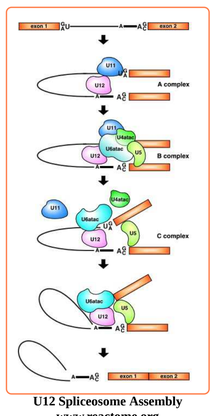
Prediction of Genes with U12-dependent Introns in the Human Genome

ALIOTO, T.S.[1*] GUIGÓ, R.[1]

1 Grup de Recerca en Informàtica Biomèdica
IMIM - Universitat Pompeu Fabra -
Centre de Regulació Genòmica
Barcelona (SPAIN)

1. U12-Dependent Introns



U12 Spliceosome Assembly
www.reactome.org

PROBLEM:

U12-dependent introns are very rare (0.3% of introns) and not very extensively studied. They come in two flavors: AT-AC and GT-AG and are characterized by highly conserved 5' donor splice [G/A]TATCCT and branch point TCCTT[G/A]AC signals. They are frequently missed by gene prediction programs, and annotation pipelines often assign splice junctions incorrectly.

STRATEGY:

- Compile a set of transcript-confirmed U12-dependent intron sequences
- Compute position weight matrices and compute log-likelihood matrices using intergenic false splice sites as background.
- Incorporate the new profiles into the `GeneID` gene prediction algorithm.
- With the new version of `GeneID`:
  - Score annotated introns for likelihood of being U12 versus U12-dependent
  - Correctly predict gene structures of genes that possess U12 introns

DATA SETS:

- 404 transcript-confirmed human U12 introns (GT-AG and AT-AC) from Levine and Durbin, 2001
- 27 introns from Burge et al., 1998

Summary

Pre-mRNA introns of most higher eukaryotic organisms can be classifi ed into two main classes, U2 and U12, according to the spliceosomal complex that excises them during RNA processing. More than 99% of eukaryotic introns are spliced by the U2-dependent spliceosome, while a minor class is spliced by the U12-dependent spliceosome. Differences between the two classes in the composition and degree of conservation at the splice sites and branch point reflect differences in base-pairing with snRNAs during spliceosome assembly.

Despite the strong consensus sequences characteristic of U12 introns, they have been consistently ignored by gene prediction programs to date. We have found that the hierarchical nature of the `GeneID` program architecture (splice site prediction ⇒ exon prediction ⇒ gene prediction) is amenable to the incorporation of multiple intron subtype profi les and, in addition, enables the explicit annotation of U12 introns in anonymous genomic sequences. We present statistics on the performance of the U12-enhanced `GeneID` on a test set of known U12-intron-possessing genes as well as results for scans of both the ENCODE regions and the entire human genome.

2. Position Weight Arrays

The kind of U12 splice signal profiles constructed for `GeneID` are called Position Weight Arrays. The matrix is addressed by nucleotide and position, where every cell contains a loglikelihood ratio between a Markov model (order k) recognizing true sites and another one, matching false ones. Thus, for every nucleotide in a candidate region, a score is computed that reflects the probability of finding the oligonucleotide (length k) ending in that nucleotide.

PWA Construction:

- Donors and Acceptors: for each class of U12 intron (GT-AG and AT-AC)
  - Construct frequency matrices for true splice sites and for false splice sites (intergenic GTs, ATs, AGs, or ACs)
  - Calculate log-likelihood matrices
- Branch points: for the combined set of GT-AG and AT-AC U12 introns
  - Construct 1st order Markov chains for U12 branch points and for random intronic and exonic 13-mers in the window -50 to +50 with respect to true 3' splice sites
  - Calculate the log-likelihood matrix

Information Content of U12 (GT-AG) Splice Signals

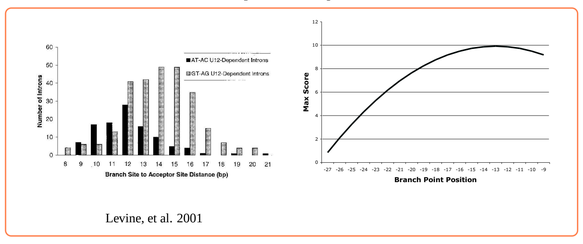
Donor	Branch	Acceptor
ATTGTATCCTTTTAGA	TACTTCCTTAACT	CTCTCCACAGCGG
ATTGTATCCTTTTAGA	CTTTTCCTTAACT	GGTTTATCAGGAT
TCCGTATCCTTTAAGAT	TTTTTCCTTAACT	TCAATGCCAGATC
ACTGTATCCTTGATGA	CCATTCCTTGACT	ATGTCTCTAGCAT
AGCGTATCCTTGAAAG	CGGTTCCTTAACA	TTTAAACTAGATT
GAAGTATCCTTATTCA	GTATTCCTTAACT	TGTCCTGTAGATG
ACTGTATCCTCTCGGG	AGTTTCCTTAACC	ACCCGCTCAGACA
GAAGTATCCTTCAGGG	TGGTTCCTTAACT	CTGGCTCCAGGCC
ACTGTATCCTTTTTCT	TTATTCCTTGACT	ACTAGCTGAGTCC
.	.	.
.	.	.
ATAGTATCCTTTTTAT	TCTTTCCTTGACC	TTCTCCCCAGACT
GCTGTATCCTTTGGCA	AGCCTCCTTAACC	AATTCTTCAGAGA
ACCGTATCCTTGATAT	GCTGTCCTTAACC	TTGGTGCGAGACG

