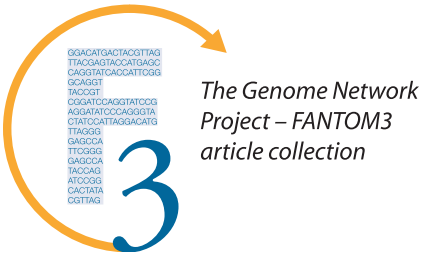
Clusters of Internally Primed Transcripts Reveal Novel Long Noncoding RNAs

Masaaki Furuno[1☯], Ken C. Pang[2,3☯], Noriko Ninomiya[4], Shiro Fukuda[4], Martin C. Frith[2,4], Carol Bult[1], Chikatoshi Kai[4], Jun Kawai[4,5], Piero Carninci[4,5], Yoshihide Hayashizaki[4,5], John S. Mattick[2], Harukazu Suzuki[4*]

1 Mouse Genome Informatics Consortium, The Jackson Laboratory, Bar Harbor, Maine, United States of America, **2** Australian Research Council Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia, **3** T Cell laboratory, Ludwig Institute for Cancer Research, Austin Health, Heidelberg, Victoria, Australia, **4** Genome Exploration Research Group (Genome Network Project Core Group), RIKEN Genomic Sciences Center, RIKEN Yokohama Institute, Yokohama, Japan, **5** Genome Science Laboratory, Discovery Research Institute, RIKEN Wako Institute, Wako, Japan

Non-protein-coding RNAs (ncRNAs) are increasingly being recognized as having important regulatory roles. Although much recent attention has focused on tiny 22- to 25-nucleotide microRNAs, several functional ncRNAs are orders of magnitude larger in size. Examples of such macro ncRNAs include *Xist* and *Air*, which in mouse are 18 and 108 kilobases (Kb), respectively. We surveyed the 102,801 FANTOM3 mouse cDNA clones and found that *Air* and *Xist* were present not as single, full-length transcripts but as a cluster of multiple, shorter cDNAs, which were unspliced, had little coding potential, and were most likely primed from internal adenine-rich regions within longer parental transcripts. We therefore conducted a genome-wide search for regional clusters of such cDNAs to find novel macro ncRNA candidates. Sixty-six regions were identified, each of which mapped outside known protein-coding loci and which had a mean length of 92 Kb. We detected several known long ncRNAs within these regions, supporting the basic rationale of our approach. In silico analysis showed that many regions had evidence of imprinting and/or antisense transcription. These regions were significantly associated with microRNAs and transcripts from the central nervous system. We selected eight novel regions for experimental validation by northern blot and RT-PCR and found that the majority represent previously unrecognized noncoding transcripts that are at least 10 Kb in size and predominantly localized in the nucleus. Taken together, the data not only identify multiple new ncRNAs but also suggest the existence of many more macro ncRNAs like *Xist* and *Air*.

Citation: Furuno M, Pang KC, Ninomiya N, Fukuda S, Frith MC, et al. (2006) Clusters of internally primed transcripts reveal novel long noncoding RNAs. PLoS Genet 2(4): e37. DOI: 10.1371/journal.pgen.0020037



The Genome Network Project – FANTOM3 article collection

Introduction

The existence of non-protein-coding RNAs (ncRNAs) has been known for many decades, and the importance of essential infrastructural ncRNAs such as ribosomal RNAs and transfer RNAs in facilitating protein synthesis has long been recognized. Recently, other ncRNAs have generated intense interest based upon their ability to regulate gene expression. Foremost among these are microRNAs (miRNAs), which are about 22 nucleotides in length and function by targeting mRNAs for cleavage or translational repression. Hundreds of miRNAs have been identified in animals, plants, and viruses, and they mediate critical regulatory functions in a range of developmental and physiological pathways [1–3]. Another prominent class of ncRNAs is the short interfering RNAs (siRNAs), which were discovered as a tool for knocking down gene expression in the lab but have subsequently been found to act as natural endogenous regulators of gene expression [1].

Given the considerable attention that these tiny ncRNAs have attracted, it would be understandable to think that regulatory ncRNAs are short. However, a small number of functional ncRNAs have also been identified that are orders of magnitude larger in size than miRNAs and siRNAs. Well-known examples of such macro ncRNAs include *Xist* and *Air*, which in mouse are approximately 18 and 108 Kb, respectively [4,5]. *Xist* plays an essential role in mammals by associating with chromatin and causing widespread gene silencing on the inactive X chromosome [6], while *Air* is

Editors: Judith Blake (The Jackson Laboratory, US), John Hancock (MRC-Harwell, UK), Bill Pavan (NHGRI-NIH, US), and Lisa Stubbs (Lawrence Livermore National Laboratory, US), together with *PLoS Genetics* EIC Wayne Frankel (The Jackson Laboratory, US)

Received August 16, 2005; **Accepted** February 1, 2006; **Published** April 28, 2006

DOI: 10.1371/journal.pgen.0020037

Abbreviations: CNS, central nervous system; ENOR, expressed noncoding region; EST, expressed sequence tag; GEV, FANTOM3 Genomic Element Viewer; GNF, Genomics Institute of the Novartis Research Foundation; Kb, kilobase(s); miRNA, microRNA; ncRNA, non-protein-coding RNA; qRT-PCR, quantitative real-time RT-PCR; siRNA, short interfering RNA; snoRNA, small nucleolar RNA; TU, transcriptional unit; UCSC, University of California Santa Cruz; UNA, unspliced, noncoding, and containing adjunct adenine-rich regions

* To whom correspondence should be addressed. E-mail: rgscerg@gsc.riken.jp

☯ These authors contributed equally to this work.

Synopsis

The human genome has been sequenced, and, intriguingly, less than 2% specifies the information for the basic protein building blocks of our bodies. So, what does the other 98% do? It now appears that the mammalian genome also specifies the instructions for many previously undiscovered "non protein-coding RNA" (ncRNA) genes. However, what these ncRNAs do is largely unknown. In recent years, strategies have been designed that have successfully identified hundreds of short ncRNAs—termed microRNAs—many of which have since been shown to act as genetic regulators. Also known to be functionally important are a handful of ncRNAs orders of magnitude larger in size than microRNAs. The availability of complete genome and comprehensive transcript sequences allows for the systematic discovery of more large ncRNAs. The authors developed a computational strategy to screen the mouse genome and identify large ncRNAs. They detected existing large ncRNAs, thus validating their approach, but, more importantly, discovered more than 60 other candidates, some of which were subsequently confirmed experimentally. This work opens the door to a virtually unexplored world of large ncRNAs and beckons future experimental work to define the cellular functions of these molecules.

required for paternal silencing of the *Igf2r/Slc22a2/Slc22a3* gene cluster [5]. Apart from their extreme length, *Xist* and *Air* share two other important features: genomic imprinting and antisense transcription. Genomic imprinting is a process by which certain genes are expressed differently according to whether they have been inherited from the maternal or paternal allele. Imprinting is critical for normal development, and loss of imprinting has been implicated in a variety of human diseases [7]. ncRNAs have been discovered at many different imprinted loci and appear to be important in the imprinting process itself [5,8]. The other feature that *Xist* and *Air* have in common is that both are members of naturally occurring *cis*-antisense transcript pairs. Previous studies have indicated the existence of thousands of mammalian *cis*-antisense transcripts [9–12]. These transcripts may regulate gene expression in a variety of ways including RNA interference, translational regulation, RNA editing, alternative splicing, and alternative polyadenylation [13,14], although the exact mechanisms by which antisense RNAs function are unknown.

In addition to well-documented ncRNAs, recent evidence from both high-density tiling arrays [15,16] and large-scale analyses of full-length enriched cDNA libraries [17] suggests that there may be thousands more ncRNAs within the mammalian transcriptome. Many of these candidates have emerged from the RIKEN Mouse Gene Encyclopedia project [17,18], and full-length sequencing and analysis by the FANTOM consortium of 102,801 cDNAs recently revealed that around one-third (34,030) lack an apparent protein-coding region as judged by manual annotation [19]. Although some of these RNAs have been shown to have biological function [20,21], the vast majority of these putative non-coding cDNAs remain of uncertain significance, especially given that many are likely to represent internally primed transcription artifacts (which arise during first-strand cDNA synthesis when oligo[dT] primers bind not to genuine polyA tails but rather to internal adenine-rich regions within longer transcripts) and are not true, full-length transcripts [22,23].

In surveying the FANTOM3 mouse cDNAs, we observed that macro ncRNAs such as *Air* and *Xist* were present not as single, full-length transcripts but rather as fragmented clusters of cDNAs, most of which were not only internally primed but also unspliced and of minimal protein-coding potential. We hypothesized that we might discover novel macro ncRNAs by conducting a genome-wide search for similar clusters of cDNAs. We subsequently identified 66 candidate ncRNA regions. A few of these overlap with known long ncRNAs, and many contain imprinted cDNA candidates, *cis*-antisense transcripts, or miRNAs. Eight regions were characterized experimentally, and the majority were found to represent previously unknown long ncRNAs that are localized to the nucleus. Taken together, the data suggest the existence of many more macro ncRNAs that, like *Xist* and *Air*, may fulfill important regulatory roles in mammalian biology.

Results

Xist and *Air* Are Represented by Clusters of Truncated Noncoding cDNAs

As part of the FANTOM3 project, we looked for the existence of known ncRNAs among the 102,801 cDNAs. We found that 16 of 43 (39%) non-small-nucleolar, non-micro reference mouse ncRNAs that are present in RNAdb, a database of mammalian ncRNAs [24], were detectable among the RIKEN cDNA collection, as judged by similarity using BLASTN (Table 1). The two longest ncRNAs detected were *Xist* and *Air*. Very long transcripts such as these create substantial difficulties for cDNA cloning protocols for a variety of well-established technical reasons [23,25]. We were therefore not surprised that examination of both loci via the FANTOM3 Genomic Element Viewer (GEV) (http://fantom32p.gsc.riken.jp/gev-f3/gbrowse/mm5) revealed that *Xist* and *Air* were represented by a cluster of truncated RIKEN and non-RIKEN cDNAs interspersed along the length of their parent transcripts. Inspection of the individual cDNAs demonstrated that the majority were unspliced, held minimal protein-coding potential, and had adjunct genomic adenine-rich regions immediately downstream of their 3′ ends, suggesting that they had been internally primed. Figure 1A illustrates transcription within the *Air/Igf2r* locus. *Air* is represented by 20 individual cDNAs dispersed along its reported length, of which 14 are unspliced, noncoding RIKEN cDNAs that contain an adjunct adenine-rich region. Figure 1B shows *Xist* and its antisense partner *Tsix*. Here, nine cDNAs are seen along the length of the spliced *Xist* transcript, of which four are unspliced, noncoding RIKEN cDNAs that contain an adjunct adenine-rich region.

