Can HMMgene Net a Gene Too? : Analysis of Alternative Splicing/Gene Prediction Algorithms

Amanda Mikels

BIOC 218

The era of genome sequencing is upon us. Everyday genetic information is entered into numerous worldwide databases with the hopes of better understanding various organisms' genetic make-ups. However, although the draft sequence of the human genome is estimated to be ~97% complete, the sequence itself remains poorly annotated. Gene annotation based on spurious EST data has necessitated the creation several gene prediction algorithms designed to uncover potential coding regions and solve the mysteries of the human genome.

One can see why annotation with a high degree of accuracy is so difficult to obtain when one considers various factors complicating gene prediction methods. Firstly, gene transfer mechanisms often introduce extra copies of genes into genomes, which then diverge through evolution. Hence, distinguishing broken *pseudo-genes* from working genes is a difficult problem. In addition, sequencing errors can hide proper donor/acceptor sites and cause apparent frame shifts. Exons can be separated by several thousand bases and assembled in multiple ways through alternative splicing. And finally, genes can overlap each other, appear in different reading frames and on different strands (16). Thus, the creation of a gene prediction algorithm that is able to take all of these factors into account is not trivial.

While current estimates predict that there are only ~30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly – these estimates exclude the mounting evidence that a great number (~60%) of human genes are alternatively spliced (1). These alternative splicing events, which have the potential to create thousands of isoforms from a single genetic locus as is the case for the *Drosphila* DSCAM gene, remain difficult to detect efficiently and accurately as alternative splicing often occurs in a cell/tissue or developmental stage specific manner. Additionally, questions regarding the underlying molecular mechanisms governing alternative splicing such as what are the primary splicing determinants and what is the effect of gene structure ie genetic frame and G+C content, on alternative splicing remain unanswered. Thus, in an attempt to determine the accuracy ie sensitivity and

specificity, of various alternative splicing/gene prediction programs, two such programs (HMMgene and NetGene2) were run against a genomic sequence containing a gene known *a priori* to be alternatively spliced into at least 6 isoforms: Hypoxia Inducible Factor 3 α (hHIF-3á).

While the gene itself remains incorrectly annotated today in the *Ensembl* genome browser, much new information about the various splice isoforms of HIF3α has been recently elucidated by Maynard et al (3). The hHIF-3á gene consists of 17 exons, which span about 46 Kb on chromosome19q13.2 (see Figure 1). Three unique exons (exons 1a, b and c) are thought to contain the transcription start sites for the six splicing variants. hHIF-3á1 begins at exon 1c and ends at exon 16. A dominant-negative regulator of mHIF-1 called mIPAS was recently identified as an alternatively spliced variant of mHIF-3á. hIPAS starts at exon 1a and ends at exon 13a without exon 1b and c and is now referred to as hHIF-3á2. cDNA entries in the GenBank with Accession Nos. AK021653 and NM_022462 were also found to be derived from the hHIF-3á locus and correspond to hHIF-3á3, which begins exon 1b and ends at exon 17.

Another full-length cDNA (Accession No. BC026308) that begins at exon 1a and ends within the imperfect intron 8 was derived from the same locus as other hHIF-3á splice variants. This isoform's intron 7 was not spliced. This cDNA which encodes for a 363 amino acid protein was termed hHIF-3á4. Finally, two additional GenBank cDNA entries were found to be splicing variants of hHIF-3á. Maynard et al named them hHIF-3á5 (Accession No.: NM_152796) and hHIF-3á6 (Accession No.: AK024095). Both isoforms start at exon 1b and lack exon 3. hHIF-3á5 contains a short exon 14c and ends at exon 15. hHIF-3á6, like hHIF-3á4 contains intron 7 and ends at intron 8. With the apparent wealth of knowledge regarding the hHIF-3á genetic locus mentioned above, this gene seemed ideally suited to be the test sequence on which to run the various alternative splicing/gene prediction algorithms.

As compared to several other currently available gene prediction programs, algorithms based on Hidden Markov Models (HMMs) such as GENSCAN and HMMgene have been found to have the highest degree of

prediction accuracy (see Figure 2) when tested against the Burset and Guigo set of data sequences (4). To calculate the accuracy statistics shown in Figure 2 at the nucleotide level, one classifies each nucleotide of a test sequence as predicted positive (**PP**) if it is in a predicted coding region, predicted negative (**PN**) otherwise, and also as actual positive (**AP**) or actual negative (**AN**) according to the sequence annotation. These assignments are then compared to calculate the number of true positives (**TP**), false positives (**FP**), true negatives (**TN**) and false negatives (**FN**). Accuracy is then measured by:

Sensitivity, **Sn** = TP / AP

Specificity, **Sp** = TP / PP

and Approximate Correlation, **AC**, defined as:

AC = ((TP/(TP+FN)) + (TP/(TP+FP)) + (TN/(TN+FP)) + (TN/(TN+FN))) / 2 - 1

At the exon level, predicted exons (**PE**) are compared to annotated exons (**AE**). True exons (**TE**) is the number of predicted exons which are exactly identical to an annotated exon (i.e. both endpoints correct). Accuracy is again measured by:

Sensitivity, **Sn** = TE / AE

Specificity, **Sp** = TE / PE

The average of **Sn** and **Sp** is typically used as an overall measure of accuracy at the exon level in lieu of a correlation measure. Two additional accuracy measures are also calculated at the exon level: Missing Exons (**ME**), the fraction of annotated exons not overlapped by any predicted exon; and Wrong Exons (**WE**), the fraction of predicted exons not overlapped by any true exon. Accuracy measures for a set of sequences are calculated by averaging the values obtained for each sequence separately, the average being taken over all sequences for which the measure is defined (17).

Because according to these measurements the HMMgene algorithm appeared to be fairly accurate, and because I did not want to run the same algorithm that the *Ensembl* human genome browser uses for its gene prediction (GENSCAN), the ~46Kb genomic sequence of HIF3α locus corresponding to bases 47191200 - 47237149 on chromosome #19 was assessed by the

HMMgene algorithm available at http://www.cbs.dtu.dk/services/ HMMgene/. The results of this scan are presented in Figure 4.

When one compares the exon/intron boundary data based on current EST data compiled in Figure 1b with the HMMgene data presented in Figure 4, one notices that the HMMgene predictions were fairly accurate. HMMgene in its "bestparse" sequence detected 13/16 true positive acceptor sites and 13/16 true positive acceptor sites. The next five suboptimal exon/intron predictions did little better at predicting other alternative true positive donor and acceptor sites and hence the "bestparse" prediction truly was the best predictor of exon structure. The HMMgene program seemed to perform better when predicting internal exons as the first exon's donor site and the last two exons' donor and acceptor sites were missed completely. However, it should be noted that the algorithm did produce false positive donor and acceptor sites for the first and last exons that were in the near vicinity of the true positive donor and acceptor sites. Thus, HMMgene was able to predict the majority of the donor and acceptor splice sites of HIF3α with a relatively high degree of accuracy.

The HIF3α genomic sequence was then analyzed via the NetGene2 algorithm; the results are listed in Figure 5. At a 95% confidence level (H), the NetGene2 algorithm was able to pick up 8/9 donor splice sites and 5/5 acceptor splice sites. When the confidence threshold was lowered to include only those sequences with a >80% probability, NetGene2 found 10/14 and 9/11 donor and acceptor splice sites respectively. Finally, when examining all sequences produced by the algorithm, it was determined that 14/72 and 14/157 donor and acceptor splice sites respectively were found.

There are several items to note from the NetGene2 output data presented in Figure 5. First, while the NetGene2 algorithm at low confidence levels yielded more true positive sequences overall (14 as opposed to the 13 produced by HMMgene), many more false positives were present in the in the data set. Also, like HMMgene, the NetGene2 algorithm was better at predicting the internal HIF3α exons than the flanking first and last exons. Thus, it appears that while the NetGene2 algorithm output at lower confidence levels had a higher degree of

specificity **(Sp),** the HMMgene algorithm was more sensitive **(Sn).** At higher confidence levels, NetGene2 was more sensitive than HMMgene but far less specific.

When considering the output data produced by the two programs, it is important to understand the differences between the two algorithms. The HMMgene algorithm utilizes a standard hidden Markov model (HMM) with coding regions being modeled by 4th order inhomogeneous Markov chains. It is a program for designed for the prediction of genes in anonymous DNA. All predictions made by the program have associated probabilities that reflect how confident the model is in the predictions. Apart from reporting the best prediction, the HMMgene program reports as an added feature the N best gene predictions for a sequence. This option is useful if the there are several equally likely gene structures and may even indicate alternative splicing. The output produced from running the HMMgene program thus is a prediction of partial or complete genes in the sequences.

In addition, the HMMgene algorithm uses conditional maximum likelihood criterion that allows for the maximization of correct predictions and is trained using a sequence set derived from human sequences taken from GenBank. The main difference between this type of HMM ab initio program versus other HMM ab initio programs such as Genscan is that HMMgene uses standard HMM while Genscan uses generalized HMM. A major advantage of the HMMgene program is that it is an integrated model (15,22).

