BMC Genomics

BioMed Central

Research article

Open Access

A systematic search for new mammalian noncoding RNAs indicates little conserved intergenic transcription

Tomas Babak[1,2], Benjamin J Blencowe[1,2] and Timothy R Hughes*[1,2]

Address: [1]Banting and Best Department of Medical Research, 112 College St., Toronto, ON M5G 1L6 Canada and [2]Department of Medical Genetics and Microbiology, 10 King's College Circle, Toronto, ON M1R 4F9 Canada

Email: Tomas Babak - tomas.babak@utoronto.ca; Benjamin J Blencowe - b.blencowe@utoronto.ca; Timothy R Hughes* - t.hughes@utoronto.ca

* Corresponding author

Published: 05 August 2005

BMC Genomics 2005, **6**:104 doi:10.1186/1471-2164-6-104

Received: 22 July 2005

Accepted: 05 August 2005

This article is available from: http://www.biomedcentral.com/1471-2164/6/104

Abstract

Background: Systematic identification and functional characterization of novel types of noncoding (nc)RNA in genomes is more difficult than it is for protein coding mRNAs, since ncRNAs typically do not possess sequence features such as splicing or translation signals, or long open reading frames. Recent "tiling" microarray studies have reported that a surprisingly larger proportion of mammalian genomes is transcribed than was previously anticipated. However, these non-genic transcripts often appear to be low in abundance, and their functional significance is not known.

Results: To systematically search for functional ncRNAs, we designed microarrays to detect 3,478 intergenic and intronic sequences that are conserved between the human, mouse, and rat genomes, and that score highly by other criteria that characterize ncRNAs. We probed these arrays with total RNA isolated from 16 wild-type mouse tissues. Among 55 candidates for highly-expressed novel ncRNAs tested by northern blotting, eight were confirmed as small, highly-and ubiquitously-expressed RNAs in mouse. Of the eight, five were also detected in rat tissues, but none were detected at appreciable levels in human tissues or cultured cells.

Conclusion: Since the sequence and expression of most known coding transcripts and functional ncRNAs is conserved between human and mouse, the lack of northern-detectable expression in human cells and tissues of the novel mouse and rat ncRNAs that we identified suggests that they are not functional or possibly have rodent-specific functions. Our results confirm that relatively little of the intergenic sequence conserved between human, mouse and rat is transcribed at high levels in mammalian tissues, possibly suggesting a limited role for transcribed intergenic and intronic sequences as independent functional elements.

Background

Comparative genomics has revealed that approximately 5% of the mammalian genome is under purifying selection [1,2]. While exons make up roughly 1.5% of the genome [3], relatively little is known about the role of the remaining 3.5% of the highly conserved genomic regions, and even less about the functional potential of evolutionarily-diverged intergenic sequences. Large-scale microarray tiling analyses (i.e. using a set of probes designed to detect all or most of a targeted genome or genomic region), as well as high-throughput cDNA sequencing efforts, have indicated that the "transcriptome" is significantly larger than was previously appreciated, although the functional significance of the vast majority of the

novel, apparently noncoding (nc) transcripts detected by these approaches has remained elusive [4-8]. To date, several studies have reported large-scale tiling efforts of the human genome [4,5,9,10]. In all cases a significantly higher proportion of transcribed sequence was reported than could be accounted for by existing exon annotation data, and much of the remainder did not appear to encode protein [4]. Comparison of datasets suggests that a high proportion of the novel transcripts are specific to tissues or cell lines [4,9]. This trend was particularly evident for cell lines, where novel cell-line specific transcripts were even more numerous than annotated cell-line specific exons [4], implying that many of these transcripts may not have endogenous functions in whole organisms. Further supporting this possibility was the observation that the majority of the novel transcripts were detected at very low levels [5].

A second source of evidence for a more extensive transcriptome arises from large-scale cDNA compilation efforts. The mouse cDNA sequencing effort led by the RIKEN Consortium identified 60,770 unique cDNA transcripts from a variety of mouse tissues and cell lines [7]. Approximately half (33,409 sequences) were derived from unique genomic locations (Transcriptional Units), of which 15,815 did not map to known or predicted coding genes in mouse [7]. Further refinement identified a set of 4,280 mRNA-like noncoding RNAs which had no homology to any known protein sequences and comprised of sequences mapped to regions located between predicted exon boundaries [6]. Many of these sequences were reported elsewhere in EST databases and displayed features of polymerase II transcripts [6]. However, unlike protein-coding mouse genes, of which 99% have homologs in the human genome [1], only 10.6% of the 4,280 apparent nc transcripts were represented by homologous sequences in the human genome [6,7]. In fact, Wang et al. [11] demonstrated that most of these transcripts are no more conserved than intergenic sequence in general, and less conserved than a comprehensive set of 321 known ncRNAs with established functional roles. In addition, expression profiling of a different but overlapping (FANTOM1) subset of cDNAs that do not map to known ESTs or protein sequences (3,388), revealed that most transcripts in this uncharacterized class were present at low abundance [12]. Collectively, these results demonstrate the transcription of uncharacterized sequence, but raise questions about the functional relevance of the novel "noncoding" set.

One possible explanation for the observed low-level expression of a much larger fraction of genomes than can be accounted for by known genes comes from the recent discovery of a nuclear posttranscriptional quality-control pathway that degrades "cryptic unstable transcripts" (CUTs) in yeast [13]. CUTs are transcribed by Pol II and are detectable by both microarrays and RT-PCR in wild-type yeast, and also appear to be frequently represented as single tags in SAGE libraries, but are undetectable by Northern blotting and do not contain significant open reading frames. However, in mutants in the quality-control pathway, they appear as a smear on Northern blots due to the fact that they have heterogeneous 3' ends [13]. The fact that a posttranscriptional quality control exists to prevent accumulation of CUTs suggests that they are aberrant and predicts that there should be little selection pressure on their expression. Moreover, these observations suggest that nonfunctional transcripts might be distinguished from *bona fide* functional transcripts on the basis of formation of a discrete species on Northern blotting, and by conservation of expression among different organisms.

In this study, we describe a systematic approach to predict and screen novel ncRNA transcripts in the mouse genome. We first identified non-exonic sequences that are most likely to encode functional ncRNAs (functional transcripts that do not encode proteins) by using the program QRNA, which searches for conserved regions with compensatory mutation patterns that are consistent with the evolutionary conservation of secondary structure in functional noncoding sequences [14]. These are hallmarks of most known functional ncRNAs, and QRNA has been used successfully to identify novel structural ncRNAs in *E. coli* [15] and *S. cerevisiae* [16]. However, even in these organisms, which have relatively compact genomes, a high false-positive prediction rate was observed [16], which presents a challenge for screening large genomes. We therefore used a custom oligonucleotide microarray [17] as an initial high-throughput screen for expression. We then tested the 55 highest-expressed candidates to ask whether they are detectable as discrete species on Northern blots. We report eight novel mouse transcripts identified using this approach. However, none of the eight appears to be expressed in humans, casting doubt on their role as independent functional elements. Taken together with the low proportion of intergenic sequences that we detected, our results suggest that much of the recently-discovered expanded transcriptome [4-7,9,10] may correspond to cryptic transcripts [13], suggesting a limited role for transcribed intergenic and intronic sequences as independent functional elements.

