary Table 1).

For the three cases with a unique probe mapping, graphs were plotted for the corresponding probe set. The probe intensities were normalized, possible spatial flaws were highlighted where appropriate and any outlying cel files were labelled. Background correction was applied by subtracting the $\log_2$ mismatch (MM) intensities from the perfect match (PM) intensities. The mean intensity of each probe over all cel files for one array type was calculated for the arrays with unique probe mappings. The mean intensity of all the mean probe intensity values was calculated to give a single value for the arrays with unique mappings (Table 1). This table provides a comparison for the probe intensity values shown in the graphs. When a unique mapping was found the ADAPT database [16] was searched to confirm that the probes do not map to any RefSeq transcript unrelated to the genes in the case. Many of the schematic images of the transcripts and probe mappings used were taken from the online ADAPT tool (http://bioinformatics.picr.man.ac.uk/adapt/ProbeToView.adapt).

3.2 Analysis of probes mapping to chimera HCC-2/HCC-1, Probe set 33789_at intensities

All sixteen perfect match (PM) probes of set 33789_at (HG_U95A, HG_U95Av2) map to the HCC-2 RefSeq NM_032965.2; probes 1-8 map also to the chimeric transcript variant NM_032964.2 and chimeric transcript variant NM_004167.3 (Figures 2B and Supplementary Figure 1B). If HCC-2 is not expressed, but at least one of the chimeric variants (NM_032964.2 and NM_004167.3) are, then probes 1-8 should be high but probes 9-16 should be low. If all three transcripts are expressed then probes 1-8 should have higher values than probes 9-16.

The PM probe intensities for probes 6-8 of set 33789_at are mainly high (top 28%) for the HG_U95A array (Figure 2A). Probes 9-16 are lower and 1-5 lower still. It is the same for the HG_U95Av2 array where the intensity of probes 6-8 are mainly high (top 12%) for the HG_U95Av2 array and the other probe intensities are lower (Supplementary Figure 1A). Probe 7 has the highest $\log_2$ intensity value after background elimination (Figures 2C and Supplementary Figure 1C). However, the $\log_2$ intensities of many of the probes after background elimination are around zero suggesting that none of the three transcripts are expressed (NM_032965.2, NM_032964.2, NM_004167.3) for either the HG_U95A array or the HG_U95Av2 array. It is difficult to determine why probe 7 is higher than the others. The Ensembl database was consulted and probe 7 does not map to an exon boundary. This probe does not contain the motif `CCTCC` which has been found to cause probes to behave as outliers [17, 18]. The probe also does not contain a run of at least four `G`s which we have found can cause probes to behave as outliers within a probe set [18, 19].

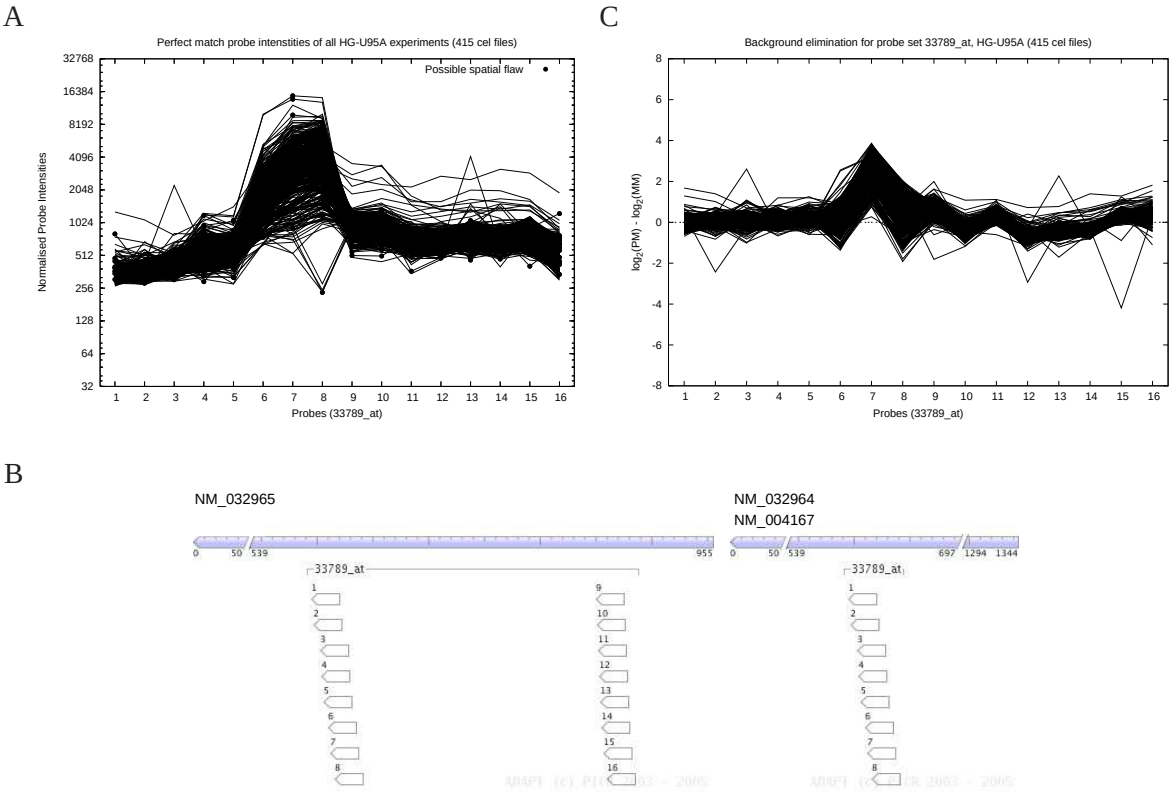


Figure 2: Probe intensities over all cel files for probe set 33789_at on the HG_U95A array (A) Perfect match intensities (log scale) (B) ADAPT probe positions on the transcripts (C) Background elimination ($\log_2$(PM) - $\log_2$(MM)).

3.3 Analysis of probes mapping to chimera SSF1-P2Y$_{11}$

3.3.1 Probe set 46597_at intensities

Probes 5-13 from set 46597_at (HG_U95B) map uniquely to the SSF1 RefSeq (NM_020230.4) and probes 1-4 of the same set map to both the SSF1 RefSeq and the chimeric transcript NM_001040664.1 (Supplementary Figure 2B). If SSF1 is not expressed but the chimera is then probes 1-4 should be high. Otherwise if both transcripts are expressed then probes 1-4 should be higher than probes 5-13. As can be seen from Supplementary Figure 2A, probes 1 and 11 have the highest intensities (top 3% for the HG_U95B array). After background elimination probes 1, 11 and 16 have the highest $\log_2$ values (Supplementary Figure 2C). However, many of the probes in the set have background-adjusted $\log_2$ PM probe intensities around zero, indicating that neither the chimera SSF1-P2Y$_{11}$ nor SSF1 are expressed.

It is interesting that probes 1, 11 and 16 have higher $\log_2$ intensities than the other probes after background elimination. Probe 1 maps to four exon boundaries from different transcripts for the same Ensembl gene. Probes 2-4 map with 25 bases to the exons that probe 1 maps to with 18 bases. So if the second exons in the pairs were spliced out then the intensities of probes 2-4 should be low but probe 1 would only map with 7 bases so would not be expected to be high. Probe 11 maps to two exons with 25 bases (ENSE00001516504 and ENSE00001408392) however these two exons are in the group of exons that probes 2-4 map to. Probe 16 maps to 18 exons with between 16 and 18 bases. These exons are different to those which probes 1 and 11 map to. Probe 16 may have higher background-adjusted $\log_2$ intensities because it maps to

many exons although the number of contiguous bases is between 16 and 18 bases.

3.3.2 Probe set 214546_s_at intensities

All eleven probes of set 214546_s_at (HG_U133A, HG_U133A_2, HG_U133_Plus_2) map to the $P2Y_{11}$ RefSeq (NM_002566.4), probes 5-11 map uniquely and probes 1-4 also map to the chimeric transcript NM_001040664.1 (Supplementary Figures 3B, 4B and 5B) . The plots in Supplementary Figures 3, 4 and 5 show that for the HG_U133A and especially for the HG_U133A_2 and HG_U133_Plus_2 arrays, the background-adjusted $\log_2$ PM probe intensities of set 214546_s_at are around or below zero, suggesting that neither transcript is expressed.

3.3.3 Probe set 33633_at intensities

Probes 1-8 of set 33633_at (HG_U95A, HG_U95Av2) map to both the $P2Y_{11}$ RefSeq (NM_002566.4) and the chimeric RefSeq (NM_001040664.1). The remaining eight probes from this set (probes 9-16) map uniquely to the $P2Y_{11}$ RefSeq (Supplementary Figures 6B and 7B). If the chimeric transcript is expressed but $P2Y_{11}$ is not then probes 1-8 should be high. If both transcripts are expressed then probes 1-8 should be higher than probes 9-16. The third quartile of the PM intensities for the HG_U95A and HG_U95Av2 arrays are 1132 and 530 respectively. Supplementary Figures 6A and 7A show that probes 1, 2, 4, 8, 9, 13, 14 and 15 have PM intensities higher than the third quartile values on the respective HG_U95A and HG_U95Av2 arrays. The high PM intensities could suggest that both transcripts ($P2Y_{11}$ and the chimeric) or just the $P2Y_{11}$ transcript are expressed.

Probes 1, 2, 7, 8 and 15 have the highest background-adjusted $\log_2$ PM probe intensities on the HG_U95A array (Supplementary Figure 6C) with probes 1, 2, 3, 4, 7, 8, 14 and 15 having the highest background-adjusted $\log_2$ values on the HG_U95Av2 array (Supplementary Figure 7C). There is no clear difference between the two sets of probes: 1-8 and 9-16. However, it is interesting to highlight the difference in intensities especially between probe 9 and probes 10-11 (Supplementary Figures 6A and 7A) because probes 9-11 map perfectly to the same exon and map to no other exon with 16 or more bases. Probe 9 contains the motif `GCCTCC` which has been found to cause probes to behave as outliers [17, 18]. This motif is complementary to the nucleotide sequence flanking the T7 primer which is used in RNA amplification in the Affymetrix protocol. The presence of `GCCTCC` in the probe sequence could explain the difference in probe intensities between probe 9 and probes 10-11.

3.4 Analysis of probes mapping to chimera MASK-BP3

3.4.1 Probe set 67520_at intensities

The chimeric mRNA MASK-BP3 comprises of MASK exon 33, the penultimate MASK exon, spliced to exon 0, a novel exon located between MASK and 4E-BP3 [7]. The intermediate exon 0 which is known as exon 34 in the chimeric transcript is only expressed in the chimeric transcript where it is spliced to exon B, the second exon of 4E-BP3. The sequences for MASK exon 33 and 4E-BP3 exon B were obtained from the Consensus CDS database

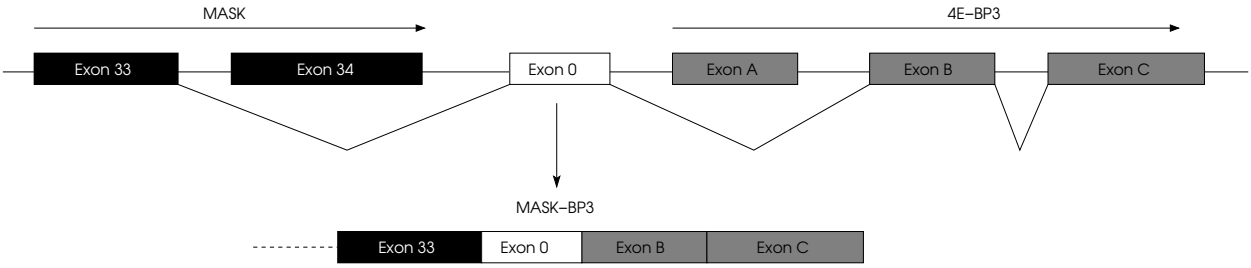


Figure 3: Schematic diagram of chimeric transcript MASK-BP3. Boxes represent exons, which are identified with numbers for MASK and letters for 4E-BP3.

