LOOK ACADEMIC PUBLISHERS
OPEN ACCESS

INTERNATIONAL JOURNAL OF ENVIRONMENTAL & SCIENCE EDUCATION
2016, VOL. 11, NO. 18, 12985-13000

Antisense Expression Repertoire of the Human Brain

Haiyang Hu[a], Xi Jiang[a], Ning Fu[b] and Philip Khaytovich[a, b, c]

[a]Shanghai Institute for Biological Sciences, CAS, Shanghai, CHINA;
[b]Skolkovo Institute for Science and Technology, Skolkovo, RUSSIA;
[c]TheImmanuel Kant Baltic Federal University, Kaliningrad, RUSSIA.

ABSTRACT

Gene expression is the most fundamental level at which the genotype gives rise to the phenotype. Therefore, when and where, why and how a gene would be expressed is of the upmost importance during biology process. One type of expression – the antisense transcription – occurs between a pair of genes that are encoded in an overlapping and opposite orientation (sense and antisense gene pair, SAS pair).

KEYWORDS
Gene; human brain; sequencing (RNA-seq)

ARTICLE HISTORY
Received 10 July 2016
Revised 22 October 2016
Accepted 5 November 2016

Introduction

Antisense transcripts were first detected in viruses, then in prokaryotes and later eukaryotes (Alfano et al., 2005; Torarinsson et al., 2006; Havgaard et al., 2005; Washietl & Hofacker, 2004; Siepel et al., 2005). With vast accumulation in expression data and advancements in high-throughput methods, especially those coupled with microarray and massively parallel sequencing, widespread existence of antisense has been recognized and reported in many species, including humans, mice, rats and chickens (Babak et al., 2007; Pollard et al., 2004; Kent, 2002), nematodes, Arabidopsis and yeast. On average, 7-30% of all genes are associated with antisense transcripts in plants and animals. And it's estimated that up to 72% of the transcripts have been demonstrated to have antisense partners in human and mouse transcriptomes.

However, as widespread as antisense transcripts are, only a small subset of antisense transcripts has been verified of possessing regulatory roles, such as X-inactivation, genomic imprinting, DNA methylation, RNA editing, and alternative splicing (Carninci et al., 2005; 2006). Recent genome-wide studies using SAGE and microarrays showed coordinated expression between sense and antisense transcription pairs (SAS pair) in human and mouse. The expression level of SAS pair decreases as the length of overlapped region increases. Exon splicing is strongly correlated to antisense gene expression (Carvunis et al., 2012; Bánfai et al., 2012; Derrien et al., 2012; Geisler et al., 2012; Georg & Hess, 2011; Grinchuk et al., 2010; Jeon et al., 2012).

Until now, studies on antisense transcripts are still very much limited compared to their prevalent existence, and certainly no studies have investigated the antisense transcription in human prefrontal cortex with age-series (Sluka et al., 2002; Griffiths-Jones et al., 2006; Argaman et al., 2001; Wassarman et al., 2001; Clote et

CORRESPONDENCE Haiyang Hu ✉ oceanhu@126.com

al., 2005; Uzilov et al., 2006; Washietl et al., 2005; di Bernardo et al., 2003). We therefore set out to explore the repertoire of antisense expression of human brain, with the objectives of assessing and characterizing overall transcriptional state of antisense, searching for potential origin, evaluating underlying effect, and hopefully providing useful information for biologist interested in verifying the actually functions of antisense transcripts.

Results and Discussion

Widespread antisense transcription

To assess the overall status of antisense transcription in human brain, we exploited the strand-specific poly-A+ transcriptome in a total of 14 individuals, with ages ranging from 2 days to 98 years, using high-throughput transcriptome sequencing (RNA-seq) on the Illumina platform. The sequencing was carried out for the prefrontal cortex (PFC) with read length 100 nucleotides (Table 1).

Table 1. Sample information.

Sample	Index	Ages		Sex	PMI*	RIN**	Ethnicity	Cause of death
		Year	Day					
2 days	1	0	2	M	3	8	Caucasian	Prematurity
4 days	2	0	4	m	5	8.8	African American	congenital heart defect
19 days	3	0	19	f	14	7.1	Caucasian	pneumonia associated with meconium aspiration
34 days	4	0	34	m	7	7.9	Caucasian	idiopathic pulmonary hemorrhage
94 days	5	0	94	m	12	7.7	Caucasian	bronchopneumania
204 days	6	0	204	m	6	8.4	African American	sudden infant death syndrome
443 days	7	1	78	m	19	7.6	African American	Asthma
787 days	8	2	57	f	21	7.5	African American	acute myocarditis
5105 days	9	13	360	m	13	8.3	caucasian	hanging
9277 days	10	25	152	m	19	9.2	African American	Asthma
19457 days	11	53	112	m	17	8.3	caucasian	ASCVD
24090 days	12	66	0	m	10	8.6	NA	ruptured abdominal aneurysm aorta
32120 days	13	88	0	m	7	7.7	NA	euthanasia
35770 days	14	98	0	m	9	7.3	NA	cardiac tamponade due to bleeding from aorta fissure

* Postmortem intervals in hours

** RNA Integrity Number determined by the Agilent Bioanalyzer assay.

In total, there were 274,927,771 reads with the average sample coverage of ~21 million and we obtained ~75% mapping uniqueness by aligning them to the human genome together with annotated junction by PalMapper by allowing up to 4 mismatches (REF & Methods, Table 2). With regard to mapping strand, by combining total 14 samples and based on relatively better-annotated gene type – protein-coding genes– from Ensembl V59, we obtained 196,335,989 sense and 11,871,650 antisense reads altogether with S/AS ratio around 16.5 when including

introns, while excluding introns, there were 170,251,420 sense and 2,993,616 antisense reads with S/AS ratio close to 57. Genes with sense expression >= 0.1 (RPKM, Methods) were considered expressed. Under this criterion we were left with 18137 protein-coding gene. Noted that other than protein-coding gene 3 other major categories containing long RNA transcripts were also included for the downstream analysis: 4601 pseudogenes, 4818 processed transcripts, and 851 lincRNAs (see Figure 1).

Table 2. Numbers of sequence reads.