In contrast to HMMgene, the NetGene2 algorithm uses a set of neural networks combined with a rule based system to predict intron splice sites (20). The networks are of the multi-layer error-back propagation type that are fully connected and have three layers: an input layer, a hidden layer, and an output layer. A two step prediction scheme, where a global prediction of the nucleotide coding potential regulates a cutoff level for a local prediction of splice sites, is refined by rules based on splice site confidence values, prediction scores, coding context, and distances between potential splice sites. In this approach, the predictions of various splice sites mutually affect each other in a non-local

manner. Thus, the combined approach in theory should reduce the large amount of false positive splice sites normally associated with splice site prediction.

Though HMMgene and NetGene2 run very different algorithms, both programs came across similar difficulties when predicting the structures of the exons contained in the HIF3α locus. While both programs were able to find several true positive donor and acceptor splice sites, neither program was able to find all of the true sites when the output data is compared to the Genbank sequence data compiled in Figure 1b. Both HMMgene and NetGene2 had problems in predicting the donor and acceptor splice sites in non-internal exons and both programs were unable to correctly predict relatively short or long exons (ex exon 1b with 46bps and exon 15 with 358bps).

Because defined sets of sequences train both programs and because these training sets are the basis for the algorithms' predictive value, one way in which to improve the sensitivity and specificity of the two algorithms would be to add various new sequences to the training sets. The HMMgene algorithm uses a training set of sequences from which it makes its predictive model. Similarly, the neural network used by NetGene2 is created from various predetermined "accurate" sequences acquired from Genbank. Thus, if one wanted to improve these two programs' ability to predict relatively short or long exons for example, it would seem that one would simply have to include more sequences that contain short and long exons into the training sets. The programs would then be able to predict the donor and acceptor splice sites of small and large algorithms with greater accuracy.

One other additional improvement to these programs would be to add a program such as *SpliceProximalCheck* as a front-end tool. *SpliceProximalCheck* is based on a decision tree approach and works under a set of discrete rules that characterize the proximal false sites as opposed to real splice sites. Thus, it is trained to discriminate true splice sites from false splice sites located in the vicinity of true splice sites. Because several of the false positive splice sites predicted for HIF3α by HMMgene and NetGene2 were indeed in close proximity

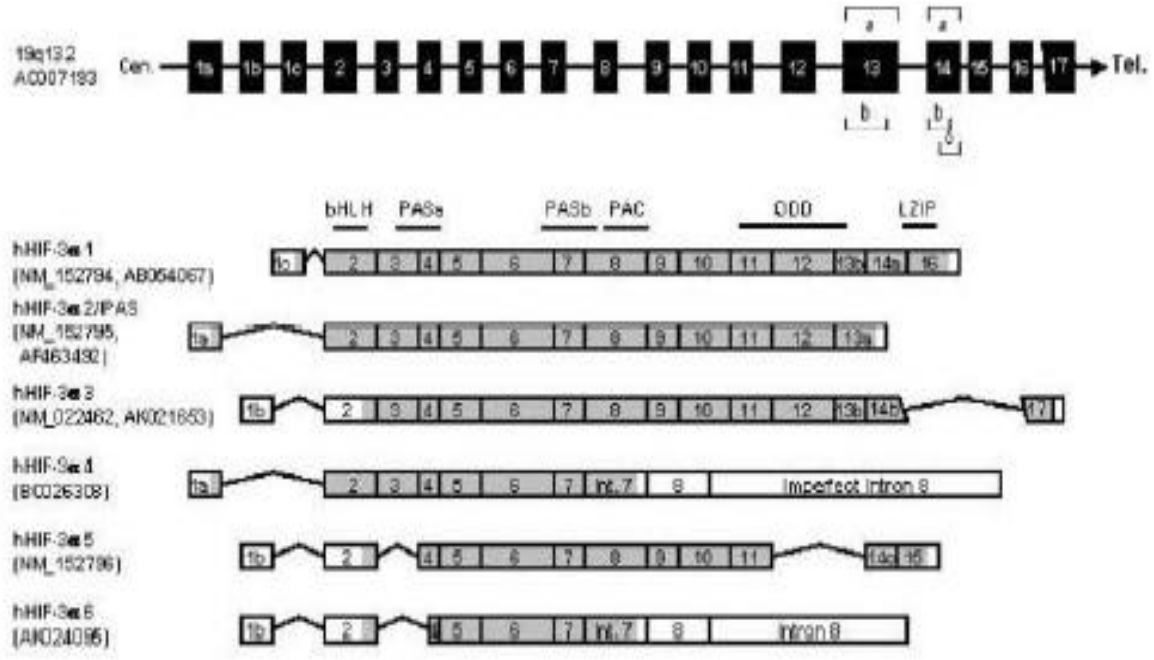
to true positive splice sites, using a front-end tool to filter out false splice sites could only increase the accuracy of the programs.

As the rate of novel EST discovery has fallen in recent years, evidence suggests that only about 80% of human genes are represented in the over two million ESTs present in public databases (13). Thus, it is glaringly obvious that methods less biased by expression level, such as gene prediction algorithms, are needed to complete the annotation of the human genome. Two alternative splicing/gene prediction programs, HMMgene and NetGene2, were tested against a genetic locus known to be alternatively spliced: HIF3α. While both programs performed well, neither was able to tease out all true positive donor and acceptor splice sites and both presented false positive data. Thus, it is clear that methods used for gene prediction, though more accurate than ever before, need to be improved before the mysteries of the human genetic code can be fully uncovered.

Figures:

Figure 1: HIF3a Locus Data

1a. HIF3a Splice Isoforms (Maynard et al, 2003).



1b: HIF3a Splice Variant Exon Boundaries.

Gene/ Isoform	Exon1a	Exon1b	Exon1c	Exon2	Exon3	Exon4	Exon5	Exon6	Exon7
HIF 3 alpha 1			6578-6631	6841-7032	8188-8334	11164-11249	11606-11719	12094-12303	15104-15211
HIF 3 alpha 2 /NM152795	0-45			6841-7032	8188-8334	11164-11249	11606-11719	12094-12303	15104-15211
HIF 3 alpha 3		1341-1458		6841-7032	8188-8334	11164-11249	11606-11719	12094-12303	15104-15211
HIF 3 alpha 5 /NM152796		1339-1458		6841-7032		11164-11249	11606-11719	12094-12303	
HIF 3 alpha 6/ ENST00244303		1375-1458				11218-11249	11606-11719	12094-12303	15104-15189
	Exon8	Exon9	Exon10	Exon11	Exon12	Exon13	Exon14	Exon15	
	15449-15597	23386-23505	24719-24910	28478-28583	32150-32422	34099-34217	37848-37930	42460-43246	
	15449-15597	23386-23505	24719-24910	28478-28583	32150-32422	34099-34457			
	15449-15597	23386-23505	24719-24910	28478-28583	32150-32422	34099-34217	37848-37930	45678-45950	
	15449-15597	23386-23505	24719-24910	28478-28583			37890-37930	41170-41470	
	15282-15423								

Figure 2: Various Programs' Accuracy Data.

2a. Accuracy statistics taken from Table 1 of Burset, M. & Guigó, R. (1996) and Genscan data where Sn=sensitivity , Sp=specificity, and AC =approx. correlation.

	Accuracy per nucleotide			Accuracy per exon				
Method	**Sn**	**Sp**	**AC**	**Sn**	**Sp**	**(Sn+Sp)/2**	**ME**	**WE**
GENSCAN	**0.93**	**0.93**	**0.91**	**0.78**	**0.81**	**0.80**	**0.09**	**0.05**
FGENEH	**0.77**	**0.85**	**0.78**	**0.61**	**0.61**	**0.61**	**0.15**	**0.11**
GeneID	**0.63**	**0.81**	**0.67**	**0.44**	**0.45**	**0.45**	**0.28**	**0.24**
GeneParser2	**0.66**	**0.79**	**0.66**	**0.35**	**0.39**	**0.37**	**0.29**	**0.17**
GenLang	**0.72**	**0.75**	**0.69**	**0.50**	**0.49**	**0.50**	**0.21**	**0.21**
GRAILII	**0.72**	**0.84**	**0.75**	**0.36**	**0.41**	**0.38**	**0.25**	**0.10**
SORFIND	**0.71**	**0.85**	**0.73**	**0.42**	**0.47**	**0.45**	**0.24**	**0.14**
Xpound	**0.61**	**0.82**	**0.68**	**0.15**	**0.17**	**0.16**	**0.32**	0.13

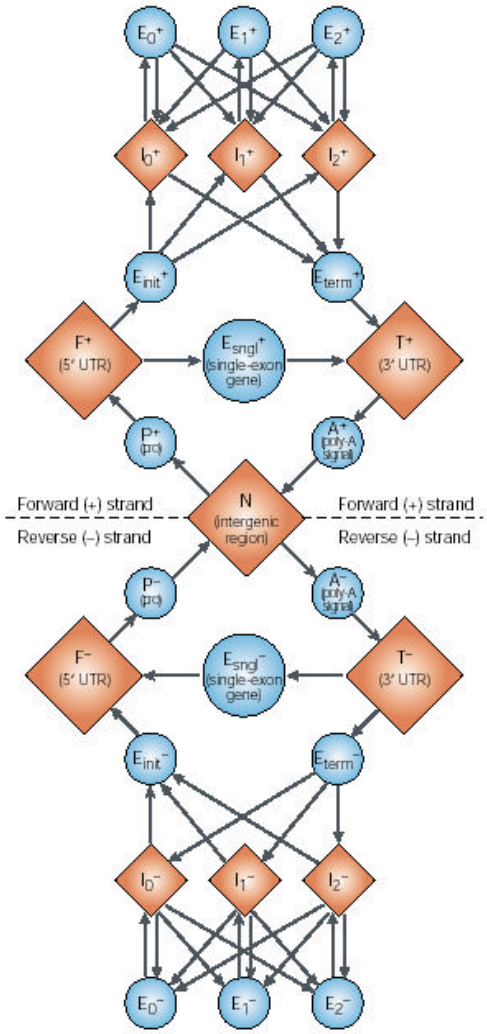
2b. Program accuracy statistics taken from Sanja Rogic's Table Page at: http://www.cs.ubc.ca/~rogic/evaluation/tablesgen.html#basic.

