Accurate Identification of Novel Human Genes Through Simultaneous Gene Prediction in Human, Mouse, and Rat

Colin Dewey, Jia Qian Wu, Simon Cawley, et al.

Genome Res. 2004 14: 661-664
Access the most recent version at doi:10.1101/gr.1939804

Accurate Identification of Novel Human Genes Through Simultaneous Gene Prediction in Human, Mouse, and Rat

Colin Dewey,[1,6] Jia Qian Wu,[5,6] Simon Cawley,[3] Marina Alexandersson,[4] Richard Gibbs,[5] and Lior Pachter[2,7]

[1]Department of Electrical Engineering, and [2]Department of Mathematics, University of California–Berkeley, Berkeley, California 94720, USA [3]Affymetrix Inc., Emeryville, California 94608, USA; [4]Fraunhofer-Chalmers Centre, SE-412 88 Gothenburg, Sweden; [5]Human Genome Sequencing Center, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA

We describe a new method for simultaneously identifying novel homologous genes with identical structure in the human, mouse, and rat genomes by combining pairwise predictions made with the SLAM gene-finding program. Using this method, we found 3698 gene triples in the human, mouse, and rat genomes which are predicted with exactly the same gene structure. We show, both computationally and experimentally, that the introns of these triples are predicted accurately as compared with the introns of other ab initio gene prediction sets. Computationally, we compared the introns of these gene triples, as well as those from other ab initio gene finders, with known intron annotations. We show that a unique property of SLAM, namely that it predicts gene structures *simultaneously* in two organisms, is key to producing sets of predictions that are highly accurate in intron structure when combined with other programs. Experimentally, we performed reverse transcription-polymerase chain reaction (RT-PCR) in both the human and rat to test the exon pairs flanking introns from a subset of the gene triples for which the human gene had not been previously identified. By performing RT-PCR on orthologous introns in both the human and rat genomes, we additionally explore the validity of using RT-PCR as a method for confirming gene predictions.

[Supplemental material is available online at http://hanuman.math.berkeley.edu/~cdewey/SLAMHMR/index.html.]

The difficulty of accurate ab initio gene finding has been well documented (e.g., Mathe et al. 2002), and yet ab initio methods remain important for identifying novel genes that may be rarely expressed, have different structures from typical genes, or do not have any known homologs and thus may have been missed by conventional methods. The completion of the mouse genome allowed for the first time a *comparative*-based annotation of the human and mouse, and several methods were developed to take advantage of the conservation between genes in order to enhance predictions (Waterston et al. 2002). Several assessments, both computational (Burset and Guigó 1996) and experimental (Guigó et al. 2003) have shown that comparative-based gene finders such as SLAM (Alexandersson et al. 2003), Twinscan (Korf et al. 2001), and SGP (Parra et al. 2003) outperform single-organism gene finders such as Genscan (Burge and Karlin 1997) and Genie (Reese et al. 2000).

Although comparative gene finders use sequence data from multiple genomes, most only predict in one genome at a time. Among gene finders that have been used to annotate entire genomes, a unique characteristic of the SLAM gene finder is its simultaneous prediction of genes having identical structure in two genomes. With the addition of a *third* genome, combining the results from two SLAM runs allows for the prediction of genes having identical structure in all three genomes. Previous studies (e.g., Rogic et al. 2002) have shown that combining gene predictions from different gene finders improves the accuracy of predictions. In this paper we combine SLAM gene predictions both to extend predictions to a third genome and to improve accuracy. We analyze the accuracy of our three-way predictions, as well as those resulting from combining predictions from other gene finders, in terms of the accuracy of predicted introns.

Finally, we show that our strategy for gene prediction using the human, mouse, and rat genomes leads to 924 novel human gene predictions (along with corresponding mouse and rat orthologs). One intron from each of a subset of these genes (48 in human and the corresponding 48 in rat) was experimentally tested by reverse transcription-polymerase chain reaction (RT-PCR) sequencing. Combined with our computational analysis, the experiments suggest that up to roughly 80% of our novel gene predictions correspond to transcribed sequence. Furthermore, the design of our experiments (simultaneous RT-PCR in both human and rat tissue) provides a method for concurrent validation of the RT-PCR technique for identifying novel gene orthologs.