Results

Predicting novel ncRNA candidates

We identified novel structural ncRNA candidates on the basis of two features: high sequence conservation and a mutation pattern consistent with sequences being under selective pressure to maintain a conserved secondary structure [see Methods for details]. Figure 1 outlines the

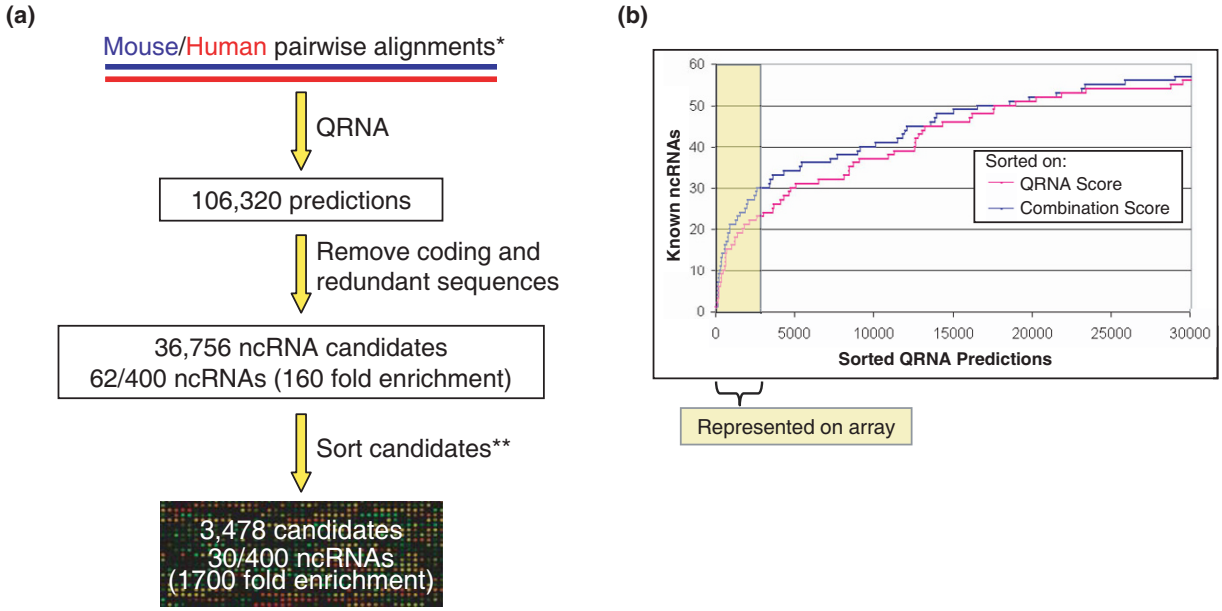

Figure 1
Summary of computational prediction and sorting of novel mammalian noncoding RNA. **(a)** Whole genome alignments were downloaded from UCSC [18], subset to regions of greater than 85% sequence identity, and analyzed with QRNA [14]. Removal of redundant and coding sequences left 36,756 ncRNA candidates containing 62 known ncRNAs (of approximately 400 known ncRNAs). Candidates were then sorted and the top 3,478 predicted RNAs, which contained 38 known ncRNAs (representing a 1,700 fold enrichment of real ncRNAs over random selection from genomic sequence), were selected for further screening. **(b)** Sorting was based on the QRNA score, stability (measured as a predicted free energy change using Mfold [44]), overlap with rat-mouse QRNA predictions, and genomic clustering (many predictions close to one another are likely the same transcript). This combination of four criteria was more powerful in identifying real ncRNAs than using the QRNA score alone. *UCSC [18], **See **(b)**.

computational screen we employed for finding novel ncRNAs. We obtained human-mouse pairwise sequence alignments from UCSC [18] and subset them to alignments with a minimum of 85% sequence identity using a 200 nt scanning window. This eliminated sequences that are unlikely to be under evolutionary selection [1,11] and reduced the dataset to a computationally manageable size. We used QRNA v.1.1 [14-16] to screen the alignments for putative ncRNAs. This generated 106,320 predicted ncRNAs. We then removed redundant sequences and predicted ncRNAs with sequences similar or identical to coding genes annotated in the Mouse RefSeq mRNA database [19], RIKEN cDNA [7], Mouse Protein NR database [20], or coding genes in other organisms annotated in GenBank [21]. The remaining 36,756 predicted ncRNAs were sorted by logistic regression using four parameters that we identified to be useful for distinguishing known ncRNAs from QRNA predictions of putative new ncRNAs: 1) the QRNA logodds RNA score; 2) the thermodynamic stability of the predicted secondary structure of each prediction; 3) a genomic clustering score of closely mapped predictions on the genome (presumed to be multiple regions of a longer ncRNA transcript); and 4) an overlap between mouse-human and mouse-rat QRNA predictions (processed similarly to mouse-human alignments). The QRNA score was the strongest indicator of defined ncRNAs (data not shown), but combining the additional parameters increased the sorting power, especially for the top 10% of the predicted RNAs (Fig. 1B). Other parameters such as GC content and length of the sequences did not improve the sorting (data not shown). To further characterize the set of predicted ncRNAs, we screened them computationally for tRNAs and snoRNAs [22,23] and searched for similar sequences in the RIKEN FANTOM2 mRNA-like ncRNA collection. For a summary of these features, see Additional file 1.

Table 1: List of tissues used in microarray expression analysis

Tissues/Organs/Cells screened for novel noncoding RNAs
Bladder
Brain
Embryonic Stem Cells
Femur
Heart
Intestine
Liver
Lung
Mammary Gland
Muscle
Stomach
Teeth
Testis
12.5-day Embryo
15-day Embryo
9.5-day Placenta

Array design

Due to the generic nature of the algorithm, QRNA has a high false positive rate, much higher than coding gene-finding algorithms, thus experimental validation is essential. Using the prediction scheme and sorting criteria described above, we designed a microarray to detect 3,478 QRNA predictions with properties most indicative of ncRNAs. The microarray contained probes for the top 9.5% of the total ncRNA predictions, and probes for 38 known ncRNAs [see Additional file 2]. The design included 20,867 complementary DNA probes to the QRNA predictions (with six probes per prediction; three for each orientation), 200 random probe sequences and 305 intergenic probes that served as negative controls, and 705 positive control probes tiled across U4, U5 and mature rRNA transcripts. A list of all microarray probes is included in Additional file 3.