Programs	***# of seq***	***Nucleotide accuracy***				***Exon accuracy***							
		Sn	***Sp***	***AC***	***CC***	***ESn***	***ESp***	***(ESn+ESp)/2***	***ME***	***WE***	***PCa***	***PCp***	***OL***
FGENES	195(5)	0.86	0.88	0.84	0.83	0.67	0.67	0.69	0.12	0.09	0.20	0.17	0.02
GeneMark	195(0)	0.87	0.89	0.84	0.83	0.53	0.54	0.54	0.13	0.11	0.29	0.27	0.09
Genie	195(15)	0.91	0.90	0.89	0.88	0.71	0.70	0.71	0.19	0.11	0.15	0.15	0.02
Genscan	195(3)	0.95	0.90	0.91	0.91	0.70	0.70	0.71	0.08	0.09	0.21	0.19	0.02
HMMgene	195(5)	0.93	0.93	0.91	0.91	0.76	0.77	0.76	0.12	0.07	0.14	0.14	0.02
Morgan	127(0)	0.75	0.74	0.70	0.69	0.46	0.41	0.43	0.20	0.28	0.28	0.25	0.07
MZEF	119(8)	0.70	0.73	0.68	0.66	0.58	0.59	0.59	0.32	0.23	0.08	0.16	0.01

Table 1: Nucleotide and exon level accuracy - # of sequences - number of sequences effectively analyzed by each program; in parentheses is the number of sequences where the absence of gene was predicted; Sn -nucleotide level sensitivity; Sp - nucleotide level specificity, AC - approximate correlation; CC - correlation coefficient; ESn - exon level sensitivity; ESp - exon level specificity; ME - missed exons; WE - wrong exons; PCa - proportion of real exons that were partially predicted (only one exon boundary correct); PCp - proportion of predicted exons that

were only partially correct; OL - proportion of predicted exons that overlap an actual exon. AC and (ESn+ESp)/2 are given with standard deviation.

Figure 3: Representative Hidden Markov Model used by HMMGene
Taken from (19)



A hidden Markov model explicitly models the probabilities for the transition from one part of a gene to another. In this model, used by the HMMGene algorithm, each circle or diamond represents a functional unit in the gene. For example, Einit– is the initial exon and Eterm– is the last.The arrows represent the probability of a transition from one part of a gene to another. The algorithm is ‘trained’ by running a set of known genes through the model and adjusting the weights of each transition to reflect realistic transition probabilities. Thereafter, test sequence data can be run through the model one base position at a time, and the model will read out the probability of a gene being present at that position. The states that occur below the dashed line

correspond to a gene in the reversed strand, and thus are symmetric with those above the line. E, exon; I, intron; UTR, untranslated region; pro, promoter.

Figure 4: HMMGene Prediction Data:

Columns

1. Sequence identifier
2. Program name
3. Prediction (see table below for the meaning).
4. Beginning
5. End
6. Score between 0 and 1
7. Strand: $+$ for direct and $-$ for complementary
8. Frame (for exons it is the position of the donor in the frame)
9. Group to which prediction belong. If several CDS's are found they will be called cds_1, cds_2, etc. `bestparse:' is there because alternative predictions will also be available (see below).

