The Transcriptional Landscape of the Mammalian Genome

P. Carninci, *et al.*

Science **309**, 1559 (2005);
DOI: 10.1126/science.1112014

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. Copyright 2005 by the American Association for the Advancement of Science; all rights reserved. The title *Science* is a registered trademark of AAAS.

only LKS stations in NH), are fully consistent with this assumption, particularly for the tropical stations. In the extratropics there are only four daytime-only stations so the MSU test is less meaningful, but the two independent estimates do agree within 0.03°C per decade.

To illustrate the importance of the heating bias, we have computed its impact δ_{sol} on the trends at LKS stations. The LKS f factors, unhomogenized trends, and trends adjusted only for solar heating are given for the middle troposphere and lower stratosphere in Table 2. In the stratosphere, our δ_{sol} is similar to the total adjustments by LKS and others, with trends moving closer to those from MSU (*13*). At the tropical tropopause (of relevance to stratospheric water vapor), δ_{sol} is somewhat smaller than LKS's. In the troposphere, however, δ_{sol} is much larger than previous adjustments. Indeed, the tropical trend with this adjustment (0.14°C per decade over 1979 to 1997) would be consistent with model simulations driven by observed surface warming, which was not true previously (*1*). One independent indication that the solar-adjusted trends should be more accurate is their consistency across latitude belts: for the period 1979 to 1997, the spread of values fell by 70% in the lower stratosphere and 25% in the troposphere.

Though this is encouraging, our confidence in these nighttime trends is still limited given that other radiosonde errors have not been addressed. SH trends from 1958 to 1997 seem unrealistically high in the troposphere, especially with the δ_{sol} adjustment, although this belt has by far the worst sampling. Previous homogenization efforts typically produced small changes to mean tropospheric trends, which could mean that other error trends cancel out δ_{sol} in the troposphere. In our judgment, however, such fortuitous cancellation of independent errors is unlikely compared to the possibility that most solar artifacts were previously either missed or their removal negated by other, inaccurate adjustments. To be detected easily, a shift must be large and abrupt, but δ_{sol} was spread out over so many stations (79% of stations during 1979 to 1997 and 90% during 1959 to 1997 experienced ΔT trends significant at 95% level), at such modest levels, and of sufficient frequency at many stations that many may have been undetectable. Most important, jumps in the difference between daytime and nighttime monthly means would be detectable at only a few tropical stations because most lack sufficient nighttime data. In any case, we conclude that carefully extracted diurnal temperature variations can be a valuable troubleshooting diagnostic for climate records, and that the uncertainty in late–20th century radiosonde trends is large enough to accommodate the reported surface warming.

References and Notes

1. B. D. Santer *et al.*, *Science* **309**, 1551 (2005); published online 11 August 2005 (10.1126/science.1114867).
2. J. K. Angell, *J. Clim.* **16**, 2288 (2003).
3. J. R. Lanzante, S. A. Klein, D. J. Seidel, *J. Clim.* **16**, 241 (2003).
4. D. E. Parker *et al.*, *Geophys. Res. Lett.* **24**, 1499 (1997).
5. P. W. Thorne *et al.*, *J. Geophys. Res.*, in press.
6. D. H. Douglass, B. D. Pearson, S. F. Singer, P. C. Knappenberger, P. J. Michaels, *Geophys. Res. Lett.* **31**, L13207 (2004).
7. D. J. Gaffen *et al.*, *Science* **287**, 1242 (2000).
8. D. E. Parker, D. I. Cox, *Int. J. Climatol.* **15**, 473 (1995).
9. M. Free, D. J. Seidel, *J. Geophys. Res.* **110**, D07101 (2005).
10. J. K. Luers, R. E. Eskridge, *J. Appl. Meteorol.* **34**, 1241 (1995).
11. I. Durre, T. C. Peterson, R. S. Vose, *J. Clim.* **15**, 1335 (2002).
12. L. Haimberger, "Homogenization of radiosonde temperature time series using ERA-40 analysis feedback information," Tech. Rep. European Center for Medium Range Weather Forecasting (2005), ERA-40 Project Report Series 23.
13. D. J. Seidel *et al.*, *J. Clim.* **17**, 2225 (2004).
14. P. R. Krishnaiah, B. Q. Miao, *Handbook of Statistics*, P. R. Krishnaiah, C. R. Rao, Eds. (Elsevier, New York, 1988), vol. 7.
15. M. Free *et al.*, *Bull. Am. Meteorol. Soc.* **83**, 891 (2002).
16. W. J. Randel, F. Wu, in preparation.
17. D. J. Seidel, M. Free, J. Wang, *J. Geophys. Res.* **110**, D09102 (2005).
18. A. Dai, K. E. Trenberth, T. R. Karl, *J. Clim.* **12**, 2451 (1999).
19. S. Chapman, R. S. Lindzen, *Atmospheric Tides* (D. Reidel, Norwell, MA, 1970).
20. D. R. Easterling *et al.*, *Science* **277**, 364 (1997).
21. D. J. Gaffen, R. J. Ross, *J. Clim.* **12**, 811 (1999).
22. W. J. Randel *et al.*, *Science* **285**, 1689 (1999).
23. K. N. Liou, T. Sasamori, *J. Atmos. Sci.* **32**, 2166 (1975).
24. R. E. Eskridge *et al.*, *Bull. Am. Meteorol. Soc.* **76**, 1759 (1995).
25. H. Riehl, *Tropical Meteorology* (McGraw Hill, New York, 1954).
26. S. C. Sherwood, *Geophys. Res. Lett.* **27**, 3525 (2000).
27. J. R. Christy, R. W. Spencer, W. B. Norris, W. D. Braswell, D. E. Parker, *J. Atmos. Oceanic Technol.* **20**, 613 (2003).
28. T. Sasamori, J. London, *J. Atmos. Sci.* **23**, 543 (1966).
29. Data files and further information on methods, uncertainty, and interpretation of our results are available as supporting material on *Science* Online.

30. S.C.S. thanks J. Risbey and K. Braganza for useful discussions. This work was supported by the National Oceanic and Atmospheric Administration Climate and Global Change Program award NA03OAR4310153, and by NSF ATM-0134893.

Supporting Online Material
www.sciencemag.org/cgi/content/full/1115640/DC1
Methods
SOM Text
Data files
References and Notes

2 June 2005; accepted 27 July 2005
Published online 11 August 2005;
10.1126/science.1115640
Include this information when citing this paper.

