Published online August 3, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 13 ***3977–3983***
doi:10.1093/nar/gkh731

The Alternative Splicing Gallery (ASG): bridging the gap between genome and transcriptome

Jeremy Leipzig, Pavel Pevzner[1] and Steffen Heber*

Department of Computer Science, College of Engineering, North Carolina State University, Raleigh, NC 27695-7566, USA and [1]Department of Computer Science & Engineering, APM 4802, University of California, San Diego, La Jolla, CA 92093-0114, USA

Received January 5, 2004; Revised March 14, 2004; Accepted July 12, 2004

ABSTRACT

Alternative splicing essentially increases the diversity of the transcriptome and has important implications for physiology, development and the genesis of diseases. Conventionally, alternative splicing is investigated in a case-by-case fashion, but this becomes cumbersome and error prone if genes show a huge abundance of different splice variants. We use a different approach and integrate all transcripts derived from a gene into a single splicing graph. Each transcript corresponds to a path in the graph, and alternative splicing is displayed by bifurcations. This representation preserves the relationships between different splicing variants and allows us to investigate systematically all possible putative transcripts. We built a database of splicing graphs for human genes, using transcript information from various major sources (Ensembl, RefSeq, STACK, TIGR and UniGene). A Web interface allows users to display the splicing graphs, to interactively assemble transcripts and to access their sequences as well as neighboring genomic regions. We also provide for each gene an exhaustive pre-computed catalog of putative transcripts—in total more than 1.2 million sequences. We found that ~65% of the investigated genes show evidence for alternative splicing, and in 5% of the cases, a single gene might produce over 100 transcripts.

INTRODUCTION

Alternative splicing is a major link between the estimated 30 000 genes and the myriad of proteins that are believed to be necessary for complex organisms like humans. Previous studies (1–3) reported that over half of all known human genes might be alternatively spliced, and some genes create a vast assortment of different transcripts. Unfortunately, existing *ab initio* gene prediction programs only infer information about one or a small number of most likely transcripts. Our goal here is to complement these programs by providing information about all putative transcripts.

Since expressed sequence tags (ESTs) and cDNAs provide direct evidence for all sampled transcripts, they are currently the most important resources to infer gene structure and alternative splicing.

Typically, these sequences are collected in the gene indices like UniGene (4), the TIGR Gene Index (5), GeneNest (6) and STACK (7). Owing to the fragmentary nature of EST sequences and their sometimes low quality, biologists often assemble them into consensus sequences before using them for further analyses (6,8,9). Several spliced alignment programs, such as sim4 (10), gap2 (11), spidey (12) and BLAT (13), are available for aligning transcripts to genomic sequence and subsequent programs (14–16) have been developed to infer gene structure and predictions about alternative splicing. All these programs represent splicing variants as a list so that a gene with n splicing variants will correspond to a list with n entries. This is hardly efficient since n is often very large, e.g. in our study, 1.7% (380) of the investigated genes have more than 500 assemblies, 0.4% (89) even more than 5000. Even more troublesome, such a representation conceals the relationships between different transcripts.

To overcome these problems, we have developed the Alternative Splicing Gallery (ASG), a web-based tool (http://statgen.ncsu.edu/asg/) that integrates transcript information from Ensembl (17), RefSeq (18), STACK (7), TIGR (5) and UniGene (4) into splicing graphs (19) in order to explore and visualize gene structure and alternative splicing, as well as to compile an exhaustive transcript catalog.

Conceptually, splicing graphs are built by 'projecting' transcribed sequences onto their genomic templates and 'overlaying' these projections (see below for a formal definition). They combine shared segments of different transcripts into single paths and display alternative splicing by bifurcations. Our approach integrates the information of all (even divergent) transcripts of a gene into a single, unambiguously defined data structure, rather than handling them separately. This distinguishes ASG from other alternative splicing databases that partition ESTs with respect to splice variants, e.g. SpliceNest (15), where UniGene clusters are partitioned by assembling them into consensus sequences, or STACK (7), where isoforms are partitioned within 'loose' EST clusters based on a multiple sequence alignment. Such a partition potentially might result in incomplete or even lost transcripts [see (19) and Figure 3]. Although the number of possible transcripts of a gene might be very large, splicing graphs

