View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

Available online at www.sciencedirect.com

SCIENCE DIRECT

Genomics 87 (2006) 437–445

GENOMICS

www.elsevier.com/locate/ygeno

An efficient and high-throughput approach for experimental validation of novel human gene predictions

Pius M. Brzoska [a], Clark Brown [a], Michael Cassel [a], Toni Ceccardi [a], Valentina Di Francisco [b], Alex Dubman [a], Jason Evans [a], Rixun Fang [a], Michael Harris [b], Jeffrey Hoover [b], Fangqi Hu [a], Charles Larry [a], Peter Li [b], Michael Malicdem [a], Sergei Maltchenko [a], Mark Shannon [a], Sarah Perkins [a], Karen Poulter [a], Marion Webster-Laig [a], Chunlin Xiao [b], Sonny Young [a], Gene Spier [a], Karl Guegler [a], Dennis Gilbert [a], Raymond R. Samaha [a,*]

[a] *Applied Biosystems, 850 Lincoln Center Drive, Foster City, CA 94404, USA*
[b] *Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA*

Received 2 August 2005; accepted 24 November 2005
Available online 9 January 2006

Abstract

A highly automated RT-PCR-based approach has been established to validate novel human gene predictions with no prior experimental evidence of mRNA splicing (ab initio predictions). Ab initio gene predictions were selected for high-throughput validation using predicted protein classification, sequence similarity to other genomes, colocalization with an MPSS tag, or microarray expression. Initial microarray prioritization followed by RT-PCR validation was the most efficient combination, resulting in approximately 35% of the ab initio predictions being validated by RT-PCR. Of the 7252 novel genes that were prioritized and processed, 796 constituted real transcripts. In addition, high-throughput RACE successfully extended the 5′ and/or 3′ ends of >60% of RT-PCR-validated genes. Reevaluation of these transcripts produced 574 novel transcripts using RefSeq as a reference. RT-PCR sequencing in combination with RACE on ab initio gene predictions could be used to define the transcriptome across all species.
© 2005 Elsevier Inc. All rights reserved.

Keywords: RT-PCR; Novel human genes

Although the act of physically sequencing of the human genome has been deemed complete for some time [1–3], the task of accurately and completely identifying all human genes still remains. The large scale of this endeavor has led researchers to rely largely on computer algorithms to predict sequences that constitute functional genes [4–6]; however, computer predictions, particularly when used without the benefit of experimental data, can often mispredict the 5′ and 3′ ends of genes and miss a lot of exons.

The problems with gene prediction algorithms are well known [7–10]. They use a variety of methods to predict the likelihood that a sequence constitutes a gene, for example by identifying homology to existing known genes, identifying consensus sequences (e.g., promoter elements, start and stop codons, splice sites, and poly(A) sites) [11], and using statistical analysis of oligonucleotide frequency, exon and intron length, and intron and exon number [8]. Algorithms that use sequence features to predict gene structure typically predict fewer than 50% of known exons and 20% of complete genes [6]. Recently, however, new algorithms that use sequence conservation between distantly related genomes have been developed. These dual genome predictors have had a higher rate of success in correctly predicting exons and known genes than single genome predictors [12,13], though to be effective, these approaches require the sequence of related organisms, which need to be separated by an appropriate evolutionary gap.

Other approaches have relied on homology to known expressed sequences (typically expressed sequence tags (ESTs)) to identify and annotate genes [14]. ESTs have been used successfully to identify genes, especially with recent advances in clustering algorithms [15–17]. Although often

* Corresponding author. Fax: +1 650 638 6893.
E-mail address: samaharr@appliedbiosystems.com (R.R. Samaha).

0888-7543/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2005.11.016

incomplete, predictions of transcripts based on expressed sequences are more accurate at predicting exons of known genes. Nevertheless, these approaches are obviously dependent on the existence of expression data and further tend to miss many predicted exons and genes expressed at low levels. They are also prone to contamination from genomic sequences and unspliced mRNAs [3,18].

Additional experimentally based approaches involve fully sequencing cDNAs and then using clustering algorithms to reconstitute individual genes. For example, Imanishi et al. sequenced over 41,000 human cDNAs that they subsequently determined represented over 21,000 genes, including over 5000 newly identified gene candidates [19]. Despite the success of this method, the cDNA approach is susceptible to the low processivity of reverse transcriptase and can result in incomplete gene sequences, particularly for long genes. Therefore, additional experimental approaches are needed to discover the true ends of genes and provide a complete picture of gene structure. To increase accuracy over automated approaches, some groups have employed expert manual curators [20]. Although this approach can be very successful, it is also dependent on availability of data and is relatively time consuming. Therefore, efficient and automated experimental approaches are critical to rapidly validate novel genes that have not been predicted by bioinformatics approaches.

Other researchers have also come to the conclusion that experimental validation of novel genes is necessary; some have taken wide, microarray-based approaches to experimentally characterize genes that have been predicted ab initio. For example, “exon” and “tiling” arrays have been used to validate and refine computational gene predictions and define full-length transcripts on the basis of coregulated expression of their exons. This first study was carried out in detail for chromosome 22q and proved reliable for the detection of multiple splice variants and for the detection of genes that are expressed under tissue- or disease-specific conditions [21]. Later studies have focused on chromosomes 21 and 22 using 25-mer oligonucleotide arrays [22] as well as chromosomes 20 and 22 using 60-mer oligonucleotide arrays [23]. These tiling approaches are typically independent of computational predictions and have been recently extended to cover a much larger portion of the human genome [23,24]. Other microarray-based approaches have used polymerase chain reaction (PCR)-amplified open reading frames to validate gene predictions in 350 kb of large human genomic clones [25]. These approaches have been quite successful in generating a wealth of expression data that not only have validated known genes but also, because of their independence from genome annotation, have identified significant transcriptional activity outside of known gene boundaries [23,24]. Such approaches, however, rely upon measuring the intensity (or differences in intensity between samples) of fluorescent probes hybridized to the array, a method with lower sensitivity, which is also prone to a high degree of false positives that are most often caused by cross-hybridization or other experimental artifacts. In addition, to generate accurate details of gene structure, a high tiling density is required, which renders these experiments prohibitively expensive. On the other hand, reverse transcription (RT)-PCR coupled with direct sequencing of generated amplicons has been used to validate predictions generated using dual genome algorithms for both human [26] and rat [27], Genscan-predicted novel genes on human chromosome 22 [10], and predicted open reading frames in the *Caenorhabditis elegans* genome with a great measure of success [28].