The Transcriptional Landscape of the Mammalian Genome

The FANTOM Consortium* and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group)*

This study describes comprehensive polling of transcription start and termination sites and analysis of previously unidentified full-length complementary DNAs derived from the mouse genome. We identify the 5′ and 3′ boundaries of 181,047 transcripts with extensive variation in transcripts arising from alternative promoter usage, splicing, and polyadenylation. There are 16,247 new mouse protein-coding transcripts, including 5154 encoding previously unidentified proteins. Genomic mapping of the transcriptome reveals transcriptional forests, with overlapping transcription on both strands, separated by deserts in which few transcripts are observed. The data provide a comprehensive platform for the comparative analysis of mammalian transcriptional regulation in differentiation and development.

The production of RNA from genomic DNA is directed by sequences that determine the start and end of transcripts and splicing into mature RNAs. We refer to the pattern of transcription control signals, and the transcripts they generate, as the transcriptional landscape. To describe the transcriptional landscape of the mammalian genome, we combined full-length cDNA isolation (*1*) and 5′- and 3′-end sequencing of cloned cDNAs, with new cap-analysis gene expression (CAGE) and gene identification signature (GIS) and gene signature cloning (GSC) ditag technologies for the identification of RNA and mRNA sequences corresponding to transcription initiation and termination sites (*2*, *3*). A detailed description of the data sets generated, mapping strategies, and depth of coverage of the mouse transcriptome is provided in supporting online material (SOM) text 1 (Tables 1 and 2). We have identified paired initiation and termination sites, the boundaries of independent transcripts, for 181,047 independent transcripts in the transcriptome (Table 3). In total, we found 1.32 5′ start sites for each 3′ end and 1.83 3′ ends for each 5′ end (table S1). Based on these data, the number of transcripts is at least one order of magnitude larger than the estimated 22,000 "genes" in the mouse genome (*4*) (SOM text 1), and the large majority of transcriptional units have alternative promoters and polyadenylation sites. The use of genome tiling arrays (*5*–*7*) in humans has also implied that the number of transcripts encoded by the genome is at least 10 times as great as the number of "genes." To extend the mouse data, two HepG2 CAGE libraries, one constructed with random primers and the other with oligo-dT primers, were combined to produce 1,000,000 CAGE tags. Mapping of these tags to the human genome identified the likely promoters and transcriptional starting site (TSS) of many of the gene models identified by tiling array, also called transfrags (*5*), and clearly indicates that the same level of transcriptional diversity occurs in humans as in mice (table S2).

The mapping of ends of transcripts can be used to identify the genomic span of the primary transcript. Figure 1A shows length distributions of the predicted genomic regions spanned by mouse cDNAs showing a bimodal distribution and compares them with one peak for unspliced and another for spliced RNAs. At the upper end of the distribution are candidate mega transcripts (transcripts originating from genomic regions in the order of millions of base pairs). For example, we located six pairs of genome signature cloning (GSC) ditags to RIKEN clone ID 9330159J16 and corresponding RIKEN expressed sequence tags (ESTs). This clone encodes for a previously unidentified large transcript that is similar to a protein tyrosine phosphatase, receptor type D (accession no. BC086654), the genomic structure of which has not been previously reported (*8*). The predicted mRNA is 2475 base pairs in length but spans a genomic region of 2.2 megabases (Mb).

We previously coined the term transcriptional units (TUs), which groups mRNAs that share at least one nucleotide and have the same genomic location and orientation (*9*). However, TU fusions can join unrelated and differently annotated transcripts (SOM text 2). Therefore, we define a transcriptional framework (TK) as grouping transcripts that share common expressed regions as well

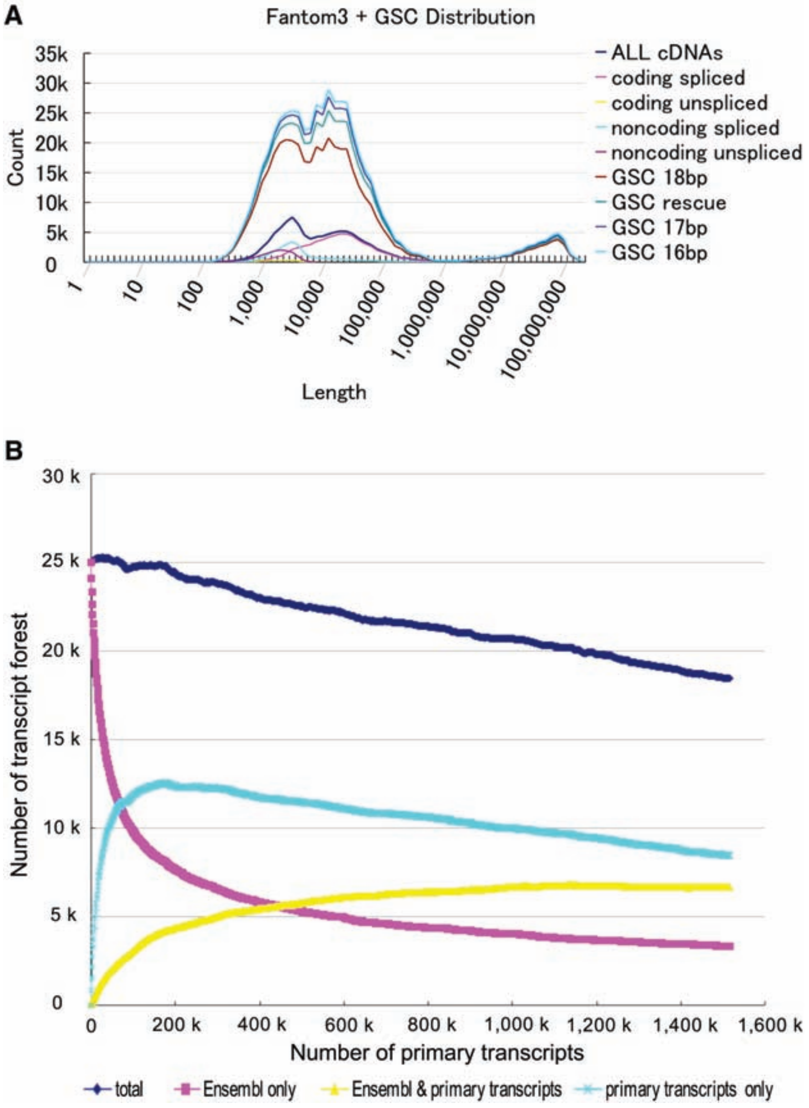


Fig. 1. Genome-transcriptome relation. (**A**) Genome span covered by full-length cDNA and GIS/GSC ditags shows similar distribution with two main peaks. Ditags mapping follows the same distribution profile at various mapping thresholds, with a minimum around 2 to 2.5 Mb. Mapping events above this genomic span are nonspecific. Count displays the number of events in the size interval. (**B**) Asymptotic unit collapse. Due to extensive overlap of the genome, transcripts overlap to the extent that they collapse to a few GFs. Simulating addition of ditags shows the collapsing rate of the known annotated genes into 9976 elements only. Primary transcripts only, GFs identified by GSC ditags only; Ensembl only, GFs produced by the 3332 Ensembl-only annotated transcripts; total, the total number of GFs.