Analysis of RNA from diverse mouse tissues

Since ncRNAs can be expressed in a tissue-or developmental-stage-specific manner [24-29] we screened 16 mouse tissues/organs encompassing a variety of tissue sources, including two embryonic stages of development (Table 1). The intensity distribution over all measurements is shown in Figure 2. As expected, the majority of QRNA-prediction intensity measurements overlapped with the negative control probe measurements, presumably due to the high false-positive rate inherent in generating the predictions. However, the distribution was skewed to the right tail of the plot (i.e. higher intensity), and twice as many of measurements from the predicted ncRNAs were above the 99% negative control threshold (generated using random sequence probes) than would be expected based on a random distribution (Fig. 2). Of the 38 known ncRNAs that were among the QRNA predicted RNAs [see Additional file 2], we detected 15, including several snoRNAs, tRNAs, a Hox antisense transcript, and miRNAs, using the same intensity cutoff used for selecting novel candidates (see below). This illustrates that the sensitivity of this technique is sufficient to detect most known ncRNA types. We have also used this technique to survey miRNA expression [24].

Validation by northern blotting

Although our microarray data indicated that many of the measurements arise from real transcripts, noise (e.g. spurious cross-hybridization) could also account for some proportion of the high-intensity measurements. Furthermore, microarray results cannot differentiate between a single RNA species and a heterogeneous population. We therefore used Northern blotting to validate our candidate novel ncRNAs. Northern blotting is more quantitative than RT-PCR, since there is no exponential amplification step. It is less sensitive for the same reason; however, using our methods, we have been able to detect all types of transcripts including structural ncRNAs, miRNAs, and mRNAs ([24,30] and data not shown). Importantly, since Northern blotting reveals the size of the RNA species detected, it can distinguish whether there is a single RNA product species and a heterogeneous transcript population. We tested all predicted ncRNAs detected by at least two of three probes (all in the same orientation) displaying signals above the 99%-negative control intensity threshold in at least one tissue. In total, this included 55 novel predicted ncRNAs, of which most appeared to be ubiquitously expressed. Northern analysis on this subset confirmed 8 novel transcripts (Fig. 3), all of which were detected ubiquitously in total RNA isolated from 16 wild-type mouse tissues. All eight transcripts were between 70 and 140 nt in length, none had tRNA or snoRNA structural or sequence characteristics, and five were located in intronic regions. It is possible that additional RNAs are expressed at low levels that are detectable by microarray and/or RT-PCR but not by Northern blotting, especially if they are heterogeneous in length [13]. We did not pursue this possibility, since it seemed that transcripts undetectable by Northern blotting are less likely to represent bona fide ncRNAs.

Expression of novel mouse transcripts is not conserved in human cells/tissues

Nearly all sequence-conserved coding genes between human and mouse have a conserved expression pattern across multiple tissues [31]. Although they have not to our knowledge been comprehensively analyzed, ncRNAs are also generally expressed similarly across related species [32] and since most are required for cell proliferation, they tend to be expressed in all tissues, as were all eight of the novel transcripts we observed. However, Northern

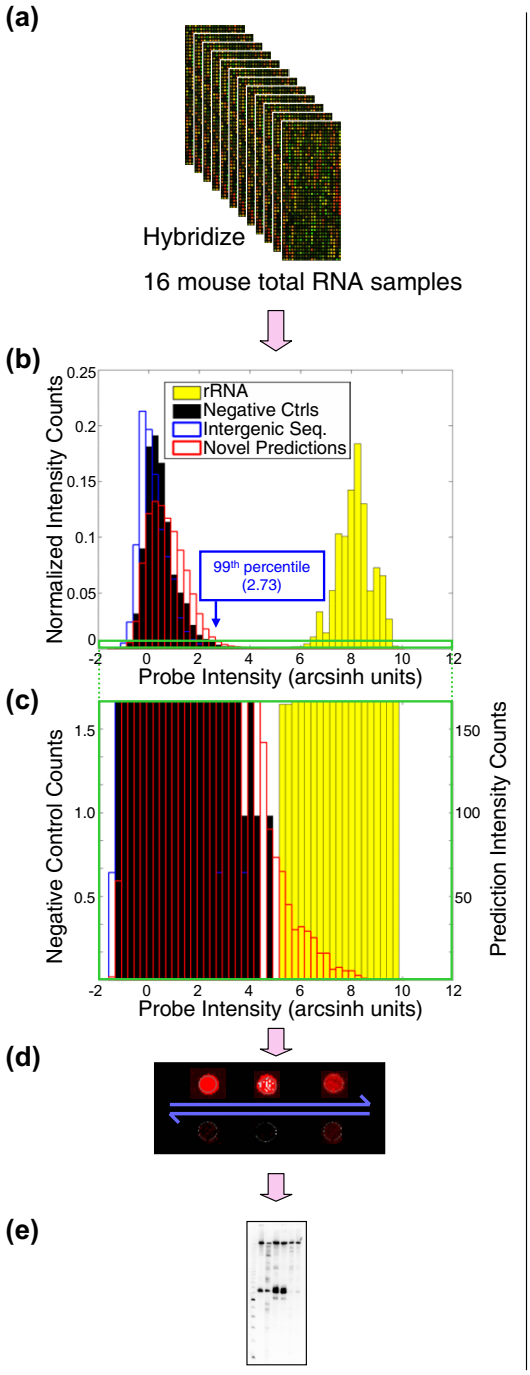

Figure 2
Screening, selection, and confirmation of novel ncRNA predictions. **(a)** Total RNA from 16 tissues was hybridized to custom Agilent microarrays containing probes to the QRNA predictions. **(b)** Most intensity measurements overlapped negative-control intensities (both random probe sequences and probes corresponding to randomly sampled intergenic regions), although a right-tailed distribution overlapping rRNA levels indicated detection of potential abundant novel transcripts. **(c)** Expanded y-axis region from **(b)**, axes denote absolute probe intensity counts. **(d)** Sample schematic of microarray spots corresponding to a transcript that was tested further by northern analysis. **(e)** 55 transcripts in total were screened by Northern analysis.