3. Modifying GeneID

The logic of GeneID:

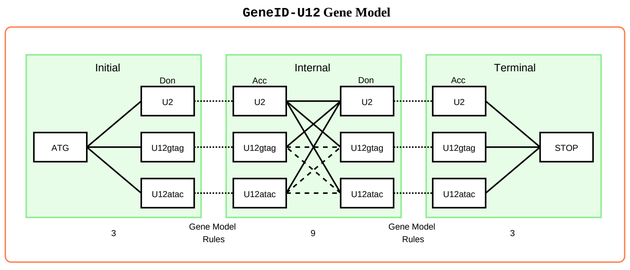
- Optional profiles (U12gtag Donor, U12gtag Acceptor, U12atac Donor, U12atac Acceptor, and U12 Branch) are read from the `GeneID` parameter file in addition to the standard U2 profiles.
- Splice Site Prediction
  - Donor sites are predicted using different profiles for each subtype.
  - Acceptor sites for each subtype are predicted. The U12 branch profile is used to find the best branch point sequence in the 40 bp upstream of each candidate U12 acceptor. The branch scores are weighted according to how close they come to the optimum branch to acceptor distance (12-13bp). The best branch site score is then added to the acceptor site score to arrive at a global acceptor score.
- Exon Prediction
  - Each combination of start, stop, donor and acceptor types is used to construct a set of all possible exons that meet a minimum coding potential score.
- Gene Prediction
  - The `Genamic` algorithm is used to find the highest scoring chain of compatible exons. Exon chaining rules are obtained from the gene model provided in the parameter file (the donor of the upstream exon must pair with a downstream acceptor of the same subtype.) Additional constraints on U12 exon-pair construction can be imposed by setting threshold splice signal and exon scores, as well as exon weighting.

Position-dependent scoring of branch site



Levine, et al. 2001

`GeneID-U12` Gene Model



4. Optimizing GeneID-U12

PARAMETER OPTIMIZATION:

The goal of this phase of `GeneID-U12` development was to evaluate the performance of different U12 profiles (0 order versus 1st order matrices), different branch site scoring methods, and different splice site and exon scoring parameters on a test set of genes known to possess U12 introns.

DATA SET:

Twenty genes were selected from the literature and extracted from Ensembl with 100 bp of flanking sequence upstream and downstream.

- 18 Ensembl genes with single U12 introns
- 2 Ensembl genes with two U12 introns

PROCEDURE:

- `GeneID-U12` was run on the set of 20 genes, varying a single parameter each time. Parameters were assumed to be independent.
- Two tests were performed on each set of predictions.
  - U12 splice sites: All predicted U12 splice sites were output and compared to the annotations. For each gene, the absolute and relative rank as well as the score of the true site with respect to the set of all predicted U12 sites was calculated.
  - U12 introns: All predicted U12 introns (in the context of full gene predictions) were compared to the annotations. The number of true positives, false negatives, and false positives were calculated.

RESULTS:

The results for the final version of `GeneID-U12` used in subsequent analyses are shown below. In general, the rankings of the true sites were high and the intron sensitivity and specificity were also very high. One caveat of these results is that the program was not run on intergenic sequence.

Performance at the splice site level:

	Donor				Acceptor			
Sequence	**Rank**	**Sites**	**Percentile**	**Score**	**Rank**	**Sites**	**Percentile**	**Score**
Chr10:93673616..93780158	1	40	100%	9.94	2	506	99.8%	9.73
Chr10:99185894..99195839	2	2	50.0%	7.58	1	44	100%	6.64
Chr11:118460703..118469567	1	2	100%	9.60	1	36	100%	8.77
Chr11:45887697..45896282	1	2	100%	7.62	2	21	95.2%	6.38
Chr17:51183313..51209833	1	10	100%	9.85	6	116	95.7%	6.11
Chr1:227189765..227244584	2	7	85.7%	8.86	7	212	97.2%	7.06
Chr1:238337476..238379188	1	7	100%	9.46	1	187	100%	10.81
Chr1:32357005..32366397	1	5	100%	7.62	3	30	93.3%	8.44
Chr2:216797596..216896609	1	27	100%	10.04	26	493	95.0%	6.40
Chr3:157741147..157755737	1	4	100%	9.26	7	89	93.3%	6.79
Chr3:180763310..180788990	1	7	100%	7.88	1	94	100%	10.43
Chr3:198150489..198157810	1	4	100%	7.70	1	41	100%	10.39
Chr5:140051102..140064590	1	2	100%	10.04	2	63	98.4%	7.14
Chr5:41765825..41906460	6	63	92.1%	7.72	1	698	100%	11.44
Chr6:88356462..88433985	1	20	100%	9.50	3	382	99.5%	9.74
Chr8:120812096..120914085	1	30	100%	9.12	1	461	100%	10.75
Chr8:30133255..30160701	1	3	100%	9.46	30	111	73.9%	4.82
Chr9:4782769..4851164	1	11	100%	9.73	3	305	99.3%	9.40

Performance at the intron level:

23 predicted U12 introns

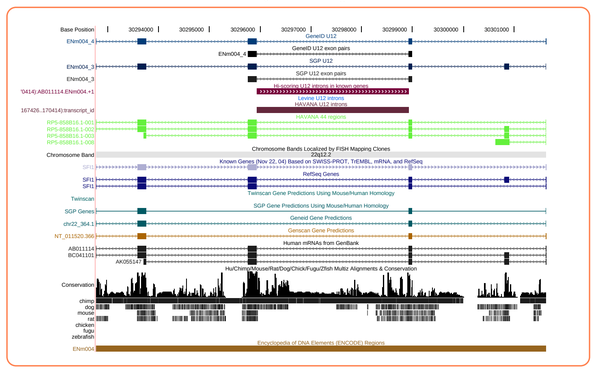
- 22 true positives
- 1 false positive
- 0 false negatives
- 100% sensitivity, 96% specifi city

5. Predicting Genes in ENCODE and the Entire Human Genome

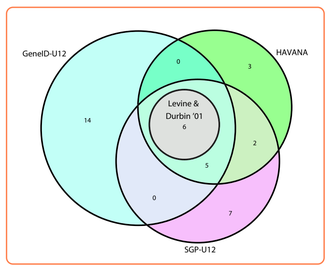
ENCODE:

44 regions of the human genome were selected for analysis by the ENCODE consortium. The HAVANA team has completed a thorough human-labor intensive annotation of these 44 regions using a battery of evidence, including RT-PCR validation of intron-exon junctions. `GeneID-U12` was run in *ab initio* mode or with the additional input of HSPs resulting from `tblastx` on syntenic mouse sequence (`SGP2-U12` mode).

Example: the SFI1 gene



U12 Introns in ENCODE



FALSE NEGATIVES:

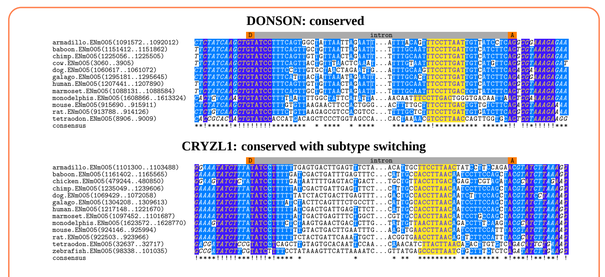
HAVANA annotations include several U12 introns not predicted by `GeneID-U12` or `SGP2-U12` gene prediction

- 3 genes with low-coding potential and low similarity across species: CTAG1A/B and CTAG2
- 1 intron in the UTR of RefSeq mRNA BC066967 (but within CDS of VEGA_Novel_transcript: AC018512.2-002, intron is RT-positive)
- 3 additional alternative acceptor sites for a U12 intron in CRYZL1 - the major isoform of which is predicted correctly by `GeneID-U12` and `SGP2-U12`
- 1 U12 intron in a minor transcript of c21orf66 - the major isoform of which is predicted correctly by `GeneID-U12` and `SGP2-U12`

TRUE vs. FALSE POSITIVES:

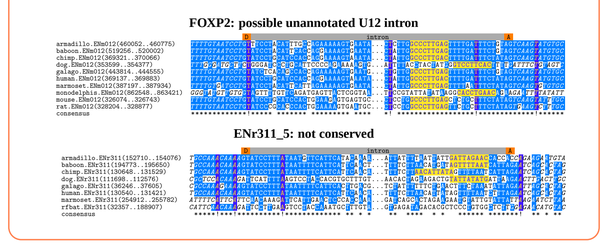
Conservation of intron positions in several divergent species is used to determine whether or not a predicted intron is "true" or "false." Predicted introns not annotated by HAVANA do not seem to be conserved across species.

