Splice Variation in Mouse Full-Length cDNAs Identified by Mapping to the Mouse Genome

Mihaela Zavolan, Erik van Nimwegen and Terry Gaasterland

Genome Res. 2002 12: 1377-1385
Access the most recent version at doi:10.1101/gr.191702

Letter

Splice Variation in Mouse Full-Length cDNAs Identified by Mapping to the Mouse Genome

Mihaela Zavolan,[1,3] Erik van Nimwegen,[2] and Terry Gaasterland[1]

[1]*Laboratory for Computational Genomics and* [2]*Center for Studies in Physics and Biology, The Rockefeller University, New York, New York 10021, USA*

We mapped the collection of The Institute of Physical and Chemical Research (Japan) (RIKEN) 21,076 full-length mouse cDNA clone sequences and the mouse RefSeq sequences to the recently completed draft of the mouse genome. Using this mapping, we identified 3674 mouse genes with multiple transcripts, of which 1098 have splice variants. All but 532 of 21,076 clones (97.5%) mapped to the genome assembly. Alignments of cDNA clone sequences with proteins show that much of the detected splice variation alters coding regions and affects the translated protein. We developed novel analytical techniques to classify observed splice variation and to assess the relation between splice variation and alternative transcription. This analysis indicates that an alternative choice of transcription start or polyadenylation signal frequently induces splice variation.

High-quality, full-length cDNA sequences mapped to a high-quality complete genome assembly are crucial for the comprehensive analysis of splice variation. Analysis of 21,076 full-length mouse cDNA clone sequences (RIKEN Genome Exploration Group Phase II Team and FANTOM Consortium. 2001) and of mouse RefSeq sequences (Pruitt and Maglott 2001) mapped to the complete mouse genome assembly (ftp://ftp.ensembl.org/pub/assembly/mouse/mgsc_assembly_3) reveals numerous and complex patterns of splice variation. We developed stringent computational filters to identify and classify splice variants while eliminating cloning, sequencing, and mapping errors. Our computational pipeline identified 3674 mouse genes with multiple transcripts, 1098 of which (30%) have splice variants (Fig. 1). A total of 971 (88%) of the genes with alternative transcripts closely matched GenBank proteins (Benson et al. 2000). The protein-to-DNA alignments indicate that most of the splice variation affects transcript-coding regions. The type of variation observed in initial and terminal exons indicates that alternative use of transcription start site and polyadenylation signals may be frequently responsible for the choice of splice signals flanking these exons.

The variant transcripts reveal many known and novel forms of proteins, including variants of the myosin light chain, phospholipase A2, and a potassium ion channel with alternative 5′ protein sequences, as well as a uridine diphosphate (UDP)-galactose transporter-related protein, variants of osmosis-responsive factor with different 5′ untranslated region (UTR) sequences, and a new form of seryl-tRNA synthase with an internal in-frame extra coding exon. These examples illustrate the breadth of protein function affected by splice variation. They also illustrate a class of variation well represented in our data set, alternative exons possibly associated with alternative start of transcription.

Prior large-scale studies of splice variation have used expressed sequence tag (EST) data to focus on two important questions. What genes in a given genome have splice variants? What signals determine the splicing pattern of a given gene in different tissues? Close to 4 million human ESTs and 3 million mouse ESTs have been deposited in GenBank as of December 2001. This large database of expressed sequence tags (Boguski et al. 1993) has been mined by various groups who attempted to estimate the frequency with which genes undergo splice variation (Mironov et al. 1999) and to identify novel forms of genes (Kan et al. 2001; Modrek et al. 2001; Zhuo et al. 2001). Croft et al. (2000) used an alternative approach for identifying novel gene forms; they compiled a dataset of well-curated spliceosomal introns and identified alternative exons by searching for coding sequences inside of this set of known intronic sequences. The signals flanking alternative exons appear to deviate from the consensus splice signals (Stamm et al. 2000). However, the mechanisms responsible for tissue-specific regulation of splicing are yet to be discovered.

ESTs generally tend to be short relative to the mature mRNA transcript, covering just a few exons, and they do not extend up to the transcription start site. Information about which ESTs come from the same clone is generally absent or incomplete. As the inference of a complete gene structure generally requires multiple ESTs, it is not possible to infer long-range correlations in the choice of splice sites. In this work, we focus on aspects of splice variation that are difficult, if not impossible, to address using EST data. We have developed new methods to use full-length cDNA sequences mapped to the genome to evaluate the impact of splice variation on encoded protein sequences, to discover long-range dependencies in the choice of splice sites, and to detect correlations between splicing and transcriptional events such as the impact of transcription start site and polyadenylation signal on the final splice form.

RESULTS

Mapping cDNAs to the Mouse Genome Assembly

All but 532 RIKEN cDNA sequences (97.5%) mapped to the public assembly of the mouse genome (ftp://ftp.ensembl.org/pub/assembly/mouse/mgsc_assembly_3). Of the 20,544 clones with some genomic match, 2714 had one or more mapping problems. Clones were rejected if internal regions could not be mapped to the genome, if the best mapping had

[3]Corresponding author.
E-MAIL mihaela@genomes.rockefeller.edu; FAX (212) 327-7765.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.191702.

