Estimating rates of alternative splicing in mammals and invertebrates

To the editor:

The sequencing of the human genome showed that humans have ~30,000 genes. This finding raised the possibility that alternative splicing, rather than an increased number of expressed genomic loci, was responsible for the functional complexity of vertebrates relative to invertebrates[1]. It has been estimated that 40–60% of all human genes[1–4] and 74% of multiexon human genes[5] are alternatively spliced. These estimates do not take into account how many different alternative splice forms exist for a given gene. Brett *et al.* examined alternative splicing in seven species, including human, using large-scale expressed-sequence tag (EST) analysis[6]. They concluded that vertebrates and invertebrates had similar rates of alternative splicing, not only with respect to the proportion of the genes affected but also with respect to the number of alternative splicing forms per gene. The method they used depends on the extent of EST coverage in the underlying data sets.

To avoid this shortcoming and to provide an alternative estimate for the number of splice variants per gene, we modified the method that Ewing and Green used to estimate the total human gene count[7]. This method requires two independent sets of incomplete gene sequence data from the organism, the first of which should be unbiased. For counting the genes in the human genome, Ewing and Green used mRNA sequences for the first data set and EST contig sequences for the second data set[7]. They required at least 100 aligned bases between a pair of sequences between the two data sets to agree in order to consider them a match. They used this low number because some of the mRNAs were incomplete or represented different alternatively spliced forms of the same gene. To count alternative splice forms using this method, we considered that a set of alternative splice forms from the same gene could be counted as different genes by increasing the minimum number of aligned bases for sequence comparison. As the minimum number of aligned bases increases, the gene count G will asymptotically approach the total number of transcriptional products. Using this approach, we estimated the extent of alternative splicing in *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus* and *Homo sapiens*. We constructed the first set of data (n_1) for each genome by selecting UniGene clusters[8] that are represented by a member of the National Center for Biotechnology Information Reference Sequences (RefSeq) database[9]. We derived the set of EST contigs (n_2) from the appropriate gene indices from The Institute for Genomic Research[10]. Further details are available in **Supplementary Methods** online.

From these comparisons, we predicted the value for G as the minimum number of aligned bases varied. As expected, G increased asymptotically as the minimum number of aligned bases increased (**Fig. 1a**). The rate of alternative splicing per gene can be estimated as the ratio of the values of G at the two extreme ends of the graph (**Fig. 1b**). These data indicate that mice and humans have a higher rate of alternative splicing than do fruit flies and nematodes.

To confirm that these observations were not the result of differing amounts of EST data, we tested the four species with same number of reference sequences (n_1). Four different subsets of EST contigs (all, one-half, one-quarter and one-eighth subsets of the data) showed nearly indistinguishable results (**Supplementary Fig. 1** online). Therefore, our predictions are independent of the extent of EST contig coverage. To test whether our results could be due to the presence of pseudogenes, we analyzed a subset of the EST contigs that hit only one

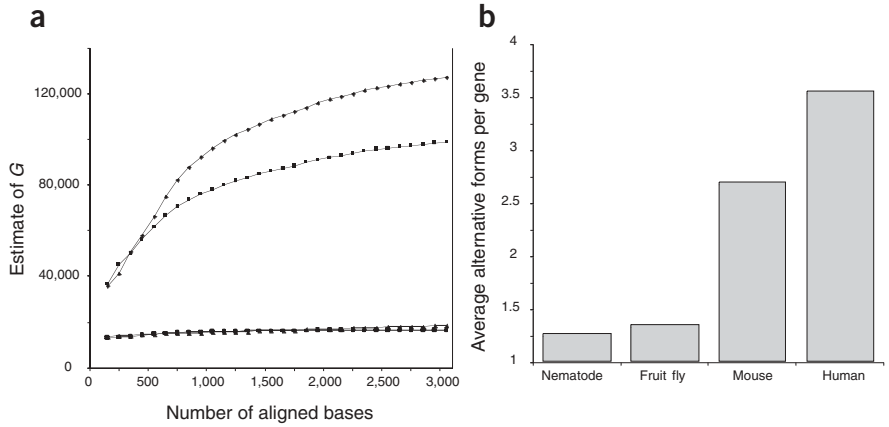

Figure 1 Extent of alternative splicing in four organisms. (**a**) Estimate of the gene count G with varying minimum number of aligned bases for *H. sapiens* (diamonds), *M. musculus* (squares), *D. melanogaster* (triangles) and *C. elegans* (circles). (**b**) Average number of alternative splice forms per gene for each organism. This was calculated by dividing the estimate for G with a minimum of 3,000 aligned bases by the estimate for G at a minimum of 100 aligned bases. Error bars show the 95% confidence intervals, assuming there are no errors in the underlying data, calculated as described in **Supplementary Methods** online.

region of the human genome. The rate of alternative splicing in the subset that excluded transcriptional variants from pseudogenes was similar to that observed using the full EST contig set (**Supplementary Fig. 2** online). This indicates that the confounding affect of the pseudogenes is not substantial. Finally, we determined whether the higher proportion of tumor libraries in human EST data (relative to the other species) confounded our results. Notably, tumor-specific transcripts seem to have a higher rate of alternative splicing in humans (**Supplementary Fig. 3** online). But the rate of alternative splicing estimated from the non-tumor-specific transcripts was similar to that estimated using a random sample of the data set including all EST contigs. (**Supplementary Fig. 3** online). This is probably because only ~10% of the contigs in the Gene Indices data set from The Institute for Genomic Research are considered tumor-specific by our criteria. Further details are available in **Supplementary Methods** online.