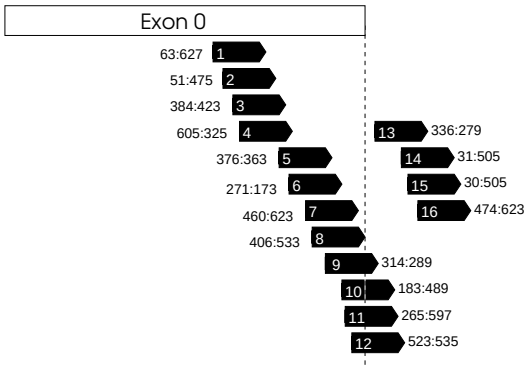


Figure 4: Schematic diagram showing probe positions of set 67520_at in relation to exon 0 of MASK-BP3.

(http://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi) and aligned to the RefSeq chimeric transcript NM_020690.4. The sequence for exon 0 was obtained from the paper by Poulin *et al.* [7] and aligned to the chimeric MASK-BP3. The alignments confirmed the structure illustrated in the literature [7] with MASK exon 33 spliced to intermediate exon 0 which is spliced at the 3′ end to the second exon of 4E-BP3 (Figure 3). All three exons sequences were also confirmed using the Ensembl database (version 48).

Eight probes from set 67520_at on chip HG_U95E map uniquely to chimera NM_020690.4 (63:627, 51:475, 384:423, 605:325, 376:363, 271:173, 460:623, 406:533). These probes map with 100% coverage and 100% sequence identity to exon 0 of the chimeric transcript. Probes 9-12 of set 67520_at also map to exon 0 with 23, 15, 14 and 12 contiguous bases respectively (Figure 4). To determine whether probes 1-8 of set 67520_at (Figure 4) map to any sequence other than MASK-BP3 (NM_020690.4), an NCBI BLASTN 2.2.17 [20] search was run. The only transcript that has 100% sequence identity and 100% coverage to each of the eight probe sequences is the chimera NM_020690.4. All of the eight probes also have 100% sequence identity and 100% coverage to genomic sequences, for example probe 1 aligns to NT_008583.16 (chromosome ten, genomic reference assembly) which is on a different chromosome to the chimera (chromosome five). Five of the probes also align to other transcripts but all with 72% or less coverage (e.g. 18 or less bases of 25 map) moreover some of these probes have less than 100% maximum sequence identity (e.g. do not map contiguously).

If the chimeric transcript is expressed, it would be expected that probes 1 to 8 would have high intensities, probes 9 to 12 decreasing intensities and probes 13 to 16 low intensities. Probes 6 to 13 display this expectation (Figure 5A), however the mean values of probes 2-4 are lower than expected (bottom 31% for the HG_U95E array) and probes 14-16 are higher than expected (top 9% for the HG_U95E array). The mean values of probes 1 and 5 look quite low but are near the

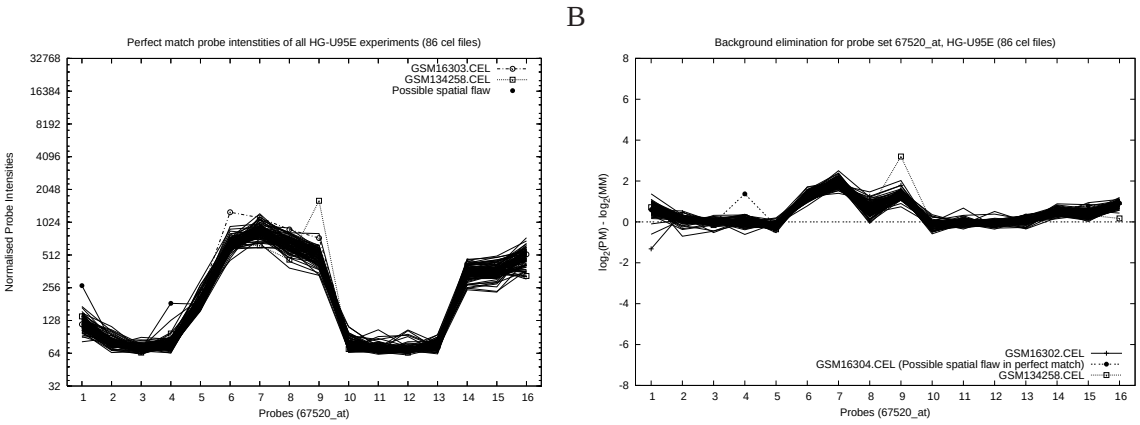


Figure 5: Probe intensities over all cel files for probe set 67520_at on the HG_U95E array (A) Perfect match intensities (log scale) (B) Background elimination (log_2(PM) - log_2(MM))

average for the HG_U95E array (132, Table 1). An investigation into the high intensity values for probes 14-16 was carried out and probes 14-15 align with 15 contiguous bases to another transcript which appears to be expressed (Supplementary document 1).

After background elimination for the probe intensities of set 67520_at (Figure 5B) it can be seen that the peaks occur at probes 6, 7 and 9. It is surprising that probes 7 and 8 display a difference in intensity levels because they only differ in sequence by one base. The only transcript that has 100% sequence identity and 100% coverage to each of the probes (1-8) is NM_020690.4 (MASK-BP3). Results from the BLASTN search show that probes 7 (460:623) and 8 (406:533) map to the same five transcripts (excluding MASK-BP3) with 15 contiguous bases (NM_024652.3, NM_021536.1, NM_021537.1, NM_000906.2 and NM_001001561.1). Therefore it is unlikely that the peak at probe 7 (Figure 5B) is due to cross-hybridisation. Another explanation could be the use of an alternative splice site where the exon is spliced internally such that the part which has alignment to probes 6 and 7 is retained in the mRNA.

In summary, the background-adjusted log_2 PM probe intensities for set 67520_at do not show the pattern that was expected if the chimeric transcript was expressed.

3.4.2 Probe set 208773_s_at intensities

Of the eleven probes in 208773_s_at (HG_U133A, HG_U133A_2, HG_U133_Plus_2), probes 4-11 map uniquely to MASK transcript variant 1, NM_017747.1 and probes 1-3 map to both NM_017747.1 and the chimera NM_020690.4 (Supplementary Figures 8B, 9B and 10B). The PM intensities of most of the probes are above average for the HG_U133A array (mean = 251, Table 1), the HG_U133_Plus_2 array (mean = 135, Table 1) and the HG_U133A_2 array (mean = 187, Table 1) (Supplementary Figures 8A, 9A and 10A). The background-adjusted log_2 PM probe intensities are greater than 0 for most of the probes in all three arrays indicating that NM_017747.1 is expressed on the HG_U133A and HG_U133_Plus_2 arrays. There is no clear difference between probes 1-3 and 4-11 so it is unlikely that the chimeric transcript is expressed (Supplementary Figures 8C, 9C and 10C). Probes 4 and 9 from set 208773_s_at map only to NM_017747.1 so therefore it is surprising that if NM_017747.1 is not expressed, these probes have background-adjusted log_2 PM probe intensities greater than zero (Supplementary Figures 8C, 9C and 10C). It is also surprising that log_2 intensities after background elimination of

probes 6, 9 and 11 are lower than 5, 7, 8 and 10 (Supplementary Figures 8C and 9C) because probes 5-11 map perfectly to only one exon (ENSE0000146780).

4 Discussion

4.1 Intergenic exons

The analysis indicates that it is easier to uncover the expression patterns of chimeric transcripts when there are novel intergenic exons present. Eight probes from set 67520_at map uniquely to the intergenic exon only present in the MASK-BP3 chimera (NM_020690.4) and therefore any changes in the intensities will reflect the expression of this chimeric transcript. In one study the splicing patterns of chimeric transcripts were observed and it was found that 12% of the predicted chimeras in their data set contain a novel exon which is positioned between the two fused genes [4]. A higher percentage of 21% was found for EST-supported TIC events containing intergenic exons [3]. For these cases at least, any probes mapping to such intergenic exons will be unique in the sense that they will not align to any of the transcripts produced by the genes independently because the intergenic exon would not be transcribed. Of all the chimeric transcripts found in the literature, only two contained intergenic exons; MASK-BP3 [7] and TSNAX-DISC1 [21]. Unfortunately, no Affymetrix probes map to the intergenic exons of the TSNAX-DISC1 chimera.

4.2 Probes aligning to the boundary of the fused genes

It was hoped that some probes mapped to the boundary of the fused exons of the two genes (e.g. exons B and 2 in Figure 1). In fact, if probes did map in this region it might be difficult to distinguish whether a *trans*-splicing event or chimeric transcript had been detected. Some authors (e.g. [3]) have argued against the hypothesis that chimeras are generated by trans-splicing. Of the few trans-splicing events that have been reported, splicing can occur between transcripts of the same gene [22]. Also, chimeras with intergenic exons are unlikely to be caused by trans-splicing because with this mechanism, the intergenic region is not transcribed [4, 3]. In any case, the probes that were studied in this paper do not map to the gene exon junctions and this is due to the design of the Affymetrix probes. The probes in the 3′ gene expression arrays are generally selected from a region 600 bases upstream of the polyadenylation (polyA) site [23]. Reverse transcription is used to label the mRNA in the Affymetrix protocol and an oligo(dT) primer complementary to the polyA tail is used in this process [24]. The reverse transcriptase copies the mRNA from the polyA tail so the farther a section of sequence is from the polyA tail, the lower the chance that it will be transcribed by the enzyme [25]. It is unlikely for exon B and especially exon 2 in Figure 1 to have probes mapping to them because they are far from the 3′ end of their respective genes (gene 1 for exon B and gene 2 for exon 2).

Probes mapping to the 3′ end of the chimeric transcript are likely to map to the independent transcript produced from the downstream gene (Figure 6). The most abundant splicing pattern for the chimeras in the [4] study was between the penultimate exon of the upstream gene and the second exon of the downstream gene. As can be seen from Figure 6, probes mapping with 3′ bias to the upstream gene transcript will not map to a chimera where the final exon of gene

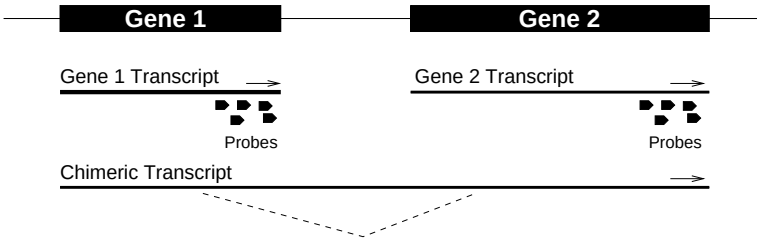


Figure 6: Hypothetical probes mapping in the 3′ region of the transcripts. The arrows show the direction of transcription. The dotted line shows where the chimeric might be spliced from the penultimate exon of Gene 1 and second exon of Gene2.

1 has been spliced out. The probes that map to the 3′ end of the upstream gene transcript are good candidates for measuring the expression of the upstream gene alone.

