

Identification of Rat Genes by TWINSKAN Gene Prediction, RT-PCR, and Direct Sequencing

Jia Qian Wu,¹ David Shteynberg,² Manimozhiyan Arumugam,² Richard A. Gibbs,¹ and Michael R. Brent^{2,3}

¹Human Genome Sequencing Center, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA; ²Laboratory for Computational Genomics, Washington University, St. Louis, Missouri 63130, USA

The publication of a draft sequence of a third mammalian genome—that of the rat—suggests a need to rethink genome annotation. New mammalian sequences will not receive the kind of labor-intensive annotation efforts that are currently being devoted to human. In this paper, we demonstrate an alternative approach: reverse transcription-polymerase chain reaction (RT-PCR) and direct sequencing based on dual-genome de novo predictions from TWINSKAN. We tested 444 TWINSKAN-predicted rat genes that showed significant homology to known human genes implicated in disease but that were partially or completely missed by methods based on protein-to-genome mapping. Using primers in exons flanking a single predicted intron, we were able to verify the existence of 59% of these predicted genes. We then attempted to amplify the complete predicted open reading frames of 136 genes that were verified in the single-intron experiment. Spliced sequences were amplified in 46 cases (34%). We conclude that this procedure for elucidating gene structures with native cDNA sequences is cost-effective and will become even more so as it is further optimized.

The publication of a draft sequence of the rat genome provides an exciting opportunity to find orthologs and novel paralogs of known human genes that can be studied in a well established physiological and pharmacological model (Rat Genome Sequencing Project Consortium 2004). It also suggests a need to rethink genome annotation. New mammalian sequences will not receive the kind of labor-intensive annotation efforts that are currently being devoted to human, and thus high-quality, automated genome-wide annotation will be required. The primary automated method used in the initial publication of the rat genome (as well as the public human and mouse genomes) was the Ensembl annotation pipeline (Hubbard et al. 2002). The core of Ensembl is GeneWise (Birney and Durbin 2000), an algorithm which (1) aligns known rat proteins from RefSeq (Pruitt and Maglott 2001) and SWISS-PROT (Bairoch and Apweiler 2000; Gasteiger et al. 2001) to their source in the rat genome, and (2) aligns known proteins from other mammals to the regions of the rat genome that could produce similar (orthologous or paralogous) proteins. In this paper, we describe a complementary approach: reverse transcription-polymerase chain reaction (RT-PCR) and direct sequencing based on dual-genome de novo predictions from TWINSKAN (Korf et al. 2001; Flicek et al. 2003; <http://genes.cse.wustl.edu/>). To test this method, we chose TWINSKAN predictions in rat with significant similarity to a well characterized set of human genes—those in the Human Gene Mutation Database (HGMD, Stenson et al. 2003)—but which were partially or completely missed by Ensembl.

Systematic RT-PCR and direct sequencing of many novel gene predictions was first reported three years ago (Miyajima et al. 2000; Das et al. 2001). Earlier this year, we published the first report of high success rates with RT-PCR and direct sequencing from hundreds of mammalian gene predictions (Guigó et al. 2003). We tested multi-exon mouse genes predicted by TWINSKAN and/or SGP2 (Parra et al. 2003) but not by Ensembl. The predictions that showed homology to a human gene prediction

with an intron in a conserved location had excellent rates of verification. For TWINSKAN, 44% of the mouse predictions that showed homology to a human prediction without an intron in a conserved location were also verified. The experiments reported by Guigó et al. 2003, Miyajima et al. 2000, and Das et al. 2001 focused on verifying the *existence* of predicted genes, rather than their complete structures. Thus, they used primers in adjacent exons flanking a single intron and judged the verification successful if a spliced product was amplified and sequenced from the primer pair. In the present study, we advanced this approach by attempting to amplify the complete open reading frame (ORF) of a subset of the genes we can verify with primers flanking a single intron. To simplify the procedure we pooled RNA from multiple tissues and sequenced all product mixtures. Finally, we developed software to automate the sequence analysis.

RESULTS

Summary of RT-PCR Results

We obtained spliced sequence from RT-PCR products for 59% of predictions for which we designed primers in adjacent exons flanking a single intron. For a subset of the single-intron successes, we designed primers in exons at or near the end of the predicted ORF to amplify most of the ORF and in the presumed untranslated regions (UTRs) to amplify the complete ORF. The success rates were 70% and 34%, respectively (Table 1).

Predictions

TWINSKAN predicted 24,490 genes comprising 182,013 exons on build 2.1 of the rat genome (January 2003). These and other statistics were similar to those obtained from the TWINSKAN annotation of human (Flicek et al. 2003)—the 8% reduction in predicted exons relative to human is not surprising given the incompleteness of this early draft of the rat genome. Our first goal was to identify a set of TWINSKAN predictions that (1) are not predicted by Ensembl, (2) are homologous to human HGMD genes, and (3) can be shown experimentally to be transcribed. Our plan was to test a single intron using primers in the flanking

³Corresponding author.