HMMgene directed strand output:

```
# SEQ: Homo 45950 (+) A:11013 C:11855 G:11183 T:11899
Homo    HMMgene1.1a     firstex 21      46      0.526   +       2
        bestparse:cds_1
Homo    HMMgene1.1a     exon_1  6842    7032    0.992   +       1
        bestparse:cds_1
Homo    HMMgene1.1a     exon_2  8189    8334    0.883   +       0
        bestparse:cds_1
Homo    HMMgene1.1a     exon_3  11165   11249   0.998   +       1
        bestparse:cds_1
Homo    HMMgene1.1a     exon_4  11607   11719   0.999   +       0
        bestparse:cds_1
Homo    HMMgene1.1a     exon_5  12095   12303   0.991   +       2
        bestparse:cds_1
Homo    HMMgene1.1a     exon_6  15105   15232   0.736   +       1
        bestparse:cds_1
Homo    HMMgene1.1a     exon_7  15450   15597   0.996   +       2
        bestparse:cds_1
Homo    HMMgene1.1a     exon_8  23387   23505   0.997   +       1
        bestparse:cds_1
Homo    HMMgene1.1a     exon_9  24720   24910   0.975   +       0
        bestparse:cds_1
Homo    HMMgene1.1a     exon_10 28479   28583   0.999   +       0
        bestparse:cds_1
Homo    HMMgene1.1a     exon_11 32151   32422   0.919   +       2
        bestparse:cds_1
```

```
Homo    HMMgene1.1a     exon_12 34100   34151   0.507   +       0
        bestparse:cds_1
Homo    HMMgene1.1a     exon_13 37849   37930   0.700   +       1
        bestparse:cds_1
Homo    HMMgene1.1a     exon_14 41770   41886   0.655   +       1
        bestparse:cds_1
Homo    HMMgene1.1a     lastex  42469   42566   0.587   +       0
        bestparse:cds_1
Homo    HMMgene1.1a     CDS     21      42566   0.044   +       .
        bestparse:cds_1
Homo    HMMgene1.1a     firstex 21      46      0.526   +       2
        subopt_1:cds_1
Homo    HMMgene1.1a     exon_1  6842    7032    0.992   +       1
        subopt_1:cds_1
Homo    HMMgene1.1a     exon_2  8189    8334    0.883   +       0
        subopt_1:cds_1
Homo    HMMgene1.1a     exon_3  11165   11249   0.998   +       1
        subopt_1:cds_1
Homo    HMMgene1.1a     exon_4  11607   11719   0.999   +       0
        subopt_1:cds_1
Homo    HMMgene1.1a     exon_5  12095   12303   0.991   +       2
        subopt_1:cds_1
Homo    HMMgene1.1a     exon_6  15105   15232   0.736   +       1
        subopt_1:cds_1
Homo    HMMgene1.1a     exon_7  15450   15597   0.996   +       2
        subopt_1:cds_1
Homo    HMMgene1.1a     exon_8  23387   23505   0.997   +       1
        subopt_1:cds_1
Homo    HMMgene1.1a     exon_9  24720   24910   0.975   +       0
        subopt_1:cds_1
Homo    HMMgene1.1a     exon_10 28479   28583   0.999   +       0
        subopt_1:cds_1
Homo    HMMgene1.1a     exon_11 32151   32422   0.919   +       2
        subopt_1:cds_1
Homo    HMMgene1.1a     exon_12 34100   34217   0.330   +       0
        subopt_1:cds_1
Homo    HMMgene1.1a     exon_13 37849   37930   0.700   +       1
        subopt_1:cds_1
Homo    HMMgene1.1a     exon_14 41770   41886   0.655   +       1
        subopt_1:cds_1
Homo    HMMgene1.1a     lastex  42469   42566   0.587   +       0
        subopt_1:cds_1
Homo    HMMgene1.1a     CDS     21      42566   0.028   +       .
        subopt_1:cds_1
Homo    HMMgene1.1a     firstex 21      46      0.526   +       2
        subopt_2:cds_1
Homo    HMMgene1.1a     exon_1  6842    7032    0.992   +       1
        subopt_2:cds_1
Homo    HMMgene1.1a     exon_2  8189    8334    0.883   +       0
        subopt_2:cds_1
Homo    HMMgene1.1a     exon_3  11165   11249   0.998   +       1
        subopt_2:cds_1
Homo    HMMgene1.1a     exon_4  11607   11719   0.999   +       0
        subopt_2:cds_1
Homo    HMMgene1.1a     exon_5  12095   12303   0.991   +       2
        subopt_2:cds_1
```

```
Homo    HMMgene1.1a     exon_6  15105   15232   0.736   +       1
        subopt_2:cds_1
Homo    HMMgene1.1a     exon_7  15450   15597   0.996   +       2
        subopt_2:cds_1
Homo    HMMgene1.1a     exon_8  23387   23505   0.997   +       1
        subopt_2:cds_1
Homo    HMMgene1.1a     exon_9  24720   24910   0.975   +       0
        subopt_2:cds_1
Homo    HMMgene1.1a     exon_10 28479   28583   0.999   +       0
        subopt_2:cds_1
Homo    HMMgene1.1a     exon_11 32151   32422   0.919   +       2
        subopt_2:cds_1
Homo    HMMgene1.1a     exon_12 34100   34151   0.507   +       0
        subopt_2:cds_1
Homo    HMMgene1.1a     exon_13 37849   37930   0.700   +       1
        subopt_2:cds_1
Homo    HMMgene1.1a     lastex  42469   42566   0.587   +       0
        subopt_2:cds_1
Homo    HMMgene1.1a     CDS     21      42566   0.016   +       .
        subopt_2:cds_1
Homo    HMMgene1.1a     firstex 21      46      0.526   +       2
        subopt_3:cds_1
Homo    HMMgene1.1a     exon_1  6842    7032    0.992   +       1
        subopt_3:cds_1
Homo    HMMgene1.1a     exon_2  8189    8334    0.883   +       0
        subopt_3:cds_1
Homo    HMMgene1.1a     exon_3  11165   11249   0.998   +       1
        subopt_3:cds_1
Homo    HMMgene1.1a     exon_4  11607   11719   0.999   +       0
        subopt_3:cds_1
Homo    HMMgene1.1a     exon_5  12095   12303   0.991   +       2
        subopt_3:cds_1
Homo    HMMgene1.1a     exon_6  15105   15211   0.264   +       1
        subopt_3:cds_1
Homo    HMMgene1.1a     exon_7  15450   15597   0.996   +       2
        subopt_3:cds_1
Homo    HMMgene1.1a     exon_8  23387   23505   0.997   +       1
        subopt_3:cds_1
Homo    HMMgene1.1a     exon_9  24720   24910   0.975   +       0
        subopt_3:cds_1
Homo    HMMgene1.1a     exon_10 28479   28583   0.999   +       0
        subopt_3:cds_1
Homo    HMMgene1.1a     exon_11 32151   32422   0.919   +       2
        subopt_3:cds_1
Homo    HMMgene1.1a     exon_12 34100   34151   0.507   +       0
        subopt_3:cds_1
Homo    HMMgene1.1a     exon_13 37849   37930   0.700   +       1
        subopt_3:cds_1
Homo    HMMgene1.1a     exon_14 41770   41886   0.655   +       1
        subopt_3:cds_1
Homo    HMMgene1.1a     lastex  42469   42566   0.587   +       0
        subopt_3:cds_1
Homo    HMMgene1.1a     CDS     21      42566   0.016   +       .
        subopt_3:cds_1
Homo    HMMgene1.1a     firstex 21      46      0.526   +       2
        subopt_4:cds_1
```

```
Homo    HMMgene1.1a     exon_1  6842    7032    0.992   +       1
        subopt_4:cds_1
Homo    HMMgene1.1a     exon_2  8189    8334    0.883   +       0
        subopt_4:cds_1
Homo    HMMgene1.1a     exon_3  11165   11249   0.998   +       1
        subopt_4:cds_1
Homo    HMMgene1.1a     exon_4  11607   11719   0.999   +       0
        subopt_4:cds_1
Homo    HMMgene1.1a     exon_5  12095   12303   0.991   +       2
        subopt_4:cds_1
Homo    HMMgene1.1a     exon_6  15105   15232   0.736   +       1
        subopt_4:cds_1
Homo    HMMgene1.1a     exon_7  15450   15597   0.996   +       2
        subopt_4:cds_1
Homo    HMMgene1.1a     exon_8  23387   23505   0.997   +       1
        subopt_4:cds_1
Homo    HMMgene1.1a     exon_9  24720   24910   0.975   +       0
        subopt_4:cds_1
Homo    HMMgene1.1a     exon_10 28479   28583   0.999   +       0
        subopt_4:cds_1
Homo    HMMgene1.1a     exon_11 32151   32422   0.919   +       2
        subopt_4:cds_1
Homo    HMMgene1.1a     exon_12 34100   34151   0.507   +       0
        subopt_4:cds_1
Homo    HMMgene1.1a     exon_13 37849   37930   0.700   +       1
        subopt_4:cds_1
Homo    HMMgene1.1a     exon_14 41770   41886   0.655   +       1
        subopt_4:cds_1
Homo    HMMgene1.1a     lastex  43992   44104   0.175   +       0
        subopt_4:cds_1
Homo    HMMgene1.1a     CDS     21      44104   0.013   +       .
        subopt_4:cds_1


Homo    HMMgene1.1a     START   21      23      0.526   +       .
Homo    HMMgene1.1a     DON     46      47      0.526   +       2
Homo    HMMgene1.1a     ACC     206     207     0.002   +       0
Homo    HMMgene1.1a     ACC     206     207     0.009   +       2
Homo    HMMgene1.1a     ACC     299     300     0.004   +       2
Homo    HMMgene1.1a     DON     316     317     0.013   +       1
Homo    HMMgene1.1a     ACC     463     464     0.002   +       0
Homo    HMMgene1.1a     DON     479     480     0.002   +       1
Homo    HMMgene1.1a     START   832     834     0.038   +       .
Homo    HMMgene1.1a     DON     844     845     0.002   +       1
Homo    HMMgene1.1a     DON     883     884     0.005   +       1
Homo    HMMgene1.1a     ACC     884     885     0.012   +       2
Homo    HMMgene1.1a     START   892     894     0.004   +       .
Homo    HMMgene1.1a     DON     913     914     0.046   +       1
Homo    HMMgene1.1a     START   1010    1012    0.002   +       .
Homo    HMMgene1.1a     ACC     1041    1042    0.003   +       2
Homo    HMMgene1.1a     DON     1064    1065    0.003   +       1
Homo    HMMgene1.1a     DON     1076    1077    0.003   +       1
Homo    HMMgene1.1a     ACC     1260    1261    0.005   +       2
Homo    HMMgene1.1a     DON     1277    1278    0.005   +       1
Homo    HMMgene1.1a     ACC     1326    1327    0.275   +       1
Homo    HMMgene1.1a     ACC     1346    1347    0.004   +       0
Homo    HMMgene1.1a     START   1426    1428    0.014   +       .
```

Homo	HMMgene1.1a	ACC	1437	1438	0.001	+	1
Homo	HMMgene1.1a	START	1456	1458	0.047	+	.
Homo	HMMgene1.1a	DON	1458	1459	0.060	+	0
Homo	HMMgene1.1a	DON	1458	1459	0.274	+	1
Homo	HMMgene1.1a	ACC	1503	1504	0.024	+	0
Homo	HMMgene1.1a	ACC	1503	1504	0.248	+	1
Homo	HMMgene1.1a	ACC	1503	1504	0.001	+	2
Homo	HMMgene1.1a	START	1537	1539	0.001	+	.
Homo	HMMgene1.1a	DON	1556	1557	0.038	+	0
Homo	HMMgene1.1a	DON	1556	1557	0.008	+	2
Homo	HMMgene1.1a	DON	1560	1561	0.102	+	1
Homo	HMMgene1.1a	DON	1579	1580	0.008	+	2
Homo	HMMgene1.1a	DON	1586	1587	0.028	+	0
Homo	HMMgene1.1a	DON	1586	1587	0.017	+	2
Homo	HMMgene1.1a	DON	1591	1592	0.069	+	2
Homo	HMMgene1.1a	DON	1594	1595	0.009	+	2
Homo	HMMgene1.1a	STOP	1593	1595	0.001	+	.
Homo	HMMgene1.1a	ACC	1744	1745	0.061	+	0
Homo	HMMgene1.1a	ACC	1744	1745	0.008	+	1
Homo	HMMgene1.1a	ACC	1768	1769	0.002	+	0
Homo	HMMgene1.1a	DON	1824	1825	0.010	+	0
Homo	HMMgene1.1a	DON	1824	1825	0.064	+	2
Homo	HMMgene1.1a	START	1928	1930	0.002	+	.
Homo	HMMgene1.1a	DON	1959	1960	0.003	+	2
Homo	HMMgene1.1a	ACC	2725	2726	0.002	+	0
Homo	HMMgene1.1a	ACC	2725	2726	0.116	+	1
Homo	HMMgene1.1a	START	2732	2734	0.001	+	.
Homo	HMMgene1.1a	DON	2745	2746	0.002	+	2
Homo	HMMgene1.1a	DON	2760	2761	0.003	+	0
Homo	HMMgene1.1a	DON	2783	2784	0.112	+	2
Homo	HMMgene1.1a	ACC	2852	2853	0.002	+	0
Homo	HMMgene1.1a	START	2928	2930	0.015	+	.
Homo	HMMgene1.1a	DON	2947	2948	0.010	+	2
Homo	HMMgene1.1a	DON	2965	2966	0.006	+	2
Homo	HMMgene1.1a	ACC	3067	3068	0.003	+	0
Homo	HMMgene1.1a	DON	3102	3103	0.003	+	2
Homo	HMMgene1.1a	ACC	3168	3169	0.001	+	1
Homo	HMMgene1.1a	DON	3187	3188	0.001	+	2
Homo	HMMgene1.1a	ACC	3228	3229	0.002	+	0
Homo	HMMgene1.1a	ACC	3240	3241	0.006	+	0
Homo	HMMgene1.1a	DON	3308	3309	0.008	+	2
Homo	HMMgene1.1a	ACC	3576	3577	0.002	+	1
Homo	HMMgene1.1a	DON	3631	3632	0.003	+	2
Homo	HMMgene1.1a	ACC	5168	5169	0.001	+	0
Homo	HMMgene1.1a	DON	5200	5201	0.002	+	2
Homo	HMMgene1.1a	ACC	5393	5394	0.025	+	0
Homo	HMMgene1.1a	ACC	5419	5420	0.009	+	2
Homo	HMMgene1.1a	START	5442	5444	0.029	+	.
Homo	HMMgene1.1a	DON	5473	5474	0.063	+	2
Homo	HMMgene1.1a	ACC	6209	6210	0.009	+	0
Homo	HMMgene1.1a	DON	6235	6236	0.009	+	2
Homo	HMMgene1.1a	ACC	6324	6325	0.001	+	0
Homo	HMMgene1.1a	DON	6386	6387	0.001	+	2
Homo	HMMgene1.1a	START	6612	6614	0.106	+	.
Homo	HMMgene1.1a	DON	6631	6632	0.106	+	2
Homo	HMMgene1.1a	ACC	6809	6810	0.004	+	0
Homo	HMMgene1.1a	ACC	6819	6820	0.003	+	1

