[RNA Biology 3:1, 40-48, January/February/March 2006]; ©2006 Landes Bioscience

Research Paper

Discrimination of Non-Protein-Coding Transcripts from Protein-Coding mRNA

Martin C. Frith[1,2]
Timothy L. Bailey[2]
Takeya Kasukawa[1]
Flavio Mignone[3]
Sarah K. Kummerfeld[4]
Martin Madera[4]
Sirisha Sunkara[5]
Masaaki Furuno[6]
Carol J. Bult[6]
John Quackenbush[7]
Chikatoshi Kai[1]
Jun Kawai[1,8]
Piero Carninci[1,8]
Yoshihide Hayashizaki[1,8]
Graziano Pesole[3]
John S. Mattick[2,*]

[1]Genome Exploration Research Group (Genome Network Project Core Group); RIKEN Genomic Sciences Center (GSC); RIKEN Yokohama Institute; Kanagawa, Japan

[2]Institute for Molecular Bioscience; University of Queensland; Brisbane, Australia

[3]Dipartimento di Scienze Biomolecolari e Biotecnologie; Universita di Milano; Milano, Italy

[4]MRC Laboratory of Molecular Biology; Cambridge, UK

[5]The Institute for Genomic Research; Rockville, Maryland USA

[6]Mouse Genome Informatics Consortium; The Jackson Laboratory; Bar Harbor, Maine USA

[7]Department of Statistics and Computational Biology; Dana-Farber Cancer Institute; Boston, Massachusetts USA

[8]Genome Science Laboratory; Discovery Research Institute; RIKEN Wako Institute; Saitama, Japan

*Correspondence to: John S. Mattick; Institute for Molecular Bioscience; University of Queensland; QLD 4072, Austrailia; Tel.: +61.7.3346.2079; Fax: +61.7.3346.2111; Email: j.mattick@imb.uq.edu.au

Received 03/24/06; Accepted 04/03/06

Previously published online as a *RNA Biology* E-publication:
http://www.landesbioscience.com/journals/rnabiology/abstract.php?id=2789

KEY WORDS

transcriptome, proteome, ncRNA, mRNA, bioinformatics

ACKNOWLEDGEMENTS

See page 47.

ABSTRACT

Several recent studies indicate that mammals and other organisms produce large numbers of RNA transcripts that do not correspond to known genes. It has been suggested that these transcripts do not encode proteins, but may instead function as RNAs. However, discrimination of coding and non-coding transcripts is not straightforward, and different laboratories have used different methods, whose ability to perform this discrimination is unclear. In this study, we examine ten bioinformatic methods that assess protein-coding potential and compare their ability and congruency in the discrimination of non-coding from coding sequences, based on four underlying principles: open reading frame size, sequence similarity to known proteins or protein domains, statistical models of protein-coding sequence, and synonymous versus non-synonymous substitution rates. Despite these different approaches, the methods show broad concordance, suggesting that coding and non-coding transcripts can, in general, be reliably discriminated, and that many of the recently discovered extra-genic transcripts are indeed non-coding. Comparison of the methods indicates reasons for unreliable predictions, and approaches to increase confidence further. Conversely and surprisingly, our analyses also provide evidence that as much as ~10% of entries in the manually curated protein database Swiss-Prot are erroneous translations of actually non-coding transcripts.

INTRODUCTION

A fundamental goal of genomics is to catalog all of the expressed products encoded in a genome. Annotations of mainly protein-coding genes have been published simultaneously with genome sequences, including those of human and mouse,[1,2] but since then evidence for large numbers of novel transcripts, many of which do not seem to encode proteins, has appeared from large-scale cDNA sequencing and interrogation of genome tiling arrays.[3-7] Establishing which transcripts encode proteins and which do not is essential for comprehending the repertoire of genomic products, but to our knowledge reliable criteria for making this distinction have not been developed, simply as a consequence of the earlier and still general assumption that most RNAs encode proteins and the fact that the large numbers of non-coding transcripts were unexpected. Moreover, different studies have used different methods to identify protein-coding sequences, and reciprocally to suggest that others are non-coding, but the reliability and congruence of these different methods have not been compared and assessed.

It is not easy to prove definitively that a transcript encodes a protein. One approach is to synthesize the protein artificially, raise an antibody against it, and use the antibody to probe whether the protein is expressed in vivo. However, this method is time-consuming and expensive. Definitive proof that a transcript does not encode a protein, on the other hand, is well-nigh impossible, since the protein might only be expressed in very rare circumstances, or the observed transcript (a cDNA for example) might represent a fragment of a longer transcript that has protein-coding capacity elsewhere.

Lacking definitive practical experimental methods, a battery of bioinformatics criteria that assess the protein-coding potential of transcripts can be applied. If most of these criteria agree, especially those that are predicated on different features or characteristics of protein-coding sequences, it would increase the confidence that the sequence in question is or is not protein-coding. To evaluate this approach, and to assess the congruence of different methods when applied to large data sets that appear to contain large numbers of non-coding sequences as well as reliably known protein-coding sequences, we applied ten computational methods to the 102,801 FANTOM mouse cDNA sequences and the Swiss-Prot database.

Open reading frame-based methods. Proteins are encoded in open reading frames (ORFs) comprised of sequences of triplet codons beginning with ATG and ending with a stop codon. Since three out of sixty-four codons encode stops, ORFs much greater than 100 codons are unlikely to appear by chance in non-coding sequences of average base composition. On the other hand, protein-coding ORFs are often, if not usually, larger than 100 codons, and so the presence of an ORF ≥100 codons is frequently taken as a rough indication of the likelihood that the sequence is, or is not, protein-coding. (In this study, we only identified full ORFs bounded by ATGs and stop codons. Relaxing this criterion to find truncated ORFs as well does not increase the concordance of this method with the others (data not shown), suggesting that the dataset has few truncated ORFs, and the sensitivity increase from finding these cases is outweighted by a decrease in specificity.)

A more sophisticated approach is *mTRANS* which strives to account for experimental errors in cDNA sequences that could corrupt open-reading frames either by introducing false frame shifts or stop codons. Sequencing errors are corrected by aligning the cDNA against the genome and constructing a "virtual cDNA" from the genomic exons. cDNA truncation and intron retention are assessed by comparing the cDNA to other transcribed sequences in the EST database. Presence of downstream A-rich potential internal priming sites and susceptibility to non-sense mediated decay are also considered as they reduce the mRNA-like characteristics of the cDNA in question. Each cDNA receives a cumulative score based on ORF size and evidence for or against the various kinds of experimental error, and cDNAs scoring above a threshold are considered coding (M. Furuno, in preparation).