*To whom correspondence should be addressed. Tel: +1 919 513 2726; Fax: +1 919 513 7315; Email: sheber@ncsu.edu

Table 1. Comparison of ASG with other databases

Database	Methodology	Statistics (human)	Organisms
ASAP (16)	Input: EST/mRNAs (UniGene) Map to genome: BLAST, dynamic programming Analyze genomic-EST-mRNA multiple alignments Tissue-specific results	68 032 EST clusters mapped to genome 44% show alternative splicing	Human Mouse
ASD (20) consists of AltExtron, AltSplice (R1), AEDB	Input: genes (Ensembl), EST/mRNAs (GenBank) AltExtron, AltSplice Computer-generated Map to genome: BLAST AEDB Manually created Literature based	AltSplice: 16 215 genes 77% of genes with >1 transcripts 61 880 transcripts	Human Mouse AltExtron: Model organisms
Ensembl (17) (V22.34d.1)	Input: ESTs (dbEST) Map to genome: Exonerate, BLAST, Est_Genome Redundancy reduction and splice site adjustment Transcript annotation: genome wise	38 581 EST genes 43% show alternative splicing 122 247 transcripts	Human Other metazoan species
PALSdb (21) (R6)	Input: EST/mRNAs (UniGene) Compare ESTs with longest mRNA in cluster No genomic reference	33 111 clusters 43% show alternative splicing	Human Mouse
ProSplicer (22) (R3.0)	Input: genes (Ensembl), mRNAs (UniGene), ESTs (dbEST), proteins (Swiss-Prot, TrEMBL) Map to genome: proteins (BLAST), EST/mRNAs (sim4)	21 786 genes	Human
GeneNest (6) & SpliceNest (15)	Input: ESTs (UniGene) Assemble UniGene clusters into contigs Map to genome: Reputer, sim4	426 178 contigs 31 185 singletons 33 431 clusters mapped to genes 45% show alternative splicing	Human Mouse Fruitfly Zebrafish Arabidopsis
STACKdb (7) (v3.1)	Input: EST/mRNAs (GenBank) Cluster ESTs and assembel cluster Post-cluster assemblies Tissue and disease-specific categories No genomic reference	270 515 cluster 850 835 singletons	Human
TAP (14)	Input: ESTs (dbEST) Map to genome: WuBLAST, sim4 Predict gene structure and Poly(A) sites	1007 multi-exon RefSeq genes 55% show alternative splicing	Human Mouse
ASG	Input: EST/mRNA data (UniGene, TIGR, STACK, RefSeq, Ensembl), genes (Ensembl) Map to genome: BLAST, sim4 Build splicing graphs	22 127 genes 65% show alternative splicing >1.2 millions transcripts	Human

display them all simultaneously. They allow us to investigate systematically all possible assemblies consistent with the input data as well as to recover the corresponding splice variants and their relationships. This complements other alternative splicing databases, which usually try to recover a minimal or most probable set of splice variants. A detailed comparison of ASG with other databases is shown in Table 1.

We analyzed and annotated ASG for alternative splicing events and constructed for each gene (except for 89 genes with more than 5000 assemblies each) an exhaustive set of transcripts. In good concordance with other studies (1,23), we found that ~65% of the genes showed evidence for alternative splicing. Surprisingly, our transcript catalog resulted in total more than 1.2 million sequences—a number that might very well explain the complexity found in humans.

We display splicing graphs with respect to transcripts and the corresponding genomic sequence. A sequence builder allows users to interactively 'assemble' transcripts. As an example, we show in Figure 1 the splicing graph of the human *CBFB* gene (Ensembl gene identifier: ENSG00000067955), which is involved in human leukemogenesis (24) and encodes the β-subunit of the heterodimeric transcription factor core-binding factor (CBF) involved in the regulation of genes important in hematopoiesis (25). The *CBFB* gene contains six exons and spans ~70 kb.

