Using surveys of Affymetrix GeneChips to study antisense expression

Olivia Sanchez-Graillet, Maria A. Stalteri, Joanna Rowsell, Graham J.G. Upton and Andrew P. Harrison*

Departments of Mathematical Sciences and Biological Sciences, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ

Summary

We have used large surveys of Affymetrix GeneChip HG-U133_Plus_2 data in the public domain to conduct a study of antisense expression across diverse conditions.
We derive correlations between groups of probes which map uniquely to the same exon in the antisense direction. When there are no probes assigned to an exon in the sense direction we find that many of the antisense groups fail to detect a coherent block of transcription. We find that only a minority of these groups contain coherent blocks of antisense expression suggesting transcription.

We also derive correlations between groups of probes which map uniquely to the same exon in both sense and antisense direction. In some of these cases the locations of sense probes overlap with the antisense probes, and the sense and antisense probe intensities are correlated with each other. This configuration suggests the existence of a Natural Antisense Transcript (NAT) pair. We find the majority of such NAT pairs detected by GeneChips are formed by a transcript of an established gene and either an EST or an mRNA.

In order to determine the exact antisense regulatory mechanism indicated by the correlation of sense probes with antisense probes, a further investigation is necessary for every particular case of interest. However, the analysis of microarray data has proved to be a good method to reconfirm known NATs, discover new ones, as well as to notice possible problems in the annotation of antisense transcripts.

1 Introduction

Our knowledge of the transcriptome is rapidly evolving and it is becoming increasingly clear that RNA plays a range of diverse roles in regulating gene expression [1]. Natural antisense transcripts (NATs) are endogenous RNAs whose sequences are complementary to other transcripts. Antisense transcripts are implicated in transcription, processing, stability, transport and translation of their complementary RNAs [2]. NATs have now been found in many organisms, but we have little knowledge of the functions of many of these transcripts [3]. Bioinformatic approaches show a large number of potential NATs in genomic sequences, but provide no information about the expression of NATs in specific cell or tissue types [4]. It is therefore important to experimentally verify the expression of NATs in order to unravel their biology.

Affymetrix GeneChip technology [5] is a widely used resource in the life sciences. GeneChips provide multiple measures of the expression level for each gene. Each probe is a

* Corresponding author: harry@essex.ac.uk

doi:10.2390/biecoll-jib-2010-114

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. The widespread popularity of GeneChips, with large data-sets stored in public repositories such as the Gene Expression Omnibus [6], makes them particularly suited for unravelling aspects of the transcriptome across many conditions, for diverse conditions, developmental stages, phenotypes and diseases. However, such studies will be limited to a sample of the transcriptome, that for which there are probes with the appropriate sequence. A huge number of expressed sequence tags (ESTs) were used in the design of the Affymetrix arrays [5] and due to the extensive use of such sequences Stalteri and Harrison [7] predicted that some probes may be mapping to exotic RNA sequences. We are not the only group to consider the use of Affymetrix data for exploring the biology of the transcriptome and there have already been searches for antisense expression using mouse arrays [8, 9].

Experimental artefacts in the preparation of targets, such as spurious synthesis of complementary strands, may act to confuse the interpretation of genome-wide experiments [2]. In some cases it is likely that a significant number of postulated NATs may be artefacts produced by genomic priming with contaminant genomic DNA during cDNA library construction [3]. Given the potential for confusion resulting from artefacts, it is imperative that care is taken in analysing Affymetrix data when searching for NATs. It is widely assumed that on a GeneChip multiple probes from within the same probeset measure the same thing. However, we find there are a number of probesets that contain probes behaving inconsistently with the rest of the probeset [10]. Some of these discrepancies may result from interesting biological processes such as alternative splicing and alternative polyadenylation [7]. However, other problems result from spatial flaws in hybridization [11]. Moreover, some probes may not measure expression reliably, due to particular sub-sequences, or motifs, within their 25 bases. Wu et al. [12] reported that probes containing runs of guanine were typically outliers in probesets and also showed abnormal binding affinities. We recently confirmed that such probes are outliers [13], but further discovered that the probes containing runs of guanine are unusually well correlated with each other across many thousands of experiments. We associate this effect with the formation of G-quadruplexes occurring on the surface of a GeneChip [13]. Studying correlations in expression across many experiments is informative because coordinated biases affecting many probes simultaneously can be identified.

The regulation of antisense transcription might be tailored to its type of action [14], and the expression patterns of NATs and their targets might indicate the regulatory mechanism that is occurring. For example, when both sense and antisense probes are correlated with each other they should measure the same thing. This might indicate bidirectional transcription, particularly as [15] discovered that antisense transcripts are mainly located in the promoter and terminator regions of genes.

2 Materials and methods

We use a pipeline to analyse tens of thousands of Affymetrix GeneChips [10] downloaded from the Gene Expression Omnibus [6]. Our pipeline brings together unique mappings of probes, quality control analysis on each GeneChip and data-mining signal intensities across many experiments.

We first identify spatial flaws in individual GeneChips [11,16,17] that leads us to blank out signals from a fraction of each chip. We group all probes aligning to the same exon together, and we calculate the correlations between each of the probe pairs. The intensities are

transformed onto a log scale and the signals are correlated across all experiments for one chip type. All the pair-wise probe correlations for each exon are collated into a matrix that is colour-coded according to the correlation value. The original correlation values are multiplied by ten, and then rounded, so that we express the correlations as integers. Heatmaps are symmetrical matrices in which the diagonal represents the perfect correlation of each probe with itself (correlation with value 10).

We wish to only study unique probes, those that only target one place on one exon [10]. This means that we only utilise a fraction of the probes available on the GeneChip, but the fraction we use will have been chosen to provide reliable measurements. We proceed by calculating the alignment "value" for each probe through multiplying the alignment length and the percentage sequence identity, e.g. a probe that aligns to a sequence with 25 bases and percentage sequence identity of 80% has an alignment value of 20 (25x0.8). A probe is considered 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 exons in the same genomic region but with different Ensembl identifiers); maps to only one place on the exon; does not map to any exon-exon junctions; does not map substantially to any other exon (i.e. does not have an alignment value between 20 and 25 for any other exon). We also identify probes that map uniquely to exons in an antisense direction. Such probes map uniquely to the reverse complement of the exon sequence. An example of an antisense probe is illustrated in Figure 1. We can observe that the reverse complement of the sequence of probe 201427_s_at:294:1093 aligns to exon ENSE00001435187. Thus if the NAT to ENSE00001435187 is expressed, it will be detected by the probe 201427_s_at:294:1093.

For the present study, we analyse the CEL files obtained from GEO for experiments that used the Human GeneChip HG-U133_Plus_2. We obtain our genomic coordinates and exon definitions from Ensembl (release 48). Probes containing the motif CCTCC or runs of four or more contiguous guanines were taken out of the exon heatmaps since they produce misleading information [18].

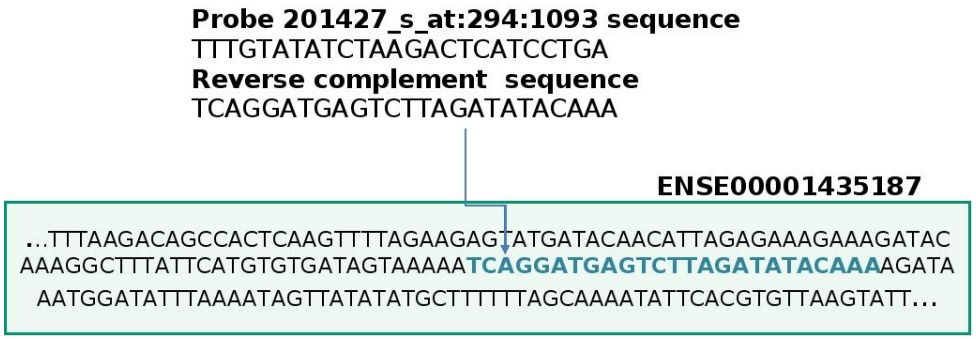


Figure 1. Probe 201427_s_at maps in an antisense direction to exon ENSE00001435187 (only a fragment is shown).

3 Results

In this section we present our analysis of the heatmaps generated for exons containing only antisense probes and for exons containing both sense and antisese probes.

3.1 Exons with antisense probes only

Almost all groups of antisense probes that map uniquely to an exon, but for which there are no sense probes mapping, showed low or negative correlation with each other. As an example, the antisense probes in Figure 2 are not correlated although they are from the same probeset (222076_at) and map to the same exon ENSE00000764836. The probes are not detecting a coherent signal. Figure 3 shows that there are RefSeq and Ensembl transcripts only on the positive strand. There are no transcripts on the negative strand, to which the probes in the probeset 222076_at map.

