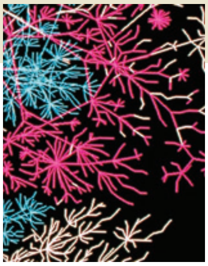

© 2005 Nature Publishing Group http://www.nature.com/naturebiotechnology

# Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome

Stefan Washietl[1], Ivo L Hofacker[1], Melanie Lukasser[2], Alexander Hüttenhofer[2] & Peter F Stadler[3,4].

**In contrast to the fairly reliable and complete annotation of the protein coding genes in the human genome, comparable information is lacking for noncoding RNAs (ncRNAs). We present a comparative screen of vertebrate genomes for structural noncoding RNAs, which evaluates conserved genomic DNA sequences for signatures of structural conservation of base-pairing patterns and exceptional thermodynamic stability. We predict more than 30,000 structured RNA elements in the human genome, almost 1,000 of which are conserved across all vertebrates. Roughly a third are found in introns of known genes, a sixth are potential regulatory elements in untranslated regions of protein-coding mRNAs and about half are located far away from any known gene. Only a small fraction of these sequences has been described previously. A comparison with recent tiling array data shows that more than 40% of the predicted structured RNAs overlap with experimentally detected sites of transcription. The widespread conservation of secondary structure points to a large number of functional ncRNAs and *cis*-acting mRNA structures in the human genome.**

The recent completion of the human genome sequence emphasizes the "need for reliable experimental and computational methods for comprehensive identification of noncoding RNAs"[1]. A variety of experimental techniques have been used to uncover the human and mouse transcriptomes, in particular tiling arrays[2–5], cDNA sequencing[6,7] and unbiased mapping of transcription factor binding sites[8]. All these studies suggest that a substantial fraction of the genome is transcribed and that a large fraction of the transcriptome consists of noncoding RNAs. It is unclear, however, which fraction is functional noncoding RNAs (ncRNAs), and which constitutes "transcriptional noise"[9].

Genome-wide computational surveys of ncRNAs, on the other hand, have been impossible until recently, because ncRNAs do not share common signals that could be detected at the sequence level. A large class of ncRNAs, however, has characteristic structures that are functional and hence are well conserved over evolutionary timescales: most of the 'classical' ncRNAs, including rRNAs, tRNAs, small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), as well as the RNA components of RNAse P and the signal recognition particle, are of this type. The stabilizing selection acting on the secondary structure causes characteristic substitution patterns in the underlying sequences. Consistent and compensatory mutations replace one type of base pair by another one in the paired regions (helices) of the molecule. In addition, there are differences in the sequence variation between loop regions and helices. These patterns can be exploited in comparative computational approaches[10–13] to discriminate functional RNAs from other types of conserved sequences. Recently, high levels of sequence conservation of noncoding DNA regions have been reported[14–17]. Here we screen the complete collection of conserved noncoding DNA sequences from mammalian genomes and provide a first annotation of the complement of structurally conserved RNAs in the human genome.

## Selection of conserved sequences and screening for structural RNAs

We start from the genome-wide alignments of vertebrate genomes provided through the UCSC Genome Browser[18]. We limit our comparative screen to the most conserved regions as annotated by the PhastCons program, which constitute 4.81% of the 3,095 MB of the human genome. It has been estimated that about 5% of the human genome is under selective pressure[19,20] but this fraction may be even higher[15]. Since we are interested in noncoding RNAs, we removed all annotated coding exons from the set and retained only the 438,788 alignments of noncoding regions that are conserved at least in the four eutherian mammals (human, mouse, rat, dog). This amounts to 82.64 MB or 2.88% of the human genome (**Table 1**).

This data set was screened for structural RNAs using RNAz[12], a program that combines a comparative approach (scoring conservation of secondary structure) with the observation[21,22] that ncRNAs are thermodynamically more stable than expected by chance. A structure conservation index (SCI) is computed by comparing the predicted minimum free energies of the sequences in an alignment with a

[1]Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse 17, 1090 Vienna, Austria. [2]Division of Genomics and RNomics, Innsbruck Medical University-Biocenter, Fritz-Pregl-Strasse 3, 6020 Innsbruck, Austria. [3]Department of Computer Science and Interdisciplinary Center of Bioinformatics, University of Leipzig, Härtelstrasse 16-18, D-04107, Leipzig, Germany. [4]Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, New Mexico 87501, USA. Correspondence should be addressed to P.F.S. (studla@bioinf.uni-leipzig.de).

Published online 4 November 2005; doi:10.1038/nbt1144

© 2005 Nature Publishing Group http://www.nature.com/naturebiotechnology

**Table 1 Genomic coverage of filtering steps and phylogenetic conservation of predicted RNAs**

| | Genome coverage | | Alignments | RNAz hits $P > 0.9$ | | |
|---|---|---|---|---|---|---|
| | Size (MB) | Fraction (%) | Number | Size (MB) | Fraction of input (%) | Number |
| Human genome | 3,095.02 | 100.00 | – | | | |
| PhastCons most conserved | 137.85 | 4.81 | 1,601,903 | | | |
| Without coding regions | 110.04 | 3.84 | 1,291,385 | | | |
| Without alignments <50 nt | 103.83 | 3.33 | 564,455 | | | |
| Set 1: 4 Mammals | 82.64 | 2.88 | 438,788 | 5.46 | 6.62 | 35,985 |
| Set 2: + Chicken | 24.00 | 0.85 | 104,266 | 1.34 | 5.50 | 8,802 |
| Set 3: + Fugu or zebrafish | 6.86 | 0.24 | 30,896 | 0.14 | 2.03 | 996 |

consensus energy, which is computed by incorporating covariation terms into a free-energy minimization computation[23]. Thermodynamic stability is quantified by means of a *z*-score that measures the folding energy relative to shuffled sequences (a regression approach replaces time-consuming shuffling methods). A support vector machine then classifies an alignment as 'structured RNA' or 'other' based on *z*-score and SCI. The significance of the classification is quantified as "RNA-class probability" *P*.