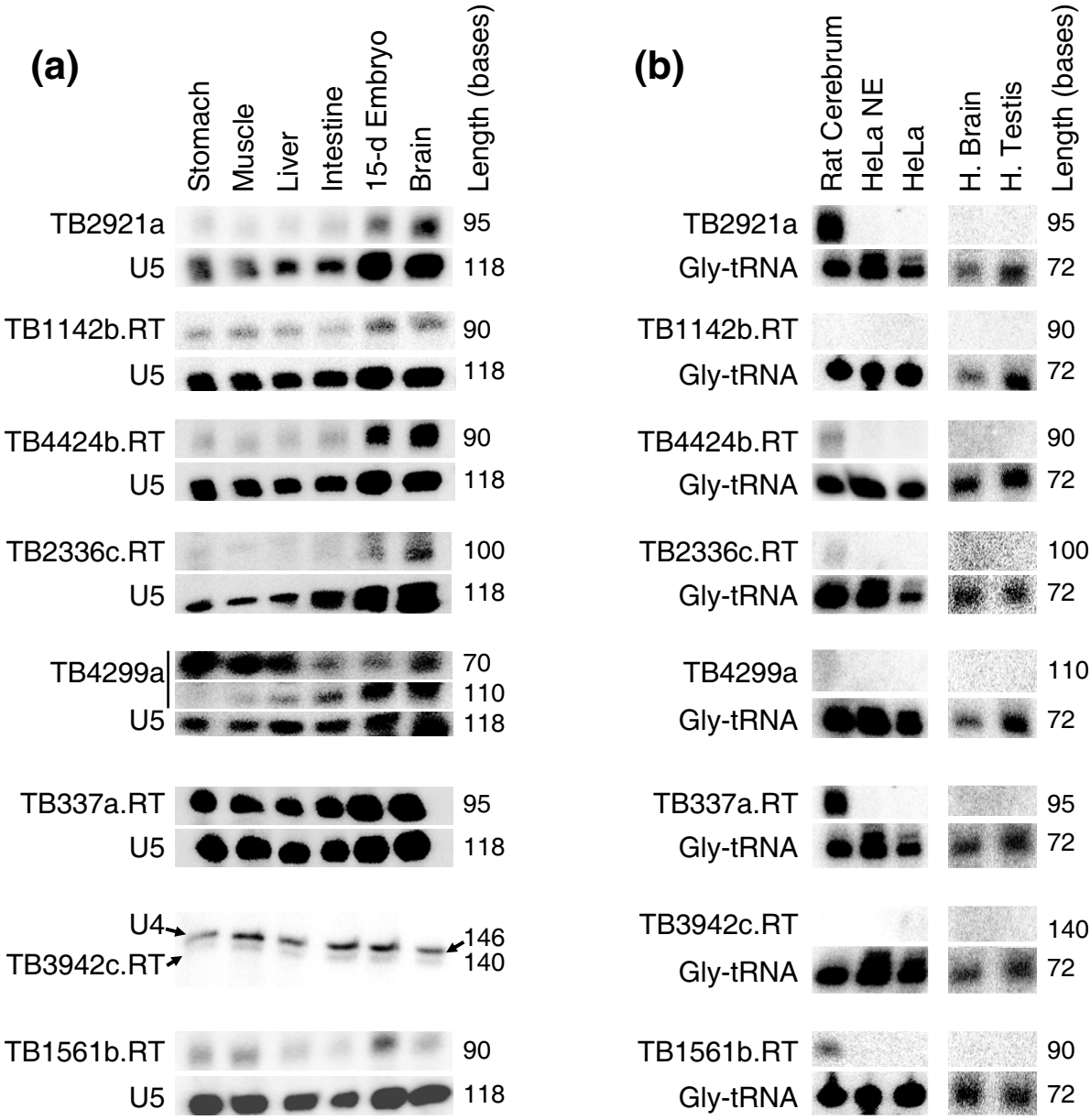

Figure 3
Abundantly and ubiquitously expressed novel mouse transcripts are not expressed in human tissues or cells. **(a)** Northern analysis of mouse transcripts using mouse-specific probes. U5 and U4 probes were used as loading controls as indicated and were co-hybridized with test probes. **(b)** No signal was detected in Northern analyses using human-specific probes. Human tissues were analyzed with a longer exposure (right panel) since short RNAs were slightly underrepresented in these commercially-obtained samples. Mouse-, and human-specific probe sequences complementary to the novel ncRNA predictions and images of all full-scale northern blots are available in Additional files 4 and 5.

blotting revealed that none of the eight novel mouse transcripts were expressed at detectable levels in HeLa cells or in human tissues (Fig. 3). Moreover, only five of these were detected in rat (Fig. 3). Images of all full-scale Northern blots shown in Figure 3 and other supporting Northern data is available in the supplementary data [see Additional file 4].

We also compared our 3,478 QRNA predictions and the RNAs we verified with recently published high-density human tiling data from Cheng et al. [9]. We subset our QRNA predictions to regions surveyed by Cheng et al. [9] (i.e. the non-repetitive regions of human chromosomes 6, 7, 13, 14, 19, 20, 21, 22, X, and Y). We considered the "transfrags" (i.e. transcribed fragments: any transcribed genomic region) described by Cheng et al. [9] in poly-A-minus RNA from HepG2 cells, which should be most comparable to our data (Fig. 4A). We confirmed that the Cheng et al. [9] "transfrags" encompassed a larger number of known noncoding RNAs and mRNA exons than was obtained from random positioning of sequences of the same length (Fig. 4B), although the vast majority of "transfrags" do not overlap any annotated features. We did not see a marked difference in the overlap between the Cheng et al. "transfrags" and our QRNA predictions (Fig. 4B). This indicates that the "transfrags" are not enriched for conserved sequence with conserved secondary structure, consistent with our data showing a lack of conserved expression of our northern-confirmed QRNA transcripts in human tissues and cells. Only one of the eight northern-confirmed novel mouse transcripts we verified mapped to the regions surveyed by Cheng et al. [9] and it did not overlap a "transfrag", also consistent with our results (Fig. 3).

Discussion

Using comparative genomics and an established ncRNA search method modified for high-throughput screening, we report eight novel mouse ncRNA transcripts that are all relatively short, ubiquitously expressed, and abundant. Despite their sequence and secondary structure conservation, none of the transcripts were expressed at detectable levels in human cells and tissues.

Given the large search space incorporated in this analysis, our results indicate that little intergenic or intronic sequence is expressed as distinct, stable transcripts at levels comparable to the expression of most known functional RNAs. This deduction contrasts somewhat with the conclusions of recent studies employing tiling arrays or large-scale cDNA sequencing [5,7,9,10]. Because results from other studies were obtained and validated in different ways, we cannot confirm or refute the basic observations of any other study. Numerous explanations exist for the breadth of the emerging transcriptome [10]. Nonetheless, we propose that claims regarding a dramatically larger transcriptome than is accounted for by current annotations should be addressed with scrutiny, particularly with regard to functional potential. Several factors indicate that a significant proportion of the newly measured transcripts may either be spurious or non-functional: 1) transcriptionally active regions identified in tiling analyses and potentially noncoding cDNAs are generally detected in low abundance [4,6]; 2) in yeast, many intergenic regions are also transcribed at low levels, apparently as heterogeneous species, and there is a specific mechanism for degrading these transcripts [13]; 3) much of the mammalian data available is from cell lines, including a high proportion of tumor-derived cell lines [5,9], which may lack the same degree of quality-control as found in normal cells and tissues; 4) most are relatively short (i.e. sequenced transcripts are shorter than the average coding gene [7], as are transcripts identified from tiling, which are on average less than 200 nt [4]); 5) potentially noncoding cDNAs correspond to regions not conserved at the sequence level [6,7]) and have evolved at a non-selective rate [11]; 6) there is little evidence for cross-species expression (only 2.6% of noncoding mouse cDNAs can be mapped to human ESTs [6,7]); 7) 70% of intergenic "transfrags" corresponding to novel transcribed regions could not be detected by northern analysis [4]); 8) the "transfrags" do not appear to be enriched in sequences with conserved secondary structures (Fig. 4), which is a hallmark feature of known structural ncRNAs [33].

