A novel class of endogenous shRNAs in human cells

Tongjun Gu[1, 2], James Q. Yin[1*], Yanwei Xu [1], Zhenghua Dai[3], Zhengge Qiu[3], Shenzhong Feng [3], Xiang Yi[1, 2] , Ling Jiang[1, 2] and Hongjie Zhang[1]

1. National Laboratory of Biomacromolecules, Center for Computing and Systems Biology, Institute of Biophysics, Chinese Academy of Sciences, 15 Datun Road, Beijing 100101
2. Graduate School of Chinese Academy of Sciences, 19 YuQuan Road, Beijing 100049
3. Key Laboratory of Computer System and Architecture, Institute of Computing Technology, Chinese Academy of Sciences, 6 Kexueyuan South Road, Beijing 100080

Key words: miRNA; shRNA; microarray; QRT-PCR, gene expression,

*To whom correspondence should be addressed.

Dr. James Q Yin
Institute of Biophysics, CAS
15 Datun Road
Chaoyang District
Beijing 100101
Tel: 86-010-64888572
Fax: 86-010-64888572
E-mail: jqwyin@sun5.ibp.ac.cn

Abstract

Until now it is still not clear how many types and amounts of small RNAs (sRNAs) exist in humans. Here we report the identification of 1258 distinct sRNAs derived from intronic regions of protein-coding genes in human with a new approach. These endogenous short hairpin RNAs (shRNAs) appear to be similar to exogenous shRNAs in structure, have a broad distribution in the stem length, and function as microRNAs (miRNAs), small interfering RNAs (siRNA) and/or piwi-interacting RNAs (piRNAs). Except for a few shRNAs, the majority of shRNAs are not phylogenetically conserved. They are differentially expressed in different cells and at diverse developmental stages. Overall, their expression levels are lower than miRNAs', but can be detected by quantitative real-time PCR and microarrays, implying that like other known sRNAs, this type of shRNAs should have important functions in modulating gene expression, and that they may exist in other genomic regions and many species.

Introduction

A number of findings indicated that introns involve many important functions including the generation of protein variability, the regulation of mRNA metabolisms, the encoding of diverse small RNA molecules and the implication of diseases [1-3]. About 127 human microRNAs (miRNAs) have been reported to locate within the introns of either non-protein-coding or protein-coding transcription units [4,5]. The expression of these miRNAs largely coincides with the transcription of their hosting genes under the control of RNA polymerase II (Pol II) [6,7]. Recently, it has been showed that some human miRNAs can be regulated by RNA Pol III through associated repetitive elements such as Alu [8].

Most of the mature miRNAs identified so far originate from the stem region of foldback structures with imperfect base-pairs whereas short interfering RNAs (siRNA) with complete complementarity are generated from long exogenous or endogenous dsRNA molecules (very long hairpins or RNA duplexes) or short hairpin RNAs (shRNA) made by in vitro chemical synthesis or in vivo from RNA polymerase III promoters [9-11]. The RNase III endonuclease Drosha and Dicer have been shown to involve the biogenesis of these hairpin molecules [12, 13]. The mature miRNA strand generated by Dicer is then loaded into a RNA-induced silencing complex (RISC) that includes a member of the Argonaute protein family at its core [11,14]. Further investigation on miRNAs and siRNA demonstrates that some pre-miRNAs have the completely complementary stem as shRNA does while some siRNA can function as miRNAs [15]. Even though the promoter based expression of shRNAs has been widely used for the stable gene silencing in mammalian cells and in vivo therapeutic application in humans [11, 16], it is still not clear how many this type of endogenous shRNAs exist within the intronic regions of human and other species, and what are their cellular functions.

Some small RNAs (sRNA) cannot be attained by molecular cloning because of their low abundance or tissue- and developmental stage-specificity affecting their representation in different tissues and cells [17, 18]. Moreover, direct miRNA cloning is not well suited to large-scale discovery efforts and may have already reached the point of diminishing returns [19]. In order to overcome this limitation, computational prediction and other experimental analysis have been developed to discover new small RNAs[17, 20]. Especially, recent advances in high-throughput sequencing technology and tiling microarrays have allowed for a more complete assessment of the global small RNA population in plants and animals [21-24]. Thus, more and more unexpected small RNAs including miRNAs in different species are being discovered, suggesting that the world of human sRNAs is larger than initially believed and is not limited only to those conserved sequences [25].

Now we add a new class of small RNAs to the ever-growing list. In this study, we have developed an integrative method that takes advantage of the characteristic hairpin structure found in pri-miRNA and pre-miRNA precursor structures and the traits of extensive complementarity in siRNA to predict the existence of any shRNAs within intronic regions of protein-encoding genes. By using custom shRNA arrays and quantitative reverse transcription polymerase chain reaction (qRT-PCR), the differential expression of these shRNAs can be detected and analyzed in different human cells.

Results

Computational prediction of endogenous shRNAs

To predict endogenous short hairpin RNAs (shRNAs) by computational methods, we defined sequence and structure properties that differentiate putative *human* shRNA sequences from random genomic sequences and other sRNA sequences, and used these properties as constraints to screen intronic regions of protein-encoding genes in human for putative shRNAs.

Six filters were used to screen a base set of intronic sequences as candidate shRNAs. (1) Sequences of these endogenous shRNA precursors have the maximum length (≤140nts) with the G+C contents ranging from 25% to 80%. (2) The predicted secondary structure of their precursors, or at least one precursor if an shRNA has multiple genomic loci, has a loop length ranging from 5 to 80 nucleotides. (3) Their stems have consecutive more than 19-bps involving the first 20-nts. For the second 10-20nts if the stem is enough long, the number of bulged or asymmetrically unpaired nucleotides is no more than 2 and the number of consecutive unpaired nucleotides is no more than 3. (4) These sequences can potentially fold into hairpin structures with the lowest minimum free energy of folding (MFE) less than –25kcal/mol, (5) All the shRNAs are complementary to their target regions such as promoter or /and coding area with at least 20 base-pairings in their 5' termini, and to the 5' or/and 3' UTR regions of target mRNAs with at least 10 base-pairings in their 5' termini. (6) In the hairpin structures formed by shRNA precursors, all mature shRNAs (<35nts) are supposed to be in either or both the stem region of the foldback structure that is similar to those observed in *exogenous sh*RNA precursors or miRNA precursors. The overall screen and validation processes were shown in Figure 1.

The human genomic sequences were oriented to the corresponding +/- DNA strands within which the intronic shRNA genes reside, and then scanned with a predefined sliding window (size of 140-nts in 1-nt steps) for potential hairpins. The intronic regions satisfying these criteria were reserved for classification. Scanning the 237999 intronic sequences identified 5888 candidate hairpin structures by using our algorithm. Of them, 2654 human hairpins could pass through all six filters. We took these sequences as an initial dataset for the further study (Supplementary Tab. 1). As shown in Figure 2A, these predicted sequences could be mapped onto all the chromosomes. Chromosomes 1 and 2 appear to contain more shRNA genes than other chromosomes whereas chromosomes 21 and Y seem to hold less shRNA genes. These shRNA genes found in this study are originated from both the sense and antisense strands of the chromosomes (Figure 2B and 2C). The sense and antisense strands appear to encode almost equal amount of shRNAs. As shown in Figure 2C, these shRNAs are different from miRNAs and snoRNAs in the distribution patterns along human chromosomes 7 and others. However, these shRNAs have similar fold-back secondary structures to miRNAs. The predicted secondary structures show that there are at least 20 nucleotides engaged in Watson–Crick base pairings between the mature shRNA and the opposite arms (shRNA*) in the hairpin structure, and all stem–loop precursors do not contain large internal loops or bulges (Figure 2D). These sequences were divided into two major classes, the repeat-derived and nonrepeat-derived ones. In each class, the sequences were further subdivided into the unique (Supplementary Tab. 1.1 and 1.2) and multiple sequence groups (Supplementary Tab.1.3 and Tab. 1.4) (Fig. 2a and 2b). After removing those homologous sequences, all distinct sequences were listed in Supplementary Table 2 for further study. Because of their small size, and the same structure as exogenous short-hairpin RNA, we referred to these novel small RNAs as endogenous shRNAs. According to the nomenclature of miRNAs, the shRNAs we investigated are abbreviated as shR-1 to shR-1258 in Table 2, and the genes encoding shRNAs are named *shr-1* to *shr-1258.* Highly homologous shRNAs are referred to by the same gene number, but followed by a lowercase letter; multiple genomic copies of an *shr* gene are annotated by adding a dash and a number.