Genome-Wide Search Reveals Multiple Clusters of Unspliced, Internally Primed Noncoding Transcripts Lying Outside Protein-Coding Loci

Based upon these observations (Table 1; Figure 1), we reasoned that it might be possible to discover novel macro ncRNAs via a genome-wide search for clusters of transcripts that were unspliced, noncoding, and contained adjunct adenine-rich regions (UNA transcripts) (Figure 2). To begin, we classified transcriptional units (TUs) into protein-coding and noncoding using the manual annotations of FANTOM3 collaborators [19], where a TU is defined as a group of transcripts that share at least one exonic nucleotide overlap and that map to the same chromosomal strand [19]. Of 37,348

Table 1. Detection of Known Mouse ncRNAs within the FANTOM3 cDNA Collection

Reference ncRNA			FANTOM3 cDNA			BLASTN Hit		
Name	**Length[a]**	**Spliced**	**Clone ID**	**Adjunct Adenine-Rich Region Present**	**Spliced**	**Length**	**Percent cDNA Length**	**Percent ncRNA Length**
Air	107,796	Unspliced	2810051F02	No	Yes	1,064	95%	1%
			6430704I19	Yes	No	2,849	100%	3%
			6720429I20	Yes	No	3,699	100%	3%
			9530009E11	Yes	No	2,160	100%	2%
			A330053J22	Yes	No	1,266	100%	1%
			A530029P07	Yes	No	1,893	100%	2%
			A530040I05	Yes	No	1,899	100%	2%
			B930018I07	Yes	No	2,054	100%	2%
			C130030E20	Yes	No	1,401	100%	1%
			C130073E08	Yes	No	1,050	100%	1%
			D130094O12	Yes	No	682	100%	1%
			D930036N23	Yes	No	2,651	99%	2%
			E130107J18	Yes	No	2,205	100%	2%
			G130203I22	Yes	No	1,644	100%	2%
Cior	2,135	Spliced	2310040O21	No	Yes	1,088	99%	51%
			7120406M20	No	Yes	2,112	99%	99%
			7120489O22	No	Yes	2,111	99%	99%
			9130011J15	No	Yes	2,112	98%	99%
			9630004F23	No	Yes	2,112	74%	99%
			E430002L08	No	Yes	2,111	99%	99%
			I830031I11	No	Yes	2,110	100%	99%
			I830072L22	No	Yes	2,110	100%	99%
			I830083L02	No	Yes	2,110	100%	99%
Ftx	676	Spliced	9530061G23	No	Yes	637	15%	98%
			B230206F22	Yes	Yes	676	100%	100%
			D430040K02	No	Yes	615	27%	100%
Gtl2	3,199	Spliced	2900058E08	Yes	No	1,229	100%	39%
			6330403F08	Yes	Yes	3,199	100%	100%
			B230342G15	No	No	418	100%	13%
H19	1,899	Spliced	1100001A04	No	Yes	868	99%	46%
			I0C0030C13	No	Yes	1,842	82%	97%
Ipw	155	Spliced	B230105C16	No	Yes	155	4%	100%
Jpx/Enox	259	Spliced	2510040I06	No	Yes	148	22%	93%
			9830107K21	No	Yes	259	7%	100%
			G370019D15	No	Yes	259	8%	100%
Kcnq1ot1/LIT1/Kvlqt1-AS	4,729	Unspliced	C130002M05	Yes	No	2,467	100%	52%
mirg	1,297	Spliced	2810474H01	No	Yes	592	100%	46%
Mit1/Lb9	1,879	Unspliced	3110055B08	No	No	240	97%	13%
Nespas	3,806	Unspliced and spliced	D030028H20	Yes	No	3,806	100%	100%
		D330038P10	No	No	629	100%	100%	
Peg13	4,419	Unspliced	F630009C06	No	No	3,334	100%	75%
			F930102O09	No	No	3,277	100%	74%
telomerase RNA	397	Unspliced	D430035J07	Yes	No	397	14%	100%
U17HG	383	Spliced	3830421G02	No	Yes	383	99%	100%
UHG	591	Spliced	A730062M15	No	Yes	476	100%	81%
Xist	17,919	Spliced	0610031I13	Yes	No	518	100%	3%
			2610022A11	Yes	No	890	100%	5%
			A430022B11	Yes	No	1,495	100%	12%
			D030072M03	Yes	No	3,420	100%	19%

FANTOM3 cDNAs were tested for similarity to a reference set of non-sno, non-micro ncRNAs from RNAdb using BLASTN (for details, see Materials and Methods).
[a]Where multiple sequences exist for the reference ncRNA (e.g., different splice variants), the indicated length is that of the longest sequence.
DOI: 10.1371/journal.pgen.0020037.t001

TUs, 20,708 were classified as noncoding TUs. We knew, however, from previous work that noncoding TUs often overlap with protein-coding genes, since they can be internally primed off long pre-mRNAs [22]. Figure 1C shows an example of this, where a cluster of five UNA cDNAs overlap with intronic regions of the large dystrophin *(Dmd)* transcript. Of 20,708 noncoding TUs, we excluded 8,228 located within intronic regions of protein-coding TUs. We then selected UNA TUs based on the following criteria: (1) an adjunct adenine-rich region was present at the TU end, (2) no major polyA signal (AATAAA/ATTAAA) was present within 100 nucleotides of the TU end, and (3) the TU was unspliced. Of 12,480 noncoding TUs, 2,699 satisfied the criteria. We then clustered these 2,699 UNA TUs by merging any two or more located within 100 Kb of one another, provided that (1) there were no intervening protein-coding transcripts or gene

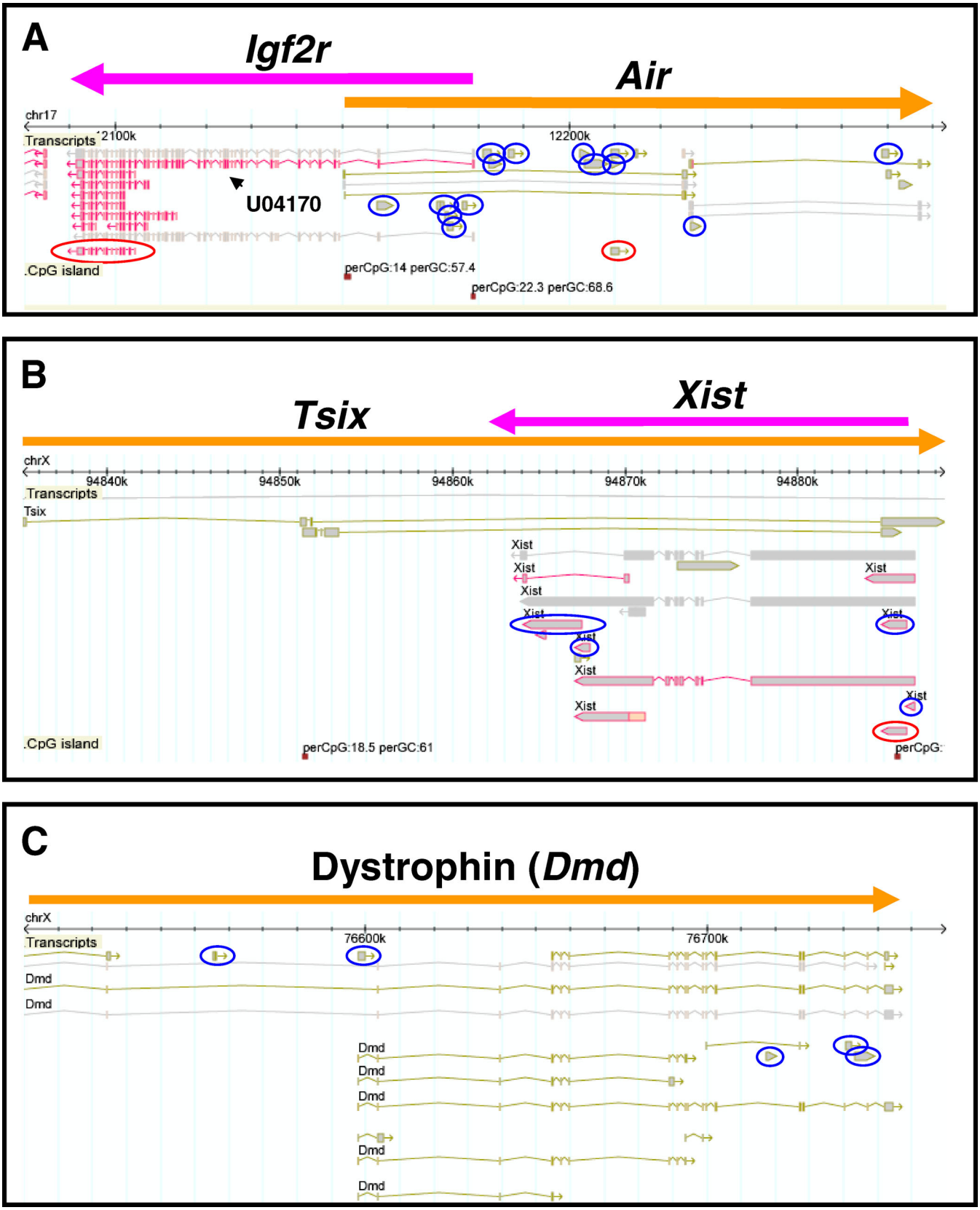


Figure 1. Snapshots of the GEV Showing Transcription

(A) The *Air*/*Igf2r* locus (Chromosome 17: 12,091,531–12,258,195).
(B) The *Xist*/*Tsix* locus (X chromosome: 94,835,096–94,888,536).
(C) The dystrophin *(Dmd)* locus (X chromosome: 76,500,000–76,754,601).
For the transcripts, cDNA sequences from the RIKEN and public databases are shown, and are colored in brown and purple depending upon their

chromosomal strand of origin. Predicted genes from Ensembl, NCBI, and RefSeq databases are shown in gray. CpG islands as defined by the UCSC Genome Browser are shown. Blue circles indicate unspliced, noncoding RIKEN cDNAs with adjunct adenine-rich regions. Red circles indicate RIKEN imprinted cDNA candidates [38].
DOI: 10.1371/journal.pgen.0020037.g001

predictions (based on either FANTOM3 annotations, NCBI RefSeq sequences, or Ensembl gene models) and (2) there were no intervening transcripts with major polyA signals and without adjunct adenine-rich regions (which would indicate likely transcript termination sites). Using this approach, we identified 191 genomic regions, containing 528 clustered UNA TUs. To increase the likelihood that these regions represented genuine long transcripts, we excluded any that were less than 10 Kb long or contained less than ten expressed sequence tags (ESTs). This left 86 regions, which were then manually inspected using the GEV to look for possible internal transcriptional start sites (e.g., CpG islands, CAGE tags, or multiple ESTs arising from the same position) or transcripts encoding small proteins not already filtered out in the discovery pipeline. Following this, 66 regions remained (Table 2). We named these long expressed noncoding regions (ENORs).

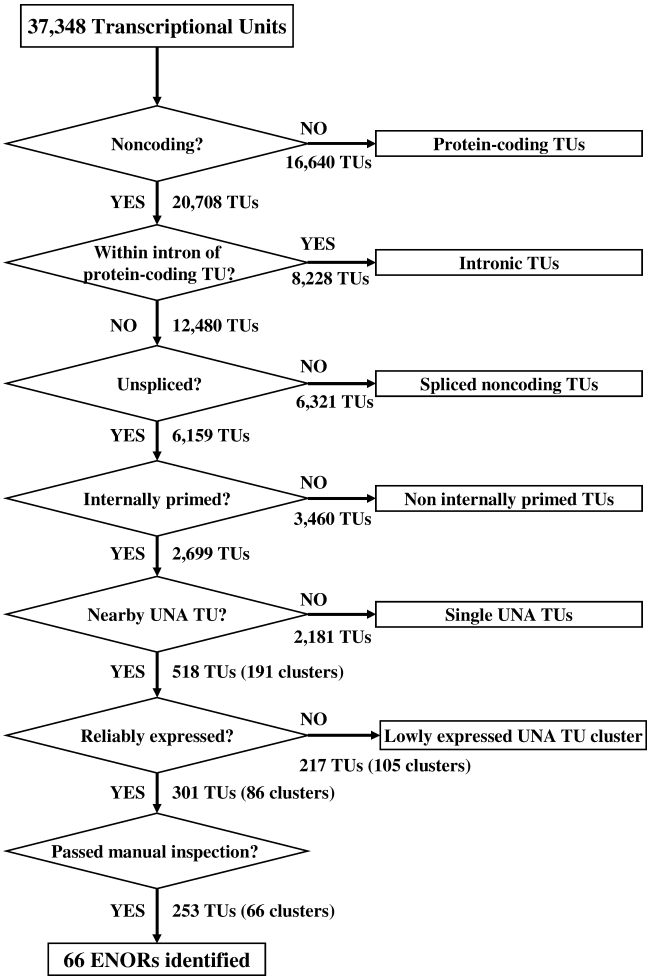


Figure 2. Discovery Pipeline for ENORs

FANTOM and public transcripts were clustered into 37,348 TUs by grouping any two or more transcripts that shared genomic coordinates. Then, the following procedures were applied. (1) Protein-coding TUs were excluded by removing any whose transcripts had an open reading frame of either 150 amino acids or more (RIKEN/MGC cDNAs) or one amino acid or more (non-RIKEN/MGC cDNAs). (2) TUs wholly encompassed within introns of protein-coding TUs were excluded to avoid possible pre-mRNA intronic transcripts. (3) Intron-containing TUs were excluded to select for unspliced transcripts. (4) TUs lacking adjunct adenine-rich regions or containing polyA signals were excluded to select for internally primed transcripts. (5) Remaining UNA TUs that mapped within 100 Kb of one another on the mouse genome (mm5) were clustered together, provided they did not overlap the genomic coordinates of a protein-coding TU/NCBI RefSeq/Ensembl gene model with a CDS of 150 amino acids or more or a noncoding TU with a polyA signal within 100 bp of the 3′ end and without an adjunct adenine-rich region. (6) Reliably expressed UNA TU clusters were selected by identifying those with at least ten supporting ESTs. (7) Selected UNA TU clusters were then manually screened and separated based upon evidence of possible internal transcription state sites (based upon CpG islands, CAGE tags, and EST clusters), resulting in the identification of 66 ENORs.
DOI: 10.1371/journal.pgen.0020037.g002

