Cell, Vol. 106, 413–415, August 24, 2001, Copyright ©2001 by Cell Press

Letter to the Editor

A Comparison of the Celera and Ensembl Predicted Gene Sets Reveals Little Overlap in Novel Genes

The recent description of the human genome and the subsequent annotation of putative novel genes has ushered in a new era in biology. One of the revelations of the human genome project was the remarkably consistent prediction that the genome harbors around 30,000 genes. This observation was based on independent analyses done by a public genome consortium (29,691 transcripts, Ensembl v0.8) (Lander et al., 2001), by work done at Celera Genomics (39,114 transcripts) (Venter et al., 2001), and by Green and colleagues using expressed sequence tag (EST) clustering incorporating quality scores (35,000 genes) (Ewing and Green, 2000). This conclusion was surprising for two reasons. First, less complex organisms like *Arabidopsis* (25,000) and *C. elegans* (19,000) have approximately the same number of genes (*C. elegans* Sequencing Consortium, 1998; *Arabidopsis* Genome Initiative, 2000). Second, earlier estimates of gene number based on EST clustering and detailed chromosomal analysis were much higher, ranging from 45,000 to 140,000 (Dunham et al., 1999; Fields et al., 1994; Liang et al., 2000; Scott, 1999). While the Celera and Ensembl annotation efforts predicted approximately the same number of genes, a direct comparison of the predicted transcript sets has not been made. If the predictions are accurate and complete, then one would expect them to be largely overlapping.

To address this point, we compared the predicted transcript sequences from the two genome efforts with each other and with a well-curated set of 11,015 reference transcripts from Refseq using BLAST (Altschul et al., 1990; Pruitt et al., 2000). Given the difficulty of precisely predicting genes, we chose a permissive clustering method that requires only a short (>100 bp) region with at least 98% identity to combine transcripts into a single cluster. Using this method, transcripts that share only a single average size exon (~140 bp; Lander et al., 2001; Venter et al., 2001) cluster together. We first compared the Celera and Ensembl transcripts with the known genes from Refseq. The combined Celera and Ensembl datasets contained a fragment (at least 100 bp) of nearly all known genes (Figure 1). More than 84% of Refseq transcripts contained a match in both datasets, with the remaining Refseq genes matching either Celera (7%) or Ensembl (5%) alone. Surprisingly, when we compared the novel gene predictions that are not represented in Refseq, we found little agreement between the two transcriptomes. Collectively nearly 80% of the 31,098 novel transcripts were predicted by only one of the groups. Further breakdown of the Celera predicted transcripts shows that nearly all Celera transcripts supported by only a single line of evidence are unique to the Celera predictions. When these are removed from the analysis, 64% of the novel transcripts are predicted by only one group. Taken in sum, these data reveal that the predicted transcripts collectively contain partial nucleotide matches to nearly all known genes, but the novel genes predicted by both groups are largely nonoverlapping.

To validate the existence of the transcript predictions, we used RNA expression profiling and a bank of 13 diverse human tissues. The commercial high-density oligonucleotide arrays used are based on Expressed Sequence Tags (ESTs) represented in Unigene (release 95). BLASTN was used to assign the transcript predictions to a Unigene cluster, and the RNA expression pattern was determined for the 8,000 known and 5,000 novel predicted genes with a corresponding Unigene cluster on the arrays (see legend to Figure 2 for details). Using these methods, we found evidence of expression for more than 80% of the known genes in at least one of the tissue samples analyzed (Figure 2A). Similarly, more than 80% of the novel predicted transcripts were detected as expressed in at least one of the 13 human tissues. Hierarchical clustering and visualization of these expression data revealed a similar fraction of tissue-restricted transcripts for both known and novel genes (Figure 2B). These data support the view that the novel transcripts predicted by both groups encode bona fide differentially expressed mRNAs. Since many of these verified transcripts were contained in only one of the two predicted transcriptomes, we conclude that the computational methods used for gene prediction by either group are inadequate and that the respective transcriptomes are *individually* incomplete.