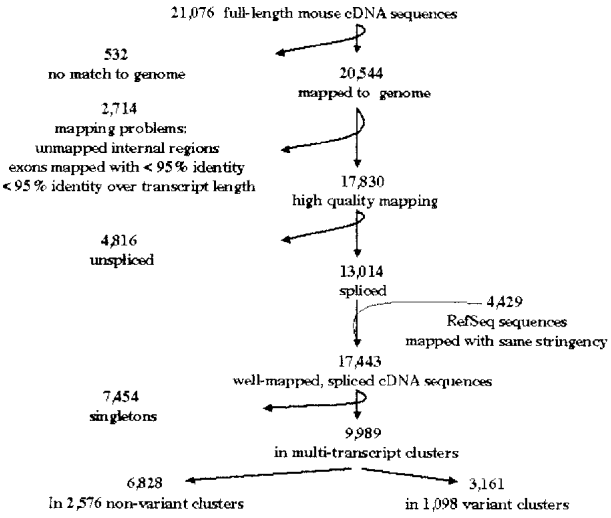

Figure 1 Flow of the data set through the analysis pipeline.

<95% identity over the entire length of the cDNA, or if some exons were mapped at <95 identity. Discarding these sequences yielded 17,830 cDNAs with very high-quality mapping. A total of 13,014 were multiexon clones on which we could study splice variation. To these we added 4429 (of a total of 7027) spliced RefSeq sequences mapped to the mouse genome with the same stringency as the RIKEN sequences. Clones were placed into the same cluster if they mapped in the same orientation and if their genomic mappings overlapped by at least one nucleotide in one exon. This clustering yielded 7454 single-clone clusters, and 9989 clones in multiple-clone clusters. The number of clones per multiclone cluster ranged from 2 to 42, with an average of 2.72. We analyzed this set of clusters for splice variation. The average number of clones was similar in clusters with splice variation and without.

Characterization of Splice Variants

For a gene whose transcripts are always identically spliced, intron and exon boundaries are sharply defined, and a single full-length transcript is sufficient to infer the gene structure. In the presence of splice variation, introns and exons may have different boundaries in different transcripts. In these situations, the identification of exons and introns at the genomic level depends on the entire set of transcripts that we have in the data set, and genomic nucleotides have a certain frequency (ranging from 0 to 1) of occurrence in the mature mRNAs of our data set. We called contiguous blocks of nucleotides with non-zero frequency of occurrence in the mature mRNAs genomic exons (Fig. 2). The number of genomic exons ranged between 2 and 59 per gene, with an average of 8.1.

Each genomic exon is associated with a number of cDNA exons, all those that map to a region of the genome overlapping the genomic exon. The number of 5′ and 3′ splice sites for each genomic exon is defined as the number of distinct 5′ and 3′ splice sites evident in the corresponding set of cDNA exons. To exclude splice variation that could be attributed to incomplete splicing, pairs of splice sites that are associated with a single splice event and that fall inside the genomic exon are not counted. A genomic exon with multiple 5′ and/or 3′ splice sites is considered a variant exon.

Additionally, we identified genomic exons that were included in some, but not all of the cDNAs whose genomic map overlapped that genomic exon. We called these exons cryptic. We made a special note of initial and terminal cryptic exons, because we cannot be certain about their 5′ and 3′ ends, respectively, and they might have been internal exons with splice variation. A total of 2576 multiple-cDNA clusters, containing a total of 6828 cDNAs, did not have any variant splice sites. A total of 1098 multiple-cDNA clusters, containing 3161 cDNAs, had indication of splice variation. Both variant and nonvariant clusters can be viewed at http://genomes.rockefeller.edu (following the link SPLICE VARIATION).

We found different choices of the intron 3′ splice sites in 259 genes, different choices of the intron 5′ splice sites in 241 genes, and cryptic exons in 386 genes. Some genes had multiple variations. Four genes had an exon flanked by both 3′ and 5′ variant splice sites. In a relatively large proportion of cases, we found that the variation occurred in initial or terminal exons that were entirely skipped in other transcripts (these are the type A splice variants defined below). A total of 339 genes had such alternative initial exons, and 273 genes had alternative terminal exons. The reading frame was preserved in 178 of the 423 cases of unique cryptic exons or exon cassettes. This proportion is higher than expected by chance ($P = 0.000196$). However, in the remaining 245 cases, the protein translation of the transcript may be affected.

We annotated the transcripts through the MAGPIE annotation system (Gaasterland and Sensen 1996) using BLASTX (ftp://ncbi.nlm.nih.gov/blast) alignments of the transcripts to the NCBI nonredundant protein database (http://www.ncbi.nlm.nih.gov) to infer functionality.

Genomic exon - contiguous genomic region whose nucleotides have a non-zero frequency of inclusion in cDNA sequences.
cDNA exon - genomic region represented in a particular cDNA (it may contain incompletely-spliced intronic sequence).
Genomic intron - contiguous genomic region whose nucleotides have zero frequency of inclusion in cDNA sequences.

Cryptic exon - cDNA exon present in at least one, but not all transcripts of a cluster.
Exon with 5′-end variation - cDNA exon whose corresponding genomic exon has multiple splice sites at the 5′ end.
Exon with 3′-end variation - cDNA exon whose corresponding genomic exon has multiple splice sites at the 3′ end.
Alternative exon - cDNA exon whose corresponding genomic exon has multiple splice sites at both 3′ and 5′ ends.

Figure 2 Definitions.

Assessing the Impact of Splice Variation on Coding Potential

Although the frequency of human (by extrapolation, mammalian) genes that undergo splice variation is reported to be high (Mironov et al. 1999; Modrek et al. 2001), it is yet unknown how much of this variation is functional. In particular, it is unknown how much of it affects the coding region. To address this ques-

tion in the context of our data set, we constructed for each gene a maximal transcript that included all of the genomic nucleotides that were represented in at least one cDNA and mapped it to the nonredundant protein database. A total of 971 of the 1098 maximal transcripts with splice variation could be mapped to a protein in the nonredundant database at >80% coverage of the protein. We considered these transcripts to be well mapped to a protein and identified their 5′ and 3′ untranslated regions as well as the coding region implied by this map. Regions with <100% frequency of inclusion in the mature mRNA, that is, regions found to be spliced differently in different transcripts, were localized as follows: 179 (11.6%) in the 3′ UTR, 315 (20.4%) in the 5′ UTR, and 1053 (68.1%) overlapping the coding region. Thus, most of the splice variation affects the protein-coding function of the mRNA. Similar numbers (4%, 22%, and 74%, respectively) have been reported by Modrek et al. (2001) in a study of alternative splicing in human genes.