**Figure 1** illustrates the strategy of our screen and shows an annotation for a 9-megabase region on chromosome 13. For details on the scanning procedure, see the Methods section. In the complete genome, we detected 91,676 (15.1% of the conserved sequence) independent RNA structures on the $P = 0.5$ level and 35,985 (6.6%) structures on the $P = 0.9$ level (**Table 1** and **Fig. 2a**,**b**).

## Estimating specificity

The specificity of RNAz is generally high, ≈99% for the $P = 0.9$ cutoff[12]. Because of the large number of input alignments, however, we have to expect a nonnegligible number of false positives. We therefore repeated the complete screen with alignments randomized by shuffling[22]. We obtained a false-positive rate of 28.9% ($P = 0.5$) and 19.2% ($P = 0.9$), respectively. As expected, the hits in the randomized data set are on average smaller than the native ones, reducing the false-positive rates to 25.7% ($P = 0.5$) and 16.3% ($P = 0.9$) in terms of sequence length. The estimate for the false-positive rate implies lower bounds of 65,000 ($P = 0.5$) and 29,000 ($P = 0.9$) for the number of structural RNA elements in the human genome. On average, we predict 21 ($P = 0.5$) and 10 ($P = 0.9$) structural elements per megabase.

Furthermore, we observed that many of the hits in randomized alignments overlap with native predictions (**Supplementary Table 1** online). This might indicate that our shuffling process does not effectively remove the signal in all cases. We also observed that the random hits are clearly enriched in highly conserved alignments. The false-positive rate of RNAz is higher in this case, because these alignments contain little covariance information so that the classification is dominated by the thermodynamic stability alone. Since many known ncRNAs are contained in this set, we decided against removing highly conserved alignments from our survey despite the increased false-positive rate.

## Detection performance on known ncRNAs

A comprehensive annotation of ncRNAs in the human genome is not available, thus it is impossible to determine the overall sensitivity of our screen. For microRNAs (miRNAs) and snoRNAs, however, a comprehensive annotation is provided in the UCSC browser.

There are 207 annotated miRNA loci (see also ref. 24) of which 45 loci are not in our set of input alignments for various reasons (see **Supplementary Table 2** online). We detect 157 (96.9%) of the remaining 162 miRNAs. The effective sensitivity is 75.8% for miRNA precursors, which are among the easiest-to-find ncRNAs (**Fig. 2c**).

Twenty-two of the 86 annotated H/ACA-box snoRNAs in the input set mostly because they are not detected by PhastCons (**Supplementary Table 3** online). We recovered 55 of the remaining 64 sequences (85.9%). We can thus relatively accurately detect this class of ncRNAs, which have resisted computational prediction so far. (Effective sensitivity: 64.0%.)

Our screen performs poorly on C/D-Box snoRNAs, however. Out of the 256 known C/D snoRNAs, about one-half (129) are missing in the input alignments. Even though we detect 39.4% of C/D snoRNAs in our set, the effective sensitivity is only 19.5%. C/D-Box snoRNAs are hard to detect computationally even with specialized approaches[25].

From these examples we estimate that the overall sensitivity of the combination of the Multiz/PhastCons alignments and RNAz is on the order of 30%.

We then compared all hits with available databases of known ncRNAs (**Table 2**). Most of the 'classical' structured ncRNAs, such as tRNAs and most snRNAs, were not contained in the input alignments because they are marked as repetitive DNA by RepeatMasker and were therefore excluded from the Multiz alignments. We did, however, detect all snRNAs of the minor spliceosome (U4atac, U6atac, U11 and U12), as well as very well conserved (although not very stable) structures within the RNAse P. We missed RNAse MRP and telomerase RNA, presumably because of the pseudoknotted structures[26,27], which are not taken into account by RNAz.

We found local secondary structure motifs in various other documented ncRNAs which do not appear to have conserved global structures (**Supplementary Table 4** online). The *XIST* gene, a 17-kb ncRNA which plays a key role in dosage compensation and X-chromosome inactivation[28], contains three independent conserved RNA secondary structures. Intriguingly, we found 8 RNAz hits in the human genome with significant sequence similarity to the *Air* RNA. This antisense transcript regulates imprinted gene expression in mouse[29] but is not conserved over its full length (≈1,000kb) in human. Seven of the eight hits corresponded to the same local secondary structure motif in *Air*. One of them can be found in an intron of *HERC2*, a locus located near the Prader-Willi imprinting center, which in turn is regulated by antisense transcripts.

The RNAdb[30] compiles collections of expressed sequences with reduced protein coding capacity. A comparison of our RNAz hits with the RNAdb identified conserved structured elements in many of these transcripts, thereby showing that they function as ncRNAs (**Table 2**).

## New members of known ncRNA families

A number of signals are novel ncRNAs that can be associated with known ncRNAs or ncRNA families through sequence similarity. Some

