534–547 Nucleic Acids Research, 2010, Vol. 38, No. 2 Published online 11 November 2009
doi:10.1093/nar/gkp954

Integrative analysis of the human *cis*-antisense gene pairs, miRNAs and their transcription regulation patterns

Oleg V. Grinchuk[1], Piroon Jenjaroenpun[1], Yuriy L. Orlov[2], Jiangtao Zhou[1] and Vladimir A. Kuznetsov[1,*]

[1]Bioinformatics Institute, 30 Biopolis Street #07-01 and [2]Genome Institute of Singapore, 60 Biopolis Street #02-01, Singapore 138672, Singapore

Received July 10, 2009; Revised October 2, 2009; Accepted October 9, 2009

ABSTRACT

***Cis*-antisense gene pairs (CASGPs) can transcribe mRNAs from an opposite strand of a given locus. To classify and understand diverse CASGP phenomena in the human we compiled a genome-wide catalog of CASGPs and integrated these sequences with microarray, SAGE and miRNA data. Using the concept of overlapping regions and clustering of SA transcripts by chromosome coordinates, we identified up to 9000 overlapping antisense loci. Four thousand three hundred and seventy-four of these CASGPs form 1759 complex gene architectures. We found that ~35% (6347/18160) of RefSeq genes are overlapped with the antisense transcripts. About 30% of Affymetrix U133 microarray initial sequences map transcripts of ~35% CASGPs and reveal mostly concordant expression in CASGPs. We found strong significant overrepresentation of human miRNA genes in loci of CASGPs. We developed a data-driven model of cross-talk between co-expressed CASGPs and DICER1-mediated miRNA pathway in normal spermatogenesis and in severe teratozoospermia. Specifically, we revealed complex SA structural–functional gene module composing the protein-coding genes, WDR6, DALRD3, NDUFAF3 and ncRNA precursors, *mir-425* and *mir-191*, which could provide downregulation of ncRNA pathway via direct targeting *DICER1* and *basonuclin 2* transcripts by *mir-425* and *mir-191* in normal spermatogenesis, but this mechanism is switched off in severe teratozoospermia. The database is available from http://globalisland.bii.a-star.edu.sg/~jiangtao/sas/index3.php?link =about**

INTRODUCTION

A *cis*-antisense gene pair (CASGP) is a pair of genes mapped to opposite strands of the same locus and therefore transcribed in opposite directions. Corresponding pairs of *cis*-antisense transcripts are mRNAs that are at least partially complementary to each other. *Cis*-antisense mRNAs naturally transcribed from CASGP are known as naturally occurrding sense–antisense (SA) RNAs. Such SA transcripts (SAT) have been observed in prokaryotes, fungi, plants and animals (1–4). The overlapping of protein coding genes is a common event in prokaryotic genomes (5,6). On the other hand, up to 32% of yeast genes (3) and up to 25% of mammalian genes (1) have been estimated to form SATs. The number of CASGPs in the human genome remains the subject of considerable debate. Conservative RefSeq-based estimates and earliest counts ranged from 100 to 500 (7,8) to more than 2500 (9,10). Sophistication of computational antisense discovery tools [e.g. Serial Analysis of Gene Expression (SAGE) (4,11), cap analysis of gene expression (12)] led to the ongoing growth of the cDNA and EST databases. As a result, new estimates of CASGPs have been raised by an order of magnitude to 4000–6000 (4). However, reliability, specificity and sensitivity of the computational estimates have not been well defined. In particular, Yelin *et al*. (10) have identified 2667 human transcripts, of which more than 1600 are predicted to be true SAT. More recently, analysis of many fully sequenced mouse cDNAs has predicted the existence of as many as 2500 distinct mammalian SATs (13). The widespread occurrence of natural SATs implies an evolutionarily advantage for this type of genome arrangement.

Natural SAT have already been found to function at several levels of molecular eukaryotic gene regulation including alternative initiation, splicing, termination (14), translational regulation (15), RNA stability, trafficking, apoptosis (4,16), genomic imprinting (17), antisense

*To whom correspondence should be addressed. Tel: +65 6478 8288; Email: vladimirk@bii.a-star.edu.sg

mediated silencing (18) as well as in development processes, such as X-inactivation (19) and eye development (20). Case studies showed that changes in CASGPs transcription could be implicated in pathological processes such as some cancer and neurology diseases (18,21).

The transcripts of CASGPs can share any exonic sequence, regardless of whether the exonic sequence is UTR (untranslated region) or protein-coding. However, many attributes of CASGPs (e.g. alternative transcription start sites, etc.), transcript isoforms and splice variants of the human genome still have not been well classified and studied. The reason for this is the absence of a uniform algorithm, which would integrate validated and predicted CASGPs. In particular, some groups identified SAT pairs from known mRNAs (22), other groups used predicted gene models or UniGene clusters (4). The reliability of predicted SA pairs was not validated by the sequences of well-characterized expression platforms.

In the mammalian genomes, CASGPs can be organized in complex SA gene architectures, in which at least one gene could share loci with two or more antisense partners (1,2,23). The study of these architectures could substantially contribute to our understanding of gene co-evolution and their association with genetic diseases. However, the complex SA structures in humans have not been systematically collected and studied. The publicly available search tools of SAT pairs, for example NATsDB [(4), last release on 7 September 2006] does not report the complex SA gene architectures and misses the gene pairs belonging to such natural SAT groups. For instance, only one gene pair of the complex SA architecture is reported by NATsDB; other pair(s) of such complex SA gene cluster were not reported and their graphic display is incorrect.

Eukaryotes produce various types of small RNAs, or small non-coding RNAs (sncRNAs) of 19–28 nt in length. sncRNAs can induce gene silencing through specific base pairing with the target molecules. Two relatively well-defined classes of small RNAs are involved in RNA silencing: short interfering RNAs (siRNAs) and micro-RNAs (miRNAs) (24). siRNAs and miRNAs are also involved in a wide range of functions such as cell growth and apoptosis, development, neuronal plasticity and remodeling. In cells, the long precursors of siRNAs are generated from long double-stranded RNAs, while miRNAs are generated from long single-strand hairpin-forming precursors.

Theoretically, both ncRNA precursors could be generated from the gene(s) of a CASGP. In case of siRNAs, such a possibility has been demonstrated in several case studies. The pioneer study reported about the protein-coding CAS*G*P represented by SRO5 and P5CDH genes in *Arabidopsis* (25). Overlapped transcripts of SRO5 and P5CDH genes can generate endogenous siRNAs, which participate in regulation of salt tolerance. Additional evidences were found in a recent report (26): on one hand, after injection of sense and antisense transcripts in *Xenopus* oocytes, processing of SA transcripts into siRNAs (SAT-siRNA) was documented. On the other hand, a possibility of a switch from antisense-oriented to sense-oriented SAT-siRNAs was shown in zebrafish embryonic development.

A fine biological regulatory circuitry involving SAT-siRNAs was recently demonstrated via mechanism that has been termed 'small RNA-induced gene activation' (or RNAa) (27,28). RNAa targeting of a CASGP could direct the transcription activation of genes in such SA pair. It was shown that suppression of the p21 antisense non-coding RNA Bx332409 with siRNA leads to a significant suppression of this antisense transcript which correlated with significant increase in expression of p21 sense mRNA (28). However, in a case study of a non-coding–protein-coding SAT pair in human cells, an association of SAT expression regulation and Dicer-mediated pathway was not confirmed (29).

Systematic analysis of occurrence of *precursors of miRNAs* in transcripts of CASGPs and relationships of regulatory pathways of miRNAs genes embedded in CASGP loci has not yet been carried out. Recent findings of a large number of unique natural SATs and miRNAs in transcriptomes of different cell types of eukaryotic organisms and discovery of interconnections in regulatory network directed by natural SAT and ncRNA precursors (1,2,4,5,8,10–13,17,18,25–28,30) necessitate their comprehensive collection, accurate mapping on the genomes and appropriate analysis.

In this work, we report and imply an integrative method for computational identification of CASGPs and their complex architectures that (i) uses RefSeq, mRNA and EST tracks; (ii) imposes stringent quality control filters on EST-to-genome mapping; (iii) provides mapping of Affymetrix U133A and U133B original target sequences and SAGE tags on the genes of SAT pairs. Finally, we analyse the co-localizations of SA genes with miRNA precursors. Ultimately, the method allows identification and characterization of ~9000 reliable SAT pairs found in the human genome including 1759 complex SA gene architectures resulting in uniformly organized collection of CASGPs stored in our new United SA Gene Pairs Pipeline (USAGP) DB. We report about our finding of 128 miRNA-containing SA genes and describe a complex transcription gene module of miRNA and SA genes co-localized on a same chromosome territory. We predict the role of the miRNA-SA gene modules in feed-back regulation of gene expression processes in development, oncogenesis and tumor-suppression activity, and finally, we present a model of cross-talk between co-expressed SATs and Dicer1-mediated miRNAs in silencing pathway in normal spermatogenesis and severe teratozoospermia.

METHODS

Definitions and classification of SA transcript pairs

Technically, we define CASGP as two oppositely transcribed genes, which share a sequence of at least one nucleotide. For mRNAs, the transcripts are considered as SAT if one of the mRNAs overlaps another by at least one nucleotide. The genes in CASGP can overlap each other in

chromosome coordinates either by 5′-ends (divergent pair) or by 3′-ends (convergent pair) (Figure 1).

Alternatively, one transcript from CASGP can be completely inside another. We define the convergent CASGP as the gene pair with overlapping ends (tail-to-tail type); the divergent CASGP as the gene pair with overlapping starts (head-to-head type); and the 'other' as oppositely directed gene pair of any other mutual configuration [e.g. 'embedded' or 'full contained' (13)] (Figure 1). Figure 2 shows an example of visual presentation of NPR2/SPAG8 CASGP in UCSC Genome Browser and our schematic presentation of the overlapping transcripts. In this figure, SATs of a full transcript of gene NPR2 and a longer isoform of mRNA of gene SPAG8 are presented.

