doi:10.1006/geno.2002.6841, available online at http://www.idealibrary.com on IDEAL

Computational Analysis of Alternative Splicing Using EST Tissue Information

Hanqing Xie,* Wei-yong Zhu, Alon Wasserman, Vladimir Grebinskiy, Andrew Olson, and Liat Mintz

Compugen Inc., 7 Centre Drive, Jamesburg, New Jersey 08831, USA

To whom correspondence and reprint requests should be addressed. Fax: (609) 655-5114. E-mail: han@cgen.com.

Expressed sequence tags (ESTs) from normal and tumor tissues have been deposited in public databases. These ESTs and all mRNA sequences were aligned with the human genome sequence using LEADS, Compugen's alternative splicing modeling platform. We developed a novel computational approach to analyze tissue information of aligned ESTs in order to identify cancer-specific alternative splicing and gene segments highly expressed in particular cancers. Several genes, including one encoding a possible pre-mRNA splicing factor, displayed cancer-specific alternative splicing. In addition, multiple candidate gene segments highly expressed in colon cancers were identified.

Key Words: alternative splicing, neoplasms, computational biology, gene expression

Introduction

The etiology of many cancers, especially those involving multiple genes or sporadic mutations, remains unknown. Expressed sequences (mRNA and ESTs [1]) from various normal or cancer tissues and cell lines have been accumulated in public sequence databases. This wealth of EST information, though a majority of it comes from heterogeneous cell types of different tissues, captures some of the changes inherent in carcinogenesis. We attempted to identify some of these changes through computational analysis of EST clustering and EST tissue information. Several earlier studies have examined EST tissue information. Counting the number of ESTs in UniGene clusters [2], with or without rigorous probability calculation [3], was used to identify endothelium-specific genes [4], disease-specific or tissue-specific polyadenylation sites [5], colon cancer-related genes [6], and genes differentially expressed in normal or cancer tissues [7,8], and to build tissue expression profiles for adult skeletal muscle [9] and retina [10]. ESTs from well-defined tissue sources were used to construct a sophisticated BodyMap [11]. However, all of these published approaches failed to consider alternative splicing—estimated to occur in over 50% of human genes [12–15]—because UniGene clusters do not have multiple alignments. In addition, ESTs in these studies were restricted to those from non-normalized libraries. SAGE [16,17] and microarray experiments [18] have been used extensively to study gene expression, but these methodologies must be linked with alternative splicing modeling to be of use for investigating alternative splicing.

Results and Discussion

Human EST and mRNA sequences were aligned against genomic sequences and clustered through Compugen's LEADS platform [19–22], which identified the boundaries of introns and predicted alternative splicing sites (Fig. 1). Modeling of alternative splicing has been reported with different degrees of sophistication [14,23,24]. The 20,301 clusters with 2.0 million ESTs contained at least one mRNA sequence, in general agreement with UniGene build #148 with 20,876 mRNA-containing clusters. The remaining EST sequences were clustered to unknown regions of known genes or to unknown genes. These ESTs were not analyzed. Table 1 provides some statistics about EST and mRNA clustering. There were 125,115 introns and 213,483 exons aligned either with an mRNA or with ESTs from at least two libraries if there was no RNA aligned to the gene segment. Alternative splicing includes exon skipping, alternative 5′ or 3′ splicing, and intron retention. All of them can be described by one simple rule, that is, a single exon connects to at least two other exons in either the 3′ end (donor site) or the 5′ end (acceptor site; Fig. 1). Table 2 lists some statistics of alternative splicing events based on this simplification.

We analyzed tissue information of ESTs in this cohort of clusters. Table 3 lists the 10 tissue types with the largest numbers of ESTs along with those from pooled or uncharacterized tissues. An alternative splicing event has at least two donor–acceptor concatenations. In more than half of the cases, concatenations between donor (exon A; Fig. 1) and proximal