Correlation Between Splice Variation in Initial and Terminal Exons and Variation in Transcription Start and Termination

Visual inspection of the multiple sequence alignments indicates that the variation in transcript length is larger for genes with splice variation. We formalized this observation in terms of the number of exons and the number of nucleotides that appear to be lost from each transcript relative to its corresponding maximal transcript. As Table 1 indicates, transcripts from clusters with splice variation appear to have lost initial and terminal exons more frequently than transcripts from clusters without splice variation. For both initial and terminal exons, the difference is highly significant ($P < 2.2 \times 10^{-16}$). The apparent exon loss is paralleled by a similar apparent loss of nucleotides; more nucleotides appear to have been lost from transcripts of genes with splice variation (Fig. 3). Because it is highly unlikely that the experimental procedure of trapping the mRNA and sequencing of the cDNA would introduce such a bias, this data suggests that the pre-mRNA transcribed from genes with splice variation vary in length, presumably due to the use of different transcription start and polyadenylation signals.

To further investigate the potential relationship between alternative transcription and alternative splicing, we focused on two types of splice variation in initial/terminal exons (Fig. 4). An initial exon of type A could have been as follows: (1) an internal skipped exon that appears to be an initial exon because some sequence was lost during cDNA cloning and sequencing; (2) an internal exon with a variant 3′ end that appears to be an initial exon due to sequence loss; (3) an alternative initial exon produced from an alternative transcription start with subsequent alternative splicing. An initial exon of type B could have been as follows: (1) an internal exon that was incompletely spliced; (2) an internal exon with 5′ end length variation; (3) an initial exon produced from an alternative transcription start. If the loss of sequence is responsible for producing all of these variant initial exons, we expect the relative number of occurrences of the two types of variants to be identical to the relative number of occurrences of their corresponding internal variants. Table 2, which contains the counts of the different splice variants in our data set, shows that this is not the case.

Table 1. Apparent Loss of Initial and Terminal Exons in Transcripts From Variant and Non-Variant Clusters

Initial exons		
No. lost exons	**Transcripts from variant clusters**	**Transcripts from nonvariant clusters**
0	1988 (83%)	4334 (91%)
>0	412 (17%)	424 (9%)
Terminal exons		
No. lost exons	Transcripts from variant clusters	Transcripts from nonvariant clusters
0	1875 (78%)	4333 (91%)
>0	525 (22%)	425 (9%)

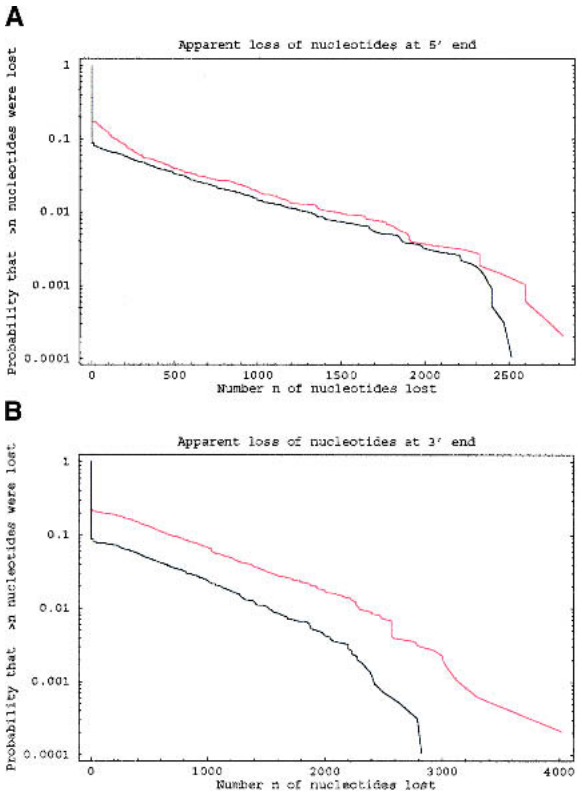

Figure 3 Distribution of the apparent base loss (nucleotides) at the 5′ (*A*) and 3′ (*B*) end of the clones in clusters with (red) and without (black) splice variation.

The data thus suggest that the use of alternative transcription start and polyadenylation sites plays a significant role. Additionally, type A splice variants, which use a cryptic splice site, must be over-represented among transcripts that used an alternative transcription start, and type B splice variants, with a skipped splice site and, thus, an extended exon, must be over-represented in transcripts that use a different polyadenylation site. We derived statistical bounds on the frequency of alternative transcription and alternative polyadenylation site usage as well as on the amount of over-representation of A and B type variants in these alternative transcripts. The results are shown in Figure 5. We infer that the frequency of alternative transcription and polyadenylation site usage in these splice variants is at least 20%. The figure also confirms that variants with alternative transcrip-