How can we distinguish *bona fide* functional transcripts, in the absence of directed genetic experimentation? Sequence conservation alone is apparently not sufficient to distinguish sequences with critical functions, as large-scale non-genic deletions encompassing highly conserved regions can be tolerated in mice without detectable fitness disadvantages [34]. The presence of conserved expression improves the likelihood that sequence-conserved regions are functional since most characterized RNA classes, including coding mRNA, and noncoding rRNA, tRNA, snRNA, snoRNA, and miRNA, generally exhibit conserved expression patterns across evolutionarily-related species [31,32,35,36]. Of the few characterized mRNA-like ncRNAs, some also have conserved expression patterns [29,32,37]. The lack of conserved expression of the eight transcripts identified in our study, despite a high sequence and structural similarity, suggests that they are not functional, although it is possible that a subset of functional transcripts have species-or lineage-specific functions despite their high degree of sequence conservation. For example, a subset of the ncRNAs detected in mouse in our study were also detected in rat tissues, suggesting the possibility of conserved functions restricted to the rodent lineage. However, in reported cases of mouse-specific

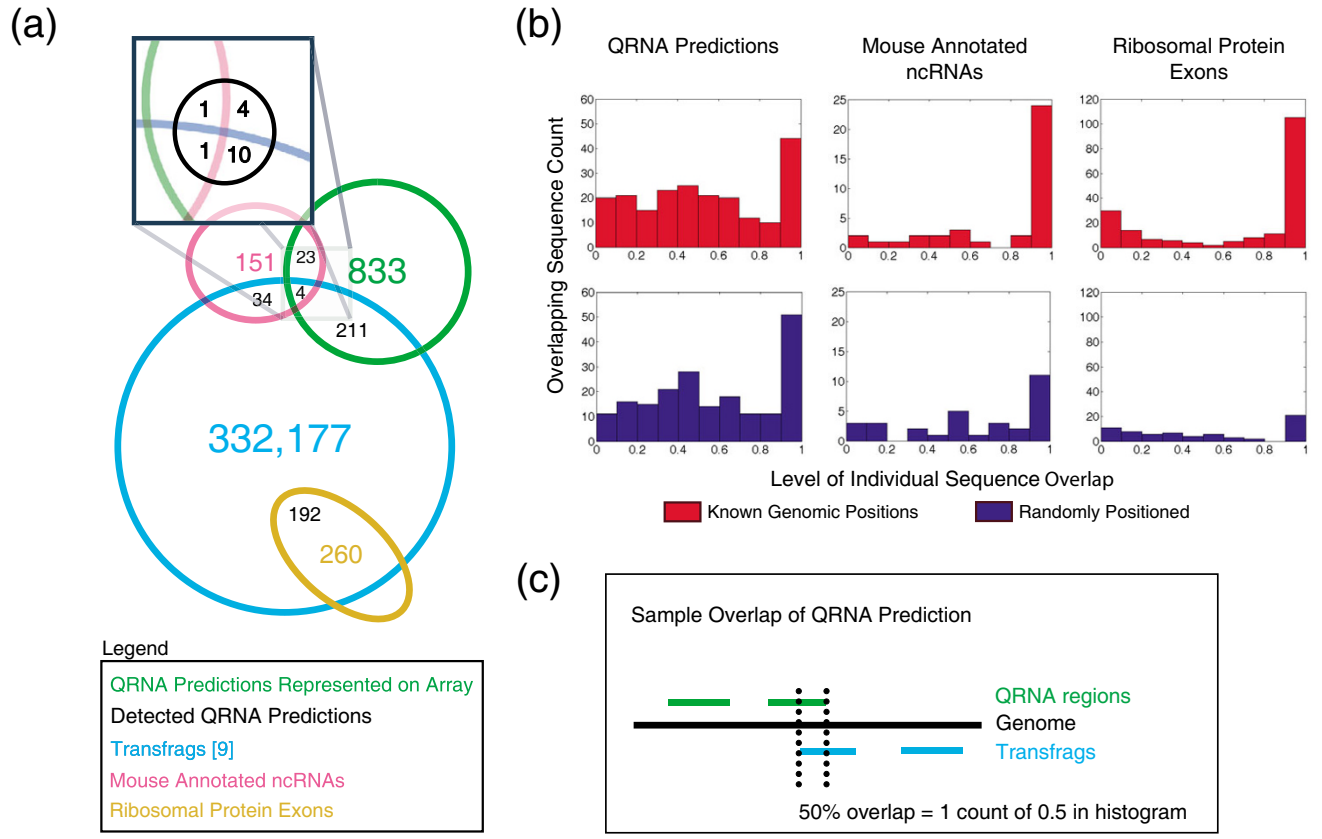

Figure 4
Overlap between known and predicted ncRNA types. **(a)** 3,478 QRNA predictions, 390 mouse annotated ncRNAs, and 716 human ribosomal exons were mapped to the human genome (hg17) regions surveyed by Cheng et al. [9] using poly-A minus RNA derived from HepG2 cells. 833 QRNA predictions, 151 mouse ncRNAs, and 260 exons are located in regions surveyed by Cheng et al. **(b)** Shown are distributions of the percentage of overlap of each sequence with a transfrag for actual genomic positions versus randomly positioned sequences in regions surveyed by Cheng et al. [9]. **(c)** Schematic of how an overlap was calculated for one QRNA prediction that overlaps a transfrag. This was repeated for all overlaps and the distributions are shown in **(b)**.

ncRNAs, such as BC1, Tsix, CIOR, and t-ncb [32], the RNAs were not conserved at the sequence level which is likely the reason for mouse-specific ncRNA differential regulation.

More cross-species and non-biased expression data is required to definitively address the likelihood of functionality of emerging transcriptomes. The most comprehensive approach will likely be an extension of whole genome tiling microarray analyses [38] using RNA derived from endogenous tissues from a variety of organisms. The approach of hybridizing covalently labeled total RNA (applied in this paper), as opposed to cRNA or cDNA derived from poly-adenylated RNA, presents a potential improvement to the unbiased nature of tiling analyses, since there is no amplification bias and strand information is retained. An added dimension of conserved expression will enable focused functional experimentation on transcripts that are likely to be important, although our data indicates that these cases will be the exception rather than the rule.