METHODS

A homology map was constructed for the human (November 2002), mouse (February 2002), and rat genomes (November 2002; Bray and Pachter, 2004). The map was designed so that the maximal segment size was 300 kb; small enough for the subsequent running of the SLAM gene finder. The number of pieces was 10,613 with a median length of 105,710 bases, and coverage was 2.6 Gb (human), 2.3 Gb (mouse), and 2.3 Gb (rat). Two whole-genome SLAM runs were performed, one using the human and mouse genomes (which we refer to as the *hm* run), and the other the human and rat genomes (the *hr* run). For each of the runs, a pairwise homology map was projected from the three-way map, and SLAM was run on each of the blocks.

In order to compare SLAM with other gene finders, whole-

[6]These authors contributed equally to this work.
[7]Corresponding author.
E-MAIL lpachter@math.berkeley.edu; FAX (510) 642-2028.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1939804.

genome gene sets were obtained from the UCSC Genome Browser Web site (Kent et al. 2002).

All available human (November 2002) gene sets from ab initio gene finders (Geneid, Genscan, SGP, and Twinscan), as well as from evidence-based methods (ENSEMBL, Known genes, and RefSeq) were obtained. The ab initio sets along with the SLAM *hm* and SLAM *hr* sets were each compared pairwise to produce "consensus" sets of gene predictions that contained genes predicted identically in human in two different sets. The accuracies of the introns of the abinitio gene prediction sets and the consensus sets were measured by comparison with the introns of the human RefSeq gene set.

The consensus set for the SLAM *hm* and *hr* runs contained 3698 genes. It is important to note that by virtue of the SLAM constraints this set consisted of genes in human, mouse, and rat, all predicted to have **exactly the same structure**. In comparison with the set of all SLAM *hm* predictions, the consensus set is enriched for single-exon genes but has a similar distribution of coding sequence length. In the interest of finding novel genes, this set was filtered for those predictions that did not overlap at all with genes in the ENSEMBL, Known genes, and RefSeq sets (Guigó et al. 2003). This final set, which we call the filtered ortholog set, consisted of 924 genes.

We set out to confirm, using RT-PCR, one pair of exons flanking an intron from each of a subset of the filtered ortholog set in order to get an experimental estimate of the accuracy of these predictions. Using Primer3 (Rozen and Skaletsky 1996) and a variety of Perl scripts, the filtered ortholog set was screened for introns at least 1000 bp in length and having flanking exons with with suitably long primers (25–30 bp) capable of producing a PCR product with a length between 150 bp and 200 bp and a suitable melting temperature (67°–73°C).

Source RNA was pooled from 20 human tissues including adrenal gland, bone marrow, brain cerebellum, brain (whole), fetal brain, fetal liver, heart, kidney, liver, lung, placenta, prostate, salivary gland, skeletal muscle, spleen, testis, thymus, thyroid gland, trachea, and uterus (Clontech Human Total RNA master panel II), and 18 rat tissues including 10–12-d embryo, adrenal gland, bladder, brain (whole), brain cerebellum, colon, heart, kidney, liver, lung, ovary, spleen, testicle, thymus (Clontech), mammary gland, pancreas, placenta, and prostate (Ambion). Reverse transcription (RT) reactions were primed by OligodT using Superscript II reverse transcriptase (Invitrogen). The RT reactions were followed by PCR using Clontech Advantage 2 PCR Enzyme System. The PCR program was set at 95°C for 30 sec, followed by 35 cycles of 95°C for 10 sec, and 68°C for 30 sec. Finally, there was an extension cycle of 72°C for 1 min. The pair of exons flanking each intron to be tested were amplified with specific primers. RT-PCR products were examined by agarose gel electrophoresis (Figure 2, below). Kodak Digital Software was used to estimate the product sizes. PCR products were purified with a QIAquick 96-well PCR purification kit (QIAGEN) and sequenced using both forward and reverse primers for each predicted gene.

The amplified sequences were compared with the original SLAM predictions to verify the identity of recovered products. Sequence alignments were computed using standard penalties (match +1, mismatch −1, gap −2, gapExtend −1) and the resulting alignments were considered "valid" if they were at least 40 bp long, overlapped the boundaries of the predicted intron with its flanking exons, and contained 75% sequence similarity (determined by counting the number of matches and dividing by the alignment length). An intron was considered to be verified if the sequenced product had a valid alignment with the predicted product. The gene predictions for which the introns were tested were also subject to further analysis in the form of BLAST alignments against standard databases, and comparison with other existing gene annotations and EST evidence.