Sample	Index	Brain Region	Total reads	Uniquely mapped reads	Unique %
2 days	1	PFC	21277649	13092207	61.53%
4 days	2	PFC	21284713	14379727	78.49%
19 days	3	PFC	20754409	11843993	73.28%
34 days	4	PFC	23722421	16285468	78.37%
94 days	5	PFC	23416250	15119297	75.59%
204 days	6	PFC	22698303	14764891	76.50%
443 days	7	PFC	23934412	16143315	76.93%
787 days	8	PFC	17759057	11401664	74.97%
5105 days	9	PFC	19901399	12479373	71.61%
9277 days	10	PFC	23201284	15203403	76.32%
19457 days	11	PFC	16019209	10396263	72.55%
24090 days	12	PFC	20948595	14199806	76.63%
32120 days	13	PFC	21032459	14104142	75.82%
35770 days	14	PFC	20255260	13994072	77.17%
Total	--	--	296205420	221601810	74.81%

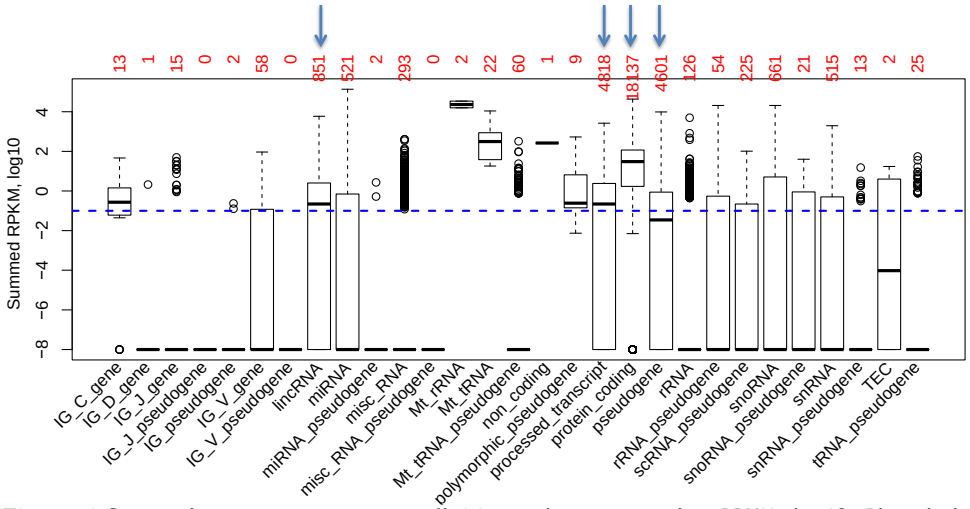


Figure 1. Summed gene expression overall 14 samples measured in RPKM, log10. Blue dash line represents the general expression cutoff at 0.1 (pseudo expression level 10-8 is added for smooth logarithmic transformation). X-axis indicates gene's biotype (Ensembl v59) with gene count after expression cutoff shown above in red.

Reads that mapped to the antisense strand of protein-coding gene (PCG) are marked as antisense reads. Combing all samples, PCG with at least 20 reads was defined as antisense-expressed protein-coding gene (ASE-PCG), which in total we collected 7636. In order to make sure that the mapped antisense reads were naturally transcribed other than artifacts generated during library preparation, PCR

cycling, etc. We first measured the expression correlation of sense and antisense of each ASE-PCG based on the assumption that antisense artifacts should grow proportional to the actual abundance of sense product. As a result, we found no correlation in any sample, while clearly PCG with high sense expression didn't necessarily have correspondingly high antisense expression (Figure 2).

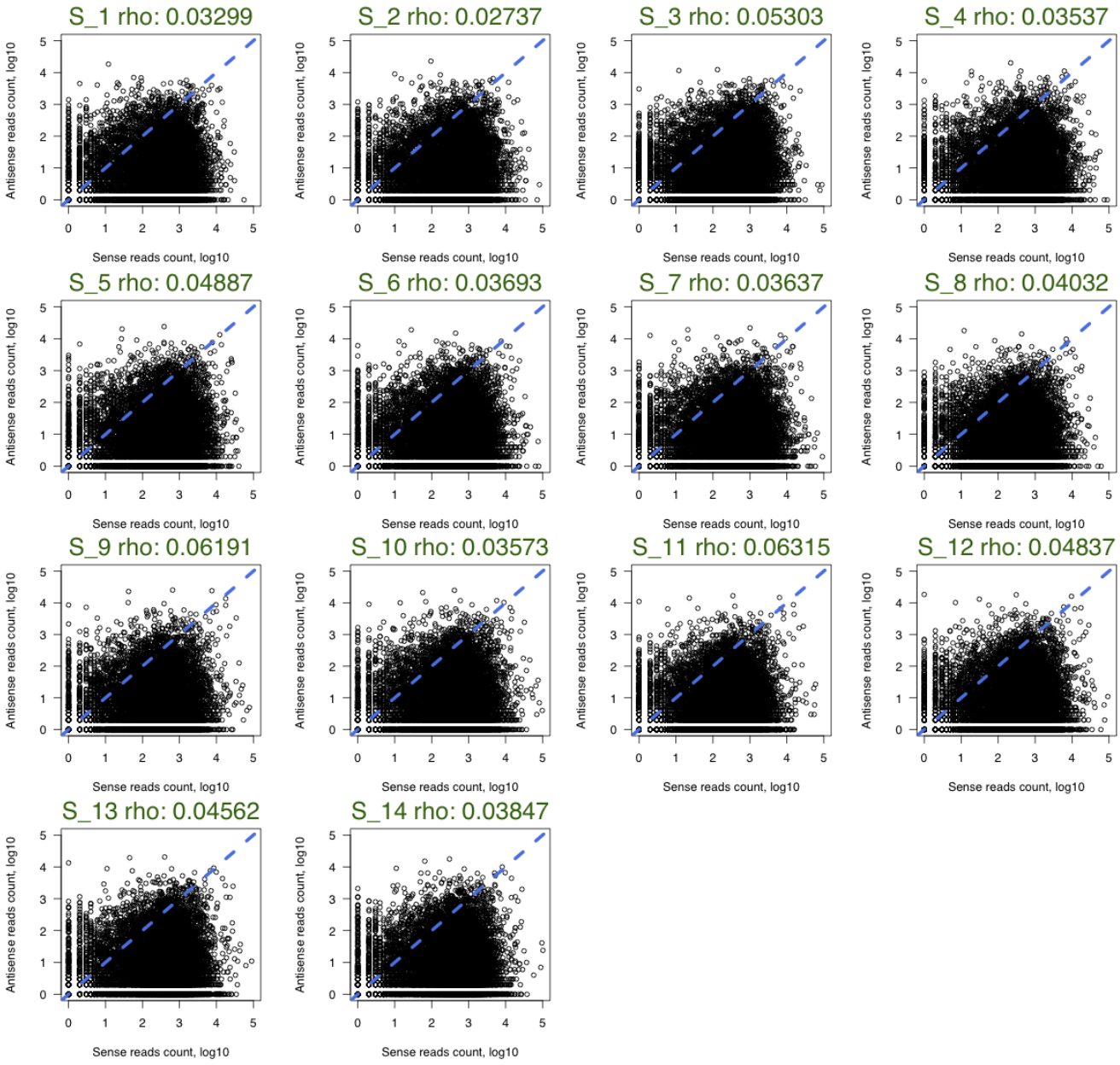


Figure 1. Expression correlation of protein-coding gene's sense/antisense. X and y axis indicates the sense and antisense reads count (log10) of protein coding gene. Green title shows sample index & spearman's correlation *rho*.

Second, as a supplement to the first, we resorted to examine junction reads with the notion that artificial antisense transcription, if it existed, would unavoidably came across the sense junction which in turn should generate reads sequence reverse

complement to that of real splicing sites (for example, GT-AG), but again we barely found any antisense reads showing that signature (Table 3). Therefore antisense reads found mostly came from bona fide transcription.

Table 3. Splice site sequence distribution over sense/antisense junction.