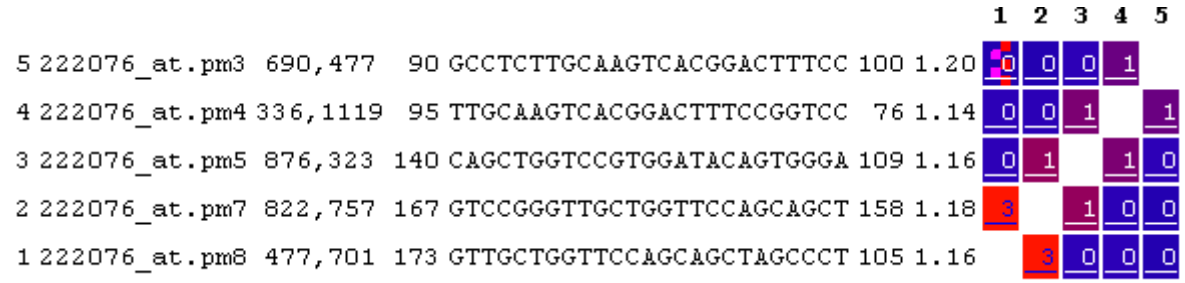


Figure 2. The antisense probes from probeset 222076_at are not correlated. The columns indicate the probe order in the heatmap, probe identifier (in which pm means perfect match, followed by the order of the probe in its probeset), x-coordinate of probe location on the array, y-coordinate of probe location on the array, interrogation position of probe on Affymetrix consensus sequence, probe sequence, geometric mean of the intensities across GEO, and standard deviation (of the logs of intensities), respectively. The numbers in each of the cells represent the rounded correlation x 10.

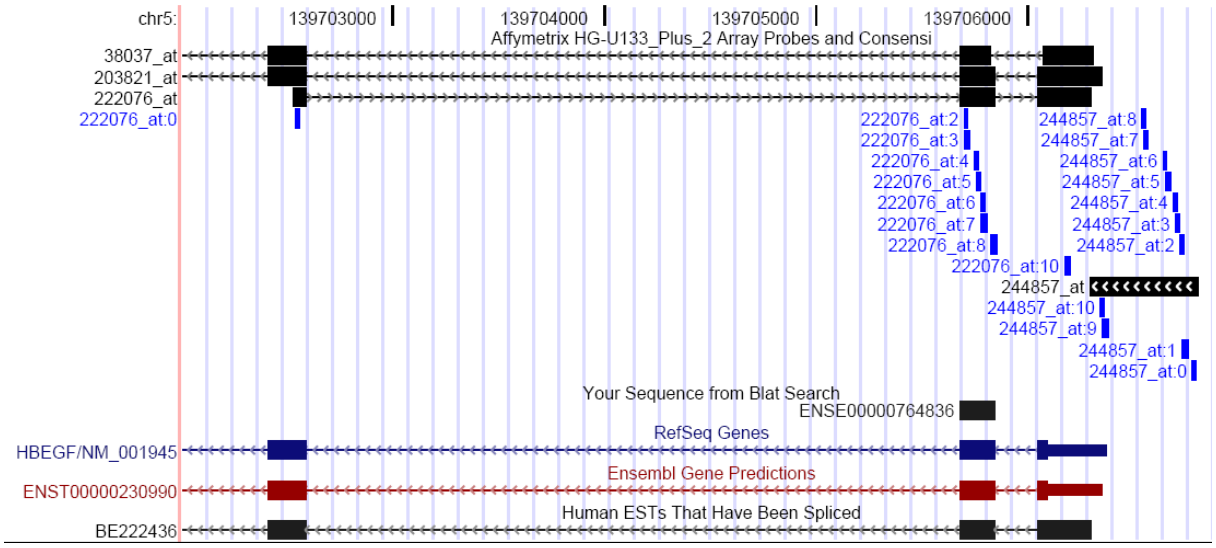


Figure 3. Screen-shot of the UCSC browser [19] shows that the probes in the probeset 222076_at are on the positive strand and exon ENSE00000764836 (in ENST00000230990) is on the negative strand. There are no transcripts aligning to the positive strand in this region.

There are only a few cases in which antisense probes mapping uniquely to the same exon (or from the same probeset) are highly correlated. As an example, the heatmap in Figure 4 refers to a group of highly correlated (average correlation 0.84) antisense probes in the probeset 201427_s_at. The probes illustrated are the only probes in the probeset that uniquely map to the exon ENSE0001435187. The high correlations among these antisense probes suggest that there may be a real biological signal. Figure 5 shows that the probes in probeset 201427_s_at

map to a region where there is overlap between the 3' ends of the CCDC152 and SEPP1 genes, which are transcribed from opposite strands. Probeset 201427_s_at aligns to the negative strand, and thus it aligns sense to the SEPP1 transcripts, and antisense to the RefSeq transcript CCDC152 and to the Ensembl transcript ENST00000361970 (through exon ENSE00001435187) that are on the positive strand. Affymetrix assigns probeset 201427_at to SEPP1, with an annotation grade of "A", i.e., at least 9 of the 11 probes perfectly match the associated transcripts [20], which in this case include the 3 RefSeq transcripts for SEPP1. The Affymetrix annotation for probeset 201427_s_at also includes several cross-hybridising transcripts assigned as having 11/11 Negative Strand Matching Probes. One of these is NM_001134848, the RefSeq transcript for CCDC152. The antisense (with respect to exon ENSE0001435187) transcription being detected is expected to be from the RefSeq transcripts corresponding to the SEPP1 gene. The NATsDB database [21] describes SEPP1 as belonging to a SA (sense-antisense) pair with the mRNA BC039102 that is on the positive strand.

							1	2	3	4	5	6
6	201427_s_at.pm11	57,837	1986	GGATACAGTACGGATTTGTCCAAAT	1769	3.97	10	10	9	9	10	
5	201427_s_at.pm10	72,441	1935	CCTGACCTCCTTTATGGTTAATACT	1410	2.90	10	10	9	10		10
4	201427_s_at.pm9	291,449	1927	CCTATAAACCTGACCTCCTTTATGG	1493	2.58	9	9	9		10	9
3	201427_s_at.pm6	627,217	1816	AAACTTGAGTGGCTGTCTTAAAAGA	705	2.81	10	10		9	9	9
2	201427_s_at.pm3	545,253	1733	AAGACTCATCCTGATTTTTACTATC	945	3.20	10		10	9	10	10
1	201427_s_at.pm2	294,1093	1722	TTTGTATATCTAAGACTCATCCTGA	795	3.21		10	10	9	10	10

Figure 4. Highly correlated antisense probes from probeset 201427_s_at. The probes map in antisense direction to exon ENSE00001435187.