Our goal was to validate experimentally ab initio predictions with low certainty scores (i.e., with little or no evidence beyond the actual computational gene prediction) to gain a better view of the completeness of the human transcriptome. Because we were ultimately interested in designing TaqMan (PCR-based) gene expression assays and microarray probes for these novel transcripts, determining the correct gene structure was critical to our endeavor. Using an approach that relies solely on microarrays was deemed incomplete and we decided instead to implement a high-throughput RT-PCR approach. Nevertheless, our previous experience validating a small number of ab initio predictions with low certainty scores using RT-PCR followed by amplicon sequencing (R. Samaha, unpublished data) indicated that scaling up this process would be prohibitively time consuming and costly, and the final percentage of validated ab initio predictions would be relatively low. Therefore, we decided to test a variety of bioinformatic and experimental approaches to prioritize the ab initio predictions that would be processed through our RT-PCR pipeline and increase our final validation rate.

Ab initio predictions were prioritized on the basis of four criteria: predicted protein classification using our proprietary Panther pipeline [29], sequence similarity compared to dog or mouse genomes, colocalization with a massively parallel signature sequencing (MPSS) tag [30], or detection of expression on microarrays (see Materials and methods). Prioritized predictions were then validated using RT-PCR on a series of mRNAs from 30 tissues followed by sequencing of the resulting amplicons, a process that we refer to as the RT-PCR pipeline. Additionally, we set up a high-throughput rapid amplification of cDNA ends (RACE) pipeline to determine additional 3′ and 5′ sequence of the predictions that had been validated by our RT-PCR pipeline. In addition to increasing our validation yields, the combination of these methods provided information about gene structure, splice variants, and gene boundaries that would not necessarily be obtainable using an array-only approach. In this paper, we report on the novel genes that were identified genome-wide from these high-throughput experiments, which add to the annotated knowledge of the transcriptome.

Results

Ultimately, after filtering, 7252 of the 98,545 ab initio predictions were processed through the RT-PCR/sequencing pipeline on 10 different tissue pools (Supplemental Data File 1). Of those, 796 (12%) ab initio predictions were validated and thus represent truly expressed, previously unidentified genes. Prioritizing by expression on microarrays resulted in the highest validation rate by RT-PCR (33%), whereas the RT-PCR

validation rates of the predictions prioritized by similarity to dog and mouse varied from 15 to 3.8% (Table 1) and, as expected, were correlated with the degree of similarity. Validation rates for predictions prioritized by Panther classification or colocalization with an MPSS tag were 21.3 and 18.5%, respectively (Table 1).

Although the in silico approaches required no wet lab work and were therefore very efficient, they yielded lower validation rates and had the disadvantage that only a small number of gene predictions passed the filtering criteria. For example, of the 21,000 predictions processed, 1684 had a Panther classification and only 313 were colocalized with an MPSS tag, severely limiting the applicability of these approaches. On the other hand, after optimizing the threshold by which we called a gene present, using TaqMan data as controls, 33% of the predictions profiled on microarrays ultimately were prioritized for RT-PCR validation. The microarray prioritization approach was made even more efficient by limiting profiling to 15 tissues.

An additional benefit of using microarrays as a prioritization step (as opposed to the other bioinformatics-based filtering methods) is their ability to reveal information about the genes' tissue distribution. The expression levels were used to determine which gene predictions were exclusively expressed in the tissues used. For example, there were more than 60 genes expressed exclusively in liver, whereas very few genes were expressed exclusively in fetal heart, fetal kidney, kidney, or lung (Supplemental Data File 2).

Using RT-PCR as the validation method also enabled us to determine the intron/exon structure of the transcripts. For the 796 ab initio gene predictions that were validated, we discovered 238 exons that were not previously identified by the prediction software programs. The resulting transcripts were compared to RefSeq, GenBank mRNA, and Ensembl cDNA. Depending on the reference set, between 505 and 574 either were entirely novel or provided novel exonic information to an otherwise known transcript (163 and 296 were novel, the remaining showed additional exonic information).

Moreover, the majority of ab initio gene predictions were small (~500 bp) compared to human curated RefSeqs (NM_ sequences) and Celera genes (hCTs) and had only one splice site, suggesting that the gene ends had not been correctly predicted (Supplemental Data File 3, unpublished data). Therefore, most transcripts that were validated by RT-PCR and sequencing were subjected to high-throughput extension by 5′ and 3′ RACE using the tissue pools we had found them to be expressed in. This enabled us to extend the ends of the transcripts and also identify new exons. Of the 687 genes subjected to RACE and sequence analysis, 385 (56%) gave a product. New 3′ ends or new 5′ ends were identified for 184 of the 385 transcripts, and 191 new exons were discovered. Overall, approximately 300,000 bp of new sequence was generated.

Table 1
RT-PCR validation rates by prioritization category

Prioritization method	RT-PCR validation rate (%)
Dog/mouse similarity >50%	15.0
Dog/mouse similarity 30–50%	17.0
Dog/mouse similarity 15–29%	3.8
Dog/mouse similarity <15%	5.2
Panther	21.3
MPSS	18.5
Arrays	32.0

Ab initio gene predictions that correlated to expressed genes on microarrays gave the highest validation rate by RT-PCR. Predictions prioritized by MPSS tag analysis and Panther software analysis were validated at intermediate levels. When dog/mouse similarity was used, predictions were validated at rates that correlated with the degree of similarity, as expected.