Several lines of studies reported that a number of mammalian miRNAs are in fact derived from transposable elements (TEs) [26]. In order to examine whether the shRNA genes also hold TE sequences, we conserved TE sequences in this study. As shown in Figure 3a, a total of 2564 shRNAs are located within 2069 host protein-coding genes and 2346 introns, the majority of which contain only one shRNAs. However, some genes or introns can hold more than 5 shRNA precursors. Careful analysis of human endogenous shRNAs revealed four major classes. The largest class (~50.4%) corresponded to repeats with multiple loci in the genome while a second class of shRNAs (15.3%) was also related to repeats mapping uniquely in the genome (Fig. 3b). In these two classes, there were short interspersed elements (SINEs) (66%), DNA-elated repeat sequences (12%), long interspersed elements (LINEs) (4%), long terminal repeat (LTR) retrotransposons

(2%) and other repeats (16%) (Supplementary Tab. 1.5). A third class was associated with unique nonrepeat-derived sequences (30.5%) while the fourth class was composed of those nonrepeat-derived sequences with multiple members. To focus our concerns to distinct shRNA sequences, we further selected some representatives out of the repeat- and nonrepeat- derived with multiple loci in the genome, and then combined them with all the repeat- and nonrepeat-derived with unique loci into the supplementary table 2.

Next we attempted to determine the identity of these sequences. The foldback precursors of these potential shRNAs are usually about 55–140 nucleotides in length (Supplementary Tab. 1). According to the difference in the length of shRNA precursors, we divided these shRNAs into five groups, and then calculated separately the minimal folding free energies (MFE) ranges and means of each group by using MFOLD [27] (Fig. 3c and Supplementary Tab. 1). These newly predicted human shRNA precursors have negative minimal folding free energies (MFE) (26–140 kcal mol) whereas the average adjusted MFE (AMFE) of 2564 shRNA precursors was -63±6.28 kcal/mol. The distribution of these parameters is even much lower than the mean values of 513 plant miRNA precursors (-45.93± 9.43 kcal /mol) because plant miRNA precursors have significantly lower AMFEs than other types of RNA, including tRNAs (–32.67 ± 6.47 kcal/mol), rRNAs (–33.10 ± 2.56 kcal/mol), and mRNAs (–30.44 ± 2.08 kcal/mol) [28]. The finding revealed that endogenous shRNAs are very stable in the secondary structure. Why do these endogenous shRNAs have such high negative AMFEs? We conducted a detailed analysis on the complementary extents of shRNA stems. The data indicated that the majority of repeat-associated shRNAs have longer base-pairings (> 35-40 bps) than nonrepeat-associated shRNAs do whereas nonrepeat-derived shRNAs prefer to have shorter stems (>20-25bps) (Fig. 3d). Our data suggest that the broad length distribution, ranging from 20–40bps, of shRNA stems may reflect the versatility of their biogenesis, biological functions and modes of action in gene silencing pathways, and that these shRNA precursors may be processed into one to two mature shRNAs with different lengths by different interactor proteins such as Drosha, Dicer, AGO3 or others[11-14].

Comparison of conserved shRNAs

For mammals it has been suggested that the more targets a microRNA has the more likely it is to be conserved [29] because of the additional constraints of having to match multiple targets. It is interesting whether the human shRNAs identified in this study are related by sequence to mouse or rat shRNAs. We used the set of human hairpin sequences to search the mouse or rat genome for corresponding hairpins (Supplementary Tab. 2.1 and 2.2). Results yielded a set of 9 hairpins with phylogenetical conservation in the stem region across human and mouse or human and rat, with >85% sequence identity between human and mouse/rat (Figure 4a). Of these conserved shRNAs, some has been composed of many members. For instance, shRNA 860 has 11 and13 members in human and mouse, respectively. The significant sequence conservation between the mouse and human *shR-860* is observed only in the 5'- and 3'-stems but not within the loop of predicted hairpin precursor (Fig. 4a). A complete stem conservation could be seen between the human *shR-399* and its corresponding sequence in rat, suggesting that an important role in biogenesis or localization. Careful observation indicated that most conserved shRNAs are repeat-associated. On the other hand, although these stem segments are conserved as mature shRNAa and//or shRNAs sequences, the other parts of shRNA precursor differ widely. Most shRNAs are on the poorly conserved with the similarity to the level of conservation observed with introns. The alignment of other mouse, rat and human shRNAs does not show such high conservation beyond the most predicted hairpin precursors. So, these data also suggest that the majority of distinct shRNAs show rapid sequence evolution or a limited number of target genes.

Validation of predicted shRNAs by other methods

In order to verify whether these predicted shRNAs are functional and expressed in human cells, we sought direct and indirect evidence to help validate the proposed set of endogenous shRNAs. In this section, we compared predicted shRNAs with piRNAs and miRNAs discovered in mammals, transfrags detected by tiling arrays, and cloned sequences in EST database (Fig 4b).

Expressed Sequence Tags (ESTs) are partial cDNA sequences of expressed genes [30]. After searching the EST database, we were able to acquire evidence for the expression of a small number of the predictions. A total of 100 ESTs were identified that contained the precursor sequences of potential shRNA homologs predicted in this study (Supplementary Tab. 3). However, this detection rate is very low. The possible explanation is that a majority of shRNA precursors are shorter in length and are rapidly processed in the nucleus so that these sequences have a lower probability of being cloned to ESTs. Of 53 human shRNAs, 52 shRNAs were found in one to three ESTs. For example, shR-265 precursor was found to reside within EST DB064688 while shR-280 sequence could be detected in ESTs AI732240 and AA579322. Human shRNA 55 exists in 25 different ESTs. These data provided the evidence that the fifty-three predicted shRNA candidates with corresponding cloned sequences in the EST database should be annotated as *bona fide* shRNAs, and that the identification of new shRNA homologs by mining the repository of available ESTs may be a useful strategy.

Recently, a report indicated that the developmentally regulated piwi-interacting RNAs (piRNAs), 26–35 nt in length, could inhibit transposons in mammals. It is not clear that the precursor form of piRNA primary transcript is single or double-stranded [31]. Recent investigation indicated that the biogenesis of piRNA might be related to the dsRNAs formed by base pairing of the terminal inverted repeats of the transposon in a fold-back structure. To investigate whether there exist any relationships between piRNAs and endogenous shRNAs, we searched for any shRNAs with inverted repeats in the piRNA dataset. Such inverted repeats may form precursors containing shRNA that initiate enzymatic processing of piRNA. As expected, we found that 20 known piRNAs were bracketed by 255 inverted repeats, i.e., the complementary segments contained the sequences of piRNAs (Supplementary Tab. 4). For example, the antisense strand of shR-693 precursor had been found to hold the whole sequence of piR-51815 and piR-54220 while piR-63025 existed in shR-132a, suggesting that shRNA derived from transposable elements may involve the biogenesis of piRNAs. Furthermore, of them some could be detected by microarray and qRT-PCR assay. It seems to imply that some piRNAs may play a role in regulating gene silencing in somatic cells beside germ cells

Subsequently, we searched for possible *shRNA* homologs of 711 miRNAs stored in the miRNA database of *H. sapiens.* We employed a Smith-Waterman algorithm to compare each predicted hairpin to each of the 711 miRNA sequences. We obtained 5 candidate homolog hairpins with exact matches to the known miRNAs (Supplementary Tab. 5), For example, human shRNA-38s was exactly the same as miR-486. Although most of these shRNAs do not bear sequence similar to the known miRNA genes, all 5 known microRNAs with the perfectly complementary stem could be selected out by our bioinformatics approach whereas other miRNAs with the imperfect base-pairing stem did not fall into our candidates, suggesting that our systematic scanning indeed detect all endogenous shRNAs derived from introns and that shRNAs may be another novel class of small RNAs different from miRNAs in structure. It may be more reasonable and suitable to reclassify those 5 known microRNAs with the perfect base-pairings into the class of shRNAs.

Sequences in the tiling array database were detected by hybridizing the capture probes with the expressed short RNA molecules [22-24]. We were therefore interested to see whether any transfrags detected by tiling microarrays contained endogenous shRNAs. After comparing our predictions located on different introns of protein-coding genes with the results obtained by tiling arrays, we found 27 of the predicted shRNAs to overlap with 74 detected sites of transcription in the tiling-microarray datasets (Supplementary Tab. 6). Among them, the whole sequences of twelve shRNA precursors were identical to or held within one or more transcripts detected by the tiles, indicating that we had correctly predicted the actual length of these shRNA precursors. For other seventeen candidates, tiling-detected sequences were shorter than their corresponding shRNA precursors or just partially matched with either arm of shRNA precursors. Moreover, the sequences of shR-55 and shR265 were also cloned in EST database. Obviously, these data suggest that a possible product-precursor links overlapping transfrags described in tiling array datasets to our predicted shRNAs.