RESULTS

The results of the SLAM gene finding runs, the comparison of the intron predictions with known introns, and the confirmation of the intron predictions by RT-PCR are summarized in Table 1. The accuracy of the introns of all available ab initio whole-genome gene prediction sets and of the consensus sets generated from each pair is shown in Figure 1. A companion Web site at http://hanuman.math.berkeley.edu/~cdewey/SLAMHMR/index.html shows the tested genes, with information about the RT-PCR experiments and subsequent analysis. The genes are clickable to show the predictions in context on the UCSC genome browser (Kent et al. 2002), and the RT-PCR results (from two separate experiments) are summarized and the alignments of the sequence products are displayed. Finally, a short description of each gene is included with highlights of peculiar features of interest.

We mention a few interesting examples: the gene M4H1U1D4r70.005 contains five exons, and has an intron that was validated only in rat. The gene appears inside the intron of a known gene (NMNAT) but in the opposite strand. The gene M16H3U2D1r112.003 (validated in both human and rat) is known only in mouse, but the human/mouse gene predictions align with 97% identity and the human rat also with 97% identity. In fact, the prediction is part of an 18-exon gene (>1000 amino acids) that was known only in mouse! This illustrates the power of the comparative method to not only identify novel genes, but to extend annotations from one organism to another.

DISCUSSION

Our analysis of the accuracy of the *introns* of currently available whole-genome ab initio gene prediction sets and consensus sets generated from them reveals several important facts. First, unsurprisingly, comparative gene finders produce more accurate intron predictions than noncomparative ab initio gene finders. Every comparative gene prediction set analyzed (SGP, SLAM *hm*, SLAM *hr*, and Twinscan) had higher intron accuracy than the noncomparative gene prediction sets (Genscan and Geneid). In terms of exact intron predictions, the noncomparative gene prediction sets had a mean accuracy of 68% whereas the comparative sets had a mean accuracy of 77%.

The consensus gene prediction sets greatly improve accuracy (up to 98% accuracy), but at a large loss of sensitivity (the most accurate sets had just below 1000 introns overlapping a RefSeq intron). Figure 1 shows that similar gene finders, such as

Table 1. Summary Statistics

# of SLAM human/mouse genes	29370
# of SLAM human/rat genes	25427
# of SLAM genes identical in human, mouse, and rat	3698
# of SLAM human/mouse/rat introns	10577
# of SLAM human/mouse/rat introns overlapping human RefSeq introns	8499
% of SLAM human/mouse/rat introns with correct structure (out of introns overlapping human RefSeq)	90%
# of novel (not overlapping with human Ensembl, RefSeq, or Known genes) SLAM human/mouse/rat genes	924
# of SLAM human/mouse/rat introns tested	48 ortholog pairs (48 human, 48 rat)
% of SLAM human/mouse/rat introns verified	73% (28 pairs verified in both human and rat, 6 verified only in rat, 1 verified only in human)

