Digital transcriptome profiling using selective hexamer priming for cDNA synthesis

Christopher D Armour[1], John C Castle[1], Ronghua Chen[1], Tomas Babak[1], Patrick Loerch[1], Stuart Jackson[2], Jyoti K Shah[1], John Dey[2], Carol A Rohl[1], Jason M Johnson[1] & Christopher K Raymond[1]

We developed a procedure for the preparation of whole transcriptome cDNA libraries depleted of ribosomal RNA from only 1 μg of total RNA. The method relies on a collection of short, computationally selected oligonucleotides, called 'not-so-random' (NSR) primers, to obtain full-length, strand-specific representation of nonribosomal RNA transcripts. In this study we validated the technique by profiling human whole brain and universal human reference RNA using ultra-high-throughput sequencing.

Large-scale transcriptome analysis has been energized in recent years by stunning technological advances in DNA sequencing. Although these new technologies obviate the need for clonal separation of cDNA fragments, library construction remains a critical component of transcriptome sequencing strategies. With an overwhelming fraction of RNA transcripts coding for structural subunits of ribosomes in prokaryotic and eukaryotic species alike, molecular techniques that enrich for more informative low-copy transcripts have been developed to maximize sequencing efficiency. In eukaryotic cells, mRNA selection has been a central feature of the most widely used methods for ultra-high-throughput sequencing[1,2].

Strategies that monitor both polyadenylated and non-polyadenylated RNA species provide an unbiased account of whole transcriptome content. The most commonly used techniques rely on affinity-based counterselection schemes to deplete ribosomal RNA (rRNA) before random-primed cDNA synthesis. Although the utility of this approach has been demonstrated for various sequencing applications in prokaryotic and eukaryotic systems[3,4], rRNA depletion involves cumbersome laboratory procedures and high sample inputs. To facilitate high-throughput whole transcriptome analysis, we developed a simple procedure that would allow generation of high-complexity rRNA-depleted cDNA libraries directly from small amounts of total RNA.

This method is based on the empirical observation that heptamer and hexamer sequences are capable of sequence-specific priming of cDNA synthesis, whereas pentamers are not[5]. We reasoned that the template discrimination of these short oligonucleotides could be exploited for the selective enrichment of non-rRNA targets by computationally subtracting rRNA priming sequences from a random hexamer library. To design such a primer set, referred to here as 'not-so-random' (NSR) primers, we aligned the full repertoire of possible hexamer sequences to human cytoplasmic 18S and 28S rRNA and mitochondrial 12S and 16S rRNA transcripts. Of 4,096 input sequences, we identified 3,347 hexamers with perfect sequence matches to at least one of the rRNA filter sequences, leaving 749 hexamers to comprise the NSR primer collection. Subsequent alignment to RefSeq mRNA transcripts[6] and a sampling of short noncoding RNAs indicated that NSR hexamers encompassed sufficient sequence complexity to obtain high-density coverage of potential target sequences, with one matching start site for every six bases of template sequence on average (**Supplementary Fig. 1**).

We then devised a simple PCR-based cDNA library construction scheme to enable short read sequencing using the Illumina GA2 platform (**Fig. 1a**). The addition of heterologous 5′ tail sequences to NSR hexamers (**Supplementary Tables 1** and **2**) during oligonucleotide synthesis allowed PCR amplification and directional sequencing without an intervening adaptor ligation step (Online Methods). After optimization of reaction conditions by diagnostic quantitative PCR (QPCR) analysis, we sequenced two cDNA libraries generated from 1 μg of universal human reference (UHR) RNA using either NSR hexamer or random primer oligonucleotide pools. Analysis of over 7 million short read sequences revealed substantial enrichment of non-rRNA transcripts in the NSR-primed library when compared to the control (**Fig. 1b**). Moreover, the abundance of each rRNA transcript targeted for depletion was specifically reduced, with cumulative rRNA amounts dropping from 78% in the control to 13% in the NSR library.