Sample XX			
#	Sequence	Frequency	Proportion
1	GTAG*	1.46E+06	0.990349
2	GCAG*	11922	0.00811089
3	ATAC*	1347	0.00091640
4	ATAT	133	9.05E-05
5	ATAG	123	8.37E-05
6	GTTG	99	6.74E-05
7	TTAG	72	4.90E-05
8	GTGG	51	3.47E-05
9	GGAC	48	3.27E-05
10	CTAC**	40	2.72E-05
11	GAAG	30	2.04E-05
12	GTCC	30	2.04E-05
13	GCCG	23	1.56E-05
14	GTTA	23	1.56E-05
15	GTCT	18	1.22E-05
16	GAAC	17	1.16E-05
17	GTAT**	17	1.16E-05
18	GTTT	17	1.16E-05
19	ATAA	13	8.84E-06
20	AAAG	10	6.80E-06
…	…	…	…
/	CTGC**	0	0

* Red coloring indicates canonical splice site sequence
** Grey coloring indicates reverse complement to canonical splice site sequence

After verifying the authenticity of antisense reads, we continued to check whether these reads were scatted randomly along the sense gene's locus or would cluster together and perhaps close enough to represent potential transcription unit. To do that, we randomly re-distribute the same number of antisense reads into the host gene's locus 1000 times, then compare between the neighboring distance of real antisense reads and that of the redistributions. If the distance distribution in real case is significantly smaller than random cases at least 99% of the time during the 1000 redistributions, the ASE-PCG is considered as containing clustered antisense reads, otherwise it was excluded from the downstream analysis. Then the mean distance of neighboring reads from the 1000 redistributions is used to connect real neighboring antisense reads along the locus. The resulting reads clusters are defined as antisense regions (ASR). But noted that ASR with less than 10 reads (combining all 14 samples) is also excluded.

Accordingly, we obtained 27640 ASR with median length ~660nt from 6404 ASE-PCG, each containing 3 ASR on average (figure 3A, B). In order to check the effectiveness of ASR definition -- whether they captured majority of antisense reads within that sense gene -- for each ASE-PCG we examined the proportion of antisense reads from its ASR to total antisense reads count and the proportion of total genomic length occupied by its ASR to entire length of the gene. As a result, we found most

genes' ASR contained large proportion of total antisense reads while only taking up small proportion of genomic length compared to the whole gene (figure 3C).

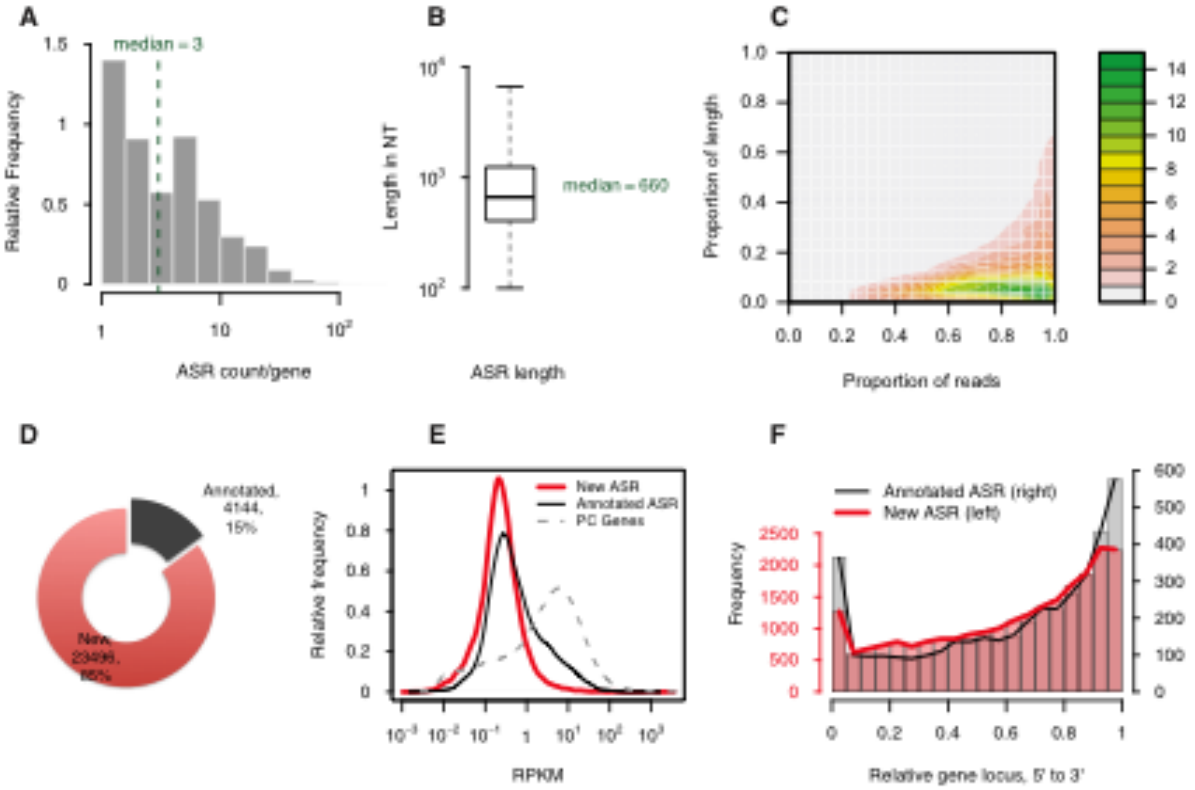


Figure 3. Characteristics of ASR. (A) ASR count distribution per/ASE-PCG, with median count 3. (B) ASR length distribution, with median length ~660nt. (C) 2D density plot showing the effectiveness of ASR definition. For each ASE-PCG, x-axis indicates the proportion of antisense reads from its ASR to total antisense reads; y-axis indicates proportion of total genomic length occupied by its ASR to length of the gene. The more density in the down-left corner, the more clustered and effective for antisense reads and defined ASR. (D) Proportion of annotated & new ASR. (E) Expression level of ASR measured in RPKM by combining all samples, red – new ASR; black – annotated ASR; grey-dash – protein-coding gene. (F) Relative genomic location of ASR, red – new ASR; Black (grey) – annotated ASR.

Extensions provide potential transcriptional origin for ASR

Antisense reads of the sense gene could either originate from un-annotated transcription unit, or come from neighboring genes of another strand with an extended 3' or 5' end, and more straightforward there were already annotated overlapping genes sitting on different strands for which ASR coming from the overlapped region could be assigned to the corresponding gene with absolute certainty. Here, out of 27640 ASR, 4144 (15%) came from annotated overlapped gene pairs and hence called annotated ASR, the rest (23496, 85%) were marked as new. Both ASR showed less expression level than protein-coding genes, which was expected considering inaccurate boundary definition. And although new ones have even less precise boundary, they exhibited expression level comparable to annotated

ones (figure 3D, E). As for genomic location, ASR basically spread the entire gene's locus, with more enrichment at both ends (figure 3F).