0888-7543/02 $35.00

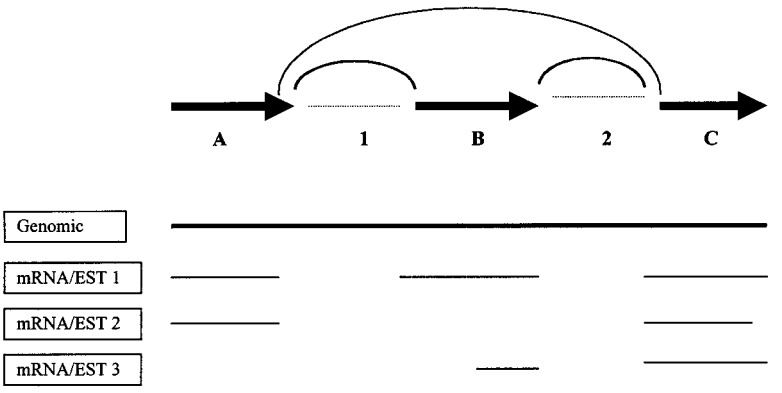

FIG. 1. Schematic representation of LEADS alignment and alternative splicing with three exons (A, B, C) and two introns (1, 2). ESTs and mRNAs were aligned to the genome and the splicing junctions were determined through the alignments. Two alternative splicing events are distinguished here. One, from the donor site, involves AB (between donor and proximal acceptor) and AC (between donor and distal acceptor), and the other, from the acceptor site, involves AC (between distal donor and acceptor) and BC (between proximal donor and acceptor).

acceptor (exon B) and between acceptor (exon C) and proximal donor (exon B) are supported by a higher number of EST libraries than those between donor (exon A) and distal acceptor (exon C) and between acceptor (exon C) and distal donor (exon A). This result indicates that exon skipping is not prevalent. Very few concatenations (three in prostate, one in lung, and one in placenta) [25] supported by ESTs from more than four libraries were restricted to a single tissue type, suggesting that the absolute tissue-specific alternative splicing might be rare among the genes analyzed. The non-quantitative nature of the current analysis method precludes the identification of alternative splicing events that are tissue-specific, yet are not restricted to a single tissue type. In these few cases, the whole genes tend to be tissue-specific, some of which have been regarded as tissue markers, and those identified events may serve as more specific tissue and diagnostic markers, as

TABLE 1: The number of clusters with different numbers of EST or mRNA after LEADS alternative splicing modeling for GenBank version 125 with genomic build #25

EST	Cluster
1	963
2–3	1457
4–7	1532
8–15	1655
16–31	1879
32–63	2500
64–127	3481
128–255	3240
256–511	1406
512–1023	422
1024–above	1766
Total	20301
RNA	Cluster
1	6527
2–3	6372
4–7	6204
8–15	1915
16–31	226
32–63	40
64 and above	17
Total	20301

Clusters must contain at least one mRNA alignment.

TABLE 2: The number of clusters with different numbers of alternatively spliced donor sites or acceptor sites

Donor site	Cluster
1	3690
2	2269
3	1348
4	760
5	435
6 and above	566
Total	9068
Acceptor site	Cluster
1	3751
2	2388
3	1511
4	799
5	508
6 and above	710
Total	9667

Donor–acceptor concatenation must be supported by at least one mRNA or by ESTs from at least two libraries. There are 8254 clusters which have alternatively spliced donor and acceptor sites. If the lower bound on the number of EST libraries supporting each donor–acceptor concatenation is increased to three, there are 13,402 alternatively spliced donor sites in 6892 clusters and 15,015 alternatively spliced acceptor sites in 7570 clusters, whereas 6111 clusters have alternatively spliced donor and acceptor sites.

doi:10.1006/geno.2002.6841, available online at http://www.idealibrary.com on IDEAL

TABLE 3: Statistics of ESTs aligned in the 20,301 clusters from the 10 most prevalent tissue types and pooled or uncharacterized tissues in the LEADS output

Tissue	Number of ESTs			Number of libraries		
	Normal	Cancer	Total	Normal	Cancer	Total
Brain	93024	87803	180827	30	25	55
Lung	35455	85596	121051	92	156	248
Placenta	86571	27291	113862	259	3	262
Uterus	30052	71521	101573	99	107	206
Colon	23796	74998	98794	274	445	719
Kidney	42628	46811	89439	9	54	63
Skin	32436	43085	75521	8	10	18
Prostate	40312	27963	68275	131	135	266
Mammary gland	26509	36638	63147	305	665	970
Head and neck	12354	50167	62521	62	800	862
Pooled	178618	992	179610	15	1	16
Uncharacterized	76193	9721	85914	778	106	884