Data sources of CASGPs

We retrieved the data associated with annotation RefSeq gene track, mRNA and EST tracks in text format using UCSC Table Browser (http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html#Download) UCSC browser filtered out mRNA track accessions of the NCBI RefSeq Build 36.1 (http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide) were used and appropriate reliable sequences were mapped on reference human genome (assembly Hg18) in both the sense and anti-sense directions. (Specifically, data that failed to map and that mapped in unexpected ways by UCSC criteria were excluded). The data arising from these three tracks could be redundant due to multiple transcripts sequenced for each gene, and also due to RefSeq records being curated by NCBI set, which could be partially derived from mRNA and EST records.

Figure 1. Three overlapping types of *cis*-antisense pairs: (i) divergent (head-to-head, 3′-ends); (ii) convergent (tail-to-tail, 5′-ends); and (iii) embedded *cis*-antisense pair.

To reduce the SA pair's catalog redundancy, we used 'gene-centric approach' to mapping the coordinates of reported antisense pairs onto reference genome Hg18. Figure 3 illustrates the schema of consequent and non-redundant 'accretion' of sequence pairs from different annotation tracks, starting from mapping and identification of RefSeq/RefSeq pairs: we mapped the sequences from track pairs onto genome coordinates, selected unique SAT pair IDs presented by the annotation tracks and removed redundant sequences as it is described in Supplementary Figures S3 and S4. We made such identification of the unique partners of SAT pairs utilizing data consequently chosen from each track. First, we constructed complete data sets of SA sequence overlapping for (i) RefSeq/RefSeq, (ii) RefSeq/mRNA and (iii) mRNA/mRNA pairs (Supplementary Figure S1) using UCSC Table Browser http://genome.ucsc.edu/cgi-bin/hgTables).

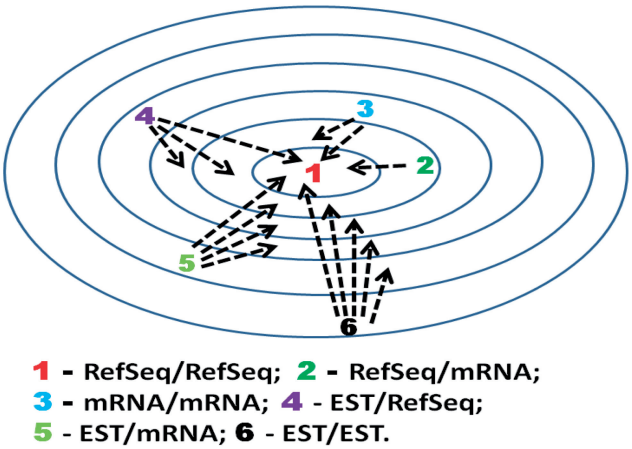

Figure 3. Intergroup redundancy removal procedure through step by step mapping–remapping of six groups of overlapping SAT pairs. The group numeration at the same time corresponds to the each step (and level) of intergroup remapping procedure. Dashed arrows: if the outer level IDs perfectly matched (Supplementary Figure S1) onto the inner level IDs, they were removed from the outer level (e.g. if both IDs of a RefSeq/mRNA SAT pair from level 2 perfectly matched onto both IDs of a RefSeq/RefSeq SAT pair from level 1, they were removed from level 2).

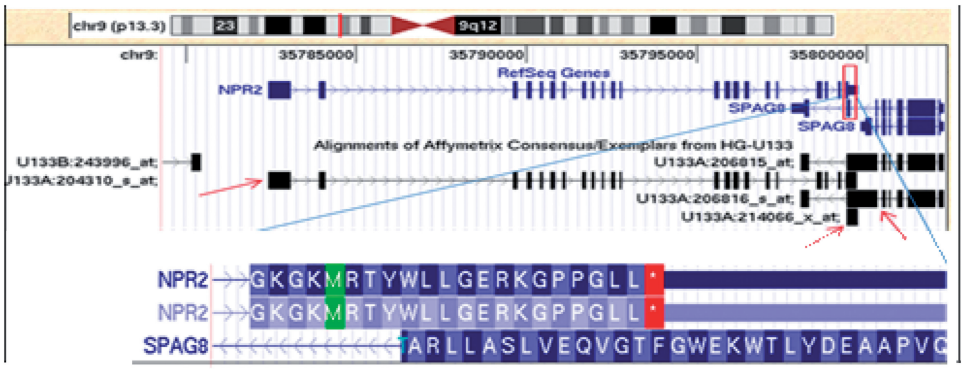

Figure 2. Convergent CASGP NPR2/SPAG8 (CDS/CDS overlap) in chromosome coordinates (UCSC Genome Browser. NPR2 encodes natriuretic peptide receptor B/guanylate cyclase B, which is expressed at moderate or high levels in many tissues. SPAG8 is a sperm-associated antigen isoform 2. This gene is moderately expressed in more than 40 tissues). According to AceView (http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/index.html), SPAG8 could be co-transcribed with NPR2 in many tissues. Red arrows: Affymetrix U133A support.

We calculated the intersections between the sequences which track accessions were located on the sense and anti-sense directions, and retrieved coordinates of DNA sequence covered all described intersected sequence pairs (e.g. mRNA/mRNA) by the track sequences. Second, we remapped and re-annotated computationally generated (7,8,10) and manually curated SA TU data sets using EST annotation track (Supplementary Figure S2). We also used NAT DB (4) as an additional independent data source of USAGP database (Supplementary Figure S2 and Supplementary Table S2). All pairs of overlapping RefSeq Gene sequences in opposite orientation in the same chromosome locus were stored in the CASGP table containing the following standard annotation information: accession numbers, gene names, gene symbol, gene orientation and chromosome coordinates. Using annotation tracks for RefSeq, mRNA and EST of UCSC Table Browser (Hg18), we identified all overlapping sequences in groups: RefSeq/RefSeq, RefSeq/mRNA, mRNA/mRNA, EST/RefSeq, EST/mRNA and EST/EST SAT pairs. Detailed workflow of processing and integration of original data and construction of SAT pairs is shown in Supplementary Figure S1.

The manually curated set of SA TU pairs (Supplementary Figure S2) was formed from several published SAT pair datasets (7,8,10) in which mapped transcripts' orientation, canonical splice sites and canonical poly-A signal have been manually curated by L. Lipovich's group (Genome Institute of Singapore/A*STAR, Singapore). A total of 6707 SA TUs were selected after manual curation. We processed this set in the following way: we computationally removed redundancy of the sequences, improved incorrect EST and mRNA sequence orientations, checked for overlaps between reference sequences using annotation by UCSC human genome assembly (Hg18). Ultimately, we ended up using SA TU 4398 pairs, 1666 of which were unique in USAGP (Supplementary Table S1).

NAT DB data that consists of the other source of our USAGP utilizes the human SA TU pairs using RNA and EST clusters (4) (Supplementary Figure S2).

Table 1 presents the counts of NCBI IDs (RefSeq, mRNA, EST; Build36.1) in both the plus and minus strands. The two rightmost columns of Table 1 present the total number of redundant IDs and the total number of non-redundant IDs after integration of data into five SAT data sets. Supplementary Table S2 shows all DB overlapping counts.

In some cases, one transcript may have two or more different antisense counterparts. Thus, the transcript may form two or more different (complex) SAT pairs with other transcripts. Due to such one-to-many and many-to-many relations, the number of unique NCBI GenBank IDs in SAT pairs (termed n) could be less than $2*n$. For example, even though we have identified 1161 RefSeq/RefSeq pairs, in total these pairs contain only 1700 unique RefSeq IDs (instead of the possible 2322 IDs) (Table 1). Detailed information on complex SAT pairs is presented in Supplementary Table S3.

Additionally, using antisense SAGE data set reported in (11), we identified 2759 SAT pairs those transcripts were mapped at least one time by 21-mer antisense SAGE sequence tags. We mapped exactly antisense SAGE tags onto 2059 SA overlap regions of studied CASGPs. The map was used for evaluation of reliability of previously selected SAT pairs.

RESULTS

Description of USAGP database

USAGP DB is organized as a table with attached tracks for search, including different data set source, number of complex SAT structures, chromosome span (for plus and minus strands), pairing type (http://globalisland.bii.a-star.edu.sg/~jiangtao/sas/index3.php?link=about). Information about SAT pair IDs, chromosome location of SAT pairs, SAT pair overlap spans, the number, type and location of complex SAT structures, long SAGE tags, RefSeq and other sequence types of CASGPs, Affymertix U133 probesets is represented in the main table of DB. The links with NCBI nucleotide DB and UCSC bioinformatics genome browsers provide quick and easy reference, specific data retrieval and visualization. Important advantage of USAGP is the presence of tracks for complex SAT architectures, direct reference to Affymetrix U133 probesets, long SAGE tags as well as presence of a unique subset of manually curated SAT pairs based on ESTs (Supplementary Figure S2).

The types of CASGPs are differentially represented in the annotation tracks

We developed USAGP based on the mapping of the latest RefSeq, GenBank mRNA, EST annotation tracks (downloaded from UCSC browser) and on the integration of these data sets with previously published SAT pair databases and SAT sequences.

Using the above described data sets, we initially compiled the redundant set of 27 074 SAT pairs. A total of 23 782 SAT pairs were non-redundant by NCBI Gene Bank IDs (Table 2). Ninety-nine (31%), 5159 (22%) and

Table 1. Number of unique IDs in five SAT data sets

SA data sets versus annotation track	SATU	RefSeq/RefSeq	RefSeq/mRNA	mRNA/mRNA	NAT DB	All data sets (redundant)	All data sets (non-redundant)
RefSeq	867	1700	5629	198	3104	11 498	7113
mRNA	4592	0	5834	12 502	2132	25 060	15 362
EST	3077	0	0	0	2174	5251	4856
Total no. of IDs	8536	1700	11 463	12 700	7410	41 809	27 331
Total no. of pairs	4398	1161	8170	9640	3705	27 074	23 782

Table 2. Pairing overlap types in five studied SAT data sets

Overlap type	Total *N*	Convergent *n* (%)	Divergent *n* (%)	Other *n* (%)
SA TU	4398	1486 (33.8)	1285 (29.2)	1627 (37.0)
RefSeq/RefSeq	1161	658 (56.7)	238 (20.5)	265 (22.8)
RefSeq/mRNA	8170	2508 (30.7)	1612 (19.7)	4050 (49.6)
mRNA/mRNA	9640	2606 (27.0)	1829 (19.0)	5205 (54.0)
NAT	3705	1340 (36.2)	1070 (28.9)	1295 (35.0)
Total non-redundant SAT pairs by ID	23 782	7399 (31.1)	5159 (21.7)	11 224 (47.2)
No. of clustered SAT pairs on distinct chr. territory	8894	2177 (24.5)	1978 (22.2)	4739 (53.3)

11 224 (47%) of these 23 782 SAT pairs were represented by 'convergent' (tail-to-tail), 'divergent' (head-to-head) and 'other' (e.g. embedded) SAT pairs, respectively (Table 2). We joined these SA pairs into 8894 SAT pair clusters that overlap regions are located on the distinct chromosome territories (see additional details in Supplementary Table S2).

