Genomics & Informatics Vol. 3(3) 80-85, September 2005

HExDB: Human EXon DataBase for Alternative Splicing Pattern Analysis

Junghwan Park[1], Minho Lee[1], Jong Bhak[2]*

[1]Department of BioSystems, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea
[2]National Genome Information Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 305-333, Korea

Abstract

HExDB is a database for analyzing exon and splicing pattern information in Homo sapiens. HExDB is useful for specific purposes: 1) to design primers for exon amplification from cDNA and 2) to understand the change of ORFs by alternative splicing. HExDB was constructed by integrating data from AltExtron which is the computationally predicted exon database, Ensemble cDNA annotation, and Affymetrix genome tile published recently. Although it may contain false positive data, HExDB is good starting point due to its sensitivity. At present, there are as many as 2,046,519 exons stored in the HExDB. We found that 16.8% of the exons in the database was constitutive exons and 83.1% were novel gene exons.

Keywords: HExDB, Exon database, Alternative splicing
Availability: HExDB is freely available at http://exonome.net
Supplementary Information: On the HExDB website.

Introduction

Alternative splicing is a versatile mechanism for producing a variety of transcripts and to regulate gene expressions for eukaryotes (Matlin *et al.*, 2005). An accurate splicing mechanism is critical for at least 15% of human genetic diseases that are caused by a splicing error (Cartegni *et al.*, 2002; Caceres *et al.*, 2002; Faustino *et al.*, 2003; Pagani *et al.*, 2004). Recent microarray data combined with ESTs suggest that 73% of human genes are alternatively spliced (Johnson *et al.*, 2003).

There are many databases of alternatively spliced genes. Broadly, they can be classified into two. One is based on experimental data while the other is based on computational predictions of the alternative splicing.

The experiment based database includes ASDB (Dralyuk *et al.*, 2000), AsMamDB (Ji *et al.*, 2001), Xpro (Gopalan *et al.*, 2004) and AEdb (Thanaraj *et al.*, 2004). The data they processed include bibliography in MEDLINE, sequence data from GenBank (Benson *et al.*, 2004) and SWISS-PROT (Bairoch *et al.*, 2004) database. The computational method based databases include AltExtron (Clark *et al.*, 2002), Asforms (Brett *et al.*, 2001), ASAP (Lee *et al.*, 2003) and TAP (Kan *et al.*, 2002). They determined splicing sites through an examination of alignments from EST and mRNA sequences. The computational approaches, however, can be error prone owing to a limited gene coverage and it does not have a good confidence measure (Thanaraj *et al.*, 2004). In 2004, the European Bioinformatics Institute (EBI) launched ASD, the Alternative Splicing Database, to combine these two approaches which is also a computation based database. Whatever the method, it is necessary for researchers to develop an efficient alternative splicing prediction method and databases. For a more efficient genome-wide disease research, a more sensitive way to find exons is needed. In other words, despite of possible false positives, detecting all the possible exons is an important starting point for alternative splicing research and a database construction.

Here, we present a new database that aims for finding maximum number of exons. It integrates many databases for exons such as transcriptional maps of ten human chromosomes published recently (Cheng *et al.*, 2005).

Methods

Integration of existing databases

To construct HExDB, AltExtron (Clark *et al.*, 2002) and Ensembl cDNA annotations (Hubbard *et al.*, 2005) are used to gather primary exon data. AltExtron is a computer- generated database by EBI. Ensembl annotation is primarily based on biological literature.