The FANTOM Consortium:
P. Carninci,† T. Kasukawa, S. Katayama, J. Gough,† M. C. Frith,† N. Maeda, R. Oyama, T. Ravasi,† B. Lenhard,† C. Wells,† R. Kodzius, K. Shimokawa, V. B. Bajic,† S. E. Brenner, S. Batalov, A. R. R. Forrest, M. Zavolan, M. J. Davis, L. G. Wilming, V. Aidinis, J. E. Allen, A. Ambesi-Impiombato, R. Apweiler, R. N. Aturaliya, T. L. Bailey, M. Bansal, L. Baxter, K. W. Beisel, T. Bersano, H. Bono, A. M. Chalk, K. P. Chiu, V. Choudhary, A. Christoffels, D. R. Clutterbuck, M. L. Crowe, E. Dalla, B. P. Dalrymple, B. de Bono, G. Della Gatta, D. di Bernardo, T. Down, P. Engstrom, M. Fagiolini, G. Faulkner, C. F. Fletcher, T. Fukushima, M. Furuno, S. Futaki, M. Gariboldi, P. Georgii-Hemming, T. R. Gingeras, T. Gojobori, R. E. Green, S. Gustincich, M. Harbers, Y. Hayashi, T. K. Hensch, N. Hirokawa, D. Hill, L. Huminiecki, M. Iacono, K. Ikeo, A. Iwama, T. Ishikawa, M. Jakt, A. Kanapin, M. Katoh, Y. Kawasawa, J. Kelso, H. Kitamura, H. Kitano, G. Kollias, S. P. T. Krishnan, A. Kruger, S. K. Kummerfeld, I. V. Kurochkin, L. F. Lareau, D. Lazarevic, L. Lipovich, J. Liu, S. Liuni, S. McWilliam, M. Madan Babu, M. Madera, L. Marchionni, H. Matsuda, S. Matsuzawa, H. Miki, F. Mignone, S. Miyake, K. Morris, S. Mottagui-Tabar, N. Mulder, N. Nakano, H. Nakauchi, P. Ng, R. Nilsson, S. Nishiguchi, S. Nishikawa, F. Nori, O. Ohara, Y. Okazaki, V. Orlando, K. C. Pang, W. J. Pavan, G. Pavesi, G. Pesole, N. Petrovsky, S. Piazza, J. Reed, J. F. Reid, B. Z. Ring, M. Ringwald, B. Rost, Y. Ruan, S. L. Salzberg, A. Sandelin, C. Schneider, C. Schönbach, K. Sekiguchi, C. A. M. Semple, S. Seno, L. Sessa, Y. Sheng, Y. Shibata, H. Shimada, K. Shimada, D. Silva, B. Sinclair, S. Sperling, E. Stupka, K. Sugiura, R. Sultana, Y. Takenaka, K. Taki, K. Tammoja, S. L. Tan, S. Tang, M. S. Taylor, J. Tegner, S. A. Teichmann, H. R. Ueda, E. van Nimwegen, R. Verardo, C. L. Wei, K. Yagi, H. Yamanishi, E. Zabarovsky, S. Zhu, A. Zimmer, W. Hide, C. Bult,† S. M. Grimmond, R. D. Teasdale, E. T. Liu,† V. Brusic, J. Quackenbush,† C. Wahlestedt,† J. S. Mattick,† D. A. Hume†
RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group):
C. Kai, D. Sasaki, Y. Tomaru, S. Fukuda, M. Kanamori-Katayama, M. Suzuki,† J. Aoki, T. Arakawa, J. Iida, K. Imamura, M. Itoh, T. Kato, H. Kawaji, N. Kawagashira, T. Kawashima, M. Kojima, S. Kondo, H. Konno, K. Nakano, N. Ninomiya, T. Nishio, M. Okada, C. Plessy, K. Shibata, T. Shiraki, S. Suzuki, M. Tagami, K. Waki, A. Watahiki, Y. Okamura-Oho, H. Suzuki, J. Kawai.
General Organizer:
Y. Hayashizaki†‡

*Affiliations can be found on *Science* Online (available at www.sciencemag.org/cgi/content/full/309/5740/1559/DC1).

†These authors are core authorship members.

‡To whom correspondence should be addressed. E-mail: yosihide@gsc.riken.jp

as splicing events, TSS, or termination events (SOM text 1).

TKs can be clustered together into transcript forests (TFs), genomic regions that are transcribed on either strand without gaps. TFs encompass 62.5% of the genome (table S1) and are separated by regions devoid of transcription, or transcription deserts. With the inclusion of GSC tags in addition to full-length cDNA and paired EST sequences, the estimated total number of transcript forests is 18,461,which will collapse further with increasing depth of coverage (Fig. 1B).