Homo	HMMgene1.1a	ACC	6841	6842	0.992	+	2
Homo	HMMgene1.1a	DON	7032	7033	0.999	+	1
Homo	HMMgene1.1a	ACC	8158	8159	0.090	+	1
Homo	HMMgene1.1a	ACC	8188	8189	0.890	+	1
Homo	HMMgene1.1a	ACC	8212	8213	0.019	+	1
Homo	HMMgene1.1a	DON	8307	8308	0.002	+	0
Homo	HMMgene1.1a	DON	8334	8335	0.993	+	0
Homo	HMMgene1.1a	DON	8373	8374	0.005	+	0
Homo	HMMgene1.1a	ACC	11164	11165	1.000	+	0
Homo	HMMgene1.1a	DON	11249	11250	0.998	+	1
Homo	HMMgene1.1a	ACC	11606	11607	0.999	+	1
Homo	HMMgene1.1a	DON	11719	11720	0.999	+	0
Homo	HMMgene1.1a	ACC	12019	12020	0.008	+	0
Homo	HMMgene1.1a	ACC	12094	12095	0.991	+	0
Homo	HMMgene1.1a	DON	12303	12304	1.000	+	2
Homo	HMMgene1.1a	ACC	15104	15105	1.000	+	2
Homo	HMMgene1.1a	DON	15211	15212	0.264	+	1
Homo	HMMgene1.1a	DON	15232	15233	0.736	+	1
Homo	HMMgene1.1a	ACC	15449	15450	0.999	+	1
Homo	HMMgene1.1a	DON	15540	15541	0.001	+	2
Homo	HMMgene1.1a	DON	15597	15598	0.997	+	2
Homo	HMMgene1.1a	ACC	16704	16705	0.002	+	2
Homo	HMMgene1.1a	DON	16841	16842	0.002	+	1
Homo	HMMgene1.1a	ACC	19931	19932	0.024	+	2
Homo	HMMgene1.1a	DON	19991	19992	0.024	+	2
Homo	HMMgene1.1a	ACC	20337	20338	0.048	+	2
Homo	HMMgene1.1a	DON	20406	20407	0.048	+	2
Homo	HMMgene1.1a	ACC	20827	20828	0.002	+	1
Homo	HMMgene1.1a	DON	20909	20910	0.002	+	2
Homo	HMMgene1.1a	ACC	21828	21829	0.003	+	2
Homo	HMMgene1.1a	DON	21912	21913	0.003	+	2
Homo	HMMgene1.1a	ACC	23322	23323	0.003	+	1
Homo	HMMgene1.1a	ACC	23386	23387	0.997	+	2
Homo	HMMgene1.1a	DON	23505	23506	1.000	+	1
Homo	HMMgene1.1a	ACC	23873	23874	0.009	+	1
Homo	HMMgene1.1a	DON	24011	24012	0.006	+	1
Homo	HMMgene1.1a	DON	24016	24017	0.004	+	0
Homo	HMMgene1.1a	ACC	24679	24680	0.002	+	0
Homo	HMMgene1.1a	ACC	24686	24687	0.002	+	1
Homo	HMMgene1.1a	ACC	24719	24720	0.995	+	1
Homo	HMMgene1.1a	DON	24853	24854	0.006	+	0
Homo	HMMgene1.1a	DON	24910	24911	0.980	+	0
Homo	HMMgene1.1a	DON	24928	24929	0.014	+	0
Homo	HMMgene1.1a	ACC	25743	25744	0.006	+	0
Homo	HMMgene1.1a	ACC	25764	25765	0.005	+	0
Homo	HMMgene1.1a	DON	25799	25800	0.011	+	2
Homo	HMMgene1.1a	ACC	26039	26040	0.011	+	2
Homo	HMMgene1.1a	DON	26121	26122	0.011	+	0
Homo	HMMgene1.1a	ACC	26359	26360	0.002	+	0
Homo	HMMgene1.1a	DON	26389	26390	0.002	+	0
Homo	HMMgene1.1a	ACC	28478	28479	0.999	+	0
Homo	HMMgene1.1a	DON	28583	28584	0.999	+	0
Homo	HMMgene1.1a	ACC	29887	29888	0.001	+	0
Homo	HMMgene1.1a	DON	29914	29915	0.001	+	0
Homo	HMMgene1.1a	ACC	30613	30614	0.002	+	0
Homo	HMMgene1.1a	DON	30709	30710	0.002	+	0
Homo	HMMgene1.1a	ACC	32024	32025	0.001	+	0

Homo	HMMgene1.1a	ACC	32150	32151	0.998	+	0
Homo	HMMgene1.1a	DON	32277	32278	0.008	+	1
Homo	HMMgene1.1a	DON	32343	32344	0.034	+	1
Homo	HMMgene1.1a	DON	32422	32423	0.920	+	2
Homo	HMMgene1.1a	STOP	32424	32426	0.038	+	.
Homo	HMMgene1.1a	ACC	32494	32495	0.003	+	2
Homo	HMMgene1.1a	DON	32547	32548	0.002	+	1
Homo	HMMgene1.1a	STOP	32595	32597	0.001	+	.
Homo	HMMgene1.1a	ACC	33077	33078	0.002	+	2
Homo	HMMgene1.1a	DON	33130	33131	0.001	+	1
Homo	HMMgene1.1a	ACC	33240	33241	0.005	+	2
Homo	HMMgene1.1a	DON	33272	33273	0.002	+	1
Homo	HMMgene1.1a	DON	33282	33283	0.001	+	2
Homo	HMMgene1.1a	START	33400	33402	0.027	+	.
Homo	HMMgene1.1a	START	33415	33417	0.006	+	.
Homo	HMMgene1.1a	ACC	33433	33434	0.004	+	1
Homo	HMMgene1.1a	START	33436	33438	0.005	+	.
Homo	HMMgene1.1a	ACC	33463	33464	0.025	+	1
Homo	HMMgene1.1a	DON	33585	33586	0.003	+	0
Homo	HMMgene1.1a	DON	33592	33593	0.022	+	1
Homo	HMMgene1.1a	ACC	33600	33601	0.052	+	2
Homo	HMMgene1.1a	DON	33602	33603	0.026	+	2
Homo	HMMgene1.1a	DON	33607	33608	0.004	+	1
Homo	HMMgene1.1a	STOP	33649	33651	0.012	+	.
Homo	HMMgene1.1a	ACC	33676	33677	0.011	+	1
Homo	HMMgene1.1a	ACC	33676	33677	0.002	+	2
Homo	HMMgene1.1a	DON	33750	33751	0.004	+	2
Homo	HMMgene1.1a	DON	33772	33773	0.003	+	0
Homo	HMMgene1.1a	DON	33772	33773	0.002	+	1
Homo	HMMgene1.1a	DON	33811	33812	0.007	+	0
Homo	HMMgene1.1a	DON	33811	33812	0.004	+	1
Homo	HMMgene1.1a	DON	33814	33815	0.017	+	0
Homo	HMMgene1.1a	DON	33814	33815	0.002	+	1
Homo	HMMgene1.1a	DON	33826	33827	0.016	+	0
Homo	HMMgene1.1a	DON	33826	33827	0.002	+	1
Homo	HMMgene1.1a	STOP	33848	33850	0.007	+	.
Homo	HMMgene1.1a	ACC	34099	34100	0.854	+	2
Homo	HMMgene1.1a	DON	34136	34137	0.002	+	0
Homo	HMMgene1.1a	DON	34151	34152	0.507	+	0
Homo	HMMgene1.1a	DON	34168	34169	0.005	+	2
Homo	HMMgene1.1a	DON	34217	34218	0.330	+	0
Homo	HMMgene1.1a	DON	34221	34222	0.011	+	1
Homo	HMMgene1.1a	ACC	36514	36515	0.034	+	0
Homo	HMMgene1.1a	DON	36567	36568	0.026	+	2
Homo	HMMgene1.1a	STOP	36569	36571	0.008	+	.
Homo	HMMgene1.1a	ACC	37130	37131	0.001	+	2
Homo	HMMgene1.1a	DON	37162	37163	0.001	+	1
Homo	HMMgene1.1a	START	37678	37680	0.001	+	.
Homo	HMMgene1.1a	ACC	37701	37702	0.127	+	0
Homo	HMMgene1.1a	ACC	37727	37728	0.051	+	2
Homo	HMMgene1.1a	ACC	37753	37754	0.043	+	1
Homo	HMMgene1.1a	ACC	37763	37764	0.004	+	2
Homo	HMMgene1.1a	ACC	37777	37778	0.001	+	1
Homo	HMMgene1.1a	ACC	37848	37849	0.710	+	0
Homo	HMMgene1.1a	ACC	37864	37865	0.007	+	1
Homo	HMMgene1.1a	DON	37930	37931	0.933	+	1
Homo	HMMgene1.1a	STOP	37936	37938	0.012	+	.