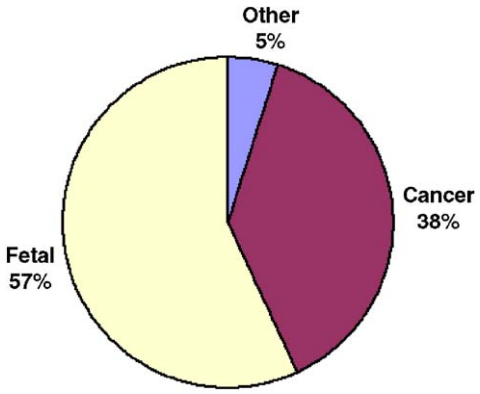

Fig. 1. Distribution of alternative splicing events among three tissue pools: cancer, fetal, and other.

We were also interested in the distribution of splice variants of the validated gene predictions in the different tissue pools that were used. These pools (Supplemental Data File 1) were combined depending on their tissue origin into three different categories: cancer template, fetal tissue template, and a normal tissue template. Splice variants specific to each category were determined by performing pair-wise comparisons of the variants between each category. As expected [31], cancer pools produced a significant number of alternative splicing events (38%). However, the largest number of alternative splicing events was observed for the fetal tissue category (57%), whereas the remaining 5% were observed in the normal tissue pools (Fig. 1). Overall, alternative splicing was identified for 253 (30%) of the 863 validated predictions.

Newly validated genes, including those that were extended by RACE, were also reclassified using the Panther protein classification software. Of the new genes, 261 had high homology to genes from different functional classes, a subset of which is listed in Table 2 (and Supplemental Data Files 4, 5, and 6), indicating that these newly discovered genes might play a significant role in important biological processes. For example, we discovered 6 novel kinases, 32 novel transcription factors, and 32 novel receptors.

Since this work was completed several months ago, we recently compared the validated transcripts against the current Ensembl v29 and NCBI RefSeq March dataset (release 10). Of 796 transcripts, 574 were still novel. We required 97% identity and 90% overlap of two sequences in a BLASTn alignment to be called identical. We found 156 sequences in Ensembl and

Table 2
Summary of validated predictions with Panther protein classifications

Panther software class	Number of genes
Receptors	32
Transcription factors	22
Defense/immunity	16
Kinases	6
Proteases	12
Chemokines/cytokines	8
Phosphatases	7
Cancer/oncogenesis	7
Cell division/cell cycle	7
Chromatin rearrangement	4
Transporters	4
Chaperones	3
Ubiquitin/proteasomes	3
Translation	3
Apoptosis	1
Splicing	1

After the ends of validated gene predictions were extended, many transcripts could be assigned new classifications by the Panther software.

189 in the NCBI RefSeq set, further demonstrating the validity of our approach.

Supplemental Data Files 4, 5, and 6 illustrate three examples of novel genes that were discovered by our approach. Using Celera's Genome Browser we superimposed the original ab initio gene prediction over the RT-PCR-generated contigs as well as the RACE-extended contigs. One novel transcript had significant homology, as determined by Panther classification, to the *Schlafen* gene family, which is involved in T cell development [32]. This prediction originally consisted of two exons that were validated by RT-PCR; however, the splice junction had been incorrectly predicted and a third new exon was identified by the RT-PCR approach. Furthermore, using RACE, we were able to extend this gene by three exons at the 3′ end.

Discussion

Predicted genes within the Celera genome have been categorized based on the level of evidence that supports their existence. Gene predictions using the OTTO algorithm [1] are sequences with considerable supporting evidence, whereas the genes predicted with the Promote software [1] have less supporting evidence. These genes are classified into one of four possible promote levels that indicate the pieces of evidence used in the prediction. For example, a promote "1" has one piece of evidence supporting it, and a promote "2" has two pieces of evidence, and so on. Evidence can be an EST, or mouse similarity, or similarity to a known protein [1]. Earlier validation work (R. Samaha et al., unpublished) focusing on promotes, using RT-PCR and sequencing, indicated that they were well annotated, with validation rates ranging from 70 to 75%. This result led us to the conclusion that, rather than validating promotes that already have supporting evidence of their existence, the greatest advancement of knowledge of the transcriptome would come from focusing on ab initio gene predictions that do not qualify as promotes and therefore currently do not have enough supporting evidence to be included in the Celera database.

Our objective was to improve the current understanding of the completeness of the human transcriptome by establishing an automated approach to identify and improve the annotation of novel transcripts. Though we were able to deduct gene structure without the use of human annotators, the data generated can also be ultimately used for further manual annotation, as performed for RefSeqs. The approach chosen should be easily scaled up to validate all available ab initio predictions. Moreover, all identified novel transcripts have been deposited in Celera's CDS database and used to design TaqMan Gene Expression Assays and were included in our Human Genome Survey Microarray. All sequences have been deposited in public databases.

RT-PCR coupled with sequencing of the generated amplicons was our method of choice for validation because of our need to determine gene structure clearly. However, it was obvious that using this approach solely would be extremely costly and time consuming and could result in lower validation yields. To address the latter problem we decided to run our RT-PCR on pools of three different tissue mRNAs, thereby reducing the number of reactions that needed to be run. More importantly, we decided to test several prioritization approaches that would filter the predictions to be processed through the RT-PCR pipeline and increase their validation yields.

Four different approaches were used to predict which sequences would have the highest validation rate by RT-PCR: expression on microarrays, presence of a Panther protein classification, similarity to dog or mouse genomic fragments, or colocalization with an MPSS tag. Of the four methods, we determined that microarray gene expression profiling on 15 single tissues provided the most accurate method for filtering out ab initio gene predictions (providing a 32% validation rate) and that the combined platform of microarray filtering with RT-PCR and amplicon sequencing as the final validation tool is ideal because it provides the most sequence information while enabling cost-effective gene processing.