AltExtron contains specific information about individual exon. However, it does not contain the exact genomic coordinates. Therefore, we transformed (i.e., matched) the sequences to the genomic coordinates using BLAST sequence search algorithm (McGinnis and Madden, 2004) and a genetic database querying. AltExtron has a file format described in its database homepage, www.ebi.ac.uk/

*Corresponding author: E-mail j@bio.cc,
Tel +82- 42- 879-8500, Fax +82- 42- 879-8519
Accepted 20 July 2005

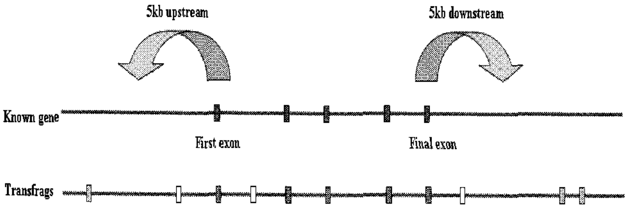


Fig. 1. Finding alternative splicing exons and novel gene exons. The first genome line is for known gene exons. The second genone line is for newly made transfrags.

Table 1. Statistics of exons from Ensembl database. NPE means Not-clearly Positioned Exons. 299,631 exons from Ensembl cDNA annotation was used to construct HExDB. About 30 thousand exons have unclear boundaries.

Chromosome Number	Number of Exons including NPE	Number of Exons Excluding NPE
1	28035	27627
2	19895	19798
3	15449	15396
4	10407	10383
5	11700	11689
6	41014	13648
7	13364	13326
8	9047	8985
9	10933	10904
10	11714	11708
11	15003	15003
12	14408	14405
13	5004	4967
14	8577	8577
15	9752	9573
16	11907	11887
17	16337	16000
18	4155	4155
19	15909	15840
20	7130	7130
21	2834	2834
22	6216	6193
X	9826	9710
Y	1015	1015
Total	299631	270753

asd/altextron/data/gene_data.html. We used the following three fields in its format GI, ACC, and AFETS.

Ensembl cDNA annotations contain many references. We filtered out much reference information to extract only the position information and the gene IDs exons belong to. The ensembl annotation is stored as EMBL file format. We used BioPerl Module Bio::SeqIO to parse annotated files. All the source codes are accessible from the supplementary material page of our database site.

Transcriptomic Map from Affymetrix

National Human Genome Research Institute launched the ENCODE project. It stands for ENCyclopedia of DNA Elements. As one of the results, Affymetrix inc. proposed 5bp resolution Transcriptome map of ten human chromosomes constructed by a microarray approach (Cheng *et al.*, 2005).

Profiles generated by the microarrays were filtered at a threshold to create exon regions. Cheng *et al.* chose the threshold level for each chip so that 94.8% of the exons were already known in the exon region assignment. Cheng *et al.* used the term 'transfrags' to denote these regions. We followed the same procedure to integrate the transfrag data with other databases collected.

Cheng *et al.* searched for transfrags on ten human chromosomes for eight kinds of cancer cell lines. Due to the imperfection of the threshold-picking algorithm, some transfrag boundaries were not clear and redundant. The redundant transfrags were reduced by clustering algorithms.

Finding alternative splicing exons and novel gene exons

Introns and both of the 5kb regions, upstream- and downstream-containing genes were obtained from GoldenPath, which is based on known gene information under the April 2003 version of human genome (NCBI v. 33). The types of HExDB exon were classified according

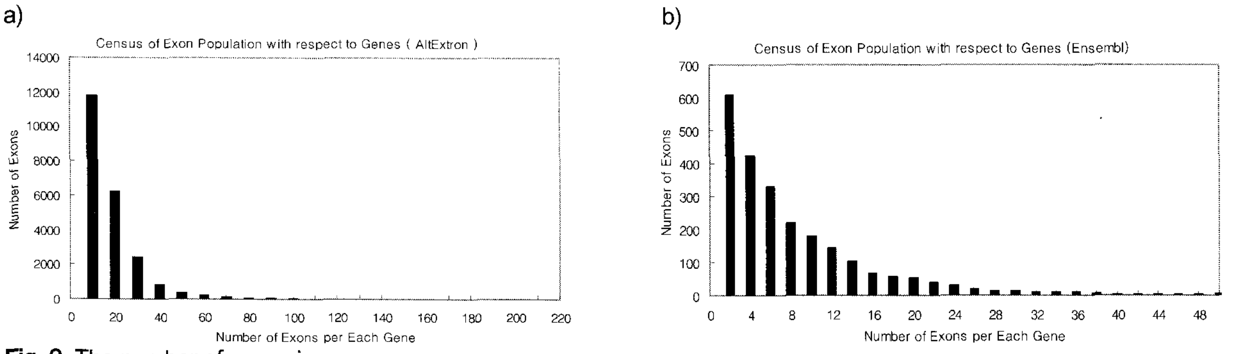


Fig. 2. The number of exons in genes.