ENORs Successfully Identify Several Known Long ncRNAs

To assess the validity of our approach, we examined whether known mouse macro ncRNAs were detected among the 66 ENORs. Notably, the cluster we had manually identified as corresponding to *Xist* was not detected. This was because one of the original *Xist* transcripts (GenBank accession number X59289) remains annotated as a hypothetical protein of 299 amino acids based upon an earlier presumption that it was translated [26]; consequently, this cluster of cDNAs was automatically classified as being protein-coding and thus rejected. We did, however, succeed in identifying *Air* (ENOR60) and several other long ncRNAs. These included the following: *Kcnq1ot1* (ENOR24), an imprinted antisense transcript of ~54 Kb [27]; *Rian* (ENOR44), a spliced 5.4-Kb imprinted transcript that spans more than 10 Kb of mouse genome and acts as a host gene for multiple small nucleolar RNAs (snoRNAs) [28,29]; and *Ube3a-ats* (ENOR22), an imprinted, ~1,000-Kb antisense transcript that is brain-specific and hosts numerous snoRNAs [30]. Additionally, we detected *Dleu2* (ENOR49), an alternatively spliced antisense ncRNA of ~1.4 Kb that spans more than 80 Kb and is a host gene for miRNAs [31,32]. Apart from *Xist,* the only other ncRNA in the RNAdb reference set longer than 5 Kb that was not detected was *Emx2os,* a 5.04-Kb antisense transcript that spans ~35 Kb [33]. Inspection of this locus showed that it contained only one UNA cDNA. Taken together, these observations indicated that our approach was able to successfully detect existing long ncRNAs, although it missed some either because of annotation errors or because the number of UNA transcripts fell below our discovery pipeline threshold. Our approach also appeared to detect shorter ncRNAs such as *Dleu2* that were spliced and spanned a long genomic region.

In Silico Characterization of ENOR Regions

Next, we sought to characterize the 66 ENORs in greater detail (Table 2). The maximum number of UNA TUs per region was 12 (ENOR59), and the average was 3.8 per 100 Kb. The region length ranged from 11 to 458 Kb, with a mean of 92 Kb. The number, length, and distribution of the ENORs across each chromosome are shown in Table S1 and Figure S1. Chromosome 8 had the highest number of ENORs (nine), with a total length of 860 Kb. Chromosome 16 had the greatest length (1,089 Kb), as represented by three ENORs.

Table 2. Bioinformatic Characterization of 66 ENORs

Region ID	Chromosomal Location[a]	Strand	Length (Kb)	Number of UNA TUs	Average TUs/100 Kb	Splicing Status	Known Sense Transcripts[b]	Antisense Transcripts (Protein-Coding)[c]	Antisense Transcripts (Noncoding)[c]	snoRNAs	miRNAs	Candidate Imprinted RIKEN cDNAs[d]	Human–Mouse Synteny[e]
ENOR1	chr1:63518734..63541420	+	23	2	8.8	Spliced	*5830412M15Rik*	*Ndufs1* (NADH dehydrogenase [ubiquinone] Fe-S protein 1)					No
ENOR2	chr1:137790614..137815760	+	25	4	15.9	Unspliced							Yes
ENOR3	chr1:137824570..137885116	+	61	6	9.9	Unspliced			*A430106G13Rik*		*mir-213*, *mir-181b-1*		Yes
ENOR4	chr1:191266590..191282191	+	16	2	12.8	Unspliced		*Ppp2r5a* (protein phosphatase 2, regulatory subunit B [B56], alpha isoform)					No
ENOR5	chr1:194886616..194938031	+	51	3	5.8	Spliced	*A330023F24Rik*	*Mcp* (membrane cofactor protein)			*mir-29b-2*, *mir-29c*	A230025O03(M), 9630038K03(P)	Yes
ENOR6	chr2:38789288..38825524	+	36	2	5.5	Unspliced		*Nr6a1* (an orphan nuclear receptor)			*mir-181a*, *mir-181b-2*		No
ENOR7	chr2:50295592..50375687	+	80	2	2.5	Spliced	B230312L02 cDNA						Yes
ENOR8	chr2:52826614..52936521	+	110	5	4.5	Unspliced	*D2Ertd295e*						Yes
ENOR9	chr2:67592786..67711002	+	118	6	5.1	Unspliced			9230005N23 cDNA				Yes
ENOR10	chr2:169143755..169218835	−	75	3	4.0	Unspliced			7530418G04 cDNA				No
ENOR11	chr3:73508279..73534688	+	26	2	7.6	Spliced	A930012C24 cDNA						No
ENOR12	chr4:47200812..47280798	−	80	3	3.8	Unspliced			4932422O18 cDNA				Yes
ENOR13	chr4:140332009..140368872	+	37	2	5.4	Spliced	B230369C19 cDNA		4932443I23 EST				Yes
ENOR14	chr4:140806332..140879255	−	73	3	4.1	Unspliced							No
ENOR15	chr5:28332435..28399702	−	67	2	3.0	Spliced	*A230098N10Rik*		F730119A01 cDNA				No
ENOR16	chr5:50478870..50535337	−	56	4	7.1	Unspliced						A330078N10(M)	Yes
ENOR17	chr6:13663958..13680754	+	17	2	11.9	Spliced	*1110019D14Rik*						No
ENOR18	chr6:15599315..15702709	−	103	2	1.9	Unspliced							Yes
ENOR19	chr6:30956520..31024430	−	68	3	4.4	Unspliced							Yes
ENOR20	chr6:52961318..52995710	−	34	2	5.8	Unspliced							Yes
ENOR21	chr6:62056407..62204002	+	148	5	3.4	Unspliced							No
ENOR22	chr7:46808311..46937814	−	130	6	4.6	Unspliced	*Ube3a-ats* (ubiquitin protein ligase E3A antisense)	*Ube3a* (ubiquitin protein ligase E3A)				A130033P09(M), B230341L10(M), 9330162G02(M), 9330156G04(M), A230073K19(M)	No
ENOR23	chr7:48744563..49025144	−	281	8	2.9	Spliced	*A330076H08Rik*, *A230057D06Rik*				*mir-344*	9630054K16(M)	No
ENOR24	chr7:130917883..131058684	−	141	6	4.3	Unspliced	*Kcnq1ot1* (Kvlqt1-as) (KCNQ1 overlapping transcript 1)	*Kcnq1* (potassium voltage-gated channel, subfamily Q, member 1)				A330049H05(M)	Yes
ENOR25	chr8:31090538..31346698	−	256	6	2.3	Unspliced			E130119F02 cDNA				Yes
ENOR26	chr8:35718163..35764584	−	46	4	8.6	Unspliced							Yes
ENOR27	chr8:49545737..49570096	−	24	3	12.3	Spliced	*1700019L22Rik*		7420404P19 cDNA				Yes
ENOR28	chr8:49895333..50079904	−	185	8	4.3	Unspliced							Yes
ENOR29	chr8:56389750..56402708	−	13	3	23.1	Unspliced							Yes
ENOR30	chr8:56424559..56484023	−	59	3	5.0	Unspliced			A930038F16 EST				Yes

Table 2. Continued

Region ID	Chromosomal Location[a]	Strand	Length (Kb)	Number of UNA TUs	Average TUs/100 Kb	Splicing Status	Known Sense Transcripts[b]	Antisense Transcripts (Protein-Coding)[c]	Antisense Transcripts (Noncoding)[c]	snoRNAs	miRNAs	Candidate Imprinted RIKEN cDNAs[d]	Human–Mouse Synteny[e]
ENOR31	chr8:56996422..57094390	−	98	5	5.1	Unspliced						A330054M17(M)	No
ENOR32	chr8:108495997..108571336	+	75	2	2.7	Spliced	*D030068K23Rik*		*4922502B01Rik*				Yes
ENOR33	chr8:114854210..114956541	−	102	3	2.9	Unspliced							Yes
ENOR34	chr9:83775886..83816085	+	40	2	5.0	Unspliced							No
ENOR35	chr9:91160523..91183087	−	23	3	13.3	Unspliced							No
ENOR36	chr9:101010585..101044467	+	34	3	8.9	Unspliced		*3222402P14Rik*					No
ENOR37	chr10:39503501..39520182	−	17	2	12.0	Spliced	AI426748	*Traf3ip2* (Traf3 interacting protein 2)					Yes
ENOR38	chr10:91880273..91926936	−	47	3	6.4	Spliced	*Dmt2* (dorso-medial telencephalon gene 2)						Yes
ENOR39	chr10:109570575..109896599	−	326	9	2.8	Spliced	*9630020C08Rik*						Yes
ENOR40	chr11:18646895..18766069	−	119	3	2.5	Spliced	*8430419K02Rik*						No
ENOR41	chr11:112711495..112810279	−	99	3	3.0	Spliced	*2610035D17Rik*						Yes
ENOR42	chr12:23903332..23925572	+	22	2	9.0	Spliced	5330419A06 cDNA						No
ENOR43	chr12:51691366..51717285	−	26	4	15.4	Spliced	*E030019B13Rik*						Yes
ENOR44	chr12:104343571..104425759	+	82	5	6.1	Spliced	*Rian* (RNA imprinted and accumulated in nucleus)			*MBII-48*, *MBII-49*, *MBII-343*	*mir-341*, *mir-370*	A230052J08(P)	Yes
ENOR45	chr13:27825037..28246820	−	422	11	2.6	Spliced	*2610307P16Rik*						Yes
ENOR46	chr13:28348660..28416295	+	68	2	3.0	Spliced	*A330102I10Rik*						No
ENOR47	chr13:35485415..35551102	−	66	3	4.6	Spliced	BC034664 cDNA						Yes
ENOR48	chr14:41377672..41452118	+	74	2	2.7	Spliced	A730014E05 cDNA		A930015C18 cDNA				No
ENOR49	chr14:53585943..53664074	−	78	2	2.6	Spliced	*Dleu2* (deleted in lymphocytic leukemia 2)	*Trim13* (tripartite motif protein 13), *Kcnrg* (potassium channel regulator)			*mir-16–1*, *mir-15a*		Yes
ENOR50	chr14:101067461..101165669	−	98	4	4.1	Unspliced						A230020C19(M)	No
ENOR51	chr15:8302259..8356966	+	55	2	3.7	Unspliced		*4933421G18Rik*					Yes
ENOR52	chr15:12030064..12041853	+	12	2	17.0	Unspliced						9330115C17(M)	No
ENOR53	chr15:20602048..20622583	+	21	2	9.7	Spliced	D130080K17 cDNA						Yes
ENOR54	chr15:68611538..68625928	−	14	2	13.9	Unspliced							Yes
ENOR55	chr15:82695945..82728384	+	32	3	9.2	Unspliced		*Cyp2d22* (cytochrome P450, family 2, subfamily d, polypeptide 22)					No
ENOR56	chr15:96691893..96703025	−	11	2	18.0	Spliced	*2610037D02Rik*						No
ENOR57	chr16:72376998..72835471	+	458	10	2.2	Unspliced						B230338E12(M), B230315D22(M)	Yes
ENOR58	chr16:75226429..75551463	−	325	7	2.2	Unspliced						A730076E23(M)	Yes
ENOR59	chr16:77741108..78046567	+	305	12	3.9	Spliced	*2810055G20Rik*				*mir-99a*, *let-7c-1*, *mir-125b-2*		Yes
ENOR60	chr17:12172664..12275448	+	103	7	6.8	Spliced	*Air* (antisense Igf2r RNA), *D17Ertd663e*	*Igf2r* (insulin-like growth factor 2 receptor)				A330053J22(M), A530040I05(M)	No
ENOR61	chr17:50015030..50064612	+	50	5	10.1	Unspliced		*Satb1* (special AT-rich sequence binding protein 1)					No