The 32% yield obtained is affected by several factors, notably cross-hybridization on the arrays; several of the predictions were quite short, preventing the design of highly specific probes. Another factor that can affect the validation rate is RT-PCR false positives caused by the highly automated nature of the process. Recently, Wu et al. [27], validating rat predictions with RT-PCR and direct sequencing, were able to increase their validation yields from 44 to 59% by repeating the RT-PCR using the same primers but slightly modifying the PCR conditions.

Of the 7252 sequences selected for prioritization that were processed by RT-PCR combined with sequencing, we identified 847 novel transcripts—an overall validation rate of 12% for ab initio gene predictions with little supporting evidence for their existence.

Our final validation rate of 12% is lower than that found in other recent RT-PCR and sequencing studies, such as a study validating 230 of the novel Genscan-predicted genes on

chromosome 22 that yielded a 27% validation rate [10] and other recent studies validating predictions generated using the recently derived dual genome-prediction program TWINSCAN [13] in humans [26] and rat [27], which yielded rates of 62 and 44–59%, respectively.

These differences can be largely explained by the prediction algorithms used to identify the genes. The new dual-genome predictors have been shown to outperform single genome predictors [33]; however, they do require sequences of related organisms separated by appropriate evolutionary distances to be fully effective. Additionally, these studies did not necessarily focus on predictions supported by weak experimental evidence or by no experimental evidence at all. For example, the 62% validation rate in Guigo et al. [26] drops to 20–25% when predictions that do not overlap with Ensembl predictions were validated. It is worth noting, however, that in the rat study the validation rate was indeed for predictions that did not overlap with the Ensembl set [27], but were homologous to human HGMD genes.

For the Das et al. study [10], the work did not focus solely on spliced transcripts but also took into account unspliced transcripts that may or may not be real. When only spliced transcripts in this study are counted, the validation rate drops from 27 to 13.5%—a rate that is very similar to our unfiltered rate. When we added filtering mechanisms to our pipeline, our validation rates went up. For example, ab initio gene predictions that had more than 50% similarity to dog and/or mouse sequences subsequently validated at a rate of 15%, and ab initio gene predictions that gave a positive signal on the microarray subsequently validated at a rate of 32%.

Given that there are more than 60,000 or so (after filtering overlapping predictions) ab initio gene predictions that remain to be processed, we predict that by employing a pipeline similar to the one described here—using microarrays on 15 single tissues to filter gene predictions before proceeding to RT-PCR validation on tissue pools and sequencing—about 4000 to 5000 new human genes remain to be discovered. This number, however, may be slightly overestimated, since only 21,000 (including the 7252 validated in this paper) of the remaining 60,000 have weak evidence supporting their existence; the remaining ones are either single-exon predictions or multiexon predictions with no supporting evidence (Supplemental Data File 7), which will likely exhibit lower validation rates.

Our RT-PCR and sequencing pipeline throughput averaged 50 96-well plates per week and could be easily increased to 200 per week by switching from a 96-well to a 384-well format, meaning that the remaining 60,000 could theoretically be processed in 3 to 4 months. Furthermore, we determined that 4 of the 10 tissue pools (pools 1, 2, 3, and 6 in Supplemental Data File 1) contained more than 90% of the genes that were successfully validated; therefore, reagent costs could be reduced and throughput could be maximized by focusing exclusively on these four tissue pools.

Mapping the transcriptome has been a challenging task underscored by the widely divergent predictions of the number of genes in the human genome. Although the number seems to have settled in the range of 20,000 to 25,000 [3], recent studies using tiling microarrays have indicated the existence of significant transcription outside of known genes [23,24]. Though a lot of these identified new genes could potentially be noncoding, a subset will surely be protein-coding genes that have not been predicted by current computational approaches. Tiling arrays, by their parallel nature, are uniquely suited to these types of genome-wide analysis; nevertheless, they do not obviate the need for RT-PCR and/or RACE for further validation and characterization of gene structure. This was demonstrated by Cheng et al. [24], who used both approaches in their 10-chromosome-wide study. Similarly in our case, as opposed to relying exclusively on computer predictions, or microarray profiling, our combined approach allowed us to gain considerable insights into the novel genes' structure; for example 238 new exons were identified and nearly 60% of all validated predictions were extended by our high-throughput RACE pipeline, indicating the ends had been mispredicted by the software-based prediction methods. Additionally, because our microarray profiling was done on individual tissues we were able to determine the novel transcripts' tissue distribution (Supplemental Data File 2).

The distribution of the alternatively spliced transcripts was also interesting. Most of the variants were identified in the cancer and fetal pools (38 and 57%, respectively). The numbers for the cancer pools make intuitive sense since alternative splicing has been associated with cancer tissues [31]; however, we were surprised by the levels identified in the fetal pools. In retrospect though, since fetal tissues are more than likely undergoing significant growth and developmental changes, they would likely require novel gene functions provided by alternative splicing events. This latter explanation is further strengthened by the significant difference in the distribution of alternative splicing events between normal and fetal tissues (5% vs 58%, respectively). We believe that our genome-wide results strongly suggest that alternative splicing is an important element in the early development process.

Overall, splice variants were identified for 30% of the validated genes. This result is slightly below the current estimates that 40–60% of human genes are alternatively spliced [34,35]; however, the discrepancy could be due to the confounding effect of using tissue pools in our RT-PCR pipeline.