Microarray-based confirmation of computationally predicted sRNAs [32-33]

To validate hundreds of computationally predicted endogenous shRNAs and also to determine their comparative expression profiles in different human cells, we performed a microarray analysis using custom arrays that contain 4000 oligonucleotides directly synthesized on the chips. These capture probes cover

740 predicted shRNAs selected out from a total of 1258 shRNA candidates (Supplementary Tab. 7). These predicted shRNAs were those sequences randomly selected from the four groups shown in table 2. Because the mature form of the shRNA/miRNA can arise from either the 5' or 3' arm or the both arms of the hairpin stem, we made antisense probes to both sides of each of the candidate hairpins (Supplementary Tab. 7) and asked whether either of these probes could detect a predicted mature shRNA in the total RNA from different normal cells and cancer cells.

As shown in Figure 5a and Supplementary Table 8, we verified 380 endogenous shRNAs, of which 108 shRNAs reside on the 5' arms of hairpins, *125* shRNAs lie on the 3' arms of their hairpins, and 147 shRNAs stem from the both arms of their fold-back structures by using shRNA microarray. This implies that the biogenetic features of these mature shRNAs may be the stable product of Dicer processing, and functional shRNAs can reside on either arm of the precursor. Comparison of the signal intensity of shRNAs and let-7 members revealed that the abundance of most shRNAs in different cells were one tenth to one hundredth of let-7 except for a few shRNAs in the K562 cells (Fig. 5b). For instance, the observation indicated that the K562 cells expressed the highest abundance of shRNAs including 116a, 57a, 393a, 179a, 38s, 714a, 67a, and 72a (Supplementary Table 9), Thus, the low abundance of shRNAs may be one cause why they could not be detected by Northern blots and cloning method in previous studies.

Subsequently, we performed a comparative expression profiling using shRNAs from human brain, liver, lung, breast, and blood normal and cancer cells. Microarray experiments in human normal cells and cancer cells resulted in 380 candidate shRNAs with significant signals ($P < 0.01$) of at least one of their two predicted mature shRNAs. Most shRNAs we detected were tissue or cell-specific, and most were expressed in cancer cells to a greater extent than in the corresponding normal cells. For example, shR-56a, shR-2a, shR-72a, shR-53s, shR-72s, shR-116s, shR-393s and shR-568a, showed increased levels in cancer cells compared to normal cells (Fig. 5b). We reasoned that data on the cell-specific expression of shRNAs might facilitate the further investigation of their functional roles and clinical implication as well as the prediction of their potential mRNA targets using sequence-based methods. Another set of highly cell-specific shRNAs was found in cancer stem cells (MCF-7S), consisting of the shR-562a, shR-674s, shR-487a, and shR-131a. Overall, the majority of shRNAs in cancer cells expressed at higher levels than the corresponding ones in cancer stem cells. This represents another case of cell-specific expression of a shRNA and indicates that shRNAs may play a regulatory role not only in cell specification but also in developmental timing (Fig. 5c). Moreover, we also examined the difference in the distribution of shRNAs between cellular cytoplasm and nucleus. The results showed that the both cytoplasm and nucleus of A549 cell could contain many shRNAs, but there were more shRNAs in the former than the latter.

Figure 5d showed a cluster-gram of the 380 shRNAs (about 50% of the 743 distinct shRNAs represented on the array) that we detected in at least one type of cells at a signal intensity ≥ 3 × (background standard deviation) (see Materials and Methods). Thus, it may be reasonable to say that many candidate shRNAs are indeed real but expressed at levels below the threshold for detection by Northern blots and cloning technology.

Validation of the expression of shRNAs by real-time RT-PCR

Real-time quantitative RT-PCR (qRT-PCR)has been successfully used to detect the expression of mature miRNAs and miRNA precursors [20, 34]. Ro S et al. [35] have proposed a poly(A)-tailed RT-PCR to detect the expression of mature miRNAs. These studies indicated that the expression of mature miRNAs was comparable to that determined by Northern blotting. Because the shRNA candidates were hard to be detected by the Northern blotting strategy (data not shown), it may either be expressed at even lower levels or may be expressed under specific environmental conditions such that they are not represented in general situations (Fig. 5). In order to overcome the low sensitivity of Northern blotting analysis and to validate the results obtained by microarrays, we employed a qRT-PCR assay to quantify the expression of the shRNAs. The 39 candidates that showed different signal intensities on the array were chosen for qRT-PCR analysis. The short sequences of 38 shRNAs selected were successfully converted to cDNA and amplified using the qRT-PCR methods.

The C_T generated from some shRNAs was lower than the C_T for other shRNAs (Table 2). Differences in one C_T unit in real-time PCR data are typical when a 2-fold difference in template is detected. Notably, a close correlation between qPCR and microarray data was found. For example, the expression of shR-116, shR-291, shR-360, shR-539, shR-57 and shR-714 in A549 lung cancer cells was determined to range from approximately 10 to 19 on a C_T scale by qPCR, as compared with about 1000 to 6000 on a signal intensity scale by microarray analysis. In case of shR-137, shR-32, shR-399 and shR-726, the signals were detected by both microarrays and qPCR. In contrast, shR-379, shR-396, and shR-55 could be detected by qPCR but not by microarray whereas some shRNAs such as shR-140, shR-399, shR-32, shR-642, and shR-726 were not found with either qPCR or microarray. Moreover, qRT-PCR, like microarrays, further verified the fact that the mature shRNAs were generated from either or both sides of a stem (Table 2). These data suggest that the real time RT-PCR employed in this study have a high specificity in amplifying endogenous shRNA sequences.

To further validate the specificity of this technology, we designed a set of controls and longer primers specific for each shRNA. As positive controls, we designed specific primers for a subset of previously reported noncoding RNAs including *mi486*, *and U6* snRNAs. As negative controls, we used a no-template control, no universal primer control and no shRNA-specific primer control. Furthemore, we designed three pairs of primers complementary to 3 abundant mRNA transcripts such as beta-actin, proliferating cell nuclear antigen (PCNA), glyceraldehyde-3-phosphate dehydrogenase (GAPDH), in human cells. For these controls, qRT-PCR assay did not detect any specific amplification. To assess the probability that the primers for the candidate shRNAs generated PCR products by adventitiously priming against rRNAs, tRNAs, and previously cloned miRNAs, we performed computational comparisons of all shRNAs and PCR primers against related databases. This comparison showed that positive results due to adventitious priming were very unlikely by showing that short sequence matches of primers against other small RNA sequences expected in the library are not enough to generate a PCR product. It suggests that it is impossible for those primers specific for shRNAs to amplify any mRNA degradation products or other known small RNAs. Thus, we propose to consider all array-positive hairpins as candidate shRNAs until independent additional confirmation.

The targets and biological functions of shRNAs

The importance of shRNAs in regulating the expression of human gene is dependent on whether they bind to their target genes and form perfect sequence complementarity as exogenous shRNAs do, or imperfect match as miRNAs do. However, computational prediction of precisely complementary RNA targets for short antisense RNAs is essentially straightforward. Using this approach it is possible to verify whether predicted shRNA candidates in human exhibit perfect base pairing with the targets, without relying on homology to other organisms. We have conducted these predictions and further functional experiments. Unexpectedly, some shRNAs have been found to have their homologous sequences residing within different target regions including promoter, 5' and 3' UTR and coding area. Details will be described in another paper of ours (manuscript in preparation). In further support of this premise, data are presented in the accompanying manuscript by Bian et al[36]. that not only can one shRNAs (shR-337) be cloned and sequenced but that it was found to have important roles in regulating the definitive differentiation of human bone mesenchymal stem cells (BMSCs) into hematopoietic stem cells (HSCs).

Discussion

Different types of small RNAs appear to have their own features in the primary and secondary structures such as tRNAs, miRNAs, snoRNAs and others. Based on this general idea and detailed observation on different sRNAs, we developed a new bioinformatics approach for *ab initio* prediction of endogenous shRNA precursors in the human introns. This method has led to the identification of 2564 shRNAs, of which 1258 shRNAs are distinct. Meanwhile, 9 new mouse and rat shRNAs were found by orthology with predicated human shRNA precursors, suggesting that this type of shRNAs may exist in other genomic regions and different species.