It was previously shown that the last 31 nt of exon 5 can be alternatively spliced (26). The splicing graph confirms this finding (node 5, marked orange) and points out to an additional (and so far unreported) alternative splicing event: skipping of exon 3 (node 3, marked red). This observation is supported by

Figure 1. Visualization of the splicing graph (gray) of the human *CBFB* gene with Ensembl gene identifier: ENSG00000067955 together with the corresponding aligned input transcripts (green) and representative transcript reconstructions (purple). Not drawn to scale! Splice sites are marked by vertical bars. Color-labeled vertices mark annotated alternative splicing events. The highlighted boxes in the sequence builder depict a transcript that skips exon 3 and uses an alternative 5′ splice site in exon 5. Transcripts are displayed with respect to their alignment with the genomic sequence as rows of boxes (aligned regions) connected by dotted lines (putative introns). Only alignments that meet our quality constraints (alignment boundaries correspond to splice sites, sequence identity >95%) are incorporated in the splicing graph.

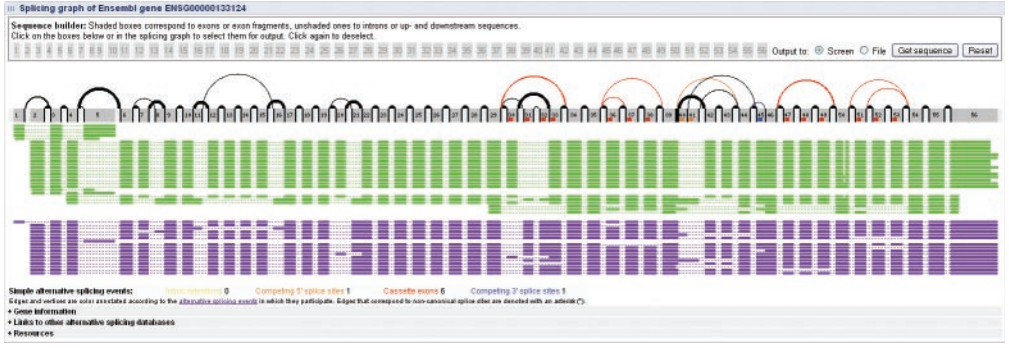

Figure 2. Visualization of the splicing graph (gray) of the human *COL4A6* gene with Ensembl gene identifier ENSG00000133124 together with the corresponding aligned input transcripts (green). Not drawn to scale!

the cDNAs BM462417 (GenBank) and BM477780 (GenBank), both derived from leiomyosarcoma tissue libraries. The splicing graph allows us to generate an exhaustive list of all possible putative transcripts by generating all paths in the graph:

$$t_1 : n_1 \rightarrow n_2 \rightarrow n_3 \rightarrow n_4 \rightarrow n_5 \rightarrow n_6 \rightarrow n_7$$
$$t_2 : n_1 \rightarrow n_2 \rightarrow n_3 \rightarrow n_4 \rightarrow n_5 \rightarrow n_7$$
$$t_3 : n_1 \rightarrow n_2 \rightarrow n_4 \rightarrow n_5 \rightarrow n_6 \rightarrow n_7$$
$$t_4 : n_1 \rightarrow n_2 \rightarrow n_4 \rightarrow n_5 \rightarrow n_7.$$

Such a list could be an invaluable starting point for subsequent research. It immediately raises questions about possible dependences between the alternative splicing events and about which of the transcripts has a biological function. Although at the moment, there is no sufficient data to answer these questions in a high-throughput setting, we consider this as one of the biggest challenges for future work.

Some splicing graphs are considerably more complicated. As an example, we show in Figure 2 the splicing graph of the human collagen, type IV, alpha 6 gene *COL4A6* (Ensembl gene identifier: ENSG00000133124). Type IV collagen is the major structural component of glomerular basement membranes, which compartmentalize tissues and provide important signals for the differentiation of the cells they support. The *COL4A6* gene maps to chromosome Xq22.3 and was found to contain two alternative promoters. The gene seems to be connected with Alport syndrome accompanied by diffuse leiomyomatosis (27,28). The gene belongs to our list of the 89 most complex genes with more than 5000 assemblies. Our annotation shows a total of eight simple alternative splicing events, but the splicing graph reveals an additional large amount of unannotated alternative splicing [alternative promoters, alternative poly(A) sites, and complex and nested events], which might be overlooked by an automated annotation procedure. It is hard to imagine, how a conventional approach could display this complex situation adequately.