4.3 Probe sets with unique probes

The most interesting cases after those with intergenic exons are ones where part of a probe set map to either the upstream or downstream transcripts (e.g. HCC-2/HCC-1, SSF1-P2Y$_{11}$ and MASK-BP3). The remaining probes map to the chimeric (and the upstream transcript) and not to the downstream transcript or any other transcript variants. For the cases with a unique part probe set, either (1) many of the background-adjusted $\log_2$ PM probe intensities are around zero (e.g. SSF1-P2Y$_{11}$: 46597_at, Supplementary Figure 2) suggesting the chimera is not expressed in these tissue samples or (2) most of the background-adjusted $\log_2$ PM probe intensities are greater than zero (e.g. MASK-BP3: 208773_s_at, Supplementary Figures 8, 9 and 10) in which case it is difficult to determine if the chimera is expressed or not. For example, the most conclusive results for a case like SSF1-P2Y$_{11}$, 46597_at would be where the background-adjusted $\log_2$ intensities are greater than zero for probes 1-4 and around zero for probes 5-13 suggesting that the chimera is expressed.

4.4 Cross-hybridisation

Alongside the issue of probes mapping to other transcripts or transcript variants of the genes involved in the chimera, probes can potentially cross-hybridise with unrelated gene transcripts. For example, for three probes in the 67520_at set, two transcripts (unrelated to the genes forming the chimeric transcript) aligned with 15/16 contiguous bases to the probes. If these transcripts (NM_015680.3 and NM_020184.3) hybridise to the probes, they may mislead the interpretation of the probe intensities for the target sequence. For probeset 208773_s_at, there are two probes (534:167 and 190:13) which align to other transcripts besides MASK and MASK-BP3 with 16 contiguous bases. Also, all probes in set 67520_at and 208773_s_at have 100% coverage and 100% alignment with genomic sequences such as NT_008583.16 so if for some reason these sequences are being transcribed and polyadenylated, they may hybridise to the probes.

4.5 Tissue samples

The tissue samples that the GeneChips are run on need to be taken into consideration when observing the expression patterns of chimeric transcripts. For example, probe set 67520_at can be used to measure the expression patterns of the MASK-BP3 chimera. Poulin *et al.* [7] analysed the expression of the 4E-BP3 and MASK-BP3 transcripts and found that they are both expressed in skeletal muscle tissues alongside other tissues. It would therefore be expected that the probes in set 67520_at that map to MASK-BP3 would display expression. However only 14 of the GEO cel files of 86 sampled normal skeletal muscle tissue.

5 Conclusions

The results here demonstrate that it is possible to use microarray data to examine the expression patterns of chimeric transcripts. However, it is often difficult to distinguish between the expression of an individual transcript and the chimeric transcript. The potential cross-hybridisation of some probes also makes the analysis difficult but cases especially where intergenic exons are present provide examples of how microarray data can be used as a method for identifying the expression of exotic transcripts. The analysis could be extended to exon arrays, thus removing the 3′ bias in the gene expression arrays. Further analysis could also be undertaken to find novel chimeric transcripts using microarray data. The results here suggest that 3′ array data may not be ideal for this work and exon array data could prove to be a better data source.

Another possible extension of the work presented here is the analysis of deep sequencing data, in particular RNA-Seq data, which can be used to detect and measure RNA expression levels. There would be a number of advantages for using this type of data, for example the interrogation of transcripts would not be restricted to only those that the microarray probes can detect [26]. Also, the analysis would not be subjected to 3′ bias in the same way as the 3′ array data. It has have already been suggested by Marioni and colleagues [26] that deep sequencing could be used to study regions between annotated genes and this would be useful for identifying chimeric transcripts with intergenic exons.

Acknowledgements

We wish to thank G. Upton for support with statistical analysis and T. Earl, R. Cummings and A. Owen for computing assistance. RdSC, WBL and MS were funded by the BBSRC grant BB/E001742/1. JR was funded by the BBSRC Strategic Studentship BBS/S/H/2005/11996A.

References

[1] B. Modrek and C. Lee. A genomic view of alternative splicing. *Nature Genetics*, 30(1):13–19, 2002.

[2] B. Tian, J. Hu, H. Zhang, and C. S. Lutz. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Research*, 33(1):201–212, 2005.

[3] G. Parra, A. Reymond, N. Dabbouseh, E. T. Dermitzakis, R. Castelo, T. M. Thomson, S. E. Antonarakis, and R. Guigó. Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Research*, 16(1):37–44, 2006.

[4] P. Akiva, A. Toporik, S. Edelheit, Y. Peretz, A. Diber, R. Shemesh, A. Novik, and R. Sorek. Transcription-mediated gene fusion in the human genome. *Genome Research*, 16(1):30–36, 2006.

[5] T. M. Thomson, J. J. Lozano, N. Loukili, R. Carrió, F. Serras, B. Cormand, M. Valeri, V. M. Díaz, J. Abril, M. Burset, J. Merino, A. Macaya, M. Corominas, and R. Guigó. Fusion of the Human Gene for the Polyubiquitination Coeffector UEV1 with Kua, a Newly Identified Gene. *Genome Research*, 10:1743–1756, 2000.

[6] B. Pradet-Balade, J. P. Medema, M. López-Fraga, J. C. Lozano, G. M. Kolfschoten, A. Picard, C. Martínez-A, J. A. Garcia-Sanz, and M. Hahne. An endogenous hybrid mRNA encodes TWE-PRIL, a functional cell surface TWEAK- APRIL fusion protein. *The EMBO Journal*, 21:5711–5720, 2002.

[7] F. Poulin, A. Brueschke, and N. Sonenberg. Gene fusion and overlapping reading frames in the mammalian genes for 4E-BP3 and MASK. *Journal of Biological Chemistry*, 278(52):52290–52297, 2003.

[8] Affymetrix. Design and Performance of the GeneChip® Human Genome U133 Plus 2.0 and Human Genome U133A 2.0 Arrays. Technical Note. Part No. 701483 Rev.2. 2003.

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

[10] T. Barrett, T. O. Suzek, D. B. Troup, S. E. Wilhite, W. C. Ngau, P. Ledoux, D. Rudnev, A. E. Lash, W. Fujibuchi, and R. Edgar. NCBI GEO: mining millions of expression profiles-database and tools. *Nucleic Acids Res*, 33(suppl 1):D562–566, 2005.

[11] P. Rice, I. Longden, and A. Bleasby. Emboss: The european molecular biology open software suite. *Trends in Genetics*, 16(6):276–277, 2000.

[12] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller. A Greedy Algorithm for Aligning DNA Sequences. *Journal of Computational Biology*, 7(1-2):203–214, 2000.

[13] W. B. Langdon, G. J. G. Upton, R. da Silva Camargo, and A. P. Harrison. A Survey of Spatial Defects in Homo Sapiens Affymetrix GeneChips. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2008.

[14] A. Pardigol, U. Forssmann, H. D. Zucht, P. Loetscher, P. Schulz-Knappe, M. Baggiolini, W. G. Forssmann, and H. Mägert. HCC-2, a human chemokine: gene structure, expression pattern, and biological activity. *Proc Natl Acad Sci USA*, 95(11):6308–6313, 1998.

[15] D. Communi, N. Suarez-Huerta, D. Dussossoy, P. Savi, and J. M. Boeynaems. Cotranscription and Intergenic Splicing of Human $P2Y_{11}$ and SSF1 Genes. *Journal of Biological Chemistry*, 276(19):16561–16566, 2001.

[16] H. S. Leong, T. Yates, C. Wilson, and C. J. Miller. ADAPT: a database of Affymetrix probesets and transcripts. *Bioinformatics*, 21(10):2552–2553, 2005.

[17] R. M. Kerkhoven, D. Sie, M. Nieuwland, M. Heimerikx, J. De Ronde, W. Brugman, and A. Velds. The T7-primer is a source of experimental bias and introduces variability between microarray platforms. *PLoS One*, 3(4)(4), 2008.

[18] G. J. G. Upton, O. Sanchez-Graillet, J. Rowsell, J. M. Arteaga-Salas, N. S. Graham, M. A. Stalteri, F. N. Memon, S. T. May, and A. P. Harrison. On the causes of outliers in Affymetrix GeneChip data. *Briefings in Functional Genomics and Proteomics*, 8(1):199 – 212, 2009.

[19] G. J. G. Upton, W. B. Langdon, and A. P. Harrison. G-spots cause incorrect expression measurement in Affymetrix microarrays. *BMC Genomics*, 9(1):613, 2008.

[20] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[21] J. K. Millar, S. Christie, C. A. M. Semple, and D. J. Porteous. Chromosomal Location and Genomic Structure of the Human Translin-Associated Factor X Gene (TRAX; TSNAX) Revealed by Intergenic Splicing to DISC1, a Gene Disrupted by a Translocation Segregating with Schizophrenia. *Genomics*, 67(1):69–77, 2000.

[22] T. Takahara, B. Tasic, T. Maniatis, H. Akanuma, and S. Yanagisawa. Delay in Synthesis of the 3′ Splice Site Promotes *trans*-Splicing of the Preceding 5′ Splice Site. *Molecular Cell*, 18(2):245–251, 2005.

[23] Affymetrix. Array Design for the GeneChip® Human Genome U133 Set. Technical Note. Part Number 701133 Rev.2., 2007.

[24] Affymetrix. GeneChip® Expression Analysis Technical Manual. Part Number 702232 Rev. 2. 2006.

[25] K. J. Archer, C. I. Dumur, S. E. Joel, and V. Ramakrishnan. Assessing quality of hybridized RNA in Affymetrix GeneChip experiments using mixed-effects models. *Biostatistics*, 7(2)(2):198–212, 2006.

[26] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517, 2008.

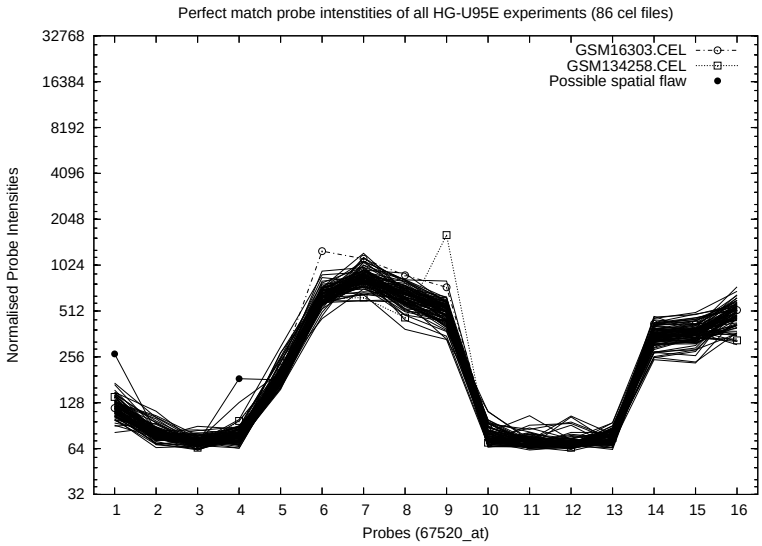
Supplementary document 1

An investigation into the high intensity values for probes 14-16, set 67520_at, chimera MASK-BP3

A NCBI BLASTN search was run on probes 14, 15 and 16. Probes 14-16 map to transcripts unrelated to MASK-BP3; probes 14 and 15 map with with 60% coverage (100% sequence identity) to transcript NM_015680.3 (15 bases of 25 map contiguously) and probe 16 maps to transcript NM_020184.3 with 64% coverage and 100% sequence identity (16 bases of 25 map contiguously). It might not be expected for probes 14-16 to display signal even if the transcripts (NM_015680.3 and NM_020184.3) are expressed because not all bases align. However, the bases are contiguous so it is possible that NM_015680.3 may hybridise to probes 14 and 15 and NM_020184.3 to probe 16.