Biological relevance of these shRNAs is supported by many different methods and syntenic data. By using public databases we were able to acquire evidence for the expression of a small number of the predictions. Further investigation with custom microarrays indicated that the expression of 380 shRNAs could be detected, of which 39 shRNAs were validated by qRT-PCR. More importantly, of these predicted shRNAs, shR-377 has been cloned and sequenced. Experimental results demonstrated that it plays a crucial role in regulating the proliferation and definitive differentiation of BMSCs. Thus, the different methods complement and support each other. All these data provide strong evidences that many of the candidates may be real but hard to be detected with conventional methods.

According to current models, intronic miRNA precursors that have the same orientation as their host gene might be processed upon cleavage of the intron by the Drosha and Dicer endonuclease or other ways [37-39]. shRNAs reported here, like microRNA precursors and exogenous shRNAs, may be generated through a similar or different mechanism and then incorporated into a sRNA-induced silencing complex (sRISC). Similarly, the Argonaute and piwi family has been found to be major interactors with sRNAs[40, 41]. Although Argonaute proteins have been implicated in many processes of gene expression in the transcriptional and translational levels [11, 42], it is unclear which of these proteins involves the biogenesis and cellular functions of shRNAs. In light of many homologous sequences of shRNAs within different regions of target genes, these shRNAs may modulate the expression of genes in a comprehensive mode.

Considered together, some shRNAs may play a role in tuning up the expression of genes at transcriptional and translational levels, others encoded by repetitive sequences may function like rasiRNAs and piRNAs to repress transposable elements or regulate a significant number of genes in the global level. Further studies on the functions of these shRNAs will help us understand the complex regulatory machinery in modulating the expression of protein-coding and nonprotein-coding genes.

Acknowledgments

This work was supported by Chinese Academy of Sciences “Hundred Talents Program” and the 863 program (Grant No: 2006AA02Z131 and 2007AA02Z181).

Correspondence and requests for materials should be addressed to J.Q.Y. (jqwyin@sun5.ibp.ac.cn).

Methods summary

Full Methods and any associated references are available in the online version of the paper.

References

1. Burge, C.B., Tuschl, T. & Sharp, P.A. Splicing of precursors to mRNAs by spliceosomes. In: The RNA World. Ed R. F. Gesteland, T. R. Cech & J. F. Atkins. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY (1999).
2. Roy, S.W. & Penny, D. Large-scale intron conservation and order-of-magnitude variation in intron loss/gain rates in apicomplexan evolution. Genome Res. 16, 1270-1275 (2006).
3. Rodríguez-Trelles, F., Tarrío, R. & Ayala, F.J. Origins and evolution of

spliceosomal introns. Annu. Rev. Genet. 40, 47-76 (2006).

4. Rodriguez, A., Griffiths-Jones, S., Ashurst, J.L. & Bradley, A. Identification of mammalian microRNA host genes and transcription units. Genome Res. 14,1902-1910 (2004).

5. Kim, Y.K. & Kim, V.N. Processing of intronic microRNAs. EMBO J. 26, 775-783 (2007).

6. Lee, Y., Kim, M., Han, J., Yeom, K.H., Lee, S., et al. MicroRNA genes are transcribed by RNA polymerase II. EMBO J. 23, 4051–4060 (2004)

7. Baskerville, S. & Bartel, D.P. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA. 11, 241-7 (2005).

8. Borchert, G.M. & Lanier, W. Davidson BL.RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol. 13, 1097-101 (2006).

9. Paddison, P.J., Caudy, A.A., Bernstein, E., Hannon, G.J. & Conklin, D.S. Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells. Genes Dev. 16, 948–958 (2002).

10. Snøve, O.Jr. & Rossi, J.J. Expressing short hairpin RNAs in vivo. Nat Methods. 3, 689-695 (2006).

11. Chapman, E.J. & Carrington, J.C. Specialization and evolution of endogenous small RNA pathways. Nat Rev Genet. 8, 884-896 (2007).

12. Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., et al. The nuclear RNase III Drosha initiates microRNA processing. Nature. 425, 415-419 (2003).

13. Hutvágner, G., McLachlan, J., Pasquinelli, A.E., Bálint, E., et al. A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. Science. 293, 834-8 (2001).

14. Mourelatos, J., Dostie, S., Paushkin, A., Sharma, B., et al. miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs, *Genes Dev.* 16**,** 720–728 (2002).

15. Doench, J.G., Petersen, C.P. & Sharp, P.A. siRNAs can function as miRNAs. *Genes Dev.* 17, 438–442 (2003).

16. Bernards, R., Brummelkamp, T.R. & Beijersbergen, R.L. shRNA libraries and their use in cancer genetics. Nat Methods. 3, 701-6 (2006).

17. Lai, E.C., Tomancak, P., Williams, R.W. & Rubin, G.M. Computational identification of Drosophila microRNA genes. Genome Biol.4, R42 (2003).

18. Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., & Bartel, D.P. Vertebrate microRNA genes. Science *299*, 1540 (2003).

19. Lagos-Quintana, M., Rauhut, R., Lendeckel, W., &Tuschl, T. Identification of novel genes coding for small expressed RNAs. Science *294*, 853–858 (2001).

20. Schmittgen, T.D., Lee, E.J., Jiang, J., Sarkar, A. et al. Real-time PCR quantification of precursor and mature microRNA. Methods. 44, 31-8 (2008).

21. Lu, S.S., Tej, S., Luo, C.D., Haudenschild, B.C. et al. Elucidation of the small RNA component of the transcriptome, *Science* 309, 1567–1569 (2005).

22. Ruby, J.G., Jan, C., Player, C., Axtell, M.J. et al. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. Cell. 127, 1193-207 (2006).

23. Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C. *et al.* The transcriptional landscape of the mammalian genome. *Science* 309,1559-1563 (2005).
24. Kapranov, P., Drenkow, J., Cheng, J., Long, J. et al. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res* 15, 987-997 (2005).
25. Bentwich, I., Avniel, A., Karov, Y., Aharonov, R. et al. Identification of hundreds of conserved and nonconserved human microRNAs. Nat Genet. 37, 766-70 (2005).
26. Piriyapongsa, J. & Jordan, I.K. A family of human microRNA genes from miniature inverted-repeat transposable elements. PLoS ONE. 2, e203 (2007).
27. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31, 3406-15 (2003).
28. Zhang, B.H., Pan, X.P., Cox, S.B., Cobb, G.P. & Anderson TA. Evidence that miRNAs are different from other RNAs. Cell Mol Life Sci. 63, 246-54 (2006).
29. Mattick, J.S. & Makunin, I.V. Non-coding RNA. Hum Mol Genet. 15 Spec No 1, R17-29 (2006).
30. Matukumalli, L.K., Grefenstette, J.J., Sonstegard, T.S. & Van Tassell, C.P. EST-PAGE – managing and analyzing EST data. Bioinformatics, 20, 286–288 (2004).
31. Betel, D., Sheridan, R., Marks, D.S. & Sander, C. Computational analysis of mouse piRNA sequence and biogenesis. PLoS Comput Biol. 3, e222 (2007).
32. Beuvink, I., Kolb, F.A., Budach, W., Garnier, A., Lange, J. et al. A novel microarray approach reveals new tissue-specific signatures of known and predicted mammalian microRNAs. Nucleic Acids Res. 35(7), e52 (2007).
33. Yin, J.Q., Zhao, R.C. & Morris, K. Exploring of miRNA expression using microarrays. Trends Biotech. 26 (2), Epub 2008 Feb.
34. Tang, F., Hajkova, P., Barton, S.C., O'Carroll, D. et al. 220-plex microRNA expression profile of a single cell. Nat Protoc. 1, 1154-1159 (2006).
35. Ro, S., Park, C., Jin, J., Sanders, K.M. & Yan, W.A. PCR-based method for detection and quantification of small RNAs. Biochem Biophys Res Commun. 351, 756-763 (2006).
36. Bian, C.J. et al. Conversion of BMSCs into HSCs by an endogenous shRNA. (Manuscript prepared for publication) (2008).
37. Ruby, J.G., Jan, C.H. & Bartel, D.P. Intronic microRNA precursors that bypass Drosha processing. Nature. 448, 83-86 (2007).
38. Kim, Y.K. & Kim, V.N. Processing of intronic microRNAs. EMBO J. 26, 775-83 (2007).
39. Lin, S.L., Kim, H. & Ying, S.Y. Intron-mediated RNA interference and microRNA (miRNA). Front Biosci. 13, 2216-30 (2008).
40. Seto, A.G., Kingston, R.E. & Lau, N.C. The coming of age for Piwi proteins. Mol Cell. 26, 603-609 (2007).
41. Aravin, A.A., Sachidanandam, R., Girard, A., Fejes-Toth, K. & Hannon, G.J. Developmentally regulated piRNA clusters implicate MILI in transposon control. Science. 316, 744-7 (2007).