E-MAIL brent@cse.wustl.edu; FAX (314) 935-7302.

Article and publication are at <http://www.genome.org/cgi/doi/10.1101/gr.1959604>.

Table 1. Summary of RT-PCR Success Rates

Target type	Number of targets	Number (%) successful
Single intron	444	260 (59%)
Most of ORF (TE primers)	152	106 (70%)
Complete ORF (UTR primers)	136	46 (34%)

Targets for TE and UTR primers were selected from the single intron successes.

exons (Fig. 1). We therefore used BLAST to divide the TWINSKAN-predicted exon pairs into two sets: those that, when joined by splicing out the intron between them, were highly similar to an mRNA predicted by Ensembl (BLASTN $E < 10^{-6}$), and those that were not (predicted single exon genes were not used). This yielded 129,757 predicted exon pairs that were similar to Ensembl mRNAs and 52,256 that were not. After translating the exon-pairs that were *not* similar to Ensembl mRNAs in the predicted reading frame, we identified the non-Ensembl exon pair that was most similar to each of the human genes in the HGMD database. We found that 859 of the HGMD genes (73%) had a highly significant match ($E < 10^{-10}$) to one or more predicted exon pair that did not match Ensembl predictions. This does not imply that Ensembl missed 73% of the HGMD orthologs—the non-Ensembl exon pairs we identified by this method had significant similarity to HGMD genes, but could have come from paralogous family members rather than orthologs. To narrow this set still further, we took only those predicted exon pairs that were the best match for one and only one HGMD gene.

Single-Intron Experiments

Of the 554 remaining exon pairs, we were able to design primers that met our criteria for 444 (see Methods for details). All 444 primer pairs were synthesized and used in RT-PCR with pooled rat RNA from diverse tissues and developmental stages. PCR products were purified and sequenced using both forward and reverse primers (see <http://genes.cse.wustl.edu/rat-data-03/> for sequences and accessions). As a control for genomic DNA contamination, the reverse transcription reaction was performed without reverse transcriptase, and the subsequent PCR procedure was carried out using primers in exons 3 and 4 of rat *p53* (NM_030989). Gel analysis did not reveal any amplification, and therefore no genomic DNA contamination is indicated.

The resulting sequences were then analyzed to determine whether the primers amplified a spliced product from the target region. If so, the experiment was categorized as successful. From the 444 experimental wells, the analysis revealed 196 successful amplifications of spliced sequences (44%). Products from another 142 experiments failed to yield 20 consecutive bases of

high-quality sequence, either because nothing was amplified or because sequencing failed. The remaining 107 primer pairs yielded sequences that did not provide a reliable indication of whether the product was spliced. Repetition of this experiment using the same primers but adding DMSO and glycerol to the PCR mixture yielded an overlapping set of 166 successes. Combining the two experiments, the total number of verified introns was 260 (59%). Among these successes, the intron boundaries were exactly as predicted by TWINSKAN in 189 (73%).

Each 96-well plate contained two tests with primers in exons 3 and 4 of rat *p53*, an intron splicing test with primers in exon 3 and intron 3, and an intron splicing test with primers in introns 7 and 8. In the initial experiment, tests with primers in exons 3 and 4 of rat *p53* yielded spliced sequence in eight cases and failures in two cases on different plates. In one failure there was probably no amplification, as both the forward and reverse reads, which were produced in different sequencing runs, contained no high-quality sequence. The other failure yielded a short sequence that matched one exon but did not cross the splice boundary. During the repetition of the experiment, three of the five positive controls did not yield sequences. This could be due to failure during high-throughput PCR amplification, purification, or sequencing. In both experiments, the intron splicing tests with primers located in intron–intron or intron–exon amplified and produced sequence that was *not* spliced. This is likely due to unspliced pre-mRNA. Therefore, amplification should not be interpreted as implying that the primers annealed to exonic sequence (see Discussion).

Full ORF Experiments

To learn more about the gene structures that gave rise to these spliced products, we wanted to amplify and end-sequence the complete predicted ORFs. Thus, we attempted to design primers in the region between 10 and 300 bases upstream of the start codon and downstream of the stop, in the presumptive UTRs. In case of failure, we wanted to be able to determine whether the predicted gene might have been only part of the real transcript, such that most of the predicted exons were correct but the gene boundaries were not. Thus, we also attempted to design primers just inside the boundaries of the ORF in the outermost of the exons with enough coding sequence to facilitate primer design (see Fig. 1 and Methods). We refer to these as the terminal exon (TE) primers, even though short initial and terminal exons were bypassed in the primer design.

We attempted to design 93 UTR and 93 TE primer pairs for predictions that yielded successful results on the first set of single-intron experiments. Product size was limited to 4 Kb, because the efficiency of the PCR methods used was expected to decline rapidly at greater product lengths. Because product length was limited and primer design is not always successful, we ended up with 93 TE primer pairs and 89 UTR primer pairs.