Homo	HMMgene1.1a	ACC	38258	38259	0.001	+	2
Homo	HMMgene1.1a	ACC	38308	38309	0.001	+	1
Homo	HMMgene1.1a	ACC	38332	38333	0.005	+	1
Homo	HMMgene1.1a	DON	38356	38357	0.007	+	1
Homo	HMMgene1.1a	DON	38491	38492	0.001	+	1
Homo	HMMgene1.1a	ACC	38651	38652	0.001	+	1
Homo	HMMgene1.1a	DON	38706	38707	0.001	+	2
Homo	HMMgene1.1a	ACC	39118	39119	0.003	+	1
Homo	HMMgene1.1a	DON	39163	39164	0.003	+	1
Homo	HMMgene1.1a	ACC	39355	39356	0.008	+	1
Homo	HMMgene1.1a	DON	39454	39455	0.009	+	1
Homo	HMMgene1.1a	ACC	39649	39650	0.001	+	1
Homo	HMMgene1.1a	DON	39717	39718	0.002	+	0
Homo	HMMgene1.1a	ACC	40054	40055	0.002	+	0
Homo	HMMgene1.1a	STOP	40172	40174	0.001	+	.
Homo	HMMgene1.1a	ACC	40239	40240	0.003	+	1
Homo	HMMgene1.1a	DON	40321	40322	0.001	+	2
Homo	HMMgene1.1a	STOP	40323	40325	0.002	+	.
Homo	HMMgene1.1a	ACC	41056	41057	0.002	+	1
Homo	HMMgene1.1a	ACC	41098	41099	0.002	+	1
Homo	HMMgene1.1a	DON	41123	41124	0.003	+	2
Homo	HMMgene1.1a	ACC	41331	41332	0.002	+	1
Homo	HMMgene1.1a	DON	41424	41425	0.003	+	1
Homo	HMMgene1.1a	ACC	41690	41691	0.010	+	0
Homo	HMMgene1.1a	ACC	41731	41732	0.005	+	2
Homo	HMMgene1.1a	ACC	41736	41737	0.010	+	1
Homo	HMMgene1.1a	ACC	41769	41770	0.665	+	1
Homo	HMMgene1.1a	ACC	41784	41785	0.008	+	1
Homo	HMMgene1.1a	DON	41886	41887	0.688	+	1
Homo	HMMgene1.1a	DON	41892	41893	0.005	+	1
Homo	HMMgene1.1a	STOP	41901	41903	0.004	+	.
Homo	HMMgene1.1a	ACC	42100	42101	0.002	+	1
Homo	HMMgene1.1a	STOP	42169	42171	0.002	+	.
Homo	HMMgene1.1a	ACC	42279	42280	0.014	+	1
Homo	HMMgene1.1a	STOP	42297	42299	0.014	+	.
Homo	HMMgene1.1a	ACC	42468	42469	0.591	+	1
Homo	HMMgene1.1a	DON	42493	42494	0.005	+	2
Homo	HMMgene1.1a	ACC	42501	42502	0.014	+	1
Homo	HMMgene1.1a	STOP	42564	42566	0.600	+	.
Homo	HMMgene1.1a	ACC	42934	42935	0.002	+	1
Homo	HMMgene1.1a	DON	42966	42967	0.002	+	0
Homo	HMMgene1.1a	ACC	43097	43098	0.003	+	0
Homo	HMMgene1.1a	ACC	43097	43098	0.078	+	1
Homo	HMMgene1.1a	ACC	43097	43098	0.012	+	2
Homo	HMMgene1.1a	STOP	43139	43141	0.077	+	.
Homo	HMMgene1.1a	STOP	43150	43152	0.012	+	.
Homo	HMMgene1.1a	ACC	43221	43222	0.001	+	1
Homo	HMMgene1.1a	STOP	43236	43238	0.004	+	.
Homo	HMMgene1.1a	ACC	43264	43265	0.002	+	1
Homo	HMMgene1.1a	STOP	43339	43341	0.002	+	.
Homo	HMMgene1.1a	ACC	43469	43470	0.002	+	1
Homo	HMMgene1.1a	STOP	43547	43549	0.002	+	.
Homo	HMMgene1.1a	ACC	43901	43902	0.002	+	1
Homo	HMMgene1.1a	ACC	43931	43932	0.018	+	1
Homo	HMMgene1.1a	STOP	43955	43957	0.019	+	.
Homo	HMMgene1.1a	ACC	43991	43992	0.177	+	1
Homo	HMMgene1.1a	ACC	44023	44024	0.002	+	1

```
Homo    HMMgene1.1a    STOP    44095   44097   0.002   +       .
Homo    HMMgene1.1a    DON     44100   44101   0.002   +       2
Homo    HMMgene1.1a    STOP    44102   44104   0.176   +       .
Homo    HMMgene1.1a    ACC     45063   45064   0.022   +       1
Homo    HMMgene1.1a    ACC     45093   45094   0.005   +       1
Homo    HMMgene1.1a    DON     45103   45104   0.002   +       2
Homo    HMMgene1.1a    STOP    45105   45107   0.025   +       .
Homo    HMMgene1.1a    ACC     45123   45124   0.003   +       1
Homo    HMMgene1.1a    STOP    45159   45161   0.004   +       .
Homo    HMMgene1.1a    ACC     45276   45277   0.002   +       1
Homo    HMMgene1.1a    STOP    45306   45308   0.002   +       .
Homo    HMMgene1.1a    ACC     45389   45390   0.001   +       1
Homo    HMMgene1.1a    ACC     45389   45390   0.001   +       2
Homo    HMMgene1.1a    STOP    45406   45408   0.001   +       .
Homo    HMMgene1.1a    STOP    45461   45463   0.002   +       .
```

Figure 5: NetGene2 Output:

```
CUTOFF values used for confidence:

Highly confident donor sites (H):     95.0 %
Nearly all true donor sites:          50.0 %

Highly confident acceptor sites (H):  95.0 %
Nearly all true acceptor sites:       20.0 %


Length: 45950 nucleotides.
24.0% A, 25.8% C, 24.3% G, 25.9% T, 0.0% X, 50.1% G+C
Donor splice sites, direct strand
---------------------------------
pos 5'->3'  phase strand  confidence  5'      exon intron    3'

    914       1     +        0.31       CGGGCAGAGG^GTGGGTTTCT
   1077       0     +        0.41       GAGGCACCCA^GTAAGCCTCA
   1459       1     +        0.94       CTGGGCGATG^GTGAGTGGGC H
   1557       0     +        0.64       GAAGTCCCTG^GTGGGTACGG
   1561       1     +        0.59       TCCCTGGTGG^GTACGGCTTG
   1592       2     +        0.32       GCCCAGTGAG^GTAGTATTCC
   2498       2     +        0.41       TATACTAGAA^GTAAGCAAAT
   2523       1     +        0.54       AGGCATGGTG^GTGAGTGTCT
   2784       2     +        0.37       CACCCAACAG^GTGTGCCACA
   3103       2     +        0.47       CTCTAGCTCC^GTGAGTGTTC
   3309       0     +        0.54       AGGGAGAACA^GTAAGTCTAA
   3632       2     +        0.44       CACACAGCTT^GTAAGTAGTG
   3648       0     +        0.47       AGTGGAGCTT^GTAAGTAGTA
   6236       1     +        0.63       ACAGGGCAGG^GTGAGTGGTA
   6387       0     +        0.34       GGCAGAGGAG^GTGATTCTGG
   6632       1     +        0.41       AAGACCACAG^GTAAAATCAG
   7033       1     +        1.00       TGCGCCGCAG^GTGAGCCCCG H
   8080       1     +        0.63       GGGGGTGCAT^GTAAGTTTCT
```