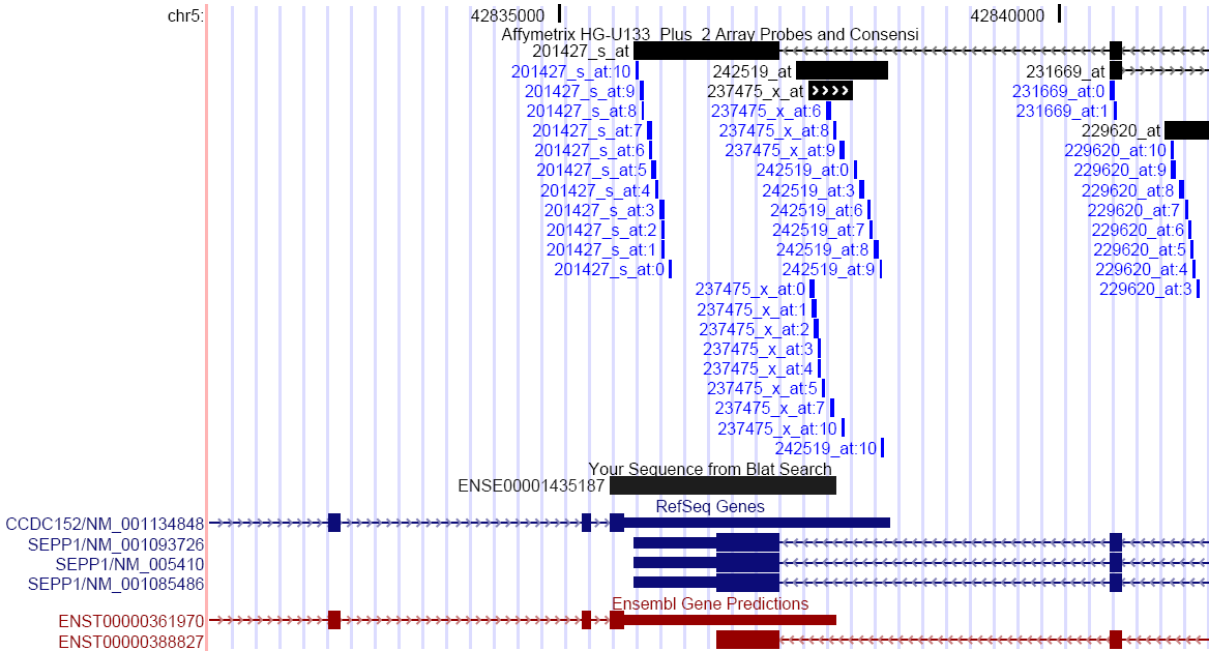


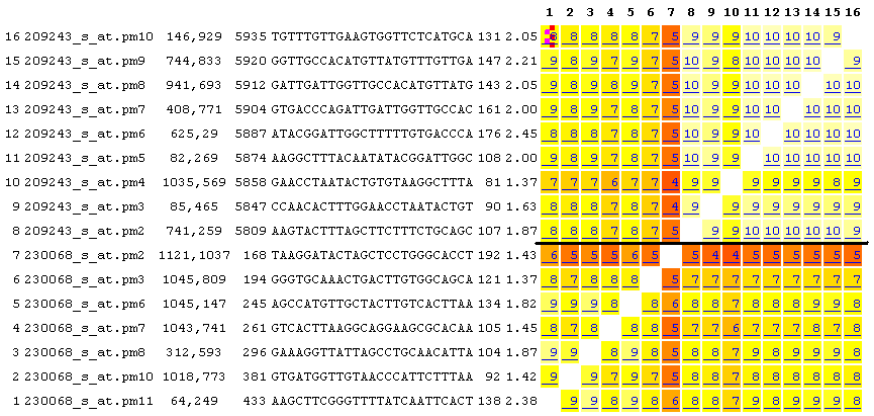
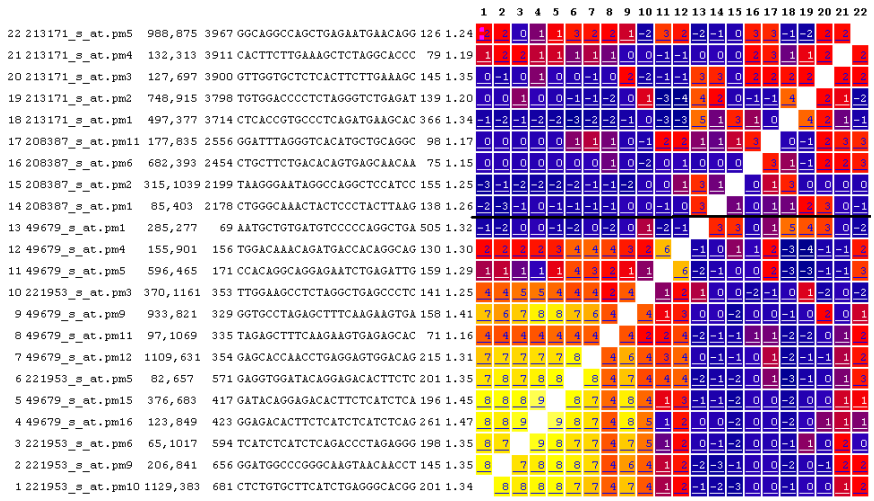
Figure 5. Screen-shot of the UCSC browser [19] showing that probes in probeset 201427_s_at map to the SEPP1 transcripts (negative strand).

3.2 Exons with antisense and sense probes

3.1.1 Classification by correlation heatmap

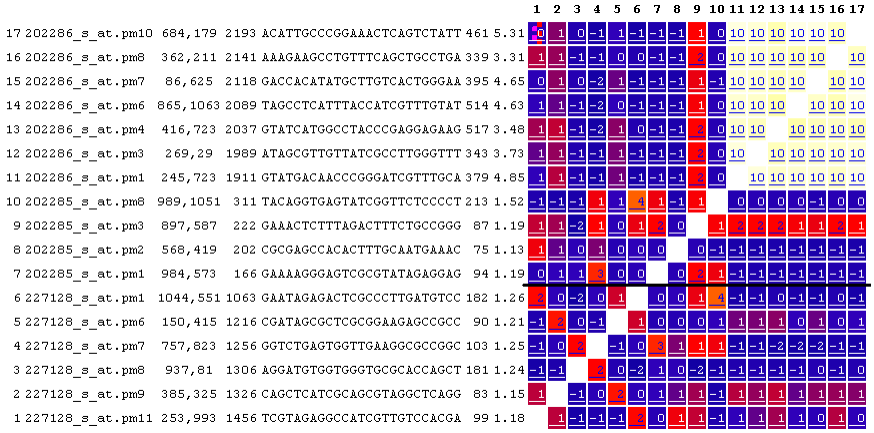
Out of 1,048 exons that have sense and antisense probes mapping uniquely to them, we selected 433 exons that have at least 4 sense probes and at least 4 antisense probes. We analysed a total of 100 exons randomly selected out of the 433 exons with probes in both senses. The exons can be classified into the general patterns shown in Table 1. Detailed descriptions of the biology of example transcripts matching these patterns are presented in the Supplementary material.

Table 1. Examples of general patterns of exons containing probes in sense and antisense directions. In each heatmap, the probes below the line align in the antisense direction and the probes above the line align in the sense direction. The antisense and sense probes do not necessarily overlap.

Pattern Description	Number of cases	Example
1: The sense and antisense probes are correlated.	14	 **ENSE00000876661**
2: Only the antisense probes are correlated.	4	 **ENSE00000860190**

3: Only the sense probes are correlated. A sub-group of sense probes may not be correlated with another sub-group of sense probes.

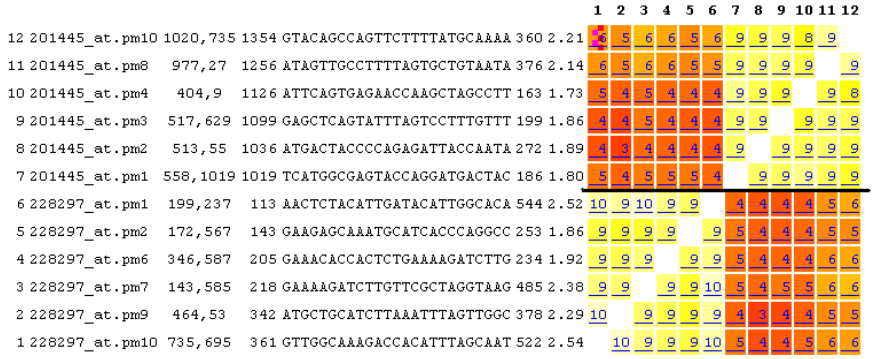
40



ENSE00001454677

4: The antisense probes are correlated, the sense probes are correlated. The antisense probes are not correlated or have low correlations with the sense probes.

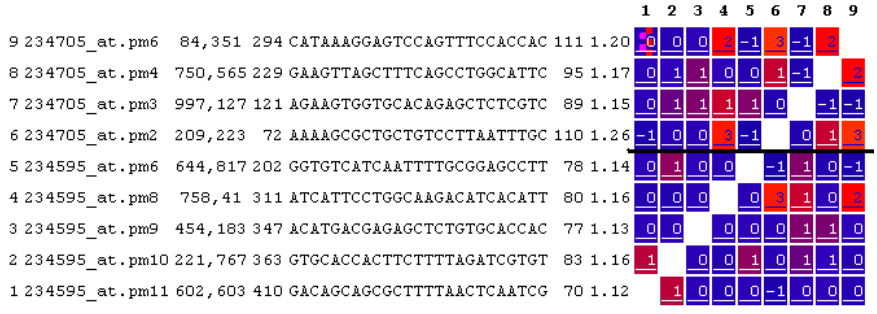
13



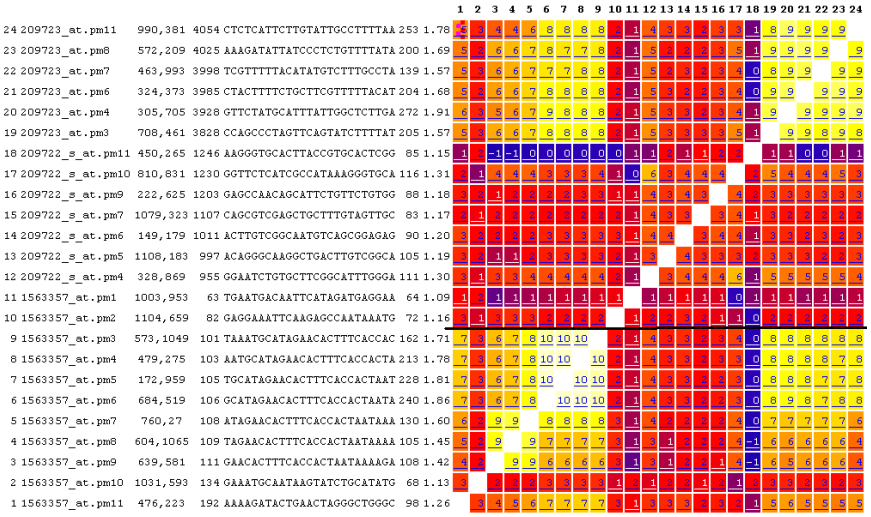
ENSE00001452067

5: The probes are not correlated or are negatively correlated.

26



ENSE00001222306



6: The antisense probes are correlated. Some of the sense probes are correlated. The antisense probes are correlated with a sub-group within the sense probes.

3

ENSE00001485956

3.1.2 Antisense and sense probes heatmaps across different GEO experiments

In order to check whether the pattern presented by exon ENSE00001452067 (in pattern 4) was constant across different GEO experiments, we generated the heatmaps for this exon in 40 GSE experiment series that had at least 10 GSM cel files each. Figure 6 depicts these heatmaps and Table 2 contains the descriptions of the types of GSE experiments corresponding to each heatmap. The first 13 heatmaps from left to right starting at the top of Figure 6 correspond to experiments related to cancer. The remaining 27 heatmaps are not related to cancer.

In general, we observe that the antisense probes are highly correlated across the different experiments either with cancer or not. The antisense probe correlations are represented by the first 6 probes on the bottom left of the heatmaps. It seems that the presence of cancer does not determine the pattern of the correlation heatmaps.