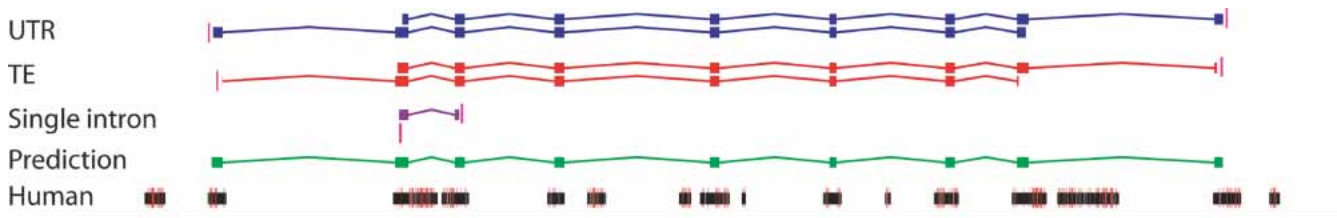


Figure 1 A TWINSKAN prediction (green, subsequently identified as rat aspartylglucosaminidase) in which sequences from the single intron (purple), terminal exon (red), and UTR (blue) primer pairs all yielded spliced alignments. Primers are shown as tall pink blocks at the same level as the sequence they yielded (the left single-intron primer did not yield high-quality sequence). TWINSKAN makes use of the blocks of alignment from the human genome (black) and the mismatches and gaps within the blocks (red) in predicting the most likely gene structure.

After the experimental procedure described above, sequence analysis revealed amplification of a spliced product overlapping the predicted gene from 64 TE experiments (69%); 13 yielded no high-quality sequence from either primer, and 16 yielded sequence that did not produce a reliable spliced alignment to the targeted region. The corresponding numbers for the UTR experiments were 26 successful (29%), 22 with no high-quality sequence, and 41 with no spliced product from the predicted region.

After the completion of this experiment, a new assembly of the rat genome was released in June 2003 (release 3.1). This assembly included 1500 new BACs yielding a net increase of 70 million bases in the assembly. In order to determine the effect of this improvement on the results described above, the 22 UTR primer pairs that yielded no high quality-sequence were remapped to the new genome sequence. This revealed that 19 of the 22 lay in regions which had been extensively revised such that the original primers would not be expected to amplify any product in the new genome sequence. We then remapped all the exon pairs that yielded positive results in the single-intron experiment to the new genome build and extracted the full-ORF predictions containing them. A new set of TE and UTR targets was selected, including 22 TE and 46 UTR targets that failed on the first UTR plate, as well as 59 new TE and 47 new UTR targets. Six of the TE targets that had been negative before came out positive (27%), as did 36 of those that had not been previously tested (61%). Five of the UTR targets that had been negative before came out positive (10%), as did 15 of those that had not been previously tested (32%). Combining both experiments, ORFs containing 152 distinct RT-PCR positive exon pairs were tested with TE primers, of which 106 were successfully amplified and end sequenced (70%). For the UTR primers, ORFs containing 136 distinct RT-PCR positive exon pairs were tested, of which 46 were successfully amplified and end sequenced (34%). All but five of the UTR successes were also TE successes.

In order to compare our sequence-based analysis to a different assay for RT-PCR success, we ran the products on a 1% agarose gel and estimated the product sizes. Products were classified according to whether they yielded a single band within 400 nucleotides (nt) of the expected size, multiple bands at least one of which was within 400 nt of the expected size, or no band within 400 nt of the expected size. Each group was divided into those that yielded a "hit" by our sequencing and analysis method and those that did not (Table 2). Most of the products that yielded a single band of the right size had been classified as hits (93%), and most of the hits yielded at least one band of the right size (65%). However, 35% of the hits yielded no visible band of the expected size, despite yielding high-quality spliced sequence that matched the expected genomic location. RT-PCR successes may have been missed by gel analysis because sequencing is more sensitive at low template concentrations, because the accuracy of

gel sizing is limited, or because of discrepancies between the predicted size and the actual size. Most of the products that were not classified as hits by sequence analysis yielded no band of the expected size (80%), but 18% yielded multiple bands including one of the expected size. Examination of sample traces from the products that yielded a band of the expected size along with other bands revealed that multiple templates were sequenced in every case, regardless of whether the product was classified as a hit by sequence analysis. This could have been the result of alternative splicing or mispriming. Five of the products that were not classified as hits by sequence analysis (2%) yielded a single band of the expected size. All of these turned out to be cases in which multiple templates were visible in the sequencing trace but not in the gel analysis.