suggested earlier [26]. For example, the short form of a prostate-specific protein, PSP57 [27], resulted from the concatenation of exon 2 and exon 4, which is supported by ESTs from 11 prostate libraries, whereas in the long form, PSP94, concatenations of exons 2, 3, and 4 were supported by ESTs from six types of tissues. We tried to identify discordant (or mutually exclusive) expressions of alternatively spliced transcripts, where the transcripts are expressed in non-overlapping sets of tissues. For that purpose, we examined alternative splicing events with only two concatenations, each of which was supported by ESTs from at least five libraries, where the two concatenations were supported by ESTs from different tissues. Six acceptor sites and seven donor sites have been identified [25]. As an example, in the gene *XRCC3*, the first 36 bp, the next 57 bp, and the subsequent 102 bp belong to three exons (here named A, B, and C). The concatenation of exon A to exon C is supported by mRNA sequence BC011725 and ESTs from five libraries originating from lymph node, cervix, muscle, mammary gland, and eye. The concatenation of exon B to C is supported by multiple mRNA and ESTs in nine libraries from nervous system, placenta, uterus, gastrointestinal tract, kidney, and pancreas. Cell heterogeneity of most tissues in EST databases precludes the identification of cell type-specific alternative splicing.

We examined alternative splicing events that are restricted to cancer tissues by looking for any donor–acceptor concatenations exclusively supported by ESTs from cancer tissues. Table 4 lists six interesting examples [25]. The gene *NONO* with BC003129 and 1496 ESTs encodes a possible splicing factor, suggesting that alternative splicing of multiple genes may be regulated during carcinogenesis. Lack of a complete and accurate inventory of transcripts hinders the identification of transcripts that are highly expressed in particular cancer tissues. However, the EST clustering and tissue information analyses can be used to identify specific gene segments that are highly represented in cancer tissues. We ranked all gene segments with probability scores and identified several possible colon cancer markers [25] (Table 5). Gene segments highly expressed in other tissues or other types of cancers can be identified similarly. During a selection of 200 gene segments for lung cancer markers, several recently published lung cancer markers [28,29], including AA033947, AA600214, AA664179, H58872, X53463, M11507, and X01060, were identified. With systematic classification of histology and oncology, such as Gene Ontology [30,31], sufficient numbers of sequences from homogenous cells, and accurate and detailed descriptions of tissue sources and library construction, the approach outlined here may identify some of the genetic alterations in carcinogenesis.

TABLE 4: Examples of putative cancer-specific alternative splicing

mRNA/EST	UniGene ID	Position	Total		Type	Specific		Non-specific			Possible function
			EST	RNA		EST	RNA	EST		RNA	
								Cancer	Normal		
BC003129	172207	123, 237	1496	8	d+	15	1	46	20	3	splicing factor candidate
NM_018035	279851	220, 301	584	2	d-	7	0	21	9	2	no known function
AL519365	21938	474, 513	162	3	s-	8	3	6	1	0	oxysterol binding
BF341144	155596	507,542	148	1	s+	6	0	7	4	1	BCL2/adenovirus E1B interacting
AB009357	7510	1372,1452	205	6	s+	7	4	2	4	2	MAPKKK 7
NM_002382	42712	57,84	165	7	s-	8		7	3	6	MAX protein

One of the mRNAs, or one of the ESTs if no mRNA contains both splicing junctions, is listed to identify the cluster. Under the "Type" column, the designation from either the donor site (d) or acceptor site (a) and to either the proximal (+) or distal (-) exon indicates the type of transcript shown to be cancer-specific. For example, "d+" indicates AB (Fig. 1) is cancer-specific. In cases of exon skipping or intron retention, the cognate donor site and acceptor site showed same or similar profile. Please note: under the "Total" column the number of ESTs or mRNAs is listed. Under "Specific" or "Nonspecific" columns, the library count is listed. All mRNA sequences under "Specific" are from cancer tissues, and there were no normal ESTs under "Specific." The numbers under the "Position" column identify the splicing junctions on the mentioned sequence.