3.1.3 General remarks

When an exon has sense and antisense probes mapping uniquely to it, there is usually a SA (sense antisense) pair in the NATsDB database [21] associated to the gene which contains that exon.

When sense and antisense probes overlap, the existing SA-pair is mainly formed by an established gene and by an mRNA or EST. This is confirmed by the work of Yelin et al. [22] who detected that the overlap between genes is frequently established by complementary ESTs, even when mRNAs are present in the clusters. They also observed that around 70% of the SA genes overlapped in their 5'-most and 3'-most exons (here called external exons) which supports the idea that SA overlap could be involved in gene regulation since the external exons contain UTRs of mRNAs. We find that ~70% of the exons containing sense and antisense probes are external exons and that ~23% of the exons containing only antisense probes are external exons.

The most frequent heatmap patterns found in our study are patterns 3 and 5. The exons which follow pattern 5 (nothing is correlated) tend to be NBD (non-bidirectional) NATs (i.e., the complementary sequences are found on the same strand going in the same direction) [23] or are not part of a SA-pair (in the NATsDB database). Pattern 3 involves exons in which only the sense probes are totally or partially correlated with each other. The fact that only sense

probes are correlated suggests that there is no transcription in the antisense gene/RNA.

Figure 6. Heatmaps with data from different experiments for exon ENSE00001452067. The heatmap representing the overall experiments is in Table 1 pattern 4 (the antisense probes are correlated, the sense probes are correlated, and the sense probes are not correlated or are weakly correlated with the antisense probes).

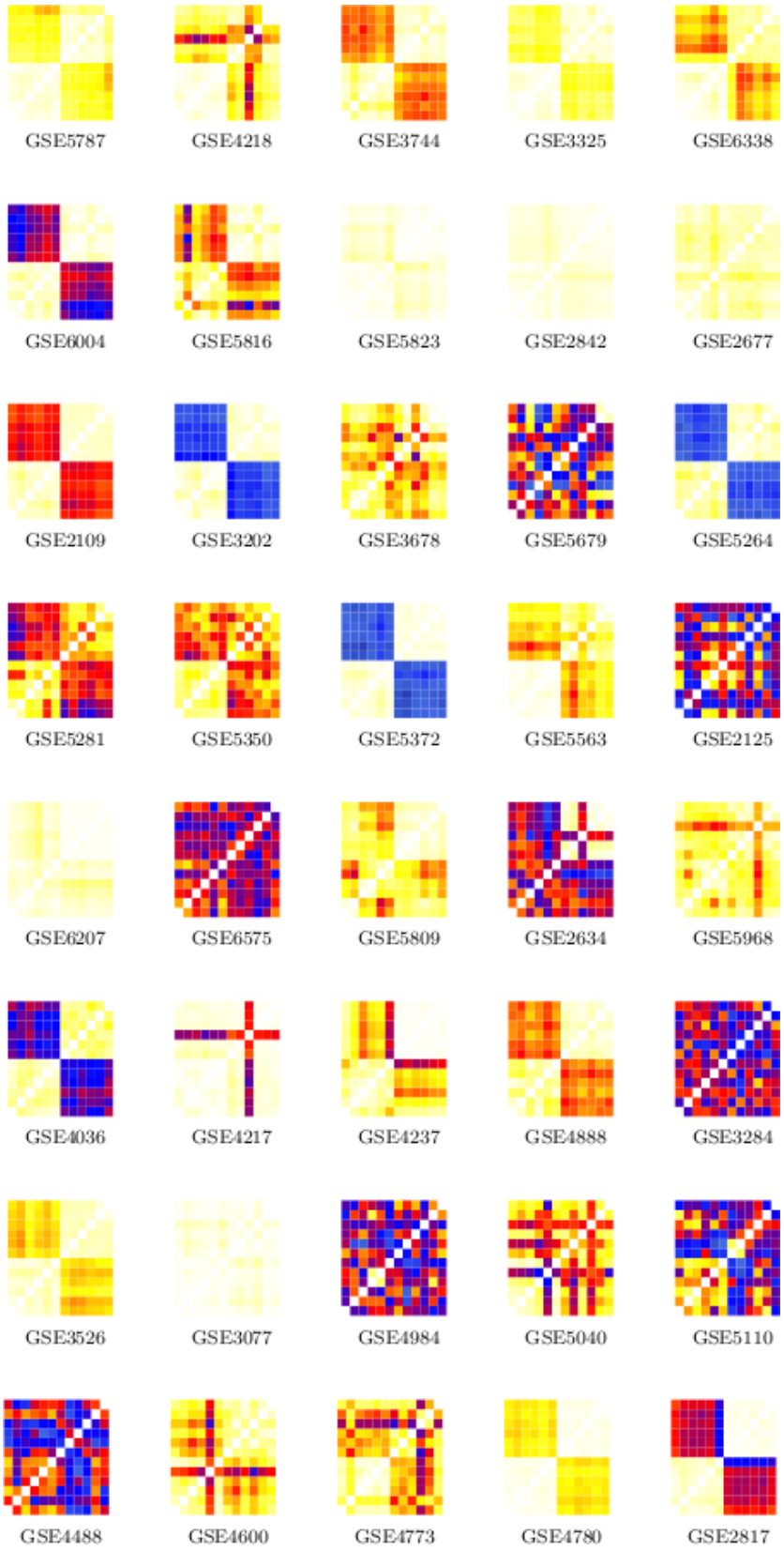


Table 2. Description of the GSE experiments

Experiment	Type	Cancer
GSE5787	cervical cancer	y
GSE4218	glioblastoma cancer cells in culture (different states)	y
GSE3744	breast cancer	y
GSE3325	prostate cancer and control	y
GSE6338	peripheral T-cell lymphoma and control	y
GSE6004	thyroid cancer and control	y
GSE5816	cancer cell lines, treatment and control	y
GSE5823	cancer cell lines, c-myc knockdown and control	y
GSE2842	childhood ALL, treated and untreated and controls	y
GSE2677	childhood ALL, treated and untreated and controls	y
GSE2109	Cancer in different tissues	y
GSE3202	non-small cell lung cancer cell lines, treatment and control	y
GSE3678	PTC (papillary thyroid carcinoma) and paired controls	y
GSE5679	dendritic cells (monocytes) ligand treatment and control	n
GSE5264	bronchial epithelial cells, differentiation time course	n
GSE5281	LCM-capture cells in Alzheimer's brain and normal controls	n
GSE5350	microarray quality control project, reference human RNA, reference human brain RNA and mixtures of the two	n
GSE5372	airway epithelial cells before and after injury	n
GSE5563	vulvar intraepithelial neoplasia and control	n
GSE2125	alveolar macrophages	n
GSE6207	human liver cell line (HepG2), transfected with miR-124, and controls	n
GSE6575	whole blood from children with autism and controls	n
GSE5809	endometrial stromal cells, treatment and control	n
GSE2634	human and non-human primate blood	n
GSE5968	HepG2 cell line transfected with PGC-1 transcription factor mutants, and controls	n
GSE4036	cerebellar tissues of schizophrenic patients and controls	n
GSE4217	spheroid formation and recovery of human foreskin fibroblasts at ambient temperature	n
GSE4237	pituitary adenomas (benign brain tumor)	n
GSE4888	endometrium sampled across the cycle in 28 normo-ovulatory women	n
GSE3284	blood leukocyte receiving inflammatory stimulus and controls	n
GSE3526	normal post-mortem tissue samples	n
GSE3077	dilution series of blood and placenta, comparison of Illumina and Affymetrix platforms	n
GSE4984	monocyte derived dendritic cells, treatment and control	n
GSE5040	lymphoblast cell lines from patients with Freidriech's ataxia and normal controls, treated and untreated	n
GSE5110	skeletal muscle biopsies from men before and after 48 h knee immobilization	n
GSE4488	blood from affected and obligatory carriers of pituitary adenoma predisposition (PAP) and controls	n
GSE4600	SH-SY5Y neuroblastoma cell line, undifferentiated, differentiated, transfected with MeCP2 decoy oligonucleotide and controls	n
GSE4773	SK-N-MC neuroblastoma cell line, treatment with rotenone and controls	n
GSE4780	benign (grade 1) and aggressive (grades 2 and 3) meningiomas	-
GSE2817	gliomas (brain tumors)	-

4 Conclusions

We find that for exons with only antisense probes mapping uniquely to them, the antisense probes are typically not correlated with each other. This suggests that the expression seen in each of these probes is not coherently detecting an antisense transcript. Just because strong signal is seen in a small fraction of the probes within an antisense probeset does not mean that the transcript is detected. Careful analysis of each probeset is required on a case by case basis.

For most of the cases where exons contain both sense and antisense probes, there is a SA-pair in which one transcript belongs to an established gene and the other is an EST or mRNA. In some cases there might be a novel antisense transcript since there are high correlations in the locus where the annotations indicate that there is not a transcript/RNA. In other cases there is a transcript in the antisense direction that does not cover the locus where the highly correlated probes align. This suggests that the transcript might be longer than the current annotation indicates. Care has to be taken when analysing the data considering that some probes might behave in different ways according to the type of experiment.