Comparison of a Sample of Confirmed Rat Genes to Known Human Genes

Querying NCBI's nonredundant protein database (nr) with the UTR-confirmed gene predictions revealed that eight of them had become provisional or curated rat RefSeqs since we picked them as targets on March 28, 2003. In all eight cases, the name of the rat RefSeq matched that of an HGMD human gene. To get a sense of the kinds of genes we had verified, we investigated the 38 that had not become rat RefSeqs (Table 3). In 33 cases the rat prediction appeared likely to be the true ortholog of the top human hit in nr. Of these, 31 were known genes and two were predictions. Of the 31 known genes, 17 were in HGMD. Most of the rat predictions were about the same length as their putative human orthologs, although alignments sometimes suggested a missed splice site or exon or an alternative splice. In six cases, however, the prediction was more than 100 amino acids shorter than the putative ortholog. Because the high-quality portions of the reads did not include both the predicted start and stop codons in these cases, the primer may have annealed to a coding exon or an intron, rather than a UTR. In five cases, the verified rat prediction did not appear to be orthologous to a known human gene, and may therefore have been a novel paralog.

DISCUSSION

In these experiments we used TWINSKAN gene prediction followed by RT-PCR and direct sequencing to confirm partial gene structures in the newly sequenced rat genome. All 444 of the targets had significant similarity to known human genes, but they were at least partially missed by the April 2003 Ensembl annotation of the rat genome. Among these, 260 were confirmed by primers spanning a single intron, 106 of the single-intron positives were also confirmed using primers in the outermost predicted exons, and 46 of the single-intron positives were also confirmed using primers in the predicted UTRs. Among the single-intron positives we tested, 70% were positive in the terminal-exon experiments and 34% in the UTR experiments.

The higher success rate with TE primers relative to UTR primers probably results from a combination of factors. First, terminal exon primers were designed in the outermost exons with *sufficient coding sequence*; the shortest exons, which are most likely to be mispredicted, were omitted from the TE experiment for reasons of primer design. Second, we designed UTR primers within 300 bp of the predicted ORF; many of these 300-bp regions may have included sequence outside the UTRs or in UTR introns. Finally, gene prediction algorithms are more accurate in identifying exons than in grouping them into transcripts; this makes it likely that some of the predicted initial and terminal exons were actually internal exons, and hence the predicted UTR regions flanking them were introns.

Table 2. Comparison of Gel Analysis Assay for RT-PCR Success to Sequence Analysis Assay

Sequence analysis	One right size band	Multiple bands \geq right size	No right size band	Total
Hit	65	35	54	154
Non-hit	5	38	166	209
Total	70	73	220	363

Products that gave high-quality sequence with a clear spliced alignment to the genome in the target region are classified as hits; all others are classified as non-hits.

Table 3. Thirty-Eight Predicted Genes That Were Partially Verified With Primers in the Predicted UTRs but Did Not Become Rat RefSeqs Between Our Experiment and Submission of This Paper