The probe sets and arrays that map to the RefSeq transcripts NM_015680.3 and NM_020184.3 are shown in Table 1. Probe sets 207511_s_at, 200070_at and 218900_at were not studied as the GEO experiments that were run on array HG_U95E (GSE1007, GSE465 and GSE5949) were not run on the U133 arrays. Cel files GSM15927, GSM16222, GSM4382 and GSM4384 were also removed from the analysis because these GEO samples were not run on the HG_U95E array. It seems that transcript NM_020184.3 is not expressed for the samples on HG_U95E since the background-adjusted $\log_2$ PM probe intensities for sets 46564_at (Figure 2), 62274_at (Figure 3) and 63395_at (Figure 4) are all around or below zero. This therefore does not explain why probe 16 in Figure 1A is higher than expected. Transcript NM_015680.3 may be expressed for the samples on HG_U95E because the PM intensities for probes 3-14 on set 34864_at are high (top 28% for the HG_U95A array, top 12% for the HG_U95Av2 array: Figures 5A and 6A). The background-adjusted $\log_2$ PM probe intensities are greater than 0 for most of the probes in set 34864_at on the HG_U95A and HG_U95Av2 arrays (Figures 5C and 6C). The expression of set 34864_at could account for the PM intensities of probes 14 and 15 of the 67520_at set being higher than expected (Figure 1A).

A



B

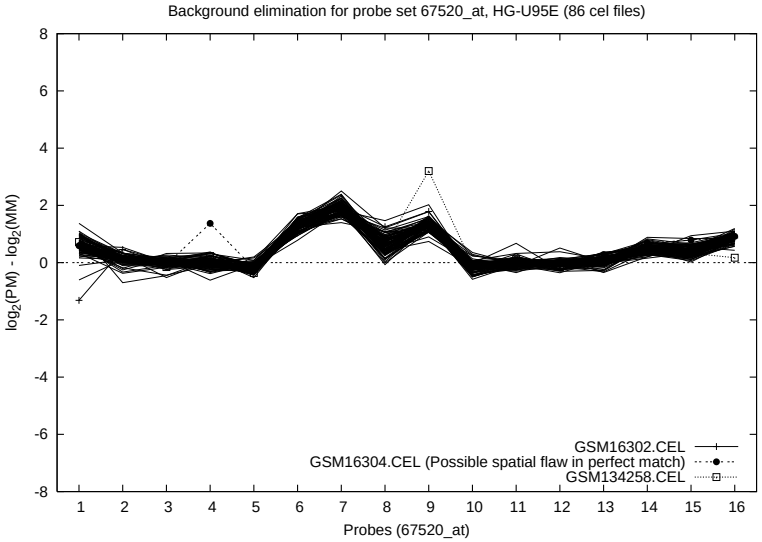
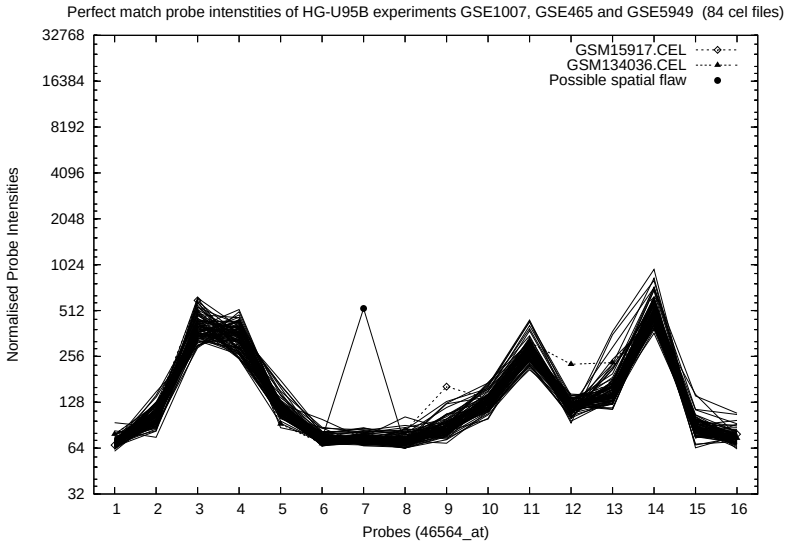


Figure 1: Probe intensities over all cel files for probe set 67520_at on the HG_U95E array (A) Perfect match intensities (log scale) (B) Background elimination (log_2(PM) - log_2(MM))

NM_015680.3	
Probe set	**Array**
207511_s_at	U133A, U133A_2, U133_Plus_2.0
200070_at	U113A, U133A_2, U113B, U133_Plus_2.0
34864_at	U95A, U95Av2
NM_020184.3	
Probe set	**Array**
218900_at	U133A, U133A, U133_Plus_2.0
46564_at	U95B
62274_at (probes 4 - 15)	U95C
63395_at (probes 1 - 8)	U95C

Table 1: Probe sets and corresponding human Affymetrix GeneChips mapping to NM_015680.3 and NM_020184.3.

A



B

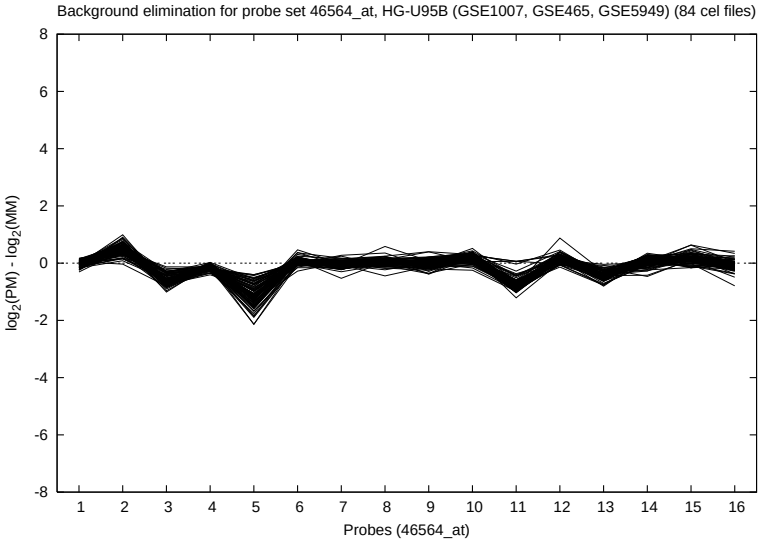
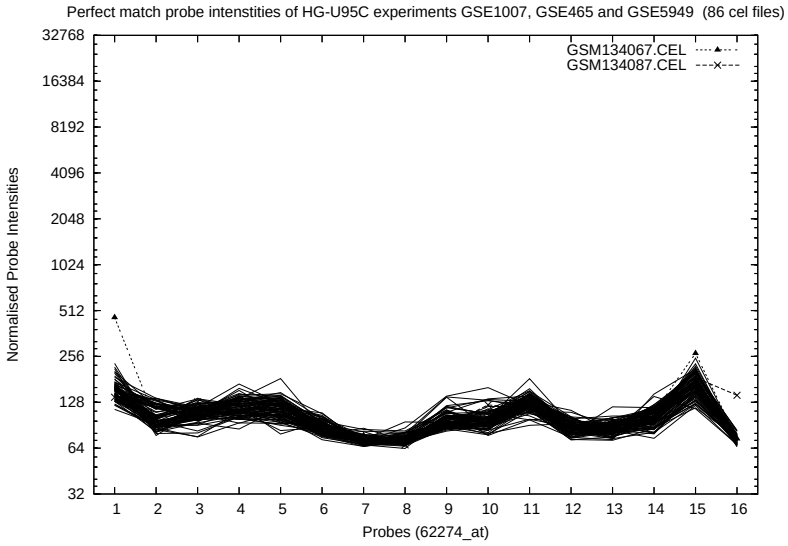


Figure 2: Probe intensities for experiments GSE1007, GSE465, GSE5949 of probe set 46564_at on the HG_U95B array (A) Perfect match intensities (log scale) (B) Background elimination ($\log_2$(PM) - $\log_2$(MM))

A



B

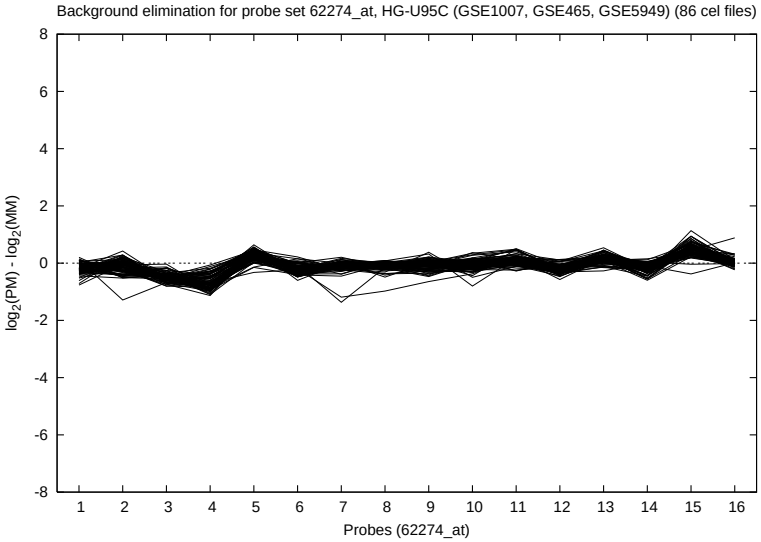
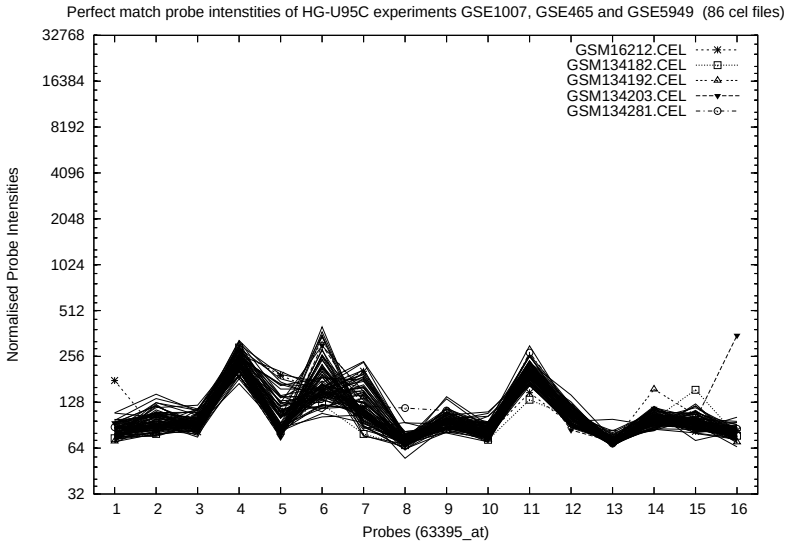


Figure 3: Probe intensities for experiments GSE1007, GSE465, GSE5949 of probe set 62274_at on the HG_U95C array (A) Perfect match intensities (log scale) (B) Background elimination ($\log_2$(PM) - $\log_2$(MM))