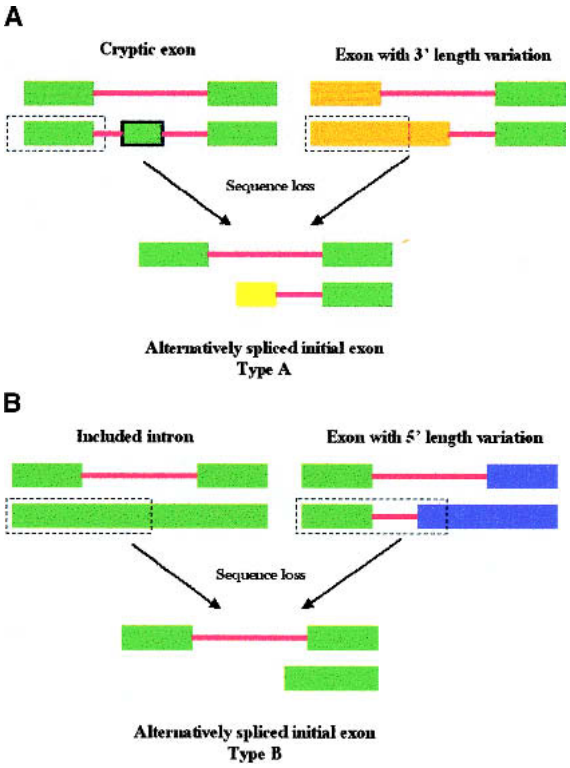

Figure 4 Types of initial exons with splice variation. (*A*) Type A; (*B*) Type B.

tion starts are biased toward the use of cryptic splice sites, whereas those with alternative polyadenylation sites are biased against such usage. The implication is that an alternative transcription start tends to generate different amino-termini, whereas alternative polyadenylation simply tends to truncate the carboxy-termini of the translated proteins.

Examples

Six examples illustrate the types of splice variants detected. The entire data set is posted at http://genomes.rockefeller.edu (link to SPLICE VARIATION) and can be explored via a browser interface.

Figure 6A shows two isoforms of phospholipase A2. The first transcript encodes the secreted form of PLA2D, and so does the RefSeq sequence (accession no: NM_011109). The second transcript, which has a differently spliced first exon, and may have used an alternative transcription start, encodes the nonsecreted form of the enzyme (SPLASH). This form is associated with lymphotoxin deficiency (Shakhov et al. 2000).

Figure 6B shows four transcripts encoding seryl-tRNA synthase. One of these transcripts contains an extra in-frame exon of 72 nucleotides and matches at 91% identity the human seryl-tRNA synthase (GenBank accession no. NP_006504). The other transcripts match the same human protein at >95% identity. We have not been able to find references indicating that seryl-tRNA synthase has known isoforms.

Figure 6C shows two alternatively spliced myosin transcripts with alternative second exons. The first exon of clone ri2310051N24 is skipped in the other clone, ri1100001J17, possibly indicating alternative transcription start. Interestingly, the start codon of the protein encoded in clone ri2310051N24 is found in this exon.

Another example of splice variation inducing proteins with different amino-termini is shown in Figure 6D. Here, one transcript, ri4930448C07, contains two leading exons that have been skipped in the other transcript, as well as in the RefSeq sequence (accession no. NM_010597). The second exon of transcript ri4930448C07 contains a start codon that, if used, allows the potassium channel protein encoded by transcript ri1200009D09 (GenBank accession AK004666) to be extended by 71 amino acids. Both forms have been submitted to GenBank.

In our data set, splice variation often leads to frameshifts. In a cluster representing an UDP-Galactose transporter-related protein (Fig. 6E) we found two transcripts, ri0710001I14 (brain library) and ri1810036N02 (pancreas library) that used the same alternative splice site, inducing a 25-nucleotide deletion, that is, a frameshift. Only the longer form of the protein is present in GenBank (accession no. NM_016752).

We also found splice variation in what seems to be the 5′ untranslated region. For example, Figure 6F shows three transcripts of the osmosis-responsive factor, each using a distinct splice site in the first exon. This exon lies entirely in the 5′ UTR.

DISCUSSION

The RIKEN set of mouse clones is the most comprehensive set of mammalian full-length cDNA clone sequences to date. We analyzed the type and frequency of splice variants present in this set combined with the set of mouse RefSeq sequences, and we found the results to be consistent with previous analyses that used ESTs. Of the genes for which multiple spliced transcripts mapped with high accuracy to the genome, fully 30% have evidence of splicing variation. Most of the variation affects the coding region. Similar results have been reported before on a large set of human ESTs mapped to the human genome (Modrek et al. 2001).

Additionally, we found that a relatively large proportion of the genes with splice variants incur splice variation in the terminal exons in such a way that the respective exon is entirely spliced out in

Table 2. Variant Exon Counts

Internal exons			
ecryptic: 438	5′ variant: 150	3′ variant: 81	incompletely spliced: 13
Initial Exons			
Type A observed 250 (87%)	Type A expected 220 (76%)	Type B observed 39 (13%)	Type B expected 69 (24%)
Terminal Exons			
Type A observed 243 (65%)	Type A expected 322 (86%)	Type B observed 131 (35%)	Type B expected 52 (14%)