Protein and domain similarity-based methods. The most common method is *BLASTX* in which transcripts (cDNAs) are translated in all three reading frames and compared to a database of known proteins.[8] If a statistically significant similarity is found, it may be deduced that the sequence encodes a related protein, or is perhaps a pseudogene. Simple repeats are filtered using the SEG program, since they violate the statistical assumptions and lead to spurious matches.[9] Some proteins in the database are partly derived from interspersed repeats and align to many cDNAs. Such alignments cannot be taken as protein-coding evidence, since most interspersed repeats are transposon fossils and do not encode proteins. Hence we also filtered interspersed repeats using RepeatMasker.[10]

The *rsCDS* method was developed to identify coding regions of known genes or those that are highly similar to known genes even if there are frameshift errors.[11] It is based on FASTY alignments of the cDNAs against a database of known proteins.

It is also possible to search for similarity to known protein domains, which is useful where the sequence may share a common domain type, but not the entire sequence, with known proteins in the database, as distantly related proteins often share functionally or structurally similar domains with subtle sequence signatures. *Pfam* is a collection of statistical models (hidden Markov models—HMMs) of such domains, which may enable identification of related proteins that are too divergent to be picked up by alignment methods such as BLAST.[12] An alternative is SUPERFAMILY, which is a library of HMMs of domain superfamilies according to the SCOP classification of protein structures.[13]

Methods based on statistical models of mRNA. *ESTScan* employs a hidden Markov model of mRNA sequences, which allows for sequencing errors and truncations.[14] The optimal (Viterbi) path through the model, which may or may not include a protein-coding region, is found for each cDNA. An alternative method, *DIANA-EST*, uses a combination of artificial neural networks and statistics for the characterization of coding regions within transcript sequences, allowing for sequencing errors.[15]

Comparative methods utilizing synonymous/non-synonymous substitution rates. We investigated two methods of this type: *CSTminer* employs a rapid alignment method (BLAT) to find homologs of each cDNA in a database of nucleotide sequences (in this case, the human, rat and dog genomes and mammalian mRNAs in RefSeq). The alignments are then recalculated using a more careful technique, and regions showing an excess of synonymous vs. non-synonymous substitutions at the nucleotide level and of conservative vs. non-conservative replacements at the amino acid level are flagged as coding.[16,17]

CRITICA, originally a bacterial gene finder, is a hybrid method that combines comparative analysis with statistical analysis of coding sequences.[18] Initial coding predictions are derived from regions with high synonymous versus non-synonymous substitution rates in nucleotide alignments. Sequence statistics of predicted coding regions are then tallied, and combined with the comparative evidence to repredict coding regions. This statistical analysis and reprediction is iterated several times.

Other methods. Our choice of methods was inevitably somewhat ad hoc, but the ten methods outlined above is a diverse and representative selection of approaches that examine different lines of evidence regarding the coding or non-coding status of transcripts. Other promising techniques include bacterial gene finders such as GeneMark,[19] dictionary-based protein identification[20] and assessment of whether predicted secondary structures of putative proteins resemble those of real proteins.[21] Comparative gene finders such as TwinScan and SGP have become popular recently, but these methods tackle the harder problem of identifying spliced protein-coding genes in DNA, and are not directly suited to analyzing RNA.[22,23]

In recent years algorithms for identifying conserved RNA structures have appeared, such as QRNA, RNAZ, DDBRNA and RNA profile.[24-27] These methods are sometimes portrayed as non-coding RNA gene predictors. However, this is misleading: they identify conserved elements of RNA secondary structure that can and do occur in mRNA as well as ncRNA.[24-27] Moreover, there is no reason why non-protein coding RNAs must have conserved secondary structures, especially if their function is based on primary sequence interactions with other RNAs, as is the case for natural antisense transcripts. Thus RNA structure analysis is not considered in this study.

The FANTOM mouse cDNA collection. The FANTOM projects obtained cDNA sequences for 102,801 mouse transcripts from many tissues and developmental stages, using the cap-trapper technique to enrich for full-length transcripts and aggressive subtraction/normalization to get rare transcripts.[4,28,29] The sequences are accurate but not error-free: 91.7% of bases have phred/phrap score ≥30, corresponding to an error rate ≤1 in 1,000[29] (FANTOM 2 data). There is a low contamination rate, with 0.26% *E. coli* DNA[30] (FANTOM 2 data). The average length is 2,146 bases. Each cDNA was manually annotated through the efforts of hundreds of curators and thousands of person-hours.[31] Protein-coding regions were annotated with the assistance of a graphical interface showing the results of various automatic predictions, including (for FANTOM 3): CRITICA, longest ORF, mTRANS, rsCDS, and Pfam. Two further CDS predictors employed during FANTOM 3, DECODER and CombinerCDS, are not considered here, because as used in FANTOM

they annotate a CDS in every clone. Here we also assess how these manual annotations compare to the ten computational methods.

cDNA sequences may be subject to artifacts such as misorientation, truncation/internal priming, and immaturity (incomplete removal of introns), so that protein-coding ORFs may be disrupted or missing. The FANTOM sequences are rarely misoriented, since 67,401 of them use the canonical GT-AG splice signal and only 199 of them exhibit the reverse signal CT-AC, and even these may be genuine.[5] A significant number were annotated as truncated and/or immature, based on their genomic coordinates relative to other transcripts. It also appears that many non-coding transcripts were internally primed at A-rich sequences,[32] but at least some of these are fragments of longer non-coding transcripts.[33] Moreover, large-scale characterization of transcript endpoints reveals more variability than previously appreciated, and suggests that many cDNAs containing partial ORFs are in fact full-length.[4] So the extent of truncation and immaturity artifacts is unclear.

The methods used here do not really allow for gross artifacts, with the notable exception of mTRANS. This study focuses on coding/non-coding discrimination rather than the orthogonal problem of artifact detection. Artifacts are entirely a function of the technology used to obtain the cDNA sequences, and associated quality control procedures, whereas coding/non-coding discrimination is a fundamental biological question. However, we have employed a conservative procedure to eliminate truncation artifacts (see Materials and Methods: Full length support), in order to derive reliable sets of coding and non-coding FANTOM sequences.

There are some transcripts that pose unique problems for coding/non-coding discrimination, because they resemble protein-coding sequence, but cannot be translated in the usual way owing to reading frame disruptions. These include transcribed pseudogenes, and more generally we term these transcripts pseudo-messenger RNA.[34] These transcripts may themselves be functional as regulatory RNAs,[35] and are best detected by special-purpose methods, as described elsewhere.[34]

MATERIALS AND METHODS

FANTOM cDNA sequences. The RepeatMasked sequences were obtained from ftp://fantom.gsc.riken.jp/FANTOM3/repeats/fantom3_total103k_r2.masked.fasta.gz.

BLASTX. The sequences were searched against the UniRef90 database (downloaded 9-1-2004) using blastall 2.2.10 with options -p blastx -e 0.01 -m9 -S1-a2 -U T.