42. Rossi, J.J. Transcriptional activation by small RNA duplexes. Nat Chem Biol. 3, 136-7 (2007).

Figure legends:

Figure 1. Schematic representation of the work-flow used to predict and verify structural short hairpin RNAs. Details are discussed in the main text.

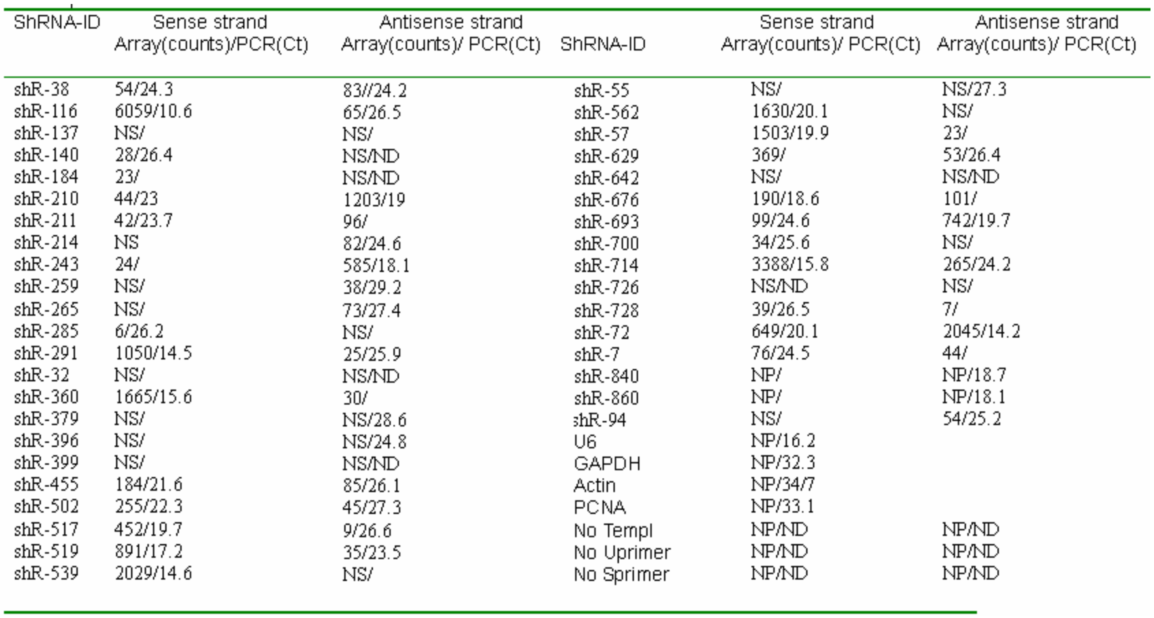
Figure 2. Genomic distribution and secondary structures of endogenous shRNAs in human. (A) A total of 2564 shRNAs predicted in this study can be mapped onto all chromosomes. (B) The bar chart depicts the comparison of endogenous shRNAs indicated as the red bar on the plus strand and as the blue bar on the minus strand. (C) shRNAs originate from long mRNA precursors transcribed from different genomic regions. Their distributions are different from those of miRNAs and snoRNAs. In this genomic view of an shRNA host gene located within chr7: 50000000–100000000, the fourth intron of this gene contains a sequence that can potentially form two shRNA fold-back structures, of which shR-248 is shown in (D). On the side of a stem, solid red lines stand for cloned shRNAs while dash red lines are referred to the shRNAs with significant signals detected by arrays. Blue boxes represent mismatched base-pairs whereas green circles show small bulges. ShR-38 and shR-243 are the same as miR-486 and miR-642, respectively. ShR-399 and shR-210 are conserved. shR-265 and shR-55 are stored in EST database. ShR-371 and ShR-337 have been cloned by us. Both arrays and qRT-PCR could detect shR116 but not shR399 in A549 cells.

Figure 3. Statlstlcal analysls of predlcted structural shRNAs . (A) Some host genes and introns appear to be capable (bottom right), with several predicted shRNAs, but most genes and introns contain only a few shRNAs (top left). (B) The fractions of shRNAs in each class (Table 1.1, 1.2, 1.3 and 1.4) found in intronic regions of protein-coding genes are shown as percentages of the total number of sRNAs in the table 1. The second pie chart illustrates the distribution of different types of repeat sequences classified in the repeat-derived shRNAs. Numbers in pie charts indicate percentage of total shRNAs for corresponding classes. (C) Each bar reflects the relationship between the number of human shRNAs or the AMFE and the given length of shRNA precursors. (D) Length distribution of shRNA stems shows that repeat-derived shRNAs have longer stems while nonrepeat-derived shRNAs contain shorter stems.

Figure 4. (a). A comparison of primary sequence conservation among human, rat and mouse shRNAs. The threshold for inclusion was 85% sequence identity in each arm. The only families shown are those that are conserved in at least two mammalian species. Identical nucleotides are marked in green while mutant nucleotides in red. Sequences were aligned with ClustalW and adjusted by hand. H, R and M are the abbreviation of human, rat and mouse, respectively. (b) The summery of evidence that support endogenous shRNAs predicted.

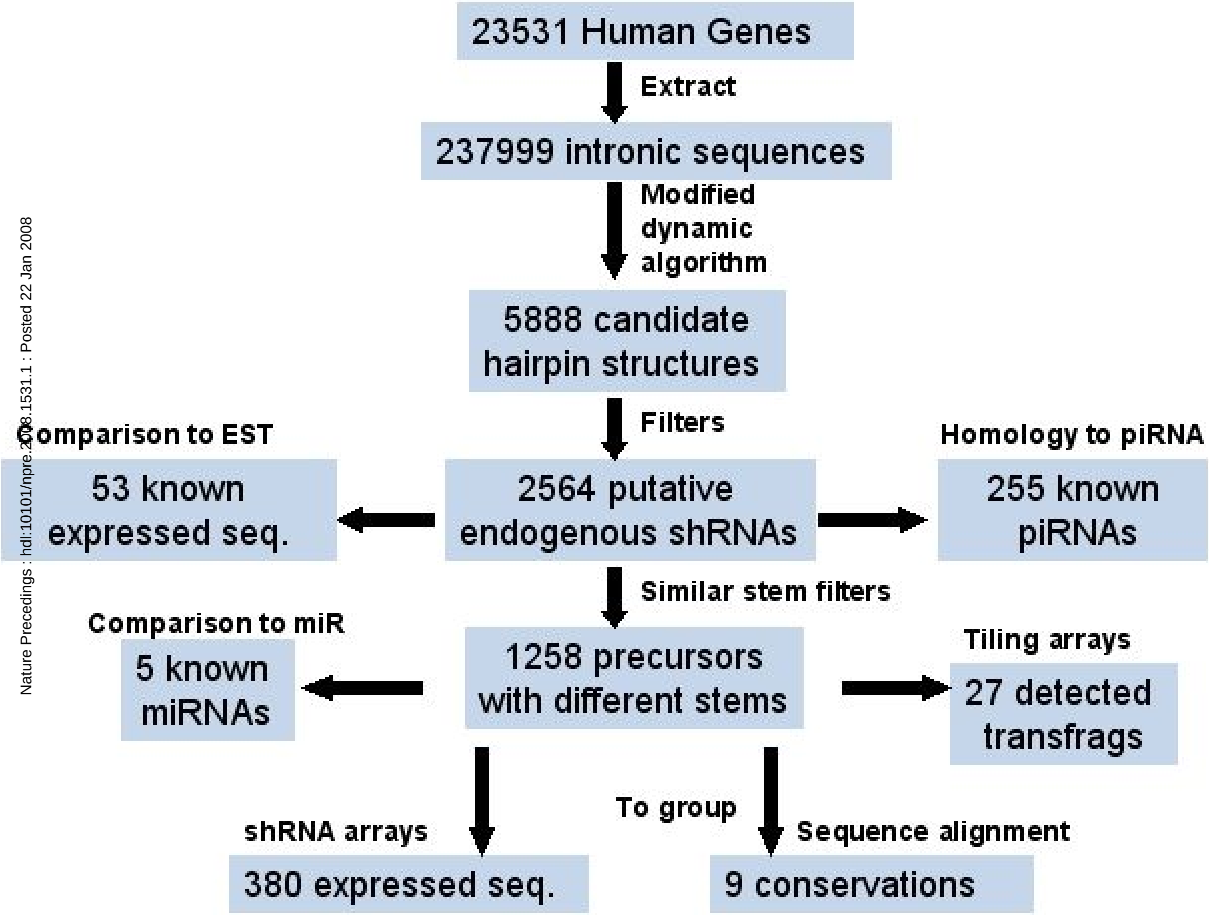
Figure 5. Expression profiles of 380 endogenous shRNAs across 12 different human cells reveal cell-specific expression of the majority of shRNAs detected. (a). The number of mature shRNAs. Blue cycle stands for mature shRNAs from sense strands, red cycle for mature shRNA from antisense strands, and overlap part for both strands of shRNAs significantly detected. (b). A heatmap of sRNA expression compared the difference in the abundance of 8 let-7 members and 10 shRNAs in 12 different cases. The relationship between the color and the expression levels is defined by the color key on the *upper* side of the figure. (c). The difference in the total signals of detectable shRNAs in 10 different cells. HB, human bronchial epithelial cells; HH, human hepatocytes; HG, human glial cells; H157, Non-Small Cell Lung Cancer; HepG2, Human hepatocellular liver carcinoma cells; U251, human glioblastma cells; MCF-7C, human breast cancer cells; MCF-7S, human breast cancer stem cells; K562, chronic myeloid leukemic cells; KG1, Human myeloblastic leukemia cells; A549-C, human lung adenocarcinoma cell cytoplasm; A549-N, human lung adenocarcinoma cell nucleus. (d). Profiling and relative abundance of 380 different shRNAs in 12 different samples. The clustering graph illustrated all the detected shRNAs with significant signal intensity. The level of expression of each shRNA in each of the samples is indicated by the color shown in (b). moderate