Conclusion

With the application of high-throughput transcriptional analyses it has been reported that more sequence is transcribed than was previously appreciated, with some estimates exceeding twice that of currently annotated transcripts. In a systematic search for sequence-conserved transcripts with hallmarks of structural ncRNAs, we identified only eight novel mouse transcripts with ubiquitous and abundant expression. This indicates that very little intergenic sequence is transcribed at high levels. Furthermore, despite meeting the stringent requirements of characterized ncRNAs, none of these eight transcripts were expressed at detectable levels in human cells or tissues. This suggests that these transcripts are unlikely to have conserved functional roles. We propose that newly-identified transcriptomes should be viewed with scrutiny, particularly with regard to function, until it is determined that they are functional or at least display properties of known functional elements.

Methods

Predicting novel ncRNAs

Whole genome pairwise human-mouse alignments were downloaded from the UCSC Genome Bioinformatics website ([18]; build 32, Nov. 2003). Repeat-masked [39] alignments were subset to segments with a minimum of 85% sequence identity. QRNA v.1.1 [14] was used to score the alignments for noncoding RNA potential using settings determined to work optimally on a test set of ncRNAs embedded in random alignments of equivalent sequence identity (default settings with -w 100 -x 50). The processing time was approximately 14 days on a 20-processor (1 GHz) linux cluster. Overlapping sequences with a positive logodds RNA score were concatenated into one sequence which was assigned the highest score of the original component sequences.

Selection of candidate ncRNAs

QRNA predictions were filtered by alignment to a variety of coding sequence databases using BLAT [40] with a default score cutoff of 30 ([alignment length] – [number of mismatches]) and a minimum sequence identity of 60% (-minIdentity = 60). The databases included: Mouse RefSeq mRNA [19] (build 29), ENSEMBL coding transcripts [41], RIKEN cDNA [7], ESTs [42] (download date: Nov. 2003), Unigene [43] (Nov. 2003), Protein NR [20] (build 29), and Genbank NT Database [21] (Nov. 2003). Redundant QRNA predictions and predictions that aligned to annotated coding sequences were removed. The remaining set was screened for tRNAs and Box C/D snoRNAs using tRNAScan SE [22] and snoscan [23] respectively. Sequences were extended by 100 bases in both directions from the genome to ensure complete coverage of potential tRNAs or snoRNAs.

QRNA predictions that did not map to annotated coding genes were sorted on a combination of criteria to maximize selection of known ncRNAs (as compiled in [24]). Sorting parameters included the minimum free energy as predicted by Mfold [44], overlap between mouse-human and mouse-rat QRNA predictions (blast, e-threshold 10^{-4}), and proximity to adjacent predictions in the genome. Genomic proximity was scored by adding the number of QRNA predictions within 1000 bp of each other in the genome. Mouse-rat alignments [18] were processed identically to mouse-human alignments. Multiple linear regression was used to assign weights to these parameters in addition to the QRNA logodds RNA score and were subsequently used to calculate an overall score for each QRNA prediction. The top 3,478 predicted RNAs (limited by space on the array) were analyzed further by microarray. These contained 38 known ncRNAs of the approximately 400 known ncRNAs, representing a 1,700-fold enrichment. The level of enrichment was calculated as the ratio of the proportion of nucleotides that are real ncRNAs in the QRNA predicted set to the proportion of nucleotides of all known ncRNAs of the mouse genome (i.e. how much more likely one could select a nucleotide belonging to a known ncRNA in the QRNA set over the whole genome).

Microarray design

Six probe sequences were allotted for each ncRNA prediction; three for each orientation. Complementary DNA probes were designed to maximize spatial coverage of each predicted sequence and were normalized by length (i.e. probe lengths were adjusted) to a uniform melting temperature of 60°C. Probe sequences were on average 26.9 nt and were concatenated to 60 nucleotides. Probe sequences were submitted to Agilent Technologies for microarray production (Palo Alto, California). The designs included 200 60-mer probes containing random sequences, which were used as negative controls, and 696 positive control probes tiled across U4 and U5 snRNAs and 18S and 28S rRNAs. Additional file 3 contains a list of all of the probe sequences.

RNA extraction, labeling, and hybridizations

HeLa nuclear extract (NE) was prepared as described previously [45]. Total RNA from HeLa cells, HeLa NE, and mouse tissues was extracted using Trizol (Invitrogen) according to the manufacturer's instructions and was treated with DNase I (Fermentas). Total RNA derived from human tissues was purchased from Clontech (BD Biosciences, Mississauga, ON) and Ambion (Austin, TX). Integrity of rRNA was confirmed on 1% agarose-formaldehyde gels. 7 μg of total RNA was chemically labeled with Ulysis Alexa Fluor 546 or Ulysis Alexa Fluor 647 (Ulysis) according to manufacturer's instructions. This protocol labels G residues [46], and there were no predicted RNAs

that lacked G residues. Samples were resuspended in 0.5 mL of hybridization buffer (1 M NaCl, 0.5% sodium sarcosine, 50 mM N-morpholino ethane sulfonate, pH 6.5, 33% formamide and 40 μg salmon sperm DNA), denatured by heating at 65°C for 5 minutes, and snap-cooled on ice prior to hybridization. Hybridizations were carried out for 16–24 h at 42°C in a rotating hyb oven. Slides were then washed (rocking ~30 seconds in 6× SSPE, 0.005% sarcosine, then rocking ~30 seconds in 0.06× SSPE) and scanned with a 4000A microarray scanner (Axon Instruments, Union City, CA).

Microarray data processing and normalization

TIFF images were quantified with GenePix 3.0 (Axon Instruments, Union City, CA). Individual channels were spatially detrended (i.e. overall correlations between spot intensity and position on the slide removed) by high-pass filtering [47] using 5% outliers. The 16 individual channels were then normalized using Variance Stabilization [48,49] and transformed to arcsinh values, which are similar to natural log values but are tolerant of negative numbers emerging from high-pass filtering.

Northern blotting

7 μg of total RNA from each tissue was separated on 10% polyacrylamide/TBE/urea gels, and electroblotted to Hybond N^+ or Hybond-XL membranes (Amersham) using asemi-dry transfer apparatus(Bio-Rad) in 0.5X TBE according to the manufacturer's instructions. The membranes were UV cross-linked using a Stratalinker (Stratagene), hybridized overnight at 42°C in Church buffer with 5'-^{32}P-end-labeled oligonucleotide probes, and washed with 2X SSC, 0.1% SDS and 0.1X SSC, 0.1% SDS for 5 minutes each at 42°C. Results were analyzed using a Phosphorimager (Bio-Rad Personal FX). Oligonucleotide probe sequences are listed in Additional file 5.