What could explain the discrepancies in the predicted

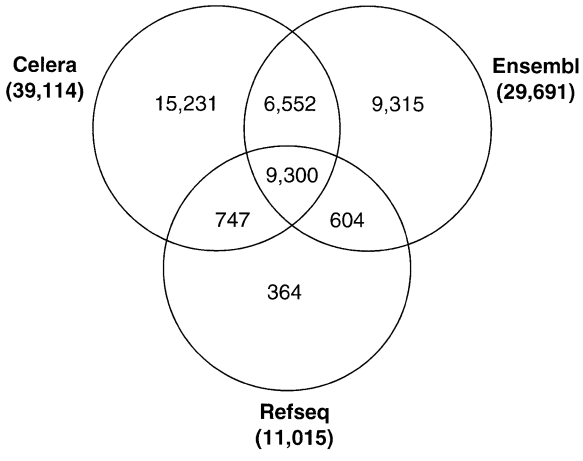

Figure 1. Nucleotide Comparison of Celera, Ensembl, and Refseq Transcripts

BLASTN was used to identify nucleotide matches between all datasets. Transcripts were clustered together when the aligned region had greater than 98% identity over at least 100 nucleotides. Numbers reflect the final cluster count, where multiple sequences from any dataset can be collapsed into one cluster. For this analysis a range of conditions from 92% to 99% identity over 100 bp was evaluated. 98% over 100 bp gave the most favorable balance of false positive and false negatives based on a BLASTN analysis of Refseq against itself.

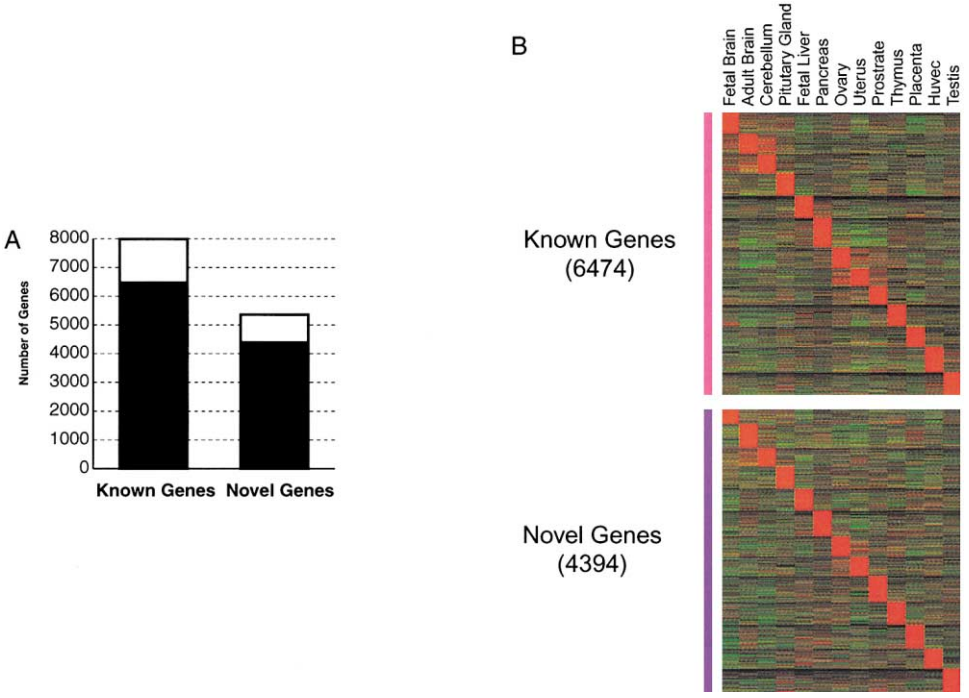

Figure 2. Use of RNA Expression Analysis to Validate Transcript Predictions

(A) Confirmation of predicted transcripts by RNA expression profiling. Known genes, novel predicted genes, and sequences represented on commercially available high-density oligonucleotide arrays were assigned to a Unigene cluster using BLASTN against a database of all human Unigene sequences (Hs.seq.all) as described in Figure 1. Based on this analysis array sequences were categorized as corresponding to known genes or novel genes. Transcript abundance was determined by RNA expression profiling using Affymetrix U95A and U95B high-density oligonucleotide arrays, as previously described (Welsh et al., 2001), across a panel of 13 tissues. A transcript was considered expressed if the average difference (AD) value as determined by the Genechip software package (Genechip v. 3.2, Affymetrix) exceeded 200 (approximately 3–5 copies per cell, Wodicka et al., 1997). Depicted is the fraction of transcripts for each class that scored as expressed (dark bars) or not expressed (light bars) in any of 13 tissues.