These histograms describe the number of exons per each gene in the data of (a) AltExtron and (b) Ensembl that are integrated into HExDB. The mean values of number of exons are 7.3 and 13.0 respectively.

Table 2. The number of transfrags on the ten chromosomes. This table show the number of exons based on transcriptome map (Cheng *et al.*, 2005). Exons in these ten chromosomes are only available by this method. The reduced number was calculated by count overlapped exons as single exon.

Chromosome Number	Raw Number of Transfrags	Reduced Number of Transfrags
13	184321	57491
14	167201	50139
19	143095	35895
20	161268	46723
21	73209	21651
22	101486	27768
6	327676	103435
7	340528	98553
X	199736	53935
Y	18515	3584
Total	1717035	499174

to the position of exons within intragenic regions and intergenic regions (Fig. 1).

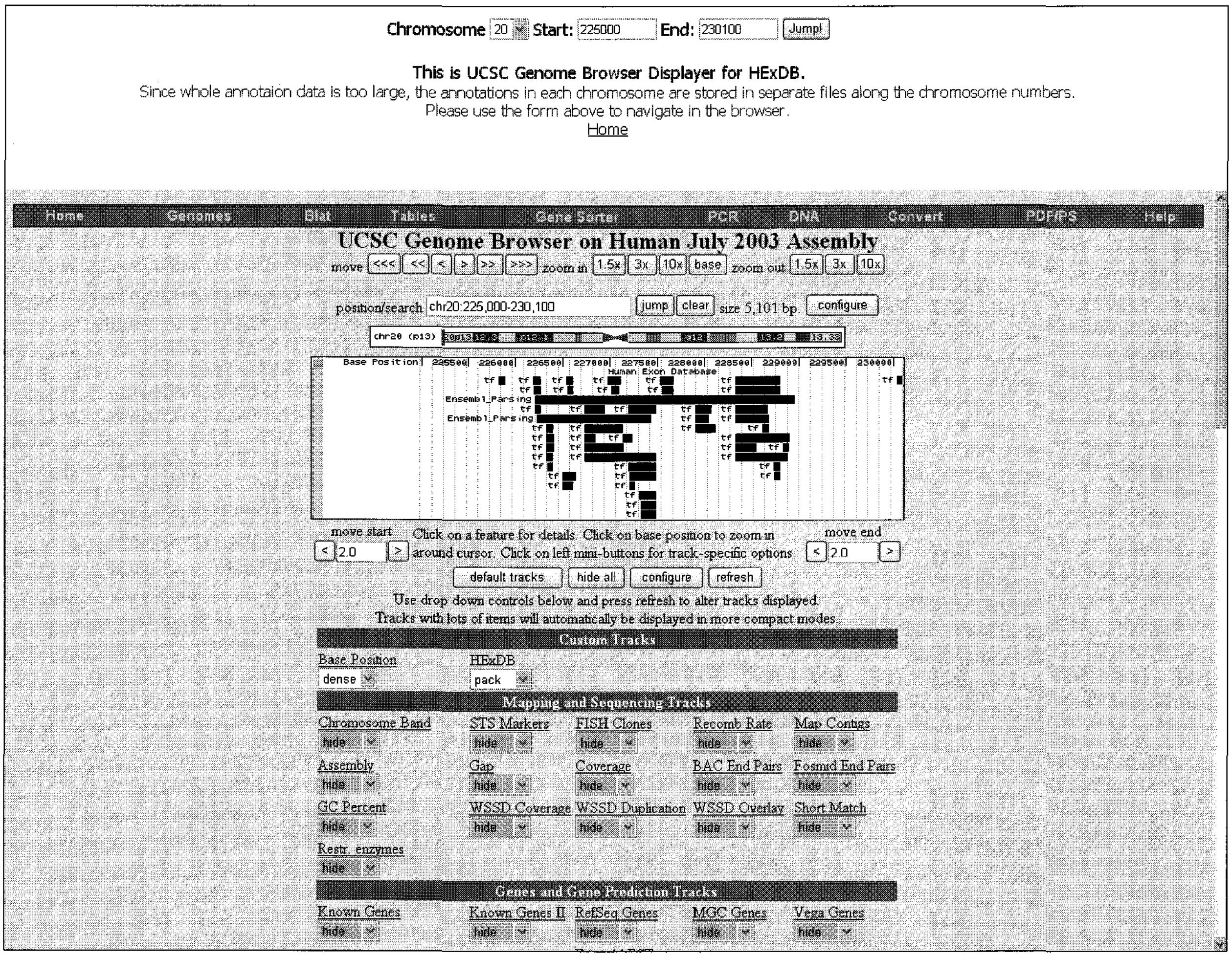
Results and Discussion

The number of Exons in HExDB

29,853 exons were integrated into HExDB from AltExtron database. They belonged to 4,113 genes. The average of number of exons was 7.3 per each gene. The distribution of exons within each gene is shown in Fig. 2(a).