A



B

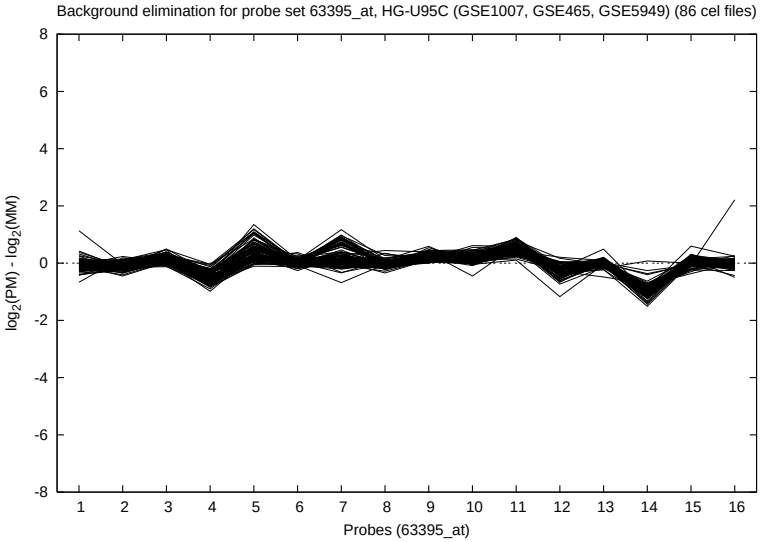
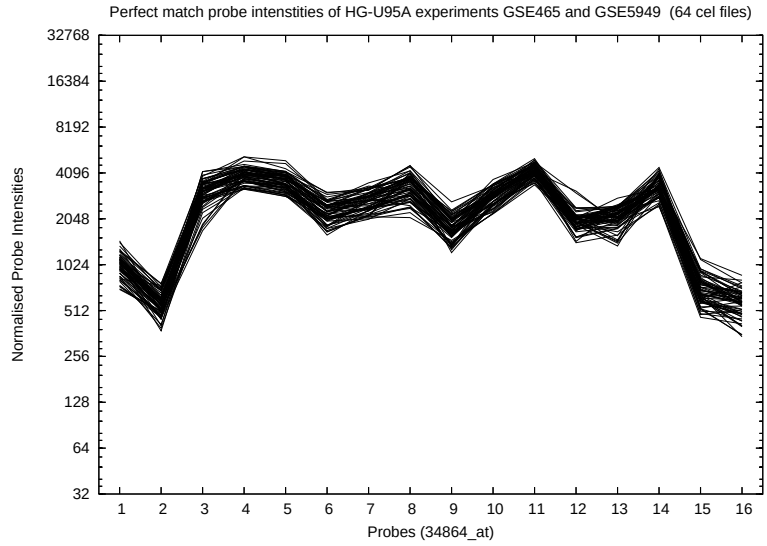


Figure 4: Probe intensities for experiments GSE1007, GSE465, GSE5949 of probe set 63395_at on the HG_U95C array (A) Perfect match intensities (log scale) (B) Background elimination ($\log_2$(PM) - $\log_2$(MM))

A



B

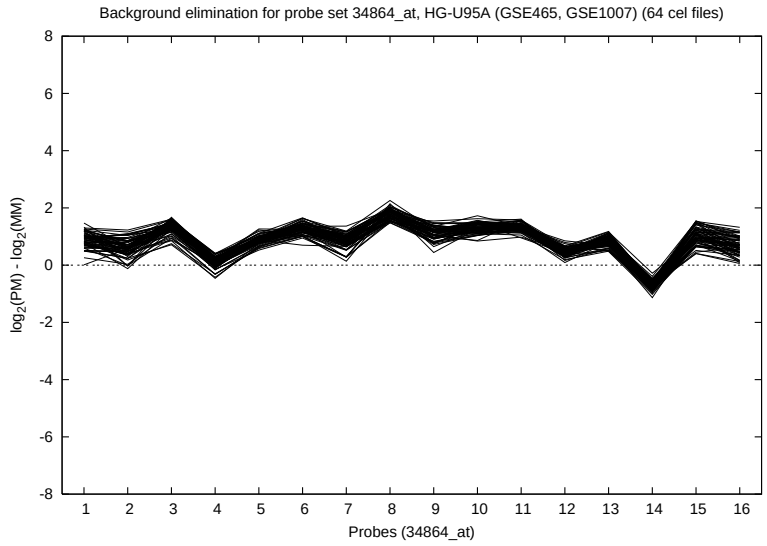
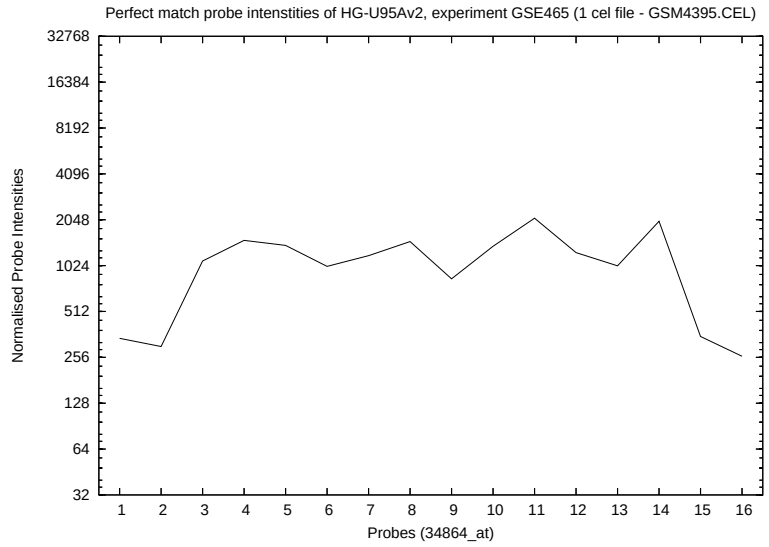


Figure 5: Probe intensities for experiments GSE1007 and GSE465 of probe set 34864_at on the HG_U95A array (A) Perfect match intensities (log scale) (B) Background elimination ($\log_2$(PM) - $\log_2$(MM))

A



B

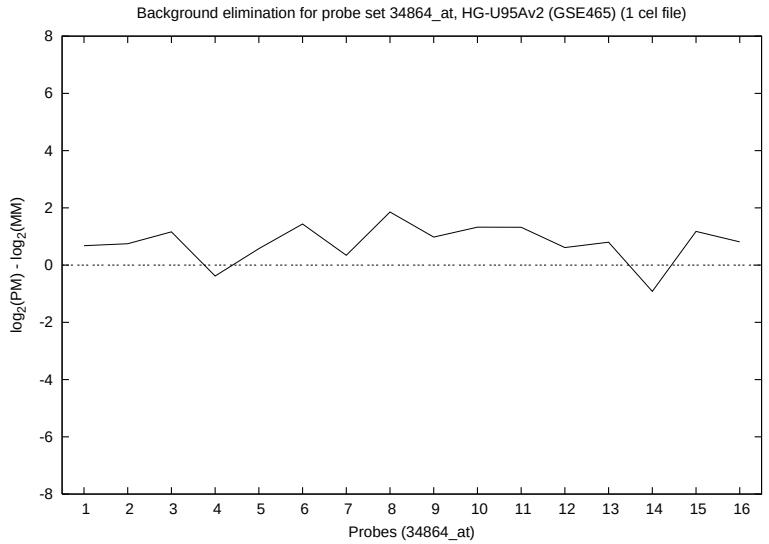
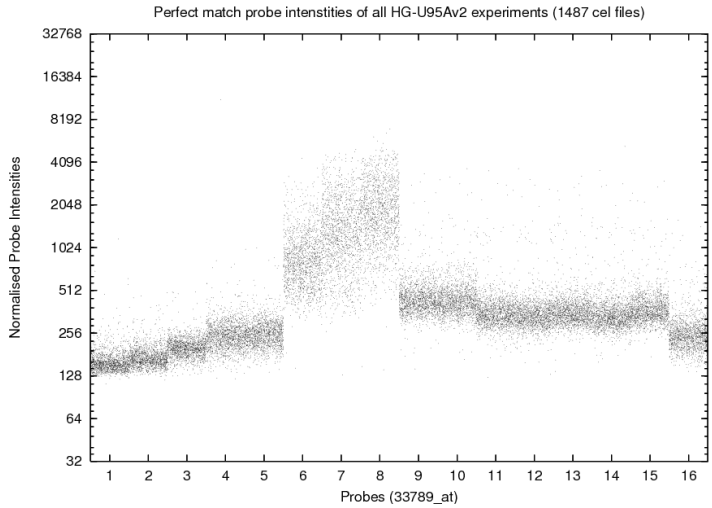
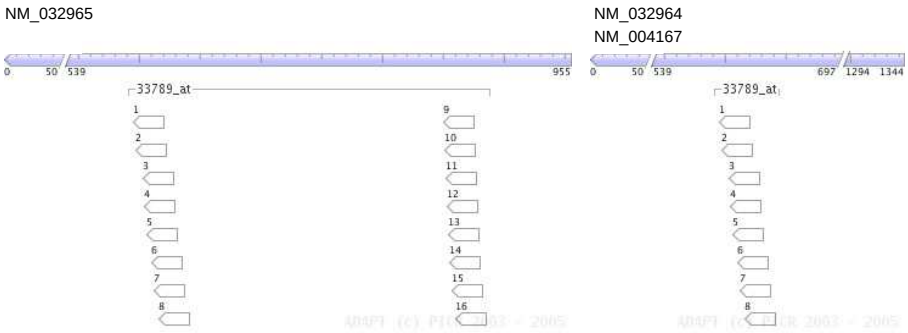


Figure 6: Probe intensities for experiment GSE465 of probe set 34864_at on the HG_U95Av2 array (A) Perfect match intensities (log scale) (B) Background elimination (log_2(PM) - log_2(MM))

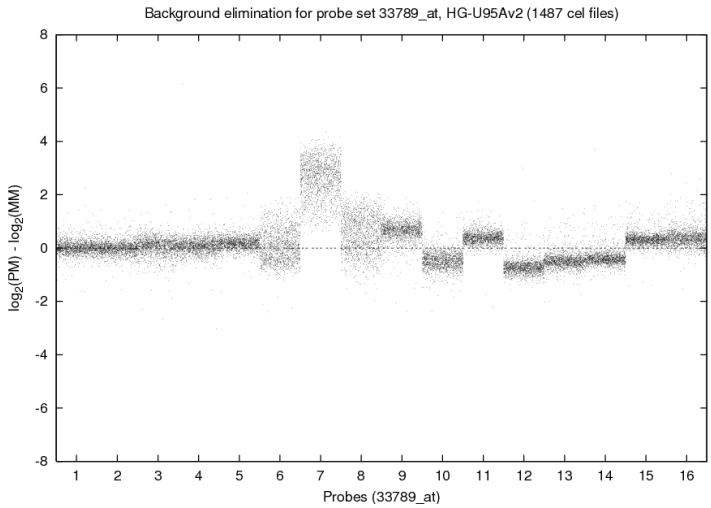
A



B

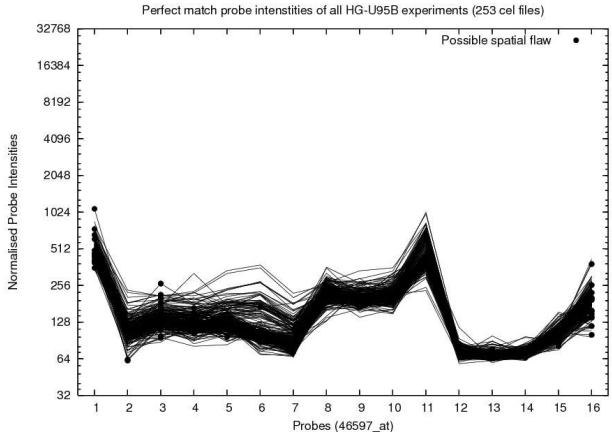


C

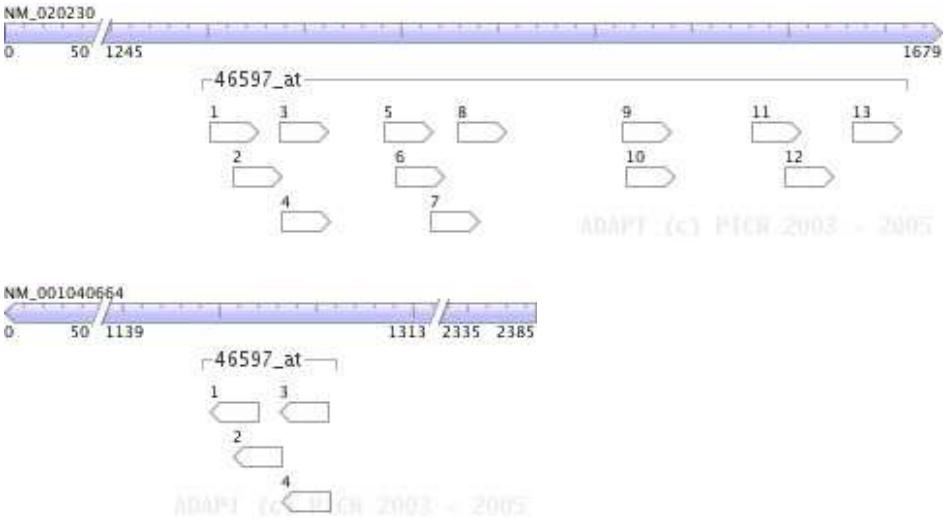


Supplementary Figure 1: Probe intensities over all cel files for probe set 33789_at on the HG_U95Av2 array (A) Perfect match intensities (log scale) (B) ADAPT probe positions on the transcripts (C) Background elimination ($\log_2$(PM) - $\log_2$(MM)).