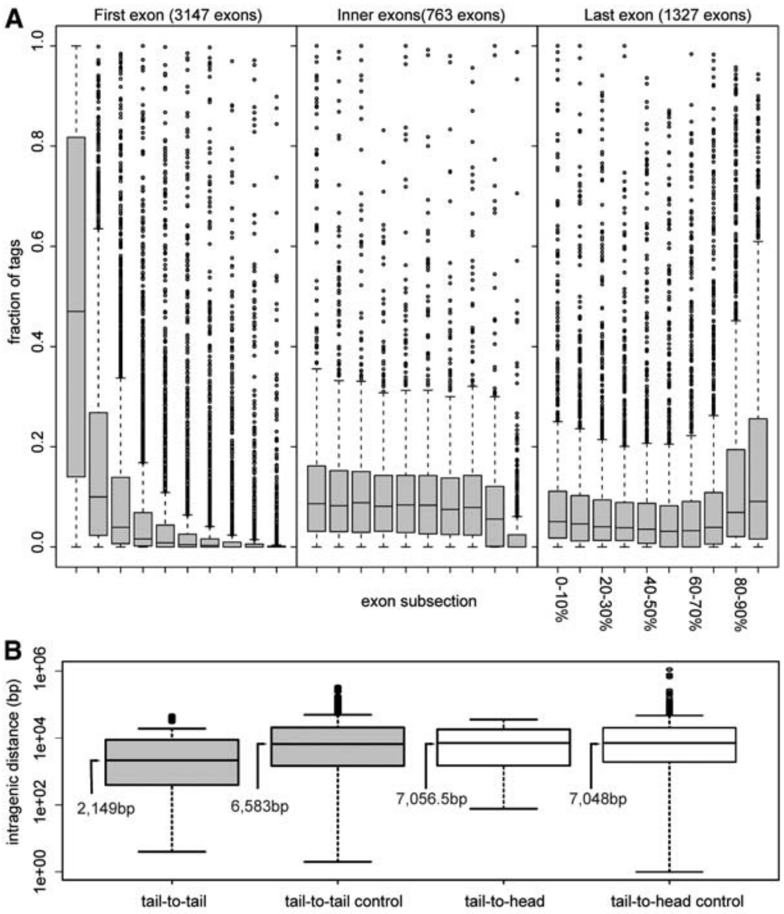


Fig. 2. Transcription originating in 3′UTRs. (**A**) For each analyzed exon, the fraction of tags mapped to 10 equally large subsections of the exon was calculated. (Left) CAGE tags mapping to the first exon are prevalently located in the first part of the exon. (Middle) CAGE tags mapping to internal exons are uniformly distributed. (Right) Last exons show a distinct overrepresentation of CAGE tags mapping close to the 3′ end. (**B**) Distance to the closest downstream gene for the set of highly expressed TUs that have extreme tag density in the 3′ of the terminal exons. Transcript pairs were grouped into tail-to-head (3′ exon and downstream TU on same strand) or tail-to-tail (3′ exon and downstream TU on opposite strand) configurations. Remaining TUs were used as control groups. For TUs with strong 3′ transcriptional activity, the distance to the next TU is significantly smaller than expected when the gene pair is in a tail-to-tail configuration ($P \leq 0.001107$, Wilcoxon test), suggesting regulatory mechanisms based on natural antisense influencing the downstream gene (*26*).

Table 1. Data set resources.

	Total	Number of libraries	Safely mapped
RIKEN full-length cDNAs	102,801	237	100,313
Public (non-RIKEN) mRNAs	56,009		52,119
CAGE tags (mouse)	11,567,973	145	7,151,511
CAGE tags (human)	5,992,395	24	3,106,472
GIS ditags	385,797	4	118,594
GSC ditags	2,079,652	4	968,201
RIKEN 5′ESTs	722,642	266	607,462
RIKEN 3′ESTs	1,578,610	265	907,007
5′/3′EST pairs of RIKEN cDNA	448,956	264	277,702

The approach used to isolate full-length cDNAs, based on library subtraction and previously unidentified 5′/3′ end selection before full-insert sequencing, was weighted toward identification of representative transcripts. Nevertheless, 78,393 different splicing variants were identified, such that 65% of TUs contain multiple splice variants (Table 2), an increase from our previous estimate (41%) (*9*). This is still expected to be an underestimate, and new approaches will be necessary for a full evaluation of exon diversity (*10*).

Transcript diversity also arises through alternative termination. Little is known about sequence motifs that control alternative polyadenylation. We identified 27 motif families with six or more nucleotides that were statistically overrepresented within 120 base pairs of the polyadenylation site of individual transcripts in our data set. These motifs represent candidate modulators of polyadenylation site for eight unconventional alternative polyadenylation signals (*1*) (table S3). In addition, we found a widespread motif family with sequence TTGTTT, which was associated with both the canonical (AAUAAA and AUUAAA) and unconventional signals (*1*, *11*).

Gene names of 56,722 transcripts that were protein coding were assigned according to annotation rules (*9*, *12*). Their encoded protein sequences were combined with the publicly available proteins supported by cDNA sequences (*8*). This generated a nonredundant set of 51,135 proteins with experimental evidence [isoform protein set (IPS)], 36,166 of which are complete (complete IPS). By comparison, the mammalian gene collection (http://mgc.nci.nih.gov) has cloned, as of July 2005, only ~16,700 transcripts (11,514 nonredundant). In the FANTOM3 data set, 16,274 protein sequences are newly described. Their splice variants were grouped together into 13,313 TKs. For 9002 of these, a previously known sequence maps to the same TK (locus), but 4311 clusters (5154 different proteins) map to new TKs (SOM text 3).

There are a total of 32,129 protein-coding TKs on the genome, of which 19,197 have only a single protein splice form, although 2525 of those do have an alternative noncoding splice variant. The SUPERFAMILY analysis of structural classification of protein database (SCOP) domain architectures (*13*) was carried out for each sequence. Of the 12,932 TKs that show variation in splicing, 8365 showed variation in SCOP domain prediction. Of the 12,932 variable TKs, 2392 produce proteins with different observed contents of InterPro entries. More than two alternatives were observed in 439 of the 2392 InterPro-variable TKs. Thus, in the majority of variable loci, splicing controls some aspect of domain content or organization. To seek evidence for such an impact in specific sets of regulatory proteins, we compared a representative protein set