Our method for studying the correlations of sense and antisense Affymetrix probes has been shown to be useful for finding possible NATs, confirming existing ones, suggesting the existence of novel transcripts, or suggesting the reannotation of transcripts.

Acknowledgements

OSG 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/2005A/11996A). We are grateful to Dr. William Langdon for the development of a number of software tools used in this research.

References

[1] J. Mattick, RNA regulation: a new genetics?, Nature Reviews Genetics, 5:316, 2004.

[2] F. Perocchi, Z. Xu, S. Clauder-Münster and L. Steinmetz, Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. Nucleic Acids Research, 35:e128, 2007

[3] P. Galante, D. Vidal, J. de Souza, A. Camargo and S. Souza, Sense-antisense pairs in mammals: functional and evolutionary considerations. Genome Biology, 8:R40, 2007.

[4] Ø. Røsok and M. Sioud, Systematic search for natural antisense transcripts in eukaryotes. International Journal of Molecular Medicine, 15:197-203, 2005.

[5] Affymetrix Inc. 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.

[6] T. Barrett, D. Troup, S. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. Kim, A. Soboleva, M. Tomashevsky and R. Edgar, NCBI GEO: mining tens of millions of expression profiles – database and tools update. Nucleic Acids Research, 35: D760-D765, 2007.

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

[8] S. Oeder, J. Mages, P. Flicek and R. Lang R. Uncovering information on expression of natural antisense transcripts in Affymetrix MOE430 datasets. BMC Genomics, 8:200, 2007.

[9] A. Werner, G. Schmutzler, M. Carlile, C. Miles and H. Peters, Expression profiling of antisense transcripts on DNA arrays. Physiol. Genomics, 28:294, 2007.

[10] O. Sanchez-Graillet, J. Rowsell, W.B. Langdon, M. Stalteri, J. Arteaga-Salas, G. Upton and A. Harrison, Widespread existence of uncorrelated probe intensities from within the same probeset on Affymetrix GeneChips. Journal of Integrative Bioinformatics, 5(2):98, 2008.

[11] J. Arteaga-Salas, H. Zuzan, W. Langdon, G. Upton and A. Harrison, An overview of image processing methods for Affymetrix GeneChips. Briefings in Bioinformatics, 9(1):25, 2008.

[12] C. Wu, H. Zhao, K. Baggerly, R. Carta and L. Zhang, Short oligonucleotide probes containing G-stacks display abnormal binding affinity on Affymetrix microarrays. Bioinformatics, 23:2566-2572, 2007.

[13] G. Upton, W. Langdon and A. Harrison, G-spots cause incorrect expression measurement in Affymetrix microarrays. BMC Genomics, 9:613, 2008.

[14] M. Lapidot and Y. Pilpel, Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. *EMBO Rep*., 7(12):1216–1222, 2006.

[15] Y. He, B. Vogelstein, V. Velculescu, N. Papadopoulos and K. Kinzler, The Antisense Transcriptomes of Human Cells. Science, 322(5909):1855–1857, 2008.

[16] G. Upton and C. Lloyd, Oligonucleotide arrays: information from replication and spatial structure. Bioinformatics, 21:4162, 2005.

[17] W. Langdon, G. Upton, R. Camargo and A. Harrison, A survey of spatial defects in Homo Sapiens Affymetrix GeneChips. Transactions on Computational Biology and Bioinformatics, TCBB.2008.108, 2008.

[18] G. Upton, O. Sanchez-Graillet, J. Rowsell, J. Arteaga-Salas, N. Graham, M. Stalteri, F. Memon, S. May and A. Harrison, On the causes of outliers in Affymetrix GeneChip data. Brief Funct Genomic Proteomic, 8(3), 199-212, 2009.

[19] W. Kent, C. Sugnet, T. Furey, K. Roskin, T. Pringle, A. Zahler and D. Haussler, The Human Genome Browser at UCSC. Genome Res., 12:996-1006, 2002.

[20] Affymetrix Inc. Affymetrix GeneChip IVT Array Whitepaper Collection. Transcript Assignment for NetAffx Annotations. Revision Date: 2006-3-24. Revision Version: 2.3. Transcript_Assignment_whitepaper.pdf, 2006.

[21] Y. Zhang, J. Li, L. Kong, G. Gao, Q.R. Liu and L. Wei., NATsDB: Natural Antisense Transcripts DataBase. Nucleic Acids Res., 35(Database issue):D156-61, 2007.

[22] R. Yelin, D. Dahary, R. Sorek, E. Levanon, O. Goldstein, A. Shoshan, A. Diber, S. Biton, Y. Tamir, R. Khosravi, S. Nemzer, E. Pinner, S. Walach, J. Bernstein, K. Savitsky, G. Rotman, Widespread occurrence of antisense transcription in the human genome. Nat Biotechnol., 21(4):371-2, 2003.

[23] J. Chen, M. Sun, W. Kent, X. Huang, H. Xie, W. Wang, G. Zhou, R. Shi, J. Rowley. Over 20% of human transcripts might form sense-antisense pairs. Nucleic Acids

Research, 32(16):4812–20, 2004.

Supplementary material to “Using surveys of Affymetrix GeneChips to study antisense expression” by Olivia Sanchez-Graillet et al.

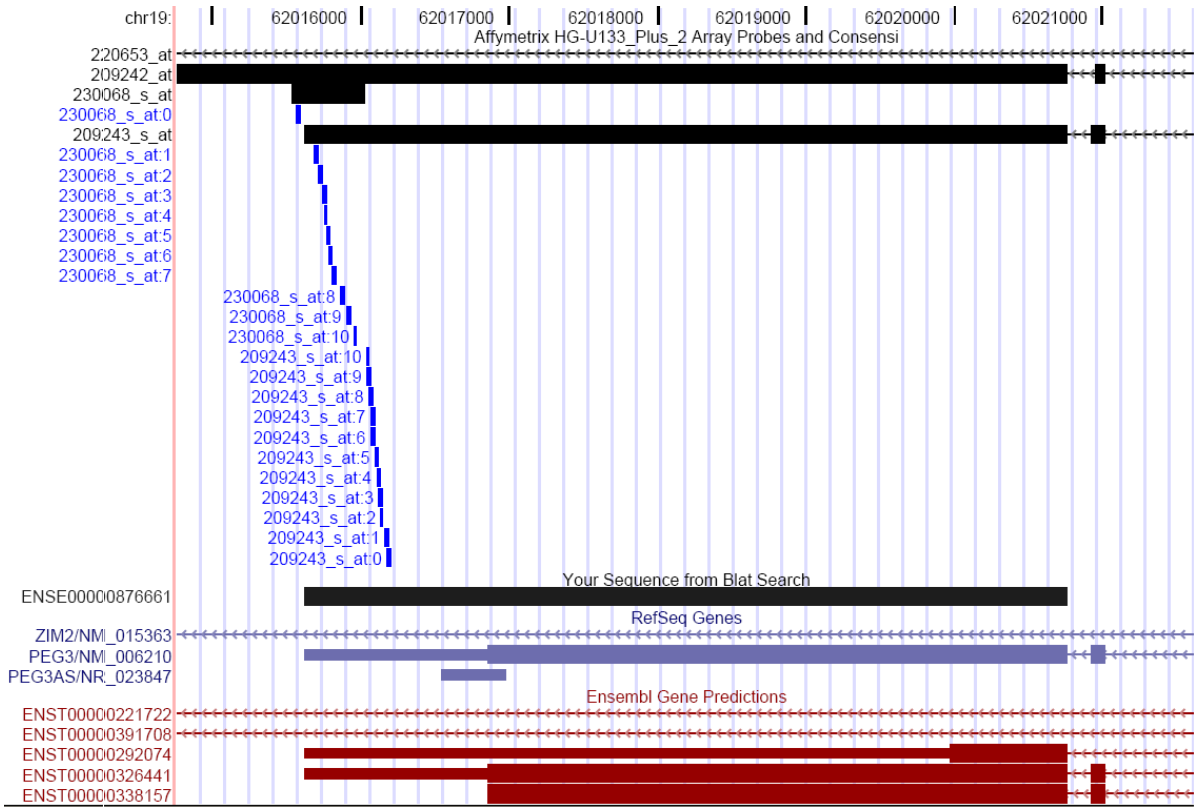
Possible explanation of antisense expression

In this supplementary section we discuss examples that follow the patterns observed in exons with groups of sense and antisense probes.

Exon ENSE00000876661, an example of pattern 1 (the sense and antisense probes are correlated with each other)

This exon is part of the Ensembl transcripts ENST00000326441 and ENST00000292074 (Ensembl gene ENSG00000198300, ZIM2, zinc finger, imprinted 2). Genes ZIM2 and PEG3 (paternally expressed gene 3), overlap on the reverse strand of chromosome 19 and share a number of exons [1]. Exon ENSE00000876661 is the 3' terminal exon of PEG3. Ensembl, however, annotates all the various transcripts as part of the ZIM2 gene. Overlapping exon ENSE00000876661 on the forward strand of chromosome 19, there is a gene PEG3AS, which produces an antisense transcript [2-3]. Glasgow et al. reported increased expression of both PEG3 and PEG3AS in rat vasopressin-magnocellular neurons (VP-MCNs) during systemic hyperosmolality. However, the antisense probes from probeset 230068_s_at are about 600 bp upstream from the 5' end of PEG3AS (Figure S1). It is possible that the 5' end of PEG3AS is further upstream than what has been documented to date, or that transcript variants with different 5' ends are produced in different tissues and that this is what the antisense probes are measuring.