Table 2. Continued

Region ID	Chromosomal Location[a]	Strand	Length (Kb)	Number of UNA TUs	Average TUs/100 Kb	Splicing Status	Known Sense Transcripts[b]	Antisense Transcripts (Protein-Coding)[c]	Antisense Transcripts (Noncoding)[c]	snoRNAs	miRNAs	Candidate Imprinted RIKEN cDNAs[d]	Human–Mouse Synteny[e]
ENOR62	chr17:69281588..69294109	+	13	2	16.0	Unspliced		*Tgif* (TG interacting factor)					No
ENOR63	chr18:55492179..55582108	+	90	2	2.2	Spliced	A730052O22 cDNA						Yes
ENOR64	chr19:20553802..20565004	+	11	2	17.9	Unspliced							No
ENOR65	chrX:7176015..7210908	–	35	2	5.7	Unspliced							No
ENOR66	chrX:151935162..152088234	–	153	6	3.9	Unspliced						9430043O19(M)	Yes

[a]Boundaries refer to the genomic coordinates of the start and end of the most 5′ and 3′ cDNA, respectively, within each ENOR.
[b]"Sense" refers to a transcript whose genomic coordinates overlapped with that of an ENOR at the pre-mRNA level and that was located on the same strand as the ENOR.
[c]"Antisense" refers to a transcript whose genomic coordinates overlapped with that of an ENOR at the pre-mRNA level and that was located on the opposite strand to the ENOR.
[d]"(P)" and "(M)" refer to paternally and maternally imprinted cDNAs, respectively.
[e]"Yes" indicates that more than 50% of the ENOR was successfully aligned to the human genome.
DOI: 10.1371/journal.pgen.0020037.t002

The total length of the 66 ENORs was 6,044 Kb, corresponding to 0.23% of the mouse genome.

We classified the 66 ENORs based upon the frequency of spliced and unspliced ESTs (Table 3). Twenty-eight regions contained numerous spliced ESTs, while the remaining 38 regions included no or very few spliced ESTs. The longest unspliced region was ENOR57, which included ten UNA TUs spanning almost 460 Kb. Interestingly, we found that *Air,* which has previously been reported as unspliced [5], overlapped with several spliced cDNAs and ESTs, suggesting that *Air* may also exist as spliced isoforms. Consistent with this idea, there is another ncRNA, *Nespas,* for which multiple spliced and unspliced forms have been reported, and the human–mouse conservation of these different isoforms suggests that they may be functionally relevant [34].

Sequence conservation between different species indirectly suggests function. To assess the conservation of the ENORs, we searched for syntenic human loci using mouse–human whole genome alignments available from the University of California Santa Cruz (UCSC) Genome Browser [35,36]. Many ENORs (38 of 66) could be successfully aligned between mouse and human over at least 50% of their length (Table 2). However, a significant minority were not well-conserved, and these included known functional ncRNAs such as *Air* and *Ube3a-ats,* which highlights that a lack of conservation does not necessarily imply a lack of function [37]. Because some long poorly conserved ncRNAs such as *Xist* retain patches of well-conserved sequence [6], we also examined ENOR conservation in short 50-nucleotide windows (Figure S2). This approach indicated not only that ENORs have patches of high conservation but also that they are more conserved than the genome average, so that while only ~45% of the mouse genome windows are alignable to the human genome, ~60% of ENOR windows are alignable.

ENORs Show Evidence of Imprinting and Antisense Transcription

Because previous studies revealed associations between macro ncRNAs and both imprinting and antisense transcription, we looked to see if our ENOR loci were associated with either of these phenomena.

To examine imprinting, we obtained 2,114 candidate imprinted mouse transcripts previously identified by Nikaido et al. [38]. By mapping these transcripts to the mouse genome (May 2004 assembly; mm5), we found that 13 ENORs (containing 20 candidate imprinted cDNAs) showed evidence of imprinting (Tables 2 and 3). This number was significantly higher than expected by chance (Chi-square, $p < 0.001$). Of the 13 ENORs identified, four contained well-characterized imprinted ncRNAs *(Rian, Air, Ube3a-ats,* and *Kcnq1ot1)* and nine represent potentially imprinted ncRNAs.

To characterize *cis*-antisense transcription, we searched for transcripts that appeared in the complementary strand of each ENOR (Tables 2 and 3). Of 28 spliced ENORs, two corresponded to known antisense ncRNAs *(Air* and *Dleu2),* and a further eight represented potentially novel antisense transcripts to either protein-coding genes *(Mcp, Ndufs1,* and *Traf3ip2)* or to noncoding transcripts. In the case of *Dleu2,* which has been suggested to play a role in the splice-site regulation of its cognate antisense partner *Trim13* [32], we also identified a potentially new antisense partner, *Kcnrg*. Of 38 unspliced ENORs, two corresponded to

Table 3. Summary Characteristics of 66 ENORs

Splicing Status		Antisense Transcription		Imprinted Candidate	miRNA Host	snoRNA Host
Category	Number	Category	Number			
Spliced	28	Antisense to protein-coding mRNA	5	2	2	0
		Antisense to ncRNA	5	0	0	0
		No antisense transcripts	18	2	3	1
Unspliced	38	Antisense to protein-coding mRNA	9	2	1	0
		Antisense to ncRNA	6	0	1	0
		No antisense transcripts	23	7	0	0

DOI: 10.1371/journal.pgen.0020037.t003

known ncRNAs antisense to *Ube3a* and *Kcnq1,* and a further 13 represented potentially novel antisense transcripts to either protein-coding genes *(Cyp2d22, Nr6a1, Ppp2r5a, Satb1, Tgif, 3222402P14Rik,* and *4933421G18Rik)* or to noncoding transcripts. Many of the protein-coding genes are involved in development and disease, and, as with *Igf2r* and *Air,* the discovery of long noncoding antisense transcripts may be very important in understanding the regulation of these genes.

ENORs Are Associated with miRNAs and Show Tissue-Specific Expression

As indicated above, a number of ENORs corresponded to known ncRNAs that act as host genes for either snoRNAs or miRNAs. We were therefore interested to see whether any other ENORs contained miRNAs or snoRNAs. We downloaded 224 known miRNAs and 175 snoRNAs from the miRBase Registry and RNAdb, respectively [24,39]. We then mapped these sequences to the mouse genome, and examined them for overlap with the 66 ENORs. We found that seven ENORs overlapped with 14 known miRNAs (14/224; 6%; Table 2), an association unlikely to have occurred by chance ($p < 0.0001$). Some of these ENORs also contained imprinted cDNA candidates, in keeping with a previously noted association between miRNAs and imprinting [40]. No new snoRNA hosts were found.

Next, we examined the expression of ENOR transcripts. Using the publicly available mouse gene expression atlas data from the Genomics Institute of the Novartis Research

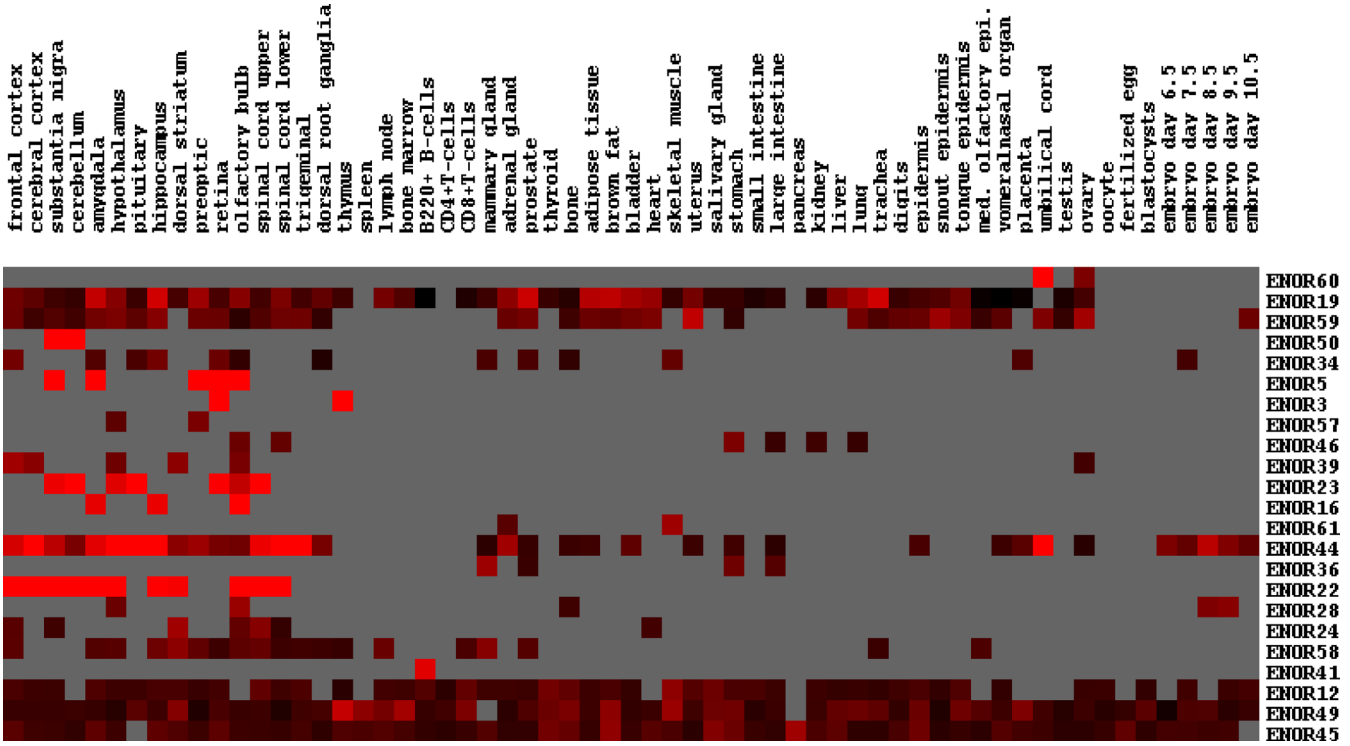
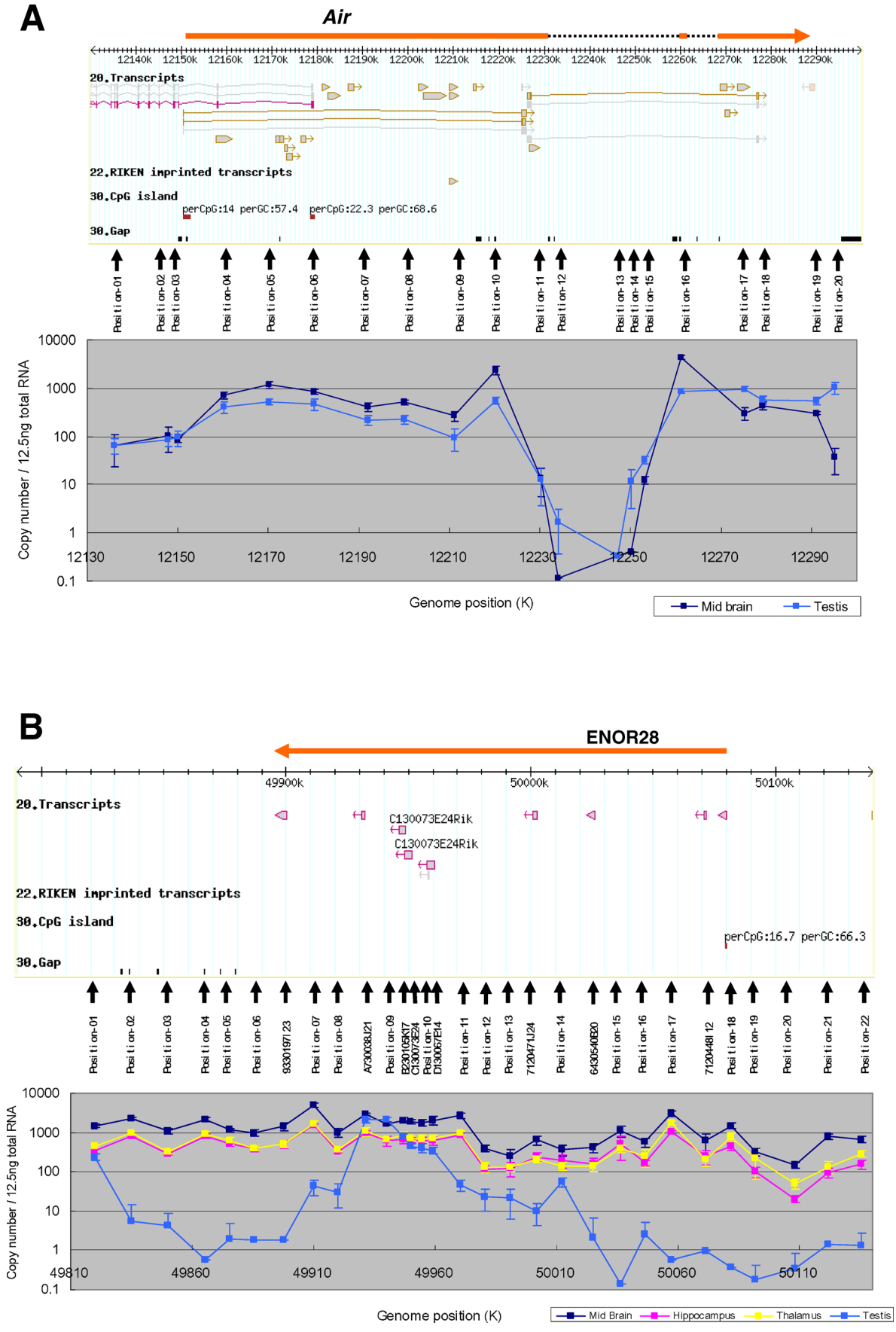