Rat prediction	AA	Top human nr hit	Accession	AA	%ID	Ortho?	Synt?	dS	HGMD hit if different	Accession	%ID
Rn21.chr10.25.002	443	GABA receptor α -1 precursor	NP_000797	456	99	Yes	Yes	0.55			
Rn21.chr7.82.002	438	Exostosin-1 (EXT1)	NP_000118	746	97	Yes	Yes	0.27			
Rn31.chr3.13.009	325	LIM homeobox TF 1 β (LMX1B)	NP_002307	372	97	Yes	Yes	1.5			
Rn21.chr1.2.39.013	517	T-box 5	NP_000183	518	96	Yes	Yes	1.13			
Rn21.chr2.176.012	747	TRIFC	AAP51206	759	95	Yes	Yes	0.46	Midline 1	NP_000372	24
Rn21.chr4.130.006	213	Microphthalmia-associated TF	NP_000239	419	94	Yes	Yes	0.51			
Rn31.chr1.200.001	572	Dihydropyrimidinase-like 4 (DPYSL4)	NP_006417	572	93	Yes	Yes	0.93	Dihydropyrimidinase	NP_001376	59
Rn31.chr9.15.002	188	G protein-coupled receptor 136	AAP72128	350	93	Yes	Yes	0.39	Retinal GPCR	NP_002912	28
Rn31.chr1.4.37.010	339	Sarcoglycan γ (dystrophin-associated)	NP_000223	318	93	Yes	Yes	1.1			
Rn21.chr1.237.016	1090	Glycine dehydrogenase (GLDC), mitochondrial precursor	NP_000161	1020	92	Yes	Yes	0.83			
Rn21.chr1.92.009	448	Human Prediction	XP_293343	518	92	Yes	Yes	0.56	Kell blood group precursor	NP_066569	42
Rn21.chr2.93.001	439	Hepatocyte nuclear factor 4 γ chain	NP_004124	408	91	Yes	Yes	1.12	Hepatocyte nuclear factor 4- α	NP_849180	64
Rn31.chr7.105.006	447	N-myc downstream-regulated gene 1	NP_006087	394	90	Yes	Yes	1.08			
Rn31.chr8.16.004	718	N-acetylated α -linked acidic dipeptidase 2	NP_005458	740	89	Yes	Yes	0.59	Transferrin receptor 2	NP_003218	28
Rn21.chr8.61.011	522	Bardet-Biedl syndrome 4	NP_149017	519	88	Yes	Yes	0.41			
Rn31.chr10.56.021	1158	PER1 (circadian pacemaker protein)	NP_002607	1290	87	Yes	Yes	0.49	PER2 (circadian pacemaker)	NP_073728	44
Rn21.chr14.79.012	629	SH3-domain binding protein	NP_003014	561	86	Yes	Yes	0.74			
Rn31.chr1.4.38.008	398	TXK tyrosine kinase, PTK4	NP_003319	527	86	Yes	Yes	0.73			
Rn31.chr8.49.002	247	Human Prediction	XP_113683	235	84	Yes	Yes	0.55	Bruton tyrosine kinase (BTK)	NP_000052	57
Rn31.chr19.25.027	436	Glutaryl-CoA dehydrogenase	NP_000150	438	84	Yes	Yes	1	Myelin protein zero	NP_000521	32
Rn21.chr1.79.002	594	Glucose phosphate isomerase/neuroleukin	NP_000166	558	82	Yes	Yes	0.85			
Rn21.chr1.6.39.002	345	Aspartylglucosaminidase	NP_000018	346	82	Yes	Yes	0.42			
Rn21.chr1.207.007	214	Iduronate 2-sulphatase (IDS)	NP_000193	550	81	No	No	1.79			
Rn21.chr1.7.48.008	292	Na/PO4 cotransporter (SLC17A4)	NP_005486	497	81	Yes	Yes	0.44	Solute carrier family 17 (anion/sugar transporter), member 5	NP_036566	46
Rn21.chr5.183.007	760	5,10-methylenetetrahydrofolate reductase (NADPH)	NP_005948	656	80	Yes	Yes	0.92			
Rn21.chr2.29.005	590	Methylcrotonoyl-co-A carboxylase 2 β	NP_071415	563	79	Yes	Yes	0.57			
Rn31.chr5.154.011	548	Selenoprotein N, 1	NP_065184	590	79	Yes	Yes	1.02			
Rn21.chr7.4.031	542	Lamin B2 (LMNB2)	NP_116126	600	78	Yes	Yes	2.17	Lamin A/C	NP_005563	52
Rn21.chr8.26.002	590	Putative β -galactosidase (hypothetical)	NP_612351	636	78	Yes	Yes	0.6	Galactosidase β	NP_000395	38
Rn21.chr4.114.014	254	Deoxyguanosine kinase isoform a precursor	NP_550438	277	78	Yes	Yes	0.48	Thymidine kinase 2, mitochondrial	NP_004605	33
Rn21.chr8.69.013	143	Tropomyosin—human (fragment)	T08796	308	76	No	No	3.24	Tropomyosin 1 α	NP_000357	65
Rn21.chr1.6.59.012	208	Werner syndrome protein (WRN)	NP_000544	1432	74	Yes	Yes	0.56			
Rn31.chr4.182.011	107	Nucleoside phosphorylase (NP)	NP_000261	289	71	No	No	0.62			
Rn21.chr1.4.17.012	338	Methylene tetrahydrofolate dehydrogenase 2 precursor (MTHFD2)	NP_006627	344	70	No	No	2.06	Methylenetetrahydrofolate dehydrogenase	NP_005947	40
Rn31.chr11.32.006	213	Interferon γ receptor 2 (IFNGR2)	NP_005525	337	68	Yes	Yes	0.67			
Rn31.chr3.4.041	311	Glycosyltransferase family 6 like	NP_892019	308	63	Yes	Yes	1.34	ABO blood group (ABO glycosyl transferase)	NP_065202	39
Rn31.chr3.5.015	351	Lysophosphatidic acid acyltransferase	NP_006403	278	59	Yes	Yes	1.63			
Rn31.chr4.29.008	159	Lipoprotein-associated coagulation inhibitor (TFPI)	NP_006278	304	36	No	No	3.41			

We determined whether the prediction is a probable ortholog of the top human hit by considering whether there is conserved synteny around the pair (Synt?) and what the estimated neutral substitution rate is since their nearest common ancestor (dS). The %ID column shows the percentage identity of the highest-scoring HSP returned when NCBI's program Blast2Sequences (bl2seq) is applied to the predicted and RefSeq proteins.

To the best of our knowledge, this is the first reported attempt to amplify complete mammalian open reading frames from de novo gene predictions on a significant scale. Previous efforts to verify many mammalian gene predictions by RT-PCR and sequencing have targeted a single intron (Miyajima et al. 2000; Das et al. 2001; Guigó et al. 2003). In *Caenorhabditis elegans*, on the other hand, an effort to amplify and clone all annotated ORFs was recently reported (Reboul et al. 2003). The greater accuracy of gene prediction in compact genomes such as that of *C. elegans* led to an impressive 25% success rate on predictions that were not supported by any expressed sequence tag (EST) or cDNA evidence.