MATERIALS AND METHODS

Data sources and preparation

We downloaded UniGene Build #160 (ftp://ftp.ncbi.nih.gov/repository/UniGene/). After pre-processing [vector trimming, poly(A) trimming and the elimination of short sequences], we assembled the UniGene clusters using CAP3 (29) with default parameters. The resulting assemblies were merged with the TIGR Human Gene Index, Version 13.0, Release October 14, 2003 (ftp://ftp.tigr.org/pub/data/tgi/Homo_sapiens/); Stackdb v3.1 of the South African National Bioinformatics Institute (SANBI) (http://www.sanbi.ac.za/Dbases.html); and the set of mRNAs of RefSeq Release 2, October 21, 2003 (http://www.ncbi.nlm.nih.gov/RefSeq/).

Transcript mapping to Ensembl genes

We mapped the above transcripts and EST contigs onto the known Ensembl genes of Ensembl Human release v18.34.1 (http://www.ensembl.org/) using a two-step approach. After masking repeats by RepeatMasker (Smit,A.F.A. and Green,P., http://ftp.genome.washington.edu/RM/RepeatMasker.html), we first identified the candidate Ensembl gene by matching each sequence with the set of Ensembl transcripts using BLASTN (30) with default parameters and *E*-value threshold $E < 10^{-50}$. To establish a match, we require an alignment with an overall identity rate of 95% over more than 100 nt; in case of multiple hits fulfilling these requirements we only use the best match. In the second step, we align the matched sequences combined with the Ensembl transcripts to the corresponding genomic region (derived from Ensembl) plus 10 kb on either end using the spliced alignment program sim4 (10) with reduced word size $W = 8$. Sequences that resulted in low-quality alignments (alignment length smaller than 100 positions, identity score <95%) or inconsistent orientation like overlapping transcripts mapping to the opposite strands of the genome, were discarded.

Splicing graph construction

Splicing graphs are constructed as follows: let $\{s_1, \ldots, s_n\}$ be the set of transcripts for a given gene. Each transcript s_i corresponds via a spliced alignment to a set of genomic positions V_i with $V_i \neq V_j$ for $i \neq j$. Define the set of all transcribed positions $\cup_{i=1}^{n} V_i$ as the union of all sets V_i. The splicing graph G is the directed graph on the set of transcribed positions V that contains an edge (v, w) if and only if v and w are consecutive positions in one of the transcripts s_i. The resulting graph is post-processed to eliminate splices that do not comply with the canonical (GT/AG) or the non-canonical (GC/AG; AT/AC) splice sites and to prune unspliced intron parts. To obtain a more compact representation, we collapse vertices that correspond to consecutive genomic positions (Figure 1).

Transcript generation

In a splicing graph, a transcript is defined as a path from a source to a sink vertex. This definition corresponds to a maximal list of consistent exons and does not capture truncated transcripts, which could result from alternative transcription initiation or termination, but such sequences could be included easily. To create an exhaustive transcript catalog, we traverse all paths from a source to a sink and report the corresponding sequences.

In contrast to conventional EST assembly approaches, splicing graphs will recover all potential putative exon combinations, regardless how often they are represented in the input data or in which order the data are processed (Figure 3). This eliminates much of the ambiguities in current EST assembly algorithms [for an overview see (19)]. On the other hand, in case of dependences between alternative splicing events, e.g. events that always coincide or are mutually exclusive of each other, this approach might combine splice variants that do not co-occur in nature, and yield overpredictions. If such spurious transcripts can be identified, they could be removed easily from our catalog. Unfortunately, to the best of our knowledge, all current methods to determine precisely which splicing events occur in individual isoforms—especially if they affect distant transcript regions—are experimental in nature and do not lend themselves to high-throughput applications [for a more complete overview see (31)].