Pfam. The sequences were searched against Pfam 12.0 using estwisedb with options -sum -quiet -pfam -dnas. Reverse-strand predictions were ignored.

SUPERFAMILY. The sequences were translated in all six frames and searched against the SUPERFAMILY database. cDNAs with *E*-values ≤0.01 were predicted as coding. Reverse-strand predictions were ignored.

ESTScan. ESTScan 2.0b was applied to the sequences with options -d -500 -i -500 -M MkTables/mm.smat (hs.smat for the human sequences). Reverse-strand predictions were ignored.

DIANA-EST. DIANA-EST was applied to the sequences as follows: est <fasta-seq-filename> -1 120. DIANA was trained as follows. All the human proteins whose starts were sequenced at the amino-acid level were manually collected and full-length mRNAs for these proteins were retrieved. Three-quarters of these were used for the extraction of the training data and one quarter for the extraction of the test data. To extract the training and test data, the EMBL database was searched for EST entries corresponding to these cDNAs. It was then determined if these ESTs were coding/non-coding, strand on which it was coding and whether it had start and stop, based on the alignment of the ESTs to the cDNAs. Finally, these ESTs were used for training and determining the parameters of the EST analysis.

CSTminer. The cDNAs' homologous sequences in the human, rat and dog genomes and mammalian RNAs in RefSeq were detected using BLAT. (The human sequences were aligned to the mouse instead of the human genome.) Then alignments of Conserved Sequence Tags (CSTs) were constructed by BLAST (wordsize = 7, expect = 1E-5). The protein coding potential of each CST was assessed through the computation of a coding potential score (CPS). CSTs were labeled as follows: CST_HCOD: high confidence coding CST, CPS≥7.67; CST_LCOD: low confidence coding CST, 6.41≤CPS≤7.67; CST_NCOD: non-coding CST, CPS < 6.41; CST_GREY: unlabelable CST (>95% sequence identity). For the binary analysis (Fig. 1A), clones classified as CST_HCOD or CST_LCOD were counted as coding, and clones classified as CST_NCOD, CST_GRAY, or None were counted as non-coding. For the trinary analysis (Fig. 1B), clones classified as CST_NCOD were counted as non-coding, clones classified as CST_HCOD that did not have a highest-scoring prediction on the reverse strand were counted as coding, and all other cases were counted as undefined.

CRITICA. The sequences were aligned with homologs in the NCBI nt database (downloaded 19-1-2004) using discontiguous MegaBLAST with options -e 1e-4 -D 1 -F "m D" -U T -J F -f T -t 18 -W 11 -A 5 0 -q -2 -G 5 -E 2. We modified CRITICA 1.05 b to handle large files, and analyzed the alignments using iterate-critica with options -no-sdscores -fraction-coding = 0.5 -genetic-code = 1 -frameshift-threshold = 10. Reverse strand predictions were ignored.

FANTOM manual annotations. The annotation results were obtained from ftp://fantom.gsc.riken.jp/fantomdb/3.0/anndata.txt.gz. cDNAs with a cds_location other than No CDS, 5'UTR or 3'UTR were considered coding; the remainder were considered non-coding.

Full-length support. The 5' and 3' ends of each FANTOM cDNA were verified by support from independent RNA sequences, 5' and 3' ESTs, CAGE tags (~20 nt sequences adjacent to the 5' cap), and GIS and GSC ditags (paired ~20 nt sequences from both ends of full-length transcripts). These datasets have been described elsewhere,[4] and are available from the FANTOM 3 website (http://fantom3.gsc.riken.jp/). cDNAs were not considered unless they mapped unambiguously to the genome, and had at most 5 nt of sequence unaligned at the end being considered. 3' ends upstream of A-rich sequences (>10 As in 20 nt immediately downstream) were discarded, since they may reflect internal priming of longer transcripts. A 5' end was verified by meeting any of the following criteria: 2 CAGE tag starts within ±15 nt, 3 CAGE tag starts within ±60 nt, 4 CAGE tag starts within ±100 nt, 1 GSC ditag start within ±0 nt, 2 GSC ditag starts within ±50 nt, 1 GIS ditag start within ±15 nt, 1 RIKEN 5'EST start within ±3 nt, 2 RIKEN 5' EST starts within ±100 nt, 1 non-RIKEN 5' EST start within ±2 nt, 2 non-RIKEN 5' EST starts within ±100 nt, 1 other FANTOM cDNA start within ±25 nt, 1 non-RIKEN RNA start within ±50 nt. A 3' end was verified by meeting any of the following criteria: 1 GSC ditag end within ±0 nt, 2 GSC ditag ends within ±50 nt, 1 GIS ditag end within ±15 nt, 1 RIKEN 3' EST end within ±2 nt, 2 RIKEN 3' EST ends within ±100 nt, 1 non-RIKEN EST end within ±7 nt, 2 non-RIKEN 3' EST ends within ±100 nt, 1 other FANTOM cDNA end within ±25 nt, 1 non-RIKEN RNA end within ±50 nt. These numbers were chosen based on the amount of each type of supporting evidence, such that each criterion has less than 1 chance in 1000 of occurring if the sites are randomly scattered across the genome. Sequences from the same clone (e.g., ESTs) were not counted as independent evidence. Both orientations were considered for GSC ditags. This method does not account for systematic truncation errors (other than internal priming).

RESULTS

The output of multiple coding/non-coding discrimination methods applied to the 102,801 FANTOM cDNAs is voluminous and complex: we initially consider simplified binary yes/no predictions listed in Table S1 and summarized graphically in Figure 1A. Each column corresponds to one method and each row to one combination of binary outcomes: red means coding and blue noncoding. A certain level of consistency among the methods is immediately apparent: 30,209 sequences (29%) are predicted as coding (24%) or non-coding (5%) by all methods, 58,967 (57%) by all but (up to)

one method, and 79,371 (77%) by all but (up to) two methods. The predictions of each pair of methods are positively correlated (Table 1), and the highest correlations are often but not always among methods based on the same underlying principle. The strongest correlation is between CRITICA and BLASTX, which are not only methods that rely on quite different criteria (sequence similarity and the pattern of synonymous/non-synonymous substitutions, respectively) but are also the two methods that agree most often with the majority vote of all methods. We suggest that the consensus of the methods reflects the true coding status of a transcript, which implies that the individual accuracy of each method is revealed by its level of concordance with the others. However, there appear to be sizeable discrepancies among the methods, casting doubt on the coding status of many cDNAs. Fortunately, most of the discrepancies stem from simple, predictable reasons for unreliability in some of the predictions.