Table 2 shows that USAGP contains a much larger number of SAT pairs (gene pairs, transcript pairs, TU pairs) than any reported estimates. The most common pattern is the complete covering (embedded in 'other' type): one large transcript contains other small ones in antisense orientation (~53%; Table 2). Comparison of the overlapping SAT gene pairs observed in five different sets of SAT pairs shows much more frequent occurrence of 3′-end/3′-end UTR patterns rather than 5′-end/5′-end patterns.

However, this distribution pattern of SA pair types could be biased because of the prevalence of the convergent type of SAT gene paring observed for the most reliable subset represented by SA RefSeq gene pairs. The fraction of 'convergent', 'divergent' and 'other' pairing types for RefSeq-RefSeq set was 56.7, 20.5 and 22.8%, respectively (Table 2). Using the Ensemble DB, Makalowska *et al.* (9) reported similar results: they found that the fraction of 791 detected 'convergent', 'divergent' and 'other' pairing type sets was 52, 25 and 23%, respectively. Similar results have been reported by Veeramachaneni *et al.* (8), for 774 pairs of human overlapping protein coding genes.

Complex SAT pair genome architectures

A gene can have several antisense partners overlapped in several disjoined loci on a given chromosome (Figure 6; Supplementary Figure S5), and hence represent overleaped SAT clusters termed as the complex SA gene architectures. Such SA gene clusters could be classified as (i) 'SAT pair chains' when SATs alternate in plus and minus strands and (ii) 'one-to-many SAT overlap structures' described in Supplementary Figure S5.

We found that 49% (4374/8894) of SAT pairs form 1759 complex gene architectures including at least three genes. The rest 51% (4520 of 8894 SAT pairs) have a single overlap with the antisense partner.

Supplementary Table S3 shows a detailed description of the distribution of the number of clustered SAT pairs in complex gene architectures. Eighty-one percent (1425/1759) of complex SAT fall into long genes with two and more distinct antisense partners transcribed from opposite strand; 19% (334/1759) complex SAT architectures are SAT pair chains. Several examples of complex SA gene architectures (SA chains) have previously been reported (1,2,25). In the human transcriptome, we found complex SA gene architectures having up to seven distinct overlaps in a SA chain and up to 21 distinct overlaps in one-to-many SAT overlap structures. To quantify types of SAT pairing, we calculated the number of overlapping blocks (exons) and the length of overlapping sequences for every SAT pair. In the most of cases studied, either only one exon overlapped one exon of the opposite strand or all exons fell into single exon of antisense transcript counterpart.

SAT overlaps are highly enriched by transcripts of protein coding–non-protein coding gene pairs

To perform functional classification of partners in SAT pairs, the clustered SAT pairs were classified based on combinations of RefSeq gene (coding for proteins) and non-Refseq genes (non-coding for proteins). According to our database, the number of clustered RefSeq-RefSeq (coding–coding) pairs in SAT overlap pairs is 677; the number of non-coding–non-coding SA pairs is 2916. Transcripts from protein coding–non-protein coding gene pairs consist of a major subset (60%; 5301/8894) of transcribed SA gene pairs. A vast majority of these SAT does not show evolutional conservation (data not presented). These findings suggest that a major fraction of CASGPs is the protein–coding transcript–non-protein coding transcript pairs and these pairs are human specific.

SAT pair clusters are highly enriched in the human genome

The proportion of protein-coding transcripts in SAT pairs may be even higher if we consider all spliced mRNAs with ORFs of large length (>100 amino acids) as protein coding. In many cases, a gene has antisense overlaps with several genes on opposite strand which increases the complexity of SAT overlap patterns. To characterize the structural and functional complexity of the SAT overlap patterns, we also counted the number of gene names associated with RefSeq DB IDs. Due to the fact that a single RefSeq gene name could often (~20%) be represented by two or more RefSeq IDs (annotation and functional isoforms), only 18 160 different gene names are associated with all annotated 24141 RefSeq IDs. We found that ~30% of RefSeq IDs (7113/24 141) included in SAT pairs represent 35% (6347/18 160) of non-redundant human gene symbol IDs. The last number suggests a high level of enrichment of protein coding genes in SAT pairs. Additionally, according to our estimates, ~51% of human gene regions (counted as a total non-redundant span of RefSeq genes in all chromosomes) are occupied by SAT gene pairs. The total length

of complex SA gene architectures covers ~21.4% of all chromosome size (excluding centromere regions).

Interestingly, the gene length of SAT pair architectures is 5% longer on average as compared to the length of random genes (out of this subset of genes). In addition, SAT pairing patterns are often associated with high-complexity genome regions; for instance, SATs are enriched on high-density gene chromosomes 19 and 13 (14,31).

Expression microarray probe sequences support functional analysis of SAT pairs on the genome level

To further validate and classify the putative SAT transcript pairs, we mapped transcripts represented by different annotation tracks by the target sequences of Affymetrix U133 A and U133 B microarrays (Supplementary Figure S2). About 29.7% (13 260/44 692) of Affymetrix initial sequences match onto RefSeq or mRNA or EST clustered SAT pairs. A total of 10 387 of 13 260 Affymerix target sequences were selected after filter-out of low-quality probesets (30). This reliable sequence mapping information allows us to provide an additional bioinformatics support of possibility to exist the transcripts formed by CASGP pairs, especially in the cases of less reliable mRNAs–ESTs and EST–EST SA pairs. After filtering out low-quality probesets (30), we found 6440 reliable Affymetrix probesets forming initial probeset sequence pairs covering both transcripts. These 6440 initial probeset sequences represent ~30% (2701/8894) clustered SA pairs. The Figure 2 shows an example of mapping of Affymetrix initial probe sequences onto CASGPs.

Thus, we can conclude that a substantial number of non-protein coding sequences in SAT pairs can be supported by Affymetrix U133 A&B probesets and these microarrays can be used in expression analysis of SAT gene pair and in identification and validation of their functions.

Using transcripts of SAT pairs supported by Affymetrix U133 A&B probesets, we found that SATs are significantly higher expressed in comparison with non-SA transcripts (Figure 4). It is important to note that Affymetrix target sequences (and corresponding Affymetrix probesets) which have been defined in (30) as unreliable target sequences, were excluded from our analysis according to the algorithm presented in (30). A set of the unreliable target sequences includes: sequences with poor homology, mis-oriented Affymetrix target sequences (those complete genomic coordinates perfectly matched the target gene on the opposite strand), Affymetrix target sequences mapped to repeat-rich regions and the sequences mapping to more than one genome locus Hg18.

Using APMA-defined reliable set of Affymetrix target sequences, we calculated correlations of SAT pairs in several normal and cancer tissues. Statistical testing of the frequency distribution function of the number of correlation coefficients (by Pearson and by Kendal) showed a common pattern: a fraction of significant positive correlations in *cis*-antisense pairs strongly

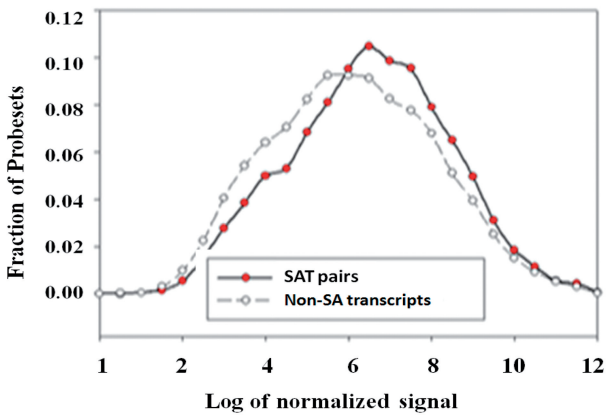

Figure 4. Comparison of the frequency distributions of Affymetrix Chip U133 hybridization signal of mRNA of the SAT and random non-SAT mRNA pairs shows preferential over-expression genes in CASGP. 21 brain samples from epilepsy patients were used (GEO ID GSE4290).

dominates over negative correlations. Figure 5 shows a typical example of our correlation analysis of microarray data on the genome scale. For relatively large ($n = 70$) and clinically homogeneous (breast cancer grade 1) data set, this figure exhibits strong preference for significant values of positive correlation coefficients versus negative correlation coefficients. For instance, the number of significant Kendal correlation coefficients at $P = 0.01$ ($r \geq 0.21$) in the right tails of the distributions is 241, while in the left tails at $P = 0.01$ ($r < -0.21$) this number is equal to only 20. The correlation analysis was done using RefSeq/RefSeq SAT pairs supported by non-redundant and high-quality (by APMA DB) Affymetrix U133A&B Chips signals. Our results are consistent with data in human and mouse (1,13) which were reported based on analysis of limited number of SAT pairs.

Gene Ontology analysis reveals specific functional categories of the genes in SAT pairs

A total of 5473 DAVID IDs of our 7113 SAT RefSeq IDs were recognized by DAVID software (http://david.abcc.ncifcrf.gov/). The term 'alternative splicing' in the category SP_PIR_KEYWORDS demonstrated the greatest significance with Bonferroni correction ($P = 1.4E{-}126$). Other most significant functional terms for the same category, such as 'phosphorilation' ($P = 2.7E{-}70$), 'nuclear proteins' ($P = 9.4E{-}64$), 'DNA-binding' ($P = 6.3E{-}28$), 'zinc-finger' ($P = 9.6E{-}20$) and 'ATP-binding' ($P = 4E{-}24$), 'coiled-coil' ($P = 1.3E{-}23$) cation binding were also related to SA phenomena. Among genes in our database, we also found statistical significant enrichment ($P < 0.05$) of genes related to breast, gastric and lung cancer cells, and Alzheimer's disease pathway.

