Commentary

Most of the Human Genome Is Transcribed

Gane Ka-Shu Wong,[1,3] Douglas A. Passey,[1] and Jun Yu[1,2]

[1]*University of Washington Genome Center, Department of Medicine, Fluke Hall, M/C 352145, Seattle, Washington 98195, USA;* [2]*Genomics and Bioinformatics Center, Chinese Academy of Science, Beijing, People's Republic of China*

Initial sequence annotations of the human genome have uncovered at least 32,000 genes (International Human Genome Sequencing Consortium 2001), or 26,000–39,000 genes (Venter et al. 2001). The mean gene size is thought to be 27 kb. Although these gene count estimates are acknowledged, by the authors themselves, to be very conservative, they are not significantly smaller than a recent estimate of 35,000 genes (Ewing and Green 2000) that was derived from a "proven" sampling technique. However, these gene count estimates are significantly smaller than the previously accepted estimates of 70,000 genes (Antequera and Bird 1993; Fields, et al. 1994). Suppose we accept the new gene counts, compromising between the two papers and settling at 35,000 genes. Let us further assume a euchromatic genome-size of 2.9 Gb. This would imply that only 33% of the genome is transcribed, and the remaining 67% is intergenic DNA between the genes.

The amount of intergenic DNA so computed contradicts an assertion (Wong, et al. 2000) that most of the human genome is transcribed. We believe that this assertion is still correct. Interestingly, the discrepancy is not due to these newer but smaller gene counts. The problem arises from the mean gene sizes, which everyone significantly underestimates because of sampling biases resulting from the lack of large genomic contigs. We note that 25%, 50%, and 75% of the public consortium's genome sequence is in contigs of sizes <21.0, <84.5, and <290.5 kb, respectively. In contrast, human genes can be much larger than these contigs. For example, the *dystrophin* gene on chromosome X is 2.3 Mb. The *neurexin-3* gene on chromosome 14 is 1.46 Mb, and one intron is 479 kb. It is impossible to determine the correct size of a large gene when its exons are scattered among smaller contigs. Insofar as estimates of mean gene size are concerned, the failure to determine the correct size of a megabase-sized gene is equivalent to the omission of a thousand of the smallest genes.

We introduce another estimate of the mean gene size—one that compensates for this contig-size bias, by focusing on the not insignificant fraction of the genome that is covered by large megabase-sized contigs. From this new estimate, we will argue that the newer but smaller gene counts are closer to being correct than the older but larger gene counts. Nevertheless, the mean gene size is so big that there is little room for intergenic DNA, and most of the human genome is transcribed.

RESULTS

It is important to recognize that, without full-length cDNA sequence, large genes are almost impossible to identify even with large genomic contigs. Protein homologies or ab initio gene-prediction algorithms might identify part of the gene, but not the entire gene. We can illustrate this problem. We have aligned all available cDNA sequences to the latest version of the human genome sequence. As a representative ab initio gene-prediction program, we selected `Genscan` (Burge and Karlin 1997) and determined the ratio of actual-to-predicted genomic extent, as a function of the gene size. It is obvious from Figure 1 that the performance of `Genscan` is severely degraded above 100 kb. This is not a criticism of the software, because the intrinsic signal-to-noise limitations involved in detecting such large genes would confound any algorithm.

However, even with full-length cDNAs, the mean gene size can be significantly underestimated if the genomic sequence is dominated by small contigs. The problem is that we can never fully account for all the large introns, because they are the ones most likely to be interrupted by breaks between the contigs. For example, *dystrophin* is known to be a 2.3 Mb gene with 79 exons (Tennyson et al. 1995). Our cDNA alignments identified 76 exons, across 16 contigs, with only 60 complete introns. Even after extrapolating for the unaligned 5′ UTR and coding regions, using the mean size of the observed introns, the extrapolated gene-size is only 1.0 Mb. We introduce here a method to minimize the effects of this contig-size bias. The basic idea is to restrict the computation of the mean gene size to those genes that are aligned to contigs above some cutoff size. We can then extrapolate to the limit of infinite contig sizes, by observing the trend in the mean gene size as a function of the cutoff size.

In practice, we can ramp the cutoff size up to 1 Mb, where there are 228 Mb of genomic sequence, primarily from chromosomes 6, 7, 14, 20, 21, and 22. Unfortunately, chromosomes 20 and 22 are GC-rich, at 0.440 and 0.477, relative to the genomewide average of 0.409. If we use every chromo-

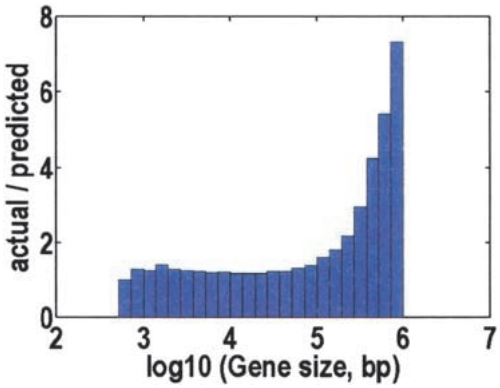

Figure 1 Distribution in ratio of actual-to-predicted genomic extent, as function of gene size. `Genscan` is given a sequence containing the cDNA-aligned exons and all intervening introns. When multiple genes are predicted, the one with the longest genomic extent is taken, as long as there is at least some overlap with the actual exons. `Genscan` performs well for small genes, but its performance is severely degraded for genes above 100 kb.