© 2005 Nature Publishing Group http://www.nature.com/naturebiotechnology

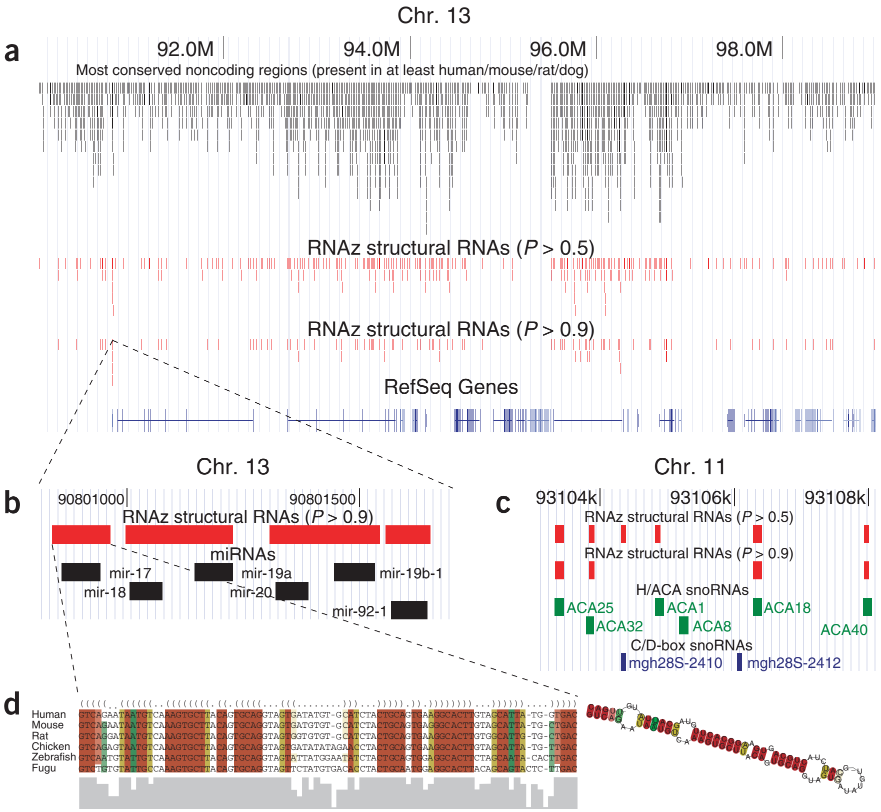

**Figure 1** Annotation procedure. (**a**) Starting from the 5% best conserved noncoding DNA as defined by Multiz alignments and PhastCons in the UCSC Genome Browser, RNAz uses a stringent filter for putative structured RNAs. These sequences are thermodynamically more stable than average and can fold into a common secondary structure. Two levels of confidence ($P > 0.5$) and ($P > 0.9$) are used. (**b**,**c**) The RNAz hits, which, depending on the confidence level, cover 6–15% of the input alignments, are highly enriched in known ncRNAs, such as the *mir17*-cluster of miRNAs (**b**) or a cluster of H/ACA and C/D box snoRNAs on chromosome 11 (**c**). (**d**) The method not only detects signals for structural RNAs but in the process of classification constructs an explicit secondary structure model from the aligned sequences (see also **Fig. 3**). Panels **a**–**c** show screen-shots of the UCSC genome-browser[18] on the human assembly hg17.

of these are additional paralogs or orthologs of known RNA genes. For example, we found more than 100 hits with sequence similarity to snoRNAs. Some of these are most likely functional snoRNAs because they are human homologs of mouse snoRNAs[31].

Another class of signals are novel members of one of the large, well-described classes of ncRNAs. A simple subscreen was performed to identify putative H/ACA box snoRNAs. We selected all RNAz hits with two stems at least 15 pairs in length and separated by an unpaired hinge, which in addition have the motif ACA in the consensus sequence in the last 20nt. We found 137 structures, of which 28 were known snoRNAs. Visual inspection shows that 30–40 additional clusters have typical H/ACA snoRNA-like secondary structure of which 15 also have the canonical H-box sequence ANANNA (**Fig. 3c**,**d**). In many known snoRNAs, only short parts of the stem are conserved in the predicted consensus structure and/or only parts of the complete structure are detected as conserved structural element. As a consequence, this subscreen is not exhaustive and a more detailed analysis can be expected to bring up even more candidates.

From the 15 candidates with H/ACA motifs, we randomly selected five candidates and tested their expression in HeLa cells using northern blot analysis. Although there is an increasing number of examples of highly regulated and tissue specific snoRNAs[32], many snoRNAs are ubiquitously expressed and should be detectable with this simple experimental setup. Indeed, we detected three of the five candidates by northern blot analysis (**Supplementary Fig. 1** online). These examples demonstrate that *de novo* prediction coupled with subsequent structure/sequence analysis is a promising approach for detection of novel ncRNAs.

Berezikov and coworkers[33] identified 975 miRNA candidates in mouse/human and mouse/rat comparisons by means of a combination of phylogenetic shadowing and selection of stable stem-loop structures. Our set of input alignments contained 642 of these candidates, of which 472 overlapped with our predictions ($P > 0.9$). Not all these stem-loops, which are stable as single sequences, were structurally conserved in all four mammals; some of them lacked a stable consensus structure. A simple filter requiring a stem with at least 20 base pairs in the consensus structure, a mean $z$-score <–3.5 and a 22-nt window with more than 0.95% pairwise sequence identity (the prospective mature miRNA sequence) retained 312 candidates, among them 109 known human miRNAs. Some of the unknown candidates showed the typical mutation pattern of miRNA (see **Figs. 1d** and **3a**,**b**). Others exhibited clear structural conservation but showed a very different mutation pattern. We speculate that these sequences are not miRNAs but belong to different, so far undescribed, classes of ncRNAs.

## Structures conserved across all vertebrates

The most highly conserved structures are of particular interest. We found 996 RNAz signals that were conserved in all four mammals, chicken and at least one of the two fish genomes (fugu, *Takifugu rubripes* and zebrafish, *Danio rerio*). Of these, 152 could be at least partially annotated: 52 were miRNAs, 16 were snoRNAs, 28 were known elements in untranslated regions (UTRs) and 56 were similar to other described RNAs. We found 42 detected regions were contained within one of the different cDNA collections and 38 overlapped with one of the 481 'ultraconserved elements' (segments longer than 200 base pairs that are identical between human, mouse and rat genomes) reported by Bejerano *et al.*[34]. A few of these could be identified as potential RNAs because of the substitution pattern in the fish and chicken sequences. For most of them, however, we cannot give a definitive classification because there is too little sequence variation in this special set of extremely conserved sequences.