Figure 3. ENOR Tissue Expression

Tissue expression information for individual ENORs was obtained using publicly available GNF Gene Expression Atlas data. GNF probes that overlapped ENORs were identified, and the corresponding relative expression ratios for 61 tissues were hierarchically clustered. Red squares indicate high expression, black squares indicate low expression, and grey squares indicate where expression was not reliably detected (based upon Affymetrix MAS5 absent/present calls). med. olfactory epi., medial olfactory epithelium.

DOI: 10.1371/journal.pgen.0020037.g003


A
Air
12140k 12150k 12160k 12170k 12180k 12190k 12200k 12210k 12220k 12230k 12240k 12250k 12260k 12270k 12280k 12290k
20.Transcripts
22.RIKEN imprinted transcripts
30.CpG island
perCpG:14 perGC:57.4
perCpG:22.3 perGC:68.6
30.Gap
Position-01
Position-02
Position-03
Position-04
Position-05
Position-06
Position-07
Position-08
Position-09
Position-10
Position-11
Position-12
Position-13
Position-14
Position-15
Position-16
Position-17
Position-18
Position-19
Position-20
Copy number / 12.5ng total RNA
10000
1000
100
10
1
0.1
12130
12150
12170
12190
12210
12230
12250
12270
12290
Genome position (K)
Mid brain
Testis
B
ENOR28
49900k
50000k
50100k
20.Transcripts
C130073E24Rik
C130073E24Rik
22.RIKEN imprinted transcripts
30.CpG island
perCpG:16.7 perGC:66.3
30.Gap
Position-01
Position-02
Position-03
Position-04
Position-05
Position-06
9330197I23
Position-07
Position-08
A730038J21
Position-09
B230105K17
C130073E24
Position-10
D130067E14
Position-11
Position-12
Position-13
7120471J24
Position-14
6430540E20
Position-15
Position-16
Position-17
7120448I12
Position-18
Position-19
Position-20
Position-21
Position-22
Copy number / 12.5ng total RNA
10000
1000
100
10
1
0.1
49810
49860
49910
49960
50010
50060
50110
Genome position (K)
Mid Brain
Hippocampus
Thalamus
Testis

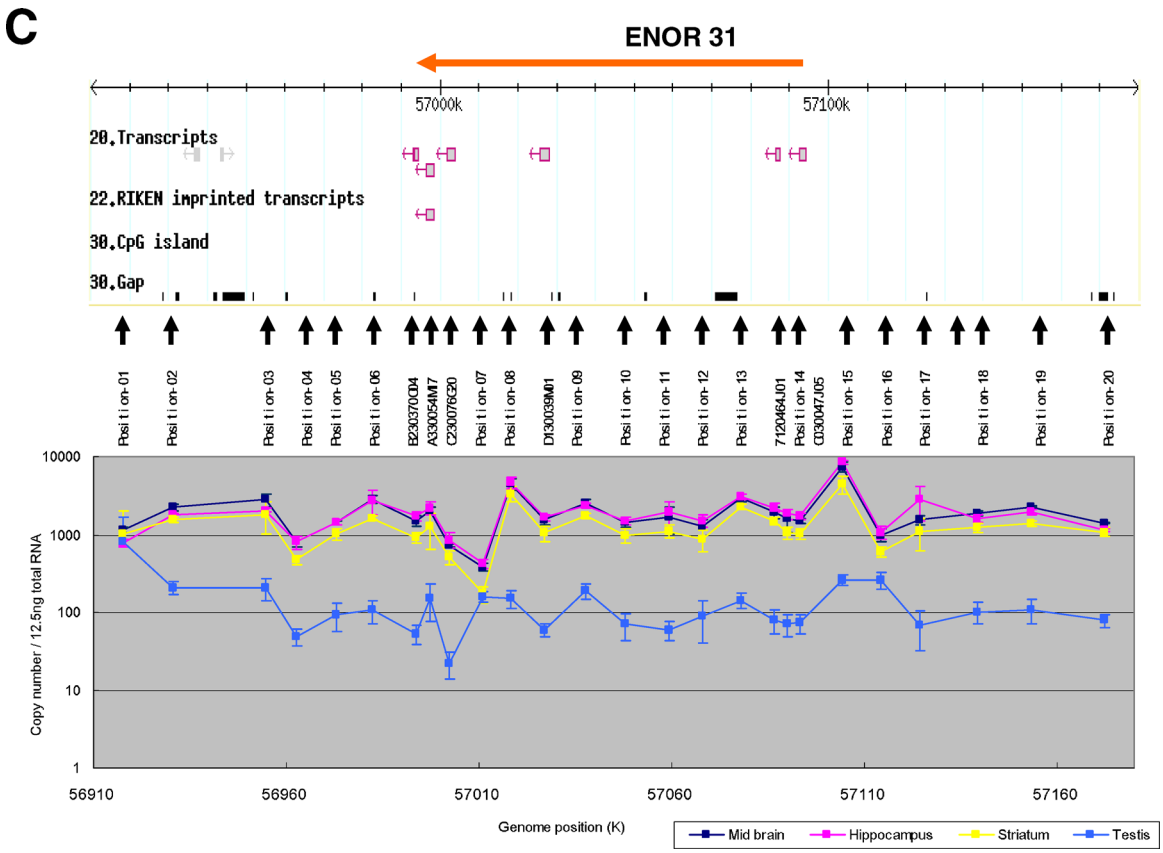


Figure 4. qRT-PCR Analysis

Analysis of (A) *Air*, (B) ENOR28, and (C) ENOR31 loci. Above in each panel, screen shots of the GEV featuring the loci around *Air*, ENOR28, and ENOR31 are shown. The orange bars indicate the regions for *Air*, ENOR28, and ENOR31. cDNA sequences from the RIKEN and public databases are shown. Sequences mapped on the plus strand and minus strand are brown and purple, respectively. Predicted genes from Ensembl, NCBI, and RefSeq databases are shown in gray. For RIKEN imprinted transcripts, imprinted cDNA candidates identified previously [38] are shown. CpG islands as defined by the UCSC Genome Browser are shown. Positions of primer pairs are marked by small vertical arrows. Below in each panel, qRT-PCR results for midbrain, hippocampus, thalamus, striatum, and testis using the corresponding primer pairs are shown.

DOI: 10.1371/journal.pgen.0020037.g004

Foundation (GNF) [41], we found that 23 ENORs were expressed in at least one of the 61 tissues examined. Thirty-three of the remaining ENORs did not have any corresponding GNF probes, while a further ten had probes whose expression was not reliably detected. Of the 23 ENORs, some were expressed almost ubiquitously, while others showed a restricted, tissue-specific expression profile (Figure 3). Notably, many ENORs were enriched in the central nervous system (CNS), and these included known brain-specific ncRNAs *Ube3a-ats* (ENOR22) and *Rian* (ENOR44) [28,30]. Because only a minority of ENORs had supporting GNF information, we also used RIKEN EST data to assess whether ENOR transcripts showed preferential expression in particular tissues. We searched for RIKEN ESTs that mapped within each ENOR and tallied the number of clones associated with the ESTs that were derived from a particular tissue (as per Edinburgh Mouse Atlas Project descriptions), and then we compared these counts with those of the entire FANTOM3 set. We found that ENOR transcripts as a whole are significantly overrepresented in a number of tissues including the CNS (Table S2). A caveat to this result is that RIKEN ESTs were derived after intensive subtraction, and their relative abundance might therefore not reflect natural tissue expression, although the EST data were in general agreement with the GNF results for a number of ENORs.

As noted earlier, spliced ENORs such as that corresponding to *Dleu2* may not necessarily represent macro ncRNAs because clusters of UNA transcripts may be derived from the introns of longer pre-mRNAs whose final product may be less than 10 Kb. To proceed, we therefore focused our attention on the unspliced class of ENORs, which we reasoned were most likely to represent novel macro ncRNAs.

Long PCR and Quantitative RT-PCR Provide Indirect Evidence of Macro ncRNAs

As proof of principle, we selected two regions for initial experimental characterization: ENOR28 and ENOR31 (Figure 4). ENOR28 (Figure 4B) was located on Chromosome 8 (49,895,333–50,079,904; mm5), appeared unspliced, spanned 185 Kb, and contained eight UNA TUs. ENOR31 (Figure 4C) was also on Chromosome 8 (56,996,422–57,094,390), appeared unspliced, spanned 98 Kb, and contained five UNA TUs, one of which was a possible imprinted transcript [38]. The majority of cDNAs in both regions were from common tissues (CNS), and this—together with their lack of splicing and greater than average length—made them good initial candidates.

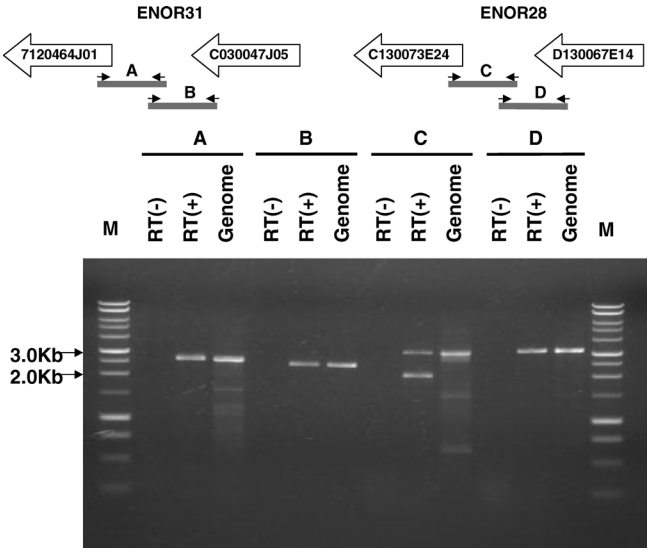


Figure 5. Presence of Transcription between Adjacent cDNAs

PCR was carried out with and without reverse transcription (RT[+] and RT[–], respectively) using midbrain total RNA and the corresponding primer pairs (see Table S3). PCR using genomic DNA was also carried out as a control. A DNA ladder (Promega; http://www.promega.com) was used as a size marker. The amplified fragments were confirmed as the expected ones by analyzing digestion pattern using several restriction enzymes. The lower band, observed in the RT(+) lane of the amplified fragment C, seems to be nonspecific, because it was amplified using only the right primer and because it showed a digestion pattern with restriction enzymes quite different from that of the upper band and the band of the genomic DNA (unpublished data).