[3]Corresponding author.
E-MAIL gksw@u.washington.edu; FAX (206) 685-7344.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.202401.

some, we run the risk of dragging the mean gene size down by a disproportionate number of abnormally small genes. Therefore, in Figure 2 we depict two different datasets for the mean gene size, one computed with and one computed without these two GC-rich chromosomes. At contig cutoffs of 1 Mb, the mean gene size is 60.0 kb or 71.9 kb, over a dataset of 829 or 498 genes, respectively. To put this large-genes issue into perspective, consider that at the 1 Mb cutoff (without chromosomes 20 and 22), 16.5%, 6.2%, and 2.8% of the genes are larger than 100 kb, 250 kb, and 500 kb in size; but 70.4%, 48.7%, and 31.5% of all the transcribed sequence is attributable to these relatively few large genes.

It will be difficult to estimate the full extent of the contig-size bias until there are more megabase-sized contigs on more of the chromosomes. Even then, it seems likely that the cDNA data are underrepresenting the larger cDNAs, which are correlated with larger genes. In other words, the true mean gene size may be even larger than we have indicated. Moreover, given all of the ongoing arguments about the gene counts, the best we can do is indicate (as in Fig. 2) how large the mean gene size must be for there to be no intergenic DNA, at a gene count of 30,000, 35,000, and 40,000. The indicator arrows include a 3.5%, 4.1%, and 4.7% correction for overlapping genes, on the reverse strand or inside the introns, based on the actual number of observed overlaps, and then also corrected for the incomplete state of our cDNA data. Despite all these uncertainties, it is abundantly clear that the intergenic fraction is much smaller than 67%.

However, what if both gene count estimates, from the public and private Human Genome Projects, are wrong? This is certainly possible, but all the ongoing arguments are trying to push the gene counts *up*, which would make our assertion—that most of the human genome is transcribed—even more likely to be correct. For example, a recent analysis of the public and private gene sets revealed little overlap between novel genes (i.e., those with no representative in RefSeq), although most of these novel genes could be confirmed by expression analysis (Hogenesch et al. 2001). Although some of the missing overlap may be attributed to the gene-prediction programs identifying different fragments of the same gene, if we take this analysis at face value it indicates that there are up to 42,000 genes in the combined gene sets (or approximately the maximum number of genes that can fit into the human genome, given our estimates of the mean gene size). Conversely, our mean gene size estimates preclude the possibility of significantly larger gene counts, such as the previously accepted estimates of 70,000 genes.

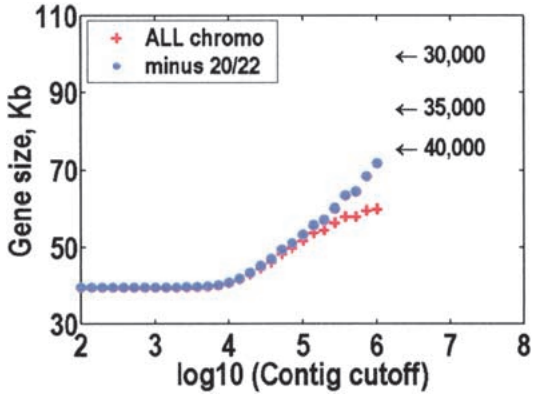

Figure 2 Trend in mean gene size, as function of contig cutoff. This analysis is restricted to cDNAs with at least some 5′ UTR sequence, because these are more likely to be full-length. When a gene is aligned to multiple contigs, the cutoff is based on the maximum contig size. In the limiting case of a 1-Mb contig cutoff, the mean gene size is 60.0 kb or 71.9 kb, depending on whether or not we include chromosomes 20 and 22. Had we based the cutoffs on the sum of the contig sizes, instead of on the maximum, the resultant gene sizes would have been even larger (65.6 kb or 81.0 kb). The arrows indicate how large the gene size must grow for there to be *no* intergenic DNA, assuming 2.9 Gb of euchromatic DNA and 30,000, 35,000, or 40,000 genes.

DISCUSSION

Why should the intergenic fraction be so small? Suppose the human genome was once compact, like *Fugu rubripes* (Venkatesh et al. 2000), and that its subsequent growth was driven by transposon activity. Assume that these transposons inserted into intronic and intergenic regions with equal probability. By transposons, we mean both common interspersed repeats, like *Alu* and L1 (Smit 1996), and most pseudogenes because 82% of pseudogenes also propagate by a transposition mechanism (Mighell et al. 2000). From our cDNA alignments, we know that there are 9.3 introns per gene, and so a reasonable lower bound for the intergenic fraction is 9.7%.