Furthermore, we separated ASR into 3 categories: potential 3' extension, 5' extension & internal transcription. For a new ASR, unlike an annotated one that had unambiguous origin, it could be the 3' extension of one gene, or the 5' extension of another gene, and certainly internal for the gene from whose locus it is defined. Noted that here by extension we only took the closest gene sitting on another strand to the ASR-containing sense gene into account. These two genes then form a pair and depending on the genomic positions and strands configuration of the pair we could easily define extension type to be 3' or 5'. For convenience, of a Sense-Antisense gene pair (SAS gene pair), we called the gene where ASR was defined as the sense gene, and the counterpart on another strand the antisense gene.

For 3' extension type, there were 2343 annotated and 23496 new ASR. Benefitting from the age-series dataset, we could conveniently check the expression correlation between new ASR and antisense gene under the assumption that different parts of the same gene should have consistent expression tendency, thus produce reasonable correlation. As negative control, we replace the antisense gene with another one, which was still the closest to the sense gene and had similar genomic location as the antisense gene (for instance, both downstream to the sense gene), except that it's on the same strand as the sense one. Be aware that annotated ASR represent unambiguous expression of the antisense gene, of which we took an advantage and used the expression correlation between annotated ASR and the rest of this antisense gene as positive control. Based on the controls, we indeed found that, although the correlation between new ASR and their antisense pair was not as good as positive control, it was significantly more positively correlated than negative ones (figure 4A). Considering the fact new ASR were still lowly expressed than annotated ones, which could lead to more noise hence worse correlation, we subsampled 100 times from new ASR so that they would have similar expression distribution to annotated ASR. Doing that, we did confirm the consistent existence of significant correlation shift ($p < 0.01$, figure 5A). And averaging from subsampling we observed an expected improvement over original signal (figure 4B). Though the distance between new ASR to their antisense pair was inevitably longer than annotated ones, we could see the shift at different distance cutoffs nonetheless (figure 5). What's more, when taking novel junction reads discovered by Tophat into consideration, we could see a much better correlation, even comparable to annotated ones, between new ASR and their antisense pairs if they could be connected by these junction reads (Figure 4A, B).

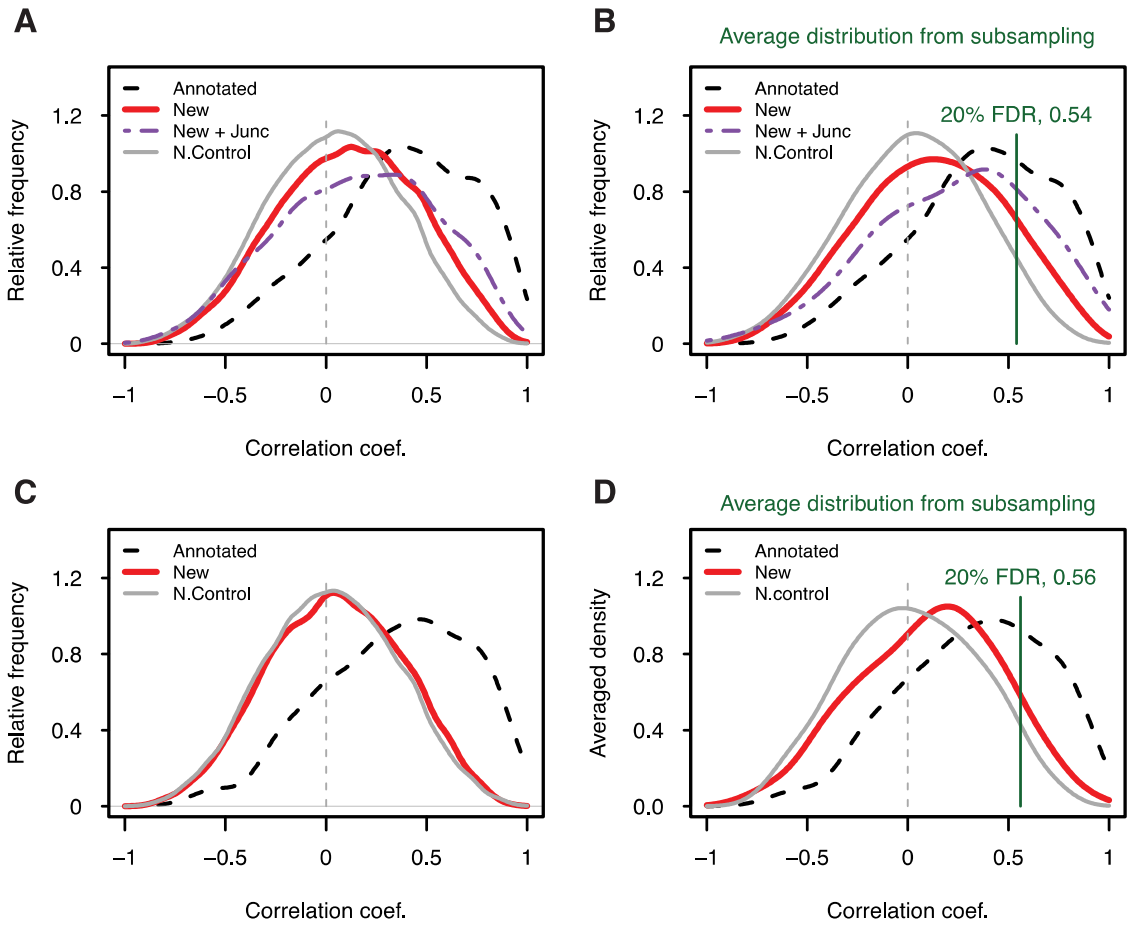


Figure 4. Expression correlation of ASR and antisense gene across aging. (A) Expression correlation of ASR and antisense gene across age in 3' extension, red - new ASR; black (dashed) - annotated ASR, positive control; purple (two dashed) - novel junction supported new ASR; grey - negative control, all measured using "spearman" correlation. (B) Curves have same meanings as (A), but are averaged from 100 subsampling to balance the expression level between new & annotated ASR, green vertical line indicates the correlation cutoff at 20% FDR. (C-D) Expression correlation in 5' extension, have same meanings as (A-B), except that no novel junction support is found.

Since it would be impossible that every new ASR and its antisense pair could be supported by novel junction, we sorted out to set a correlation cutoff at 0.56 which gives 20% FDR and 2635 new ASR as potential 3' extension to 1233 antisense genes (figure 4B). As expected, these correlated ASR have more novel junction support compared to total (Figure 5).