Figure S1. UCSC screen shot for ENSE00000876661



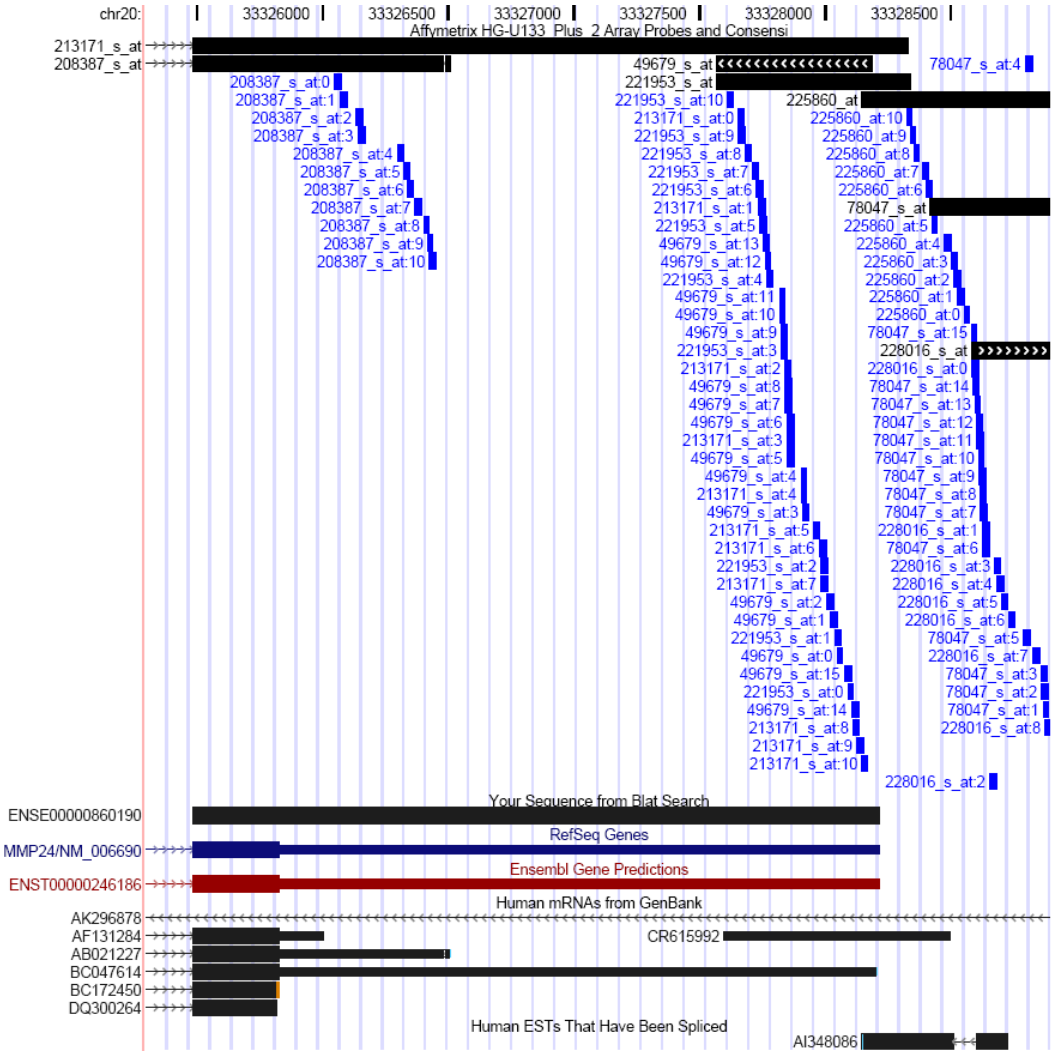
Exon ENSE00000860190, an example of pattern 2 (only the antisense probes are correlated)

In this exon several of the antisense probes overlap; sense (probesets 208387_s_at and 213171_s_at) and antisense probes (probesets 221953_s_at and 49679_s_at) overlap, except for probeset 208387_s_at, which is upstream from the other 3 probesets (Figure S2). The antisense transcript CR615992 that the antisense probes map to is not very reliable, i.e., the UCSC entry for the mRNA CR615992 carries the following warning: "*CR615992 is from the InVitroGen/Genoscope full-length library. Some of the entries associated with this dataset appear to have been aligned to the reference genome and the sequences subsequently modified to match the genome. This process may have resulted in apparent high-quality alignments to pseudogenes. Care should be taken in using alignments of this sequence as evidence of transcription.*"[1].

A lot of the antisense probes overlap, so are not independent, although the ones with the greatest overlap are not necessarily the most highly correlated ones. Some antisense probes have runs of 3 G's. It is not clear why the sense probes are not correlated.

[1] http://genome.ucsc.edu/cgi-bin/hgc?hgsid=131985429&o=33327601&t=33328502&g=mrna&i=CR615992&c=chr20&l=33327601&r=33328502&db=hg18&pix=620

Figure S2. UCSC screen shot for ENSE00000860190



Exon ENSE00001454677, an example of pattern 3 (only the sense probes are correlated)

Exon ENSE00001454677 is the only exon of the Ensembl transcript ENST00000371225 from gene ENSG00000184292, also known as TACSTD2 (Tumour-associated calcium signal transducer 2 Precursor[2]). TACSTD2 is also known as Pancreatic carcinoma marker protein GA733-1 and Cell-surface glycoprotein Trop-2. TACSTD2 is an intronless gene on the reverse strand of chromosome 1.

TACSTD2 encodes a carcinoma-associated antigen defined by the monoclonal antibody GA733[3]. High expression of TACSTD2 is associated with a poor prognosis in pancreatic cancer. A chimeric CYCLIN D1-TROP2 mRNA has been isolated from various types of cancer cells, in the absence of chromosome rearrangements [4]. Mutations of TACSTD2 result in gelatinous drop-like corneal dystrophy, an autosomal recessive disorder characterized by severe corneal amyloidosis leading to blindness.

[2] http://www.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000184292
[3] http://www.ncbi.nlm.nih.gov:80/nuccore/166795235?report=genbank

There are no entries in NATsDB that overlap this region of chromosome 1[4]. There are three probesets with probes that map uniquely to exon ENSE00001454677 (Figure S3). Probesets 202285_s_at and 202286_s_at align sense to the exon, while probeset 227128_s_at aligns antisense to the exon. The sense probes and the antisense probes do not overlap. Probeset 202285_s_at maps to the 5' UTR region of the exon, while probeset 202286_s_at maps to the 3' UTR[5].

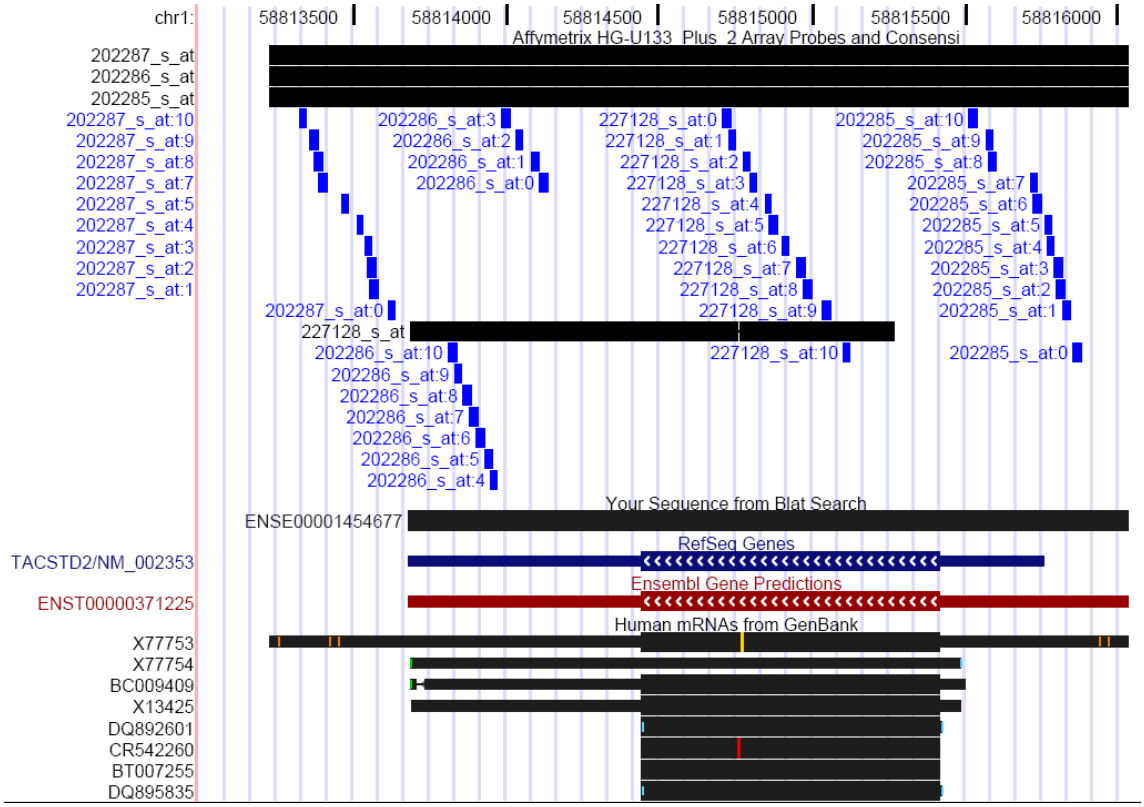
The probes which show strong correlations are the probes from probeset 202286_s_at, which map to the 3' UTR. Probeset 202285_s_at aligns about 1800 bp upstream (along the transcript) from probeset 202286_s_at. It is possible that probeset 202285_s_at is too far from the poly A tail of the transcript to be detected by the assay used with the Affymetrix 3' expression arrays, which uses a poly dT primer and reverse transcriptase[6].