299,646 exons were integrated into HExDB from Ensembl database annotation information. Most of them were well positioned on chromosomes. Table 1 and Fig. 2(b) show the statistics of them for each chromosome. However, there were 1572 exons that did not belong to proper chromosome contigs.

By Cheng *et al.*(2005)'s Transcriptomic map analysis,



Chromosome 20 Start: 225000 End: 230100

This is UCSC Genome Browser Displayer for HExDB.

Since whole annotaion data is too large, the annotations in each chromosome are stored in separate files along the chromosome numbers.

Please use the form above to navigate in the browser.

UCSC Genome Browser on Human July 2003 Assembly

position/search chr20:225,000-230,100 size 5,101 bp.

move start 2.0 move end 2.0

Click on a feature for details. Click on base position to zoom in around cursor. Click on left mini-buttons for track-specific options.

Use drop down controls below and press refresh to alter tracks displayed.

Tracks with lots of items will automatically be displayed in more compact modes.

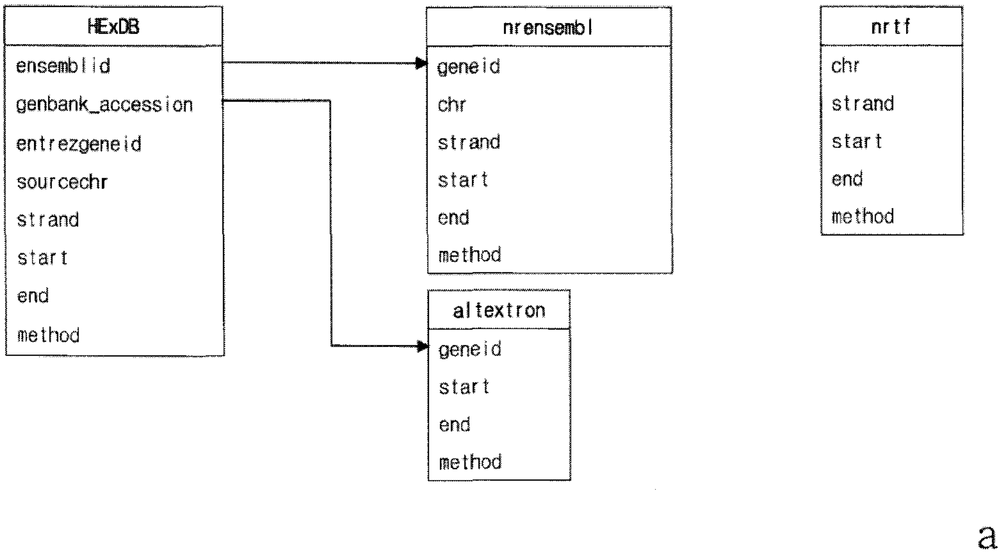
Custom Tracks: Base Position dense; HExDB pack