DOI: 10.1371/journal.pgen.0020037.g005

Initially, we looked for the presence of transcription between neighboring cDNAs by long PCR (all of the PCR primers are shown in Table S3). Figure 5 shows that we successfully amplified transcripts between cDNAs 7120464J01 and C030047J05 in ENOR31, and between cDNAs C130073E24 and D130067E14 in ENOR28. These results suggested that directly adjacent cDNAs arise from a common transcript.

If each region represents one continuous transcript under the control of a single promoter in a given tissue, we reasoned, then across multiple tissues the levels of expression for each ENOR cDNA should remain consistent with those of the other cDNAs in the region. Total RNA was therefore isolated from eight different tissues, and each ENOR cDNA expression profile was examined by quantitative real-time RT-PCR (qRT-PCR). We found that the expression of all cDNAs (apart from A730038J21 in ENOR28) was highly intercorrelated ($R > 0.9$; Table S4), thus providing indirect evidence that not only directly adjacent cDNAs but also those more remote from one another were from the same transcript.

Northern Blots Directly Confirm Existence of Multiple Novel Macro ncRNAs

Northern blot analysis is a direct means to demonstrate the existence of very large RNAs. We therefore selected eight ENORs (ENOR2, ENOR14, ENOR16, ENOR28, ENOR31, ENOR54, ENOR61, and ENOR62), which together were representative of a broad range of lengths, chromosomes, and EST abundance, and tested them by northern blot using specific probes (Table S3). As a positive control, we also examined ENOR60, which corresponds to *Air*. Figure 6 shows that *Air* was readily detected as a band greater than 10 Kb in size. Similarly, probes against ENOR2, ENOR16, ENOR28, ENOR31, and ENOR54 all detected clearly visible bands greater than 10 Kb. Other ENORs gave less clear results. ENOR61 had a broad signal that appeared as a smear originating from the upper reaches of the gel, and it was unclear whether this was due to degradation of a large transcript or was nonspecific. ENOR62 produced a similar result. Probes for ENOR14, on the other hand, detected one major product larger than 7.5 Kb and possibly another larger than 10 Kb. Thus, in six of nine cases, we were able to successfully demonstrate macro RNAs larger than 10 Kb and in the remaining three cases the results were equivocal.

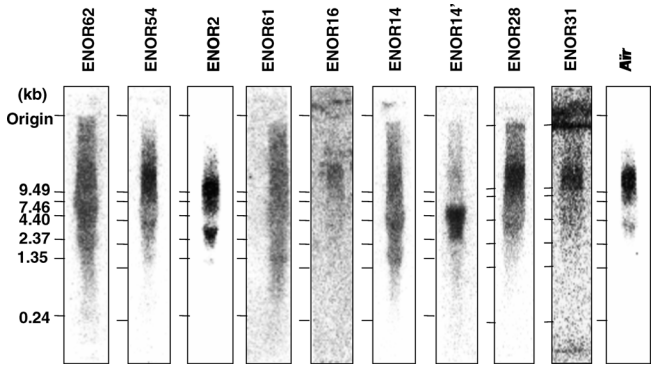


Figure 6. Northern Blot Analysis of ENOR Transcripts

Mouse whole brain total RNA (10 μg/lane) was used for the analysis except for ENOR2 and ENOR61, where mouse thymus total RNA was used. DNA fragments without any predicted repeated sequences were PCR-amplified from cDNAs in ENORs (Table S3), labeled with ^{32}P-dCTP (Amersham Biosciences), and then used as probes. RNA size was estimated with an RNA ladder (Invitrogen). ENORs are listed in increasing order based on the estimated length of each region.

DOI: 10.1371/journal.pgen.0020037.g006

Detailed qRT-PCR Analysis Reveals That ENORs Might Contain Multiple Long Transcripts

Northern blots do not accurately resolve the size of transcripts larger than 10 Kb. For this reason, the 108-Kb *Air* transcript in Figure 6 appears only just above 10 Kb (which is similar to the original northern blot result obtained by Lyle et al. [42]), and the actual size of the other macro ncRNAs cannot be successfully determined from the blots. To gain a better understanding of the true extent of transcription across our regions, we therefore performed further qRT-PCR analysis across our original candidate ENORs, ENOR28 and ENOR31. Specific primer pairs were designed before, after, and along the length of the region, incorporating individual cDNAs as well as the areas in between (Figure 4B and 4C; Table S3). As a control, ENOR60, containing *Air*, was analyzed in a similar manner (Figure 4A; Table S3).

To begin, we extracted total RNA from different CNS tissues (midbrain, hippocampus, corpus striatum, and thalamus) and from testis, and assessed the level of expression of at least 20 separate subregions spanning the length of both ENORs as well as *Air*. Figure 4A demonstrates that, in the *Air* locus, expression arises from downstream of a CpG island, as previously reported [43], and then remains relatively constant for the next 70–80 Kb. Beyond this, expression falls below 100 transcripts per 12.5 ng of total RNA for the next 30–40 Kb

(primer pairs 11–15) then rises and plateaus again for a further 30 Kb. Examination of the alignment between the genomic DNA sequence of *Air* (GenBank accession number AJ249895) and the genome assembly revealed that there are two inserted sequences in the genome assembly (dotted lines in Figure 4A). These sequences are disconnected by gaps in the genome assembly, indicating that the transient fall in expression is an artifact. Overall, then, this result was in keeping with the presence of a continuous macro ncRNA ~100–110 Kb in size, and provided evidence that the qRT-PCR-based strategy employed here was able to successfully detect such long transcripts and to provide a reasonable estimate of their size.

Figure 4B illustrates expression across the ENOR28 locus. In CNS tissues, the overall expression pattern was similar to that of *Air,* with sustained expression over tens of kilobases at transcript copy numbers in the hundreds to thousands (per 12.5 ng of total RNA). Looking more closely and starting from upstream of the 5′ end, expression levels are at their lowest around primer pair 20, dipping below 100 copies; next, transcript levels for primer pairs 12–18 are intermediate; finally, from primer pairs 1–11 (a distance larger than 100 Kb), expression is highest of all, is relatively constant, and extends well beyond the previously defined 3′ ENOR boundary. Our previous experience using primer pairs against different positions of the same protein-coding genes had indicated that the expected differences in transcript copy number are generally less than 2-fold for the same transcript (unpublished data). Assuming such results can be applied here, the roughly 10-fold variation in CNS expression across ENOR28 challenges our original hypothesis of a single promoter driving expression across the entire region. Rather, it is possible that a number of separate transcripts are present, the longest of which spans primer pairs 1–11 and appears to be larger than 100 Kb. Interestingly, testis expression fell below detection threshold at both the 5′ and 3′ ends of ENOR28, suggesting the existence of a shorter testis-specific transcript.

Figure 4C shows that expression in the ENOR31 locus was relatively constant and extensive, with transcript copy numbers in CNS greater than 1,000 per 12.5 ng of total RNA not only within but also up- and downstream of the original ENOR boundaries. Approximately 10-fold expression spikes at primer pairs 15–16 and 7–8 suggested the possibility of up to three separate transcripts larger than 50 Kb. Testis expression gave a similar pattern but was much lower than CNS expression. Overall, then, assuming that a 10-fold variation in transcript levels between primer pairs is indicative of separate transcripts, both the ENOR28 and ENOR31 loci appear to produce not one but several macro ncRNAs (all of which are enriched in brain). However, it is worth noting that our data for *Air* (Figure 4A) (excluding regions with assembly gaps) also showed ~10-fold variation in transcript levels (e.g., primer pairs 9–10). Since *Air* is generally acknowledged to be a continuous transcript spanning ~108 Kb [5], it seems plausible that a 10-fold variation in transcript levels between primer pairs need not indicate multiple transcripts. If that is true, then the data for ENOR28 and ENOR31 would support the alternative conclusion that each region gives rise to a single macro ncRNA larger than 100 Kb.

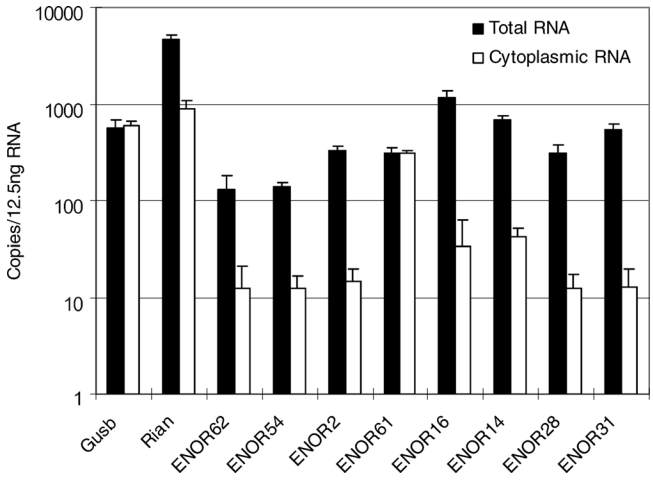


Figure 7. Localization of ENOR Transcripts

qRT-PCR was carried out using total and cytoplasmic RNA from mouse whole brain and the corresponding primer pairs (Table S3). ENORs are listed in increasing order based on the estimated length of each region. Apart from the results shown, we also examined the localization of other mRNAs *(β-actin* and *GAPDH)* and additional regions of *Rian* and other ENORs, and these results were consistent with the rest (unpublished data).
DOI: 10.1371/journal.pgen.0020037.g007

ENOR Transcripts Predominantly Localize to the Nucleus

Subcellular localization may provide clues to the function of ENOR transcripts. For instance, *Xist* exerts its chromosomal silencing effect within the nucleus [6]. We therefore examined the localization of the same eight ENORs (ENOR2, ENOR14, ENOR16, ENOR28, ENOR31, ENOR54, ENOR61, and ENOR62) we previously had characterized via northern blot by comparing brain expression levels from cytoplasmic and total RNA (the latter consists of both cytoplasmic and nuclear RNA). To validate our method, we initially tested β-glucuronidase *(Gusb)* mRNA, a housekeeping gene, and the *Rian* ncRNA (ENOR44), which preferentially localize to the cytoplasm and nucleus, respectively [28]. Figure 7 shows that, in keeping with our expectations, the copy number for *Gusb* mRNA was similar in cytoplasmic and total RNA (which suggests that there is a negligible nuclear RNA component) while *Rian* exists in cytoplasmic RNA at much lower levels than in total RNA (which suggests that the nuclear component predominates). Interestingly, when we examined the eight ENOR transcripts in an identical manner (Figure 7), seven of them (ENOR2, ENOR14, ENOR16, ENOR28, ENOR31, ENOR54, and ENOR62) showed much higher expression in total RNA, suggesting that they are localized in the nucleus. ENOR 61, on the other hand, appeared to be cytoplasmic.

Discussion

The analysis of full-length enriched cDNA libraries has been of vital importance in improving our understanding of the mammalian transcriptome. In this regard, however, unspliced noncoding cDNAs are often viewed with skepticism because they can arise as truncation artifacts of cDNA library construction. Here, we have shown that such artifacts cluster within very long, functionally important ncRNAs such as *Air* and *Xist,* and, rather than summarily dismissing these cDNAs as worthless, we have employed a strategy that uses them to identify long ncRNAs genome-wide. The resulting list of 66

candidate ENORs—itself almost certainly an underestimate—potentially expands several-fold the number of known mouse ncRNAs larger than 10 Kb in size, which, to date, includes only a few examples such as *Xist, Air, Kcnq1ot1,* and *Ube3a-ats,* most of which were successfully detected with our methods. In the past, such macro ncRNAs have been discovered experimentally on an ad hoc basis, and it has not been possible to systematically identify large ncRNAs by bioinformatics means, since most existing tools are limited to the discovery of smaller ncRNAs with conserved primary sequences and/or secondary structures [44]. Our strategy offers a solution to this problem.