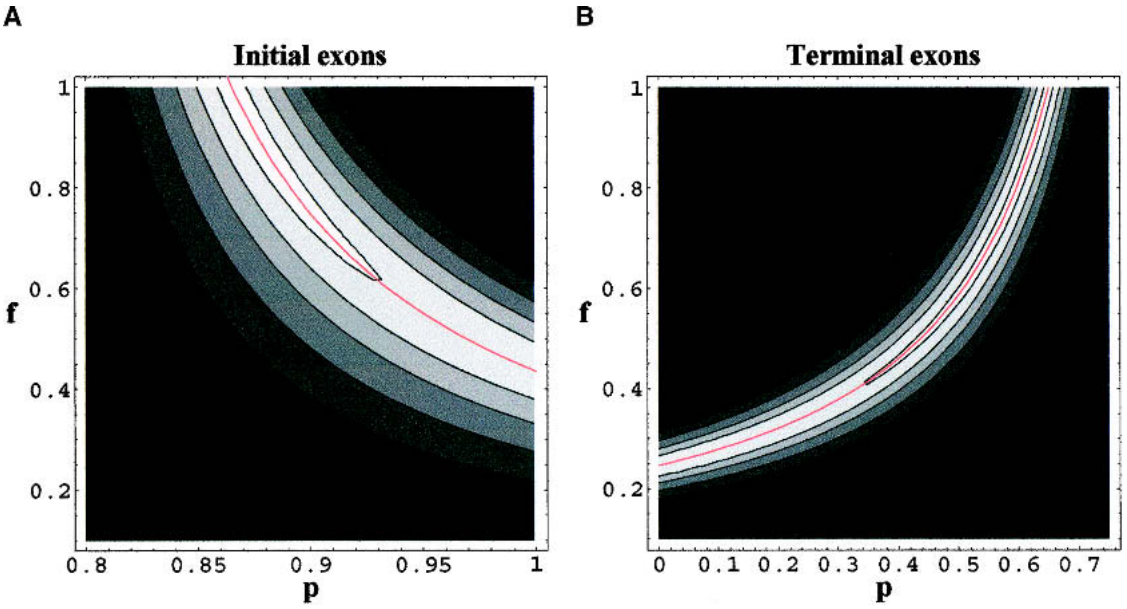

Figure 5 Contour plots of the posterior distribution of *f*, the frequency of using an alternative transcription start (*A*) or polyadenylation (*B*) site, and *p*, the frequency of type A and type B variants (see text) generated. The maximum likelihood solutions are shown in red lines. The contours are drawn at 0.95, 0.75, 0.5, 0.25, 0.1, and 0.05 of the maximum-likelihood value.

another transcript of the same gene. A total of 31% of the genes with splice variation (339/1098) have alternative initial exons, and 25% (273/1098) have alternative terminal exons (some genes have both types of variants). These numbers are comparable with the number of genes with cryptic internal exons (35%, 386/1098). The apparent loss of sequence during the sequencing process is more extensive in the case of genes with splice variation. This suggests that the choice of alternative transcription start and polyadenylation signals frequently alters the splicing pattern. In particular, alternative transcription start biases the splicing process toward the use of a cryptic splice site, whereas alternative polyadenylation tends to generate terminal exons that extend beyond a 3′ splice site.

The extent to which alternative transcription and subsequent alternative splicing shape the proteome is unknown. Accumulated evidence indicates that transcription and splicing are coupled, with the carboxy-terminal domain of RNA polymerase II being implicated at various steps of pre-mRNA processing (Cramer et al. 2001). Our data provides a rich set of candidates for studying this phenomenon.

To eliminate apparent splice variation due to cloning artifacts, we carefully filtered the clones that contributed to our analysis. We only included clones for which the mapping to the genome was complete, allowing small differences in the ends of the clones to accommodate sequencing errors. This conservative approach was taken to reduce the rate of false positives. Once we identified splice variation in initial exons, we checked (using BLASTN) the entire region on which the multiclone cluster was mapped to try to identify alternative mappings of this exon. In none of the cases were we able to identify an alternative mapping. This indicates that the mapping of the variant exon is not an artifact of choices made by the alignment program in situations in which competing choices were available.

The splice sites predicted by SIM4 were used as the basis for splice variant identification. SIM4 attempts to find GT-AG splice signals, which do account for 98.71% of the splice signals (Burset et al. 2001). Also, previous studies (Florea et al.1999) showed that SIM4 performs very well on aligning ESTs to genomic sequences from the same species. If the real splice signals were different, there is the potential of incorrectly identifying the splice sites. However, it is unlikely that the different transcripts in which such a splicing event has taken place would be mapped differently. Moreover, our analysis focused mostly on variations that could not be due to such artifacts.

The complete sequence of a genome provides the basis for understanding the protein functions encoded in that genome. In metazoans, pre-mRNA splicing sometimes produces multiple distinct spliced mRNA sequences from a single transcribed gene, considerably increasing the complexity of the proteome compared with that of the transcribed pre-mRNA pool. The important questions that are now emerging are as follows. Is the variation observed at the transcript level functional; will truncated proteins or proteins with alternative translation be produced in vivo? Is the observed variation regulated, or is it a manifestation of stochasticity in splicing?

Our computational study reveals a high frequency of splice variation in mouse full-length cDNA sequences, similar to that estimated previously for human genes (Mironov et a1. 1999; Modrek et al. 2001). Additionally, the work reported here provides new tools to characterize computationally the functional impact of splice variation. This information will have to be ultimately incorporated into the genome annotation.

METHODS

Data

The RIKEN cDNA sequence dataset downloaded from GenBank contains 21,076 clones sequenced at 99.1% accuracy

Figure 6 (*Continued*)

(RIKEN Genome Exploration Group Phase II Team and FANTOM Consortium 2001). The average sequence length is 1257 nucleotides, and the maximum is 6327 nucleotides. Clones were selected from 160 normalized, subtracted cDNA libraries prepared from a variety of tissues and developmental stages and enriched for full-length clones using biotinylated cap trapping. Clones were sequenced only if an initial 3′ end sequencing read was <95% identical to the 3′ read of any other sequenced clone. This clone-selection strategy ensured that even if clones corresponding to the same gene were sequenced multiple times, the clones would differ at the 3′ end, thus potentially enriching the dataset for 3′ end transcription variation and potentially increasing the number of genes represented in the clone sequences.