Table 2. Comparison of the expression of 39 representative shRNAs in A549 cells recorded by custom arrays and qRT-PCR.



ShRNA-ID	Sense strand	Antisense strand		Sense strand	Antisense strand
	Array(counts)/PCR(Ct)	Array(counts)/ PCR(Ct)	ShRNA-ID	Array(counts)/ PCR(Ct)	Array(counts)/ PCR(Ct)
shR-38	54/24.3	83//24.2	shR-55	NS/	NS/27.3
shR-116	6059/10.6	65/26.5	shR-562	1630/20.1	NS/
shR-137	NS/	NS/	shR-57	1503/19.9	23/
shR-140	28/26.4	NS/ND	shR-629	369/	53/26.4
shR-184	23/	NS/ND	shR-642	NS/	NS/ND
shR-210	44/23	1203/19	shR-676	190/18.6	101/
shR-211	42/23.7	96/	shR-693	99/24.6	742/19.7
shR-214	NS	82/24.6	shR-700	34/25.6	NS/
shR-243	24/	585/18.1	shR-714	3388/15.8	265/24.2
shR-259	NS/	38/29.2	shR-726	NS/ND	NS/
shR-265	NS/	73/27.4	shR-728	39/26.5	7/
shR-285	6/26.2	NS/	shR-72	649/20.1	2045/14.2
shR-291	1050/14.5	25/25.9	shR-7	76/24.5	44/
shR-32	NS/	NS/ND	shR-840	NP/	NP/18.7
shR-360	1665/15.6	30/	shR-860	NP/	NP/18.1
shR-379	NS/	NS/28.6	shR-94	NS/	54/25.2
shR-396	NS/	NS/24.8	U6	NP/16.2	
shR-399	NS/	NS/ND	GAPDH	NP/32.3	
shR-455	184/21.6	85/26.1	Actin	NP/34/7	
shR-502	255/22.3	45/27.3	PCNA	NP/33.1	
shR-517	452/19.7	9/26.6	No Templ	NP/ND	NP/ND
shR-519	891/17.2	35/23.5	No Uprimer	NP/ND	NP/ND
shR-539	2029/14.6	NS/	No Sprimer	NP/ND	NP/ND

Signals considered as absent in arrays are described in NS. NP stands for not available probe while ND is referred to no fluorescent intensity. No Templ is for no cDNA template, No Uprimer for no universal primer; No Sprimer for no specific primer.



23531 Human Genes

Extract

237999 intronic sequences

Modified
dynamic
algorithm

5888 candidate
hairpin structures

Filters

2564 putative
endogenous shRNAs

Comparison to EST

53 known
expressed seq.