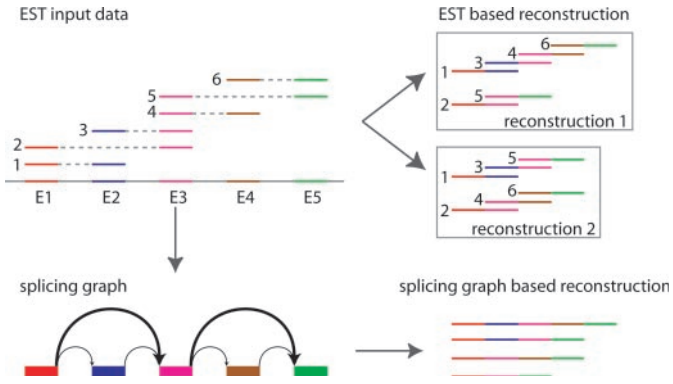

Figure 3. In the presence of alternative splicing, conventional EST-based transcript reconstruction is often incomplete. For example, given the set of displayed ESTs, there are two different ways of assembling (partitioning) all input ESTs into consensus sequences. Both reconstructions are equally computable from the data and explain all ESTs, but each one consists of only two sequences. Dependent on the order of the processed ESTs, a conventional approach might result in either reconstruction and miss the other. In contrast, a splicing graph-based approach does not partition the data but reports exhaustively all four different putative transcripts. However, in the presence of dependences between alternative splicing events, this approach runs the risk of overpredictions by grouping together splicing events that might not co-occur in nature.

Our method does not require that a given alternative splice form is detected in multiple transcripts, but we complement our predictions by a quality value [similar to the approach described in (32)], which tries to assess the degree to which a prediction corresponds to a potential real transcript. Our quality value (range: 0–1, where 0 is bad and 1 is good) penalizes the occurrence of non-consensus splice sites and transcript regions with poor EST support, and it rewards a high overall EST coverage. The precise combination of these paramaters into a single score is heuristically determined based on the inspection of individual transcripts.

RESULTS

In total, approximately 500 000 EST consensus sequences and mRNAs were used to build 22 127 splicing graphs. ASG shows splicing graphs with respect to their corresponding genomic sequence and the input sequences (Figures 1 and 2). Exons or exon fragments are depicted as rectangular nodes, and splices as circular edges between nodes that correspond to non-consecutive genomic positions. Splice sites are marked by vertical bars and non-canonical splice sites are highlighted by an asterisk. Alternative splicing is indicated by positions of in-degree or out-degree larger than 1. The splicing graphs are automatically analyzed and four simple main types of alternative splicing [single and multiple cassette exons, retained introns, competing 5′ and 3′ splice sites; see Figure 4 and (31)] are highlighted by colors. We perform this analysis by identifying graph patterns similar to those in Figure 4. In a splicing graph, exons correspond to adjacent vertices that map to consecutive genomic positions bordered by splice sites. Now, for example, to determine an exon-skipping event, we look for two

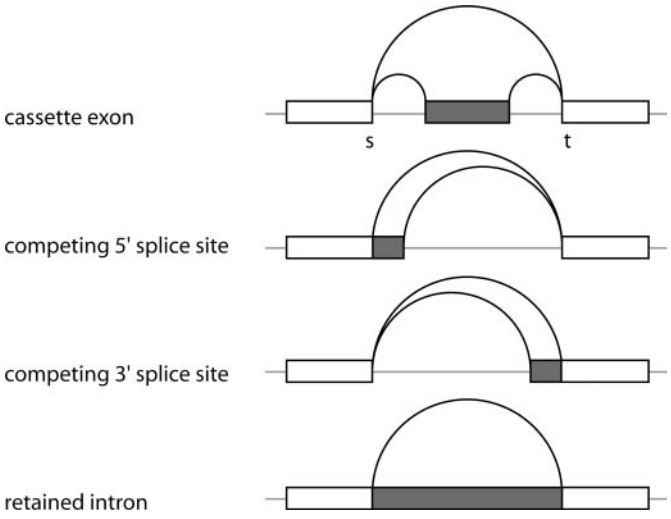

Figure 4. Types of alternative splicing annotated in the splicing graph gallery. Boxes represent exons or exon fragments. Retained introns are often caused by incompletely spliced ESTs and should be interpreted very carefully.