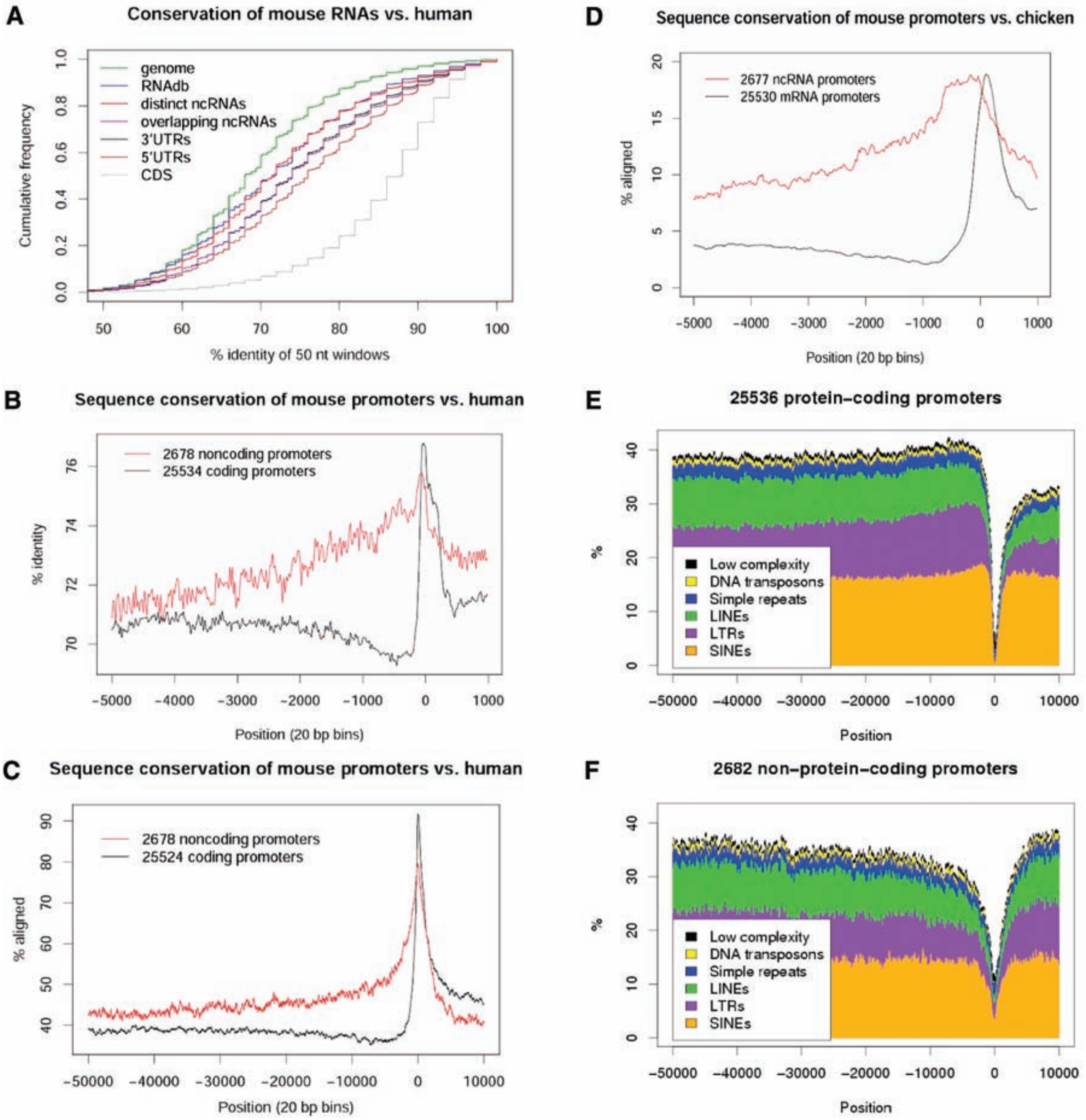


Fig. 3. Noncoding RNA promoters are highly conserved. (**A**) Human-mouse conservation of coding and noncoding RNAs compared with random genome sequence. (**B** and **C**) Promoters conservation of noncoding and coding mRNA evaluated (B) by identity and (C) by alignment. (**D**) Overlap of promoters of ncRNAs. (**E** and **F**) Promoters of coding mRNAs contain a larger fraction of low complexity and repeats than noncoding promoters. LINE, long interspersed nuclear elements; LTR, long terminal repeats; SINEs, short interspersed nuclear elements.

(RPS) and a variant protein set (VPS) of phosphatases and kinases that have been comprehensively annotated (*14*) by looking at domain composition counts (table S4). These phosphoregulators could be functionally modulated through alteration in their intracellular location. Among the 21 receptor tyrosine phosphatase loci, we identified 23 variant transcripts from 14 loci with predicted changes to the subcellular localization and function of the encoded peptides. Of these, we identified two noncatalytic classes: secreted (10) and tethered (3). Furthermore, we identified two catalytic classes that lack the extracellular domains: catalytic only (5) and tethered catalytic (5). Similarly, among the 77 receptor kinase loci, we identified 41 variant transcripts from 33 loci which encode secreted (16), tethered (10), catalytic only (7), or other tethered catalytic (8) peptides. We then analyzed the membrane organization splicing variants class within the full set of TUs (table S5), which revealed 1287 TUs that exhibit alternative initiation, splicing, and termination, likely to yield variant isoforms of membrane proteins that differ in their cellular location.

Of the 102,281 FANTOM3 cDNAs, 34,030 lack any protein-coding sequence (CDS) and are annotated as non–protein coding RNA (ncRNA) (*6*, *15*) (table S1). Many putative ncRNAs were singletons in the full-length cDNA set. Among the FANTOM3 cDNA set there was additional support from ESTs, CAGE tags, or other cDNA clones overlapping both the starting and termination sites for 41,025 cDNAs, of which only 3652 were ncRNAs. This supported ncRNA set includes many known ncRNAs (SOM text 4), and many are dynamically expressed (SOM text 5). Following these same criteria, 3012 from 8961 cDNAs previously annotated as truncated CDS were supported as genuine transcripts and are believed to be ncRNA variants of protein-coding cDNAs.

Many ncRNAs appear to start from initiation sites in 3′ untranslated regions (3′UTRs) of protein-coding loci (*16*). The normalized distribution of CAGE tags along annotated exons of known transcripts with more than 300 mapped tags each is shown in Fig. 2A. As expected, the highest tag density on average occurs at the 5′ end, but there is also a substantial increase of tags in the last one-fifth of the 3′UTR. Strong evidence of 3′ end initiation was correlated with a short intergenic distance when in tail-to-tail orientation with a neighboring gene (Fig. 2B), suggesting a possible role in an intergenic regulatory interaction.

The function of ncRNAs is a matter of debate (*17*). Some ncRNAs are highly conserved even in distant species: 1117 out of 2886

Table 2. Transcript grouping and classification. The extent of splice variation was calculated by excluding T-cell receptor and immunoglobulin genes from the transcripts. The remaining 144,351 transcripts were grouped in 43,539 TUs, of which 18,627 (42.8%) consist of single-exon transcripts, 8110 (18.6%) contain a single multiexon transcript, and the remaining 16,802 TUs (38.6%) contain at least two spliced transcripts. Among these TUs, 5862 (34.9%) show no evidence of splice variation, whereas 10,940 (65.1%) contain multiple splice forms.

	Total	Average per TU cluster	Average per TK cluster
Total number of transcripts	158,807	7.59	7.30
RIKEN full-length	102,801		
Public (non-RIKEN) mRNAs	56,006		
GFs	25,027	1.20	1.15
Framework clusters	31,992	1.53	1.47
TUs	44,147	2.11	2.03
With proteins	20,929	1.00	0.96
Without proteins	23,218	1.11	1.07
TK	45,142	2.16	2.07
With proteins	21,757	1.04	1.00
Without proteins	23,385	1.12	1.07
Splicing patterns	78,393	3.75	3.60

Table 3. Determination of transcripts start/end accuracy. Two pieces of evidence (cDNA, tags, ditags, EST, and 5′-3′ EST pairs) are required when TSS/terminations lie inside larger transcripts, and one piece of evidence is required when they extend or identify new transcripts. Reliable indicates that both ends are associated with reliable tag clusters.