The mouse genome sequence was assembled by the international Mouse Genome Sequencing Consortium from >33

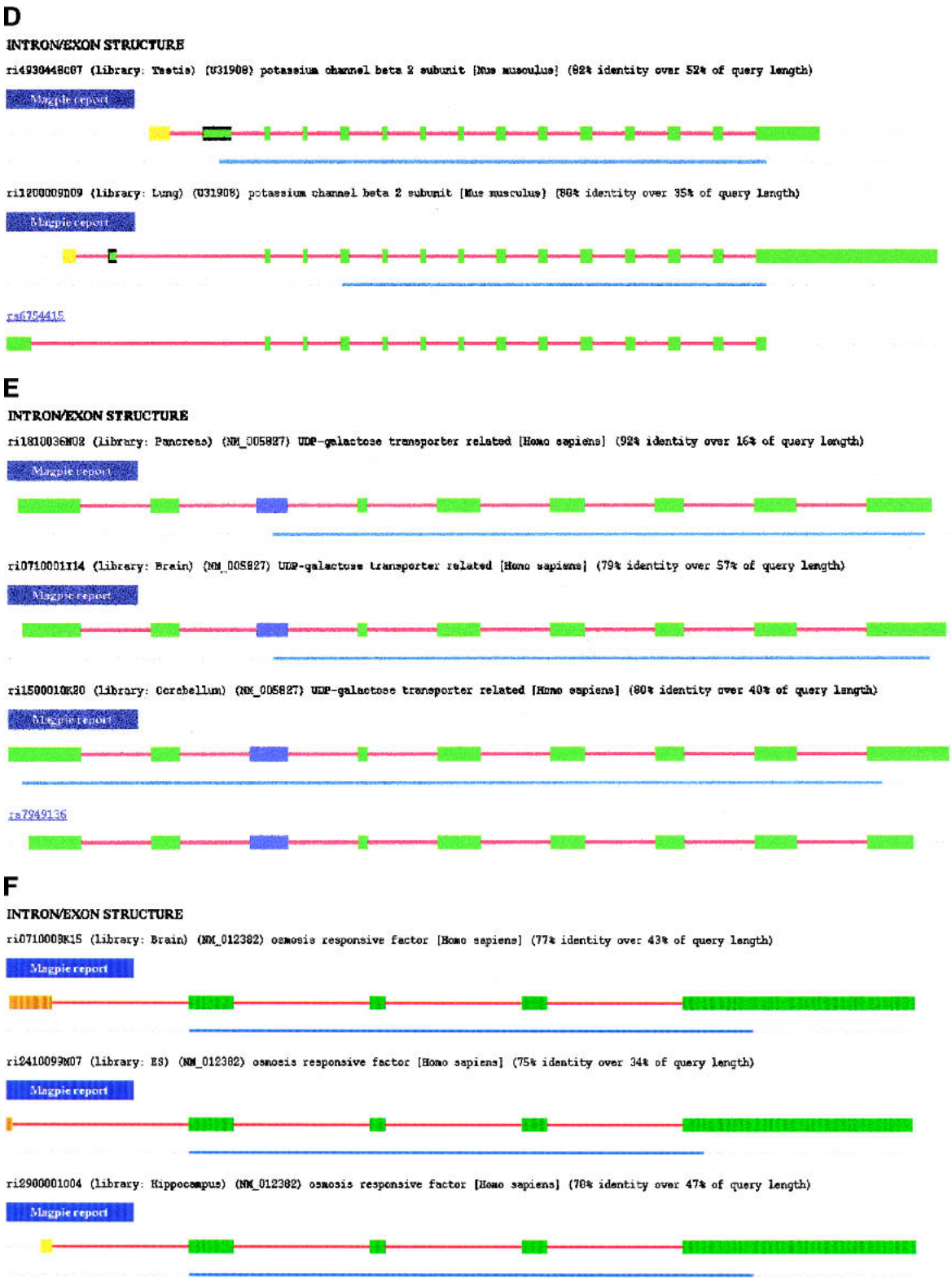

Figure 6 (*A*) Splice variants of phospholipase A2. (*B*) Splice variants of seryl-tRNA synthase. (*C*). Splice variants of myosin light chain. (*D*) Splice variants of a potassium channel. (*E*) Splice variants of UDP-Galactose transporter related-protein. (*F*) Splice variation in osmosis-responsive factor.

million reads, representing approximately sevenfold coverage of the genome. The sequencing strategy combined clone-based, hierarchical shotgun, and whole-genome shotgun. It has been estimated that the resulting draft sequence covers

>96% of the mouse euchromatic genome, nearly all contained in 89 ultracontigs with a typical size of 50 megabases each. The data is available for download at ftp://ftp.ensembl.org/pub/assembly/mouse/mgsc_assembly_3.

Mapping cDNAs to the Mouse Genome Assembly

To speed up the mapping process, we first subjected the entire set of RIKEN clone sequences to BLAST (Altschul et al. 1990) for alignment to the genomic sequence. We determined the coverage of each clone by each potential locus (Kondo et al. 2001), and then performed a more sensitive alignment of the clones to their best genomic targets using the SIM4 program (Florea et al. 1999). SIM4 was designed to accurately align cDNA sequences to genomic DNA, and assumes that mismatches are due to sequencing error or introns (in the genomic sequence). It searches for regions of homology, presumably exons, between the cDNA and the genome, and then looks for GT-AG splice signals at the boundaries of the homologous regions.

Characterization of Splice Variants

We clustered the clones on the basis of their mapping to the genome. Clones were placed into the same cluster if their genomic mappings overlapped by at least one nucleotide in one exon.