In our experience, and based on the previous work cited here, we believe that experimentally filtering de novo predictions generated by single or dual genome predictors (or a combination of both) followed by systematic validation using an automated RT-PCR/amplicon sequencing pipeline will be a cost-effective approach to characterizing the transcriptome further. In addition to providing a better picture of the status of the human transcriptome, identifying and annotating a complete and accurate set of human genes have important implications for large-scale genetic research that begins with known human genes, for example gene expression surveys using microarrays and gene mapping experiments within genomic regions linked to disease through linkage mapping

studies. In both cases, even if these unidentified genes have a significant effect on the biological process being studied, they will be missed because they are simply not represented on the array or in the annotation of the genomic region being studied. The addition of these genes (and their respective homologies, gene structure, and splice forms) to the annotation of the human genome is an important step in ensuring that future biological research is complete.

Materials and methods

Identifying gene predictions and promotes

Approximately 98,454 ab initio gene predictions from three prediction programs—Genscan [36], GRAIL2 [37], and FgenesH [38]—did not have exon overlap with existing curated Celera transcripts (i.e., they had insufficient evidence to be instantiated as genes by our manual curators). From these, 39,951 predictions with very weak similarity to existing protein or EST sequences were initially selected for further validation. In many cases, the software programs produced overlapping or redundant gene predictions. To avoid redundant validation, gene predictions with exon overlap were clustered and a single cluster representative was chosen; pseudogenes were also filtered out at this stage. Cluster representatives were selected primarily to maximize the number of exons and secondarily to maximize the length of the gene prediction. After clustering, approximately 31,000 gene predictions emerged, made up of 10,000 single-exon gene predictions and 21,000 multiexon gene predictions. The 21,000 multiexon gene predictions were then considered to be candidates for RT-PCR validation (Supplemental Data File 7).

Prioritizing predictions

The complete process for prioritizing and processing ab initio gene predictions is diagrammed in Fig. 2.

Similarity to dog and mouse genomic fragments. Gene predictions were first ranked based on their percentage of similarity (over 50% of their length) at the DNA level to dog and mouse genomic fragments. Predictions were binned by percentage similarity: >50%, 30 to 50%, 15 to 29%, and less than 15% similarity. Of the predictions (mostly from the higher similarity bins) 7252 were chosen for RT-PCR validation. This method was based on the hypothesis that exons from ab initio predictions with significant similarity to dog and mouse sequences have a higher probability of being real exons.

Panther protein classification. All 21,000 ab initio predictions were also processed through Celera's Panther protein classification software pipeline [29]. The software pipeline was able to assign a classification to 1684 of the predictions, and 1127 of those predictions were processed through the RT-PCR validation pipeline.

Colocalization with MPSS tags. Three hundred thirteen ab initio predictions of the 7252 nonredundant predictions also colocalized with a 17-mer MPSS tag [30] from the 88,782 present in Celera's database. MPSS tags were mapped to the Celera human genome, and only perfectly mapped tags were used for this analysis. An MPSS tag was considered "colocalized" with a predicted gene if the tag overlapped with the exon region of the predicted gene. The RT-PCR validation rates of these predictions were analyzed independently.

Detection of expression by microarrays. Additionally, the expression patterns of 557 randomly chosen predictions (of the total 7252 predictions) were analyzed using microarrays. Three to six 60-mer probes for each ab initio gene prediction were designed using a custom bioinformatics pipeline developed at Applied Biosystems (Foster City, CA, USA). These probes were custom synthesized onto Agilent 60-mer Custom in Situ Oligonucleotide Microarrays (8.4K (G2508A) and 22K (G2509A); Agilent, Palo Alto, CA, USA). mRNA expression values were generated from the arrays using dual-color experiments in which Stratagene Universal human reference RNA (Stratagene, La Jolla, CA, USA) was labeled with Cy3, and mRNA from 15 individual tissues (BD Biosciences, Palo Alto, CA, USA) was labeled with Cy5. RNA integrity was confirmed with the Agilent bioanalyzer or 12.5% Mops denaturing agarose gel electrophoresis. A two-step amino allyl–cDNA labeling strategy [39] was used to label the RNA samples with the following modifications: briefly, 1 μg mRNA and 20 μg of Universal total RNA was used for each array. Bacterial RNAs (*bioB, bioC, bioD, Cre, Dap*, and *Phe*) were spiked into the RNA to monitor the cDNA synthesis efficiency as well as the hybridization success. cDNA synthesis was primed with a mixture of oligo(dT) primers and random nanomers. Cy3- and Cy5-labeled cDNAs

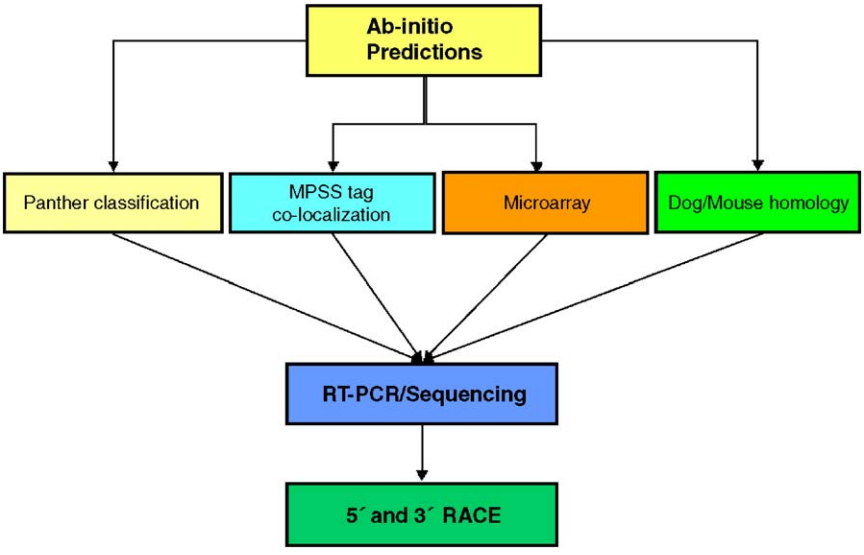

Fig. 2. Bioinformatic and experimental approaches for prioritizing and validating ab initio gene predictions. Ab initio gene predictions were prioritized for progression to RT-PCR and sequencing using four different methods: Panther classification, MPSS tag colocalization, microarray analysis, and level of similarity to dog and/or mouse genes. Some ab initio gene predictions were subjected to more than one method of prioritization. Most predictions that passed the prioritization step and went on to be validated by RT-PCR and sequencing were also subjected to 5′ and 3′ RACE analysis.

were purified using a GFX PCR purification kit (Amersham Biosciences, Piscataway, NJ, USA). The incorporation of Cy3 and Cy5 was monitored using a scanning UV spectrophotometer (200–600 nm). Hybridization, washing, array scanning, and data extraction systems were performed according to the Agilent Technologies user's guide. A minimum of three chip replicates were used for each tissue.