An unexpected outcome of our experiments was the ease with which pre-mRNAs can be amplified from primers in introns. In particular, primers complementary to introns of *p53* yielded amplification and high-quality sequence that, when aligned to the genome, revealed no splicing. Although pre-mRNA intermediates of *p53* are probably particularly abundant (Khochbin et al. 1992), we were also able to amplify rat *hypoxanthine-guanine phosphoribosyl transferase* (X62085) from primers in introns. This was not the result of genomic contamination, because PCR with the same primers did not amplify anything when the reverse transcriptase was omitted. These results suggest that amplification of an unspliced sequence provides relatively little information about gene structure beyond the fact that the two primers anneal to cDNA from the same primary transcript.

Much more is learned when alignment of the amplicon sequence to the genome reveals the locations of one or more introns. Furthermore, when neither primer anneals to a particular intron, cDNAs in which it is retained are not selectively amplified. Thus, when alignment of the amplicon sequence to the genome reveals an aligned sequence bounded by two introns that have been spliced out, that sequence is almost certainly an exon of the mature mRNA. Using this analysis, the complete set of TE and UTR experiments determined the locations of 598 complete exons (both splice sites found), 743 introns, and 1486 splice boundaries. Because each complete exon is bounded by two inferred introns, the number of introns identified is necessarily greater than the number of complete exons. TWINSKAN correctly predicted 87% of the complete exons, 83% of the introns, and 91% of the splice boundaries.

The exon pairs we targeted were selected for both similarity to a human disease gene from the HGMD database and lack of similarity to an Ensembl prediction on rat. We expected that Ensembl and its protein mapping engine GeneWise would map known human genes to all locations on the rat genome that could produce highly similar proteins. Therefore, we expected that the TWINSKAN predictions that did not match Ensembl's rat predictions would be distant paralogs of known human genes. However, most of the predicted ORFs that were amplified and sequenced turned out to be orthologs of known human genes (see Table 3). Of the predictions that were verified by TE primers, 71% were partially missed by Ensembl whereas 29% were missed completely (no stretch of 100 consecutive nt identical to any Ensembl predicted mRNA). For example, TWINSKAN correctly predicted the structure of the rat *aspartylglucosaminidase* (AGA; Fig. 1), deficiency of which causes the lysosomal storage disease aspartylglucosaminuria. The rat and human proteins align over their entire lengths with 84% amino acid identity. Ensembl only predicts two exons from this nine-exon gene, apparently because GeneWise aligns a fragmentary 40-amino acid rat protein (SWISS-PROT P30919) in preference to the complete human protein. Future versions of Ensembl will likely overcome the blocking of complete xeno-proteins (e.g., human AGA) by fragments of native proteins (e.g., rat AGA), but this example highlights the importance of comparing systems for genome-wide annotation

to one another, not to the structures that could be derived by an expert focusing on a small number of genes.

One key to the scalability of our approach is the automation of the sequence analysis. In principle, aligning the experimental sequence to the genome and checking for introns ought to be relatively straightforward. When we implemented this procedure and checked the automated analyses manually, we discovered a number of special cases that needed to be incorporated in the program. For example, in several of the experiments that our program classified as failures, further inspection revealed that we had discovered an exon within a gap in the early draft assembly of the genome.

Optimizing the protocol for amplifying full ORFs from UTR primers will almost certainly improve the yield. For example, the likelihood that primers anneal to UTR exons can be increased by optimizing the simple primer placement approach used here and, ultimately, by incorporating UTR prediction into algorithms such as TWINSKAN. Sequencing the entire PCR product would eliminate failures due to end sequences that do not reach a splice site. Finally, the success rates for products under 2.5 Kb (55%) was much higher than for those over 2.5 Kb (23%). Thus, amplifying longer ORFs in two overlapping segments would likely give a much higher yield.

Even without optimization, our approach is attractive compared to obtaining new clones by traditional library construction, screening ESTs, and sequencing promising clones to determine whether they contain a complete ORF. The number of ESTs required to identify each new, full ORF clone has recently increased dramatically, indicating that the traditional approach is reaching saturation. On the other hand, the declining cost of primer synthesis will likely make RT-PCR even more attractive in the future. At the same time, the sequencing of more mammalian genomes will improve the accuracy of comparative de novo prediction methods such as TWINSKAN.