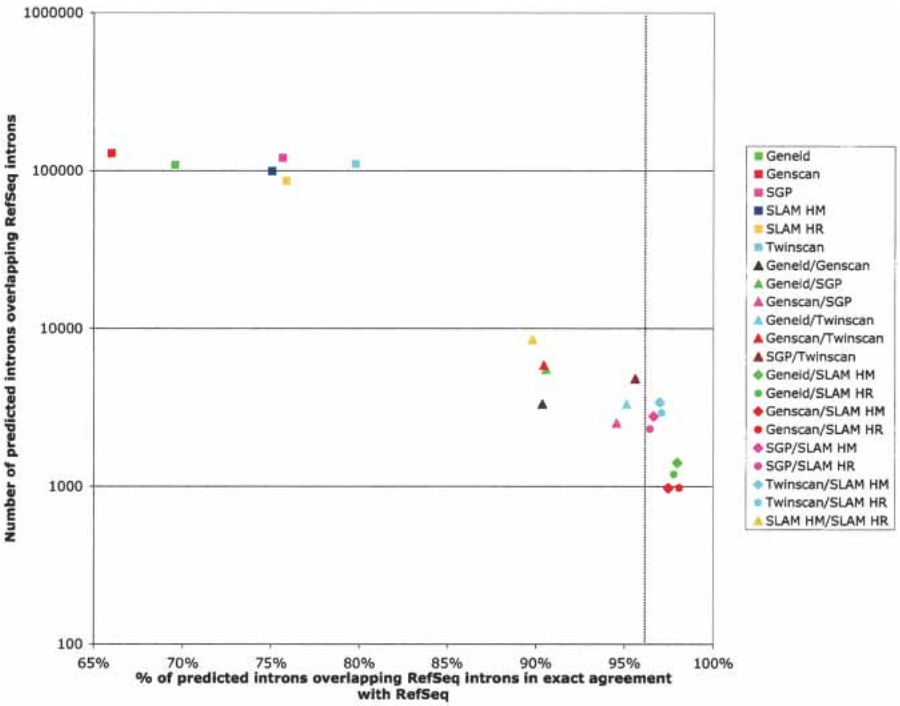

Figure 1 Enrichment for intron accuracy by gene prediction consensus. Intron accuracy of gene predictions by SLAM, other ab initio programs, and from consensus sets was measured by agreement with human RefSeq annotations. Comparative gene predictions utilizing two genomes were more accurate than those from noncomparative gene finders (Genscan and Geneid). Genscan, a noncomparative gene finder, had the greatest sensitivity, with close to 130,000 intron predictions overlapping RefSeq introns. Combining gene predictions to form consensus sets greatly increases accuracy while significantly reducing sensitivity. The vertical dotted line indicates the limit of the accuracy of consensus sets not involving SLAM.

Twinscan and Genscan or SGP and Geneid, give rise to larger consensus sets with lower accuracy. It follows from this observation that the most dissimilar gene finders can be combined to give the most accurate consensus sets. A case of combining the results from extremely similar gene finders is the SLAM *hm*/SLAM *hr* consensus set, which utilizes the same exact gene finder but different comparative data sets. Of all the consensus sets, this set is the largest and has the lowest accuracy. It strikes a good compromise between accuracy and sensitivity, with about 8500 intron predictions overlapping RefSeq introns and an accuracy of 90%, a 10% improvement over the accuracy obtained by the best comparative gene finders by themselves. The largest consensus set with accuracy above 91% (Twinscan/SGP) has only 4806 predictions. It is important to note that the SLAM *hm*/SLAM *hr* set is also unique in that it represents simultaneous predictions of orthologs in the human, mouse, and rat genomes.

The fact that SLAM is required for all consensus sets with accuracy greater than 96% indicates that SLAM is quite different than other gene finders. The most accurate consensus sets had accuracies up to 98% and were those resulting from combining SLAM with the two noncomparative gene finders, Genscan and Geneid. Some aspect of SLAM's comparative nature must account for its uniqueness, as SLAM and Genscan are based on very similar gene models. As the consensus sets involving the other comparative gene finders had lower accuracy than those involving SLAM, it is likely that it is SLAM's ability to predict simultaneously in two genomes (Pachter et al. 2002) that sets it apart from the other gene finders and allows for more accurate consensus sets. Unfortunately, this analysis shows that there is much room for improvement in gene-finding models; we are still far from having a gene finder that it is both sensitive and accurate in terms of intron structure, as the sets with accuracy of at least 90% all had less than 9000 intron predictions (the RefSeq annotation set contains around 150,000 introns). This reflects our continued lack of understanding of the biological mechanisms underlying transcription and posttranscriptional modification such as splicing.

In other computational analyses where we have analyzed exon and whole-gene structure accuracy, the results are essentially the same as for the intron analysis (see Supplemental Data). Consensus sets including SLAM have the highest accuracies: up to 95% at the exon level, and 83% at the whole-gene level. Interestingly, consensus-set intron accuracy was greater than consensus-set exon accuracy, and nonconsensus-set intron accuracy was lower than nonconsensus-set exon accuracy. This suggests that introns are generally harder to predict accurately than exons, but that by using gene-finder consensus, the task becomes much easier. This seems important in light of the fact that RT-PCR experiments validate introns and not exons.