Our results disagree with those of Brett *et al.*[6]. We believe that our method provides a more accurate answer, as it does not depend on the extent of EST coverage. Our results indicate that, in accordance with expectations given that there are only ~30,000 genes in humans, there is a greater amount of alternative splicing in mammals than in invertebrates. Enumerating these different forms and understanding their roles in contributing to biological complexity is a vast area for future research.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

This work was supported by a grant from BioGreen 21 Program of the Korean Rural Development Administration (to H.K.), by the Brain Korea 21 Project of the Ministry of Education (to H.K.), by a grant from the US National Human Genome Research Institute (to J.O.) and by a grant from the US National Institute of Mental Health (to J.O.).

Heebal Kim[1], Robert Klein[2], Jacek Majewski[2] & Jurg Ott[2]

[1]School of Agricultural Biotechnology, Seoul National University San 56-1, Sillim-dong, Gwanak-gu, Seoul 151-742, Korea. [2]Laboratory of Statistical Genetics, Rockefeller University 1230 York Avenue, New York, New York 10021, USA. Correspondence should be addressed to H.K. (heebal@snu.ac.kr).

1. Modrek, B. & Lee, C. *Nat. Genet.* **30**, 13–19 (2002).
2. Mironov, A.A. *et al. Genome Res.* **12**, 1288–1293 (1999).
3. Modrek, B. *et al. Nucleic Acids Res.* **29**, 2850–2859 (2001).
4. Kan, Z. *et al. Genome Res.* **11**, 889–900 (2001).
5. Johnson, J.M. *et al. Science* **302**, 2141–2144 (2003).
6. Brett, D. *et al. Nat. Genet.* **30**, 29–30 (2002).
7. Ewing, B. & Green, P. *Nat. Genet.* **25**, 232–234 (2000).
8. Wheeler, D.L. *et al. Nucleic Acids Res.* **31**, 28–33 (2003).
9. Pruitt, K.D. & Maglott, D.R. *Nucleic Acids Res.* **29**, 137–140 (2001).
10. Quackenbush, J. *et al. Nucleic Acids Res.* **28**, 141–145 (2000).

In reply:

Kim *et al.* provide an interesting, albeit somewhat indirect, method to estimate the rate of alternative splicing per gene in different organisms based on expressed-sequence tag (EST) and mRNA data. Their results disagree with our earlier conclusions[1] and fit better with the general expectation that the rate of alternative splicing increases with organismal complexity. Furthermore, the authors argue that their method is superior to direct EST matching approaches as it is independent of the number of ESTs used. The problem with both methods, as with many other studies based on 'omics' data, is that hidden biases and flaws in data and methods can heavily affect the outcome of an estimation. The results of Kim *et al.* agree with our results before normalizing for EST redundancy[1] (*i.e.*, more ESTs record more splice variants; *e.g.* refs. 2,3 and references therein) and correlate with the length distribution of EST contigs in the species analyzed (**Fig. 1a**). This prompted us to study the method of Kim *et al.* in detail and attempt to reproduce their results (see **Supplementary Note** online for a discussion of the difficulties of reproducing the method of Kim *et al.*). To test Kim *et al.*'s claim that their method is independent of EST coverage, we included the rat in our analysis, as it has an EST coverage similar to that of invertebrates. Notably, the estimated rate of alternative splicing in the rat was closer to that of invertebrates than to that of the mouse (**Fig. 1b**), suggesting that EST coverage does bias the results of Kim *et al.* This idea is further supported by the correlation between EST and RefSeq sequence coverage per organism and the rate of alternative splicing estimated by Kim *et al.* (**Fig. 1b**).