Mapping and Sequencing Tracks: Chromosome Band hide; STS Markers hide; FISH Clones hide; Recomb Rate hide; Map Contigs hide; Assembly hide; Gap hide; Coverage hide; BAC End Pairs hide; Fosmid End Pairs hide; GC Percent hide; WSSD Coverage hide; WSSD Duplication hide; WSSD Overlay hide; Short Match hide; Restr. enzymes hide

Genes and Gene Prediction Tracks: Known Genes hide; Known Genes II hide; RefSeq Genes hide; MGC Genes hide; Vega Genes hide

Fig. 3. HExDB genome browser.

This browser is constructed by using UCSC genome browser and is similar with it. The top frame should be used to navigate the genome instead of UCSC browser itself because a custom-track of a certain chromosome is separated from that of the others due to the huge size of HExDB. Where the data is integrated from is annotated in the browser.



HExDB

ensemblid	genbank_accession	entrezgeneid	sourcechr	strand	start	end	method
NULL	AF187320	NULL	NULL	plus	10614	10716	AltExtron_Parsing
NULL	J02843	NULL	NULL	plus	12838	12979	AltExtron_Parsing
...	...	...	...	...	...	...	...
NULL	NULL	NULL	21	plus	13960004	13960074	tf
NULL	NULL	NULL	20	plus	48851523	48851597	tf
...	...	...	...	...	...	...	...
ENSG00000108342	NULL	NULL	17	plus	35425442	35425624	Ensembl Parsing
ENSG00000118308	NULL	NULL	12	minus	25222607	25222716	Ensembl Parsing
...	...	...	...	...	...	...	...

altextron

Geneid	start	end	method
A06939	808	908	AltExtron Parsing
A06939	1035	1138	AltExtron Parsing
A06939	1231	1348	AltExtron Parsing
...	...	...	...

nrtf

chr	strand	start	end	method
6	plus	144665184	144665325	tf
6	plus	40374747	40374828	tf
14	plus	93482470	93482526	tf
...	...	...	...	...

nrensembl

geneid	chr	strand	start	end	method
ENSG00000176395	19	minus	49716348	49716626	Ensembl Parsing
ENSG00000105982	7	plus	155950354	155950422	Ensembl Parsing
ENSG00000162517	1	minus	31779535	31779558	Ensembl Parsing
...	...	...	...	...	...

b

Fig. 4. HExDB schema.

These figures show detailed information about the MySQL database of HExDB. (a) This simple diagram shows the data inclusion relationships among tables. Since table 'nrtf' has no ID system, table 'HExDB' and 'nrtf' have no common key field. (b) These tables show some example entries. 'method' field cotains the information about which method is used to get the corresponding row. If the method is 'AltExtron Parsing', the only supported ID system is Genbank accession number. For 'Ensembl Parsing', ensembl id is supported.

1,717,035 exons were found. However, due to unclear boundaries, redundant entries were assigned to the same exons. To reduce this redundancy, we treated overlapped exons as single exon. The final exon number became 499,174. It means the true total exon number is perhaps between the two numbers. Therefore, a further research on accurately reducing redundancy is necessary to get the exact number of human exons. The numbers of transfrags on the ten chromosomes are shown in Table 2.

The total number of exons in our HExDB combining all the above source databases was 2,046,519. This figure contains some redundancy from the integration of the source databases (see Table 2). However, we suggest that this is about the upper limit of exons in the human genome. In contrast to the average number of exons per each gene in AltExtron and Ensembl, if we suppose human genome contains around 30,000 genes, 68.2, the number of exons per each gene in HExDB is far larger than that in others.