To further evaluate NSR performance, we analyzed non-rRNA tag sequences obtained from one UHR library and two independently prepared libraries generated from whole brain RNA. Of 54 million 32 nucleotide (nt) reads aligning to the genome, 77% mapped unambiguously to single genomic sites. We determined mRNA representation by mapping NSR reads to ~21,000 RefSeq transcripts. Over 92% of transcripts were represented by ten or more reads in at least one of the samples queried, and 75% were represented by ten or more reads in all three libraries. Comparison of transcript levels across libraries revealed high reproducibility among

[1]Departments of Molecular Informatics and [2]Scientific Computing, Rosetta Inpharmatics LLC, a wholly owned subsidiary of Merck and Co., Inc., Seattle, Washington, USA. Correspondence should be addressed to C.D.A. (christopher_armour@merck.com).

RECEIVED 23 MARCH; ACCEPTED 9 JULY; PUBLISHED ONLINE 9 AUGUST 2009; DOI:10.1038/NMETH.1360

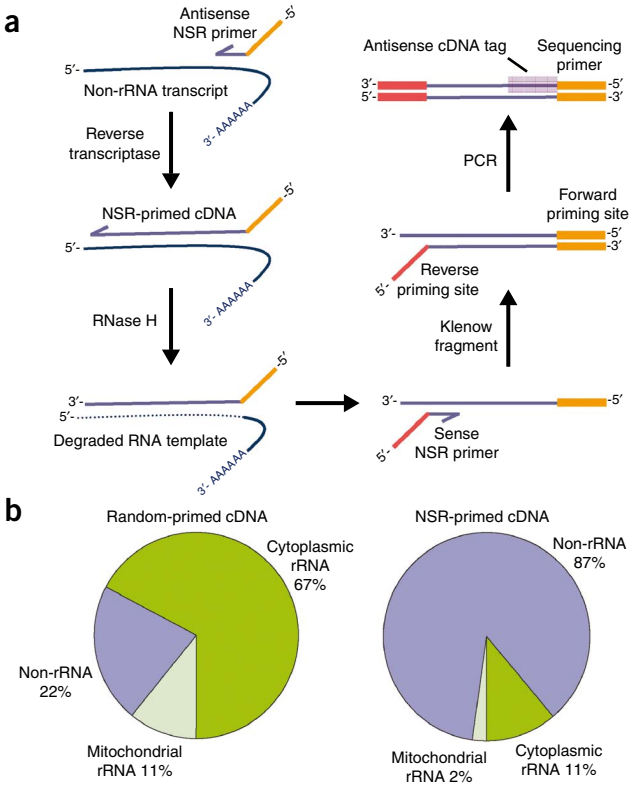

Figure 1 | Construction of NSR-primed whole transcriptome cDNA libraries. (**a**) To generate libraries, oligonucleotides with a universal 5′ tail sequence were either synthesized in the antisense orientation to initiate first-strand cDNA synthesis or made in the sense orientation for second-strand DNA synthesis. Distinct barcode sequences introduced during first- and second-strand synthesis allow strand orientation to be preserved throughout the procedure. Short tag sequences are generated in the antisense direction. (**b**) Comparison of rRNA content in random-primed and NSR-primed sequence libraries.

technical replicates ($R = 0.97$) and accurate reporting of relative abundance in brain and UHR libraries when compared to published qPCR[7] ($R = 0.93$) and RNA-seq[8] ($R = 0.90$) data (**Supplementary Fig. 2**). Moreover, greater than 99% of reads mapping to protein-coding exons were in agreement with annotated strand orientation, indicating that our directional library construction process did not result in artifactual second-strand priming from randomly assorted cDNA end sequences as has been reported for other methods[9].

NSR primer sites were distributed across whole transcript lengths at densities comparable to those observed for RNA-seq[8] (**Supplementary Fig. 3**), evidence of the robust priming potential of NSR hexamers despite the lower sequence complexity relative to random primers. However, several positional biases were evident in our data. First, our strand-specific, single end sequencing approach resulted in a noticeable coverage deficit at extreme 5′ sites, an effect not observed with nondirectional RNA-seq or paired-end expressed sequence tag data (**Supplementary Fig. 4**). Second, we observed variation in read frequency within a single transcript, and although this effect is commonly associated with random-priming techniques, NSR read coverage was less uniform than produced by cDNA libraries produced by RNA-seq protocols[8] (**Supplementary Fig. 5**). Neither the replacement of NSR sequences with random hexamers nor the fragmentation of RNA before cDNA synthesis improved coverage uniformity (data not shown). Additional investigation revealed an enrichment of universal PCR primer sequences in some genomic sites immediately upstream of NSR reads, suggesting that partial annealing of tail sequences enhances the priming efficiency at specific template sites (**Supplementary Fig. 6**). A slight increase in G+C content in the NSR hexamer site was also apparent. Despite these shortcomings, the coverage biases described here were highly reproducible and independent of expression amount, allowing robust comparison of the same transcript sites in different samples.