Why do subsets of the EST contigs show the same rate of alternative splicing as the full data set? We believe that the use of The Institute for Genomic Research (TIGR) EST contigs[4] rather than ESTs themselves leads to misleading statistics. TIGR contigs are derived data that merge redundant ESTs, and, in theory, each splice form of a gene should be represented by a single contig, including the forms represented by RefSeq sequences. The rate of alternative splicing per gene measured by Kim *et al.* ultimately depends on the ratio of the total number of contigs that support alternative splicing forms to the number of contigs that support RefSeq forms (for details see **Supplementary Note** online). When Kim *et al.* used subsets of TIGR contigs to test the dependency of their method on EST coverage, they affected the number of RefSeq and alternative splicing forms equally. Therefore, the ratio (and the rate)

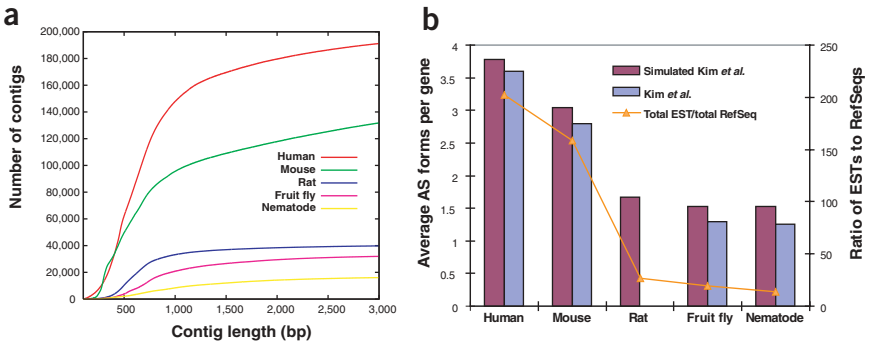

Figure 1 Dependence of the method of Kim *et al.* on EST coverage. (**a**) Cumulative length distribution of TIGR EST contigs for each organism used in this analysis. The similarity between this graph and the graph by Kim *et al.* showing their estimate of *G* implies that the method indirectly measures the length of contigs (**Supplementary Note** online). (**b**) Relationship between EST coverage and the estimated rates of alternative splicing (AS) per gene in each organism. To adjust EST coverage to the context of this analysis, it is given as ratio of the total number of ESTs to the number of RefSeqs available for that organism. Limited reproducibility and hidden parameter choices are inherent in large-scale analyses, and we had difficulty reproducing the method of Kim *et al.* We therefore used simulations to reproduce their figures approximately (simulated Kim *et al.*; **Supplementary Note** online).

remains the same with any subset of the data. Thus, this test is not valid to prove independence from EST redundancy, which we argue is biasing the results of Kim *et al.*

Independent of the hidden biases that affect the method of Kim *et al.*, the use of EST data might not be optimal for comparative analysis: they might not reflect global gene expression patterns adequately[5] and they have various other flaws[6]. The limitations of EST coverage is best illustrated by the *Drosophila melanogaster* gene *Dscam*, which is only represented by ~20 ESTs, even though it seems to express >38,000 alternative splice variants[7]. The estimates of the rate of alternative splicing also strongly depend on the treatment of the data; just one parameter choice, the number of base pairs at the end of the ESTs to be ignored to account for sequencing errors, can substantially influence the estimation of the rate of alternative splicing (**Supplementary Note** online). Finally, the use of EST data to compare distant organisms is not trivial, as the protocols and sources used to produce ESTs vary greatly between organisms. In mammals, ESTs are heavily sampled in a subset of tissues, which can lead to biases (*e.g.*, human brain ESTs are over-represented, and brain is the tissue with the highest rate of alternative splicing[8]), whereas ESTs from invertebrates are often taken from whole organisms. It is a conceptual and practical challenge to normalize for these and other heterogeneities when comparing the extent of alternative splicing between different species.

The study by Kim *et al.* is important as it indicates that no estimate should be considered absolute when extrapolating from data with many hidden biases. Nevertheless, we doubt that the method in general and the implementation in particular is superior to direct EST matching (**Supplementary Note** online). The opportunity now exists to consider and carefully evaluate a variety of data and methods to approach an understanding of the impact of alternative splicing in each organism. The art is to avoid hidden biases in derived data and their processing as much as possible.

E D Harrington[1], S Boue[1], Juan Valcarcel[2], Jens G Reich[3] & Peer Bork[1,3]

[1]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [2]ICREA and Centre de Regulació Genòmica, Passeig Marítim, 37-49,08003 Barcelona, Spain. [3]Bioinformatics Section, Max Delbrück Center for Molecular Medicine, Berlin, Germany. Correspondence should be addressed to P.B. (bork@embl.de).

Note: Supplementary information is available on the Nature Genetics website.

1. Brett, D. *et al. Nat. Genet.* **30**, 29–30 (2002).
2. Kan, Z. *et al. Genome Res.* **12**, 1837–1845 (2002).
3. Boue, S. *et al. BioEssays* **25**, 1031–1034 (2003).
4. Quackenbush, J. *et al. Nucleic AcidsRes.* **29**, 159–164 (2001).
5. Kapranov, P. *et al. Science* **296**, 916–919 (2002).
6. Modrek, B. *et al. Nucleic Acids Res.* **29**, 2850–2859 (2001).
7. Neves, G. *et al. Nat. Genet.* **36**, 240–246 (2004).
8. Xu, Q. *et al. Nucleic Acids Res.* **30**, 3754–3766 (2002).