Our RT-PCR intron validation rates (60% in human, 71% in rat, 66% overall, and 73% for intron pairs when requiring validation in only one organism) are encouraging compared to the rates obtained in a previous study by Guigó et al. (2003). In that study, RT-PCR validation rates obtained for predicted mouse introns were 62% for all introns and 76% for introns predicted by two different programs (SGP and Twinscan). We are encouraged by this comparison, as we were looking for novel transcripts in the human genome (and not in mouse). Because the human genome is much better characterized, it seems that it should be harder to detect novel genes, and indeed many of our products were only expressed in a small fraction of tissues. The fact that only seven out of 48 tested genes were validated in only one organism reinforces the idea that RT-PCR is a legitimate method for validating genes. In three of these seven cases, an amplified product similar to the predicted product was sequenced in the other organism, but did not have a good enough alignment with the predicted product (perhaps due to sequence quality issues) to be considered validated. These results confirm the intuition that SLAM gene predictions are correct in both orthologs or in neither. It also suggests that genes rarely expressed in one organism may be rarely expressed in another. Despite the correlated results, our tests confirm that multiple RT-PCR in different organisms improves gene validation rates.

Although we did not explicitly study alternative splicing here, we have implemented a sampling strategy for SLAM in which it is possible to sample alternative orthologous transcripts instead of obtaining just one prediction (Cawley and Pachter 2003). Preliminary testing suggests that the human–mouse, human–rat prediction SLAM run combination described in this paper should yield alternative transcripts conserved between the mammals. At this point, the extent of such conservation is still unclear (Modrek and Lee 2003; Nurtdinov et al. 2003; Thanaraj et al. 2003).

By using our intron accuracy rates obtained both computationally and experimentally, we can make an estimate of the

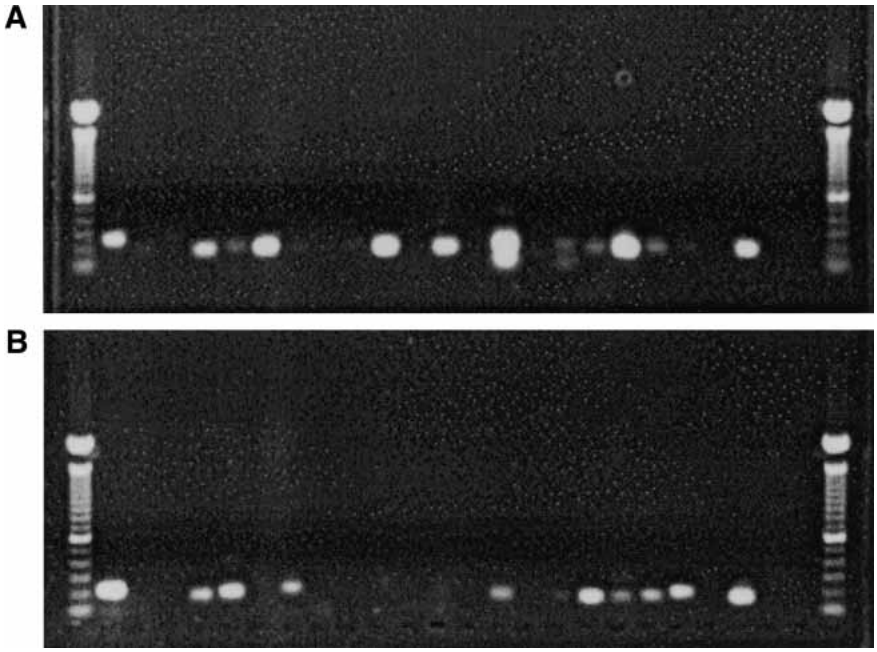

Figure 2 Gels used as part of RT-PCR validation of predicted introns. RT-PCR was used to validate one intron from 48 predicted novel human genes and their simultaneously predicted rat orthologs. The gels contain the RT-PCR sequenced products from human (*A*) and rat (*B*). Each column in the gel contains the sequenced product for one intron tested, with the same columns in the human and rat gels containing products from orthologous predictions. The *leftmost* and *rightmost* columns contain the ladders used to determine the lengths of the sequenced products. The second column contains a positive control (RT-PCR of an exon-pair from actin), and the third column contains a negative control (no reverse transcriptase used).

number of novel gene predictions obtained by our method. If we assume that the RefSeq gene annotation set is somewhat representative of the entire human gene set, then we estimate from our computational analysis that given that a SLAM *hm*/SLAM *hr* predicted intron overlaps with a real intron, it will be 100% accurate 90% of the time. Because of our RT-PCR procedures, we could only validate predicted introns that were 100% accurate. Therefore, given a 73% validation rate for our predicted introns (where we consider an intron validated if the predicted product is sequenced in either organism), we estimate that 0.73/0.90 = 81% of the introns in the SLAM *hm*/SLAM *hr* set overlap with a real gene. With 322 genes from the filtered ortholog set (potentially novel genes) possessing one or more introns, a rough estimate of the number of novel multiexon genes that can be discovered and validated (by RT-PCR validation of an intron) through this method is 322 × 81% = 260.