Next, we analyzed the abundance of known noncoding RNAs to measure the ability of NSR priming to capture non-polyadenylated transcripts. Transcripts from diverse functional classes such as small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs) and small cytoplasmic RNAs (scRNAs) were tenfold more abundant in NSR libraries than in random-primed mRNA libraries (**Supplementary Fig. 7**). Expression levels of

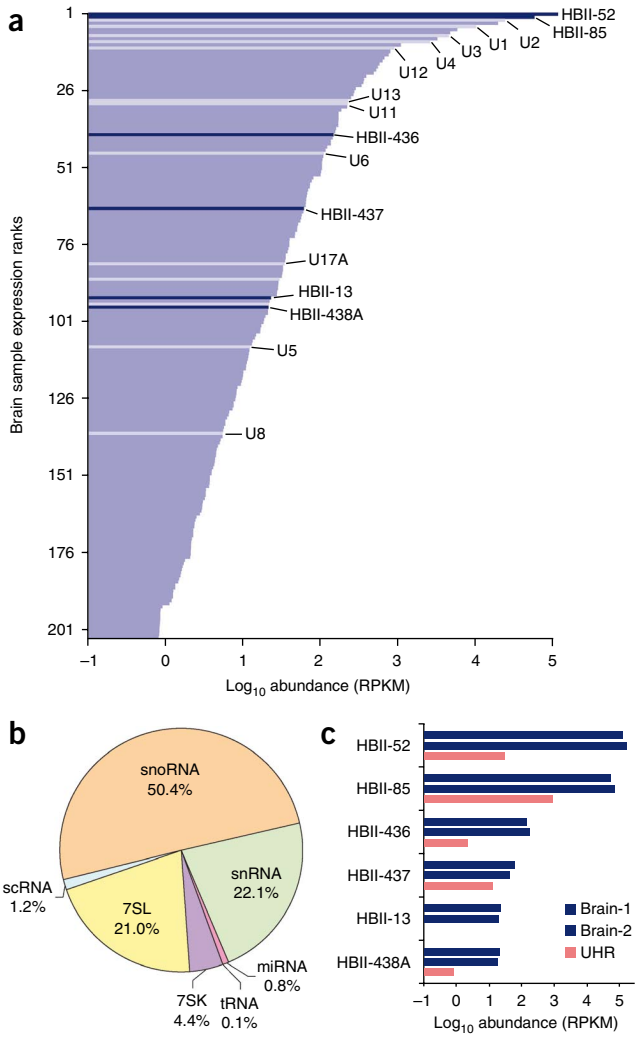

Figure 2 | Detection of poly(A)$^-$ noncoding RNAs in NSR-primed cDNA libraries. (**a**) Rank-ordered expression of noncoding RNA transcripts represented by at least two NSR tag sequences in whole brain sample. Members of the snRNA class (light blue) and known brain-specific C/D box snoRNAs (dark blue) are highlighted. RPKM, reads per kilobase per million reads. (**b**) Proportion of noncoding RNA reads mapping to selected functional classes. (**c**) Enrichment of snoRNAs encoded in the chromosome 15 Prader-Willi disease locus in whole brain sample relative to the UHR sample.

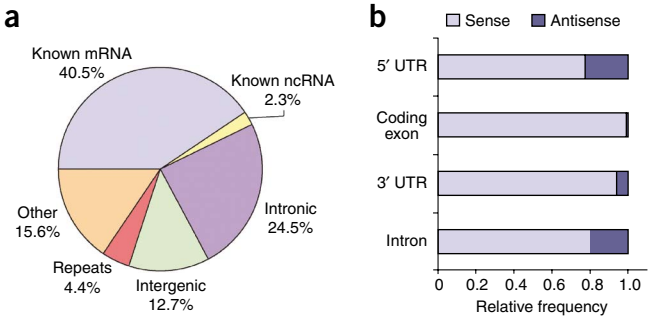

Figure 3 | Classification of global transcriptional activity. (**a**) Non-rRNA reads from UHR and whole brain libraries were assigned one of the six nonoverlapping categories shown. The mRNA, intronic and intergenic categories were defined by the genomic coordinates of UCSC known genes and include only cDNAs that map to unique locations. Reads mapping to multiple genomic sites were classified as noncoding RNA (ncRNA), repeats or other. (**b**) Strand orientation of whole brain and UHR reads across genomic subregions of UCSC known gene loci.