	Total	Reliable
Total 5′/3′-end pair sequence	1,507,122	1,336,397
5′/3′-end pair cluster	313,821	181,047

overlap chicken sequences, of which 780 do not overlap known CDS and 438 do not overlap known mRNAs on either strand, whereas 68 out of 2886 have BLAST-like alignment tool (BLAT) alignments to the Fugu genome, of which 40 do not overlap known CDS on either strand. These ncRNAs are at least as conserved as a reference set of known ncRNAs (Fig. 3A), contrary to a previous study (*17*). However, ncRNAs are slightly less conserved on average than 5′ or 3′UTRs. In contrast, the promoter regions of ncRNAs are generally more conserved than the promoters of the protein-coding mRNA, not only between human and mouse but also down in the evolutionary scale to chicken (Fig. 3, B to F), and they contain binding sites for known transcription factors (*18*). We conclude that the large majority of ncRNAs that we analyzed display positional conservation across species. In considering function, one might conclude that the act of transcription from the particular location is either important or a consequence of genomic structure or sequence (for example, enhancers such as that of the globin locus can act as promoters), the transcript may function through some kind of sequence-specific interaction with the DNA sequence from which it is derived, or many noncoding RNAs have other targets but are evolving rapidly (*19*, *20*).

New databases have been created for cDNA annotation, expression, and promoter analysis (http://fantom3.gsc.riken.jp/db/ and SOM text 6). The databases integrate common gene and tissue ontologies like eVOC mouse developmental ontologies (*21*), cross mapped to Edinburgh Mouse Atlas Project (EMAP) ontology terms (*22*). These eVOC terms allow analysis standardization of RNA samples used for cDNA and CAGE libraries in both mouse and human and were included into the DNA Database of Japan (DDBJ) data submission (*23*).

Analysis of the output of FANTOM2 suggested that there were many more transcripts still to be discovered (*24*). Here, we have confirmed that the majority of the mammalian genome is transcribed, commonly from both strands. Such transcriptional complexity implies caveats in interpretation of microarray experiments (*25*) and genome manipulation in mice, because these will commonly interrupt or interrogate more than one TK. Although the current overview gives us an indication of the complexity of the mammalian transcriptional landscape and a new set of tools to begin to understand transcriptional control (for example a very large set of promoters that can be ascribed to distinct classes) (*16*), we also gain insight into the scale of the task that remains. The ditag data indicate the existence of very long transcripts whose isolation and sequencing will require new cloning and sequencing strategies. Although we have isolated and sequenced many putative ncRNAs, the FANTOM3 collection only contains 40% of those already known. Finally, the focus has been on polyadenylated mRNAs that are processed and exported to the cytoplasm. Recently, Gingeras and colleagues (*5*) have shown that the set of nonpolyadenylated nuclear RNAs may be very large, and that many such transcripts arise from so-called intergenic regions (*7*). The future can only reveal additional complexity in the mammalian transcriptome.

References and Notes

1. P. Carninci *et al.*, *Genome Res.* **13**, 1273 (2003).
2. T. Shiraki *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 15776 (2003).
3. P. Ng *et al.*, *Nat. Methods* **2**, 105 (2005).
4. R. H. Waterston *et al.*, *Nature* **420**, 520 (2002).
5. D. Kampa *et al.*, *Genome Res.* **14**, 331 (2004).
6. P. Bertone *et al.*, *Science* **306**, 2242 (2004).
7. J. Cheng *et al.*, *Science* **308**, 1149 (2005).
8. R. L. Strausberg *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 16899 (2002).
9. Y. Okazaki *et al.*, *Nature* **420**, 563 (2002).
10. A. Watahiki *et al.*, *Nat. Methods* **1**, 233 (2004).
11. V. Bajic, in preparation.
12. N. Maeda, R. Oyama, in preparation.
13. J. Gough, in preparation.
14. A. R. Forrest *et al.*, *Genome Res.* **13**, 1443 (2003).
15. Materials and methods are available as supporting material on *Science* Online.
16. P. Carninci *et al.*, in preparation.
17. J. Wang *et al.*, *Nature* **431**, 1 p following 757; discussion following 757 (2004).
18. S. Cawley *et al.*, *Cell* **116**, 499 (2004).
19. T. Ravasi, D. A. Hume, in *Encyclopedia of Genetics, Genomics, Proteomics, and Bioinformatics*, L. B. Jorde, P. F. R. Little, M. J. Dunn, S. Subramaniam, Eds. (John Wiley & Sons, Chichester, UK, in press), part 2.3.
20. J. S. Mattick, I. V. Makunin, *Hum. Mol. Genet.*, in press.
21. J. Kelso *et al.*, *Genome Res.* **13**, 1222 (2003).
22. R. A. Baldock *et al.*, *Neuroinformatics* **1**, 309 (2003).
23. All sequences (CAGE, and cDNA) are available through DDBJ to other public databases. The cDNA clones are available.
24. Y. Okazaki, D. A. Hume, *Genome Res.* **13**, 1267 (2003).
25. E. Marshall, *Science* **306**, 630 (2004).
26. RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) and the FANTOM Consortium, *Science* **309**, 1564 (2005).

27. We thank H. Atsui, A. Hasegawa, Y. Hasegawa, K. Hayashida, H. Himei, F. Hori, T. Iwashita, S. Kanagawa, C. Kawazu, M. Aoki, K. Murakami, M. Murata, H. Nishida, M. Nishikawa, K. Nomura, M. Ohno, Y. Onodera, N. Sakazume, H. Sato, Y. Shigemoto, N. Suzuki, Y. Takeda, Y. Tsujimura, K. Yoshida for discussion, encouragement, and technical assistance. We thank A. Wada, T. Ogawa, M. Muramatsu, and all the members of RIKEN Yokohama Research Promotion Division for support and encouragement. We also thank the Laboratory of Genome Exploration Research Group for secretarial and technical assistance, Yokohama City University for providing human samples, and computational resources of the RIKEN Super Combined Cluster (RSCC). This work was mainly supported by Research Grant for the Genome Network Project from MEXT, the RIKEN Genome Exploration Research Project from MEXT (Y.H.), Advanced and Innovational Research Program in Life Science (Y.H.), National Project on Protein Structural and Functional Analysis from MEXT (Y.H.), Presidential Research Grant for Intersystem Collaboration of RIKEN (P.C. and Y.H.) and a grant from the Six Framework Program from the European Commission (P.C.).