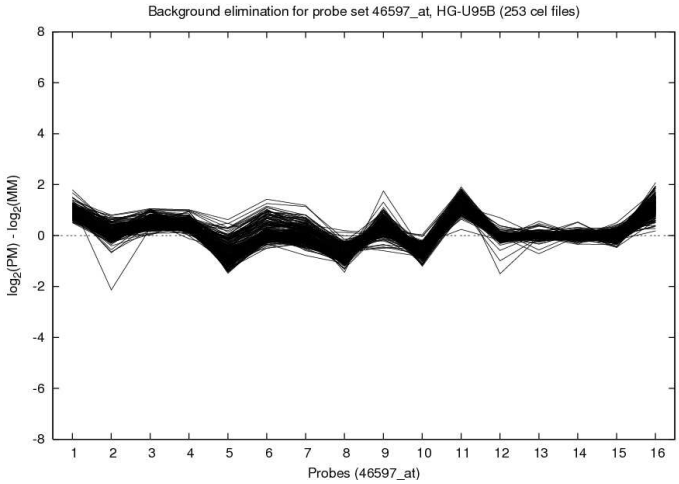
A



B

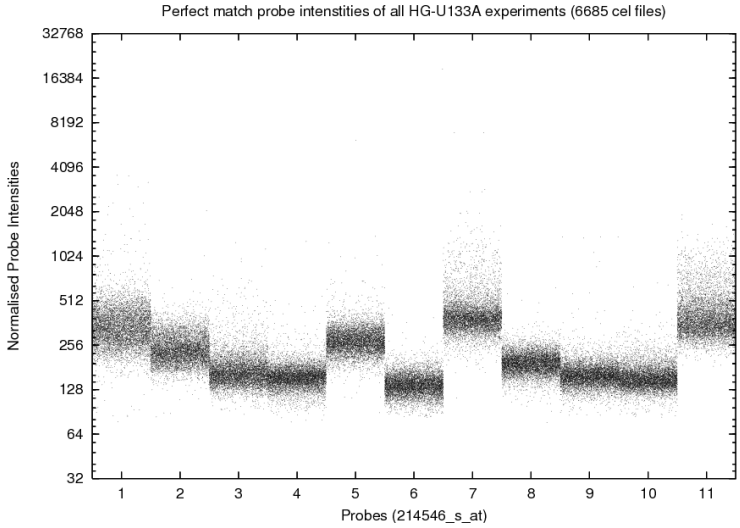


C

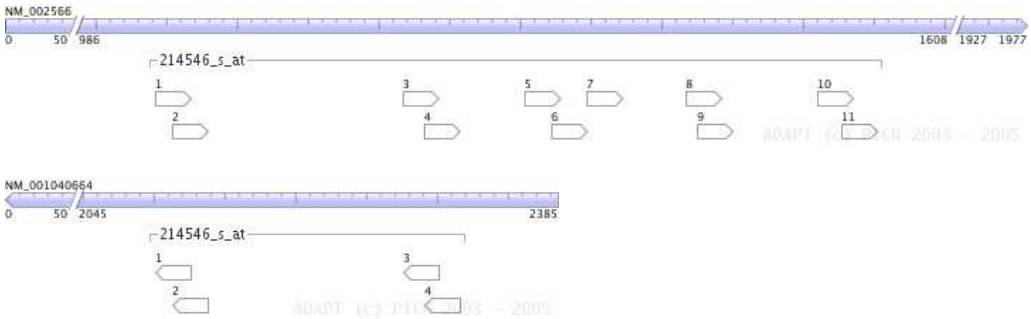


Supplementary Figure 2: Probe intensities over all cel files for probe set 46597_at on the HG_U95B array (A) Perfect match intensities (log scale) (B) ADAPT probe positions on the transcripts (C) Background elimination (log_2(PM) - log_2(MM)).

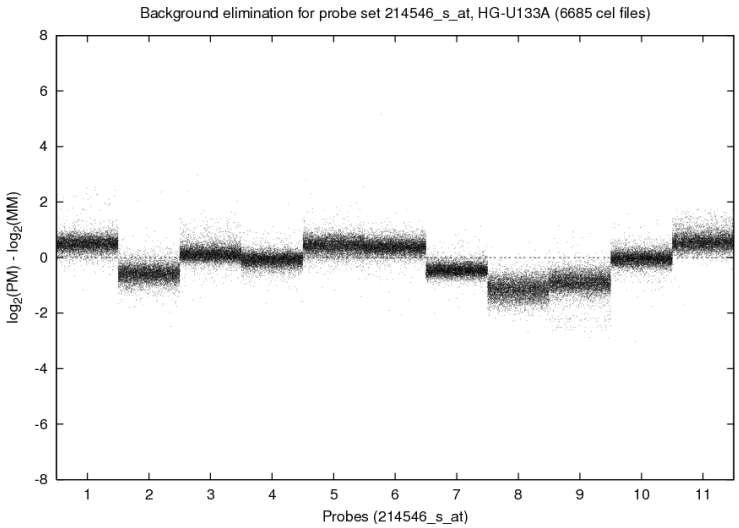
A



B

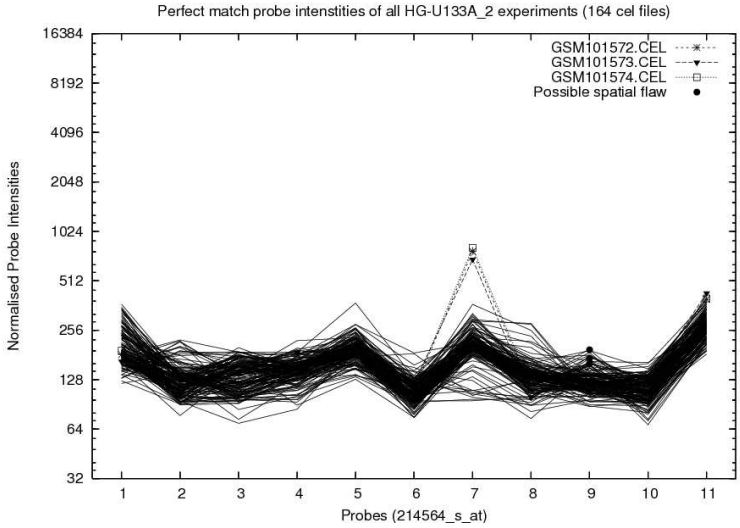


C

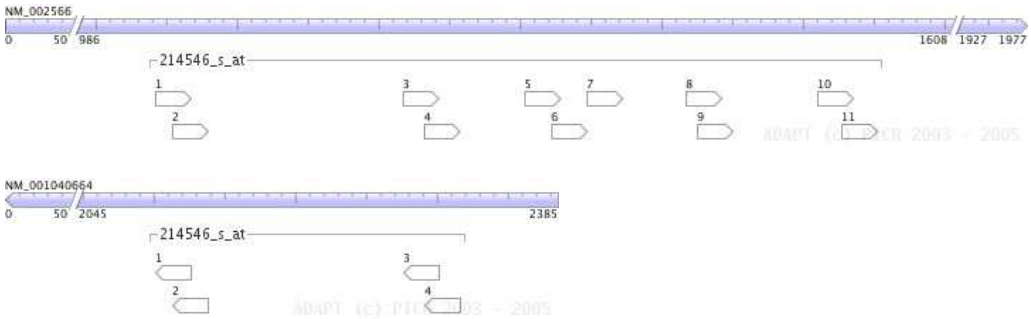


Supplementary Figure 3: Probe intensities over all cel files for probe set 214546_s_at on the HG_U133A array (A) Perfect match intensities (log scale) (B) ADAPT probe positions on the transcripts (C) Background elimination (log_2(PM) - log_2(MM)).

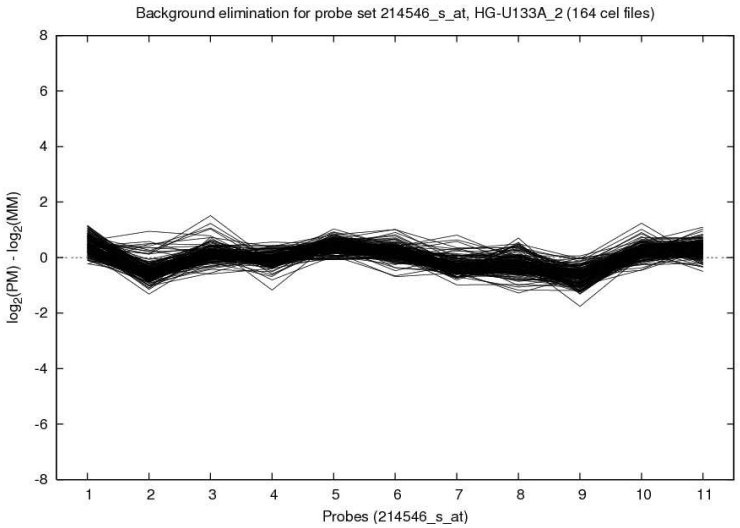
A



B

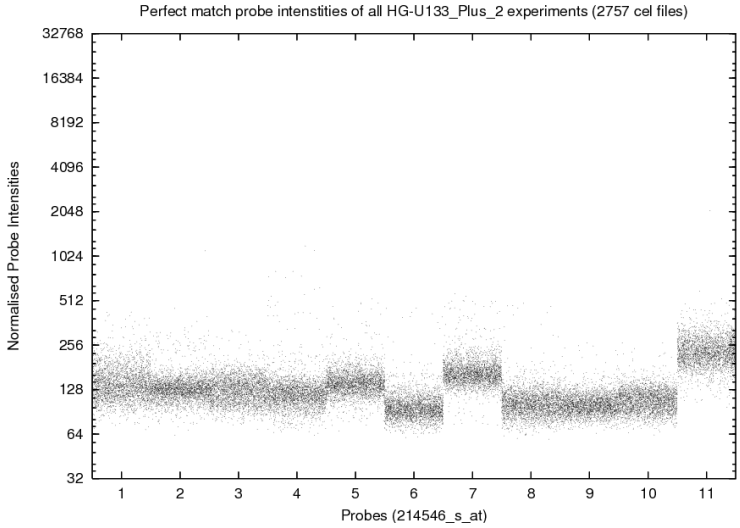


C

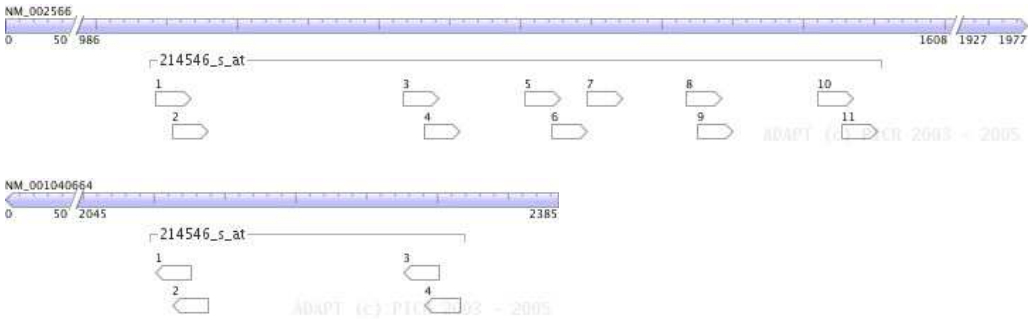


Supplementary Figure 4: Probe intensities over all cel files for probe set 214546_s_at on the HG_U133A_2 array (A) Perfect match intensities (log scale) (B) ADAPT probe positions on the transcripts (C) Background elimination ($\log_2$(PM) - $\log_2$(MM)).