Enrichment of SAT pairs by protein-coding genes generating miRNAs

We have mapped 701 putative precursor miRNAs regions (MiRBase Annotation System, release12; (http://

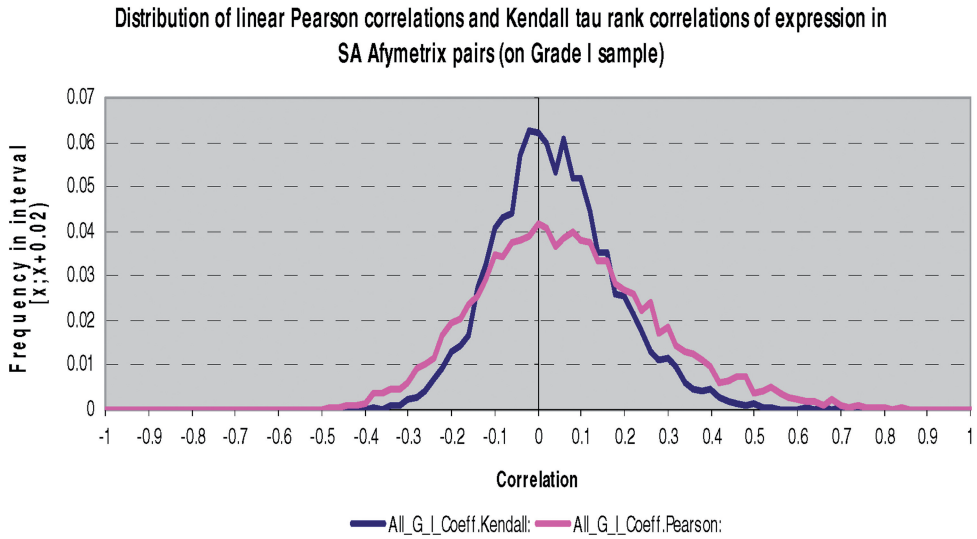

Figure 5. Frequency distributions of Kendal correlation and Pearson correlation values of gene pairs expressed in SA partners show overrepresentation of positive (co-regulation) patterns versus negative (mutually negative) patterns at the global genome scale. Identification of reliable probesets detecting SAT pairs was defined based on APMA DB (30). Chip U133A&B microarray expression and clinical data (70 breast cancer patients with grade 1) were described by Ivshina *et al.*, (32) and were presented in GEO NCBI DB (ID: GSE4922).

microrna.sanger.ac.uk/) onto RefSeq annotation track (Hg18). Detail description of this mapping and summary statistics of AS gene pairs containing miRNAs is presented in Supplementary Table S4. Four hundred and thirty-six distinct RefSeq IDs containing one or more putative precursor miRNA in the host RefSeq gene region were identified. Two hundred and eighty-eight of these 436 RefSeq IDs corresponded to unique gene symbols. One hundred and twenty-eight of these 288 genes (44%) included from 1 to 3 miRNA precursors into a given gene region. We termed these miRNA precursors containing genes the *SAT-miRNA* genes. Detailed annotation of SAT-miRNA genes is presented in Supplementary Table S4A. Considering the fact that at least 35% of total number of non-redundant RefSeq gene symbols are involved in SAT pairing ('Results'), a strong enrichment (44%) of miRNA putative precursor regions in such defined protein-coding genes involved in SAT pairs was observed ($P = 0.0004$, Fisher's test). Moreover, 21% (150/701) of 701 MiRBase miRNA precursors are derived from one of the 128 Refseq genes consisting >0.7% of Refseq genes reported in the human genome. Interestingly, the long non-coding RNAs could be also the sources of miRNAs, for instance for the members of let-7 family (Suplementary Table S7).

Among the 128 genes, the genes of mRNA transcription regulation, nucleic acid binding, tRNA metabolism, transferases and other transcribed genes are strongly enriched by Panther Bioinformatics DB. By DAVID DB, 33/128 genes are involved in developmental process (GOTERM_BP_ALL; $P = 0.0039$, FDR = 0.073) including 'anatomical structure development' (24 genes; $P = 0.006$; FDR = 0.11). Protein binding (63 genes; $P = 0.00064$; FDR = 0.011), RNA polymerase II transcription factor activity (8 genes, $P = 0.00098$; FDR = 0.017) and nucleotide binding (26 genes, $P = 0.0038$; FDR = 0.066), including 22 genes related to purine nucleotide binding) are also reported by David DB. DGCR8, encoding a key protein of pre-miRNA processing is also presented in the list of SAT-miRNA precursors. This finding suggests a direct cross-talk between SAT and miRNA pathways. Thus, our results suggest a role of CASGPs containing miRNA precursors in regulation of gene expression, nucleic acid binding and biogenesis including ncRNA biogenesis (see below).

miRNA precursors could be included into complex SA pair-driving structural–functional regulatory module

miRNA precursors could be found in SA gene pairs and architectures. For instance, we found two miRNA precursors (*mir-425* and *mir-191*) within the three-gene complex genome architecture composed of WDR6, DALRD3 and C3orf60 (NDUFAF3) located on 3p21.31. Figure 6 shows that WDR6, DALRD3 and NDUFAF3 genes are organized in the human genome into a complex SAT chain. WDR6 is able to interact and synergize with the well-known tumor suppressor–serine/threonine kinase STK11 in cell cycle G1 arrest in cancer cells (33). The functions of DALRD3 are unknown. Protein encoded by DALRD3 contains DALR anti-codon binding domain. The gene NDUFAF3 encodes a mitochondrial complex I assembly protein that interacts with complex I subunits. Mutations in this gene cause mitochondrial complex I deficiency (a fatal neonatal disorder of the oxidative phosphorylation system) (http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide). Orthologous WDR6-DALRD3 SAT pair is also presented in the mouse and the rat; however, DALRD3-NDUFAF3 SAT pair is found only in the human (defined by http://genome.ucsc.edu/).

Both *mir-425* and *mir-191* precursors are embedded in the 1st intron of DALRD3 gene (upper red box and

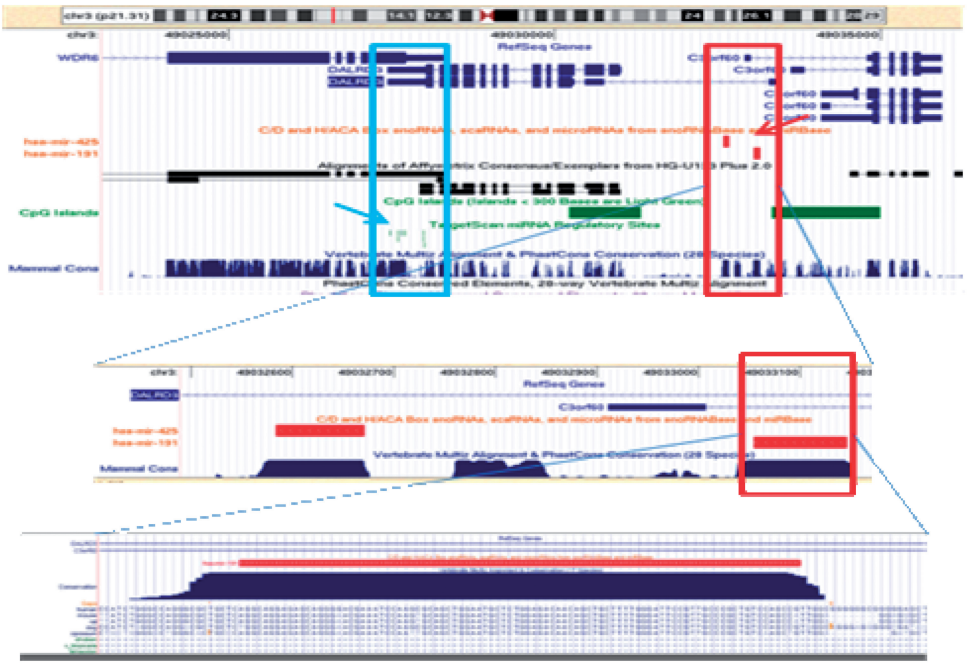

Figure 6. Complex sense antisense gene architecture represents a structural-functional regulatory module. Top panel: exon–intron structure of CASGP protein-coding genes in the cluster WDR6, DALRD3 and NDUFAF3 and precursors mir-425 and mir-191. Upper red box and red arrow: the 1-st intron of DALRD3 gene includes the mir-495 and mir-191 precursor genes. Upper blue box and blue arrow: putative miRNA target sites (mir-331, mir-205, mir-433-3p, mir-125/351, mir-299-5p, mir-409-3p, mir-543, mir-1/206) located on 3′ UTR's of WDR6 and DALRD3 genes in the region of their SAT overlap. Mammalian genome conservation regions are presented along the bottom line of upper panel. Green rectangle: CpG island. Intermediate and bottom panels show the detail structure of UCSC annotation tracks presented in the red box of the top panel. See details in the text. View in UCSC bioinformatics browser.

arrow, Figure 6) and due to common transcription event could be expressed concordantly, because miRNA biogenesis is initiated by transcription with RNA polymerase-II and their primary transcripts (pri-miRNAs) harbor a local hairpin structure that is then cropped by a nuclear RNase III, Drosha, into ~70 nt precursor-miRNAs (pre-miRNAs). Drosha functions in nucleus in a complex known as miRNA Microprocessor that also contains a dsRNA-binding protein, DGCR8 (DiGeorge syndrome chromosomal region 8). Expression of *mir-191* miRNA was observed in many human cancer cells including HL-60 leukemia cells (34). It was shown that biogenesis of pre-miRNA-191 from pri-miRNA in leukaemia cell line K-562 is under strong positive control of Drosha microprocessor complex (35). Over-expression of *mir-191* was also documented in studies of several types of solid tumors including breast, colon and lung cancers (36). Significant abandance and coexpression of *mir-191* and *mir-425* were demonstrated in various cancer cell lines (37). Strong expression in the human cancer cells and conservation of *mir-425* and *mir-191* sequences in mammals (human, mouse, dog and opossum; see bottom red box on Figure 6) supports the idea that they are functional and could play important regulatory role in the cells of high eukaryotic organisms.