## Comparison with protein-gene annotations and transcriptional maps

The majority of the 35,989 structured RNA features detected with $P > 0.9$ were of completely unknown function (see **Fig. 3e–h** for a few selected examples). We compared the location of the hits with the protein coding gene annotations provided by the UCSC genome browser. About half of the predicted structures were located far away from any

© 2005 Nature Publishing Group http://www.nature.com/naturebiotechnology

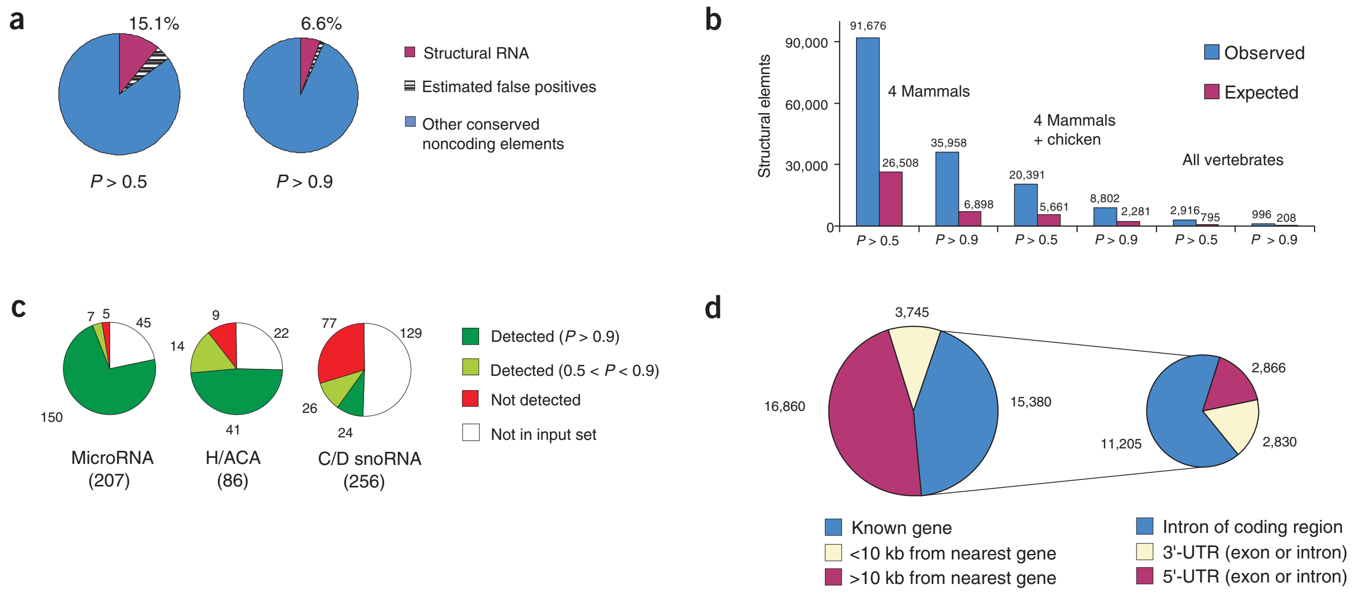

**Figure 2** Statistical analysis of predicted structural RNAs. (**a**) The RNAz method classifies, depending on the user-defined confidence cut-off, a small fraction of the conserved noncoding regions as structural RNA. (**b**) A second screen on shuffled data demonstrates that the effective false positive rate of the entire screen is well below 25% and decreases with increasing confidence level and phylogenetic conservation. (**c**) The sensitivity is estimated for known miRNAs and snoRNAs. Between one-quarter and two-thirds of the known ncRNAs are not contained in the input alignments due to insufficient accuracy in the alignments, incomplete sequences and removal of repeated DNA. This is the most severe limitation at present. (**d**) Comparison of the detected hits (shown here for the $P > 0.9$ level) with current protein-gene annotations.

known protein coding gene, the other half was associated with known genes. Two thirds of the latter were located in introns. One sixth can be mapped to annotated UTRs.

In a recent study, sites of transcription of polyadenylated and non-polyadenylated RNAs for ten human chromosomes were mapped at 5-bp resolution in eight cell lines using tiling array technology[5]. We compared our predictions located on the ten chromosomes with the cumulative 1-in-8 map, in which a positive probe needs to appear in at least one of eight cell lines. We found 40.7% of the predicted RNAs to overlap with detected sites of transcription (45.0% including signals in exons or introns of known UTRs). This is significantly (~10%) higher than the background and comparable to the detection rate of well-known ncRNAs: 45.2% of known microRNAs and 56.7% of known snoRNAs were detected on this transcriptional map.

A list compiling 50 examples of RNAz hits that are transcribed according to the tiling array experiment and that are strong candidates

**Table 2 Comparison of predicted RNAs with ncRNAs from the literature**

| Database | Ref. | $P > 0.5$ | $P > 0.9$ |
|---|---|---|---|
| Rfam | 24 | 267 | 189 |
| NONCODE | 41 | 273 | 177 |
| RNAdb | 30 | 446 | 327 |
| miRNA Registry | 40 | 176 | 168 |
| UTRdb | 42 | 388 | 159 |
| **Curated** | | 984 | 563 |
| hinv | 7 | 478 | 205 |
| Fantom | 6 | 1,908 | 781 |
| chr7 | 43 | 180 | 90 |
| Antisense pipeline | 30 | 149 | 59 |
| **cDNA collections** | | 2,539 | 1,056 |
| Total | | 3,441 | 1,585 |

We compared the human sequences detected by RNAz with all major RNA databases using Blast. Hits with $E < 10^{-6}$ are reported. Apart from curated databases we compared our data with four collections of cDNAs. Fantom contains more than 15,000 unique, putative ncRNAs from mouse that do not contain a significant coding sequence. H-invitational (hinv) contains more than 2,000 transcripts with ORFs <80 amino acids that passed several additional filters designed to exclude likely protein-coding genes. The human chromosome 7 annotation project (chr7) has described over 350 putative ncRNAs derived from computer-based annotation in conjunction with extensive laboratory experimentation. Note that many ncRNAs may function, for example, as antisense regulators without exhibiting a conserved, functionally important, secondary structure. Also, it is unclear which fraction of these sequences are functional RNAs and not transcriptional artifacts. Thus, one cannot use these databases to estimate the sensitivity of the RNAz screen. Nevertheless, we find that many of these transcribed putative ncRNAs exhibits evolutionarily conserved structures.