(B) Similar differential expression of known and novel genes. Quantitative average difference values for the expressed (AD $>$ 200) Unigene clusters from each of the two classes (Known and Novel genes) were visualized using Treeview (Michael Eisen, Stanford University [Eisen et al., 1998]) as previously described (Welsh et al., 2001). Genes from each group are on the vertical axis, with the indicated tissues arranged on the horizontal axis. Transcripts enriched for a given tissue are colored red and those that are repressed in a given tissue are colored green.

transcriptomes? The lack of overlap in the predicted transcripts may reflect differences in the underlying genome assembly, or the algorithms and types of evidence used for transcript prediction. While the draft nature of the human genome sequence may contribute to some of these discrepancies, similar findings from the annotation of the finished *Drosophila* genome support the view that the evidence-based gene prediction methods used may be too conservative (Gopal et al., 2001). An alternative approach would be to overpredict and use experimental methods such as RNA expression analysis to validate predicted transcripts (Reboul et al., 2001). Our initial results using high-density DNA arrays support such an approach. Collectively, these studies suggest caution in the use of the current predicted transcript sets and cast doubt on these latest estimates of human gene numbers. We conclude that an integrated approach combining computational predictions, human curation, and experimental validation will be required to complete a finished picture of the human transcriptome.

John B. Hogenesch,[1] Keith A. Ching,[1] Serge Batalov,[1] Andrew I. Su,[2] John R. Walker,[1] Yingyao Zhou,[1] Steve A. Kay,[1] Peter G. Schultz,[1,2] and Michael P. Cooke[1,3]

[1]The Genomics Institute of the Novartis Research Foundation
San Diego, California 92121
[2]Department of Chemistry
The Scripps Research Institute
La Jolla, California 92037

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature *408*, 796–815.

C. elegans Sequencing Consortium (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. Science *282*, 2012–2018.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., et al. (1999). The DNA sequence of human chromosome 22. Nature *402*, 489–495.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA *95*, 14863–14868.

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.

Gopal, S., Schroeder, M., Pieper, U., Sczyrba, A., Aytekin-Kurban, G., Bekiranov, S., Fajardo, J.E., Eswar, N., Sanchez, R., Sali, A., and Gaasterland, T. (2001). Homology-based annotation yields 1,042 new candidate genes in the Drosophila melanogaster genome. Nat. Genet. *27*, 337–340.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. International Human Genome Sequencing Consortium. Nature *409*, 860–921.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S.L., and Quackenbush, J. (2000). Gene index analysis of the human genome estimates approximately 120,000 genes. Nat. Genet. *25*, 239–240.

Pruitt, K.D., Katz, K.S., Sicotte, H., and Maglott, D.R. (2000). Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. Trends Genet. *16*, 44–47.

Reboul, J., Vaglio, P., Tzellas, N., Thierry-Mieg, N., Moore, T., Jackson, C., Shin-i, T., Kohara, Y., Thierry-Mieg, D., Thierry-Mieg, J., et al. (2001). Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in C. elegans. Nat. Genet. *27*, 332–336.

Scott, R. (1999). The future in understanding the molecular basis of life. In 11th International Genome Sequencing and Analysis Conference (Miami, FL).

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.

Welsh, J.B., Zarrinkar, P.P., Sapinoso, L.M., Kern, S.G., Behling, C.A., Monk, B.J., Lockhart, D.J., Burger, R.A., and Hampton, G.M. (2001). Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proc. Natl. Acad. Sci. USA *98*, 1176–1181.

Wodicka, L., Dong, H., Mittmann, M., Ho, M.H., and Lockhart, D.J. (1997). Genome-wide expression monitoring in Saccharomyces cerevisiae. Nat. Biotechnol. *15*, 1359–1367.

[3] Correspondence: cooke@gnf.org