As we showed above (Figure 5), gene co-expression is also common regulatory pattern of SA gene pairs. Thus, biogenesis of miRNA(s) might tightly couple by co-transcription pattern of both genes involved in SA gene pair. Therefore, due to common RNA polymerase-II-mediated transcription event the precursors of *mir-425* and *mir-191* could be concordantly processed by Drosha microprocessor complex and, finally, the mature *mir-425* and *mir-191* could be concordantly expressed with mRNAs of DALRD3 and NDUFAF3 genes (see below). A theoretical possibility of the structural and functional integrity and cross-talk of Dicer1-mediated miRNA pathway with DALRD3–NDUFAF3 SA pair could also be supported by multiple miRNA target sites located on 3′ UTRs of WDR6 and DALRD3 genes in the region of their SAT overlap (Figure 6). In addition to these findings, presence of CpG islands in promoter regions of WDR6, DALRD3 and NDUFAF3 genes (Figure 6) suggests that expression of the complex SAT gene architecture and the miRNA precursors could be under specific epigenetic control. This type of complex genome cluster is considered as a novel SA pair-coupled structural–functional regulatory module (SFRM) including both SA protein-coding genes and the two known miRNA precursors.

Specific functions and processes associated with *miR-425* and *mir-191* target genes

To search the functions of predicted SFRM, we used TargetScan microRNA prediction software, (http://www.targetscan.org; release 5.0). We identified 120 conserved gene targets of *miR-425* and 34 conserved targets of

mir-191 (Supplementary Table S5A and B). We did functional annotation of predicted target genes by DAVID bioinformatics software and observed a statistically significant enrichment for the term 'nucleus' in the category GOTERM_CC for both miRNA targets sets: for *mir-425*—42 out of 120 genes (35.7%, $P = 0.012$); for *mir-191*—14 out of 34 genes (42.4%, $P = 0.015$) (Supplementary Table S5A and B). However, several more specific terms related to transcription regulation and genes related to miRNA biogenesis were found (Supplementary Table S5A and B). Specifically, we found that a large number of *mir-425* target genes are involved into negative RNA biosynthesis (GOTERM_BP – 'negative regulation of transcription', $P = 0.0005$) and gene expression silencing (Supplementary Table S5A). By TargetScan' prediction, the transcripts of the nuclease DICER1 could also be potential targets for *mir-425* (Supplementary Table S5A). DICER1 cleaves double-stranded pre-micro RNA into siRNAs or miRNAs and plays a key role in miRNA gene silencing pathway. SMAD2 is also a potential target of *mir-425*. SMAD2 is a member of SMAD gene family. The proteins of this family can control DROSHA-mediated miRNA maturation required in the first step of miRNA processing and acting in the nucleus. *Mir-191* also has potential target genes directly involved in miRNA gene silencing pathway. Its putative target gene, SOX4 can directly regulate expression of three components of the RNA-induced silencing complex, namely DICER1, Argonaute 1 (RANSAP) and RNA Helicase A (38).

Moreover, both *mir-425* and *mir-191* could regulate the same predicted target gene—*basonuclin 2* (BNC2). Although the exact function of BNC2 in miRNA biogenesis is not clear, in testis and epidermis of human newborns BNC2 virtually completely co-localizes with splicing factor SC35 (39). SC35 is a specific marker of subnuclear organelles known as SC35-containing foci, or splicing 'speckles'. Interesting model was proposed of SC35 foci as dynamic 'hubs', which facilitate the expression of highly active Pol II-transcribed genes by spatially connecting their synthesis with the recycling of numerous proteins involved in mRNA metabolism (40). Noteworthy, a certain fraction of *pri-miRNA* may escape processing by DROSHA-DGCR8 microprocessor complex at sites of highly active transcription. This pri-miRNA fraction can also accumulate in SC35-containig foci, where it may be degraded or retained, or processed into pre-miRNA and exported into cytoplasm (41).

Thus, our GO and literature analyses suggest an intriguing possibility of negative feedback loop control of miRNA pathway and coherent targeting DICER1 as well as BNC2 by *mir-425* and *mir-191,* the miRNAs whose expression level is directly associated with co-transcription of DALRD3 and NDUFAF3.

DALRD3 and NDUFAF3 are co-expressed in testis and may be involved in testis-specific regulation of ncRNA pathway

Using DAVID Bioinformatics Gene ontology (GO) tool, we carried out a tissue enrichment analysis of the 128 SAT-miRNA genes. We found that DGCR8, DALRD3 and NDUFAF3 are preferentially expressed in testis (GNF_U133A_QUARTILE: Testis Germ Cell_3rd). When we carried out DAVID Bioinformatics GO tissue enrichment analysis of 120 putative target genes reported for *mir-425*, we found a high enrichment of the genes in testis (Supplementary Table S6). We used DAVID category GNF133A_ QUARTILE and identified three highly significant terms for testis ('Testis_3rd', $P = 0.02$), testis interstitial cells ('Testis interstitial_3rd', $P = 0.014$) and testis germ cells tissues (Testis germ cell_3rd'', $P = 0.04$, Supplementary Table S6A). Moreover, after analysis of the joined lists of the 120 putative targets of *mir-425* and the 34 putative target genes for *mir-191*, we observed stronger enrichment scores (and *P*-value) for testis ('Testis_3rd', $P = 0.002$, Supplementary Table S6B). These facts suggest an association of DALRD3, NDUFAF3, *mir-425* and *mir-191* function with testis and spermatogenesis.

We have studied these associations in terms of microarray gene expression, and finally found interesting expression data sets (42–44) which allowed us (i) to integrate our findings and (ii) to propose a new molecular model of SA-miRNA-gene driving regulation of normal spermatogenesis (N_S) and to suggest a role of the SFRM in the molecular mechanism of human severe teratozoospermia (T_Z) (Figure 7). In the report (42), the expression profile of 582 human miRNAs in sperm cells of reproductively normal infertile individuals and sperm cells of patients with non-obstructive azoospermia (NOA) have been studied. NOA is a condition which is characterized by meiotic errors leading to spermatogenic arrest and the production of aneuploid gametes. A total of 173 miRNAs were found to be differentially expressed in NOA compared to normal control, and 90% of these 173 miRNAs were downregulated in NOA (42). The shorter list of top-down regulated mRNA includes *miRNA-191**, *miRNA-425* and three members of *let-7* mRNA (Supplementary Table S7).

Teratozoospermia is one of the most critical morphological parameters associated with sperm aneuploidy that affects fertility in males. In (43), the authors have used U133-2plus and Illumina Sentrix WG6 microarray expression platforms to examine transcription profiles of 13 reproductively normal spermatozoal RNA samples, N_S, and 8 spermatozoal RNA samples from severe teratozoospermia, T_Z.

Based on NCBI GEO data (GSE6872) reported in (43), we found that DALRD3 and NDUFAF3 are strongly co-expressed in sperm cells of normal individuals (Figure 7). Additionally, these two SA genes showed significant positive correlation of mRNA levels (Kendal coefficient = 0.77, $P < 0.01$) in N_S; however, both genes were switched-off in severe T_Z. We found similar gene expression pattern for miRNA *let-7d* which is represented in Affymetrix data sets by the probe 227793_at. Finally, transcriptional concordance of *mir-425, mir-191* and *miRNA let-7* family was reported (44). Moreover, expression of *mir-425, mir-191* is similar to that of the mRNA from the host gene, DALRD3 (44) and inversely correlated with expression of many target genes. It was

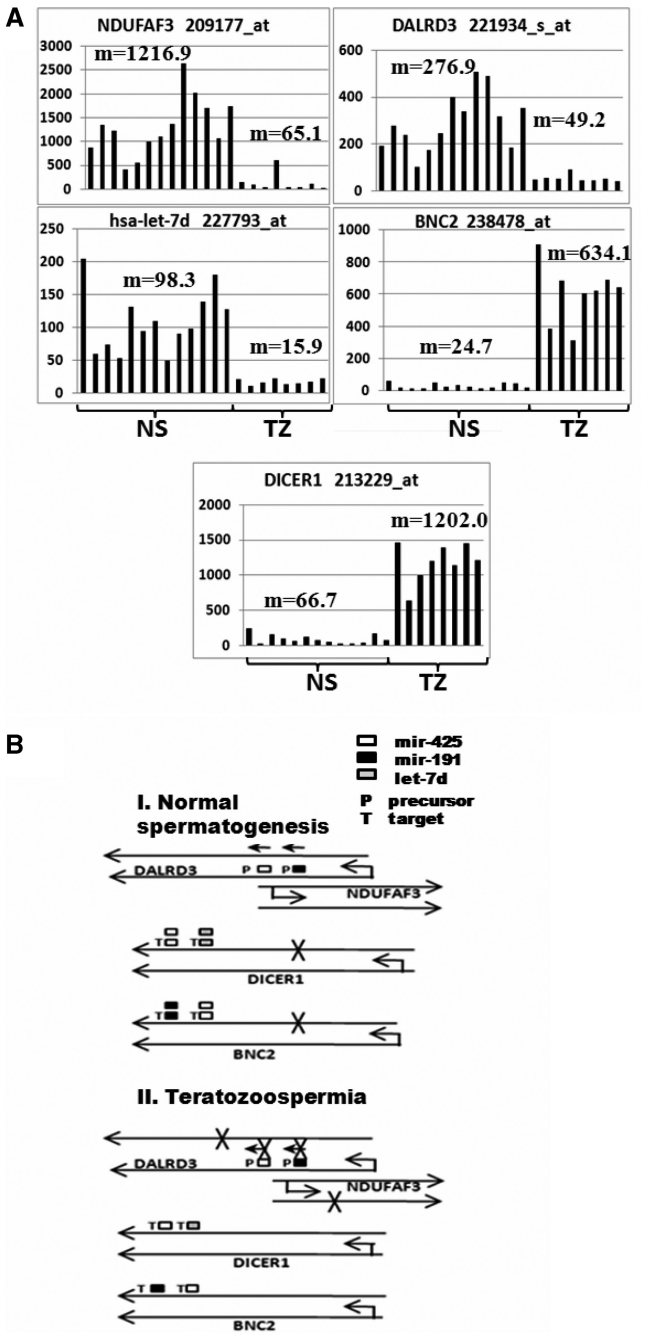

Figure 7. Data analysis and hypothetical model of transcriptional regulation of *miRNA-425* and *mirRNA-191* associated with co-expression of NDUFAF3-DALRD3 SA gene pair in normal spermatogenesis and teratozoospermia. (**A**) Expression profiles for NDUFAF3, DALRD3, *hsa-mir-let7d*, DICER1 and BNC2 in N_S and T_Z retrieved from Genome Omnibus platform GDS1665. N_S, patients with normal spermatogenesis ($n = 13$); T_Z, patients with teratozoospermia ($n = 8$). (**B**) Scheme of hypothetical mechanism by which in normal spermatogenesis local activation of transcription of DALRD3/NDUFAF3 CASGP may also include biogenesis of precursors of *miRNA-425* and *miRNA-191*.

documented that miRNA of *let-7* family are able to suppress DICER1 activity (45,46).