© 2005 Nature Publishing Group http://www.nature.com/naturebiotechnology

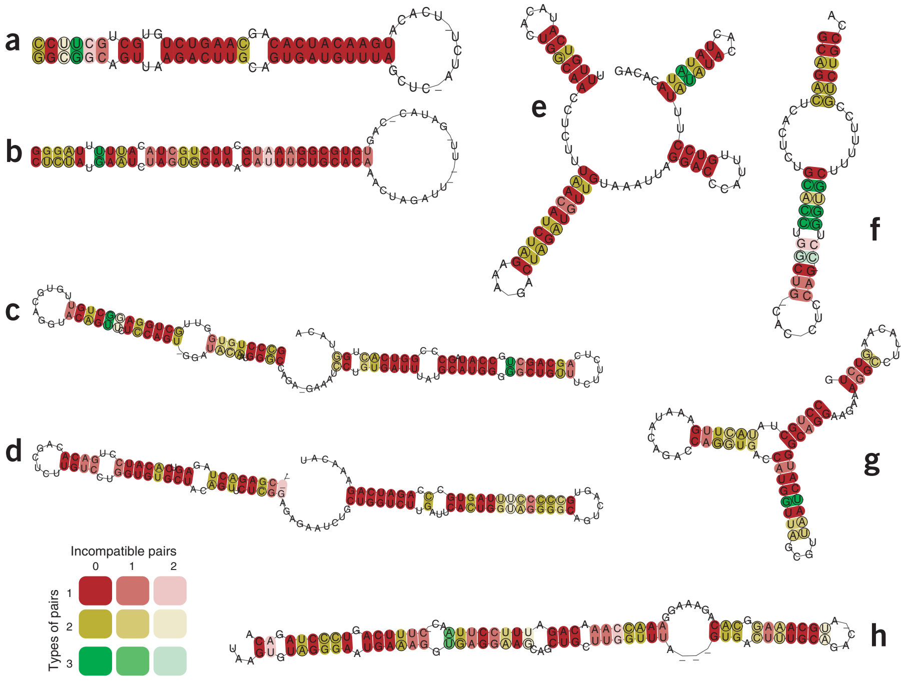

**Figure 3** Selected examples of candidates for novel structural RNAs detected with $P > 0.9$. Predicted consensus structures with annotation of consistent and compensatory mutations are shown. Circles indicate variable positions in stems, colors indicate the number of different types of base pairs that support stabilizing selection on the structure. (**a**,**b**) These structures conserved across all vertebrates can be unambiguously identified as microRNA precursors on the basis of several characteristic features: (i) a stable hairpin consensus structure; (ii) the sequence of one arm of the stem is highly conserved over 22 nt (the putative mature miRNA); (iii) the opposite stem is also conserved but not that strictly; (iv) the loop sequence is diverged due to the absence of functional constraints in this region; (v) compensatory, or at least consistent, mutations are found in the outer parts of the stem where only structure but not sequence is important for function. See also the alignment in **Figure 1d** illustrating these typical microRNA features. (**c**,**d**) Candidates for novel H/ACA snoRNAs identified by secondary structure and primary sequence motifs. Both candidates fold into the typical bipartite hairpin secondary structure. We observe H-box motifs ANANNA in the hinge regions and ACA motifs in the tail regions. (**e**–**h**) Novel structural RNA elements. The sequence of structure **e** has similarity to a transcript in the Chr. 7 set of RNAdb and is conserved in mammals. We found more than 50 conserved secondary structures throughout the genome with sequence similarity to this transcript. Within these hits, we could identify this structural motif seven times by visual inspection. The structures **f**–**h** are conserved across all vertebrates and have particularly strong RNAz signals. Structure **f** is located near an intron/exon boundary and EST data suggests alternative splicing events in this region. Structure **h** is also located in an intron of a coding gene. It is an extremely stable stem-loop structure, which is longer than a typical microRNA precursor and also shows a different mutational signature. Additional RNAz hits in the close vicinity suggest that this is a local substructure of a longer RNA. Genomic locations of all examples (based on hg17 assembly): (**a**) chr.20:33,041,857 (intron of a mysine protein gene, AB040945); (**b**) chr.15:43,512,536 (UTR region of FOAP-11, AF228422); (**c**) chr.9:92,134,300 (intron of Isoleucine-tRNA synthetase, D28473); (**d**) chr.16:2,786,411 (near a pseudogene of ribosomal protein 27A, flanked by LINE elements); (**e**) chr.12:74,595,654 (intergenic); (**f**) chr.22:18,488,478 (in intron of RAN binding protein 1, D38076); (**g**) chr.8:57,457,661 (intergenic); (**h**) chr.5:32,415,412, (intron of zinc-finger RNA binding protein, AJ314790).

for independent ncRNAs is provided as **Supplementary Table 5** online. All these conserved RNA structures are at least 10 kb away from the nearest known gene and also do not appear to be part of other plausibly predicted protein coding genes.

## Interpreting the results

In this computational study, we sought to draw a map of significant RNA secondary structures in the human genome. We applied a comparative approach making use of the recently sequenced mammalian/vertebrate genomes and screened the most conserved noncoding DNA for characteristics of functional RNAs, that is, (i) evolutionary conservation and (ii) thermodynamic stability of secondary structure. Our predictions are therefore restricted to RNAs whose spatial structure is of functional importance.

At the highest significance level, we predict structural RNA elements in 6.6% of the most conserved noncoding DNA (~36,000 structural elements throughout the genome). Our prediction covers 0.2% of the complete genome. The initial analysis underlines the value of this prediction: RNAz recovers hundreds of known structural RNAs (both ncRNAs and structural elements in UTRs of mRNAs), it identifies additional members of known ncRNA families, and detects previously undescribed conserved structural elements in some known ncRNAs.

The most intriguing result of our study, however, is the number of predicted structures that could not be assigned to known RNAs. Using randomized controls, we tried to assess the significance of these predictions. We estimate a false-positive rate of 1.1% and thus observe an overall signal-to-noise ratio of 6:1, implying that the majority of the predictions are biologically relevant. This estimate relies on a computational estimate of the background, which in turn is based on shuffled sequences. Clearly, any such approach can only approximate the true genomic background and hence cannot rule out the possibility that nonrandom sequence patterns that have not been described in the past could cause spurious hits resembling stable and conserved RNA structures. Although we are currently not aware of such effects, the estimated false-positive rate must therefore be seen as a lower bound.