The antisense probeset, 227128_s_at, maps antisense to the CDS[7]. There are no annotated genes or mRNAs antisense to exon ENSE00001454677. Therefore it is likely that the antisense probes are not correlated because they are not measuring anything.

Probeset 227128_s_at is assigned to TACSTD2 by Affymetrix, but it is given an E grade annotation (NetAffx release 28, Mar. 16, 2009). The probeset design was based on ESTs, and the Affymetrix annotation pipeline assigns it to an EntrezGene identifier based on the UniGene cluster. The ESTs were included in at the time of array design (Affymetrix annotation whitepaper).

[4] http://natsdb.cbi.pku.edu.cn/download/Latest_Release/sa_hs
[5] http://genome.ucsc.edu/cgi-bin/hgTracks; Affymetrix HG-U133_Plus_2 Array Probes and Consensi
[6] Affymetrix, Inc. 2001. Array design for the GeneChip® human genome U133 set. Technical note. Part No. 701133 Rev. 1. http://www.affymetrix.com/support/technical/technotes/hgu133_design_technote.pdf
[7] http://genome.ucsc.edu/cgi-bin/hgTracks; UCSC Genome Browser on Human Mar. 2006 Assembly

Figure S3. UCSC screen shot for ENSE00001454677



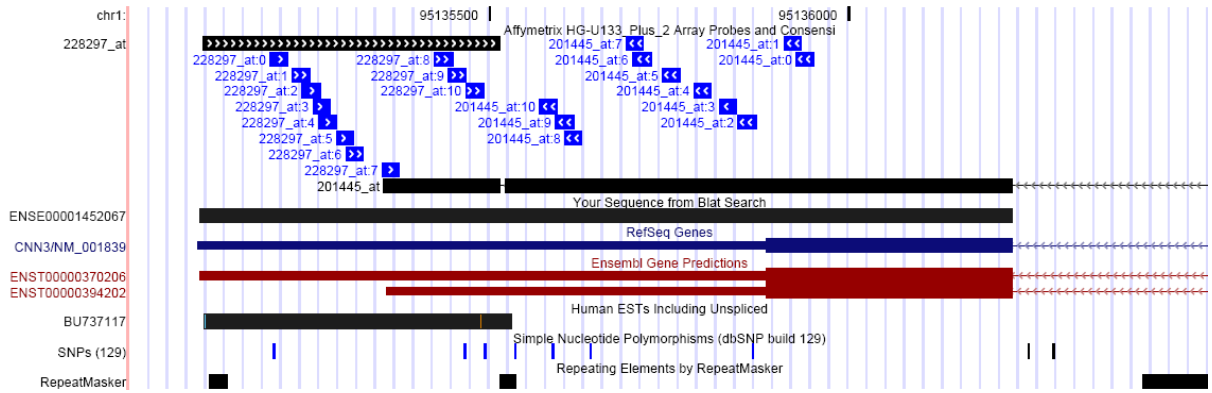
Exon ENSE00001452067, an example of pattern 4 (the sense and antisense probes are not correlated, the sense probes are correlated and the antisense probes are correlated)

Exon ENSE00001452067 is the 3' terminal exon of transcript ENST00000370206 from gene ENSG00000117519, CNN3[8], Calponin-3. CNN3 is on the reverse strand of chromosome 1.

The antisense probes map to the EST BU737117, Figure S4. The sense probes map to NM_001839 (CNN3). The SA(sense-antisense)-pair of NM_001939 and BU737117 is listed in NATsDB. The antisense probes are correlated because they are detecting expression of BU737117.

[8]http://www.ensembl.org/Homo_sapiens/Transcript/Exons?db=core;g=ENSG00000117519;r=1:95135097-95165294;t=ENST00000370206

Figure S4. UCSC screen shot for ENSE00001452067



Exon ENSE00001222306, an example of pattern 5 (probes are not correlated or are negatively correlated)

This exon is the 5' exon of the Ensembl transcript ENST00000320761 (gene ENSG00000179093, AC009503.3-201[9]). AC009503.3-201 is a novel pseudogene on the reverse strand of chromosome 2. It has two exons, which are separated by an intron that is just 2 bp long. AC009503.3-201 is on the opposite strand from an intron in gene NCK2, which is transcribed from the forward strand of chromosome 2 [5-6][10][11]. NCK2 is a member of the NCK family of adapter proteins. The protein contains three SH3 domains and one SH2 domain, and has been shown to bind and recruit various proteins involved in the regulation of receptor protein tyrosine kinases[12]. Recent work suggests that NCK2 may be associated with a genetic predisposition for normal tension glaucoma [7].

The antisense probes align to an intronic region of NCK2, so presumably there is no transcription from that region on that strand. The sense probes align to gene AC009503.3-201 and exon ENSE00001222306 (Figure S5). The sense probes and antisense probes overlap. The consensus sequences of the two probesets span an identical region, but on opposite strands. This suggests Affymetrix was not sure which strand the transcript was from, and so tiled probesets on both strands. Affymetrix labels probeset 234705_at as an A-grade probeset (Affymetrix NetAffx release 28, March 16, 2009), and assigns it to the transcript AF083117. Probeset 234595_at is also assigned to the mRNA AF083117 by Affymetrix, but it is labelled as an E-grade annotation (Affymetrix NetAffx release 28, March 16, 2009) with the warning that "*This assignment is strictly based on mapping accession IDs from the original UniGene design cluster to the latest UniGene design cluster*"[13].

[9] http://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000179093

[10] Ensembl 48

[11] http://www.ensembl.org/Homo_sapiens/Transcript/Exons?db=core;g=ENSG00000179093;r=2:105865912-105866276;t=ENST00000320761

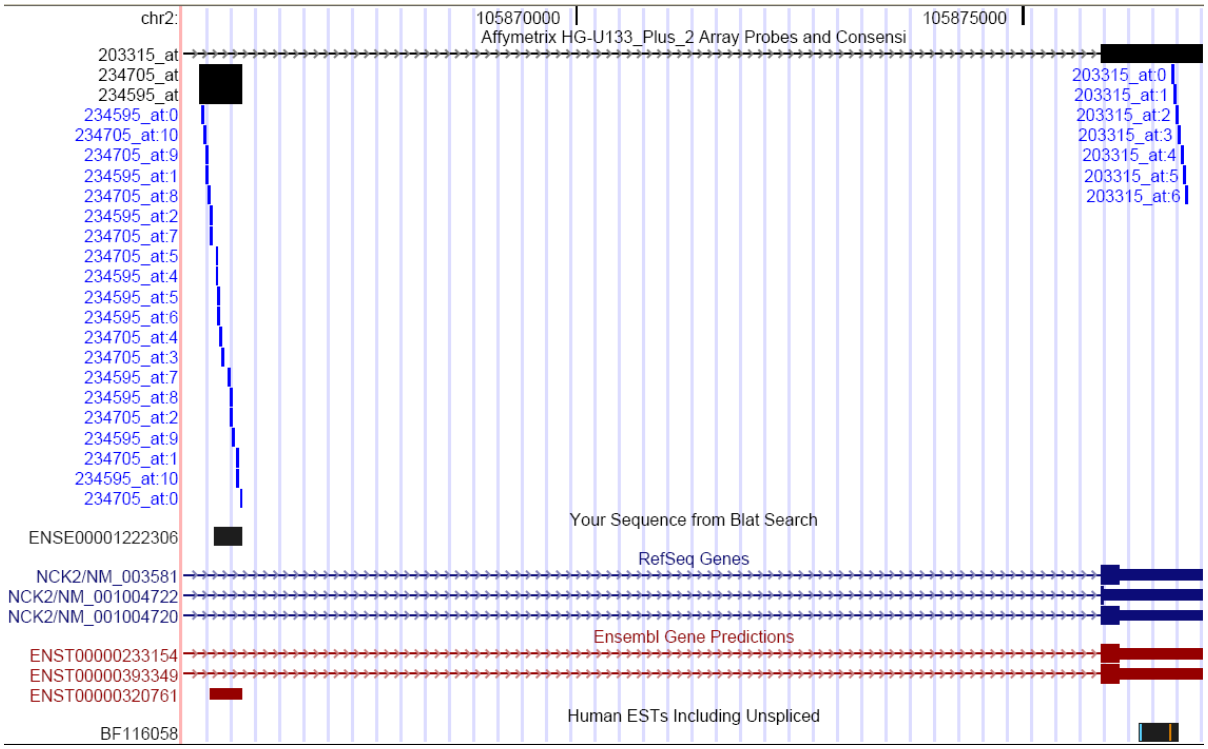
[12] http://www.ncbi.nlm.nih.gov/nuccore/52630422?report=GenBank

[13] https://www.affymetrix.com/analysis/netaffx/fullrecord.affx?pk=HG-U133_PLUS_2%3A234595_AT

NATsDB also includes a sense-antisense pair involving NCK2, NCK2 and the EST BF116058. However, BF116058 is antisense to the 3' terminal exon of NCK2, and does not overlap with exon ENSE00001222306 or either of the two probesets.

