Supplementary Information 14

A. TU/bases

	Protein-coding	Non-coding
representative sequences (TU)	17594	15815
bases	37804289	29853506
average size	2148.70	1887.67

B. PolyA signal

	Protein-coding		Non-coding	
AATAAA	6766		3720	
ATTAAA	1880	49.14%	1155	30.83%
No polyA signal	8948	50.86%	10940	69.17%

C. PolyA tail (genome mapping: >=95%ID, >=100bp, 3'end aligned up to the end)

		Protein-coding		Non-coding	
No/short polyA tail		59	0.34%	27	0.17%
with polyA tail	not mapped	889	5.05%	785	4.96%
	# of "A"= 0- 9/20bps in genome	11983 (1,478)	68.11%	6047 (3,385)	38.24%
	# of "A"=10-14/20bps in genome	2573 (632)	14.62%	3729 (3,097)	23.58%
	# of "A"=15-20/20bps in genome	2090 (906)	11.88%	5227 (4,579)	33.05%

( ): unspliced

D. Spliced (genome mapping: >=95%ID, >=100bp)

	Protein-coding		Non-coding	
not mapped	244	1.39%	404	2.55%
spliced	14192	80.66%	4148	26.23%
unspliced/single exon	3158	17.95%	11263	71.22%

E. Sequence quality (Phred/Phrap value)

	Protein-coding		Non-coding	
0-9	303553	0.80%	246053	0.82%
10-19	791687	2.09%	608153	2.04%
20-29	1927910	5.10%	1325928	4.44%
30-39	3758918	9.94%	2612022	8.75%
40-49	4973085	13.15%	3868318	12.96%
50-59	5448644	14.41%	4496829	15.06%
60-69	4285038	11.33%	3729881	12.49%
70-79	5118264	13.54%	4337335	14.53%
80-89	3829462	10.13%	3289031	11.02%
>90	7367728	19.49%	5339956	17.89%

F. Sequence quality (genome alignment 2: >=100bp)

	Protein-coding		Non-coding	
aligned sequences	17496	99.44%	15500	98.01%
aligned bases	36614430		29080199	
matched bases	36530010	99.77%	29012867	99.77%
mismatched bases	63353	0.17%	48792	0.17%
inserted bases	21067	0.06%	18540	0.06%
deleted bases	18435		13963	

These tables show comparison of sequence information for protein-coding and non-coding transcripts in the RTPS. These analyses were done for FANTOM2 representative 33,409 sequences since not all information can be obtained for public sequences. (A) This table shows the number of representative sequences (TUs), total bases and average sizes. There are no significant differences between the two groups. (B) This table shows how many sequences have polyA signals in protein-coding and non-coding transcripts. The protein coding RNA has a larger proportion of polyA signals, whilst a comparable number of non-coding RNAs have polyA signals. (C) We have checked the 20bp of the genomic region of the 3'end of each clone that was mapped to the mouse genome with ≥ 95%ID, ≥ 100bp and aligned down to the 3' end. '# of "A" = 15-20' means that there are 15-20 bp of 'A' sequences in the 20 bp region scanned. The number of clones containing polyA in the genomic region is 2090 (11.88 %) vs 5227 (33.05%). Although a larger proportion of non-coding RNAs have the genome encoding

polyA, it unlikely these are the contamination of the genome. For we strategically excluded the sequences for further analyses if they did not contain polyA with the restriction enzyme sites used for the cloning procedure. (D) The ratio of splicing is evaluated for clones mapped to the mouse genome with ≥ 95%ID, ≥ 100bp. There is higher ratio of unspliced sequences in the non-coding RNAs. It is likely that most of the non-coding RNAs are not spliced like *Air*. (E) The distribution of Phred/Phrap scores for each base between protein coding vs non-coding transcripts. The unit is bp. Each number shows how many bps are categorized for each sequence quality. The sequence quality between the two classes did not differ. (F) This is another way of looking at the quality of sequencing by comparison with the genomic sequence of the mouse genome (MGSCv3). The genome sequence was used as a standard. This comparison used best-hit alignments with ≥ 100bp. No differences in the quality of the sequences between the two classes denying the possibility that the failure to annotate a CDS is the reason for the annotation of non-coding RNAs.