SUPERFAMILY and Pfam make the fewest coding predictions, but their coding predictions usually agree with most other methods. This outcome is not surprising, since Pfam and SUPERFAMILY do not cover all protein domains that exist. So their coding predictions are reliable, but their non-coding predictions are not since we might be dealing with proteins that lack well-characterized domains, of which there remain many. These methods do occasionally make coding predictions that are contradicted by most other methods, some of which can be explained by dubious matches to repetitive sequences. For example Pfam identifies a protamine P1 domain in a low-complexity region of clone 4930432I21 and a Gag domain in an LTR element of clone I1C0023K22, while SUPERFAMILY finds a ribonuclease H-like domain in an endogenous retroviral sequence of clone 1200015M12. These hits to retrotransposons are understandable because active retrotransposons encode functional proteins, and their many inactive copies contain disabled homologs, i.e., pseudogenes.

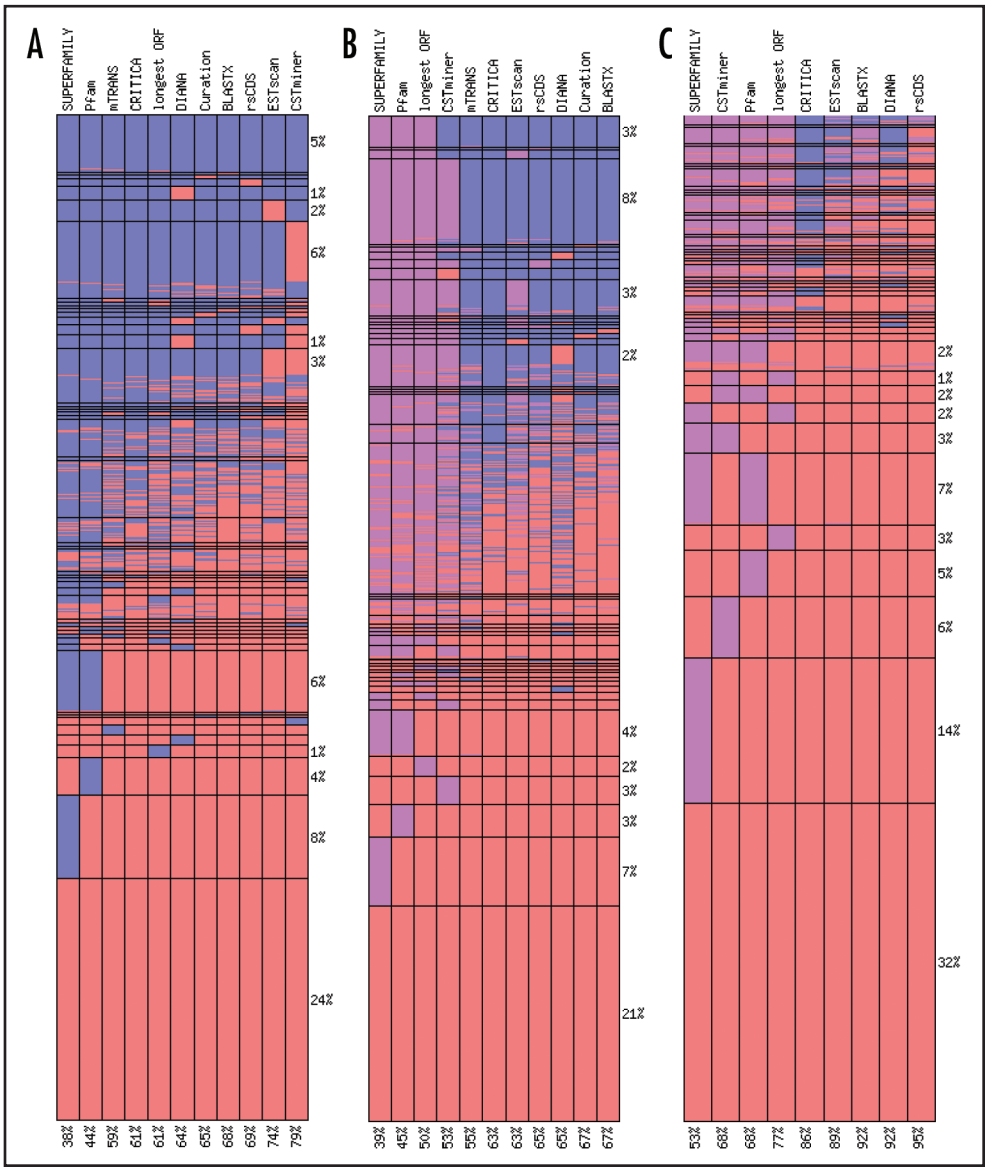

Figure 1. Comparisons of methods to discriminate coding from non-coding transcripts. Each row corresponds to one set of outcomes from each method: red is "coding", blue is "non-coding", and purple is "no confident prediction". The height of each row is proportional to the number of sequences with that set of outcomes. The percentage of sequences predicted as coding by each method is indicated at the base of each column. (A) Eleven sets of binary coding/non-coding predictions for the 102,801 FANTOM cDNAs. (B) Eleven sets of trinary coding/non-coding predictions for the 92,122 FANTOM cDNAs that remain after excluding pseudo-messenger RNAs. (C) Seven sets of trinary coding/non-coding predictions for 1,078 human mRNA sequences linked to the Swiss-Prot protein database.

CSTminer, ESTScan and DIANA make a substantial number of coding predictions that are not supported by most other methods. For ESTScan and CSTminer, these unsupported predictions have markedly lower scores than average (Fig. 2A and B), so their agreement with the other methods is actually better than it appears from the binary results. Unsupported DIANA predictions also have below-average scores (Fig. 2C), but there is not such a clean separation. If a high-scoring ESTScan or CSTminer prediction contradicts most other methods, further investigation may be warranted, but low CSTminer scores and ESTScan scores below about 300 can be regarded as weak and unreliable predictions.

CSTminer actually produces complex predictions (including high-confidence coding, low-confidence coding, non-coding, and undefined) that are greatly simplified in Figure 1A. Of the 6,208 clones designated coding by CSTminer alone, 3,852 (62%) are low-confidence predictions. In comparison, only 19,407 clones in total (19%) are low-confidence coding. In a further 1,287 out of 6,208 cases (21%), the strongest coding prediction is on the reverse strand; the corresponding figure for all clones is 8,777 (9%). These cases may reflect non-coding transcripts that are cis-antisense to protein-coding exons.