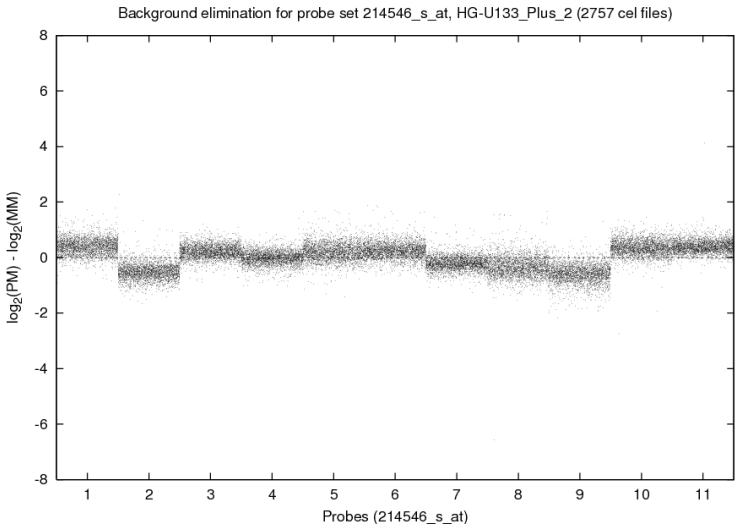
A



B

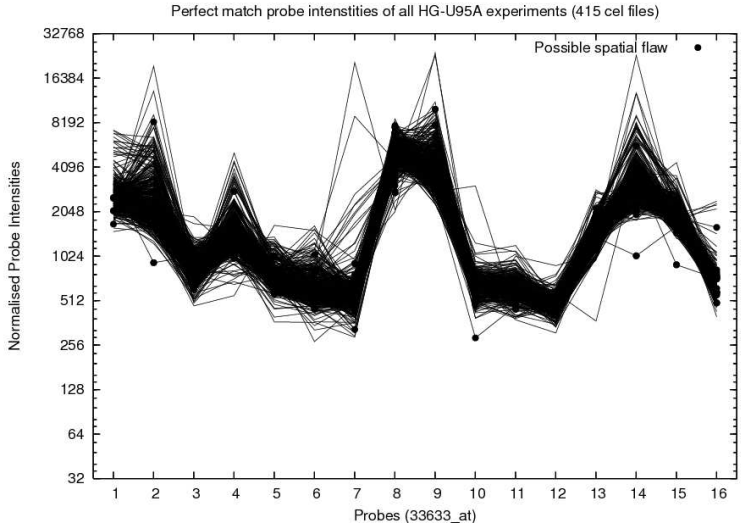


C

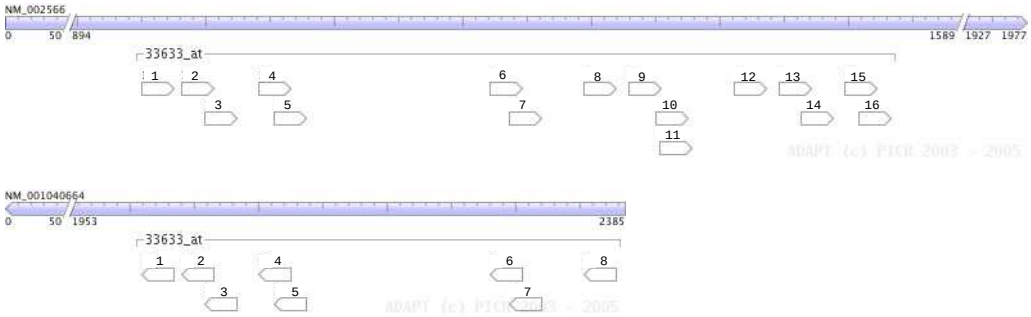


Supplementary Figure 5: Probe intensities over all cel files for probe set 214546_s_at on the HG_U133_Plus_2 array (A) Perfect match intensities (log scale) (B) ADAPT probe positions on the transcripts (C) Background elimination ($\log_2$(PM) - $\log_2$(MM)).

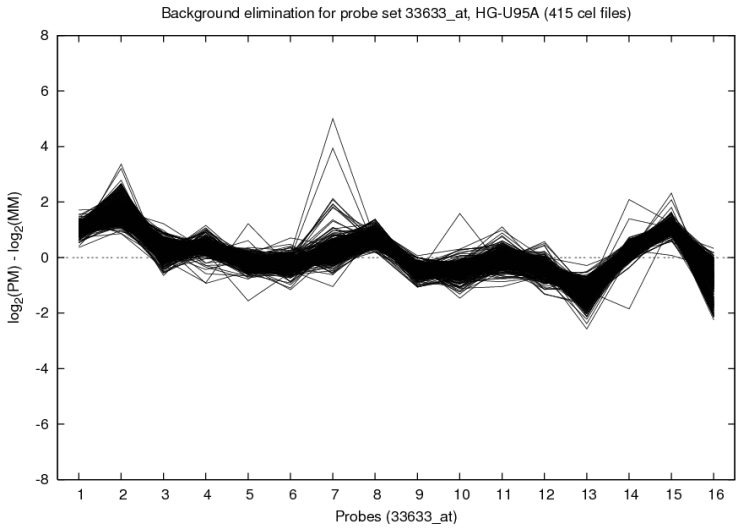
A



B

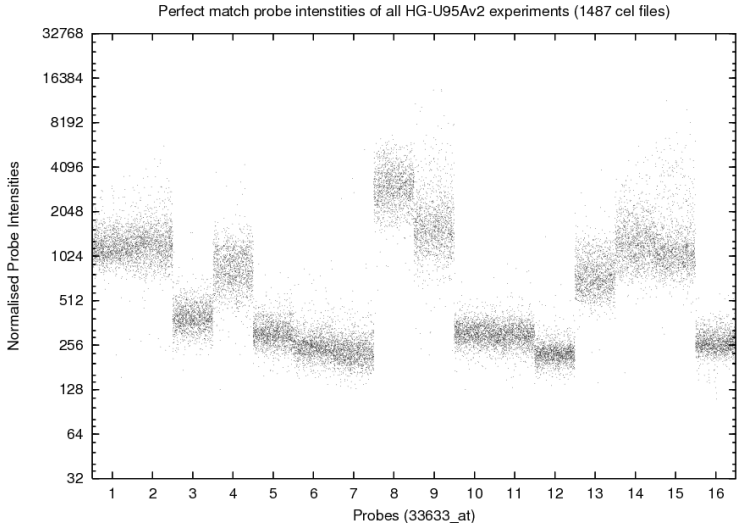


C

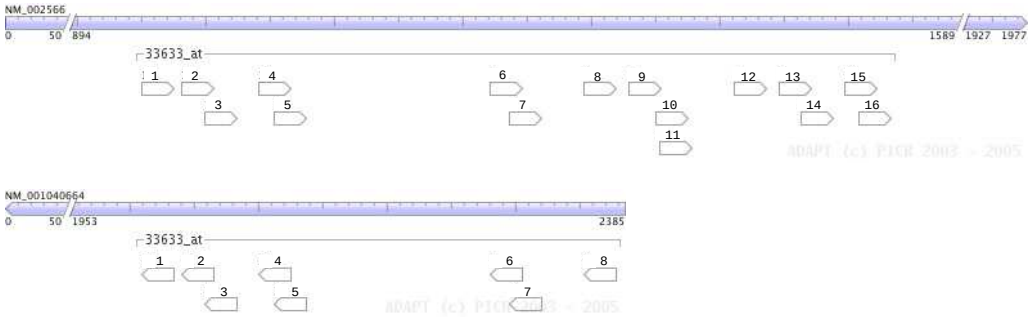


Supplementary Figure 6: Probe intensities over all cel files for probe set 33633_at on the HG_U95A array (A) Perfect match intensities (log scale) (B) ADAPT probe positions on the transcripts (C) Background elimination (log_2(PM) - log_2(MM)).

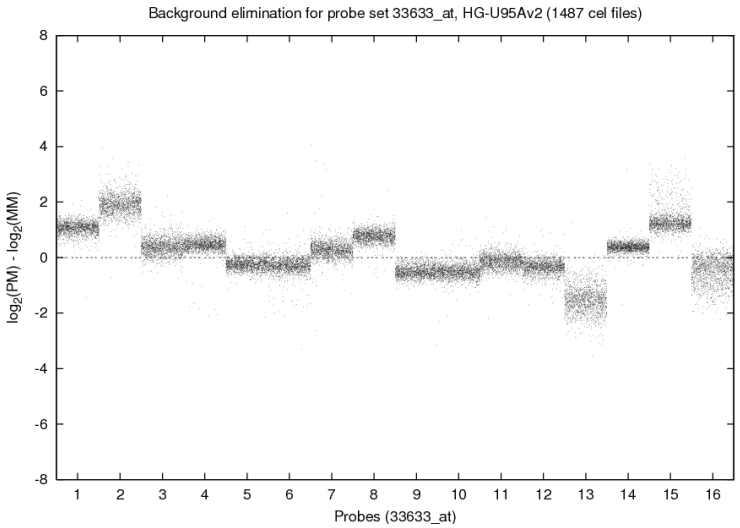
A



B

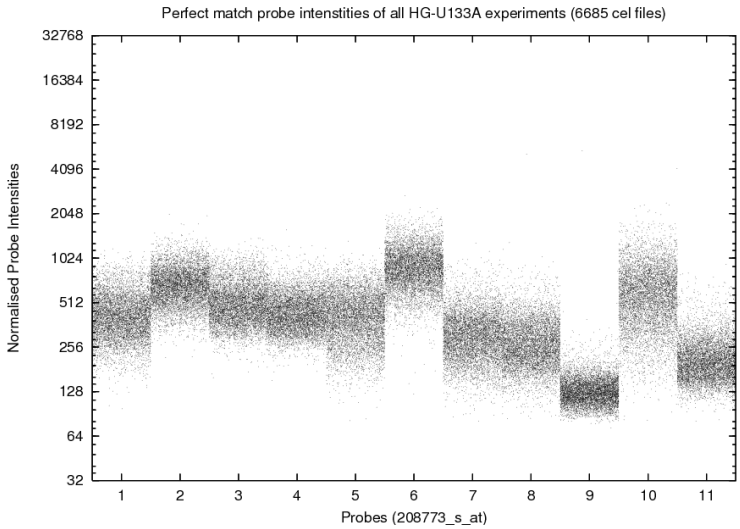


C

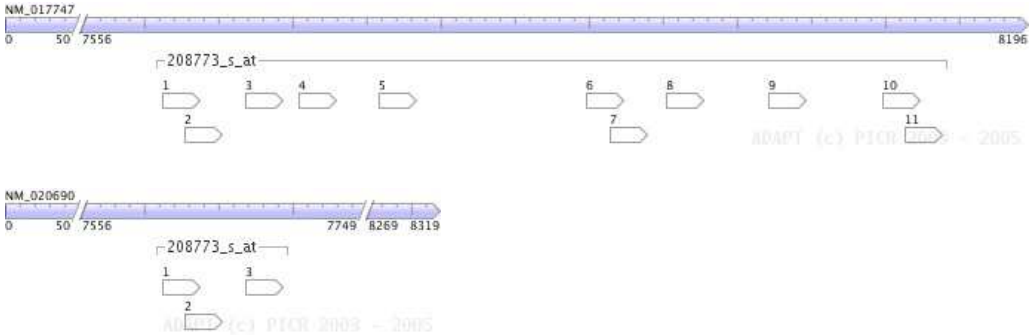


Supplementary Figure 7: Probe intensities over all cel files for probe set 33633_at on the HG_U95Av2 array (A) Perfect match intensities (log scale) (B) ADAPT probe positions on the transcripts (C) Background elimination (log_2(PM) - log_2(MM)).