Summarizing these findings, we could suggest that in N_S, high expression of DALRD3 and NDUFAF3 might provide high and stable expression and robust biogenesis of *mir-425* and *mir-191* (derived by DROSHA/DGCR8 microprocessor from DALRD3 mRNA) which together with the members of *let-7* family could suppress a transcription of DICER1 and BNC2 genes in spermatogenesis. Inversely, due to our model, the expression switch-off of DALRD3, NDUFAF3, *mir-425, mir-191* and *let-7d* in T_Z [original data from (42,43)], suggests the switch-on in DICER1 and BNC2 transcription. Figures 7A and B show that the expression patterns of DICER1 and BNC2 in N_S and T_Z follow the model.

Moreover, the expression patterns of the proteins of ncRNA biogenesis machineries (DROSHA, DGCR8, AGO2 and PIWIL2) are concordant with expression patterns of DALRD3, NDUFAF3, *mir-425, mir-191* and *let-7d* in T_Z and N_S (Supplementary Table S8). Similar expression pattern were found for several markers of normal spermatogenesis (e.g. SEMG1, SEMG2) as well as factors of etiology and pathogenesis of T_Z (SPATA16, DAZAP1 and DAZL) (47,48), precursors of long protein-non-coding RNAs (e.g. MALAT1) (49) (Supplementary Table S8). However, BNC2, DICER1 and stem cell markers (Nanog, Sox4) express opposite transcription pattern (Figure 7; Supplementary Table S8).

Thus, results of our analysis suggest that pri-miRNA and pre-miRNA processing in the nucleus of T_Z sperm cells might be dissociated from miRNA maturation steps coupled with DICER1 activity. We speculate that (i) high or moderate expression of *mir-191*, *mir-425* might be modulated and coordinated by DALRD3/NDUFAF3 and (ii) these DALRD3/NDUFAF3-driven miRNAs in their turn could lead to suppression of DICER1 and BNC2 (Figure 7A and B).

DISCUSSION

We have developed a comprehensive and reliable pipeline which enables to store, update and analyze heterogeneous information on SAT gene and transcript pairs overlapped on opposite strands of the same locus in the human genome. We found that non-redundant integration of SATU, RefSeq/RefSeq, RefSeq/mRNA, mRNA/mRNA, NAT data sets in USAGP data set provides 23 782 pairs of SAT transcripts (Table 1). For the first time, we identified and classified 1759 complex SA gene structures, most of which have not been described in the literature. Using our clustering procedure, we joined these pairs and identified 5977 SAT over-lapped regions (Table 1) with the total (non-redundant) length of SAT overlaps equal to 35 451 847 bp. These regions cover ~1.15% of the human chromosome length (3076 Mb). Interestingly, this estimate is comparable with the total length of protein coding regions in the human genome (~2%).

Our united SA pair database differs from published databases in the following ways:

(i) Completeness. DB is constructed using genome annotation tracks and data tables independently compiled by different scientific groups using different algorithms.

(ii) Non-redundancy. The major unit is the pair of genes based on chromosome coordinates, but not on TUs or EST clusters.
(iii) Classification of SAT chains and complex structures description. DB contains explicit information on complex SA structures (e.g. three and more SAT overlapping genes).
(iv) Detailed description of exon–exon and exon–intron overlaps for protein–coding genes.
(v) Hierarchical organization of the data by reliability of annotation: from more reliable to less reliable (RefSeq → mRNA → EST) and from several data sources to one data source. We found that 1608/2667 TU pairs from the data set collected by Compugen (10) are present in our database. Interestingly, ~1600 predicted SA TUs, have been considered as the reliable pairs (10). We also compared our data sets with ANTICODE DB (http://www.anticode.org/) which integrates data from three sources (10,31,50). Overall, our database is ~30% larger than ANTICODE database (Supplementary Table S9). The vast majority of transcript pairs skipped in USAGP were either not mapped onto the latest human genome release or did not pass quality control or consisted of not validated EST pairs. It is noteworthy that the last updated (November, 2006) NAT DB contains 7356 SAT pairs (51) in contrast to 3915 pairs previously announced by the same group (4). The number of SAT pairs in present release of USAGP (8894 pairs) exhibits an increment of more than ~1500 SAT pairs in comparison with this update of the NAT DB. Recently, Galante *et al.* (52) have reported 10 077 SA transcript pairs in the human genome. However, only 5408 of these pairs were left following the mapping of corresponding transcripts on the NCBI Build 36.1 and removing redundant mRNA sequences.
(vi) Expression sequence support. SATs are directly mapped by target sequences of well-established microarray platform (Affymetrix U133A&B). In agreement with (53), we report here that commercial microarrays can be used to detect ~30% of 8894 human SAT pairs. Additionally, 28% of the SAT pairs are supported by SAGE tags.

By computational analyses of *cis*-antisense data (1,4,8,10), from 15% to 25% of mammalian genes located on the opposite chromosome strands can be overlapped on the same loci, giving rise to pairs of sense and antisense RNAs. However, the number and structure of *cis*-antisense transcription units in the databases could be quite sensitive to the working definition of SAT pairs, gene model and to genome release assembly. We found that ~30% of RefSeq genes (7113/24121, UCSC Table Browser, March 2006) have one or more antisense transcript on the opposite strand. These SAT genes correspond to 35% of non-redundant gene symbol names (6347 of the 18 160 counted), and thus tend to be alternatively spliced (GO analysis in Results section). We increased the estimates of SAT gene pairs up to 35% of RefSeq protein coding genes. We have been able to derive such result due to our advanced computational algorithm, improvement in genome annotation and due to our integration of a large number of databases on SATs. However, it is still not possible to take a human genome sequence and accurately predict which RNA species it will produce (54). In particular, many anti-sense transcripts of protein-coding genes are still unknown (11,54–56). Based on our estimates, additional ~1500 SAT in annotated mRNA–EST, Refseq–EST and EST–EST pairs should be predicted. The start and end position of many coding and non-coding genes are poorly defined (54). Thus, further gene cartography, gene classification and more complete transcript identification will lead to improvement of annotation systems and correction of the number of CASGP and SAT in USAGP and the database completeness.

We and others showed that a protein-coding–non-protein coding transcript pairing pattern dominates in SAT pairs overlaps. According to our estimates, the numbers of SAT with at least one protein coding transcript (defined by RefSeq ID) is 5978, which is two times higher than the number of SAT pairs observed for non-protein–coding transcript pairs (2916 pairs) and 10 times higher than the number of pairs with protein coding transcripts on both strands (677 pairs). Statistics of SA overlaps show a variety of pairing types and protein coding/non-coding transcript combinations. Our GO analysis of the most reliable set of SAT overlap genes (SAT pairing RefSeq genes) shows a strong enrichment of the genes essential in many important regulatory processes of cells (protein binding, nucleotide binding, DNA binding, metal binding, coiled-coil, ATP binding, phosphorylation, transcription, chromosome translocation, disease mutation, etc.). Specifically, GO analysis demonstrates the essential contribution of genes of the SAT gene pair overlapping regions to alternative splicing ($P = 1.4E–126$). The mutations and splicing associated with SA pairing could be important sources of functional diversity and instability of transcriptional and post-transcriptional events involved in cancer and other diseases (14,20,21). Over-expression of oncogenes in SAT pairs can be in synergetic manner correlated with aggressiveness of cancer patients (57).

The present study revealed strong domination of convergent protein-coding SAT gene pairs (RefSeq/RefSeq) over other types of SAT pairs (example in Figure 2). This finding is consistent in several reports (1,52,53). The number of SAT gene pairs in NAT DB (4) with both representative genes mapped to HomoloGene is 520 in human genome and 480 in mouse genome. Among them, 155 human SAT gene pairs also overlap in mouse. 120 of those overlapping gene pairs maintain the convergent overlapping pattern and only eight maintain divergent patterns (see also distribution of total non-redundant RefSeq/RefSeq SAT pairs and such pairs in NAT DB in Supplementary Tables S1 and S2). Is such a bias related to SAT gene expression regulation of overlapping genes? To address this question, Sun *et al.* (58), observed that putative SAT pairs overlapping at 3′-UTRs have a much higher evolutionary rate between

human and mouse genomes than do those overlapping at their 5′-UTRs. They suggested that the function of many 3′-UTR-to-3′-UTR overlaps might be related to SAT pair regulation. These analyses allow us to suggest that for functional protein-coding gene pairs, the 3′-UTRs of the gene pairs (in comparison to 5′-UTRs) may often contain the evolutionarily conserved reverse complementary regions co-evolved in the positive selection manner and thus provide co-expression regulation of corresponding CASGPs.

However, detailed analysis of other types of SAT pairs shows more diverse patterns. For pairing of RNA and ESTs, however, our analysis provides a more uniform distribution of by convergent, divergent and others types, whereas in case of the SAT pairs containing mRNA and EST ('mRNA/Any' and 'EST/Any') divergent and embedded ('other') types of SAT pairing might dominate.