8125	2	+	0.41	CTAGCCCAGG^GTCAGTCCAT	
8335	0	+	1.00	CCTCAGTCAG^GTGAGAGGAG	H
8452	1	+	0.45	GGGGATATAG^GTATGTCACT	
8623	1	+	0.82	CGGTCAAAGG^GTATGGAGAA	
8699	0	+	0.53	AGGCCAAGAG^GTGAGGAAAA	
9253	2	+	0.47	GATCCTCTAG^GTACGGCATG	
9266	0	+	0.34	CGGCATGGAG^GTAGATTTGG	
9831	0	+	0.41	CATTCACATG^GTGAGCCAAT	
10205	2	+	0.31	CCACTGGCCA^GTAAGCTCTT	
10365	0	+	0.44	GGGCAACATG^GTGAGACCCC	
11250	1	+	0.95	CCCCAGCAGA^GTGAGTTCCC	H
11720	0	+	1.00	CACCTGGAAG^GTGCGTGGGG	H
11842	0	+	0.34	GCTCACTCAG^GTCAGGTCTA	
12304	2	+	0.91	GTGACGACAG^GTGGGCAGGG	H
12938	0	+	0.41	GGAGGCCAAG^GTGGGTGGAT	
13636	1	+	0.81	GAGCAGCTCA^GTGAGTTTCC	
14873	1	+	0.37	CTTTAGCCAC^GTAAGAGACC	
15212	1	+	0.31	ATCCACACCT^GTATGTATCC	
15233	1	+	0.94	ATTTCCCCAG^GTGCGAAGCC	H
15598	2	+	0.67	TTTTAATCAG^GTAAGCAGGA	
15657	0	+	0.41	GTGTGGACAG^GTGTGTGTGT	
15775	2	+	0.37	GTAAATGCCG^GTGTGTGTGT	
15799	2	+	0.50	ATGGACACAG^GTATGTGTAT	
18944	0	+	0.41	GGAGTCCAAG^GTGGGTGGAT	
19248	0	+	0.37	CATCAGTAAC^GTATGGGGGT	
20008	0	+	0.47	GGGACTATAG^GTGCGCACTA	
21220	2	+	0.74	CATTAAAAAG^GTAAACGGAA	
21241	0	+	0.36	GCTGGGCATG^GTAACTCACA	
21913	2	+	0.41	ACACCACTGT^GTAAGTAGCG	
23506	1	+	0.38	GACAGCCTTG^GTATGGGGCA	
24911	1	+	0.76	TCCTCTTTCG^GTAAGCCATC	
25306	1	+	0.37	GATTTCACAG^GTTGGGCCTC	
25805	1	+	0.85	ATTGAGTGAG^GTATGGAGGA	
28584	0	+	1.00	TATAGCTCAG^GTAAGGGCTG	H
29716	0	+	0.53	AAAAGGAGGG^GTAAGAGGGG	
32423	0	+	0.82	CTCGGAAGAG^GTGAGCCACA	
32888	1	+	0.44	CTGGGCAATG^GTGAGACCCT	
33972	2	+	0.32	GGATGGACTG^GTGGGTGTAT	
34500	2	+	0.31	GCAAATACAT^GTAAGATGGC	
34511	1	+	0.47	TAAGATGGCA^GTAAGTCAGC	
36568	2	+	0.70	AGGAACACAA^GTAGGATGAC	
37931	1	+	0.93	GAGCCCCTGG^GTGAGTAGCA	H
38357	0	+	0.41	GGGATCACAG^GTGTGTGCCA	
38707	1	+	0.62	AGAGGGCAGG^GTGAGTGTTT	
38840	0	+	0.47	TAAAAACCTG^GTGAGTGTGG	
39455	0	+	0.39	GGGACTGCAG^GTGGGCACCA	
39718	0	+	0.47	ACCACCTGAG^GTGAGGAGTT	
40175	0	+	0.39	AGGCCCTTAG^GTGGGGACTG	
40557	2	+	0.83	GGAGAGAAAG^GTGAGGGCTG	
41124	2	+	0.41	CTGACACAAA^GTAGGTGCTC	
41485	0	+	0.31	CATAAATGAA^GTATGTGTGA	
41491	0	+	0.41	TGAAGTATGT^GTGAGTATCC	
42898	2	+	0.41	GGAATCAGGG^GTGAGGAGGG	
45104	2	+	0.47	ACCTCATTCC^GTAAGTTCCC	

Acceptor splice sites, direct strand

pos 5'->3'	phase	strand	confidence	5' intron exon 3'
206	0	+	0.28	CCACCCTTAG^GGACTGCAGG
463	0	+	0.33	TTGCCACCAG^GTGCCTGGGA
942	0	+	0.33	TGCCCAGCAG^GTGCAGCTCT
1114	1	+	0.17	GGCTTCTCAG^GCCTGATTGG
1188	1	+	0.31	TCTATTCCAG^GCCCTGGCTC
1326	1	+	0.25	CCCACCCAAG^GCCGGCCCTT
1503	1	+	0.82	CTTCCTCTAG^CTGGGGCTGG
1744	0	+	0.56	GGTCTCTCAG^GACACCTCTC
1986	0	+	0.15	CTGGTTGGAG^GGGGGCACTG
2017	0	+	0.16	CCATTCTCAG^GCCTCACCCT
2335	2	+	0.18	ATTGTGTCAG^TGACTGCCGC
2725	1	+	0.69	ACCACCCCAG^ATCAAGATGA
3067	0	+	0.16	TGCTCAATAG^GGTAGTAGTA
3168	1	+	0.26	TCCTTTTAAG^ATGCTTGTGT
3298	0	+	0.25	ATACGCACAG^AGGGAGAACA
4179	0	+	0.26	CTGTCTCCAG^ACAGGAAAAT
4183	1	+	0.31	CTCCAGACAG^GAAAATATCC
4303	1	+	0.14	CCCCTCCTAG^AGGTTCTGCT
4832	0	+	0.33	TCCCTCCTAG^GGGTCCCCTG
6298	2	+	0.71	CCCCTCTTAG^GGCCCCTGTT
6515	1	+	0.43	TGTCCCCCAG^GCGTCCCTAG
6681	1	+	0.26	CCAGGCCCAG^GGAGCGCCGC
6809	0	+	0.85	CTTTCCCGAG^TCACCACCAG
6819	1	+	0.45	TCACCACCAG^TGAATGCTGC
6841	2	+	0.96	ATGCCCTCAG^GTCGACCACG
6854	0	+	0.17	GACCACGGAG^CTGCGCAAGG
7728	1	+	0.27	ACCCCCTTAG^AAGTCTGTTC
7761	2	+	0.15	TCTTCTCTAG^AGTTCTGTCC
7912	1	+	0.43	GTCCTCCTAG^GAGGCCCACC
7968	2	+	0.16	ACTTCCTTAG^AATTCTGCCA
8188	1	+	0.19	CGGGCACCAG^GGGAGTGGAA
8193	0	+	0.31	ACCAGGGGAG^TGGAACCAGG
8202	0	+	0.34	GTGGAACCAG^GTGGGAGCAG
8209	1	+	0.44	CAGGTGGGAG^CAGGGGGAGA
8212	1	+	0.44	GTGGGAGCAG^GGGGAGAACC
8244	0	+	0.14	CTACCTGAAG^GCCCTGGAGG
8547	1	+	0.07	CCTGACCCAG^CCTGAGTGGG
8553	1	+	0.17	CCAGCCTGAG^TGGGAATGGA
8564	0	+	0.19	GGGAATGGAG^GGCTTCCTGG
8576	0	+	0.18	CTTCCTGGAG^GAGAAGATAT
8579	0	+	0.17	CCTGGAGGAG^AAGATATCTG
9064	0	+	0.56	ATCCTCCTAG^GTGCAGAGTG
9162	2	+	0.71	TTATCTCCAG^GGCCCCGGGG
9874	2	+	0.29	CTCCTCCCAG^CACCACGCTC
9990	0	+	0.49	CTCTTTCTAG^AATACTCTTA
10006	1	+	0.27	CTTACTCCAG^GGGGCCATAA
11164	0	+	0.92	TCCCTGGCAG^CTGGAGCTCA
11170	0	+	0.43	GCAGCTGGAG^CTCATTGGAC
11184	2	+	0.34	TTGGACACAG^CATCTTTGAT
11495	0	+	0.33	CTGGTTCCAG^GTCCCAATTG
11606	1	+	0.34	GCCTCCCCAG^CCCTGTCCAG
11616	2	+	0.76	CCCTGTCCAG^GAGGAAGGTG