TABLE 5: A select group of gene segments highly expressed in colon cancer

mRNA/EST	UniGene ID	EST	RNA	Position	Score	Number of ESTs			Number of Libraries			Function
						Cancer	Normal	Total	Cancer	Normal	Total	
NM_032044	105484	105	3	484–589	57	11	25	46	4	18	29	gastrointestinal secretory
NM_033049	5940	92	3	1566–2866	44	9	33	73	6	15	38	mucin 13
NM_002083	2704	220	7	256–564	42	46	34	196	9	17	70	glutathione peroxidase
AK000683	273321	216	2	1343–2306	37	11	30	164	8	13	49	unknown
NM_006408	91011	256	4	388–452	36	13	37	165	7	17	71	XCG homolog
NM_002273	242463	1081	9	1451–1509	35	33	90	562	8	25	146	keratin 8
NM_005814	143131	26	1	1171–2670	33	7	10	20	5	7	15	glycoprotein A33

One of the mRNAs, or one of the ESTs if no mRNA contains the segment, is listed to identify the cluster. Under "EST" or "RNA," the number of ESTs or mRNAs is listed. The ranges in the "Position" column identify the selected segment from the mRNA or EST sequence listed under "mRNA/EST." Scores are the – logarithmic of the probabilities, calculated under binomial distribution.

MATERIALS AND METHODS

DATA and LEADS alternative splicing modeling. GenBank version 125 with genomic build #25 from the National Center for Biotechnology Information (NCBI) was input to the LEADS platform as described [19–22]. The LEADS process aligned about 80,000 mRNA and 3.7 million ESTs with genomic sequences. The mRNAs and ESTs supporting any expressed genomic segment or donor–acceptor concatenations were identified. Results from the application of the LEADS platform have also been reported [19–22]. UniGene Build #146 and libraryQuest.txt were obtained from the NCBI and Cancer Genome Anatomy Project (CGAP) and the National Cancer Institute (NCI), respectively.

EST tissue information. Most ESTs have tissue library information. The information is also available in web form in Library Browser or Library Finder in NCBI or in the flat file libraryQuest.txt. The file lists 53 tissue sources, five histological states (cancer, multiple histology, normal, pre-cancer, and uncharacterized histology), six types of tissue preparations (bulk, cell line, flow-sorted, microdissected, multiple preparation, and uncharacterized), and brief descriptions for each library. The ORESTES set with 5000 libraries has close to one million ESTs [32]. The 5318 libraries were from bulk tissue preparation (including 5000 ORESTES libraries), 329 were from cell lines, 37 were flow-sorted, 66 were microdissected, 5 were multiple preparations, and 1121 were from uncharacterized preparations. Excluding ORESTES libraries, 507 libraries were designated as "non-normalized" and 100 were designated "normalized" or "substracted" indicating the pretreatment of mRNA before cDNA library construction. A small number of libraries were derived from the same original sample. They were not considered separately. Library counts of ESTs rather than direct EST counts are used to provide semi-quantitative measurements of expression level, as EST counts in some cases reflect the prevalence of ESTs in one or a few particular libraries, and library counts provide better indications across different tissue types when both normalized and non-normalized libraries were analyzed. Such tissue information analyses are limited to those tissues with a sufficient number of libraries. The inclusion of normalized cDNA libraries allowed the examination of genes expressed at low levels.

To exclude possible genomic contamination in expressed sequences and other EST problems, only donor–acceptor concatenations or gene segments aligned with at least one mRNA or ESTs from at least two libraries were considered. The ESTs from "pooled tissue" or "uncharacterized tissue" were considered non-conforming in order to maintain the robustness of the results. In addition, 139,243 ESTs that had no library information were considered non-conforming in investigating tissue- or cancer-specific alternative splicing events, and were not considered in gene segment selection.

Simple probability scoring based on the binomial distribution was used to rank gene segments highly expressed in particular tissues. In the case of colon cancer (Table 5), the colon tumor library number, the colon library number, and the total library number for each gene segment were considered.

Supplementary data for this article are available on IDEAL (http://www.idealibrary.com).

ACKNOWLEDGMENTS

We thank Sarah Pollock, Raveh Gill-More, Avner Magen, Eyal Fink, Ariel Scolnicov, Guy Kol, Eran Halperin, Eitan Rubin, Avi Rosenberg, Yuval Cohen, Ohad Shoshany, David Lehavi, Alex Golubev, Tomer Zecharia, Gil Dogon, Dror Dotan, and Avishai Vaaknin (Compugen Ltd., Tel Aviv, Israel), who developed the LEADS platform.

RECEIVED FOR PUBLICATION MAY 2; ACCEPTED JULY 10, 2002.