Calculating overlap with human tiling analyses

The 3,478 QRNA predictions analyzed by microarray were mapped to the mouse-human UCSC genomic alignments (mm6-hg17) and were subset to the same regions analyzed by Cheng *et al*. [9] (i.e. not repetitive regions, for example), which were determined from the probe positions used in the tiling analysis (coordinates were converted to the hg17 genome release using the UCSC LiftOver tool [17]). The tiling dataset we focused on was generated using nuclear poly-A minus RNA derived from HepG2 cells. For QRNA predictions that overlapped with a transfrag, the degree of overlap was calculated as a percentage of the length of the QRNA prediction that overlaps the transfrag. The distribution of QRNA overlaps was compared to overlaps from randomly positioned QRNA predictions in the human-surveyed regions. The random set consisted of a set of sequences identical in length to the actual QRNA predictions, but with randomized positions in the human surveyed regions. The same analysis was repeated using 390 (151 mapped to human surveyed regions) mouse annotated miRNAs, snoRNAs, snRNAs, and tRNAs downloaded from NONCODE [47] and Rfam [48] databases, and 716 human ribosomal protein exons (260 mapped to human surveyed regions) annotated in the Refseq database [18].

Data availability

All supplementary data is available at http://hugheslab.med.utoronto.ca/Babak/ncRNA/. The microarray design has been submitted to NCBI GEO in MIAME format under accession GSE2366. The 8 novel transcripts have been submitted to GenBank [Genbank:AY954743 – Genbank:AY954751].

Authors' contributions

TB carried out the data compilation, microarray analysis, and data analysis. BB and TH coordinated the study. All authors contributed to preparation of the manuscript.

Additional material

Additional File 1

Summary of array-tested ncRNA predictions.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-6-104-S1.zip]

Additional File 2

List of known mouse ncRNAs represented on array.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-6-104-S1.zip]

Additional File 3

Microarray probe sequences.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-6-104-S3.zip]

Additional File 4

Whole-blot northern data.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-6-104-S4.pdf]

Additional File 5

List of northern probe sequences.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-6-104-S5.pdf]

Acknowledgements

We would like to thank Janet Rossant, Wen Zhang, and Eric Sat for sharing mouse tissues and dissection expertise. We would also like to thank Susan

McCracken for sharing HeLa nuclear extract. This work was supported by grants to T.R.H. and B.J.B. from CIHR and CFI. T.B. was supported by an NSERC graduate scholarship.

References

1. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420(6915):**520-562.
2. Chiaromonte F, Weber RJ, Roskin KM, Diekhans M, Kent WJ, Haussler D: **The share of human genomic DNA under selection estimated from human-mouse genomic alignments.** *Cold Spring Harb Symp Quant Biol* 2003, **68:**245-254.
3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409(6822):**860-921.
4. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, Tammana H, Gingeras TR: **Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22.** *Genome Res* 2004, **14(3):**331-342.
5. Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR: **Large-scale transcriptional activity in chromosomes 21 and 22.** *Science* 2002, **296(5569):**916-919.
6. Numata K, Kanai A, Saito R, Kondo S, Adachi J, Wilming LG, Hume DA, Hayashizaki Y, Tomita M: **Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection.** *Genome Res* 2003, **13(6B):**1301-1306.
7. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, Yamanaka I, Kiyosawa H, Yagi K, Tomaru Y, Hasegawa Y, Nogami A, Schonbach C, Gojobori T, Baldarelli R, Hill DP, Bult C, Hume DA, Quackenbush J, Schriml LM, Kanapin A, Matsuda H, Batalov S, Beisel KW, Blake JA, Bradt D, Brusic V, Chothia C, Corbani LE, Cousins S, Dalla E, Dragani TA, Fletcher CF, Forrest A, Frazer KS, Gaasterland T, Gariboldi M, Gissi C, Godzik A, Gough J, Grimmond S, Gustincich S, Hirokawa N, Jackson IJ, Jarvis ED, Kanai A, Kawaji H, Kawasawa Y, Kedzierski RM, King BL, Konagaya A, Kurochkin IV, Lee Y, Lenhard B, Lyons PA, Maglott DR, Maltais L, Marchionni L, McKenzie L, Miki H, Nagashima T, Numata K, Okido T, Pavan WJ, Pertea G, Pesole G, Petrovsky N, Pillai R, Pontius JU, Qi D, Ramachandran S, Ravasi T, Reed JC, Reed DJ, Reid J, Ring BZ, Ringwald M, Sandelin A, Schneider C, Semple CA, Setou M, Shimada K, Sultana R, Takenaka Y, Taylor MS, Teasdale RD, Tomita M, Verardo R, Wagner L, Wahlestedt C, Wang Y, Watanabe Y, Wells C, Wilming LG, Wynshaw-Boris A, Yanagisawa M, Yang I, Yang L, Yuan Z, Zavolan M, Zhu Y, Zimmer A, Carninci P, Hayatsu N, Hirozane-Kishikawa T, Konno H, Nakamura M, Sakazume N, Sato K, Shiraki T, Waki K, Kawai J, Aizawa K, Arakawa T, Fukuda S, Hara A, Hashizume W, Imotani K, Ishii Y, Itoh M, Kagawa I, Miyazaki A, Sakai K, Sasaki D, Shibata K, Shinagawa A, Yasunishi A, Yoshino M, Waterston R, Lander ES, Rogers J, Birney E, Hayashizaki Y: **Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs.** *Nature* 2002, **420(6915):**563-573.
8. Shabalina SA, Spiridonov NA: **The mammalian transcriptome and the function of non-coding DNA sequences.** *Genome Biol* 2004, **5(4):**105.
9. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, Sementchenko V, Piccolboni A, Bekiranov S, Bailey DK, Ganesh M, Ghosh S, Bell I, Gerhard DS, Gingeras TR: **Transcriptional Maps of 10 Human Chromosomes at 5-Nucleotide Resolution.** *Science* 2005, **308(5725):**1149-1154.
10. Johnson JM, Edwards S, Shoemaker D, Schadt EE: **Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments.** *Trends Genet* 2005, **21(2):**93-102.
11. Wang J, Zhang J, Zheng H, Li J, Liu D, Li H, Samudrala R, Yu J, Wong GK: **Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs.** *Nature* 2004, **431(7010):**1 p following 757; discussion following 757.
12. Bono H, Yagi K, Kasukawa T, Nikaido I, Tominaga N, Miki R, Mizuno Y, Tomaru Y, Goto H, Nitanda H, Shimizu D, Makino H, Morita T, Fujiyama J, Sakai T, Shimoji T, Hume DA, Hayashizaki Y, Okazaki Y: **Systematic expression profiling of the mouse transcriptome**