In different tissues (including normal and cancerous brain, breast and others) the co-regulation (concordant) expression pattern was dominated versus anti-regulation pattern (Figure 4). These results are consistent with our observation reported for human breast cancer subtypes (57) and are consistent with several other reports (10,13,52). However, mechanisms of such bias are mostly unknown.

USAGP includes a unique subclass of *cis*-antisense CDS/CDS overlapping genes coding for proteins. Figure 2 shows an example of CDS/CDS SA pair (NPR2 and SPAG8). Such class of CASGPs have not been described yet.

This is the first report of structural and functional association between SATs and Dicer-associated ncRNA biogenesis in the human sperm cells. We identified a new type of testis-specific SFRM which is composed of SA protein-coding genes (DALRD3-NDUFAF3) and two ncRNA genes (*miRNA-425* and *miRNA-191*) embedded into DALRD3 gene. Based on GO analysis of putative targets of the miRNAs, we suggest that this module might be involved in miRNA-mediated control of normal spermatogenesis and could be inactivated in severe teratozoospermia as a result of severe genome defects (e.g. deletion and/or epigenomic modification). We suggest a model, in which a transcription of this SFRM is not only coordinated with expression of DROSHA (RNASEN) and PIWI complex pathways, but also (together with miRNA *let-7d*) does strong association with silencing of DICER1 gene expression.

The molecular pathways of genetic defects in spermatogenesis are not known. In conditional knockout mouse in which DICER1 was specifically deleted in germ cells of testis, spermatogenesis was retarded at early stage of proliferation and/or early differentiation (59). In severe teratozoospermia, we found switch-off/suppression of all components of the SFRM together with the DROSHA and PIWI pathways, biogenesis of miRNA and long non-coding RNAs (MALAT1). DICER1 and basonuclein-2 switch-on in severe teratozoospermia; probably, they play a pathobiological role associated with maturation abnormalities in spermatogenesis (43,47) and expression of embryonic stem cell-like genotype (Nanog, Sox4).

Recently, a mechanism for a miRNA/DICER1 auto-regulatory negative feedback loop has been described in which *let-7* miRNA plays a role as an important miRNA for implementing the tightly regulated, equilibrated state of DICER1 and various miRNAs (45,46). It was experimentally demonstrated that the members of *let-7* miRNA family directly target the DICER1 within its coding sequence (46). We suggest that *mir-425* targeting DICER1 transcription could play an important role in auto-regulatory negative feedback of ncRNA biogenesis. Interestingly, switch-on and switch-off functions of miRNAs in DICER1 and other proteins of miRNA-induced silencing mathinery can often be found in cancer cells (60).

Due to our analysis, DALRD3-NDUFAF3 SFRM may be involved in tissue-specific regulation of transcription of many essential genes (e.g. mRNA biogenesis, cancer-associated genes, oncogenesis and cell differentiation genes) (Suplementary Tables S5–S8) and thus, it may form the tightly interconnected network driving ncRNA biogenesis.

One would ask a question whether the pool of miRNA precursors located in gene spans of SATs have any specific functional characteristics compared to other genome-wide pool of miRNA genes? This question deserves detailed theoretical consideration and special experimental validation and will be addressed in our future studies. USAGP database could be used as a valuable source for further detailed and comprehensive study of these mechanisms.

In summary, our study provides integrative and deep quantitative insight showing that SATs are specifically different in comparison to transcripts of other genes and therefore these transcripts could be considered as distinct *anti-sense sub-transcriptome*. Our integrated data analysis provides strong support of this suggestion due to (i) non-uniform chromosomal localization of SAT (1,4), (ii) aggregation of SAT into complex genome architectures, (iii) association of transcription of protein-coding and non-protein coding RNAs with antisense gene pairing, (iv) miRNA precursors production, (v) high diversity of RNA isoforms and splice variants produced by antisense genes and (vi) higher probability of CASGPs being coordinately expressed comparing to other (random) gene pairs. The mechanisms by which SAT are co-regulated and act as local and distant SFRMs in different cell types could be very diverse (2,16,18,55,61). Understanding these mechanisms demand further development of annotation systems, integrative data analysis and systematic experimental validation.

For the first time biological significance of NDUFAF3-DALRD3 SA genes as well as *mir-425* and *mir-191* in normal spermatogenesis and theratozoospermia is suggested. By our data-search algorithm and data-driven model, in normal spermatogenesis a strong co-expression and co-regulation pattern of *mir-425* and *mir-191* precursors together with NDUFAF3-DALRD3 SA genes are observed. The expression of this structural–functional module could be functionally related to (i) observed silencing of DICER1-mediated and

BNC2-mediated pathways (which are putative targets of *mir-425* and by *miRNA 191*) and (ii) co-expression of key genes of pre-ncRNA biogenesis (DROSHA, DGCR8, AGO2, PIWi, SOX4). In teratozoospermia condition, these key pre-ncRNA biogenesis genes and ncRNA biogenesis are mostly suppressed, but DICER1 and BNC2 expression is occurred.

Many other severe genome abnormalities in expression of tissue-specific and important regulatory genes might be associated with alteration in DALRD3, NDUFAF3, *mir-425, mir-191* expression pattern. It is anticipated that our hypothetical model of SAT-ncRNA auto-regulation of miRNA biogenesis will prove to be valuable for understanding of normal spermatogenesis and male infertility.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful to L. Lipovich for kindly providing manually curated data on SA transcription units. They thank V. Tanavde, A. Yarmishev, I. Kurochkin and A. Giannakakis for review of the manuscript and helpful comments and critiques.

FUNDING

Biomedical Research Council of A*STAR (Agency for Science, Technology and Research), Singapore. Funding for open access charge: Bioinformatics Institute, A*Star, Singapore.

Conflict of interest statement. None declared.

REFERENCES

1. Engstrom,P.G., Suzuki,H., Ninomiya,N., Akalin,A., Sessa,L., Lavorgna,G., Brozzi,A., Luzi,L., Tan,S.L., Yang,L. *et al.* (2006) Complex Loci in human and mouse genomes. *PLoS Genet*., **2**, e47.
2. Kapranov,P., Willingham,A.T. and Gingeras,T.R. (2007) Genome-wide transcription and the implications for genomic organization. *Nat. Rev. Genet*., **8**, 413–423.
3. Kuznetsov,V.A., Knott,G.D. and Bonner,R.F. (2002) General statistics of stochastic process of gene expression in eukaryotic cells. *Genetics*, **161**, 1321–1332.
4. Zhang,Y., Liu,X.S., Liu,Q.R. and Wei,L. (2006) Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species. *Nucleic Acids Res*., **34**, 3465–3475.
5. Johnson,Z.I. and Chisholm,S.W. (2004) Properties of overlapping genes are conserved across microbial genomes. *Genome Res*., **14**, 2268–2272.
6. Rogozin,I.B., Spiridonov,A.N., Sorokin,A.V., Wolf,Y.I., Jordan,I.K., Tatusov,R.L. and Koonin,E.V. (2002) Purifying and directional selection in overlapping prokaryotic genes. *Trends Genet*., **18**, 228–232.
7. Shendure,J. and Church,G.M. (2002) Computational discovery of sense-antisense transcription in the human and mouse genomes. *Genome Biol*., **3**, RESEARCH0044.
8. Veeramachaneni,V., Makalowski,W., Galdzicki,M., Sood,R. and Makalowska,I. (2004) Mammalian overlapping genes: the comparative perspective. *Genome Res*., **14**, 280–286.
9. Makalowska,I., Lin,C.F. and Makalowski,W. (2005) Overlapping genes in vertebrate genomes. *Comput. Biol. Chem*., **29**, 1–12.
10. Yelin,R., Dahary,D., Sorek,R., Levanon,E.Y., Goldstein,O., Shoshan,A., Diber,A., Biton,S., Tamir,Y., Khosravi,R. *et al.* (2003) Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol*., **21**, 379–386.
11. Ge,X., Wu,Q., Jung,Y.C., Chen,J. and Wang,S.M. (2006) A large quantity of novel human antisense transcripts detected by LongSAGE. *Bioinformatics*, **22**, 2475–2479.
12. Seno,S., Takenaka,Y., Kai,C., Kawai,J., Carninci,P., Hayashizaki,Y. and Matsuda,H. (2006) A method for similarity search of genomic positional expression using CAGE. *PLoS Genet*., **2**, e44.
13. Katayama,S., Tomaru,Y., Kasukawa,T., Waki,K., Nakanishi,M., Nakamura,M., Nishida,H., Yap,C.C., Suzuki,M., Kawai,J. *et al.* (2005) Antisense transcription in the mammalian transcriptome. *Science*, **309**, 1564–1566.
14. Enerly,E., Sheng,Z. and Li,K.B. (2005) Natural antisense as potential regulator of alternative initiation, splicing and termination. *In Silico Biol*., **5**, 367–377.
15. Henderson,C.M., Anderson,C.B. and Howard,M.T. (2006) Antisense-induced ribosomal frameshifting. *Nucleic Acids Res*., **34**, 4302–4310.
16. Orfanelli,U., Wenke,A.K., Doglioni,C., Russo,V., Bosserhoff,A.K. and Lavorgna,G. (2008) Identification of novel sense and antisense transcription at the TRPM2 locus in cancer. *Cell Res*., **18**, 1128–1140.
17. Gallagher,E., Mc Goldrick,A., Chung,W.Y., Mc Cormack,O., Harrison,M., Kerin,M., Dervan,P.A. and Mc Cann,A. (2006) Gain of imprinting of SLC22A18 sense and antisense transcripts in human breast cancer. *Genomics*, **88**, 12–17.
18. Yu,W., Gius,D., Onyango,P., Muldoon-Jacobs,K., Karp,J., Feinberg,A.P. and Cui,H. (2008) Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature*, **451**, 202–206.
19. Ogawa,Y. and Lee,J.T. (2002) Antisense regulation in X inactivation and autosomal imprinting. *Cytogenet. Genome Res*., **99**, 59–65.
20. Alfano,G., Vitiello,C., Caccioppoli,C., Caramico,T., Carola,A., Szego,M.J., McInnes,R.R., Auricchio,A. and Banfi,S. (2005) Natural antisense transcripts associated with genes involved in eye development. *Hum. Mol. Genet*., **14**, 913–923.
21. Guo,J.H., Cheng,H.P., Yu,L. and Zhao,S. (2006) Natural antisense transcripts of Alzheimer's disease associated genes. *DNA Seq*., **17**, 170–173.
22. Li,Y.Y., Qin,L., Guo,Z.M., Liu,L., Xu,H., Hao,P., Su,J., Shi,Y., He,W.Z. and Li,Y.X. (2006) In silico discovery of human natural antisense transcripts. *BMC Bioinformatics*, **7**, 18.
23. Coriton,O., Lepourcelet,M., Hampe,A., Galibert,F. and Mosser,J. (2000) Transcriptional analysis of the 69-kb sequence centromeric to HLA-J: a dense and complex structure of five genes. *Mamm. Genome*, **11**, 1127–1131.
24. Kim,V.N. (2005) Small RNAs: classification, biogenesis, and function. *Mol. Cells*, **19**, 1–15.
25. Borsani,O., Zhu,J., Verslues,P.E., Sunkar,R. and Zhu,J.K. (2005) Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis. *Cell*, **123**, 1279–1291.
26. Carlile,M., Nalbant,P., Preston-Fayers,K., McHaffie,G.S. and Werner,A. (2008) Processing of naturally occurring sense/antisense transcripts of the vertebrate Slc34a gene into short RNAs. *Physiol. Genomics*, **34**, 95–100.
27. Li,L.C., Okino,S.T., Zhao,H., Pookot,D., Place,R.F., Urakami,S., Enokida,H. and Dahiya,R. (2006) Small dsRNAs induce transcriptional activation in human cells. *Proc. Natl Acad. Sci. USA*, **103**, 17337–17342.
28. Morris,K.V., Santoso,S., Turner,A.M., Pastori,C. and Hawkins,P.G. (2008) Bidirectional transcription directs both transcriptional gene activation and suppression in human cells. *PLoS Genet*., **4**, e1000258.
29. Faghihi,M.A. and Wahlestedt,C. (2006) RNA interference is not involved in natural antisense mediated regulation of gene expression in mammals. *Genome Biol*., **7**, R38.