individual transcripts spanned more than five orders of magnitude (**Fig. 2a**) with snRNA and snoRNA families accounting for most of the transcriptional activity attributed to known noncoding RNAs (**Fig. 2b**). Over half of the transcripts detected in this study were enriched in whole brain samples (**Supplementary Table 3**) including C/D box snoRNAs located in the chromosome 15 Prader-Willi susceptibility locus previously reported to be highly expressed in the central nervous system[10] (**Fig. 2c**). In contrast, the complexity and abundance of microRNA or microRNA hairpins were severely compromised in NSR libraries: only 2% (11/471) of miRNA species queried were represented by at least five NSR tag sequences in any library. Poor miRNA representation was probably due to molecular weight constraints applied during library construction, whereas hairpin detection was likely impeded by the short half-life of precursor molecules or inefficient priming of their stable secondary structures.

To obtain an overview of global transcriptional activity, we classified NSR tag sequences from both samples into mutually exclusive categories based on current genome annotations (**Fig. 3a**). Although the majority (65%) of transcript sequences mapped to regions situated within the boundaries of previously identified protein-coding genes, over one-third of these events were transcribed from unannotated intronic regions. When we considered intronic and intergenic events together, we found that 37% of NSR reads mapped to genomic sequences not included in conventional transcript models. We also investigated the prevalence of antisense expression among University of California Santa Cruz (UCSC) Genome Browser known gene loci, which has been purported to be widespread in mammalian systems[11]. Although only a small fraction (11%) of reads mapping within gene coordinates were oriented in the antisense direction, nearly all were localized in untranslated and intronic regions (**Fig. 3b**). Anecdotal evidence suggests that at least some of these events correspond to overlapping divergently transcribed genes, which are indistinguishable by alternative nondirectional sequencing approaches (**Supplementary Fig. 8**).

Though the use of short nonrandom oligonucleotides for selective cDNA synthesis has been described previously[12], the NSR method is to our knowledge the first to use computational design criteria to deplete specific transcripts from heterogeneous total RNA mixtures. An important feature of NSR priming is the capability to quantitatively monitor polyadenylated and non-polyadenylated transcripts in parallel. Notably, this technique allows the direct observation of mammalian lincRNA expression[13] and provides an open system for the identification and characterization of additional classes of new non-polyadenylated RNAs[14]. Although here we focus on the utility of NSR priming for expression profiling in human tissue, primer design and library construction can be easily modified to accommodate other applications and model systems. For instance, the hexamer sequences we tested here have been used successfully in other mammalian systems[15], but the computational primer selection strategy outlined here can be used to design primers that are specifically tailored to more divergent organisms. Moreover, priming specificity can be refined by combining empirical data with computational selection for NSR hexamer design. NSR libraries can also be constructed by commonly used adaptor ligation methods and paired-end sequencing configurations to overcome the coverage biases described earlier. We have begun to exploit NSR-priming selectivity to mitigate the effects of high globin content for whole blood profiling, to explore global operon expression in prokaryotic systems and to monitor host-pathogen gene expression in parallel.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank T. Fare, L. Lim, D. Haynor, P. Lum and E. Schadt for valuable input, M. Biery and H. Bouzek for technical assistance, and G. Schroth and M. Schlador for advice.

