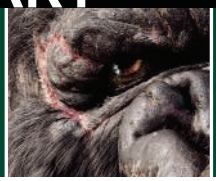
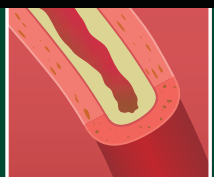
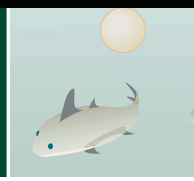
King Kong epic ape 1714



Lower cholesterol for a lifetime 1721



Organic matter and oxygen reactions 1723



LETTERS | BOOKS | POLICY FORUM | EDUCATION FORUM | PERSPECTIVES

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