Table 2. Tabulation of simple alternative splicing events and number of genes where they occurred in the ASG consisting of 22 127 Ensembl genes

	Total number	Percentage of genes	Number of genes
Cassette exons	10 940	30.3	6701
Competing 5′ splice sites	4808	17.1	3783
Competing 3′ splice sites	5211	17.8	3935
Retained introns	12 777	31.0	6856

In addition to these numbers, we found over 10 000 more complex or nested alternative splicing events, which did not fall in the above classification, 5879 genes showed evidence for multiple promoters or multiple poly(A) sites. Only ~35% of the genes did not show evidence for alternative splicing.

bifurcation vertices, *s* and *t*, which correspond to a 5′ and a 3′ splice site on the border of different exons, and which are connected by a single edge as well as by a path traversing one or more exons. Similar searches are performed for retained introns, competing splice sites, multiple promoters and poly(A) sites (the latter two are not displayed in the graph), and a detailed description is given in our Web page. Currently, our automated annotation does not identify mutual exclusive exons, complex or nested splicing events, or transcript truncations. The results of our alternative splicing analysis of human genes are summarized in Tables 2 and 3.

A sequence builder allows users to construct and retrieve interactively any transcript supported by the splicing graph—as well as neighboring upstream/downstream regions and introns—by simply selecting the corresponding elements in the splicing graph or the sequence builder. In addition, we provide for each gene (except for 89 genes that produced more than 5000 different assemblies each) a pre-computed exhaustive set of putative transcript reconstructions (i.e. paths in the graph)—in total more than 1.2 million sequences. We also provide a usually much smaller set of representative assemblies that 'cover' the splicing graph. The representative assemblies were chosen by selecting for each splice (i.e. each graph edge) an assembly of maximal length. For each splice site, we generated a probe by concatenating the splice site flanking 30mers of the splicing graph. We used MegaBLAST (33) to search dbEST (34) with these probes. The probe set and the GenBank identifiers of the found ESTs, which support the splice sites, can be downloaded from our Web page. A list of the genes, which produced more than 5000 different assemblies, is provided as Supplementary Material.

Table 3. Distribution of the number of transcript reconstructions per gene in the Alternative Splicing Gallery consisting of 22127 Ensembl genes

Transcripts per gene	Percentage of genes	Number of genes
1	34.9	7722
2	15.7	3471
3–4	14.8	3282
5–10	11.5	2547
11–20	8.1	1781
21–50	6.9	1518
51–100	3.1	694
101–200	1.8	400
201–500	1.5	332
501–5000	1.3	291
>5000	0.4	89

We display for each gene basic information like genomic position, gene description, Gene Ontology annotation (http://www.geneontology.org/), OMIM annotation (4), known PFAM domains (35) and provide links to other alternative splicing databases [ASD (20), ASAP (16), HASDB (23), PALSdb (21), ProSplicer (22) and SpliceNest (15)]. ASG can be queried by using source database identifiers or by a BLAST (30) search.

DISCUSSION

ASG is a compact genome-based representation of the huge quantity of EST and cDNA data—designed as a starting point for the systematic investigation of gene structure and the transcriptome. We integrated transcript data from RefSeq, Ensembl, UniGene, STACK and TIGR with respect to the set of Ensembl genes into splicing graphs. Combining these various data sources has several advantages. We get a more complete overview, and reduce potential bias introduced by different EST clustering strategies, an important point, since the large amount of missed real splice forms is a big disadvantage of any method that maps EST data to genomic sequence (2). In addition, by merging EST data with full-length mRNAs and model sequences we overcome the problem of coverage gaps in gene structure and transcript prediction. Since splicing graphs combine reoccurring transcript segments into single paths and display alternative splicing as bifurcations, they yield a compact and biologically meaningful visualization, which highlights potential splice variants. We automatically annotate the main simple types of alternative splicing, and in contrast to most other alternative splicing databases; we also display other more complex events. The essential advantage of splicing graphs over conventional representations is that they preserve the relationships between splice variants and therefore allow us to

systematically generate and analyze all putative transcripts represented by the input data. This is an important prerequisite for the analysis and quantification of complex splicing patterns, the investigation of mRNA splicing regulation and for cataloging the transcriptome.