Homology to piRNA

255 known
piRNAs

Similar stem filters

1258 precursors
with different stems

Comparison to miR

5 known
miRNAs

Tiling arrays

27 detected
transfrags

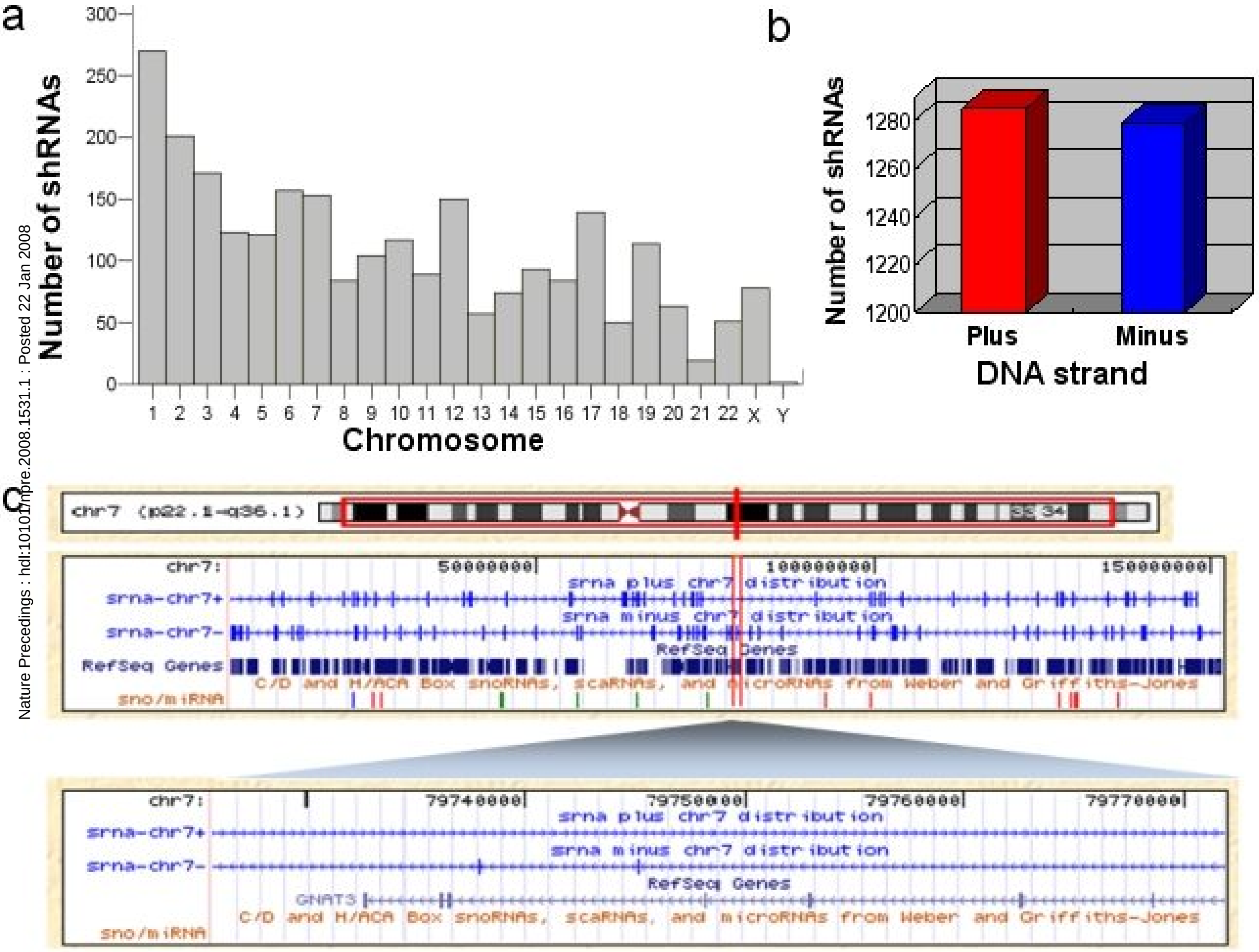
shRNA arrays

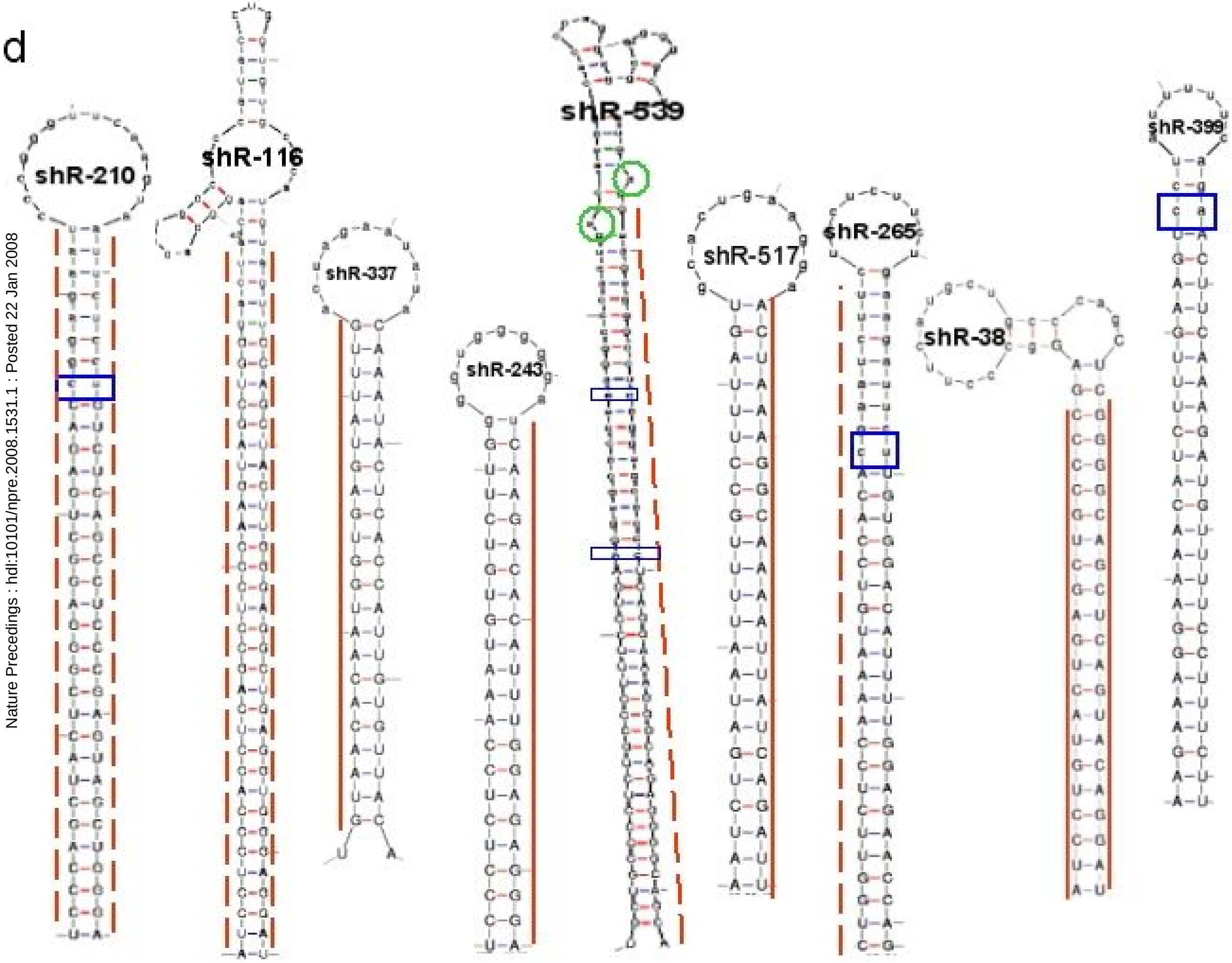
380 expressed seq.

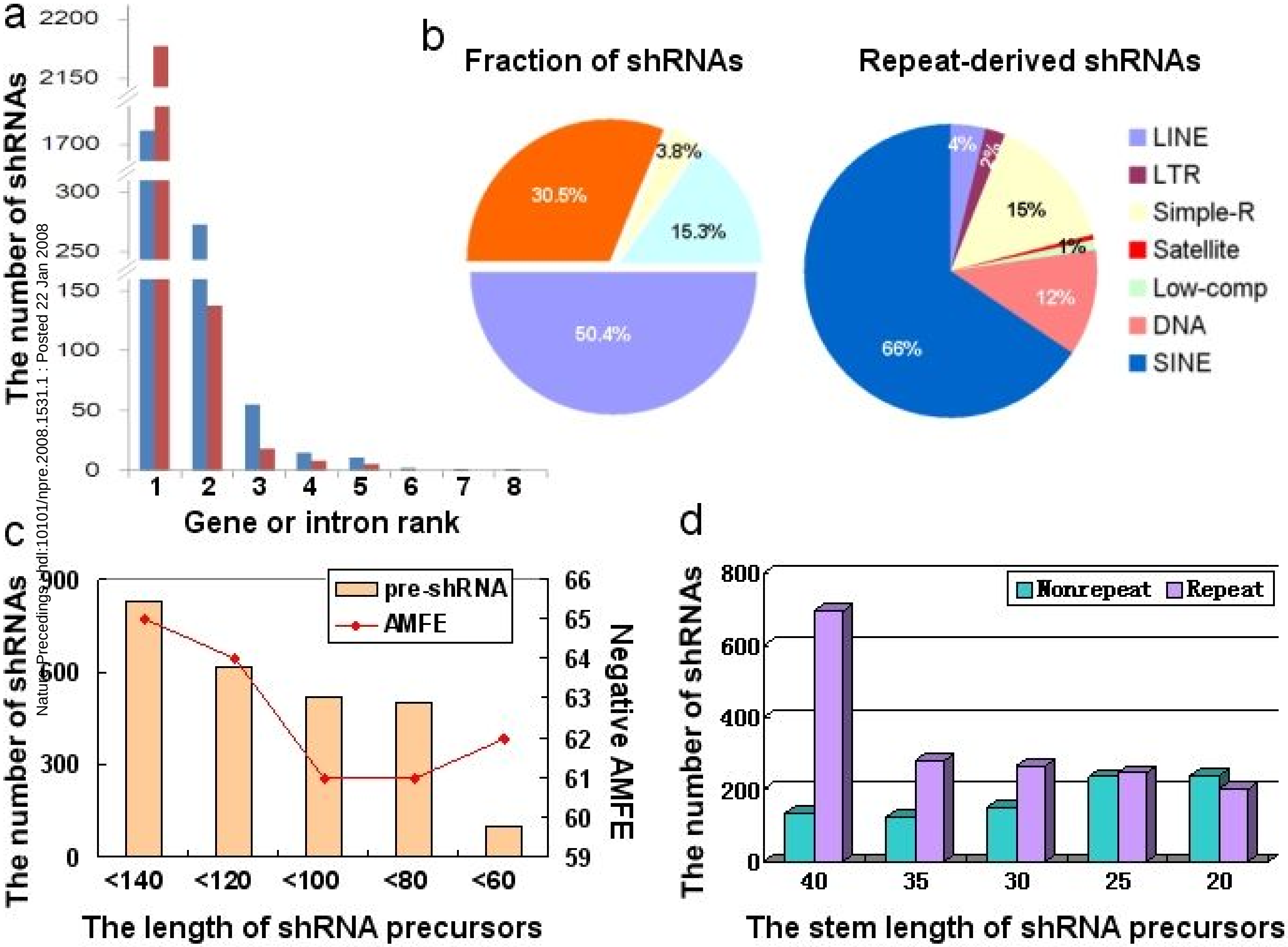
To group

Sequence alignment

9 conservations


a
300
250
200
150
100
50
0
Number of shRNAs
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
Chromosome
b
1280
1260
1240
1220
1200
Number of shRNAs
Plus
Minus
DNA strand
c
chr7 (p22.1-q36.1)
33 34
chr7:
50000000
100000000
150000000
srna plus chr7 distribution
srna-chr7+
srna minus chr7 distribution
srna-chr7-
RefSeq Genes
RefSeq Genes
C/D and H/ACA Box snoRNAs, scaRNAs, and microRNAs from Weber and Griffiths-Jones
sno/miRNA
chr7:
79740000
79750000
79760000
79770000
srna plus chr7 distribution
srna-chr7+
srna minus chr7 distribution
srna-chr7-
RefSeq Genes
GNAT3
C/D and H/ACA Box snoRNAs, scaRNAs, and microRNAs from Weber and Griffiths-Jones
sno/miRNA


d
shR-210
shR-116
shR-337
shR-243
shR-539
shR-517
shR-265
shR-38
shR-399


a
2200
2150
1700
300
250
150
100
50
0
The number of shRNAs
1
2
3
4
5
6
7
8
Gene or intron rank
b
Fraction of shRNAs
30.5%
3.8%
15.3%
50.4%
Repeat-derived shRNAs
4%
2%
15%
1%
12%
66%
LINE
LTR
Simple-R
Satellite
Low-comp
DNA
SINE
c
pre-shRNA
AMFE
900
600
300
0
The number of shRNAs
66
65
64
63
62
61
60
59
Negative AMFE
<140
<120
<100
<80
<60
The length of shRNA precursors
d
Nonrepeat
Repeat
800
600
400
200
0
The number of shRNAs
40
35
30
25
20
The stem length of shRNA precursors