In conclusion, the antisense probes are not correlated because they are not detecting any signal. The fact that the sense probes are not correlated would suggest that the pseudogene they align to, AC009503.3-201, is not being expressed.

Figure S5. UCSC screen shot for ENSE00001222306



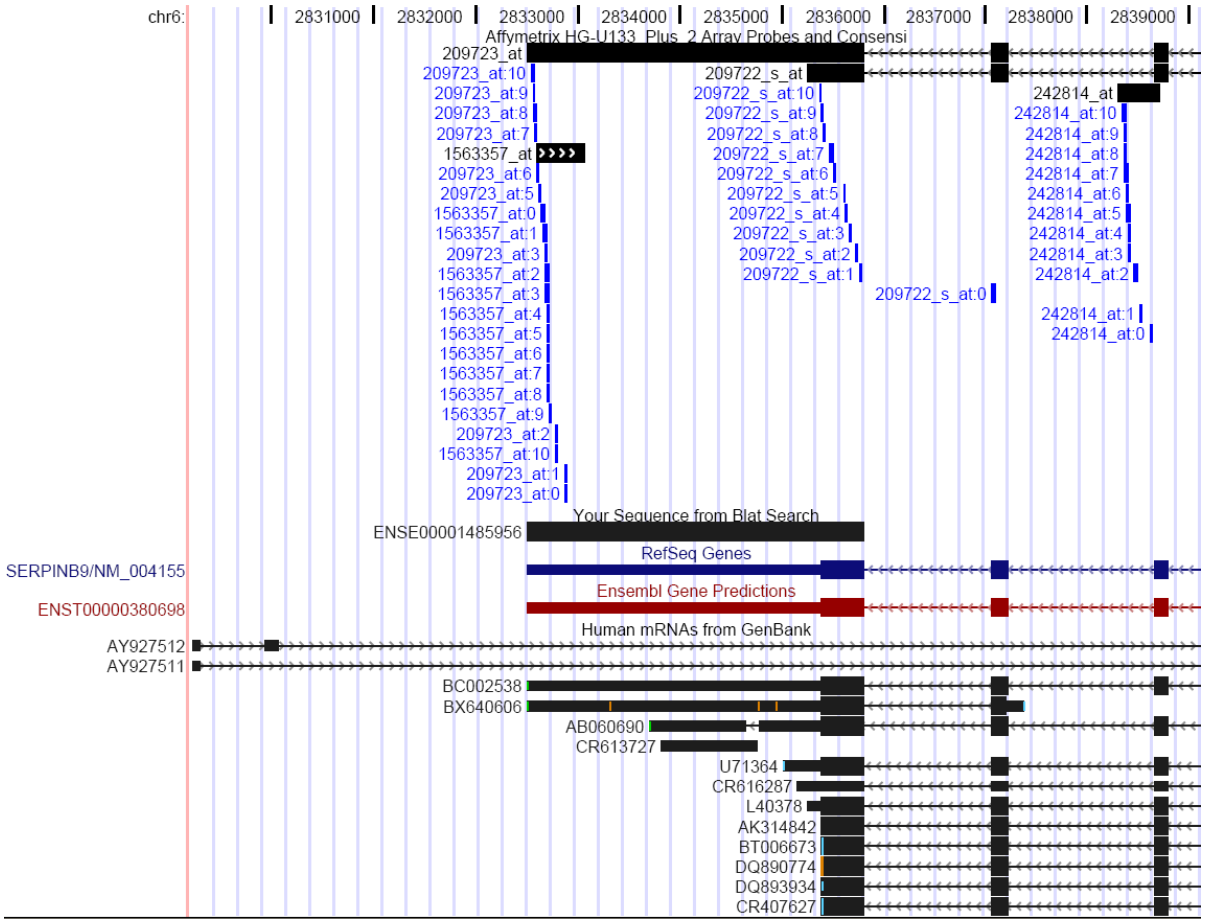
Exon ENSE00001485956, an example of pattern 6 (antisense probes are correlated with a sub-group within the sense probes)

ENSE00001485956 is the 3' terminal exon of the Ensembl transcript ENST00000380698 (Ensembl gene ENSG00000198300, SERPINB9). SERPINB9 belongs to the large superfamily of serine proteinase inhibitors (serpins), which bind to and inactivate serine proteinases. SERPINB9 is on the reverse strand of chromosome 6. The mRNAs AY927511 and AY927512 are on the forward strand of chromosome 6, antisense to SERPINB9, but the region of these transcripts overlapping exon ENSE00001485956 is an intron. AY927511 and AY927512 are novel transcripts, which were discovered using Affymetrix tiling arrays [8]. The probes from the sense probeset 209722_s_at, which are not correlated with the other sense and antisense probesets, map to the region of exon ENSE00001485956 which is in the CDS, and do not overlap the other two probesets. The probes from the sense probeset 209723_s_at and the antisense probeset 1563357_at map to the region of exon ENSE00001485956 which is in the 3' UTR (Figure S6). The probes from the sense probeset 209723_s_at and the antisense probeset 1563357_at overlap. Furthermore, the antisense probe 1563357_s_at pm11 and the sense probe 209723_at pm3 have a complementary region spanning 23 of the 25 bases. However, the correlation between these two probes is not that

strong (0.5). The antisense probes from probeset 1563357_s_at also have a large degree of overlap with each other, with each of probes 1563357_s_at pm4 – pm9 is offset from the previous probe by only one or two bases. Probe 1563357_s_at pm11, which is not as highly correlated with the sense probes, has two runs of GGG. The correlated sense and antisense probes all have runs of T's. The sense probes from probeset 209722_s_at measure transcription from the part of exon ENSE00001485956 which is in the coding sequence, whereas the sense probes from probeset 209723_s_at measure transcription from the 3' UTR.

The lack of correlation between the two probesets suggests that perhaps there may be an alternative polyA site between the two probesets. It is difficult to explain what the antisense probes are measuring, since they align to an intronic region of the AY927511 and AY927512 transcripts.

Figure S6. UCSC screen shot for ENSE00001485956



Supplementary references

[1] J. Kim, A. Bergmann and L. Stubbs, Exon sharing of a novel human zinc-finger gene, ZIM2, and paternally expressed gene 3 (PEG3). Genomics, 64(1):114-118, 2000

[2] E. Glasgow, S.L. Ryu, M. Yamashita, B.J. Zhang, N. Mutsuga, and H. Gainer, APeg3, a novel paternally expressed gene 3 antisense RNA transcript specifically expressed in

vasopressinergic magnocellular neurons in the rat supraoptic nucleus., Brain Res. Mol. Brain Res., 137(1-2):143-151, 2005

[3] J.H. Choo, J.D. Kim and J. Kim, Imprinting of an evolutionarily conserved antisense transcript gene APeg3., Gene, 409(1-2):28-33, 2008

[4] E. Guerra, M. Trerotola, R. Dell' Arciprete, V. Bonasera, B. Palombo, T. El-Sewedy, T. Ciccimarra, C. Crescenzi, F. Lorenzini, C. Rossi, G. Vacca, R. Lattanzio, M. Piantelli and S. Alberti, Cancer Res., 68:8113-21, 2008

[5] W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler and D. Haussler, The Human Genome Browser at UCSC. Genome Res., 12:996-1006, 2002

[6] W.J. Kent, BLAT - The BLAST-Like Alignment Tool. Genome Res., 12:656-664, 2002

[7] M. Akiyama, K. Yatsu, M. Ota, Y. Katsuyama, K. Kashiwagi, F. Mabuchi, H. Iijima, K. Kawase, T. Yamamoto, M. Nakamura, A. Negi, T. Sagara, N. Kumagai, T. Nishida, M. Inatani, H. Tanihara, S. Ohno, H. Inoko and N. Mizuki. Microsatellite analysis of the GLC1B locus on chromosome 2 points to NCK2 as a new candidate gene for normal tension glaucoma. Br J. Opthalmol. 92:1293-6, 2008

[8] J. Cheng, P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, S. Patel, J. Long, D. Stern, H. Tammana, G. Helt, V. Sementchenko, A. Piccolboni, S. Bekiranov, D. Bailey, M. Ganesh, S. Ghosh, I. Bell, D. Gerhard and T. Gingeras. Transcriptional Maps of 10 Human Chromosomes at 5-Nucleotide Resolution. Science, 308(5725):1149–1154, 2005