In what kinds of cells might these large genes be important? We believe that the common denominator will be a long cell cycle. For example, we know that 16 h are needed to transcribe the 2.3 Mb *dystrophin* gene (Tennyson et al. 1995). This does not necessarily imply that expression levels are reduced, because multiple polymerases can operate on the gene simultaneously. Indeed, an analysis of the EST data indicates that expression levels are independent of gene size (Wong et al. 2000). However, when the cell cycle is too short, the first mRNA will not be able to exit the nucleus before the cell has to undergo mitosis again. Therefore, proteins from large genes will be found mostly in terminally-differentiated cells, as in muscles and in the brain. Many of the largest genes that we found fit this description (e.g., *dystrophin*, *neurexin*, and many neurotransmitter receptors). One could argue that intron-size modulation is a form of gene regulation.

Most of the human genome is transcribed. However, the contrary idea—that little of the human genome is transcribed—is embedded deep in the popular consciousness. In fact, it is rarely discussed in the papers that make such an assumption. There may be significant implications for models of evolutionary biology, particularly given that although there is little intergenic DNA in the animal genomes, there is plenty of intergenic DNA in the plant genomes, including the compact genome of *Arabidopsis thaliana* (The *Arabidopsis* Genome Initiative 2000).

METHODS

Our cDNA-to-genomic alignment software was discussed in a previously published work (Wong et al. 2000). A set of 10,309 cDNAs was derived from 11,001 RefSeq cDNAs, by removing immune -system related and redundant cDNAs, and reinstating another 170 cDNAs that were not available at ftp://ncbi.nlm.nih.gov/refseq/H_sapiens on March 2, 2001. Genomic data were downloaded from ftp://ncbi.nlm.nih.gov/genomes/H_sapiens on February 27, 2001. Because scaffold joins do not include estimates of the gap size, we ripped these contigs apart at every instance of 20 or more N's, as gaps (or extended low-quality regions) are represented in the genomic sequence by long strings of N's. Significant hits were found

for 9142 of the cDNAs. After removing low-quality, ambiguous, and redundant alignments, we had 8281 usable alignments.

ACKNOWLEDGMENTS

We thank Maynard Olson, Lars Bolund, and Lee Rowen for their comments and suggestions. This analysis was partially supported by a grant from the National Institute of Environmental Health Sciences (1 RO1 ES09909).

NOTE ADDED IN PROOF

Recent high-density array analysis of human chromosome 21 has revealed that, even in gene "deserts" with absolutely no annotated genes, there are nevertheless exon-sized segments that are conserved between human-mouse and human-dog (Frazer et al. 2001). We believe that at least some of these segments are exons, which were not annotated because of the gene-prediction programs' inability to handle large genes.

REFERENCES

Antequera, F. and Bird, A.P. 1993. Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci.* **90:** 11995–11999.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Ewing, B. and Green, P. 2000. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.* **25:** 232–234.

Fields, C., Adams, M.D., White, O., and Venter, J.C. 1994. How many genes in the human genome? *Nat. Genet.* **7:** 345–346.

Frazer, K.A., Sheehan, J.B., Stokowski, R.P., Chen, X., Hosseini, R., Cheng, J.F., Fodor, S.P., Cox, D.R., and Patil, N. 2001. Evolutionarily conserved sequences on human chromosome 21. *Genome Res.* **11:** 1651–1659.

Hogenesch, J.B., Ching, K.A., Batalov, S., Su, A.I., Walker, J.R., Zhou, Y., Kay, S.A., Schultz, P.G., and Cooke, M.P. 2001. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **106:** 413–415.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Mighell, A.J., Smith, N.R., Robinson, P.A., and Markham, A.F. 2000. Vertebrate pseudogenes. *FEBS Lett.* **468:** 109–114.

Smit, A.F. 1996. The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6:** 743–748.

Tennyson, C.N., Klamut, H.J., and Worton, R.G. 1995. The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. *Nat. Genet.* **9:** 184–190.

The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.

Venkatesh, B., Gilligan, P., and Brenner, S. 2000. *Fugu*: A compact vertebrate reference genome. *FEBS Lett.* **476:** 3–7.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

Wong, G.K.S., Passey, D.A., Huang, Y.Z., Yang, Z., and Yu, J. 2000. Is "junk" DNA mostly intron DNA? *Genome Res.* **10:** 1672–1678.

Most of the Human Genome Is Transcribed

Gane Ka-Shu Wong, Douglas A. Passey and Jun Yu

Genome Res. 2001 11: 1975-1977
Access the most recent version at doi:10.1101/gr.202401