REFERENCES

1. Adams, M. D., Soares, M. B., Kerlavage, A. R., Fields, C., and Venter, J. C. (1993). Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nat. Genet.* **4:** 373–380.
2. Boguski, M. S., and Schuler, G. D. (1995). ESTablishing a human transcript map. *Nat. Genet.* **10:** 369–371.
3. Audic, S., and Claverie, J. M. (1997). The significance of digital gene expression profiles. *Genome Res.* **7:** 986–995.
4. Huminiecki, L., and Bicknell, R. (2000). In silico cloning of novel endothelial-specific genes. *Genome Res.* **10:** 1796–1806.
5. Beaudoing, E., and Gautheret, D. (2001). Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. *Genome Res.* **11:** 1520–1526.
6. Brett, D., *et al.* (2001). A rapid bioinformatic method identifies novel genes with direct clinical relevance to colon cancer. *Oncogene* **20:** 4581–4585.
7. Schmitt, A. O., *et al.* (1999). Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. *Nucleic Acids Res.* **27:** 4251–4260.
8. Bortoluzzi, S., d'Alessi, F., Romualdi, C., and Danieli, G. A. (2000). The human adult skeletal muscle transcriptional profile reconstructed by a novel computational approach. *Genome Res.* **10:** 344–349.
9. Bortoluzzi, S., d'Alessi, F., and Danieli, G. A. (2000). A novel resource for the study of genes expressed in the adult human retina. *Invest. Ophthalmol. Vis. Sci.* **41:** 3305–3308.
10. Scheurle, D., *et al.* (2000). Cancer gene discovery using digital differential display. *Cancer Res.* **60:** 4037–4043.
11. Kawamoto, S., *et al.* (2000). BodyMap: a collection of 3′ ESTs for analysis of human gene expression information. *Genome Res.* **10:** 1817–1827.
12. International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
13. Sorek, R., and Amitai, M. (2001). Piecing together the significance of splicing. *Nat. Biotechnol.* **19:** 196.
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. Modrek, B., and Lee, C. (2001). A genomic view of alternative splicing. *Nat. Genet.* **30:** 13–19.
16. Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial analysis of gene expression. *Science* **270:** 484–487.
17. Caron, H., *et al.* (2001). The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291:** 1289–1292.
18. Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270:** 467–470.
19. Shoshan, A., *et al.* (2001). Designing oligo libraries taking alternative splicing into account. In *Proc. SPIE Microarrays: Optical Technologies and Informatics* (M. L. Bittner, Y. Chen, A. N. Dorsel, and E. D. Dougherty, Eds.), pp. 86–95. SPIE, Bellingham, WA.
20. Matloubian, M., David, A., Engel, S., Ryan, J. E., and Cyster, J. G. (2000). A transmembrane CXC chemokine is a ligand for HIV-coreceptor Bonzo. *Nat. Immunol.* **1:** 298–304.
21. David, A., *et al.* (2002). Unusual alternative splicing within the human kallikrein genes KLK2 and KLK3 gives rise to novel prostate-specific proteins. *J. Biol. Chem.* **277:** 18084–18090.
22. Sorek, R., Ast, G., and Graur, D. Alu-containing exons are alternatively spliced. *Genome Res.* (in press).
23. 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.
24. 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.
25. Web supplement. Detailed EST membership in LEADS contigs and sequence alignments are available at http://www.cgen.com/~han/Genomics_paper/supplement.htm.
26. Caballero, O. L., de Souza, S. J., Brentani, R. R., and Simpson, A. J. (2001). Alternative spliced transcripts as cancer markers. *Dis. Markers* **17:** 67–75.
27. Xuan, J. W., *et al.* (1995). Alternative splicing of PSP94 (prostatic secretory protein of 94 amino acids) mRNA in prostate tissue. *Oncogene* **11:** 1041–1047.
28. Garber, M. E., *et al.* (2001). Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl. Acad. Sci. USA* **98:** 13784–13789.
29. Nacht, M., *et al.* (2001). Molecular characteristics of non-small cell lung cancer. *Proc. Natl. Acad. Sci. USA* **98:** 15203–15208.
30. The Gene Ontology Consortium. (2000). Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25:** 25–29.
31. Xie, H., *et al.* (2002). Large scale protein annotation through Gene Ontology. *Genome Res.* **12:** 785–794.
32. Camargo, A. A., *et al.* (2001). The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. *Proc. Natl. Acad. Sci. USA* **98:** 12103–12108.