30. Orlov,Y.L., Zhou,J., Lipovich,L., Shahab,A. and Kuznetsov,V.A. (2007) Quality assessment of the Affymetrix U133A&B probesets by target sequence mapping and expression data analysis. *In Silico Biol.*, **7**, 241–260.
31. Chen,J., Sun,M., Kent,W.J., Huang,X., Xie,H., Wang,W., Zhou,G., Shi,R.Z. and Rowley,J.D. (2004) Over 20% of human transcripts might form sense-antisense pairs. *Nucleic Acids Res.*, **32**, 4812–4820.
32. Ivshina,A.V., George,J., Senko,O., Mow,B., Putti,T.C., Smeds,J., Lindahl,T., Pawitan,Y., Hall,P., Nordgren,H. *et al.* (2006) Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res.*, **66**, 10292–10301.
33. Xie,X., Wang,Z. and Chen,Y. (2007) Association of LKB1 with a WD-repeat protein WDR6 is implicated in cell growth arrest and p27(Kip1) induction. *Mol. Cell Biochem.*, **301**, 115–122.
34. Kasashima,K., Nakamura,Y. and Kozu,T. (2004) Altered expression profiles of microRNAs during TPA-induced differentiation of HL-60 cells. *Biochem. Biophys. Res. Commun.*, **322**, 403–410.
35. Nakamura,T., Canaani,E. and Croce,C.M. (2007) Oncogenic All1 fusion proteins target Drosha-mediated microRNA processing. *Proc. Natl Acad. Sci. USA*, **104**, 10980–10985.
36. Volinia,S., Calin,G.A., Liu,C.G., Ambs,S., Cimmino,A., Petrocca,F., Visone,R., Iorio,M., Roldo,C., Ferracin,M. *et al.* (2006) A microRNA expression signature of human solid tumors defines cancer gene targets. *Proc. Natl Acad. Sci. USA*, **103**, 2257–2261.
37. Landgraf,P., Rusu,M., Sheridan,R., Sewer,A., Iovino,N., Aravin,A., Pfeffer,S., Rice,A., Kamphorst,A.O., Landthaler,M. *et al.* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**, 1401–1414.
38. Scharer,C.D., McCabe,C.D., Ali-Seyed,M., Berger,M.F., Bulyk,M.L. and Moreno,C.S. (2009) Genome-wide promoter analysis of the SOX4 transcriptional network in prostate cancer cells. *Cancer Res.*, **69**, 709–717.
39. Vanhoutteghem,A. and Djian,P. (2006) Basonuclins 1 and 2, whose genes share a common origin, are proteins with widely different properties and functions. *Proc. Natl Acad. Sci. USA*, **103**, 12423–12428.
40. Hall,L.L., Smith,K.P., Byron,M. and Lawrence,J.B. (2006) Molecular anatomy of a speckle. *Anat. Rec. A Discov. Mol. Cell Evol. Biol.*, **288**, 664–675.
41. Pawlicki,J.M. and Steitz,J.A. (2009) Subnuclear compartmentalization of transiently expressed polyadenylated pri-microRNAs: processing at transcription sites or accumulation in SC35 foci. *Cell Cycle*, **8**, 345–356.
42. Lian,J., Zhang,X., Tian,H., Liang,N., Wang,Y., Liang,C., Li,X. and Sun,F. (2009) Altered microRNA expression in patients with non-obstructive azoospermia. *Reprod. Biol. Endocrinol.*, **7**, 13.
43. Platts,A.E., Dix,D.J., Chemes,H.E., Thompson,K.E., Goodrich,R., Rockett,J.C., Rawe,V.Y., Quintana,S., Diamond,M.P., Strader,L.F. *et al.* (2007) Success and failure in human spermatogenesis as revealed by teratozoospermic RNAs. *Hum. Mol. Genet.*, **16**, 763–773.
44. Wang,Y.P. and Li,K.B. (2009) Correlation of expression profiles between microRNAs and mRNA targets using NCI-60 data. *BMC Genomics*, **10**, 218.
45. Forman,J.J., Legesse-Miller,A. and Coller,H.A. (2008) A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. *Proc. Natl Acad. Sci. USA*, **105**, 14879–14884.
46. Tokumaru,S., Suzuki,M., Yamada,H., Nagino,M. and Takahashi,T. (2008) let-7 regulates Dicer expression and constitutes a negative feedback loop. *Carcinogenesis*, **29**, 2073–2077.
47. Poongothai,J., Gopenath,T.S. and Manonayaki,S. (2009) Genetics of human male infertility. *Singapore Med. J.*, **50**, 336–347.
48. Xu,M., Xiao,J., Chen,J., Li,J., Yin,L., Zhu,H., Zhou,Z. and Sha,J. (2003) Identification and characterization of a novel human testis-specific Golgi protein, NYD-SP12. *Mol. Hum. Reprod.*, **9**, 9–17.
49. Wilusz,J.E., Freier,S.M. and Spector,D.L. (2008) 3′ end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell*, **135**, 919–932.
50. Lehner,B., Williams,G., Campbell,R.D. and Sanderson,C.M. (2002) Antisense transcripts in the human genome. *Trends Genet.*, **18**, 63–65.
51. Zhang,Y., Li,J., Kong,L., Gao,G., Liu,Q.R. and Wei,L. (2007) NATsDB: Natural Antisense Transcripts DataBase. *Nucleic Acids Res.*, **35**, D156–161.
52. Galante,P.A., Vidal,D.O., de Souza,J.E., Camargo,A.A. and de Souza,S.J. (2007) Sense-antisense pairs in mammals: functional and evolutionary considerations. *Genome Biol.*, **8**, R40.
53. Werner,A., Schmutzler,G., Carlile,M., Miles,C.G. and Peters,H. (2007) Expression profiling of antisense transcripts on DNA arrays. *Physiol. Genomics*, **28**, 294–300.
54. Buratowski,S. (2008) Transcription gene expression–where to start? *Science*, **322**, 1804–1805.
55. He,Y., Vogelstein,B., Velculescu,V.E., Papadopoulos,N. and Kinzler,K.W. (2008) The antisense transcriptomes of human cells. *Science*, **322**, 1855–1857.
56. Peters,B.A., St Croix,B., Sjoblom,T., Cummins,J.M., Silliman,N., Ptak,J., Saha,S., Kinzler,K.W., Hatzis,C. and Velculescu,V.E. (2007) Large-scale identification of novel transcripts in the human genome. *Genome Res.*, **17**, 287–292.
57. Kuznetsov,V.A., Zhou,J., George,J. and Orlov,Y.L. (2006) Genome-wide co-expression patterns of human cis-antisense gene pairs. *Proceedings of the Fifth International Conference on Bioinformatics of Genome Regulation and Structure*, Vol. 1. Novosibirsk, Inst. of Cytology & Genetics, pp. 90–93.
58. Sun,M., Hurst,L.D., Carmichael,G.G. and Chen,J. (2005) Evidence for a preferential targeting of 3′-UTRs by cis-encoded natural antisense transcripts. *Nucleic Acids Res.*, **33**, 5533–5543.
59. Hayashi,K., Chuva de Sousa Lopes,S.M., Kaneda,M., Tang,F., Hajkova,P., Lao,K., O'Carroll,D., Das,P.P., Tarakhovsky,A., Miska,E.A. *et al.* (2008) MicroRNA biogenesis is required for mouse primordial germ cell development and spermatogenesis. *PLoS ONE*, **3**, e1738.
60. Sotiropoulou,G., Pampalakis,G., Lianidou,E. and Mourelatos,Z. (2009) Emerging roles of microRNAs as molecular switches in the integrated circuit of the cancer cell. *RNA*, **15**, 1443–1461.
61. Munroe,S.H. and Zhu,J. (2006) Overlapping transcripts, double-stranded RNA and antisense regulation: a genomic perspective. *Cell Mol. Life Sci.*, **63**, 2102–2118.