11619	2	+	0.34	TGTCCAGGAG^GAAGGTGGAG
11623	0	+	0.34	CAGGAGGAAG^GTGGAGGCCC
11629	0	+	0.31	GAAGGTGGAG^GCCCCCACGG
12094	1	+	0.82	CGGCCCCCAG^GTGCTGAACT
12142	0	+	0.17	ACCTGCGCAG^ACTTCTCCAG
12152	1	+	0.19	ACTTCTCCAG^CTGGGAGCCC
12159	2	+	0.19	CAGCTGGGAG^CCCTGACTCA
12170	1	+	0.19	CCTGACTCAG^AGCCCCCGCT
12172	0	+	0.19	TGACTCAGAG^CCCCCGCTGC
12184	0	+	0.18	CCCGCTGCAG^TGCCTGGTGC
12221	1	+	0.25	CCCCACCCAG^GCAGCCTGGA
13485	2	+	0.23	TCATGCTTAG^GAATGGTGAC
14469	0	+	0.33	CCTTTTCCAG^CTGCTAGAGC
15066	0	+	0.28	TCTCACCCAG^TAGCATCCTG
15069	0	+	0.34	CACCCAGTAG^CATCCTGACC
15081	0	+	0.31	TCCTGACCAG^ACCCCACTCC
15104	2	+	0.96	CCACCCCCAG^GATTGCAGAA H
15112	1	+	0.34	AGGATTGCAG^AAGTGGCTGG
15115	1	+	0.31	ATTGCAGAAG^TGGCTGGCTA
15128	2	+	0.18	CTGGCTATAG^TCCCGATGAC
15449	1	+	1.00	CCCTCCTCAG^TGCTGAGCAA H
15456	2	+	0.34	CAGTGCTGAG^CAAGGGCCAG
15460	0	+	0.32	GCTGAGCAAG^GGCCAGGCAG
15466	0	+	0.19	CAAGGGCCAG^GCAGTAACAG
15470	1	+	0.19	GGCCAGGCAG^TAACAGGGCA
15476	1	+	0.07	GCAGTAACAG^GGCAGTATCG
17910	1	+	0.43	CTTTTTCCAG^AGGGTCACAT
19790	0	+	0.17	TTCTTTGAAG^AAGGTTATCT
20337	0	+	0.56	CCCTTCTCAG^AGCCCTTCTG
21899	1	+	0.16	CATATTCAAG^ATCACACCAC
21956	2	+	0.14	CTGACTACAG^AGTCCTTGCT
22184	0	+	0.33	TATCTTTCAG^GGGAATAAGG
22420	1	+	0.16	GTCTCTGCAG^CCAGAGACTG
22724	2	+	0.53	CTGTTCCTAG^GTCTCCTCCC
22793	0	+	0.27	CTCTGCCCAG^GCTGTGCTAA
22837	0	+	0.56	CCTTTTACAG^AAGGAGCTTG
23306	2	+	0.26	TCTGTAATAG^GACATCACCT
23386	2	+	0.44	TGTACCCCAG^CCAGGTGGAA
23390	0	+	0.76	CCCCAGCCAG^GTGGAAGAGA
23397	1	+	0.32	CAGGTGGAAG^AGACCGGAGT
23873	1	+	0.17	TGGCTTCCAG^TAATCTCGAT
23941	0	+	0.07	CCTGGATGAG^CCTCCCTCCC
24136	1	+	0.31	TTCTCCTAAG^GTGAAAGCTC
24208	0	+	0.25	GACCTTTCAG^AGCCTCACCA
24686	1	+	0.16	CAGTCACCAG^CCCGTGTCCT
24719	1	+	0.96	TGTCCTGCAG^ACACCCCTGG
25344	0	+	0.25	TGCTCTTCAG^ACACCAGTCA
25954	1	+	0.33	CTTCCCCCAG^GGTATCGGGC
26359	0	+	0.27	CTTTGGGCAG^GTGAAACGAG
26864	1	+	0.30	CCTCCCACAG^TGTGTGGGAA
28371	1	+	0.33	TGTGCCCCAG^ATGCCTGGCA
28438	2	+	0.17	TACCTCCCAG^GCTCAAGATT
28456	2	+	0.07	TTTCCTGAAG^TTTCTACTCC
28478	0	+	1.00	CTCTCCACAG^GCTGATCTCC H
28491	1	+	0.77	GATCTCCCAG^ATGAACTACC
28514	0	+	0.07	GGGCACCGAG^AATGTGCACA
29535	1	+	0.43	CCCCTTACAG^CATCAGCTTG

30846	1	+	0.56	TTCTCCTCAG^GGACATGTTG
31519	1	+	0.25	ATAACTGCAG^GCAACAGTGT
31852	0	+	0.29	TTTTGTGTAG^GTGGGGGTCT
32000	1	+	0.31	TGTTCTTCAG^GGAGTCCTTA
32024	0	+	0.15	CATTTTACAG^ATGAGGAAAC
32058	1	+	0.15	TTGACCACAG^GCACAGAACT
32150	0	+	1.00	CCCTCCGCAG^GATGCTGATG H
32174	0	+	0.20	GGATTTGGAG^ATGCTGGCCC
32536	1	+	0.15	TCTGCTCCAG^CCAGGCCCTC
32540	2	+	0.33	CTCCAGCCAG^GCCCTCGGTG
34060	0	+	0.15	ACTTGCCCAG^GTTTGCTGTG
34077	1	+	0.18	GTGTCCTGAG^ATTCTTGCCT
34099	2	+	1.00	TTTCCTCCAG^GACCCTGGCC H
34112	0	+	0.18	CCTGGCCCAG^AGCTCAGAGG
34114	2	+	0.07	TGGCCCAGAG^CTCAGAGGAC
34213	1	+	0.31	TTCCTCTCAG^CCTGGTGTGT
34286	2	+	0.33	TGTTTTATAG^ATAGGAAACC
36514	0	+	0.56	CTGTGTACAG^GTCCTCGGGC
37763	2	+	0.07	CCCCCGAAAG^TGCTGGGATT
37777	1	+	0.18	GGGATTACAG^TCACCACTCC
37848	0	+	0.25	TCTCCCACAG^AGTTTCCTTC
37850	2	+	0.29	TCCCACAGAG^TTTCCTTCTG
37864	1	+	0.43	CTTCTGACAG^GAGGACCAGC
39118	1	+	0.26	CTCACTGCAG^GCTCTGCCCC
41602	1	+	0.33	CTCACAACAG^AGGAGTAAAT
41769	1	+	0.66	TTCACCGCAG^CCCTGAACAC
41784	1	+	0.18	AACACCCCAG^AGAGGTTCAG
41786	0	+	0.18	CACCCCAGAG^AGGTTCAGAG
41788	2	+	0.07	CCCCAGAGAG^GTTCAGAGAG
42279	0	+	0.16	TTTGTTCCAG^TGTGCTCTCA
42468	1	+	0.53	ATCTTTGCAG^GCCTGGGCCC
42501	1	+	0.33	CCGTACTCAG^ACGAGGACAC
42603	0	+	0.77	CCTCCCCCAG^AAAGGACCTC
42815	1	+	0.33	CTTCTCTCAG^GATTTCTCTT
42934	1	+	0.33	GTGTTTCCAG^GTTCTGGGGA
43097	0	+	0.77	CCCTCTTCAG^GTGTCAGTGC
43469	1	+	0.29	TTTTTGACAG^AGTCTTGCTC
43991	1	+	0.25	CTTGTTTTAG^ATCTCACACC
44088	1	+	0.17	TTGGTTCCAG^TTCTCCTGAT
44469	2	+	0.27	TTTGTTTTAG^ACCTTATGAG
44500	1	+	0.25	TCTTCTGCAG^TTCTCTATAA
44920	1	+	0.17	GTTCTTCTAG^AACCAGACGA
44947	1	+	0.25	GTTCTCCTAG^ATTCTGAACC
45139	0	+	0.16	TCATTTCCAG^CTCCTCCTCA
45780	1	+	0.31	TACCCTCTAG^GCTGTGCTGC
45866	1	+	0.00	CCCACCCCAG^CCCTCCAGAC
45874	0	+	0.00	AGCCCTCCAG^ACATGCACTT

Literature Cited:

1) Lander ES et al, *Nature* 409, 860 - 921 (2001).

2) Venter JC et al, *Science* 291 (5507) 1304-1351 (2001).

3) Maynard MA, et al *J Biol Chem* (2003).

4) Moises Burset and Roderic Guigo, *Genomics* 34, 353-367 (1996).

5) TA Thanaraj and Alan J Robinson, *Briefings in Bioinformatics* 1 (4) 343-356 (2000).

6) Francis Clark and TA Thanaraj, Human Molecular Genetics 11 (4) 451-464 (2002).

7) TA Thanaraj, *Nucleic Acids Research* 28 (3) 744-754 (2000).

8) Modrek B et al, *Nucleic Acids Research* 29 (13) 2850-2859 (2001).

9) Douglas L Black, *Cell* 103 367-370 (2000).

10) Christopher WJ Smith and Juan Valcarcel, *TIBS* 25 (2000).

11) Mironov AA et al, *Genome Research* 9 1288-1293 (1999).

12) Brenton R Graveley, *Trends in Genetics* 17 (2) 2001.

13) Manjula Das et al, *Genomics* 77 (1-2) 71-78 (2001).

14) Roderic Guigo et al, *Genome Research* 10 1631-1642 (2000).

15) Chris Burge and Samuel Karlin, *J. Mol. Biol*. 268 78-94 (1997).

16) CSE 549 - Gene Prediction (Lectures 10-12) @ http://www.cs.sunysb.edu/~skiena/549/lectures/geneprediction/node1.html

17) http://genes.mit.edu/ GENSCAN.html

18) http://www.ebi.ac.uk/~thanaraj/MZEF-SPC.html

19) Lincoln Stein, *Nature Reviews Genetics* 2 493-505 (2001).

20) S.M. Hebsgaard, P.G. Korning, N. Tolstrup, J. Engelbrecht, P. Rouze, S. Brunak: 15, *Nucleic Acids Research* 24 (17) 3439-3452 (1996).

21) Brunak, S., Engelbrecht, J., and Knudsen, S, *Journal of Molecular Biology* 220 49-65 (1991).

22) Krogh, A, *Proc Int Conf Intell Syst Mol Biol* 5 179-86 (1997).