AUTHOR CONTRIBUTIONS

C.D.A., J.C.C. and C.K.R. contributed to the conceptual development and experimental design. C.D.A. and C.K.R. constructed libraries and generated sequencing data. S.J. and J.D. analyzed images and managed the base calling pipeline. J.C.C., R.C., T.B., P.L. and J.K.S. performed sequence alignments and genome analysis. C.A.R. and J.M.J. provided analysis support and project management. C.D.A. and C.K.R. prepared the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

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

1. Cloonan, N. *et al. Nat. Methods* **5**, 613–619 (2008).
2. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. *Nat. Methods* **5**, 621–628 (2008).
3. Morin, R. *et al. Biotechniques* **45**, 81–94 (2008).
4. Yoder-Himes, D.R. *et al. Proc. Natl. Acad. Sci. USA* **106**, 3976–3981 (2009).
5. Raymond, C.K., Roberts, B.S., Garrett-Engele, P., Lim, L.P. & Johnson, J.M. *RNA* **11**, 1737–1744 (2005).
6. Pruitt, K.D., Tatusova, T. & Maglott, D.R. *Nucleic Acids Res.* **33**, D501–D504 (2005).
7. Shi, L. *et al. Nat. Biotechnol.* **24**, 1151–1161 (2006).
8. Wang, E.T. *et al. Nature* **456**, 470–476 (2008).
9. Wu, J.Q. *et al. Genome Biol.* **9**, R3 (2008).
10. Cavaille, J. *et al. Proc. Natl. Acad. Sci. USA* **97**, 14311–14316 (2000).
11. Katayama, S. *et al. Science* **309**, 1564–1566 (2005).
12. Gonzalez, J.M. & Robb, F.T. *J. Microbiol. Methods* **71**, 288–291 (2007).
13. Guttman, M. *et al. Nature* **458**, 223–227 (2009).
14. Amaral, P.P., Dinger, M.E., Mercer, T.R. & Mattick, J.S. *Science* **319**, 1787–1789 (2008).
15. Babak, T. *et al. Curr. Biol.* **18**, 1735–1741 (2008).

ONLINE METHODS

Oligonucleotides Accession numbers for rRNA sequences used as computational filters for NSR primer selection were obtained from NCBI: 12S (NC_001807, nt 650–1603), 16S (NC_001807, nt 1673–3230), 18S (U13369.1, nt 3657–5527) and 28S (U13369.1, nt 7935–12969). Each NSR oligonucleotide shown in **Supplementary Tables 1** and **2** was synthesized individually by Operon Biotechnologies Inc. Oligos were desalted and resuspended in water to 100 μM before pooling at equimolar concentrations. NSR hexamers were synthesized with a 5′ amplification annealing site for first-strand (5′-TCCGATCTCTN-(NSR reverse complement)-3′) and second strand (5′-TCCGATCTGAN-(NSR)-3′) priming events. A collection of random hexamers was also synthesized with these tail sequences for generation of control libraries. The same forward and reverse primers (**Supplementary Table 4**) were used for PCR amplification of both NSR- and random-primed cDNA libraries. Amplification primer sequences were designed to be compatible with sequencing on the Illumina GAII sequencing platform.

Library generation. RNA from whole brain samples was obtained from the FirstChoice human total RNA survey panel (Applied Biosystems). The UHR cell line RNA was purchased from Stratagene Corp. Total RNA was converted into cDNA using Klenow Fragment (New England Biolabs Inc.). Second-strand synthesis was carried out with 3′ to 5′ exo− Klenow Fragment (New England Biolabs Inc.). DNA was amplified using the Expand High Fidelity^{PLUS} PCR system (Roche Diagnostics Corp.).