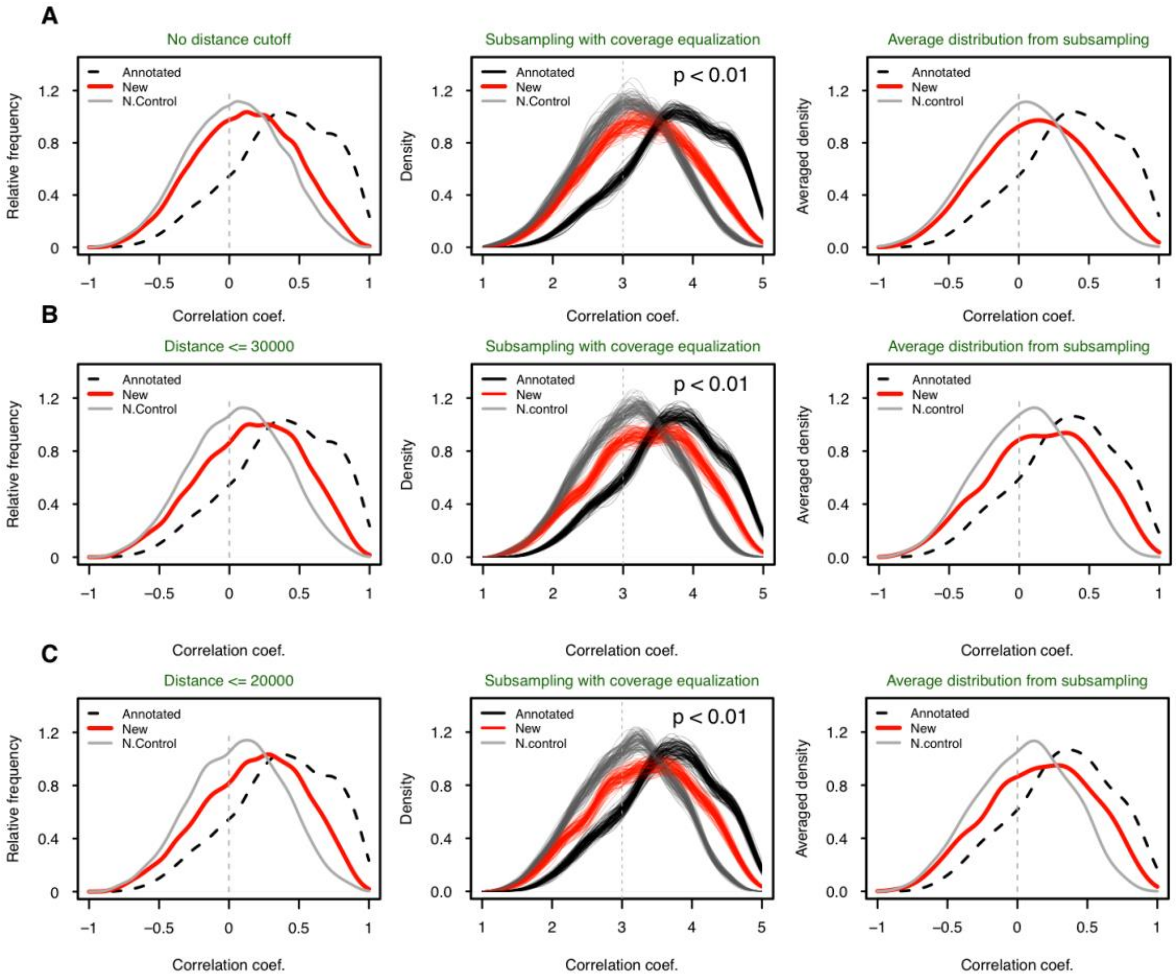


Figure 5. Expression of ASR and antisense gene in 3' extension with distance cutoff & expression equalization. (A) Without distance cutoff, left panel shows the original correlation shift; middle panel shows 100 time subsampling to balance the expression between new and annotated ASR (p-value is calculated by counting out of 100 subsampling how many times new ASR show significant positive correlation shift compared to negative control); right panel shows the averaged curve of subsampling. (B) Same as (A), with distance cutoff at 30k NT. (C) Same as (A), with distance cutoff at 20k NT.

For 5' extension, we had 681 annotated and the same 23496 new ASR as in 3' type. After applying the same analysis, we found that unlike 3' type, the overall correlation between new ASR and their antisense pair didn't stand out from the negative control while positive one still exhibits great correlation as expected (Figure 4C). Considering that 5' UTR are in general much shorter than 3' UTR no matter having splicing or not, distance could be playing a more important role here than in 3' type. Therefore, we set distance cutoff at 50k bp, with a stepwise decrease of 10k. From 50k to 30k, there is no clear signal, but at 20k bp we started to observe a significant positive correlation shift, which gets even better at 10k (Figure 6). Therefore, we applied distance cutoff at 20k, set FDR at 20%, and get 350 new ASR as potential 5' extension to 211 antisense genes. However this time, no novel junction reads could be found to support these connections (Figure 4D).