For such a crude method, the longest ORF shows a surprisingly high level of concordance with the others. To our knowledge, this is the first evidence that the traditional 100 aa threshold is a relatively good choice. Almost all the coding predictions made by longest ORF alone or by longest ORF and mTRANS alone (i.e., that are not supported by other methods) are just above the threshold of 100 codons (Fig. 2E): these are almost certainly non-coding since we expect a significant number of ORFs slightly greater than 100 codons to occur by chance in such a large dataset. Conversely, some cDNAs predicted as non-coding by longest ORF but coding by most other methods exhibit frameshifts or truncations that disrupt

Table 1 **Correlation coefficients among eleven sets of coding/non-coding predictions for the 102,801 FANTOM cDNAs**

	Longest ORF	mTRANS	BLASTX	rsCDS	Pfam	SUPERFAMILY	ESTscan	DIANA	CSTminer	CRITICA	Curation
Longest ORF		0.709	0.661	0.626	0.525	0.464	0.533	0.523	0.360	0.684	0.676
mTRANS	0.709		0.661	0.623	0.551	0.477	0.555	0.578	0.381	0.703	0.692
BLASTX	0.661	0.661		0.772	0.557	0.513	0.616	0.573	0.395	0.822	0.779
rsCDS	0.626	0.623	0.772		0.523	0.461	0.564	0.527	0.365	0.723	0.705
Pfam	0.525	0.551	0.557	0.523		0.599	0.476	0.499	0.332	0.623	0.561
SUPERFAMILY	0.464	0.477	0.513	0.461	0.599		0.424	0.449	0.293	0.561	0.498
ESTscan	0.533	0.555	0.616	0.564	0.476	0.424		0.524	0.332	0.663	0.592
DIANA	0.523	0.578	0.573	0.527	0.499	0.449	0.524		0.328	0.639	0.575
CSTminer	0.360	0.381	0.395	0.365	0.332	0.293	0.332	0.328		0.440	0.386
CRITICA	0.684	0.703	0.822	0.723	0.623	0.561	0.663	0.639	0.440		0.772
Curation	0.676	0.692	0.779	0.705	0.561	0.498	0.592	0.575	0.386	0.772	
% agreement with majority vote	87.8	88.7	92.8	89.6	79.4	74.1	85.7	85.0	75.5	94.8	92.4

the reading frame, and others encode known proteins that are shorter than 100 aa such as Cox8a and Cox6b. As expected, mTRANS makes quite similar predictions to longest ORF. Coding predictions by mTRANS that are unsupported by most other methods have borderline scores (Fig. 2F), indicating that mTRANS scores below about 500 are less reliable.

We expected BLASTX to behave similarly to Pfam and SUPERFAMILY, i.e., to identify proteins with homologs in the UniRef90 database accurately, but to miss many proteins without known homologs. In fact we observe the opposite: BLASTX almost always makes a coding prediction when most other methods do so, but it also makes a number of coding predictions that are not supported by most other methods. This suggests that almost all mouse proteins have recognizable homologs in the database. The database includes proteins derived from FANTOM 1 and 2, but not FANTOM 3.

Given the BLASTX *E*-value threshold of 0.01, some false positive coding predictions are expected. BLASTX coding predictions that are unsupported by most other methods do have higher *E*-values than average (Fig. 2G), but many of them are still rather significant with *E*-values $<10^{-8}$, so this is not the whole explanation. Some cases appear to be artifacts of the *E*-value calculation, e.g., a nine-codon region of clone 1700122E12 aligns with thirteen separate regions of a repetitive viral protein, receiving a collective *E*-value of 10^{-12}. Other cases are variants of protein-coding genes that include very small portions of the protein-coding region, e.g., clone 0610023I12 includes 26 codons of TFF-I interacting peptide 20: the correct classification in such cases is not clear.

Other false positives arise from dubious entries in the protein database. For example, clone 1110021L09 aligns to just one protein, Q8BTD6, with 100% identity. This protein entered the database via an earlier annotation of this clone, so this is a self-referential alignment without predictive value. Among 701 clones predicted as coding by BLASTX and rsCDS, and non-coding by all others except curation and CSTminer, 242 (35%) only have alignments with 100% identity. There might also be hits to false proteins from other species with less than 100% identity. This problem reveals that extremely good alignments with "known" proteins are not necessarily reliable indicators of protein-coding capacity.

As expected, rsCDS makes similar predictions to BLASTX. The lack of repeat masking in this method leads to false coding predictions: in the cDNAs predicted as coding by rsCDS but non-coding by most other methods, the predicted coding regions are largely covered by repeats (Fig. 2D).

CRITICA shows the highest degree of concordance with the other methods, and makes the fewest recognizable mistakes. A few false positives are expected using the default p value cutoff of 10^{-4}, and indeed the nine cDNAs predicted as coding by CRITICA alone have borderline p values. Some false negatives are apparent among the few cDNAs predicted as non-coding by CRITICA but coding by most other methods. For example clone 1200010C05 encoding Pex7 has a single nucleotide insertion in the protein-coding region near the ATG, and there is no alternative in-frame ATG. Although CRITICA initially finds arbitrary regions with high synonymous/non-synonymous rates, it only reports those that it can extend to an open reading frame beginning with an ATG and ending with a stop codon. Thus it fails on 1200010C05 but makes a coding prediction for clone 2510005J23, which encodes the same protein but lacks the frameshift.

The curated results generally agree well with the consensus of the computational methods, but there are a few outliers, which often look like human errors. For instance, a 14 aa protein was annotated in clone 0710008P21, but all the computational methods agree that it is non-coding and it overlaps the 5'UTR of another gene, *Slc20a1*. Some variation in annotation quality is almost inevitable in distributed projects, and this comparison offers a powerful way to flag suspect annotations for rechecking.

By examining discrepancies between the methods, we have found situations where coding or non-coding predictions are unreliable. This knowledge allows us to obtain more realistic, trinary predictions from the methods: "coding", "non-coding", or "no confident prediction". We assign "no confident prediction" in the following situations: ESTScan score <300, mTRANS score <500, longest ORF <50 codons (including <100 codons), rsCDS predictions >20% repetitive, BLASTX hits with 100% identical alignments only, and low confidence CSTminer predictions (see Supplementary Materials). Lack of SUPERFAMILY or Pfam hits is also treated as "no confident prediction" rather than "non-coding". Finally, we identified and removed 10,679 pseudo-messenger RNAs.[34] These modifications reveal greater concordance among the methods (Fig. 1B, Table 2, Table S2): 60,453 clones (66%) are predicted as coding or non-coding by all methods that make a confident prediction, 79,177 (86%) by all but one method, and 87,340 (95%) by all but two methods.