About one-sixth of the predicted structures can be found in untranslated regions of annotated protein-coding genes and are thus potential *cis*-acting regulatory elements of the mRNA. A third of our hits are located in introns of annotated protein-coding genes. This finding strongly supports the notion that a plethora of functional RNAs are expressed from intronic DNA[35]. It is also conceivable that the structures play a regulatory role in the pre-spliced mRNA. Recently, a well conserved RNA secondary structure was shown to regulate alternative splicing in the homothorax gene in *Drosophila melanogaster*[36].

In some cases, the structures we detect might also be part of the coding region of mRNAs, since we excluded only well-annotated coding exons. Less than 6% of the RNAz hits, however, overlap with computationally predicted coding exons, which suggests that this accounts for only a small fraction of our hits.

One-half of the detected structures are located in intergenic regions at least 10 kb away from any known protein-coding gene. Given that the current protein-gene annotation of the human genome is fairly complete, one can assume that most of these hits are unrelated to mRNAs of protein-coding genes and thus are candidates for independent functional ncRNAs.

Our study points to thousands of so far unrecognized functional ncRNAs in the human genome. It provides a strong basis for further theoretical and experimental studies. A systematic analysis and classification of all detected RNA structures, anticipated and dubbed 'structural RNomics' a few years ago[37], together with rationally designed expression studies are promising strategies towards a better understanding of ncRNA function on a genome-wide scale. It is plausible to assume that the expression of many ncRNAs is difficult to verify directly in vertebrates because, like many known ncRNAs, their transcription is limited to specific tissues and/or developmental stages. Low concentrations, furthermore, might make it necessary to employ more sensitive detection techniques[38].

Despite the promising results reported above, the sensitivity and accuracy of *de novo* ncRNA prediction in the human genome is still nowhere as good as the predictions for protein-coding genes. At present the method is limited by the quality of the input data, both in terms of the number of available genomes at suitable evolutionary distances, and in particular in terms of alignment quality.

Although our computational predictions indicate that, consistent with previous estimates[8,35], the number of ncRNAs in vertebrate genomes is at least comparable to that of protein coding genes, we believe that the picture is still too incomplete to attempt a quantitative estimate.

*While this article was in production, a paper describing theory and applications of the PhastCons program was published by Siepel et al. (Genome Res.* ***15****, 1034–1050, 2005). Using a probabilistic method to model conserved RNA structures, the authors report strong statistical evidence for enrichment of RNA secondary structures in highly conserved noncoding regions of vertebrates. Meanwhile, the predicted microRNA in* ***Figure 3a*** *was cloned and named hsa-mir-499 (Nat. Genet.* ***37****, 766–770, 2005).*

© 2005 Nature Publishing Group http://www.nature.com/naturebiotechnology

## METHODS

**Alignments.** Genome-wide alignments of vertebrates ('multiz8way') were downloaded from the UCSC genome browser[18]. The alignments included sequences of up to eight species: human (hg17), chimp (panTro1), mouse (mm5), rat (rn3), dog (canFam1), chicken (galGal2), zebrafish (danRer1) and fugu (fr1). The chimp sequences were removed from the alignments because human and chimp are so similar that sequence differences between them provide essentially no information on RNA structure conservation.

**Selection of the most conserved noncoding regions.** We started from the 'Most Conserved' track generated by the PhastCons program. This track was edited as follows. (i) Adjacent conserved regions that are separated by <50 nucleotides were joined because many known ncRNAs are not conserved over the full length but only contain shorter fragments of highly conserved regions (in microRNA precursors, for example, the two sides of the stems are detected as conserved whereas the loop region in between is not). (ii) Conserved regions (after the joining step) with a length <50 nucleotides were removed because shorter RNA secondary structures are below the detection limit of RNAz. (iii) All regions with any overlap with annotated coding exons according to the 'Known Genes' and 'RefSeq Genes' annotation tracks were removed.

The initial set of alignments consisted of all Multiz alignments corresponding to regions in the modified 'Most Conserved' track. After the processing steps described below, we only considered alignments which were conserved at least in the four mammals ('input alignments').

**RNAz screen.** The input alignments were screened for structural RNAs using RNAz (version 0.1.1)[12]. Alignments with <200 columns were used as a single block. Alignments with length >200 were screened in sliding windows of length 120 and slide 40. This window size, on the one hand, appears long enough to detect local secondary within long ncRNAs and, on the other hand, is small enough to detect short ncRNAs (~50–70 nucleotides) without losing the signal in a much too big window.

The individual alignment blocks presented to RNAz were further processed in the following way. (i) We discarded alignments in which the human sequence contained masked positions by RepeatMasker. The vast majority of repeats was already filtered out in the input alignments; either they were not aligned by Multiz or not detected by PhastCons. (ii) Some alignments in the input set contained a large fraction of gaps resulting from a documented problem of PhastCons when treating missing data. We therefore further edited the alignments and removed sequences with more than 25% gaps. The region was regarded as not conserved in this species. If the human reference sequence contained more than 25% gaps, the complete alignment was discarded. (iii) The classification model of RNAz is currently only trained for up to six sequences. Therefore, we removed one sequence from alignments that were conserved in all seven species. One of the two sequences in the most similar pair of sequences in the alignment was removed because this pair provides the least comparative information. For the same reason only one representative was retained if two or more sequences in the alignment were 100% identical. (iv) Columns of gaps were removed from the reduced alignments.

The resulting alignments were scored with RNAz using standard parameters. All alignments with classification score $P > 0.5$ were stored. Finally, overlapping hits (resulting from hits in overlapping windows and/or hits in both the forward and reverse strand) were combined into clusters. The corresponding region in the human sequence was annotated as 'structured RNA' with the maximum $P$ value of the single hits in the cluster.

Clustering of RNAz hits with Blastclust yields only small groups of RNAs or isolated sequences. This rules out the possibility that a substantial fraction of the RNAz hits are derived from pseudogenes or belong to repeat families that so far have not been annotated.