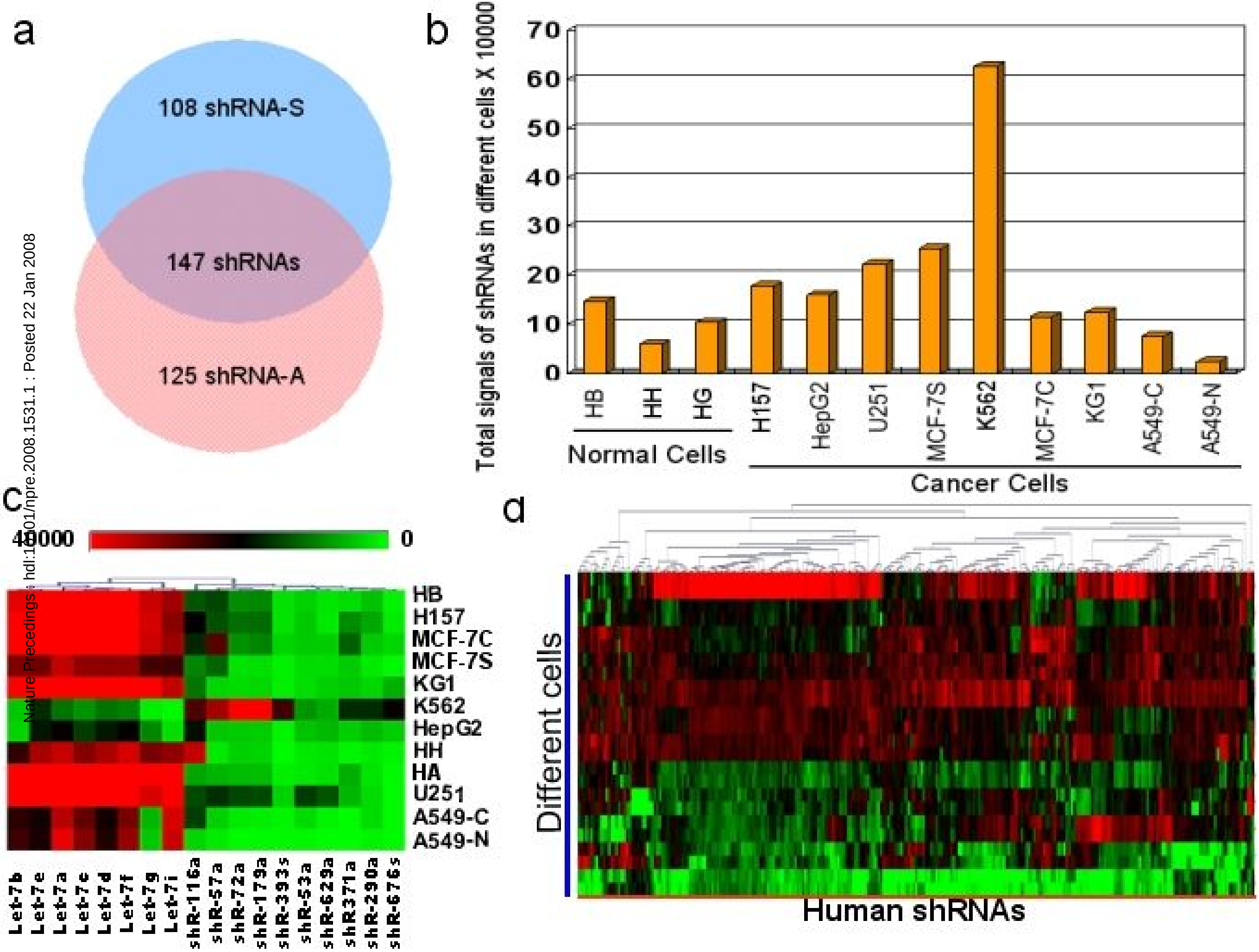
a

```
CUCGGCCUCCCAAAGUGCUGGGAUU------UCCCAGCUACUCGGGAGGCUGAG-H.shR-860
CUCUGCCUCCC-AAGUGCUGGG-----------CCCAGC-ACUUGGGAGGCAGAG-R
CUCUGCCUCCC-AAGUGCUGGGAUU------UCCCAGC-ACUCGGGAGGCAGAG-M
 CCAGCCUGGGCGACAGAGCGAGAC------GUCUCGCUCUGUCGCCCAGGCUGG-H.shR-850
 CCAGCCUGGUCUACAGAGUGAGUU------GUCUCACUCUGUAGCCCAGGCU---R
 CCAGCCUGGGCUACAGAGUGAGUU------AACUCACUCUGUAGACCAGGCUGG-M
 GCCUCGGCCUCCCAAAGUGCUGGA--------CCAGUACUUUGGGAGGCUGAGGC-H.shR-830
 GCCUCUGCCUCCC-AAGUGCUGGGG------CCAGCACUU-GGGAGGCAGAGGC-R
 GCCUCUGCCUCCC-AAGUGCUGGG--------CCAGCACUU-GGGAGGCAGAGGC-M
 AAGAAAGGAAAACAUCUUUGAAGU------ACUUCAAAGAUGUUUUCCUUUCUU-H.shR-399
 AAGAAAGGAAAACAUCUUUGAAGU------ACUUCAAAGAUGUUUUCCUUUCUU-R
AGCCUGGGCGACAGAGCGAGAC-----------GUCUCGCUCUGUCGCCCAGGCU-H.shR-840
AGCCUGGGCUACAGAGUGAGAC-----------GUCUCACUCUGUAGCCCAGGCU-R
AUACAUACAUACAUACACACA-------------UGUGUGUAUGUAUGUAUGUAU-H.shR-800
AUACACACACACACACACACA-------------UGUGUGUGUGUGUGUGUAUAU-R
 UCCCAGCUACUCGGGAGGCUGAG-------- CUCAGCCUCCCGAGUAGCUGGGA-H.shR-210
 UCCCAGC-ACUUGGGAGGCAGAG-------- CUCUGCCUCCCAAGU-GCUGGGA-M
CUGGGAUUACAGGCGUGAGCCACC------GGUGGCUCACGCCUGUAAUCCCAG-H.shR-810
CUGGGAUUAAAGGCGUGUGCCACC------GGUGGCACACGCCUUUAAUCCCAG-M
GUGUGUAUAUAUGUGUGUGUAUA---------UAUACACACACAUAUAUACACAC-H.shR-820
GUGUGUGUGUGUGUGUGUGUAUA---------UAUACACACACACACACACACAC-M
```

b

Table 1. Summary of evidence that supports predicted endogenous shRNAs in human.

Category	Number of shRNA candidates	Reference
Total predicted distinct shRNAs	1256	Suppl. tab. 2
Previously known miRNAs	5	Suppl. tab. 5
Evolutional conservation	9	Suppl. Tab. 2.1, 2.2
Expressed Sequence Tags	53	Suppl. tab. 3
Known human piRNA homologs	255	Suppl. tab. 4
Tiling microarrays	27	Suppl. tab. 6
shRNA arrays	380	Suppl. tab. 8
qRT-PCR	39	Tab. 2
Cloning and sequencing	2	Fig. 2d
Functional confirmation	1	Ref. 36


a
108 shRNA-S
147 shRNAs
125 shRNA-A
b
Total signals of shRNAs in different cells X 10000
70
60
50
40
30
20
10
0
HB
HH
HG
H157
HepG2
U251
MCF-7S
K562
MCF-7C
KG1
A549-C
A549-N
Normal Cells
Cancer Cells
c
40000
0
HB
H157
MCF-7C
MCF-7S
KG1
K562
HepG2
HH
HA
U251
A549-C
A549-N
Let-7b
Let-7e
Let-7a
Let-7c
Let-7d
Let-7f
Let-7g
Let-7i
shR-116a
shR-57a
shR-72a
shR-179a
shR-393s
shR-53a
shR-629a
shR371a
shR-290a
shR-676s
d
Different cells
Human shRNAs