It is important to note that RT-PCR validations of predicted genes, as undertaken here as well as by Guigó et al. (2003), are conditioned on the ability to select suitable primer pairs. Although our success rate suggests that we can identify many new genes in the human genome, a direct extrapolation is likely inacurrate because of the difficulty in selecting primers in some of the examples. RT-PCR validations of predicted genes are also affected by the gene expression locations and time. Some genes may not have been validated due to our limited RNA source. Furthermore, our results on the validation of novel genes should be interpreted more strictly as validation of transcribed regions, because the experiments do not directly measure protein concentration levels.

ACKNOWLEDGMENTS

L.P. and C.D. were partially supported by NIH grant R01 HG2362-2. The whole-genome SLAM runs were performed on the Affymetrix computing cluster. R.G. and J.Q.W. were partially supported by grants from the NHGRI/NHLBI (1 U54 HG02345) and NCI/SAIC (20XS182A).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Alexandersson, M., Cawley, S., and Pachter, L. 2003. SLAM—Cross-species gene finding and alignment with a generalized pair hidden markov model. *Genome Res.* **13:** 496–502.

Bray, N. and Pachter, L. 2004. MAVID: Constrained ancestral alignment of multiple sequences. *Genome Res.* (this issue).

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

Burset, M. and Guigó, R. 1996. Evaluation of gene structure prediction programs. *Genomics* **34:** 353–367.

Cawley, S. and Pachter, L. 2003. HMM sampling and applications to gene finding and alternative splicing. *Bioinformatics* **19:** ii36–ii41

Guigó, R., Dermitzakis, E.T., Agarwal, P., Ponting, C., Parra, G., Reymond, A., Abril, J.F., Keibler, E., Lyle, R., Ucla, C., et al. 2003. Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proc. Natl. Acad. Sci.* **100:** 1140–1145.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The Human Genome Browser at UCSC. *Genome Res.* **12:** 996–1006.

Korf, I., Flicek, P., Duan, D., and Brent, M.R. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* **1:** S1–S9.

Mathe, C., Sagot, M-F., Schiex, T., and Rouze, P. 2002. SURVEY and SUMMARY: Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.* **30:** 4103–4117.

Modrek, B. and Lee, C.J. 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.* **34:** 177–180.

Nurtdinov, R.N., Artamonova, I.I., Mironov, A.A., and Gelfand, M.S. 2003. Low conservation of alternative splicing patterns in the human and mouse genomes. *Hum. Mol. Genet.* **12:** 1313–1320.

Pachter, L., Alexandersson, M., and Cawley, S. 2002. Applications of generalized pair hidden Markov models to alignment and gene finding problems. *J. Comp. Biol.* **9:** 389–399.

Parra, G., Agarwal, P., Abril, J.F., Wiehe, T., Fickett, J.W., and Guigó, R. 2003. Comparative gene prediction in human and mouse. *Genome Res.* **13:** 108–117.

Reese, M.G., Kulp, D., Tammana, H., and Haussler, D. 2000. Genie—Gene finding in *Drosophila Melanogaster*. *Genome Res.* **10:** 529–538.

Rogic, S., Ouellette, B.F.F., and Mackworth, A.K. 2002. Improving gene recognition accuracy by combining predictions from two gene-finding programs. *Bioinformatics* **16:** 1034–1045.

Rozen, S. and Skaletsky, H.J. 1996. Primer3. Code available at http://www-genome.wi.mit.edu/genome_software/other/primer3.html

Thanaraj, T.A., Clark, F., and Muilu, J. 2003. Conservation of human alternative splice events in mouse. *Nucleic Acids Res.* **31:** 2544–2552.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

WEB SITE REFERENCES

http://hanuman.math.berkeley.edu/~cdewey/SLAMHMR/index.html; Supplemental data.

Received November 5, 2003; accepted in revised form January 26, 2004.