We focused on splice variation at the level of genomic exons. These were defined as contiguous regions of the genome whose nucleotides were represented with non-zero frequency in the set of transcripts. We tabulated the following types of variants (illustrated in Figs. 2 and 7): exons present in one transcript, but spliced out in another; exons with splice variation at the 5′ end, due to alternative choice of the flanking 3′ splice site in different transcripts; exons with splice variation at the 3′ end, due to alternative choice of the flanking 5′ splice site in different transcripts; exons flanked by both alternative 3′ and alternative 5′ splice sites. When we tabulated the splice sites found within and at the boundaries of a genomic exon, we did not include those splice sites that were partners in a splicing event, a situation that is encountered when entire introns are included in a transcript. We cannot distinguish between functional inclusion and mere incomplete pre-mRNA splicing, and we chose to discard these cases. Visual inspection of our clusters revealed a very small number of intron inclusions.

Assessing the Impact of Splice Variation on Coding Potential

We constructed the maximal transcript of a multiclone cluster as the concatenation of all genomic nucleotides with non-zero frequency of inclusion in the mRNA. We then searched the nonredundant protein database for the protein that best matches the maximal transcript, using the BLASTX algorithm, and aligned this protein to the maximal transcript using the GENEWISE (Birney and Durbin 2000) program. For genes for which we found a protein in the nonredundant database that was mapped over at least 80% of its length to the maximal transcript, we identified the 5′ and 3′ untranslated regions, as well as the coding region implied by this map.

Assessing Splice Variation in Initial and Terminal Exons

The mapping of each transcript was compared with the inferred mapping of its corresponding maximal transcript to determine the number of initial and terminal exons that appear to be lost. We then tabulated this number separately for transcripts of genes with and without splice variation. The results are presented in Table 1.

Under the assumption that loss of initial and terminal exons is caused solely by loss of nucleotides in the experimental procedure, we can estimate the number of lost nucleotides as follows. We construct the maximal transcript as defined above (and shown in Fig. 7). We then sum the number of nucleotides that appear to be lost from the 5′ and 3′ ends of a given transcript compared with its corresponding maximal transcript. Because genomic exons may be instantiated differently in different transcripts, this number may not be uniquely defined. We chose to calculate a lower bound on the number of lost nucleotides; the length of each genomic exon was assumed to be the length of the shortest cDNA exon that overlaps the given genomic exon. By pooling data from genes with and without splice variation, we constructed cumulative probability distributions for the number of lost nucleotides in genes with and without splice variation. The results are shown in Figure 3.

As discussed above, we distinguish two types of initial and terminal exons with splice variation. Here we will discuss only variation of initial exons; the case of terminal exons is treated entirely analogously. Figure 4 shows the two types of initial exons. Given that some of the nucleotides at the 5′ end may have been lost, the variants of types A and B are consistent with several splicing patterns in the original transcript. Alternatively, such exons might be the result of choosing an alternative transcription start site downstream of the start site of the longer transcript. We want to assess how much evidence our data provides for the occurrence of alternative transcription, and to what extent alternative transcription biases the splice variation in initial exons toward type A or type B. We describe the splice variation in initial exons by the following model. Each transcript with splice variation in the initial exon has a probability f to have used an alternative transcription start site (downstream from the transcription start apparent in the maximal transcript). Given that an alternative transcription start has been used, transcripts have a probability p to generate initial exons of type A [and $(1-p)$ to generate initial exons of type B]. With probability $(1-f)$, the transcript with splice variation in the initial exon did not undergo alternative transcription, but instead underwent nucleotide loss. These transcripts have a probability q to be of type A. Combining these, the probability P_A that a transcript with variation in the initial exon is of type A is thus

$$P_A = fp + (1-f)q, \qquad (1)$$

whereas the probability to be type B is

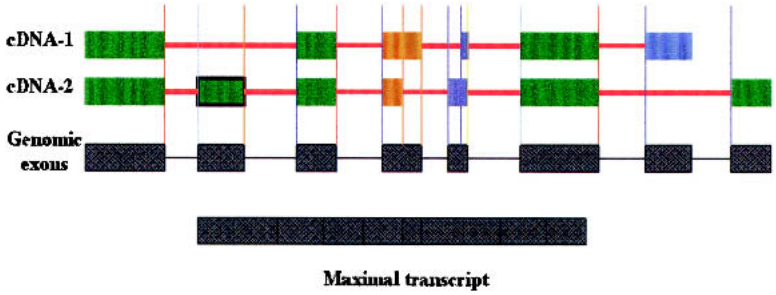

Figure 7 Types of splice variants. Vertical lines indicate splice sites, blue for 3′ splice sites and gold for 5′ splice sites. Genomic exons are shown at *bottom*. Exons are colored to indicate type of variation as follows: green exons have conserved splice sites; blue exons have multiple splice sites at the 5′ end; gold exons have multiple splice sites at the 3′ end. Cryptic exons are shown with a black frame. Terminal exons occurring in the intron of another clone are shown in blue.

$$P_B = f(1-p) + (1-f)(1-q). \quad (2)$$

The probability to observe n_A A type variants and n_B B type variants is then

$$P(n_A, n_B|f,p,q) = (fp + (1-f)q)^{n_A} (f(1-p) + (1-f)(1-q))^{n_B}. \quad (3)$$

We now further assume that the probability q is the same for initial exons as for internal exons. Let m_A be the number of internal genomic exons that have been found to be either cryptic or to have 3′-end variation. Let m_B be the number of internal exons for which we observed 5′-end variation or incomplete splicing/intron inclusion. The probability to observe these numbers given q is

$$P(m_A, m_B|q) = q^{m_A} (1-q)^{m_B}. \quad (4)$$

The probability for all our data given f, p, and q is now

$$P(n_A, n_B, m_A, m_B|f,p,q) = (fp + (1-f)q)^{n_A} (f(1-p) + (1-f)(1-q))^{n_B} q^{m_A} (1-q)^{m_B}. \quad (5)$$