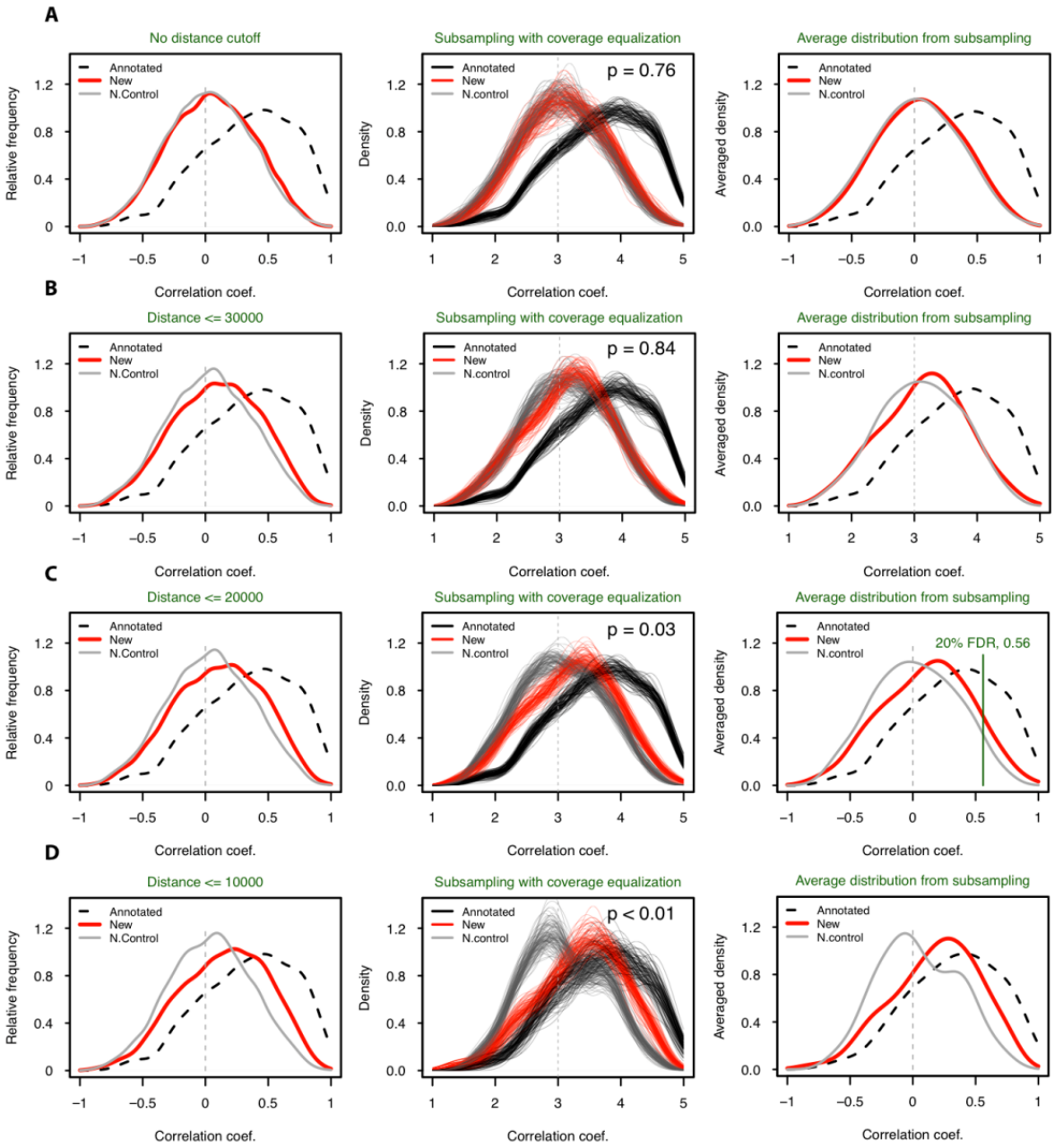


Figure 6. Expression of ASR and antisense gene in 5' extension with distance cutoff & expression equalization. (A) Without distance cutoff, left panel shows the original correlation shift; middle panel shows 100 time subsampling to balance the expression between new and annotated ASR (p-value is calculated by counting out of 100 subsampling how many times new ASR show significant positive correlation shift compared to negative control); right panel shows the averaged curve of subsampling. (B) Same as (A), with distance cutoff at 30k NT. (C) Same as (A), with distance cutoff at 20k NT. (D) Same as (A), with distance cutoff at 10k NT.

Effect on gene's expression

Until now, we have found 6404 ASE-PCG, 20% of which could be explained as having SAS gene pair overlap. Before going on to check the consequences resulting from overlap, we first evaluated the global effect of antisense transcription by comparing the expression level of ASE-PCG & non-ASE-PCG. In addition, surprisingly ASE-PCG did have a significantly higher expression though not so dramatic than non-ASE-PCG (by combing all 14 samples, Figure 7). Then again by taking the advantage of the age-series dataset, we examined the expression change during aging of all expressed PCG, and find 7014 significantly age-related PCG that fall into 8 expression clusters (Figure 8). After that, we tried to check whether ASE-PCG would have any enriched expression patterns within these 8 clusters, but no significant result came out.

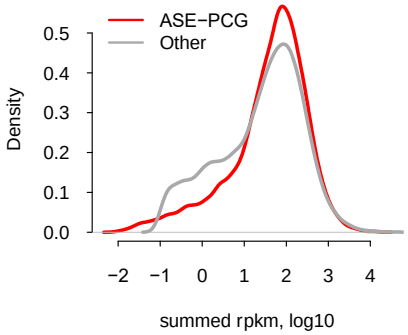


Figure 7. Expression comparison between ASE-PCG and non-ASE-PCG. X-axis represents summed expression measured in RPKM over all 14 samples. One tail Wilcox rank sum test on ASE-PCG(red) are higher expressed than non-ASE-PCG (grey), p-value < 2e-16.

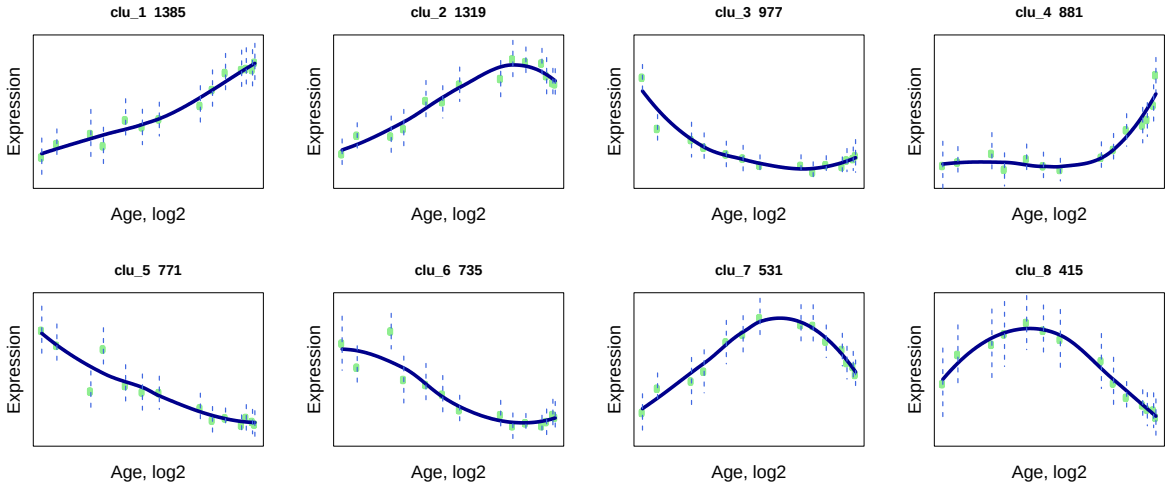


Figure 8. Expression patterns of age-related protein-coding gene.

Nonetheless, ASE-PCG formed overlapping gene pairs from different strand still held quite a lot interest considering transcribing one gene from one strand would inevitably loose up the local chromatin structure thus influencing nearby genes, and potential PolII collision events coming from overlapped transcripts could also lead to complex interference. Therefore, to infer the relationship between overlapped SAS gene pair, we filtered out overlapped sense/antisense gene pairs that had either been

annotated in Ensembl or defined by us as having extension into another gene's locus supported by either novel junction or correlated ASR. And then examined the expression correlation across aging for these pairs. Remaining expressed non-overlapping sense/antisense gene pairs were selected as negative control.

For 3' extension, we observed excess of both positive and negative correlation between theses pairs compared to negative control. (Figure 9A, B). Previously, overlapping length and expression level were shown to have an effect on the gene pairs, thus we went on to check these differences between positively and negatively correlated pairs (selected at correlation cutoff 0.6 & -0.6 where >= 0.6 were marked as positive pairs and <= -0.6 as negative pairs). Although we found no significant distinction between the two types in terms of overlapping length, we did observe that positively correlated pairs tend to have longer overlapping length than negatively or non-correlated pairs when cutoff became more stringent from 0.6 ~ 0.8/-0.6 ~ -0.8D (Figure 10). Besides, we also discovered that no matter whatever cutoff being used, positively correlated pairs always show significantly lower expression compared to negatively correlated gene pairs while remaining significantly higher than non-correlated gene pairs. Further, we found genes from both positively and negatively correlated pairs are significant underrepresented in cluster 7 (BH corrected p-value 0.00872 and 0.00256); genes from positively correlated pairs are overrepresented in Cluster 5 (BH corrected p-value 0.03496).