**Estimating specificity.** The specificity of RNAz tested on shuffled alignments was found to be ≈99% and ≈96%, for $P = 0.9$ and $P = 0.5$, respectively[12]. For benchmarking RNAz, we used a defined set of high quality Clustal W alignments of 2–4 sequences and 60–100% mean pairwise identity. In this screen, however, we used automatically generated genome-wide alignments essentially based on Blast hits. It was therefore not clear if the specificity is the same on these alignments and how other parameters (e.g., the sliding window) affects the false-positive rate. We therefore estimated the false-positive rate for this particular special screen. To this end, we repeated the complete screen in exactly the same manner on randomized alignments. Alignments <200 columns were randomized as a whole, alignments >200 were randomized in nonoverlapping windows of 200 before they were sliced in windows for scoring as described above for the true data.

For randomization, we used a slightly modified version of the program shuffle-aln.pl (available on request) which is described in detail in reference 22. This program shuffles the positions in an alignment to remove any correlations arising from a native secondary structure. It takes care not to introduce randomization artifacts and generates random alignments of the same length, the same base composition, the same overall conservation, the same local conservation and the same gap pattern.

This procedure is very conservative and we found that it cannot remove the signal in all cases. The number of possible permutations is reduced if all of the alignment characteristics mentioned above are strictly preserved. Furthermore, the typical mutation pattern of noncoding RNAs is not removed by shuffling of the columns. The number of 'compatible' columns that can form a base pair in the consensus structure remains the same. This might be one reason why we observe a number of random hits overlapping with native hits (**Supplementary Table 1** online). Another reason for this effect might be that some alignments display special properties that cause an increased false-positive rate. As mentioned in the text, we observed this for highly conserved alignments with little covariance information.

In a screen of the urochordate *Ciona intestinalis* based on pairwise alignments[39], RNAz detected more than 300 tRNAs (about 55% of the tRNAscan-SE predictions) but found at $P > 0.5$ only 2 out of the more than 600 tRNA-pseudogenes predicted by tRNAscan-SE. This shows that RNAz very efficiently distinguishes between RNA secondary structures that are under stabilizing selection and similar sequences for which the selection pressure has been relaxed.

**Sensitivity on microRNAs and snoRNAs.** We used the 'sno/miRNA' track created from the microRNA Registry[40] and the snoRNA-LBME-DB maintained at the *Laboratoire de Biologie Moléculaire Eucaryote*. The track contained 207 unique microRNA loci, 86 H/ACA snoRNA and 256 C/D snoRNAs. We compared our predictions with the annotation tracks using the 'Table browser' feature of the UCSC Genome Browser. Loci overlapping with our predictions were counted as detected. Loci that did not show any overlap with our input alignments were counted as 'Not in input set' (**Fig. 2c**). We found that most of the microRNAs and snoRNAs are missed in our screen because they are not in our input set. To optimize future screens, and in particular sub-screens for miRNAs and H/ACA snoRNAs, we investigated in detail why miRNAs and H/ACA snoRNAs were missed in our selection of input alignments (**Supplementary Tables 2** and **3** online). MicroRNAs are mainly missed because they overlap with repeats or because they are not strictly conserved in all four mammals. (It is more likely that the corresponding sequences are simply missing in one of the unfinished draft assemblies, in particular of the rat genome.) H/ACA snoRNAs are not well conserved on sequence level and PhastCons cannot detect conserved regions >50 nucleotides in many of them. In the case of C/D snoRNAs, the problem is even more pronounced. Out of the 129 C/D snoRNAs not in our set, 63 are completely missed by PhastCons, in most of the other cases only short regions <50 are detected. Moreover, many snoRNAs which are contained in our set are not conserved over the full length. Given the fact the C/D snoRNAs in general do not exhibit very stable structures, the detection for RNAz is even more difficult if significant portions of the structure are missing in the input alignments.

© 2005 Nature Publishing Group http://www.nature.com/naturebiotechnology

**Noncoding RNA annotation.** We compared all hits to available databases of noncoding RNAs: Rfam (release 6.1, August 2004)[24], RNAdb (August 2004)[30], NONCODE (release 1.0, March 2004)[41], microRNA registry (release 5.0, September 2004)[40], UTRdb (April 2004)[42].

We generated Blast libraries for each of the databases and matched the human sequence of all the detected RNAz clusters against them. In the case of the UTRdb we used the EMBL formatted files from ftp://bighost.ba.itb.cnr.it/pub/Embnet/Database/UTR/data/ and extracted all annotated UTR elements >20 with flanking regions of 30 to build the Blast library. **Table 2** reports Blast hits with E-values < $10^{-6}$.

**Annotation relative to protein coding genes.** For annotating the RNAz hits relative to known protein coding genes (**Fig. 2d**), we used the 'Known Genes' and 'RefSeq Genes' annotation tables from UCSC genome browser. The UTR annotation is partly ambiguous. As a result, some hits in the second pie chart in **Figure 2d** are classified both as an intron of a coding region and an UTR. Counting only unambiguous annotations, 9,825, 2,095 and 1,987 hits are annotated as an intron of coding region, 3′-UTR and 5′-UTR, respectively.

**Comparison with tiling array transcriptional maps.** We downloaded the 11 'transfrag' annotation tracks for all cell lines and RNA fractions from http://transcriptome.affymetrix.com/. The annotation tracks were combined into one and the coordinates were converted from the 'hg15' assembly to 'hg17' using the liftOver tool and chain files provided by UCSC (http://hgdownload.cse.ucsc.edu/downloads.html). We then compared our annotation (Set 1, $P > 0.9$) on chromosomes 6, 7, 13, 14, 19, 20, 21, 22, X and Y with the transcription map using the 'Table browser' and determined the fraction of overlapping annotations. To estimate the significance, we generated randomized annotation tracks. For each predicted structural RNA, we randomly chose a nonrepeat region of the same length, on the same chromosome with the same annotation. We distinguished the following three annotation types: intergenic <10 kb from the nearest gene, intergenic >10 kb from the nearest gene and intronic. We did not consider regions in UTRs for this comparison. We compared five of such random tracks with the transcriptional map and found on average 29.6% overlapping annotations (the maximum overlap of all five tracks was 30.0%). To assess the detection performance on known miRNAs and snoRNAs, we used the annotation tracks described above.