We derived for each gene a small set of representative putative transcripts as well as an exhaustive catalog. Since our approach explores all possible compatible splice variations it might overpredict the number of transcripts in the case of dependences between alternative splicing events of a gene. If identified, the spurious transcripts could be removed easily from our catalog. Unfortunately, current high-throughput techniques in general cannot determine such dependences with much certainty (31). Since our goal was to complement existing techniques by an algorithm that provides an exhaustive transcript catalog, we refrained from applying additional filtering steps at this stage to avoid omitting a real variant. We did, however, complement our transcript reconstructions by a quality value, which ranks transcripts with respect to the occurrence of non-standard splice sites and regions of poor EST support. This allows users to further filter and prioritize our predicitions. In addition to the above transcript catalogs, ASG offers a sequence builder that allows users to interactively assemble exons, and to retrieve upstream and downstream regions, or introns by simply 'clicking' on the corresponding elements. This feature is especially helpful for investigating gene structure or for finding regulatory sequences. Following the suggestion of a very helpful anonymous referee, we interconnected our database with other alternative splicing databases. This allows users to compare and combine our results with other approaches, as well as to complement ASG with additional information.

Quantifying the number of different transcripts that originate from a single gene under certain conditions is a fascinating and sparsely addressed dimension of the hidden transcriptome, which exceeds simply cataloging alternative splicing events. Our database is only a first step toward this direction. Although we neither expect that each of our *in silico* reconstructions corresponds to a biological functional transcript nor that we reconstructed all such transcripts, our database highlights a set of genes which potentially produce hundreds of different proteins.

To illustrate the biological importance of such genes, we investigated in a preliminary study (data not shown) their frequency among the genes involved in inherited human disease, which are stored in OMIM (4). We found a highly significant (P-value = 2.2×10^{-16}) overrepresentation of genes with multiple transcripts, in average OMIM genes produced over 50% more transcripts than others. Although this has to be interpreted very carefully—one could for example argue that due to the high interest in desease genes databases are biased or that desease genes might have a higher transcription levels which could result in more biologically non-functional erroneous transcripts—we hypothesize that genes with multiple transcripts are of fundamental biological importance and therefore more likely to be involved in desease. Screening these genes in different tissues and under different conditions as well as investigating them with respect to their function, evolution and involvement in diseases are interesting challenges for future research.

Future work

The current version of ASG does not display annotations of coding sequences, promoters, polyadenylation sites, the strength of splice sites and transcript truncations. We plan to include these features together with an accompanying protein section in a future edition of the gallery.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Chris Smith and Dr Christopher Basten for computer support and very valuable discussions.

REFERENCES

1. Mironov,A., Fickett,J. and Gelfand,M. (1999) Frequent alternative splicing of human genes. *Genome Res.*, **9**, 1288–1293.
2. Modrek,B. and Lee,C. (2001) A genomic view of alternative splicing. *Nature Genet.*, **30**, 13–19.
3. Graveley,B. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, **17**, 100–107.
4. Wheeler,D., Church,D., Federhen,S., Lash,A., Madden,T., Pontius,J., Schuler,G., Schriml,L., Sequeira,E., Tatusova,T. and Wagner,L. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
5. Quackenbush,J., Cho,J., Lee,D., Liang,F., Holt,I., Karamycheva,S., Parvizi,B., Pertea,G., Sultana,R. and White,J. (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.*, **29**, 159–164.
6. Haas,S., Beissbarth,T., Rivals,E., Krause,A. and Vingron,M. (2000) GeneNest: automated generation and visualization of gene indices. *Trends Genet.*, **16**, 521–523.
7. Christoffels,A., vanGelder,A., Greyling,G., Miler,R., Hide,T. and Hide,W. (2001) STACK: Sequence Tag Alignment and Consensus Knowledgebase. *Nucleic Acids Res.*, **29**, 234–238.
8. Burke,J., Wang,H., Hide,W. and Davison,D. (1998) Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Res.*, **8**, 276–290.
9. Zhuo,D., Zhao,W., Wright,F., Yang,H., Wang,J., Sears,R., Baer,T., Kwon,D., Gordon,D., Gibbs,S., Dai,D., Yang,Q., Spitzner,J., Krahe,R., Stredney,D., Stutz,A. and Yuan,B. (2001) Assembly, annotation, and integration of UNIGENE clusters into the human genome draft. *Genome Res.*, **11**, 904–918.
10. Florea,L., Hartzell,G., Zhang,Z., Rubin,G. and Miller,W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, **8**, 967–974.
11. Huang,X., Adams,M., Zhou,H. and Kerlavage,A. (1997) A tool for analyzing and annotating genomic sequences. *Genomics*, **46**, 37–45.
12. Wheelan,S.J., Church,D.M. and Ostell,J.M. (2000) Spidey: a tool for mRNA-to-genomic alignments. *Genome Res.*, **11**, 1952–1957.
13. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
14. Kan,Z., Rouchka,E.C., Gish,W.R. and States,D.J. (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res.*, **11**, 889–900.
15. Coward,E., Haas,S. and Vingron,M. (2002) SpliceNest: visualizing gene structure and alternative splicing based on EST clusters. *Trends Genet.*, **18**, 53–55.
16. Lee,C., Atanelov,L., Modrek,B. and Xing,Y. (2003) ASAP: the Alternative Splicing Annotation Project. *Nucleic Acids Res.*, **31**, 101–105.
17. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.