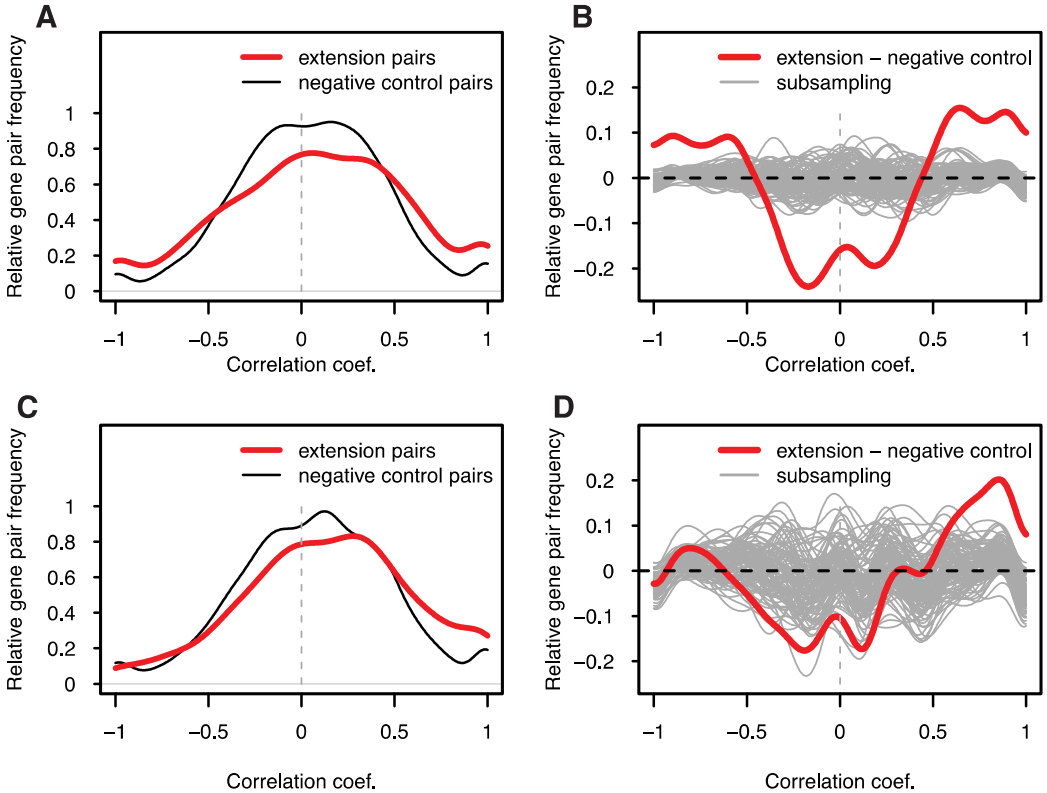
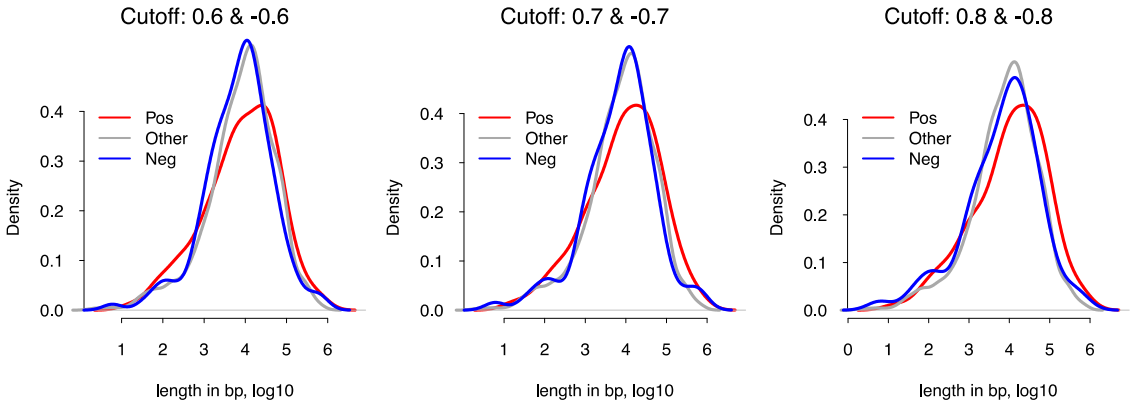


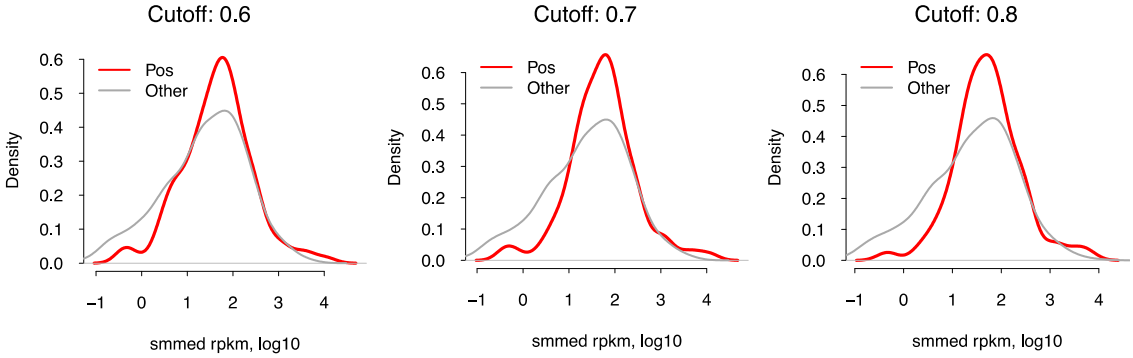
Figure 9. Expression correlation of overlapped gene pairs across aging. (A) Excess of positive and negative correlation between overlapped 3' extension gene pairs in (red) compared to negative control (other non-overlapping sense/antisense gene pairs, black). (B) Shows the difference between density distribution of overlapped 3' extension gene pairs and the background, grey lines represent 100 subsampling from the background with the same number of overlapped 3' extension gene pairs. (C, D) Same as (A, B) for 5' extension.



Wilcox rank sum test	Cutoff: 0.6 & -0.6	Cutoff: 0.7 & -0.7	Cutoff: 0.8 & -0.8
Pos > Neg **P-value**	0.149	0.134	0.065
Pos > Other **P-avlue**	0.450	0.205	0.045

Figure 10. Comparison of overlapped length between differently correlated gene pairs in 3' extension. Up panel shows distribution of overlapped length between differently correlated gene pairs in 3' extension. Lower table shows p-value of one tail Wilcox rank sum test on 1. Positively correlated gene pairs have longer distance than negatively correlated pairs; 2. Positively correlated gene pairs have longer distance than non-correlated pairs.

For 5' extension, there existed excess only for positive correlation (Figure 9C, D). Hence we separated these pairs into positively and non-positively correlated, and examined the difference in gene expression. As a result, we observed that positively correlated gene pairs exhibited higher expression than non-positively correlated pairs (Figure 11).



Wilcox rank sum test	Cutoff: 0.6	Cutoff: 0.7	Cutoff: 0.8
Pos > Other **P-avlue**	0.0037	0.0008	0.0064

Figure 11. Comparison of expression level between differently correlated gene pairs in 5' extension. Up panel shows distribution of expression level (summing all 14 samples) between differently correlated gene pairs in 5' extension. Lower table shows p-value of one tail Wilcox rank sum test on: Positively correlated gene pairs have higher expression than non-positively correlated pairs.