Since it is inconvenient to apply so many computational tools, we investigated how well the results can be reproduced by a majority vote of just

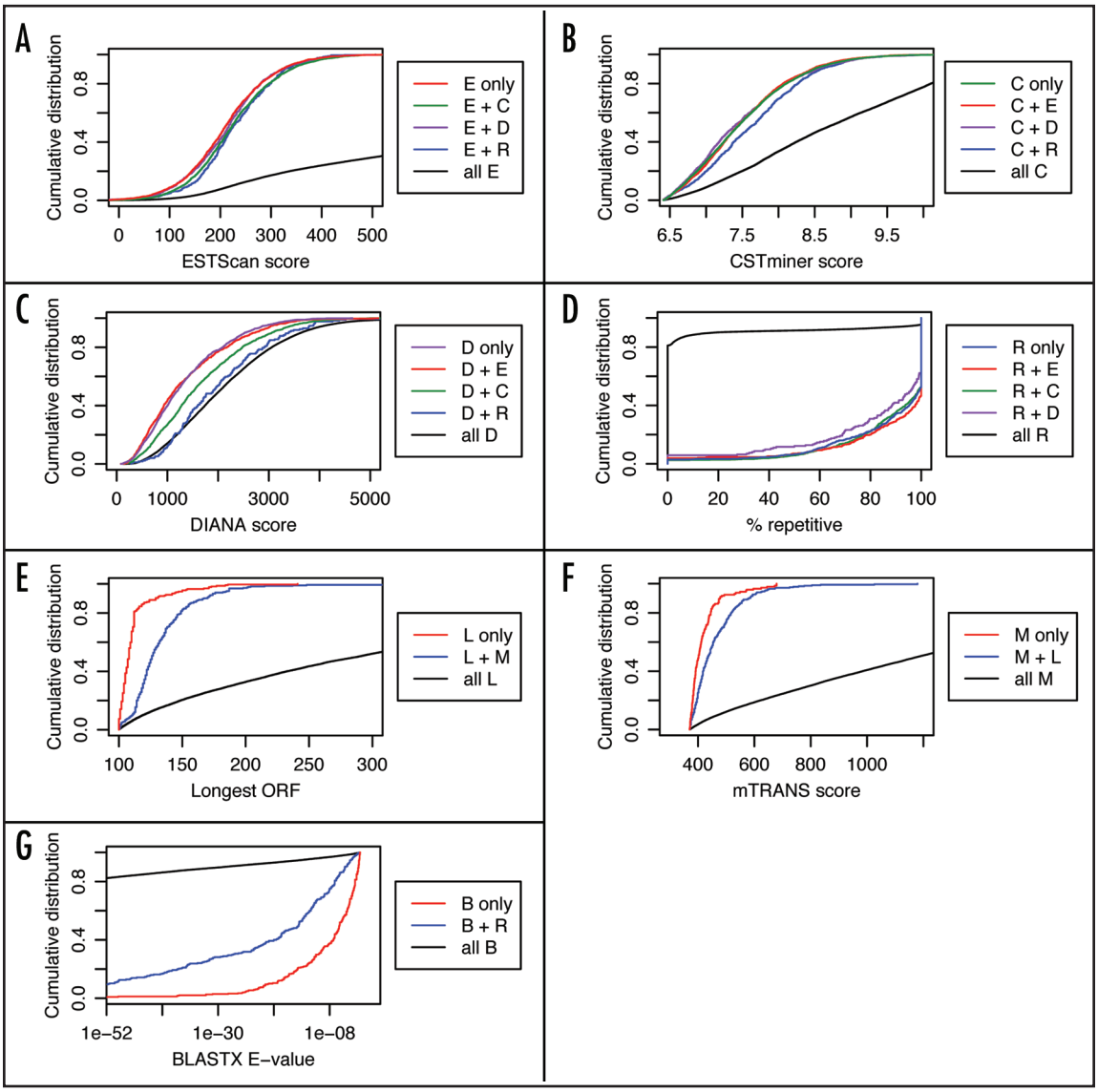

Figure 2. Analysis of minority protein-coding predictions. In these plots the y-value indicates the proportion of cDNAs with quantity labeled on the x-axis < the x-value. E, ESTScan; C, CSTminer; D, DIANA; R, rsCDS; L, longest ORF; M, mTRANS; B, BLASTX. "Only," cases predicted to be coding by this method but non-coding by all other methods; "+," cases predicted to be coding by these two methods but non-coding by all other methods; "all," all coding predictions by this method.

three methods. Up to 96% agreement with the majority vote of all methods can be achieved, for example using CRITICA, BLASTX, and mTRANS (Table 3). There are many combinations of methods that perform close to optimally, although they tend to include CRITICA, and tend to involve three different underlying principles. Remarkably, methods such as Pfam that individually have a low concordance with the majority can appear in some of the best trios. This implies that Pfam supplies useful information complementary to the other methods, and can be used to increase the confidence that a sequence is protein-coding, but not to increase the confidence that it is not. Nonetheless, CRITICA on its own reflects the consensus of the methods almost as well as the best trio (Table 1).

Reliable discrimination of FANTOM coding and non-coding transcripts. In order to identify coding and non-coding transcripts reliably, it is necessary to consider that some cDNAs may be artifactual. To address this problem, we utilized the massive datasets of RNA 5' and 3' sequence tags from FANTOM 3 and other sources to find independent experimental support for the 5' and 3' end of each cDNA (see Materials and Methods: Full-length support). In total, 41,025 cDNAs (of which 2,541 are pseudo-messenger RNAs) have support for both ends, and thus are likely to represent real, full-length transcripts. The remaining cDNAs are not necessarily artifactual: evidence for their being full-length is simply lacking. Combining the full-length data with the coding/non-coding discrimination results gives conservative numbers of coding and non-coding transcripts (Table 2). The proportion of non-coding RNAs with full-length support is smaller, but still several thousand.

Non-coding transcripts in Swiss-prot. We also applied the methods (apart from mTRANS, which is tuned for high-throughput mouse cDNAs) to 1,078 human transcripts encoding proteins listed in the Swiss-Prot database,[36] partly to test our approach on known protein-coding transcripts, and partly to look for incorrectly annotated non-coding transcripts. As expected, the methods indicate a clear coding consensus for most transcripts, but they also suggest that about 10% of the sequences listed in Swiss-Prot are in fact non-coding (Fig. 1C, Table S3). Tellingly, the coding and non-coding predictions are highly correlated between the methods, which enhances their credibility. Note that BLASTX and rsCDS rely on comparison to a database of known proteins, which includes Swiss-Prot, so they naturally have high rates of coding predictions. The transcripts with a non-coding consensus invariably correspond to poorly characterized Swiss-Prot proteins. For example,

Table 2 **Numbers of FANTOM cDNAs predicted as coding and non-coding with different degrees of confidence**

Prediction	Predicted by	All cDNAs*	Confidently full-length cDNAs*
Coding	All methods†	44,722 (49%)	28,686 (75%)
	All methods† but one	53,396 (58%)	32,860 (85%)
	All methods† but two	57,385 (62%)	34,112 (89%)
Non-coding	All methods†	15,731 (17%)	1,799 (5%)
	All methods† but one	25,781 (28%)	2,949 (8%)
	All methods† but two	30,059 (33%)	3,497 (9%)
Total cDNAs		92,122	38,484

*Excluding 10,679 pseudo-messenger RNAs. †Excluding methods that make no confident prediction.