Expression studies produced a number of interesting observations. First, in silico analysis indicated that some ENORs cluster together within the genome and are coexpressed. For example, ENOR22 and ENOR23 are located within 2,300 Kb of each other on Chromosome 7 and are specifically expressed in CNS. One possible explanation for this coexpression is that these regions share a common chromatin domain. Second, we found that the majority of ENOR transcripts were predominantly nuclear, similar to functional ncRNAs such as *Xist* and *Tsix*. ncRNAs like these are increasingly being recognized as important in altering chromatin structure [45,46], and it is tempting to speculate that the ENOR transcripts might also function in this way. Third, qRT-PCR studies of the ENOR28 and ENOR31 loci (Figure 4) indicated that the actual transcribed regions are almost certainly underestimated based upon current ENOR boundaries. This is not surprising, since the boundaries were estimated using internally primed transcript coordinates, and reflects that our discovery pipeline was not designed to capture transcription start and end sites. Lastly, despite the possible existence of multiple macro ncRNAs in ENOR28 and ENOR31, expression correlation between the individual cDNAs was extremely high (average $R = 0.96$). This indicates that even if there are separate transcripts arising from each region they appear to be under the influence of similar regional promoters, enhancers, or chromatin domains. Fluorescence in situ hybridization studies might prove useful to visualize the ENOR transcripts and their surrounding chromatin structure (via the use of histone-specific antibodies), and may also directly demonstrate in which specific cell types and subcellular compartments ENORs are localized. For instance, knowing exactly which groups of neurons in the brain express ENOR28 and ENOR31 transcripts might provide indirect information as to their function. Understanding how the expression of these transcripts is regulated will also be important. For instance, fine-detailed mapping of transcript copy number by qRT-PCR using more primer pairs might better define the relevant transcriptional start sites and promoter regions.

Macro ncRNAs can function in a variety of ways, and some clues to the possible function of the ENORs can be gleaned from their association with antisense transcription, candidate imprinting domains, and miRNAs. Antisense transcripts exert regulatory effects in a number of ways, as mentioned earlier. Some of these effects (e.g., RNA interference and translation regulation) can be mediated by small miRNAs and siRNAs, and it is unclear if longer antisense transcripts—such as those identified in this study—are required to function in certain regulatory contexts. Of course, long antisense transcripts might be processed into smaller functional RNAs, although there has been no evidence that *Xist* or *Air,* for instance, work in this manner. Macro ncRNAs can also regulate genomic imprinting. *Ube3a-ats, Kcnq1ot1,* and *Air* have all been implicated in the imprinting control of their antisense transcripts. These three ncRNAs are themselves imprinted, a fact correctly predicted by the methods we used here. These same methods suggest that a further nine ENORs might represent potentially imprinted ncRNAs, which, if confirmed, would add substantially to the number of imprinted ncRNAs currently characterized. Finally, in silico analysis detected overlap between ENORs and more than 5% of known mouse miRNAs, suggesting that one of the possible functions of some of these regions may be to act as miRNA host genes. Given a recent report indicating that many mammalian miRNAs are still to be discovered [47], the possibility exists that more ENORs will be associated with novel miRNAs in the future.

Lacking any direct evidence of ENOR function, we also acknowledge the possibility that some of these regions do not play any functional role as RNAs. It has been shown, for instance, that expression of the yeast noncoding RNA *Srg1* is necessary for the repression of its downstream gene, *Ser3,* but this appears to be due to the act of *Srg1* transcription (causing *Ser3* promoter interference) rather than any direct action of the *Srg1* RNA itself [48]. Meanwhile, Wyers et al. found that intergenic transcripts in yeast are rapidly degraded by a specific nuclear quality control pathway and are therefore likely to be nonfunctional [49]. Another recent report in which megabase deletions of noncoding DNA were engineered and failed to produce any detectable phenotype in mice [50] suggests that large noncoding regions of the genome may not have function. It should be noted, however, that the regions targeted in this deletion study lacked evidence of transcription, in direct contrast to the regions we have characterized. A suggestion has also been made that many noncoding transcripts simply represent useless by-products of "leaky transcription" [51]. Based upon our expression studies of ENOR28 and ENOR31, transcripts from both these regions appear to be clearly expressed in brain (estimated at 1–8 copies/cell based upon our previous work [52], which is similar to *Air* [Figure 4] and to most mRNAs [53]), suggesting that in these cases, at least, transcripts are controlled. To demonstrate the importance (or otherwise) of the ENORs, it will ultimately be necessary to test their function directly. This, together with efforts to better understand the gene structure, expression, and regulation of individual transcripts within each region, is the challenge that lies ahead.

Materials and Methods

Identification of known mouse ncRNAs within the FANTOM3 cDNA collection. Non-sno, non-micro reference mouse ncRNA sequences were downloaded from RNAdb, a database of mammalian ncRNAs (http://jsm-research.imb.uq.edu.au/rnadb) [24]. BLASTN was used to assess the similarity between the 102,801 FANTOM3 cDNAs and the reference ncRNAs using an initial *E*-value cutoff of 0.01, and any resulting hits with 98% or greater identity across 90% or more of the length of either a query cDNA or reference ncRNA sequence were considered significant matches. Repetitive sequences were identified in the FANTOM3 sequences using the union of RepeatMasker (http://www.repeatmasker.org) and runnseg predictions, and BLAST options –F "m" –U T were used to ignore repeats in the seeding but not the extension stage of the alignment.

Genome-wide search for clusters of internally primed cDNAs. We used the TU data prepared for the FANTOM3 project (ftp://fantom.gsc.riken.jp/RTPS/fantom3__mouse/primary__est__rtps/TU), which were generated by clustering the following mouse cDNA and EST sequences: (1) 56,006 mRNA sequences from GenBank (Release 139.0 and daily [2004–1–27]), (2) 102,597 RIKEN cDNAs from the FANTOM3 set, (3) 606,629 RIKEN 5′-end ESTs (5′-end set), (4)

907,007 RIKEN 3′-end ESTs (3′-end set), and (5) 1,569,444 GenBank EST sequences. Figure 2 summarizes the subsequent search pipeline, a full description of which was provided in the Results.

Bioinformatic analysis of candidate clusters. To judge whether ENOR sequences were spliced or unspliced, we searched for all TUs that overlapped with the chromosomal boundaries of each ENOR and were on the same strand. We included any spliced TUs whose intronic area overlapped with a region. We then counted ESTs associated with the TUs and classified the regions as follows: spliced, if spliced ESTs were more than 10% of total ESTs; otherwise unspliced. We used the threshold of 10% since a certain number of ESTs can be expected to be inappropriately mapped onto the genome and may therefore appear as falsely spliced ESTs. To find transcripts on the sense or antisense strand, we searched for TUs that overlapped with the regions on the same or opposite strand based on genomic coordinates. We searched for the gene name associated with these TUs, as defined by the RTPS pipeline used for FANTOM2 and FANTOM3 [54], and selected appropriate names manually. For the spliced ENORs, we selected the gene name of major transcripts on the same strand. For the unspliced ENORs, we used only informative gene names because uninformative names such as the RIKEN clone IDs were associated with unspliced cDNAs that covered only short regions. We also searched for gene names on the MGI database (http://www.informatics.jax.org) and used official gene symbols if available.

To examine ENOR conservation, we used blastz axtNet alignments from UCSC (ftp://hgdownload.cse.ucsc.edu/goldenPath/mm5/vsHg17) to identify blocks in the mouse genome that successfully align with the human genome. We classified individual ENORs as conserved if the total length of alignable blocks was greater than 50% of the ENOR length. To determine the overall conservation levels of ENOR sequences, the mouse genome was divided into 50-nucleotide windows, and the number of identically matching nucleotides in each window in the human genome was counted for both the ENORs and the genome as a whole.

Information on candidate imprinted cDNAs was provided by Nikaido et al. ([38]; http://fantom2.gsc.riken.go.jp/imprinting), and lists of miRNA and snoRNAs were downloaded from the miRBase Registry (http://www.sanger.ac.uk/Software/Rfam/mirna) and RNAdb, respectively, and mapped to the mouse genome (mm5) using MEGABLAST with options –F F –D 1 –J F. We then searched these imprinted cDNAs, miRNAs, and snoRNAs for overlap with the ENOR loci based on genomic coordinates. To determine whether the association between candidate imprinted cDNAs and ENOR loci was likely to have occurred by chance, we randomly sampled 66 regions with an average cDNA density equal to that of the ENORs (five RIKEN cDNAs per region) and determined the number of regions that contained at least one candidate imprinted cDNA; this procedure was repeated 100 times, and the significance was determined using a Chi-square test. To determine the significance of the association between miRNAs and ENOR loci, we performed the following calculation: given that ENORs cover 0.23% of the genome, the probability that a miRNA lies in an ENOR on the same strand is 0.0023×0.5. Using the binomial distribution, the probability that 14 or more out of 224 miRNAs lie in ENORs is about 3×10^{-20} (i.e., $p < 0.0001$).

To examine ENOR expression, we identified GNF Gene Expression Atlas (http://expression.gnf.org/cgi-bin/index.cgi) probes that overlapped with the genomic loci of the ENORs via the UCSC Genome Browser (http://genome.ucsc.edu), then downloaded the relevant expression data (http://symatlas.gnf.org). Affymetrix MAS5 software absent/present calls were used to identify probes with detectable expression in at least one of the 61 tissues tested. Log 2 ratio expression data for these probes were then hierarchically clustered via average linkage clustering using Cluster software [55]. Additionally, we downloaded the list of RIKEN libraries and their corresponding Edinburgh Mouse Atlas Project (http://genex.hgu.mrc.ac.uk) tissue descriptions, then searched for RIKEN ESTs that mapped within an ENOR region on the same strand, and tallied the number of ESTs that were derived from each tissue library. We counted 5′ EST and 3′ EST sequences derived from a same clone only once. Library information for some ESTs could not be used because of uninformative tissue descriptions.

Primers. Primer pairs were designed using Primer3 software [56], with an optimal primer size of 20 bases and annealing temperature of 60 °C (see Table S3). The uniqueness of the designed primer pairs was checked by a BLAST search (http://www.ncbi.nlm.nih.gov/BLAST) so that homologous regions were not cross-amplified by the same primer pair.

Preparation of RNA samples. Adult male C57BL/6J mice were killed according to the RIKEN Institute's guidelines, and the tissues were removed. Total RNA was extracted by the acid phenol-guanidinium thiocyanate-chloroform method [57]. Cytoplasmic RNA was prepared as described elsewhere [58]. RNA was checked by agarose gel electrophoresis and was treated with DNaseI before RT-PCR as described elsewhere [52].

RT-PCR analysis of candidate clusters. First-strand cDNA synthesis (5 μg of total RNA per 20-μl reaction) was carried out using a random primer and the ThermoScript RT-PCR System (Invitrogen; http://www.invitrogen.com) in accordance with the manufacturer's protocol. qRT-PCR was carried out with first-strand cDNA corresponding to 12.5 ng of total RNA per test well using the tailor-made reaction [52]. The PCR reactions were performed with an ABI Prism machine (Applied Biosystems; http://www.appliedbiosystems.com) using the following cycling protocols: 15-min hot start at 94 °C, followed by 40 cycles of 15 s at 94 °C, 30 s at 60 °C, and 30 s at 72 °C. The threshold cycle (Ct) value was calculated from amplification plots, in which the fluorescence signal detected was plotted against the PCR cycle. The number of transcripts was calculated from the slope of the standard curve using genomic DNA.

Long PCR. Long PCR was carried out with first-strand cDNA corresponding to 500 ng of total RNA and KOD DNA polymerase (Toyobo; http://www.toyobo.co.jp/e) per 50-μl reaction according to the manufacturer's protocol. We also used 200 ng of mouse genomic DNA, instead of first-strand cDNA, to amplify the fragments from the genome. The PCR reactions were performed with an ABI9700 (Applied Biosystems) using the following cycling protocols: 2-min hot start at 94 °C, followed by 35 cycles of 15 s at 94 °C, 30 s at 60 °C, and 5 min at 68 °C. One to two microliters of sample was subjected to 1% agarose gel electrophoresis.

Northern blot. Total RNA was denatured by formaldehyde/formamide and electrophoresed in a 1% agarose gel. RNA was transferred onto Hybond-N+ nylon membrane (GE Healthcare Life Sciences; http://www4.amershambiosciences.com). Hybridization was carried out using ^{32}P-labeled DNA probe and ExpressHyb hybridization solution (BD Biosciences; http://www.bdbiosciences.com) according to the manufacturer's protocol. The hybridization signal was detected using a BAS2500 image analyzer (Fujifilm; http://www.fujifilm.com).