A gene was called present on the array if 69% of all probes representing that gene on all replicates gave expression levels above a determined threshold. This 69% cutoff was derived by comparing the gene expression levels of a subset of positive and negative control genes as determined by quantitative real-time PCR (TaqMan Gene Expression Assays; Applied Biosystems) and microarray experiments. These results were used to identify the average expression level at which array results were likely to correlate with TaqMan results. The threshold was calculated using two approaches. The first was a background-oriented algorithm that calculates a noise threshold level based on mean and standard deviation values that do not change at a certain low noise level range [40]. The second approach used the signal generated by random 60-mer negative controls to help determine the threshold above which we would call a gene absent or present. The threshold was determined to be the mean signal generated by the random oligos plus 2 standard deviations. These oligos were designed to have no homology to human sequences (as determined by BLAST) and this fact was subsequently validated by profiling their expression pattern on microarrays using mRNA for 17 different tissue sources. Any outliers were removed from the set of controls (data unpublished). Both approaches gave similar thresholds, and the method based on the negative controls was predominantly used to generate the results presented in this paper. Additionally, the ratio of the Cy5 to Cy3 channel was used to determine tissue distribution.

RT-PCR

cDNA preparation. Pooled total RNA was made from equal amounts of each human tissue listed in Supplemental Data File 1. In addition, each pool was spiked with RNA from *bioC* (100 pM *Escherichia coli* (Migula)), *Cre* (10 pM bacteriophage P1 *Cre* gene for recombinase protein), and *pheB* (1 pM *Bacillus subtilis* phenylalanine biosynthesis-associated protein) as positive controls. A reaction with no template was used as the negative control. cDNAs were constructed from these pools (100 μg) using the High Capacity cDNA Archive Kit (Applied Biosystems).

SYBR green PCR. Primers (2.5 μl of 2.5 μM) were mixed with 0.15 μl of the cDNA preparation and 6.25 μl of SYBR Green PCR Master Mix (Applied Biosystems) in a total volume of 12.5 μl. Positive control primers for *bioC, Cre*, and *pheB* were included. A reaction without template was used as the negative control.

Forward and reverse 10 μM gene-specific primers (Biosource International, Camarillo, CA, USA) were combined with the cDNA pools and reactions were carried out in ABI Prism 96-well optical reaction plates (Applied Biosystems) that were sealed with optical adhesive covers (Applied Biosystems) and thermal cycled on the GeneAmp PCR System 9700 (Applied Biosystems) using the following cycling parameters: 10 min at 95°C; 40 cycles of 95°C for 30 s, 64°C for 45 s, 68°C for 90 s; and 10 min at 68°C.

Fluorescence end-point reads of SYBR/ROX were determined using an ABI Prism 7900HT sequence detection system (Applied Biosystems). In addition, one plate of every five was loaded on a 4% gel as a control, and data were documented using a Gel Doc 2000 gel documentation system (Bio-Rad Laboratories, Hercules, CA, USA).

RT-PCR was conducted on 10 different pools of cDNAs derived from multiple tissues (Supplemental Data File 1).

PCR cleanup. Single-stranded DNA was degraded using 2.5% exonuclease I (USB, Cleveland, OH, USA) and dephosphorylated using 25% shrimp alkaline phosphatase (USB) in 72.5% $T_{10}E_{0.1}$ buffer (Teknova, Half Moon Bay, CA, USA) The SYBR green reaction plate was heat sealed with Eazy Peel AB-3739 (Abgene, Rochester, NY, USA). After a quick spin down, the plate was placed into a 96-well GeneAmp PCR System 9700 (Applied Biosystems) and run for 60 min at 37°C, and then 15 min at 72°C.

Sequencing

Reactions from the cleaned PCR plates were sequenced using the original forward and reverse primers. Reactions were carried out using 5 μl of a mix containing 60% BigDye Mix from the BigDye Terminator Cycle Sequencing Ready Reaction Kit, version 1.0 (Applied Biosystems), 30% PCR H_2O, and 10% 5× sequencing buffer (Applied Biosystems), combined with 4 μl of PCR product and 1 μl of 3.3 μM primer. The 96-well PCR plates containing these reactions were sealed, spun down, and placed on the 96-Well GeneAmp PCR System 9700 for 1 min at 96°C and then 50 cycles of 96°C for 1 min, 52°C for 10 s, 60°C for 4 min.

Sequencing reactions were cleaned using the Performa DTR Gel Filtration Systems 96 (Edge BioSystems, Gaithersburg, MD, USA) according to manufacturer's specifications. The collection plates were dried down using the SpeedVac concentrator system (Savant Instruments, Holbrook, NY, USA), rehydrated in Hi-Di Formamide (Applied Biosystems), and run on the ABI Prism 3700 DNA analyzer (Applied Biosystems) using POP-6 performance-optimized polymer (Applied Biosystems).