Table 3 **Agreement between the majority vote of three coding/non-coding discrimination methods and the majority vote of all eleven methods: top ten trios**

Method 1	Method 2	Method 3	% Agreement with Majority Vote of all Methods
mTRANS	BLASTX	CRITICA	96.4
mTRANS	rsCDS	CRITICA	96.3
mTRANS	CRITICA	Curation	96.3
DIANA	CRITICA	Curation	96.2
rsCDS	CRITICA	Curation	96.2
ESTScan	CRITICA	Curation	96.1
Longest ORF	CRITICA	Curation	96.1
BLASTX	CRITICA	Curation	96.0
Longest ORF	BLASTX	CRITICA	96.0
BLASTX	Pfam	Curation	95.9

sequence AK024977 is said to encode a protein of unknown function called C21orf97, but this transcript overlaps the 3'UTR of another gene (*PDXK*) and there is little evidence for its coding status. The methods also highlighted eight testis transcript Y mRNAs, which are described as "apparently non-coding" in the original publication.[37] Hence the methods reveal non-coding RNAs that have been mistakenly annotated as protein-coding.

DISCUSSION

The concordance of computational methods based on diverse underlying principles allows coding and non-coding transcripts to be discriminated with high confidence. The main drawback of using a battery of computational methods is the inconvenience of obtaining and applying all of these algorithms. Fortunately, the results can be well approximated using just three methods based on different principles, or even just using CRITICA, although confirmation from independent methods would add confidence. We recommend using CRITICA for the initial analysis, and increasingly more methods if more confidence is desired. We found minor flaws in some of the methods that could be fixed to increase accuracy and convenience; for example CRITICA could be made more robust to frameshift errors, and its current implementation does not scale well to large datasets.

Our results do not provide a definitive picture of which methods are "better" than others: to do so was not the aim of this study. This is partly because we did not control for differences in various parameters of the methods: for instance, it would be interesting to compare CRITICA and CSTminer using the same homolog detection step for both. More fundamentally, the methods are optimized for different aspects of annotation, e.g., rsCDS was designed to annotate known proteins accurately in the presence of frameshifts. The aim of this study was to assess whether a battery of diverse methods can be used to discriminate coding from non-coding transcripts, not to judge which method is best.

This study confirms that the FANTOM cDNA set divides into about two thirds protein-coding and one third non-coding (Table 2).[4] When considering only transcripts with experimental support for both ends, the proportion of non-coding RNA decreases but remains non-negligible (Table 2). This result is slightly at odds with tiling array experiments, which suggest that there are at least as many non-coding as coding transcripts.[5,6] However, noncoding RNA may well be underrepresented in the FANTOM set because they are expressed at low levels, too short (e.g., miRNA) or too long (e.g., AIR) for the cloning procedure,[33] or because they lack polyA tails.[5]

We cannot rule out the possibility that some transcripts designated non-coding by all the methods encode highly unusual proteins. Short, rapidly evolving proteins with unusual codon usage patterns could be invisible to all the methods. Extremely short proteins, e.g., <10 a.a., would not provide enough statistical signal to be detected by any method. Conversely, the presence of a clear protein-coding region does not guarantee that a transcript gets translated: there could be splice variants that combine protein-coding exons with translation inhibition signals.

We have shown that the protein databases contain a small fraction of erroneous translations from non-coding sequences, and this must explain some proportion of so-called orphan proteins that lack similarity to any well-characterized protein. The approach presented here could be used to screen protein databases systematically for such errors.

This reclassification of previously known transcripts from coding to non-coding, and the large numbers of transcripts that are reliably predicted not to encode proteins (i.e., are not mRNAs), also suggest a reevaluation of the role of non-coding RNA in biology, particularly eukaryotic biology.[38,39] While it is sometimes suggested that these non-coding RNAs may be the consequence of "transcriptional noise", i.e., background transcription from illegitimate and irrelevant promoters, there is little evidence that this actually occurs, and some evidence to the contrary,[40] as opposed to the stochastic firing of legitimate promoters that has been well-documented in a number of systems and is also referred to as transcriptional noise.[41,42] It is also worth noting that many putative non-coding RNAs appear to show differential expression in different tissues,[32] suggesting that their expression is purposeful. Longer functional non-coding RNAs (as opposed to shorter ones like sno- and miRNAs) are not highly conserved at the primary sequence level,[43] suggesting that these

sequences, which presumably have mainly regulatory functions, evolve more fluidly than protein-coding sequences that are tightly constrained by strict analog structure-function relationships.[44,45]

The answer to the question posed in the introduction is that we can indeed distinguish coding from non-coding transcripts with high confidence, using a battery of computational analyses, provided the transcript sequences are accurate and full-length. This need not have been the case: it is conceivable that the methods might have had a high degree of discordance with no simple explanation, or that careful manual annotation considering published experimental data might have revealed frequent incorrect predictions. Strikingly, the consensus of the methods turned out to be more believable than two resources that we had considered using as "gold standards" to test our approach: the FANTOM manual annotations and the respected, manually curated Swiss-Prot database. This establishes a more principled and rigorous approach to answering the most basic question about the transcripts produced from a genome: whether or not they encode proteins.

Acknowledgements

We are grateful to Jinfeng Liu for providing the Swiss-Prot dataset, and to Julian Gough, Ken Pang and Pär Engstrom helpful advice. This work was funded by a Research Grant for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government to Y.H., a Research Grant for Advanced and Innovational Research Program in Life Science to Y.H., a grant of the Genome Network Project from the Ministry of Education, Culture, Sports, Science and Technology, Japan to Y.H., a Grant for the Strategic Programs for R&D of RIKEN to Y.H., and Research Grants for Preventure Program C of the Japan Science and Technology Agency (JST) to Y.H. J.S.M. and T.L.B. are supported by the Queensland State Government and the Australian Research Council. M.C.F. is a University of Queensland Postdoctoral Fellow, and J.S.M. is a Federation Fellow of the Australian Research Council.

Note

A description of the mTRANS algorithm is provided in the additional file entitled mTRANS-F3. In addition, there are three supplementary data files: Table S1: Eleven sets of binary coding/non-coding predictions for the 102,801 FANTOM cDNAs. Table S2: Eleven sets of trinary coding/non-coding predictions for the 92,122 FANTOM cDNAs that remain after excluding pseudo-messenger RNAs. Table S3: Nine sets of trinary coding/non-coding predictions for 1,078 human mRNA sequences linked to the Swiss-Prot protein database.

