

# A comparative analysis of noncoding and protein-coding RNAs in the mammalian transcriptome.

Tim R. Mercer<sup>1</sup>

t.mercer@imb.uq.edu.au

Marcel E. Dinger<sup>1</sup>

m.dinger@imb.uq.edu.au

John S. Mattick<sup>1</sup>

j.mattick@imb.uq.edu.au

<sup>1</sup>ARC Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, University of Queensland, St Lucia QLD 4072, Australia

## Abstract

Recent genome-wide studies have revealed the existence of thousands of long noncoding transcripts, whose function and significance are unclear. The discovery of this hidden transcriptome has made the need to distinguish between mRNAs and ncRNAs both more pressing and more complicated [1]. We have developed an automated pipeline to annotate long noncoding RNAs [2]. This annotation has been applied to a range of transcriptomic resources, including genomic tiling arrays, cDNA and RNA-seq libraries to describe the landscape of long ncRNAs. We found long ncRNAs are the most abundant classes of transcripts, and while a significant number are found in intergenic regions, the majority surround and overlap most protein-coding genes. We intersected our annotated noncoding transcriptome with large-scale microarray or *in situ* hybridization (sourced from the Allen Brain Atlas) expression data to show these transcripts exhibit, in the main, specific expression, often correlating with distinct processes in developmental programs such as embryonic stem cell differentiation or with neuroanatomical structures within the adult mouse brain [3, 4]. We also compared the expression of long ncRNAs with associated protein-coding genes, often finding complex relationships of expression.

The complexity apparent in the organization of the mammalian genome, which contains overlapping protein-coding and noncoding transcripts, presents a challenge to unambiguously annotating RNA, and indeed we exposed a large class of ambiguous transcripts that could not be reliably categorized as coding or noncoding. For example, we identified a class of long ncRNAs that reside within 3'UTRs, which we term uaRNAs. Further analysis showed that uaRNAs may be linked to, or expressed independently to the upstream protein-coding mRNA sequence. Moreover the expression of the uaRNAs is on occasion discordant to the upstream coding sequence and subject to independent developmental regulation. More broadly, the finding that some transcripts can function both intrinsically at the RNA level and also encode proteins further confounds the distinction between coding and noncoding transcript and has prompted us to suggest the functionality of any transcript at the RNA level should not be discounted. Together this study contributes towards our shifting understanding of the nature of a gene with RNA increasingly appreciated as an intrinsically functional significant and significant component of the genomes expression [5].

**Keywords:** noncoding RNA, messenger RNA, 3' untranslated region, brain, embryonic stem cell

## References

- [1] Dinger ME, Pang KC, Mercer TR, Mattick JS. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Computational Biology* 4 (11). 2008
- [2] Dinger ME\*, Pang KC\*, Mercer TR, Crowe ML, Grimmond SM and Mattick JS. NRED: a database of long noncoding RNA expression. *Nucleic Acid Research Database issue*. 2008
- [3] Dinger ME, Amaral PP, Mercer TR, Pang KC, Bruce SJ, Gardiner BB, Askarian-Amiri ME, Ru K, Soldà G, Simons C, Sunkin SM, Crowe ML, Grimmond SM, Perkins AC, Mattick JS. Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Research* 18: 1433-1445. 2008
- [4] Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS. Specific expression of long noncoding RNAs in the adult mouse brain. *Proc Natl Acad Sci USA* 105: 716-721. 2008
- [5] Amaral PP, Dinger ME, Mercer TR, Mattick JS. The eukaryotic genome as an RNA machine. *Science* 319: 1787-1789. 2008