*All annotation data can be downloaded from http://www.tbi.univie.ac.at/papers/SUPPLEMENTS/ncRNA/.*

*Note: Supplementary information is available on the Nature Biotechnology website.*

ACKNOWLEDGMENTS
This work was supported in part by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Project No. P15893, by the German *DFG* Bioinformatics Initiative BIZ-6/1-2, and by the Austrian *Gen-AU bioinformatics integration network* sponsored by bm:bwk.

COMPETING INTERESTS STATEMENT
The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. The Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
2. Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242–2246 (2004).
3. Kampa, D. *et al.* Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* **14**, 331–342 (2004).
4. Johnson, J.M., Edwards, S., Shoemaker, D. & Schadt, E.E. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.* **21**, 93–102 (2005).
5. Cheng, J. *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**, 1149–1154 (2005).
6. Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).
7. Imanishi, T. *et al.* Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biology* **2**, 0856–0875 (2004).
8. Cawley, S. *et al.* Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499–509 (2004).
9. Hüttenhofer, A., Schattner, P. & Polacek, N. Non-coding RNAs: hope or hype? *Trends Genet.* **21**, 289–297 (2005).
10. Hofacker, I.L. *et al.* Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucleic Acids Res.* **26**, 3825–3836 (1998).
11. Rivas, E., Klein, R.J., Jones, T.A. & Eddy, S.R. Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr. Biol.* **11**, 1369–1373 (2001).
12. Washietl, S., Hofacker, I.L. & Stadler, P.F. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA* **102**, 2454–2459 (2005).
13. Moulton, V. Tracking down noncoding RNAs. *Proc. Natl. Acad. Sci. USA* **102**, 2269–2270 (2005).
14. Shabalina, S.A. & Kondrashov, A.S. Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. *Genet. Res.* **74**, 23–30 (1999).
15. Shabalina, S.A., Ogurtsov, A.Y., Kondrashov, V.A. & Kondrashov, A.S. Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet.* **17**, 373–376 (2001).
16. Margulies, E.H., Blanchette, M., Haussler, D. & Green, E.D. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13**, 2507–2518 (2003).
17. Dermitzakis, E.T. *et al.* Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science* **302**, 1033–1035 (2003).
18. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
19. International Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
20. Cooper, G.M. *et al.* Characterization of evolutionary rates and constraints in three mammalian genomes. *Genome Res.* **14**, 539–548 (2004).
21. Le, S.V., Chen, J.H., Currey, K.M. & Maizel, J.V., Jr. A program for predicting significant RNA secondary structures. *Comput. Appl. Biosci.* **4**, 153–159 (1988).
22. Washietl, S. & Hofacker, I.L. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J. Mol. Biol.* **342**, 19–30 (2004).
23. Hofacker, I.L., Fekete, M. & Stadler, P.F. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* **319**, 1059–1066 (2002).
24. Griffiths-Jones, S. *et al.* Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121–D124 (2005).
25. Accardo, M.C. *et al.* A computational search for box C/D snoRNA genes in the *D. melanogaster genome. Bioinformatics* **20**, 3293–3301 (2004).
26. Childs, J.L., Poole, A.W. & Turner, D.H. Inhibition of *Escherichia coli* RNase P by oligonucleotide directed misfolding of RNA. *RNA* **9**, 1437–1445 (2003).
27. Lin, J. *et al.* A universal telomerase RNA core structure includes structured motifs required for binding the telomerase reverse transcriptase protein. *Proc. Natl. Acad. Sci. USA* **101**, 14713–14718 (2004).
28. Avner, P. & Heard, E. X-chromosome inactivation: counting, choice, and initiation. *Nat. Rev. Genet.* **2**, 59–67 (2001).
29. Rougeulle, C. & Heard, E. Antisense RNA in imprinting: spreading silence through Air. *Trends Genet.* **18**, 434–437 (2002).

30. Pang, K.C. *et al.* RNAdb — comprehensive mammalian noncoding RNA database. *Nucleic Acids Res.* Database issue. **33**, D125–D130 (2005).
31. Hüttenhofer, A. *et al.* RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J.* **20**, 2943–2953 (2001).
32. Bachellerie, J.-P., Cavaillé, J. & Hüttenhofer, A. The expanding snoRNA world. *Biochimie* **84**, 775–790 (2002).
33. Berezikov, E. *et al.* Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120**, 21–24 (2005).
34. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
35. Mattick, J.S. RNA regulation: a new genetics? *Nat. Rev. Genet.* **5**, 316–323 (2004).
36. Glazov, E.A., Pheasant, M., McGraw, E.A., Bejerano, G. & Mattick, J.S. Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mrna splicing. *Genome Res.* **15**, 800–808 (2005).
37. Doudna, J.A. Structural genomics of RNA. *Nat. Struct. Biol.* **7**, 954–956 (2000).
38. Hartig, J.S., Grüne, I., Najafi-Shoushtari, S.H. & Famulok, M. Sequence-specific detection of microRNAs by signal-amplifying ribozymes. *J. Am. Chem. Soc.* **126**, 722–723 (2004).
39. Missal, K., Rose, D. & Stadler, P.F. Non-coding RNAs in *Ciona intestinalis. Bioinformatics* **21**, Suppl 2, ii77–ii78 (2005).
40. Griffiths-Jones, S. The microRNA Registry. *Nucleic Acids Res.* **32**, D109–D111 (2004).
41. Liu, C. *et al.* NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res.* Database issue. **33**, D112–D115 (2005).
42. Pesole, G. *et al.* UTRdb and UTRSite: specialized databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs. Update 2002. *Nucleic Acids Res.* **30**, 335–340 (2002).
43. Scherer, S.W. *et al.* Human chromosome 7: DNA sequence and biology. *Science* **300**, 767–772 (2003).

© 2005 Nature Publishing Group http://www.nature.com/naturebiotechnology