Supporting Information

Figure S1. Genomic Distribution of 66 ENORs

Found at DOI: 10.1371/journal.pgen.0020037.sg001 (68 KB PPT).

Figure S2. ENORs Are More Conserved than the Genome Average

Found at DOI: 10.1371/journal.pgen.0020037.sg002 (68 KB PPT).

Table S1. ENORs on Each Chromosome

Found at DOI: 10.1371/journal.pgen.0020037.st001 (17 KB XLS).

Table S2. EST Tissue Data for 66 ENORs

Found at DOI: 10.1371/journal.pgen.0020037.st002 (33 KB XLS).

Table S3. Primer Pairs

Found at DOI: 10.1371/journal.pgen.0020037.st003 (39 KB XLS).

Table S4. Expression Correlation between cDNAs within ENOR28 and ENOR31

Data for (A) ENOR28 and (B) ENOR31.

Found at DOI: 10.1371/journal.pgen.0020037.st004 (19 KB XLS).

Accession Numbers

The MGI (http://www.informatics.jax.org) accession numbers for the sequences described in this paper are *3222402P14Rik* (2442104), *4933421G18Rik* (1913976), *Air* (1353471), *Cyp2d22* (1929474), *Dleu2* (1934030), *Dmd* (94909), *Emx2os* (3052329), *Gusb* (95872), *Igf2r* (96435), *Kcnq1ot1* (1926855), *Kcnrg* (2685591), *Mcp* (1203290), *Ndufs1* (2443241), *Nespas* (1861674), *Nr6a1* (1352459), *Ppp2r5a* (1929474), *Rian* (19222995), *Satb1* (105084), *Slc22a2* (18339), *Slc22a3* (1333817), *Tgif* (1194497), *Traf3ip2* (2143599), *Trim13* (1913847), *Tsix* (1336196), *Ube3a* (105098), and *Xist* (98974). The SGD (http://www.yeastgenome.org) accession number for yeast *Srg1* is S000029010. The NCBI EntrezGene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene) accession number for yeast *Ser3* is 856814.

Acknowledgments

Author contributions. MF, KCP, PC, YH, and HS conceived and designed the experiments. NN, SF, CK, and HS performed the experiments. MF, KCP, MCF, and HS analyzed the data. MF, CK, JK,

PC, YH, and HS contributed reagents/materials/analysis tools. MF, KCP, CB, JSM, and HS wrote the paper.

Funding. This study was supported by research grants from the Australian Research Council to JSM and from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government to YH for: (1) the Genome Network Project, (2) the RIKEN Genome Exploration Research Project, and (3) the Strategic Programs for R&D of RIKEN. MF is supported by The Jackson Laboratory Postdoctoral Fellowship award. KCP is supported by a National Health and Medical Research Council Medical Postgraduate Scholarship. MCF is supported by a University of Queensland Postdoctoral Fellowship.

Competing interests. The authors have declared that no competing interests exist. ■

References

1. Mattick JS, Makunin IV (2005) Small regulatory RNAs in mammals. Hum Mol Genet 14 (Suppl 1): R121–R132.
2. Pfeffer S, Zavolan M, Grasser FA, Chien M, Russo JJ, et al. (2004) Identification of virus-encoded microRNAs. Science 304: 734–736.
3. Bartel DP (2004) MicroRNAs: Genomics, biogenesis, mechanism, and function. Cell 116: 281–297.
4. Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, et al. (1992) 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 71: 515–526.
5. Sleutels F, Zwart R, Barlow DP (2002) The non-coding Air RNA is required for silencing autosomal imprinted genes. Nature 415: 810–813.
6. Wutz A (2003) Xist RNA associates with chromatin and causes gene silencing. In: Barciszewski J, Erdmann VA, editors. Noncoding RNAs: Molecular biology and molecular medicine. Georgetown (Texas): Landes Bioscience. pp. 49–65.
7. Delaval K, Feil R (2004) Epigenetic regulation of mammalian genomic imprinting. Curr Opin Genet Dev 14: 188–195.
8. Kelley RL, Kuroda MI (2000) Noncoding RNA genes in dosage compensation and imprinting. Cell 103: 9–12.
9. Chen J, Sun M, Kent WJ, Huang X, Xie H, et al. (2004) Over 20% of human transcripts might form sense-antisense pairs. Nucleic Acids Res 32: 4812–4820.
10. Yelin R, Dahary D, Sorek R, Levanon EY, Goldstein O, et al. (2003) Widespread occurrence of antisense transcription in the human genome. Nat Biotechnol 21: 379–386.
11. Kiyosawa H, Yamanaka I, Osato N, Kondo S, Hayashizaki Y (2003) Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. Genome Res 13: 1324–1334.
12. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, et al. (2005) Antisense transcription in the mammalian transcriptome. Science 309: 1564–1566.
13. Lavorgna G, Dahary D, Lehner B, Sorek R, Sanderson CM, et al. (2004) In search of antisense. Trends Biochem Sci 29: 88–94.
14. Jen CH, Michalopoulos I, Westhead DR, Meyer P (2005) Natural antisense transcripts with coding capacity in *Arabidopsis* may have a regulatory role that is not linked to double-stranded RNA degradation. Genome Biol 6: R51.
15. Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, et al. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell 116: 499–509.
16. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, et al. (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308: 1149–1154.
17. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, et al. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420: 563–573.
18. Numata K, Kanai A, Saito R, Kondo S, Adachi J, et al. (2003) Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. Genome Res 13: 1301–1306.
19. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, et al. (2005) The transcriptional landscape of the mammalian genome. Science 309: 1559–1563.
20. Willingham AT, Orth AP, Batalov S, Peters EC, Wen BG, et al. (2005) A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. Science 309: 1570–1573.
21. Mattick JS (2005) The functional genomics of noncoding RNA. Science 309: 1527–1528.
22. Ravasi T, Suzuki H, Pang KC, Katayama S, Furuno M, et al. (2006) Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. Genome Res 16: 11–19.
23. Nam DK, Lee S, Zhou G, Cao X, Wang C, et al. (2002) Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. Proc Natl Acad Sci U S A 99: 6152–6156.
24. Pang KC, Stephen S, Engstrom PG, Tajul-Arifin K, Chen W, et al. (2005) RNAdb—A comprehensive mammalian noncoding RNA database. Nucleic Acids Res 33: D125–D130.
25. Carninci P, Waki K, Shiraki T, Konno H, Shibata K, et al. (2003) Targeting a complex transcriptome: The construction of the mouse full-length cDNA encyclopedia. Genome Res 13: 1273–1289.
26. Borsani G, Tonlorenzi R, Simmler MC, Dandolo L, Arnaud D, et al. (1991) Characterization of a murine gene expressed from the inactive X chromosome. Nature 351: 325–329.
27. Engemann S, Strodicke M, Paulsen M, Franck O, Reinhardt R, et al. (2000) Sequence and functional comparison in the Beckwith-Wiedemann region: Implications for a novel imprinting centre and extended imprinting. Hum Mol Genet 9: 2691–2706.
28. Hatada I, Morita S, Obata Y, Sotomaru Y, Shimoda M, et al. (2001) Identification of a new imprinted gene, Rian, on mouse chromosome 12 by fluorescent differential display screening. J Biochem (Tokyo) 130: 187–190.
29. Cavaille J, Seitz H, Paulsen M, Ferguson-Smith AC, Bachellerie JP (2002) Identification of tandemly-repeated C/D snoRNA genes at the imprinted human 14q32 domain reminiscent of those at the Prader-Willi/Angelman syndrome region. Hum Mol Genet 11: 1527–1538.
30. Landers M, Bancescu DL, Le Meur E, Rougeulle C, Glatt-Deeley H, et al. (2004) Regulation of the large (approximately 1000 kb) imprinted murine Ube3a antisense transcript by alternative exons upstream of Snurf/Snrpn. Nucleic Acids Res 32: 3480–3492.
31. Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A (2004) Identification of mammalian microRNA host genes and transcription units. Genome Res 14: 1902–1910.
32. Corcoran MM, Hammarsund M, Zhu C, Lerner M, Kapanadze B, et al. (2004) DLEU2 encodes an antisense RNA for the putative bicistronic RFP2/LEU5 gene in humans and mouse. Genes Chromosomes Cancer 40: 285–297.
33. Noonan FC, Goodfellow PJ, Staloch LJ, Mutch DG, Simon TC (2003) Antisense transcripts at the EMX2 locus in human and mouse. Genomics 81: 58–66.
34. Williamson CM, Skinner JA, Kelsey G, Peters J (2002) Alternative non-coding splice variants of Nespas, an imprinted gene antisense to Nesp in the Gnas imprinting cluster. Mamm Genome 13: 74–79.
35. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome Res 12: 996–1006.
36. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D (2003) Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A 100: 11484–11489.
37. Pang KC, Frith MC, Mattick JS (2006) Rapid evolution of noncoding RNAs: Lack of conservation does not mean lack of function. Trends Genet 22: 1–5.
38. Nikaido I, Saito C, Mizuno Y, Meguro M, Bono H, et al. (2003) Discovery of imprinted transcripts in the mouse transcriptome using large-scale expression profiling. Genome Res 13: 1402–1409.
39. Griffiths-Jones S (2004) The microRNA Registry. Nucleic Acids Res 32: D109–D111.
40. Seitz H, Royo H, Bortolin ML, Lin SP, Ferguson-Smith AC, et al. (2004) A large imprinted microRNA gene cluster at the mouse Dlk1-Gtl2 domain. Genome Res 14: 1741–1748.
41. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 101: 6062–6067.
42. Lyle R, Watanabe D, te Vruchte D, Lerchner W, Smrzka OW, et al. (2000) The imprinted antisense RNA at the Igf2r locus overlaps but does not imprint Mas1. Nat Genet 25: 19–21.
43. Wutz A, Smrzka OW, Schweifer N, Schellander K, Wagner EF, et al. (1997) Imprinted expression of the Igf2r gene depends on an intronic CpG island. Nature 389: 745–749.
44. Eddy SR (2002) Computational genomics of noncoding RNA genes. Cell 109: 137–140.
45. Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, et al. (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. Cell 120: 169–181.
46. Navarro P, Pichard S, Ciaudo C, Avner P, Rougeulle C (2005) Tsix transcription across the Xist gene alters chromatin conformation without affecting Xist transcription: Implications for X-chromosome inactivation. Genes Dev 19: 1474–1484.
47. Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, et al. (2005) Identification of hundreds of conserved and nonconserved human microRNAs. Nat Genet 37: 766–770.
48. Martens JA, Laprade L, Winston F (2004) Intergenic transcription is required to repress the *Saccharomyces cerevisiae* SER3 gene. Nature 429: 571–574.
49. Wyers F, Rougemaille M, Badis G, Rousselle JC, Dufour ME, et al. (2005) Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. Cell 121: 725–737.
50. Nobrega MA, Zhu Y, Plajzer-Frick I, Afzal V, Rubin EM (2004) Megabase deletions of gene deserts result in viable mice. Nature 431: 988–993.
51. Dennis C (2002) The brave new world of RNA. Nature 418: 122–124.
52. Suzuki H, Okunishi R, Hashizume W, Katayama S, Ninomiya N, et al. (2004)

Identification of region-specific transcription factor genes in the adult mouse brain by medium-scale real-time RT-PCR. FEBS Lett 573: 214–218.
53. Hastie ND, Bishop JO (1976) The expression of three abundance classes of messenger RNA in mouse tissues. Cell 9: 761–774.
54. Kasukawa T, Katayama S, Kawaji H, Suzuki H, Hume DA, et al. (2004) Construction of representative transcript and protein sets of human, mouse, and rat as a platform for their transcriptome and proteome analysis. Genomics 84: 913–921.
55. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95: 14863–14868.
56. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365–386.
57. Chomczynski P, Sacchi N (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem 162: 156–159.
58. Carninci P, Nakamura M, Sato K, Hayashizaki Y, Brownstein MJ (2002) Cytoplasmic RNA extraction from fresh and frozen mammalian tissues. Biotechniques 33: 306–309.