18. Pruitt,K. and Maglott,D. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res*., **29**, 137–140.
19. Heber,S., Alekseyev,M., Sze,S.H., Tang,H. and Pevzner,P.A. (2002) Splicing graphs and EST assembly problem. *Bioinformatics*, **18** (Suppl. 1), 181–188.
20. Thanaraj,T.A., Stamm,S., Clark,F., Riethoven,J.J., Le Texier,V. and Muilu,J. (2004) ASD: the Alternative Splicing Database. *Nucleic Acids Res*., **32**, 64–69.
21. Huang,Y.H., Chen,Y.T., Lai,J.J., Yang,S.T. and Yang,U.C. (2002) PALS db: putative alternative splicing database. *Nucleic Acids Res*., **30**, 186–190.
22. Huang,H.D., Horng,J.T., Lee,C.C. and Liu,B.J. (2003) ProSplicer: a database of putative alternative splicing information derived from protein, mRNA and expressed sequence tag sequence data. *Genome Biol*., **4**, R29.
23. 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.
24. Liu,P., Tarle,S., Hajra,A., Claxton,D., Marlton,P., Freedman,M., Siciliano,M. and Collins,F. (1993) Fusion between transcription factor CBF beta/PEBP2 beta and a myosin heavy chain in acute myeloid leukemia. *Science*, **261**, 1041–1044.
25. Wang,S., Wang,Q., Crute,B.E., Melnikova,I.N., Keller,S.R. and Speck,N.A. (1993) Cloning and characterization of subunits of the T-cell receptor and murine leukemia virus enhancer core-binding factor. *Mol. Cell. Biol*., **13**, 3324–3339.
26. van der Reijden,B.A., Lombardo,M., Dauwerse,H.G., Giles,R.H., Muhlematter,D., Bellomo,M.J., Wessels,H.W., Beverstock,G.C., van Ommen,G.J., Hagemeijer,A. *et al*. (1995) RT–PCR diagnosis of patients with acute nonlymphocytic leukemia and inv(16)(p13q22) and identification of new alternative splicing in CBFB-MYH11 transcripts. *Blood*, **86**, 277–282.
27. Zhang,X., Zhou,J., Reeders,S.T. and Tryggvason,K. (1996) Structure of the human type IV collagen COL4A6 gene, which is mutated in Alport syndrome-associated leiomyomatosis. *Genomics*, **33**, 473–479.
28. Zhou,J., Mochizuki,T., Smeets,H., Antignac,C., Laurila,P., de Paepe,A., Tryggvason,K. and Reeders,S.T. (1993) Deletion of the paired alpha 5(IV) and alpha 6(IV) collagen genes in inherited smooth muscle tumors. *Science*, **261**, 1167–1169.
29. Huang,X. and Madan,A. (1999) CAP3: a DNA sequence assembly program. *Genome Res*., **9**, 868–877.
30. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
31. Roberts,G.C. and Smith,C.W.J. (2002) Alternative splicing: combinatorial output from the genome. *Curr. Opin. Chem. Biol*., **6**, 375–383
32. Gupta,S., Zink,D., Korn,B., Vingron,M. and Haas,S.A. (2004) Genome wide identification and classification of alternative splicing based on EST data. *Bioinformatics*, April 29 (Epub ahead of print).
33. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol*., **7**, 203–214.
34. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST-database for ''expressed sequence tags''. *Nature Genet*., **4**, 332–333.
35. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res*., **30**, 276–280.