using RIKEN cDNA microarrays. *Genome Res* 2003, **13(6B):**1318-1323.
13. Wyers F, Rougemaille M, Badis G, Rousselle JC, Dufour ME, Boulay J, Regnault B, Devaux F, Namane A, Seraphin B, Libri D, Jacquier A: **Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase.** *Cell* 2005, **121(5):**725-737.
14. Rivas E, Eddy SR: **Noncoding RNA gene detection using comparative sequence analysis.** *BMC Bioinformatics* 2001, **2(1):**8.
15. Rivas E, Klein RJ, Jones TA, Eddy SR: **Computational identification of noncoding RNAs in E. coli by comparative genomics.** *Curr Biol* 2001, **11(17):**1369-1373.
16. McCutcheon JP, Eddy SR: **Computational identification of noncoding RNAs in Saccharomyces cerevisiae by comparative genomics.** *Nucleic Acids Res* 2003, **31(14):**4119-4128.
17. Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, Kobayashi S, Davis C, Dai H, He YD, Stephaniants SB, Cavet G, Walker WL, West A, Coffey E, Shoemaker DD, Stoughton R, Blanchard AP, Friend SH, Linsley PS: **Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer.** *Nat Biotechnol* 2001, **19(4):**342-347.
18. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ: **The UCSC Genome Browser Database.** *Nucleic Acids Res* 2003, **31(1):**51-54.
19. NCBI: **ftp://ftp.ncbi.nih.gov/genomes/M_musculus/RNA/.** .
20. NCBI: **ftp://ftp.ncbi.nih.gov/genomes/M_musculus/protein/.** .
21. NCBI: **ftp://ftp.ncbi.nih.gov/genbank/.** .
22. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res* 1997, **25(5):**955-964.
23. Lowe TM, Eddy SR: **A computational screen for methylation guide snoRNAs in yeast.** *Science* 1999, **283(5405):**1168-1171.
24. Babak T, Zhang W, Morris Q, Blencowe BJ, Hughes TR: **Probing microRNAs with microarrays: tissue specificity and functional inference.** *Rna* 2004, **10(11):**1813-1819.
25. Cavaille J, Vitali P, Basyuk E, Huttenhofer A, Bachellerie JP: **A novel brain-specific box C/D small nucleolar RNA processed from tandemly repeated introns of a noncoding RNA gene in rats.** *J Biol Chem* 2001, **276(28):**26374-26383.
26. Inoue A, Kobayashi Y, Ishizuka M, Hirose S, Hagiwara H: **Identification of a novel osteoblastic gene, inducible by C-type natriuretic peptide, whose transcript might function in mineralization as a noncoding RNA.** *Calcif Tissue Int* 2002, **70(2):**111-116.
27. Lagos-Quintana M, Rauhut R, Yalcin A, Meyer J, Lendeckel W, Tuschl T: **Identification of tissue-specific microRNAs from mouse.** *Curr Biol* 2002, **12(9):**735-739.
28. Takeda K, Ichijo H, Fujii M, Mochida Y, Saitoh M, Nishitoh H, Sampath TK, Miyazono K: **Identification of a novel bone morphogenetic protein-responsive gene that may function as a noncoding RNA.** *J Biol Chem* 1998, **273(27):**17079-17085.
29. Brannan CI, Dees EC, Ingram RS, Tilghman SM: **The product of the H19 gene may function as an RNA.** *Mol Cell Biol* 1990, **10(1):**28-36.
30. Hiley SL, Babak T, Hughes TR: **Global analysis of yeast RNA processing identifies new targets of RNase III and uncovers a link between tRNA 5' end processing and tRNA splicing.** *Nucleic Acids Res* 2005, **33(9):**3048-3056.
31. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci U S A* 2004, **101(16):**6062-6067.
32. Szymanski M, Erdmann VA, Barciszewski J: **Noncoding regulatory RNAs database.** *Nucleic Acids Res* 2003, **31(1):**429-431.
33. Eddy SR: **Computational genomics of noncoding RNA genes.** *Cell* 2002, **109(2):**137-140.
34. Nobrega MA, Zhu Y, Plajzer-Frick I, Afzal V, Rubin EM: **Megabase deletions of gene deserts result in viable mice.** *Nature* 2004, **431(7011):**988-993.
35. Eddy SR: **Noncoding RNA genes.** *Curr Opin Genet Dev* 1999, **9(6):**695-699.
36. Sempere LF, Freemantle S, Pitha-Rowe I, Moss E, Dmitrovsky E, Ambros V: **Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation.** *Genome Biol* 2004, **5(3):**R13.
37. Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, Cooper PJ, Swift S, Rastan S: **The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus.** *Cell* 1992, **71(3):**515-526.
38. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M: **Global identification of human transcribed sequences with genome tiling arrays.** *Science* 2004, **306(5705):**2242-2246.
39. Smit AFA, Green P: **http://ftp.genome.washington.edu/RM/RepeatMasker.html.** .
40. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res* 2002, **12(4):**656-664.
41. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E: **EnsMart: a generic system for fast and flexible access to biological data.** *Genome Res* 2004, **14(1):**160-169.
42. NCBI: **ftp://ftp.ncbi.nih.gov/blast/db/.** .
43. NCBI: **ftp://ftp.ncbi.nih.gov/repository/UniGene/.** .
44. Mathews DH, Sabina J, Zuker M, Turner DH: **Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure.** *J Mol Biol* 1999, **288(5):**911-940.
45. Dignam JD, Lebovitz RM, Roeder RG: **Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei.** *Nucleic Acids Res* 1983, **11(5):**1475-1489.
46. Wiegant JC, van Gijlswijk RP, Heetebrij RJ, Bezrookove V, Raap AK, Tanke HJ: **ULS: a versatile method of labeling nucleic acids for FISH based on a monofunctional reaction of cisplatin derivatives with guanine moieties.** *Cytogenet Cell Genet* 1999, **87(1-2):**47-52.
47. Shai O, Morris Q, Frey BJ: **Spatial Bias Removal in Microarray Images.** *University of Toronto Technical Report, http://hugheslabmedutorontoca/Zhang/tech%reportpdf* 2003.
48. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression.** *Bioinformatics* 2002, **18 Suppl 1:**S96-104.
49. Zhang W, Morris QD, Chang R, Shai O, Bakowski MA, Mitsakakis N, Mohammad N, Robinson MD, Zirngibl R, Somogyi E, Laurin N, Eftekharpour E, Sat E, Grigull J, Pan Q, Peng WT, Krogan N, Greenblatt J, Fehlings M, van der Kooy D, Aubin J, Bruneau BG, Rossant J, Blencowe BJ, Frey BJ, Hughes TR: **The functional landscape of mouse gene expression.** *J Biol* 2004, **3(5):**21.