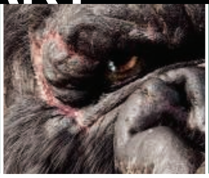
Supporting Online Material
www.sciencemag.org/cgi/content/full/309/5740/1559/DC1
Materials and Methods
SOM Text
Figs. S1 to S4
Tables S1 to S10
References
DDBJ Accession Codes

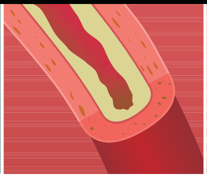
9 March 2005; accepted 4 August 2005
10.1126/science.1112014

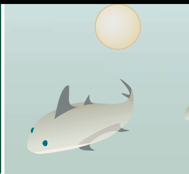
ERRATUM

Post date 24 March 2006

Reports: "The transcriptional landscape of the mammalian genome," by The FANTOM Consortium *et al.* (2 Sept. 2005, p. 1559). On page 1561, column 3, lines 40-46 should read: "In the FANTOM3 data set, 11,559 protein sequences are newly described. Their splice variants were grouped together into 7445 TKs (transcriptional frameworks). For 5453 of these, a previously known sequence maps to the same TK (locus), but 1992 clusters (2222 different proteins) map to new TKs (see SOM text 3)."







LETTERS

edited by Etta Kavanagh

How Many New Genes Are There?

IN THEIR REPORT "THE TRANSCRIPTIONAL LANDSCAPE OF THE MAMMALIAN GENOME" (2 SEPT. 2005, p. 1559), the RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) and the FANTOM Consortium claim to have found 5154 new proteins in the mouse genome not encoded by previously known mRNA sequences, which could potentially correspond to a considerable number of new protein-coding genes (4311, following clustering) (*1*). This claim contrasts dramatically with the view of the International Human Genome Sequencing Consortium (*2*), which estimated that there are 20,000 to 25,000 protein-coding genes. Since there are already 22,287 genes in the Ensembl 34d catalog, this implies 0 to 2713 new genes. RIKEN/FANTOM's estimate contrasts even more strikingly with our results using exon microarrays (*3*), in which the number of new multi-exon protein-coding genes was estimated to be at most in the hundreds. We analyzed the putative new FANTOM proteins (*4*), first by comparing their sequences with RefSeq release 13 from NCBI, NIH, and including only those transcripts that are linked to a reference published no later than 1 May 2005, thus excluding all the new FANTOM proteins. Restricting our analysis to the transcripts that have strong experimental evidence (labeled Provisional, Validated, or Reviewed), we found that 2917 (56.6%) of the FANTOM proteins are in fact splice isoforms of known RefSeq transcripts, with the majority of them (2716) corresponding to exon-skipping events. By then including predicted RefSeq transcripts (labeled Genome Annotation, Inferred, Model, Predicted) in our analysis, 3568 (69.2%) were found to be splice isoforms of known transcripts. By including GenBank mRNAs linked to publications before 1 May 2005, we found an extra 303 splice isoforms, bringing the total of already-annotated genes to 3871 (75.1%). Moreover, of the 5154 FANTOM proteins, our microarray analysis detected 2293 (by two or more exons), 144 of which are among the remaining 1283 FANTOM proteins and most (131) of which are associated with known genes. We next asked whether the remaining 1193 putative proteins could be accounted for as false detections. The median open reading frame (ORF) size in this set is 119 amino acids (aa), significantly shorter than that of all the FANTOM proteins (330 aa). Although many real proteins have a length less than 119 aa, we hypothesized that such a short ORF length can arise in noncoding transcripts by chance. The FANTOM Consortium identified 23,218 nonoverlapping, noncoding transcripts, so to test this hypothesis we generated a set of 20,000 random cDNAs of 2000 bases (typical gene length) and found that 1247 of them had ORFs of 119 aa or more. Therefore, it is possible that a large portion of the remaining 1193 putative proteins arose at random from noncoding transcripts and may not encode functional polypeptides. On the basis of this analysis, the number of completely new protein-coding genes discovered by the FANTOM Consortium is at most in the hundreds, consistent with current estimates based on both sequence and microarray analysis (*2*, *3*).

> "...the number of completely **new protein-coding genes** discovered by the FANTOM Consortium is **at most in the hundreds...**"
>
> —Lee *et al.*

LEO J. LEE,[1] TIMOTHY R. HUGHES,[2] BRENDAN J. FREY[1]

[1]Department of Electrical and Computer Engineering and [2]Banting and Best Department of Medical Research, University of Toronto, 10 King's College Road, Toronto, ON M5S 3G4, Canada.

References and Notes

1. FANTOM3 cDNA sequences are not provided, but the protein sequences can be downloaded from http://fantom3.gsc.riken.jp/.
2. International Human Genome Sequencing Consortium, *Nature* **431**, 931 (2004).
3. B. J. Frey *et al.*, *Nat. Genet.* **37**, 991 (2005).
4. See www.psi.toronto.edu/TransLand for details.

Response

LEE *ET AL.* POINT OUT THAT THE NUMBER OF reported protein sequences in FANTOM3 that map to new positions on the genome appears to be too large. We are grateful to them for highlighting this discrepancy, which we investigated and thus discovered an error. For a detailed description of the correction, see the Corrections and Clarifications section in this issue. The effect of the error is somewhat less than suggested by Lee *et al.* In particular, our estimate of the number of new protein-coding genes found by us has been revised from 5154 to 2222, a reduction of more than half, but much less than the order of magnitude suggested by Lee *et al.* As correctly pointed out, the rest of the 5154 cDNAs are mainly alternatively spliced isoforms.

Lee *et al.* present three forms of evidence: sequence similarity, exon microarray data, and ORF size. (i) The sequence homology data largely reflect the revision to the number that we mention above, except that Lee *et al.* used a recent RefSeq database, whereas we used Genbank (7 January 2004). There is no evidence that all RefSeq sequences correspond to real transcribed RNAs because they often include ab initio predicted exons (*1*). Our strategy was to construct the transcriptional frameworks entirely based on real RNA transcripts, rather than in silico reconstruction of putative gene structures. (ii) The exon microarray data concern less than 3% of the number

> "...the number of new protein-coding genes found by us has been **revised from 5154 to 2222...**"
>
> —FANTOM Consortium

Published by AAAS

of discussed proteins and do not have any impact on the global message of a project of the scale of FANTOM3. Despite Frey *et al.*'s impressive computational reconstruction of gene structure by analyzing expression patterns of ab initio predicted exons (*2*), we argue that this does not prove the physical structure of each mRNA and the complexity of the transcriptome with the same resolution achieved by sequencing libraries derived from mRNAs. In fact, our data show that "genes" have multiple starting and termination sites: We have conservatively identified at least 181,000 different RNA transcripts. Additionally, Frey *et al.* (*2*) used only computationally predicted exons. Rare, newly discovered transcripts are unlikely to have been in the training sets of ab initio exon identification tools, and their sensitivity to predict rare transcriptional events is not obvious. (iii) As for ORF size, 119 amino acids is a perfectly respectable size for a protein and within the bounds of statistical variation we expect. In this regard, we have further identified in the FANTOM3 dataset at least 1100 proteins shorter than 100 amino acids (*3*). Also, all of the novel FANTOM3 transcripts have been manually curated by individual researchers to distinguish them from novel noncoding RNAs. In any case, our final understanding of the number of protein-coding mRNAs will derive from experimental validation with full-length cDNA clones (*3*) rather than computational inferences. We direct interested parties to the relevant section of the FANTOM3 Web site (http://fantom.gsc.riken.jp) where the updated files are available, and we thank Lee *et al.* for helping us to improve and update our analysis.