We have shown for the first time that high success rates can be obtained in RT-PCR and sequencing of many predicted mammalian ORFs. There is ample opportunity to improve this method, and trends in traditional cDNA library sequencing, primer cost, and gene prediction accuracy will inevitably increase its role in genome annotation. Ultimately, this will lead to a fundamental improvement in annotation—the most reliable annotations will be those that are completely supported by native cDNA sequence obtained from an experiment designed to address a specific hypothesis—the gene prediction. The road forward in annotating open reading frames is clear; the challenges for the future are alternative splices and 5' UTRs.

METHODS

Predictions

Initial gene predictions were made by running TWINSKAN version 1.1 on release 2.1 of the rat genome sequence (January 2003) as repeat-masked by the University of California Santa Cruz (UCSC; <http://genome.ucsc.edu/goldenPath/rnJan2003/chromosomes/>). The informant database was human genome Build 31. All parameters and database preparation were as described by Flicek et al. (2003). A subsequent round of predictions was made by running TWINSKAN version 1.2 on release 3.1 of the rat genome (June 2003) as repeat-masked by UCSC (<http://genome.ucsc.edu/goldenPath/rnJun2003/chromosomes/>) using human genome Build 33 as the informant database. The TWINSKAN predictions for each assembly can be obtained from the corresponding UCSC rat genome Web sites.

Primer Design

All primer pairs used in these experiments were designed using Primer3 (Rozen and Skaletsky 2000) in conjunction with custom

Perl scripts that generated input files for Primer3 and parsed its output.

The goal of the single-intron experiment was to obtain high-quality sequence around the target intron. To achieve this, primers were designed in the exons surrounding the target intron but no less than 30 bp away from it. Primer3 parameters were as follows: primer optimal size 27, primer optimal Tm 70, primer minimum Tm 67, primer maximum Tm 73, single primer maximum mispriming score of 27, primer pair maximum mispriming score of 51.3. All other parameters were left at default values. The mispriming library consisted of all Provisional and Reviewed Rat RefSeqs (NM accessions) available on March 28, 2003.

The aim of the full-ORF experiment was to produce high-quality sequence from both ends of the predictions that covered as much of the gene structure as possible. Predictions with more than 4 Kb of coding sequence were not considered for this experiment, because the high-throughput PCR methods used here have significantly reduced efficiency beyond this length (though we can amplify much longer transcripts using different conditions for different lengths). Two primer pairs were generated for each targeted prediction: one pair in the outermost exons among those with at least 75 coding bases (100 in the second TE experiment), and one pair in the putative UTRs just outside the ORF. The UTR primers were between 10 and 300 bases of genomic sequence from the predicted start and end of translation. Primer3 parameters were as above except for: primer optimal Tm 72.5, primer minimum Tm 65, primer maximum Tm 80, and primer pair maximum mispriming score of 45.9. The mispriming library was as above except that 300 bp of genomic sequence surrounding each RefSeq (analogous to the primer-design region of the targeted predictions) was also included.

Primer sequences for all experiments can be found at <http://genes.cse.wustl.edu/rat-data-03/>.

PCR and Sequencing

For all but the first set of UTR and TE experiments, Poly-A RNA from 10 rat tissues (Sprague-Dawley and Wistar strains) including 10–12-day-old embryo, brain, heart, kidney, liver, lung, ovary, spleen, testicle, and thymus (Ambion) was pooled. For the first set of UTR and TE experiments, rat total RNA (Sprague-Dawley and Wistar strains) from 18 tissues including 10–12-day-old 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) was pooled. First-strand cDNA was generated using Superscript II reverse transcriptase by Oligo-dT priming (Invitrogen). RT was followed by PCR amplification using Clontech Advantage 2 PCR Enzyme System. For single-intron primer pairs, the PCR program was 95°C for 30 sec, followed by 35 cycles of 95°C for 10 sec, 68°C for 30 sec, and concluded by an extension cycle of 72°C for 1 min. For TE and UTR primer pairs, PCR experiments were performed using touchdown PCR (Don et al. 1991) combined with an autosegment extension PCR program, with 35 PCR cycles. The initial annealing temperature was set at 75°C. During the following 5–7 cycles, the annealing temperature was reduced by 2°C each cycle. The rest of the amplification cycles used this annealing temperature. The autosegment extension PCR program uses auto-extend cycles: at the 15th and subsequent cycles, the extension time is increased by 15 sec each cycle. The full-length PCR was concluded by an extension cycle of 72°C for 10 min. 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.

All sequencing traces were submitted to the NCBI trace archive (<http://www.ncbi.nlm.nih.gov/Traces/>) and given trace id numbers 310742461–310743085 (single-intron experiments) and 310742175–310742460 (TE and UTR experiments).

Sequences are also available on the auxiliary data Web site: <http://genes.cse.wustl.edu/rat-data-03/>.