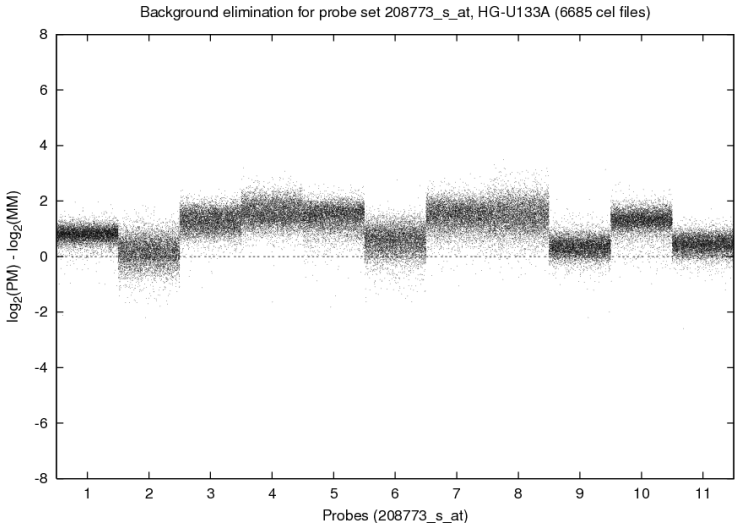
A



B

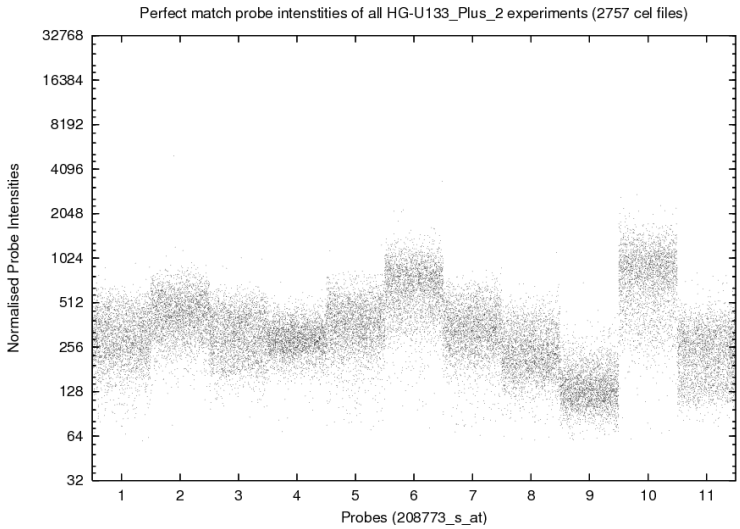


C

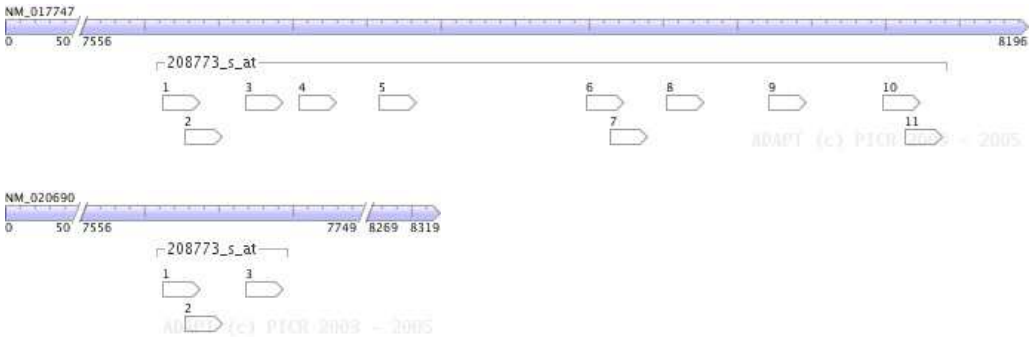


Supplementary Figure 8: Probe intensities over all cel files for probe set 208773_s_at on the HG_U133A array (A) Perfect match intensities (log scale) (B) ADAPT probe positions on the transcripts (C) Background elimination (log_2(PM) - log_2(MM)).

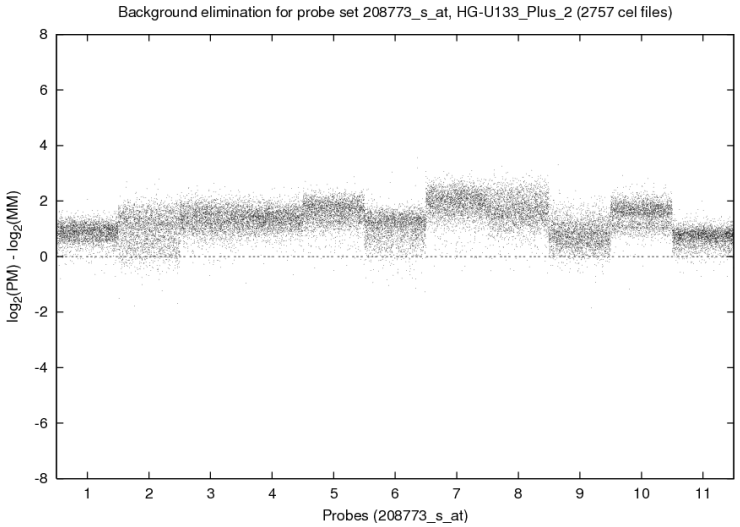
A



B

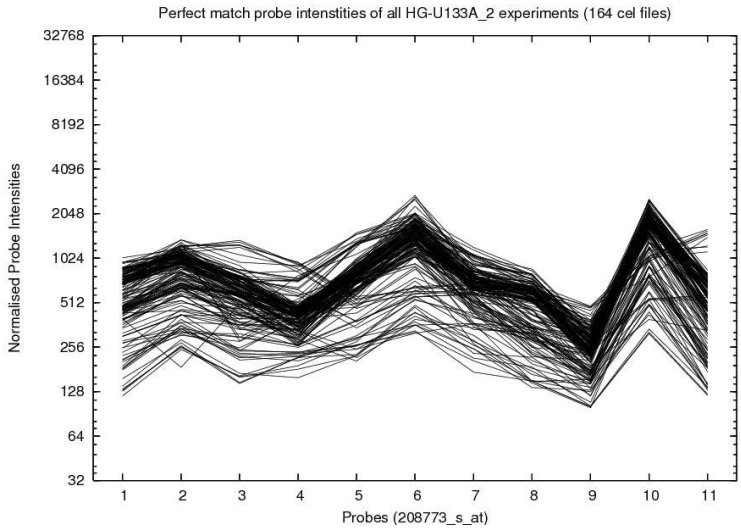


C

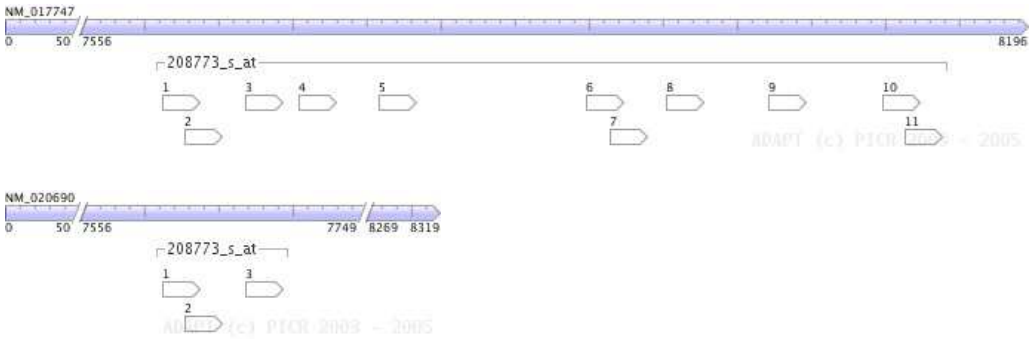


Supplementary Figure 9: Probe intensities over all cel files for probe set 208773_s_at on the HG_U133_Plus_2 array (A) Perfect match intensities (log scale) (B) ADAPT probe positions on the transcripts (C) Background elimination ($\log_2$(PM) - $\log_2$(MM)).

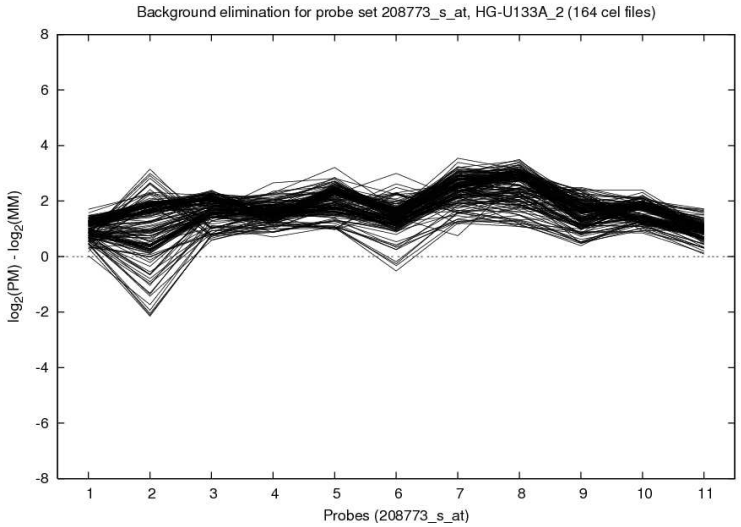
A



B



C



Supplementary Figure 10: Probe intensities over all cel files for probe set 208773_s_at on the HG_U133A_2 array (A) Perfect match intensities (log scale) (B) ADAPT probe positions on the transcripts (C) Background elimination ($\log_2$(PM) - $\log_2$(MM)).

Fusion	Gene 1	Gene 2	Chimera
HCC-2/HCC-1	NM_032965.2	NM_032963.2	NM_032964.2 (Var1) NM_004167.3 (Var2)
SSF1-P2Y$_{11}$	NM_020230.4	NM_002566.4	NM_001040664.1
MASK-BP3	NM_017747.1 (Var1) NM_017978.1 (Var2) NM_024668.2 (Var3)	NM_003732.2	NM_020690.4

Table 1: RefSeq transcript accession numbers of the chimeras and individual transcripts of the adjacent genes. The third and fourth columns show the RefSeq accession numbers for each of the individual gene transcripts and any variants. The fifth column shows the RefSeq accession numbers for the chimeric transcripts.