PIERO CARNINCI,[1,2] JULIAN GOUGH,[1] TAKEYA KASUKAWA,[1,3] YOSHIHIDE HAYASHIZAKI[1,2]

[1]Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan. [2]Genome Science Laboratory, Discovery and Research Institute, RIKEN Wako Institute, 2-1 Hirosawa, Wako, Saitama, 351-0198, Japan.[3]NTT Software Corporation, Teisan Kannai Building 209, Yamashita-cho, Naka-ku, Yokohama, Kanagawa, 231-8551, Japan.

References

1. X. Pruitt *et al.*, *Nucleic Acids Res.* **33,** D501 (2005).
2. B. Frey *et al.*, *Nat. Genet.* **37**, 991 (2005).
3. M. Frith *et al.*, *Plos Genet.*, in press.

Why Suicide Rates Are High in China

WE READ WITH INTEREST G. MILLER'S ARTICLE describing a discrepancy between Chinese rates of suicide and depression ("China: healing the metaphorical heart," News Focus, 27 Jan., p. 462). However, we feel that Miller, by concentrating on fatal self-harm rather than all acts of self-harm, misses an opportunity to understand the discrepancy he notes.

High rates of suicide and low rates of depression are not restricted to China. Many countries of the Asian "suicide belt" have suicide rates higher than those of China (*1*, *2*).

Suicide rates result from the incidence of self-harm and the resulting fatality rate among those individuals. Our research in Sri Lanka indicates that high rates of suicide from self-poisoning are due to a high fatality rate rather than a high incidence of self-harm itself (*3*). A useful contrast can be made with the UK.

Self-poisoning in the UK is very common, with an annual incidence of presentation to hospital of around 300 per 100,000. However, self-poisoning is rarely lethal, with a fatality rate per 1000 incidents normally less than 0.5% (*4*). Self-poisoning is also common in Sri Lanka, with an estimated incidence of around 363 per 100,000 in one rural district. However, the fatality rate is significantly higher at ~7.4%—at least 15 times higher than in the UK (*3*). The reason for this higher fatality rate in Sri Lanka, as in China, is the common use of highly toxic poisons such as pesticides. Sri

Lankan self-poisoners are not more keen to die—they simply have easier access to pesticides than do the residents of the UK (*5*).

The high suicide rate in Sri Lanka and China is not due to higher levels of mental illness or rates of self-harm, but to a higher lethality of self-harm acts. Concentrating solely on rates of mental illness in Asia will not explain the high rate of suicide in this region.

MICHAEL EDDLESTON[1]* AND DAVID GUNNELL[2]

[1]Centre for Tropical Medicine, Nuffield Department of Clinical Medicine, University of Oxford, Oxford OX3 9DU, UK. [2]Department of Social Medicine, University of Bristol, Bristol BS8 2PR, UK.

*To whom correspondence should be addressed. E-mail: eddlestonm@eureka.lk

References

1. P. Brown, *New Sci.*, 22 Mar. 1997, p. 34.
2. A. Joseph *et al*., *Br. Med. J.* **326**, 1121 (2003).
3. M. Eddleston *et al*., *Bull. World Health Org.*, in press.
4. D. Gunnell, D. D. Ho, V. Murray, *Emergency Med. J.* **21**, 35 (2004).
5. M. Eddleston *et al*., *Clin. Toxicol*., in press.

CORRECTIONS AND CLARIFICATIONS

BOOKS *ET AL.*: "Humanity usurps nature" by B. Chameides (10 Mar., p. 1379). The affiliation information is incorrect. Bill Chameides is at Environmental Defense, 257 Park Avenue South, New York, NY 10010, USA. E-mail: BChameides@environmentaldefense.org

REPORTS: "The transcriptional landscape of the mammalian genome" by the FANTOM Consortium *et al.* (2 Sept., p. 1559). On page 1561, column 3, lines 40–46 should read: "In the FANTOM3 data set, 11,559 protein sequences are newly described. Their splice variants were grouped together into 7445 TKs (transcriptional frameworks). For 5453 of these, a previously known sequence maps to the same TK (locus), but 1992 clusters (2222 different proteins) map to new TKs (see SOM text 3)."

NEWS FOCUS: "With energy to spare, an engineer makes the case for basic research" by E. Kintisch (10 Mar., p. 1369). The National Superconducting Cyclotron Laboratory at Michigan State University was incorrectly identified as being supported by the Department of Energy. The NSCL is a campus-based national user facility funded by the National Science Foundation.

TECHNICAL COMMENT ABSTRACTS

Comment on "Changes in Tropical Cyclone Number, Duration, and Intensity in a Warming Environment"

Johnny C. L. Chan

Analyses of tropical cyclone records from the western North Pacific reveal that the recent increase in occurrence of intense typhoons reported by Webster *et al*. (Reports, 16 September 2005, p. 1844) is not a trend. Rather, it is likely a part of the large interdecadal variations in the number of intense typhoons related to similar temporal fluctuations in the atmospheric environment.

Full text at www.sciencemag.org/cgi/content/full/311/5768/1713b

Response to Comment on "Changes in Tropical Cyclone Number, Duration, and Intensity in a Warming Environment"

P. J. Webster, J. A. Curry, J. Liu, G. J. Holland

Although Chan makes several valid points, his analysis confuses relationships associated with the long-term variations with those associated with shorter term variability (interannual and decadal). We present an analysis that clarifies the observations from the western North Pacific.

Full text at www.sciencemag.org/cgi/content/full/311/5768/1713c

Letters to the Editor

Letters (~300 words) discuss material published in *Science* in the previous 6 months or issues of general interest. They can be submitted through the Web (www.submit2science.org) or by regular mail (1200 New York Ave., NW, Washington, DC 20005, USA). Letters are not acknowledged upon receipt, nor are authors generally consulted before publication. Whether published in full or in part, letters are subject to editing for clarity and space.

Published by AAAS