For NSR-primed cDNA synthesis, 2 μl of 100 μM first-strand NSR primer mix was combined with 1 μl of template and 7 μl of water in a PCR strip-cap tube (Genesee Scientific Corp.). The primer-template mix was heated at 65 °C for 5 min and chilled on ice before adding 10 μl of high dNTP reverse transcription master mix (3 μl of water, 4 μl of 5× buffer, 1 μl of 100 mM DTT, 1 μl of 40 mM dNTPs and 1.0 μl of SuperScript III enzyme). The high dNTP concentration during first strand cDNA synthesis is critical to the priming specificity of the protocol. The 20 μl reverse transcription reaction was incubated at 40 °C for 30 min, 70 °C for 15 min and cooled to 4 °C. RNA template was removed by adding 1 μl of RNase H (InVitrogen Corp.) and incubating at 37 °C for 20 min, 75 °C for 15 min and cooling to 4 °C. DNA was subsequently purified using the QIAquick PCR Purification kit and eluted from spin columns with 30 μl of elution buffer (Qiagen, Inc.). For second-strand synthesis, 25 μl of purified cDNA was added to 65 μl of Klenow master mix (46 μl of water, 10 μl of 10× NEBuffer 2, 5 μl of 10 mM dNTPs, 4 μl of 5 units μl^{-1} exo$^-$ Klenow fragment; New England Biolabs, Inc.) and 10 μl of 100 μM second-strand NSR primer mix was added. The 100 μl reaction was incubated at 37 °C for 30 min and cooled to 4 °C. DNA was purified using QIAquick spin columns and eluted with 30 μl of elution buffer. For PCR amplification, 25 μl of purified second-strand synthesis reaction was combined with 75 μl of PCR master mix (19 μl of water, 20 μl of 5× Buffer 2, 10 μl of 25 mM $MgCl_2$, 5 μl of 10 mM dNTPs, 10 μl of 10 μM forward primer, 10 μl of 10 μM reverse primer, 1 μl of Expand^{PLUS} enzyme; Roche Diagnostics Corp.). Samples were denatured for 2 min at 94 °C and followed by 2 cycles of 94 °C for 10 s, 40 °C for 2 min, 72 °C for 1 min; 8 cycles of 94 °C for 10 s, 60 °C for 30 s, 72 °C for 1 min; 15 cycles of 94 °C for 15 s, 60 °C for 30 s, 72 °C for 1 min with an additional 10 s added at each cycle; and 72 °C for 5 min to polish ends before cooling to 4 °C. Double-stranded DNA was purified using QIAquick spin columns as described earlier. The scheme used for the random primer control library was the same as the one described here, except the final dNTP concentration during reverse transcription was 0.5 mM (rather than 2.0 mM) as recommended by the manufacturer (InVitrogen Corp.).

Protocol note: both the temperature and duration of cDNA synthesis appear to influence NSR-priming selectivity and read coverage. Since this study was completed, we have found that lengthening the reverse transcription reaction to 90 min at 40 °C results in more robust rRNA depletion and a modest improvement in coverage uniformity across target transcripts. Elevating the temperature to 45 °C is effective for rRNA reduction, but read coverage is distributed less evenly across non-rRNA sequences owing to tail effects.

Sequencing and read classification. Purified PCR products were used without additional manipulation to generate clusters for sequencing-by-synthesis using the Illumina GA2 platform. Single-end sequencing produced 36-nucleotide antisense reads containing a dinucleotide barcode sequence (CT) at the 5′ terminus. Novoalign (Novocraft) was used to align the reverse complements of base positions 3–34 against NCBI human genome release 36.1 (UCSC March 2006 release (hg18)) and a collection of splice junctions generated from Refseq genes, ENSEMBL genes and UCSC known genes. Predicted splice junctions from expressed sequence tags, Genscan and N-scan predictions were also considered in regions that lack coding gene models. All possible splice sites spanning up to two exon skipping events in gene or transcript models above were represented. A minimum of 5 nt overlap per flanking junction sequence was required for alignment to be considered. All reads that aligned uniquely to the genome or splice sites and redundantly mapped reads that overlap unique reads in only one genomic location were retained for subsequent analysis. Alignment to rRNA and noncoding RNA sequences was carried out using BLAST. To generate digital expression profiles, read counts were converted to RPKM (reads per kilobase per million reads) as described[2] except normalization was carried out with mapped non-rRNA reads rather than all mapped reads. For global classification, reads were aligned to the noncodingRNA and repeat databases with alignments to multiple reference sequences permitted. Reads mapping to single genomic sites were classified into mRNA, intron and intergenic categories using coordinates defined by UCSC known genes (http://genome.ucsc.edu/). Sequences that mapped to multiple genomic sequences excluding repeats or ncRNAs were binned into a separate category (other). Ribosomal RNA sequences were obtained from RepeatMasker (http://www.repeatmasker.org/) and Genbank (NC_001807). Noncoding RNA sequences were compiled from Sanger RFAM (http://www.sanger.ac.uk/Software/Rfam/), Sanger miRBASE (http://microrna.sanger.ac.uk/), snoRNABase (http://www-snorna.biotoul.fr/) and RepeatMasker. Repetitive elements were obtained from RepeatMasker. Short read sequences generated by RNA-seq[8] were processed as described above.

doi:10.1038/nmeth.1360