This Supplemental Material can be found at:
www.landesbioscience.com/journals/rnabiology/supplement/frith-supdata.zip

References

1. Finishing the euchromatic sequence of the human genome. Nature 2004; 431:931-45.
2. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. Initial sequencing and comparative analysis of the mouse genome. Nature 2002; 420:520-62.
3. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, et al. Global identification of human transcribed sequences with genome tiling arrays. Science 2004; 306:2242-6.
4. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. The transcriptional landscape of the mammalian genome. Science 2005; 309:1559-63.
5. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 2005; 308:1149-54.
6. Frith MC, Pheasant M, Mattick JS. Genomics: The amazing complexity of the human transcriptome. Eur J Hum Genet 2005; 13:894-7.
7. Stolc V, Gauhar Z, Mason C, Halasz G, van Batenburg MF, Rifkin SA, Hua S, Herreman T, Tongprasit W, Barbano PE, et al. A gene expression map for the euchromatic genome of *Drosophila melanogaster*. Science 2004; 306:655-60.
8. Gish W, States DJ. Identification of protein coding regions by database similarity search. Nat Genet 1993; 3:266-72.
9. Altschul SF, Boguski MS, Gish W, Wootton JC. Issues in searching molecular sequence databases. Nat Genet 1994; 6:119-29.
10. Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. 1996-2004, (www.repeatmasker.org).
11. Furuno M, Kasukawa T, Saito R, Adachi J, Suzuki H, Baldarelli R, Hayashizaki Y, Okazaki Y. CDS annotation in full-length cDNA sequence. Genome Res 2003; 13:1478-87.
12. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, et al. The Pfam protein families database. Nucleic Acids Res 2004; 32:D138-41.
13. Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol 2001; 313:903-19.
14. Lottaz C, Iseli C, Jongeneel CV, Bucher P. Modeling sequencing errors by combining Hidden Markov models. Bioinformatics 2003; 19:II103-12.
15. Hatzigeorgiou AG, Fiziev P, Reczko M. DIANA-EST: A statistical analysis. Bioinformatics 2001; 17:913-9.
16. Mignone F, Grillo G, Liuni S, Pesole G. Computational identification of protein coding potential of conserved sequence tags through cross-species evolutionary analysis. Nucleic Acids Res 2003; 31:4639-45.
17. Castrignano T, Canali A, Grillo G, Liuni S, Mignone F, Pesole G. CSTminer: A web tool for the identification of coding and noncoding conserved sequence tags through cross-species genome comparison. Nucleic Acids Res 2004; 32:W624-7.
18. Badger JH, Olsen GJ. CRITICA: Coding region identification tool invoking comparative analysis. Mol Biol Evol 1999; 16:512-24.
19. Hirosawa M, Ishikawa K, Nagase T, Ohara O. Detection of spurious interruptions of protein-coding regions in cloned cDNA sequences by GeneMark analysis. Genome Res 2000; 10:1333-41.
20. Shibuya T, Rigoutsos I. Dictionary-driven prokaryotic gene finding. Nucleic Acids Res 2002; 30:2710-25.
21. Liu J, Gough J, Rost B. Distinguishing protein-coding from non-coding RNAs through support vector machines. PLoS Genetics 2006; 2:e29.
22. Korf I, Flicek P, Duan D, Brent MR. Integrating genomic homology into gene structure prediction. Bioinformatics 2001; 17(Suppl 1):S140-8.
23. Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW, Guigo R. Comparative gene prediction in human and mouse. Genome Res 2003; 13:108-17.
24. Rivas E, Eddy SR. Noncoding *RNA* gene detection using comparative sequence analysis. BMC Bioinformatics 2001; 2:8.
25. Washietl S, Hofacker IL, Stadler PF. Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci USA 2005; 102:2454-9.
26. di Bernardo D, Down T, Hubbard T. ddbRNA: Detection of conserved secondary structures in multiple alignments. Bioinformatics 2003; 19:1606-11.
27. Pavesi G, Mauri G, Stefani M, Pesole G. RNAProfile: An algorithm for finding conserved secondary structure motifs in unaligned RNA sequences. Nucleic Acids Res 2004; 32:3258-69.
28. Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, Arakawa T, Hara A, Fukunishi Y, Konno H, et al. Functional annotation of a full-length mouse *cDNA* collection. Nature 2001; 409:685-90.
29. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length *cDNAs*. Nature 2002; 420:563-73.
30. Carninci P, Waki K, Shiraki T, Konno H, Shibata K, Itoh M, Aizawa K, Arakawa T, Ishii Y, Sasaki D, et al. Targeting a complex transcriptome: The construction of the mouse full-length *cDNA* encyclopedia. Genome Res 2003; 13:1273-89.
31. Maeda N, Kasukawa T, Oyama R, Gough J, Frith MC, Engstrom PG, Lenhard B, Aturaliya RN, Batalov S, Beisel KW, et al. Transcript annotation in FANTOM 3: mouse gene catalog based on physical CDNA clones. PLoS Genetics 2006; 2:e62.
32. Ravasi T, Suzuki H, Pang KC, Katayama S, Furuno M, Okunishi R, Fukuda S, Ru K, Frith MC, Gongora MM, et al. Experimental validation of the regulated expression of large numbers of noncoding *RNAs* from the mouse genome. Genome Res 2006; 16:11-9.
33. Furuno M, Pang K, Ninomiya N, Fukuda S, Frith MC, Bult C, Kai C, Kawai J, Carninci P, Hayashizaki Y, et al. Clusters of internally primed transcripts reveal novel long noncoding RNAs. PLoS Genetics 2006; 2:e37.
34. Frith MC, Wilming LG, Forrest AR, Kawaji H, Tan SL, Wahlestedt C, Bajic VB, Kai C, Kawai J, Carninci P, et al. Pseudo-messenger RNA: Phantoms of the transcriptome. PLoS Genetics 2006; 2:e23.
35. Hirotsune S, Yoshida N, Chen A, Garrett L, Sugiyama F, Takahashi S, Yagami K, Wynshaw-Boris A, Yoshiki A. An expressed pseudogene regulates the messenger-*RNA* stability of its homologous coding gene. Nature 2003; 423:91-6.
36. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003; 31:365-70.

37. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 2003; 423:825-37.
38. Mattick JS. Challenging the dogma: The hidden layer of nonprotein-coding RNAs in complex organisms. Bioessays 2003; 25:930-9.
39. Mattick JS. RNA regulation: A new genetics? Nat Rev Genet 2004; 5:316-23.
40. Dahary D, Elroy-Stein O, Sorek R. Naturally occurring antisense: Transcriptional leakage or real overlap? Genome Res 2005; 15:364-8.
41. Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. Science 2002; 297:1183-6.
42. Blake WJ, M KA, Cantor CR, Collins JJ. Noise in eukaryotic gene expression. Nature 2003; 422:633-7.
43. Pang KC, Frith MC, Mattick JS. Rapid evolution of noncoding RNAs: Lack of conservation does not mean lack of function. Trends Genet 2006; 22:1-5.
44. Mattick JS, Makunin IV. Small regulatory RNAs in mammals. Hum Mol Genet 2005; 14:R121-32.
45. Mattick JS, Makunin IV. Non-codong RNA. Hum Mol Genet 2006; 15:R17-29.