Other than extensions, a third overlapping type – internal – when one gene was completely nested within another from opposite strand, also presented themselves in the annotation and our datasets. And all the remaining ASR that didn't have an explanation by extension also tended to represent novel transcripts yet to be annotated. However when we examined the expression correlation between these internal ASR and their host gene on different strand, no significant difference could separate them from background. The situation held true when checking the expression correlation of internal type on genes level.

Conclusion

Taken together, here we showed that the widespread existence of antisense was even higher than previously reported with Sense/Antisense reads ratio around 16.5 based on the better-annotated gene type – protein-coding genes. Additionally, antisense reads could be efficiently clustered into larger transcription units – Antisense Region (ASR) – which prevailed within 1/3 of all expressed protein-coding genes. In contrast to the sense gene that had determined structure with reads only enriched in exons, ASR were preferentially placed at the both end of the sense gene's locus and mainly lay within the corresponding intron. This two-strands overlaps could raise substantial concern in non-strand-specific studies when quantifying expression or reporting intron retention.

Of all the newly defined ASR, more than 10% could be assign to the neighboring gene on the different strand by evaluating the expression correlation throughout aging and checking connectivity with novel junction. Considering the difficulty in measuring ASR's expression accurately without precise annotation and in identifying novel junction without prior knowledge, this proportion could be greatly underestimated. Furthermore, majority of defined extension events came from 3', which was understandable since as long as the essential transcription start had been determined, extension could be more easily regulated by splicing machinery, and the elongated part could also present themselves as alternative source for splicing or provide more regulatory flexibility through miRNA. Here, compared to earlier study, which could only focus on investigating intergenic region for lacking strand information, we demonstrated that as far as expression was concerned there was necessarily no obstacle in transcription elongation. Hopefully there would be more and more strand-specific studies in the future to better help us understand the gene structure, transcription start site determination as well as transcription termination.

As the result of extending into other gene's locus, we observed significantly more positive and negative expression correlation of overlapping SAS gene pairs across aging compared to non-overlapping pairs for 3' overlap. This was basically consistent with what has long been suspected as the general regulatory effect of antisense transcription. However, as for the difference between positively and negatively correlated pairs, we didn't find much notable changes. On 5' extension, only significant positive correlation was detected, possibly concordant with bi-directional promoter or opened local chromatin structure leading to coordinated expression.

Acknowledgement

Thanks. This study was supported by The Federal Targeted Programme for Research and Development in Priority Areas of Advancement of the Russian Scientific and Technological Complex for 2014-2020 (the Ministry of Education and Science of the Russian Federation), grant № 14.615.21.0002, the Unique identifier of the agreement: RFMEFI61515X0002

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Haiyang Hu, PhD, specialist at the Shanghai Institute for Biological Sciences, CAS, Shanghai, China.

Xi Jiang, specialist at the Shanghai Institute for Biological Sciences, CAS, Shanghai, China.

Ning Fu, PhD, specialist at the Skolkovo Institute for Science and Technology, Skolkovo, Russia.

Philip Khaytovich, PhD, specialist at the Shanghai Institute for Biological Sciences, CAS, Shanghai, China; Skolkovo Institute for Science and Technology, Skolkovo, Russia; The Immanuel Kant Baltic Federal University, Kaliningrad, Russia.

References

Alfano, G., Vitiello, C., Caccioppoli, C., Caramico, T., Carola, A., Szego, M. J. (2005). Natural antisense transcripts associated with genes involved in eye development. Hum. Mol. Genet., *14,* 913-923.

Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E.G., Margalit, H., Altuvia, S. (2001). Novel small RNA-encoding genes in the intergenic regions of Escherichia coli.Curr Biol., *11* (12), 941-950.

Babak, T., Blencowe, B. J., Hughes, T. R. (2007). Considerations in the identification of functional RNA structural elements in genomic alignments. BMC Bioinformatics, *8,* 33-38.

Bánfai, B., Jia, H., Khatun, J., Wood, E., Risk, B., Gundling, W. (2012). Long noncoding RNAs are rarely translated in two human cell lines. Genome Res., *22*, 1646-1657.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N. (2005). The transcriptional landscape of the mammalian genome. *Science*, *309*, 1559-1563.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet*., *38,* 626-635.

Carvunis, A. R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N. (2012). Proto-genes and de novo gene birth. *Nature*, *487*, 370-374.

Clote, P., Ferre, F., Kranakis, E., Krizanc, D. (2005). Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. RNA, *11* (5), 578-591.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res*., *22*, 1775-1789.

di Bernardo, D., Down, T., Hubbard, T. (2003). ddbRNA: detection of conserved secondary structures in multiple alignments. *Bioinformatics, 19* (13), 1606-1611.

Geisler, S., Lojek, L., Khalil, A. M., Baker, K. E., Coller, J. (2012).Decapping of long noncoding RNAs regulates inducible genes. Mol. Cell, *45*, 279-291.

Georg, J., Hess, W. R. (2011). Cis-antisense RNA, another level of gene regulation in bacteria. Microbiol. *Mol. Biol. Rev*., *75*, 286-300.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., Enright, A.J. (2006). MiRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*., *34*, 140-144.

Grinchuk, O. V., Jenjaroenpun, P., Orlov, Y. L., Zhou, J., Kuznetsov, V. A. (2010). Integrative analysis of the human cis-antisense gene pairs, miRNAs and their transcription regulation patterns. *Nucleic Acids Res*., *38*, 534-547.

Havgaard, J.H., Lyngso, R.B., Stormo, G.D., Gorodkin, J. (2005). Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics, 21 (*9), 1815-1824.

Jeon, Y., Sarma, K., Lee, J. T. (2012). New and Xisting regulatory mechanisms of X chromosome inactivation. Curr.Opin.Genet. Dev., *22*, 62-71.

Kent, W.J. (2002). BLAT – the BLAST-like alignment tool. *Genome Res*., *12* (4), 656-664.

Pollard, D.A., Bergman, C.M., Stoye, J., Celniker, S.E., Eisen, M.B. (2004).Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics*, *5,* 6-9.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., Weinstock, G.M., Wilson, R.K., Gibbs, R.A., Kent, W.J., Miller, W., Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*., *15* (8), 1034-1050.

Sluka, P., O'Donnell, L., Stanton, P.G. (2002). Stage-specific expression of genes associated with rat spermatogenesis: characterization by laser-capture microdissection and real-time polymerase chain reaction. *BiolReprod., 67*, 820-828.

Torarinsson, E., Sawera, M., Havgaard, J.H., Fredholm, M., Gorodkin, J. (2006). Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res*., *16* (7), 885-889.

Uzilov, A.V., Keegan, J.M., Mathews, D.H. (2006). Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. *BMC Bioinformatics*, *7 (*1), 173-176.

Washietl, S., Hofacker, I.L. (2004). Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J Mol Biol*., *342* (1), 19-30.

Washietl, S., Hofacker, I.L., Stadler, P.F. (2005).Fast and reliable prediction of noncoding RNAs. *Proc Natl AcadSci USA, 102* (7), 2454-2459.

Wassarman, K.M., Repoila, F., Rosenow, C., Storz, G., Gottesman, S. (2001).Identification of novel small RNAs using comparative genomics and microarrays. Genes Dev., *15* (13), 1637-1651.