Primer design, data storage, and validation criteria

RT-PCR primers were designed using Primer3 software [41]. RT-PCR primers were tiled across the entire predicted transcript with a target amplicon size of 300–500 bp and 50 to 100 bp overlap of the amplicons. Ten primer pairs were generated for each tile and compared against all human Celera and public transcripts using the BLAST program. On average, 90% of the predicted transcripts were covered by primers. The primer pairs with the lowest predicted cross-hybridization were picked. Cross-hybridization was estimated based on the sequence homology of the primers to the transcripts, and mismatches toward the 3′ end of a primer were considered more detrimental (i.e., weighted higher) than mismatches in the middle or at the 5′ end. Sequence reads were generated using the RT-PCR primers as sequencing primers. Bases were called using Phred software [42,43]. Sequences were curated using in-house software and low-quality nucleotides were masked. The resulting sequence reads were assembled using Phrap assembly software [44].

The assembled sequence reads were then aligned to genomic sequences using Celera's proprietary in-house EstMapper software to determine their gene structure. To be determined "validated" a transcript was required to have at least one exon/intron boundary when the corresponding transcript was aligned to the genome. The one exon/intron boundary limit was chosen because so many of the ab initio gene predictions were very small (<500 bp) and likely contained only one exon.

Following RT-PCR and direct sequencing, assembled sequences with corresponding genomic alignments were stored in an ORACLE database (HercDB). This database was designed and developed to serve as a hub for the data generated by the various assay design and data capture pipelines. HercDB stored and tracked all information generated by the project, including transcript sequences, primer designs, amplicon/contig sequences, and validated prediction sequences.

RACE

A subset of gene predictions that had been validated by RT-PCR and sequencing were subjected to RACE [45] analysis to obtain sequences beyond their predicted 5′ and 3′ ends. RACE was performed using Marathon cDNA and Advantage 2 polymerase mix (BD Biosciences) in accordance with the manufacturer's instructions except that all reactions were scaled down to 10 μl. Marathon cDNAs were pooled to match the compositions of templates used for RT-PCR as described above. RACE was first carried out using a transcript-specific primer (P1) in combination with an adaptor primer (AP1) and at least one template pool from which the transcript was validated by RT-PCR. Primary RACE product was diluted 1:15 for secondary RACE, which used a nested transcript-specific primer (N1) in combination with the other adaptor primer (AP2). The melting temperature of each primer typically ranged from 69 to 71°C. Primary RACE reactions utilized the following method: 3 min at 94°C; 5 cycles of 5 s at 94°C, 4 min at 72°C; 5 cycles of 5 s at 94°C, 10 s at 70°C, 4 min at 72°C; 25 cycles of 5 s at 94°C, 10 s at 68°C, 4 min at 72°C; 7 min at 72°C. Secondary RACE reactions utilized the following method: 3 min at 94°C; 25 cycles of 5 s at 94°C, 10 s at 68°C, 4 min at 72°C; 7 min at 72°C. Secondary RACE products were treated with exonuclease I and SAP as described above

and sequenced directly with N1 primer and a nested sequencing primer (S1) as described above.

Alternative splicing bioinformatics pipeline

The algorithm described below was developed to use the data generated from the RT-PCR and direct sequencing experiments and perform an automatic search for alternative splicing events; 847 validated gene predictions were analyzed by this method. Several rounds of manual curation of results were performed and pipeline parameters were adjusted accordingly:

1. Collect contiguous assembled sequences (contigs) per tissue pool per validated gene prediction.
2. Use above contigs to mine a database for corresponding genomic alignments showing a "split" of the contig in a following manner: e1(p1, p2), e2(p3, p4), e3(p5, p6), where (a) e1, e2, and e3 are three exons derived from a contig; (b) pX are the positions of individual exons on a genome and $p1 < p2 < p3 < p4 < p5 < p6$; and (c) $p3 - p2 > 15$ and $p5 - p4 > 15$ ("introns").
3. Alternative splicing events are defined as follows: (a) "Missing/extra exon": if tA = e1a + e2a + e3a and tB = e1b + e2b, then e1a(p1,p2) = e1b(p1,p2) and e3a(p5,p6) = e2b(p3,p4), where tA and tB are two alternative forms arising from a validated gene. (b) "Alternative 5′-exon position": if tA = e1a + e2a and tB = e1b + e2b, then $[e2a(p3) - e1a(p3)] > 10$, where tA and tB are defined as in 3(a) above. (c) "Alternative 3′-exon position": if tA = e1a + e2a and tB = e1b + e2b then $[e1a(p3) - e2a(p3)] > 10$, where tA and tB are defined as in 3(a) above.

Acknowledgment

We thank Mignon Fogarty for assistance with the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ygeno.2005.11.016.