Classification of Sequencing Results

The sequence obtained from each experiment was aligned to the genomic sequence containing the prediction by using EST_GENOME (Mott 1997) version 5.1 (http://www.well.ox.ac.uk/~rmott/est_genome.shtml). Default parameters were used except for: splice_penalty=10 intron_penalty=20 minscore=10. Sequences lacking 20 consecutive called bases were classified as “bad.” The remaining sequences were considered “hits” if their EST_GENOME alignment met all of the following criteria:

1. It had at least 75% identity over the entire alignment,
2. It fell entirely within the genomic region between the primers,
3. It contained at least one intron,
4. All 10-bp sequences flanking introns in the alignment contained at least eight matches, and
5. The experimental sequence did not contain 20 or more contiguous unaligned bases with a quality value above 30.

When EST_GENOME produced a spliced alignment that failed criterion 1 or 4, the alignment and sometimes the trace were inspected manually. If a convincing explanation for the poor alignment could be found, the experiment was classified as positive. Examples of convincing explanations include an experimental exon that falls into a gap in the genome assembly or a miscalled base near a splice site.

ACKNOWLEDGMENTS

We thank the Baylor College of Medicine Human Genome Sequencing Center and the Rat Genome Sequencing Project for the genome sequence, without which this project would not have been possible. Thanks to Jeltje van Baren and James Kent for useful discussions, and to Chris Ponting and Mikhail Velikanov for suggestions about ortholog determination. D.S., M.A., and M.R.B. were partially supported by grants HG02278 from the National Human Genome Research Institute and DBI-0091270 from the National Science Foundation. R.G. and J.Q.W. were partially supported by grants from the NHGRI/NHLBI (1 U54 HG02345) and the 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

- Bairoch, A. and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**: 45–48.
- Birney, E. and Durbin, R. 2000. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* **10**: 547–548.
- Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**: 78–94.
- Das, M., Burge, C.B., Park, E., Colinas, J., and Pelletier, J. 2001. Assessment of the total number of human transcription units. *Genomics* **77**: 71–78.
- Don, R.H., Cox, P.T., Wainwright, B.J., Baker, K., and Mattick, J.S. 1991. “Touchdown” PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Res.* **19**: 4008.
- Flicek, P., Keibler, E., Hu, P., Korf, I., and Brent, M.R. 2003. Leveraging the mouse genome for gene prediction in human: From whole-genome shotgun reads to a global synteny map. *Genome Res.* **13**: 46–54.
- Gasteiger, E., Jung, E., and Bairoch, A. 2001. SWISS-PROT: Connecting biomolecular knowledge via a protein database. *Curr. Issues Mol. Biol.* **3**: 47–55.
- 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.
- Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. 2002. The Ensembl genome database project. *Nucleic Acids Res.* **30**: 38–41.
- Khochbin, S., Brocard, M.P., Grunwald, D., and Lawrence, J.J. 1992. Antisense RNA and p53 regulation in induced murine cell

- differentiation. *Ann. NY Acad. Sci.* **660**: 77–87.
- Korf, I., Flicek, P., Duan, D., and Brent, M.R. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* (Suppl. 1) **17**: S140–S148.
- Miyajima, N., Burge, C.B., and Saito, T. 2000. Computational and experimental analysis identifies many novel human genes. *Biochem. Biophys. Res. Commun.* **272**: 801–807.
- Mott, R. 1997. EST_GENOME: A program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.* **13**: 477–478.
- Parra, G., Agarwal, P., Abril, J.F., Wiehe, T., Fickett, J.W., and Guigo, R. 2003. Comparative gene prediction in human and mouse. *Genome Res.* **13**: 108–117.
- Pruitt, K.D. and Maglott, D.R. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29**: 137–140.
- Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the Brown Norway Rat yields insights into mammalian evolution. *Nature* (in press).
- Reboul, J., Vaglio, P., Rual, J.F., Lamesch, P., Martinez, M., Armstrong, C.M., Li, S., Jacotot, L., Bertin, N., Janky, R., et al. 2003. *C. elegans* ORFeome version 1.1: Experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat. Genet.* **34**: 35–41.
- Rozen, S. and Skaletsky, H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**: 365–386.
- Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S., Abeyasinghe, S., Krawczak, M., and Cooper, D.N. 2003. Human Gene Mutation Database (HGMD(R)): 2003 update. *Hum. Mutat.* **21**: 577–581.

WEB SITE REFERENCES

- <http://genome.ucsc.edu/goldenPath/rnJan2003/chromosomes/>; Source for rat genome build 2.1.
- <http://genome.ucsc.edu/goldenPath/rnJun2003/chromosomes/>; Source for rat genome build 3.1.
- http://www.well.ox.ac.uk/~rmott/est_genome.shtml; EST_GENOME alignment program.
- <http://www.ncbi.nlm.nih.gov/Traces/>; NCBI Trace Archive.
- <http://genes.cse.wustl.edu/>; TWINSKAN Web site.
- <http://genes.cse.wustl.edu/rat-data-03/>; Auxiliary data for this paper, including sequences and accessions.

Received September 11, 2003; accepted in revised form November 17, 2003.