From this we can infer likely values for f, p, and q. In particular, the maximum likelihood values satisfy

$$q^* = \frac{m_A}{m_A + m_B} \equiv w, \quad (6)$$

and

$$f^*p^* + (1-f^*)q^* = \frac{n_A}{n_A + n_B} \equiv u, \quad (7)$$

in which we have introduced the fraction of A types in internal (w) and initial (u) exons to simplify the expressions. Solving for f^*, we find

$$f^* = \frac{u-w}{p^*-w}. \quad (8)$$

Because both f and p have to fall between 0 and 1, we can immediately derive bounds on f^* and p^*. If $u > w$ (more type A in initial exons), we find that $f^* \geq (u-w)/(1-w)$ and $p^* \geq u$. If $u < w$, we have $f^* > (w-u)/w$ and $p^* < u$. In our data set, we have $n_A = 250$, $n_B = 39$, $m_A = 438 + 81$, and $m_B = 150 + 13$. Using those numbers, we find that $f^* > 0.435$ and $p^* > 0.865$. The derivation for terminal exons is entirely analogous to the derivation above for initial exons. There, we find bounds $f^* > 0.246$ and $p^* < 0.65$.

The full posterior distribution $P(f,p|n_A,n_B,m_A,m_B)$ is given by

$$P(f,p|n_A, n_B, m_A, m_B) = \frac{\int_0^1 dq P(n_A, n_B, m_A, m_B|f,p,q)}{\int_0^1 dp \int_0^1 df \int_0^1 dq P(n_A, n_B, m_A, m_B|f,p,q)}. \quad (9)$$

This distribution can be easily obtained numerically and a contour-plot representation of it is shown in Figure 5. Essentially, all probability is contained in a strip of roughly constant width around the maximum likelihood solution for f and p (the red lines). We also note that after integrating over q, these distributions do take on a unique maximum (at $f^* = 1$ for both initial and terminal exons), with very slow decrease of probability along the red curves.

ACKNOWLEDGMENTS

The authors are supported by National Cancer Institute Grant R33-CA84699 and National Science Foundation Grant DBI9984882 to Terry Gaasterland, and by The Rockefeller University Lita Annenberg Hazen Presidential Fellowship to Mihaela Zavolan. We thank The RIKEN Institute for sharing the cDNA data, and Shinji Kondo for help with some of the mappings. 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

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

Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., Rapp, B., and Wheeler, D. 2000. GenBank. *Nucleic Acids Res.* **28:** 15–18.

Birney, E. and Durbin, R. 2000. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* **10:** 547–548.

Boguski, M., Lowe, T., and Tolstoshev, C. 1993. dbEST-database for "expressed sequence tags". *Nat. Genet.* **4:** 332–333.

Burset, M., Seledtsov, I., and Solovyev, V. 2001. SpliceDB: Database of canonical and non-canonical mammalian splice sites. *Nucleic Acids Res.* **29:** 255–259.

Cramer, P., Srebrow, A., Kadener, S., Webajh, S., de la Mata, M., Melen, G., Nogues, G., and Kornblihtt, A. 2001. Coordination between transcription and pre-mRNA splicing. *FEBS Lett.* **498:** 179–182.

Croft, L., Schandorff, S., Clark, F., Burrage, K., Arctander, P., and Mattick, J. 2000. ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nat. Genet.* **24:** 340–341.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G., and Miller, W. 1999. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8:** 967–974.

Gaasterland, T. and Sensen, C. 1996. Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture. *Biochimie* **78:** 302–310.

Kan, Z., Rouchka, E., Gish, W., and States, D. 2001. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res.* **11:** 889–900.

Kondo, S., Shinagawa, A., Saito, T., Kiyosawa, H., Yamanaka, I., Aizawa, K., Fukuda, S., Hara, A., Itoh, M., Kawai, J., et al. 2001. Computational analysis of full-length mouse cDNAs compared with human genome sequences. *Mamm. Genome* **12:** 673–677.

Mironov, A., Fickett, J., and Gelfand, M. 1999. Frequent alternative splicing of human genes. *Genome Res.* **9:** 1288–1293.

Modrek, B., Resch, A., Grasso, C., and Lee, C. 2001. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* **29:** 2850–2859.

Pruitt, K. and Maglott, D. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29:** 137–140.

RIKEN Genome Exploration Group Phase II Team and FANTOM Consortium, 2001. Functional annotation of a full-length mouse cDNA collection. *Nature* **409:** 685–690.

Shakhov, A., Rubtsov, A., Lyakhov, I., Tumanov, A., and Nedospasov, S. 2000. SPLASH (PLA2IID), a novel member of phospholipase A2 family, is associated with lymphotoxin deficiency. *Genes Immunol.* **1:** 191–199.

Stamm, S., Zhu, J., Nakai, K., Stoilov, P., Stoss, O., and Zhang, M. 2000. An alternative-exon database and its statistical analysis. *DNA Cell Biol.* **19:** 739–756.

Zhuo, D., Zhao, W., Wright, F., Yang, H., Wang, J., Sears, R., Baer, T., Kwon, D., Gordon, D., Gibbs, S., et al. 2001. Assembly, annotation and integration of UNIGENE clusters into the human genome draft. *Genome Res.* **11:** 904–918.

WEB SITE REFERENCES

ftp://ftp.ensembl.org/pub/mouse-5.3 Assembly of the mouse whole genome sequence data.

http://www.ncbi.nlm.nih.gov; National Center for biotechnology information.

http://genomes.rockefeller.edu; Laboratory of Computational Genomics.

Received February 15, 2002; accepted in revised form July 18, 2002.