Exon data of HExDB can be accessed by using interactive genome browser as shown in Fig. 3 or by downloading MySQL database file (see Fig. 4). The genome browser was constructed by using custom-track feature of the UCSC genome browser.

Conclusion

The purpose of HExDB is to list all the possible exons that can be predicted and annotated by current technology. There can be some redundancy due to this. However, it can give us the estimation and the most number of exon data possible. To do research on alternative splicing, to list all possible exons are needed. The contribution of HExDB, therefore, is to provide biologists useful tool for research on alternative splicing as well as the most comprehensive exon information. We found that there were about two million exons in the human genome from the existing exon data. This number is by no means definite or accurate and will be adjusted in the future. However, we predict that it is close to the upper limit of the total number of human exons. To our surprise, there were a great number of unknown exons. This indicates that the actual number of genes in the genome can be as high as 100,000 due to the high number of exons.

Acknowledgements

This work was supported by CHUNG Moon Soul Center for BioInformaion and BioElectronics (CMSC). Computational resources are provided in part by the IBM SUR Grant. JB was supported by Biogreen21 program of RDA and R01-2004-000-10172-0 grant of KOSEF.

References

Bairoch, A., Boeckmann, B., Ferro, S., and Gasteiger, E. (2004). Swiss-Prot: juggling between evolution and stability. *Brief Bioinform.* 5, 39-55

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. (2004). GenBank update. *Nucleic Acids Res.* 32, D23-D26.

Brett, D., Pospisil, H., Valcarcel, J., Reich, J., and Bork, P. (2001). Alternative splicing and genome complexity. *Nat. Genet.* 30, 29-30.

Caceres, J.F. and Kornblihtt, A.R. (2002). Alternative splicing : multiple control mechanisms and involvement in human disease. *Trends Genet.* 18, 186-193.

Cartegni, L., Chew, S.L., and Krainer, A.R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat. Rev. Genet.* 3, 285-298.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., Sementchenko, V., Piccolboni, A., Bekiranov, S., Bailey, D.K., Ganesh, M., Ghosh, S., Bell, I., Gerhard, D.S., and Gingeras, T.R. (2005). Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* 308, 1149-1154.

Clack, F. and Thanaraj, T.A. (2002). Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum. Mol. Genet.* 11, 451-464.

Dralyuk, I., Brudno, M., Gelfand, M.S., Zorn, M., and Dubchak, I. (2002). ASDB : database of alternatively spliced genes. *Nucleic Acids Res.* 28, 296-297.

Faustino, N.A. and Cooper. T.A. (2003). Pre-mRNA splicing and human disease. *Genes Dev.* 17, 419-437.

Gopalan, V., Tan, T.W., Lee, B.T., and Ranganathan, S. (2004). Xpro : database of eukaryotic protein-encoding genes. *Nucleic Acids Res.* 32, D59-D63.

Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., *et al.* (2005). Ensembl 2005, *Nucleic Acids Res.* 33, D447-D453.

Ji, H., Zhou, Q., Wen, F., Xia, H., and Li, Y. (2001). AsMamDB : an alternative splice database of mammals. *Nucleic Acids Res.* 29, 260-263.

Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., and Shoemaker, D.D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* 302, 2141-2144.

Kan, Z., States, D., and Gish, W. (2002). Selecting for functional alternative splices in ESTs. *Genome Res.* 12, 1837-1845.

Lee, C., Atanelov, L., Modrek, B., and Xing, Y. (2003). ASAP: the Alternative Splicing Annotation Project. *Nucleic Acids Res.* 31, 101-105.

McGinnis, S. and Madden T.L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* 32, W20-W25.

Matlin, A.J., Clark, F., and Smith, C.W. (2005). Understanding alternative splicing : towards a cellular code. *Nat. Rev. Mol. Cell Biol.* 6, 386-398.

Pagani, F. and Baralle, F.E. (2004). Genomic variants in exons and introns: identifying the splicing spoilers. *Nat. Rev. Genet.* 30, 13-19.

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, D64-D69.