References

[1] J.C. Venter, et al., The sequence of the human genome, Science 291 (2001) 1304–1351.
[2] E.S. Lander, et al., Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921.
[3] International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome, Nature 431 (2004) 931–945.
[4] M.R. Brent, R. Guigo, Recent advances in gene structure prediction, Curr. Opin. Struct. Biol. 14 (2004) 264–272.
[5] M. Snyder, M. Gerstein, Defining genes in the genome era, Science 300 (2003) 258–260.
[6] M.Q. Zhang, Computational prediction of eukaryotic protein-coding genes, Nat. Rev. Genet. 3 (2002) 698–709.
[7] A.A. Salamov, V.V. Solovyev, Ab initio gene finding in Drosophila genomic DNA, Genome Res. 10 (2000) 516–522.
[8] I. Biju, Gene identification in silico: still a long way to go, In National Workshop on Genomics and Proteomics (2001), Chandigarh, India, http://imtech.res.in/bic/workshop/nwgp/chap4.htm.
[9] I. Dunham, et al., The DNA sequence of human chromosome 22, Nature 402 (1999) 489–495.
[10] M. Das, C.B. Burge, E. Park, J. Colinas, J. Pelletier, Assessment of the total number of human transcription units, Genomics 77 (2001) 71–78.
[11] S. Rogic, A.K. Mackworth, F.B. Ouellette, Evaluation of gene-finding programs on mammalian sequences, Genome Res. 11 (2001) 817–832.
[12] P. Flicek, E. Keibler, P. Hu, I. Korf, M.R. Brent, Leveraging the mouse genome for gene predictions in human: from whole-genome shotgun reads to a global synteny map, Genome Res. 13 (2003) 46 (54).
[13] G. Parra, P. Agarwal, J.F. Abril, T. Wiche, J.W. Fickett, R. Guigo, Comparative gene prediction in human and mouse, Genome Res. 13 (2003) 108–117.
[14] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, et al., The Ensembl genome database project, Nucleic Acids Res. 30 (2002) 38–41.
[15] J.P. Wang, et al., EST clustering error evaluation and correction, Bioinformatics 9 (2004) 9.
[16] J. Parkinson, M. Blaxter, Expressed sequence tags: analysis and annotation, Methods Mol. Biol. 270 (2004) 93–126.
[17] R.T. Miller, et al., A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base, Genome Res. 9 (1999) 1143–1155.
[18] P. Bork, R. Copley, The draft sequences: filling in the gaps, Nature 409 (2001) 818–820.
[19] T. Imanishi, et al., Integrative annotation of 21,037 human genes validated by full-length cDNA clones, PLoS Biol. 2 (2004) 856–875.
[20] J.L. Ashurst, J.E. Collins, Gene annotation: prediction and testing, Annu. Rev. Genom. Hum. Genet. 4 (2003) 69–88.
[21] D.D. Shoemaker, et al., Experimental annotation of the human genome using microarray technology, Nature 409 (2001) 922–927.
[22] P. Kapranov, S.E. Crawley, J. Drenkow, S. Bekiranov, R.L. Strausberg, S.P. Fodor, T.R. Gingeras, Large scale transcriptional activity in chromosomes 21 and 22, Science 296 (2002) 916–919.
[23] E.E. Schadt, S.W. Edwards, D. GuhaThakurta, D. Holder, et al., A comprehensive transcript index of the human genome generated using microarrays and computational approaches, Genome Biol. 5 (2004) R73.
[24] J. Cheng, P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, S. Patel, J. Long, et al., Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution, Science 308 (2005) 1149–1154.
[25] S.G. Penn, D.R. Rank, D.K. Hanzel, D.L. Barker, Mining the human genome using microarrays of open reading frames, Nat. Genet. 26 (2000) 315–318.
[26] R. Guigo, et al., Comparison of mouse and human genomes followed by experimental verification yields an estimated 1019 additional genes, Proc. Natl. Acad. Sci. USA 100 (2003) 1140–1145.
[27] J.Q. Wu, D. Shteynberg, M. Arumugam, R.A. Gibbs, M.R. Brent, Identification of rat genes by TWINSCAN gene prediction, RT-PCR and direct sequencing, Genome Res. 14 (2004) 665–671.
[28] J. Reboul, et al., Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in C. elegans, Nat. Genet. 27 (2001) 332–336.
[29] P.D. Thomas, et al., PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification, Nucleic Acids Res. 31 (2003) 334–341.
[30] S. Brenner, et al., Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays, Nat. Biotechnol. 18 (2000) 630–634.
[31] Q. Xu, C. Lee, Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences, Nucleic Acids Res. 31 (2003) 5635–5643;
O.L. Caballero, S.J. de Souza, R.R. Brentani, A.J. Simpson, Alternative spliced transcripts as cancer markers, Dis. Markers 17 (2001) 67–75.
[32] D.A. Schwarz, C.D. Katayama, S.M. Hedrick, Schlafen, a new family of growth regulatory genes that affect thymocyte development, Immunity 9 (1998) 657–668.
[33] M.R. Brent, R. Guigo, Recent advances in gene structure prediction, Curr. Opin. Struct. Biol. 14 (2004) 264–272.
[34] D.L. Black, Mechanisms of alternative pre-messenger RNA splicing, Annu. Rev. Biochem. 72 (2003) 291–336;
Z. Kan, E.C. Rouchka, W.R. Gish, D.J. States, Gene structure prediction and alternative splicing analysis using genomically aligned ESTs, Genome Res. 11 (2001) 889–900.
[35] B. Modrek, A. Resch, C. Grasso, C. Lee, Genome-wide detection of alternative splicing in expressed sequences of human genes, Nucleic Acids Res. 29 (2001) 2850–2859.
[36] C. Burge, S. Karlin, Prediction of complete gene structures in human genomic DNA, J. Mol. Biol. 268 (1997) 78–94.
[37] E.C. Uberbacher, R.J. Mural, Locating protein-coding regions in human

DNA sequences by a multiple sensor-neural network approach, Proc. Natl. Acad. Sci. USA 88 (1991) 11261–11265.
[38] V.V. Solovyev, A.A. Salamov, C.B. Lawrence, Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames, Nucleic Acids Res. 22 (1994) 5156–5163.
[39] T.R. Hughes, Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer, Nat. Biotechnol. 19 (2001) 342–347.
[40] D.M. Rocke, B. Durbin, A model for measurement error for gene expression arrays, J. Comput. Biol. 8 (2001) 557–569.
[41] S. Rozen, H. Skaletsky, Primer 3 1996–1998, Online at http://www-genome.wi.mit.edu/genome_software/other/primer3.html.
[42] B. Ewing, L. Hillier, M.C. Wendl, P. Green, Base-calling of automated sequencer traces using Phred. I. Accuracy assessment, Genome Res. 8 (1998) 175–185.
[43] B. Ewing, P. Green, Base-calling of automated sequencer traces using Phred: II. Error probabilities, Genome Res. 8 (1998) 186–194.
[44] P. Green, Phrap 1994–1999, Seattle, WA, Online at http://www.phrap.org/phredphrapconsed.html.
[45] M.A. Frohman, M.K. Dusk, G.R. Martin, Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer, Proc. Natl. Acad. Sci. USA 85 (1988) 8998–9002.