DONSON: conserved



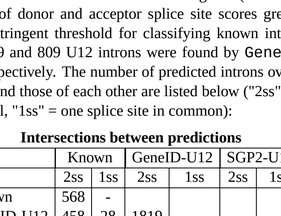
CRYZL1: conserved with subtype switching

FOXP2: possible unannotated U12 intron



ENr311_5: not conserved

U12 Intron Scores for Known Genes in hg17



THE HUMAN GENOME:

Predictions of genes on the human genome were made with `GeneID-U12` and `SGP2-U12`. `GeneID-U12` was also used to score the splice sites of all non-redundant introns in the CDS of known genes (UCSC known gene track.) A sum of donor and acceptor splice site scores greater than 9.0 was used as a stringent threshold for classifying known introns as U12-dependent. 1819 and 809 U12 introns were found by `GeneID-U12` and `SGP2-U12`, respectively. The number of predicted introns overlapping annotated introns and those of each other are listed below ("2ss" = both splice sites are identical, "1ss" = one splice site in common):

Intersections between predictions

	Known		GeneID-U12		SGP2-U12	
	2ss	1ss	2ss	1ss	2ss	1ss
Known	568	-				
GeneID-U12	458	28	1819	-		
SGP2-U12	475	15	544	38	809	-

Bibliography

[Abril *et al.*, 2005] Abril,J.F., Castelo,R. and Guigo,R. (2005) Comparison of splice sites in mammals and chicken. *Genome Res*, **15** (1), 111–119.

[Burge *et al.*, 1998] Burge,C.B., Padgett,R.A. and Sharp,P.A. (1998) Evolutionary fates and origins of U12-type introns. *Mol Cell*, **2** (6), 773–785.

[Clark and Thanaraj, 2002] Clark,F. and Thanaraj,T.A. (2002) Categorization and characterization of transcript-confi rmed constitutively and alternatively spliced introns and exons from human. *Hum Mol Genet*, **11** (4), 451–464.

[Consortium., 2004] Consortium.,E.P. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306** (5696), 636–640.

[Dietrich *et al.*, 1997] Dietrich,R.C., Incorvaia,R. and Padgett,R.A. (1997) Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns. *Mol Cell*, **1** (1), 151–160.

[Dietrich *et al.*, 2001] Dietrich,R.C., Peris,M.J., Seyboldt,A.S. and Padgett,R.A. (2001) Role of the 3' splice site in U12-dependent intron splicing. *Mol Cell Biol*, **21** (6), 1942–1952.

[Levine and Durbin, 2001] Levine,A. and Durbin,R. (2001) A computational scan for U12-dependent introns in the human genome sequence. *Nucleic Acids Res*, **29** (19), 4006–4013.

[McConnell *et al.*, 2002] McConnell,T.S., Cho,S.J., Frilander,M.J. and Steitz,J.A. (2002) Branchpoint selection in the splicing of U12-dependent introns in vitro. *RNA*, **8** (5), 579–586.

[Parra *et al.*, 2003] Parra,G., Agarwal,P., Abril,J.F., Wiehe,T., Fickett,J.W. and Guigo,R. (2003) Comparative gene prediction in human and mouse. *Genome Res*, **13** (1), 108–117.

[Parra *et al.*, 2000] Parra,G., Blanco,E. and Guigo,R. (2000) GeneID in Drosophila. *Genome Res*, **10** (4), 511–515.

[Sharp and Burge, 1997] Sharp,P.A. and Burge,C.B. (1997) Classifi cation of introns: U2-type or U12-type. *Cell*, **91** (7), 875–879.

[Tarn and Steitz, 1996] Tarn,W.Y. and Steitz,J.A. (1996) A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. *Cell*, **84** (5), 801–811.

[Zhu and Brendel, 2003] Zhu,W. and Brendel,V. (2003) Identifi cation, characterization and molecular phylogeny of U12-dependent introns in the Arabidopsis thaliana genome. *Nucleic Acids Res*, **31** (15), 4561–4572.

Conclusions

1. The distinct consensus sequences and high information content of U12 splice signals are suffi cient for discrimination of U12 and U2 introns.
2. The modular architecture of `GeneID` is amenable to the incorporation of multiple splice signal profi les, although in its current implementation, exon types are predetermined (i.e. hard-coded).
3. It is possible to increase gene prediction sensitivity to minor intron subtypes without affecting specifi city signifi cantly.
4. While no clear novel U12 introns were found in the ENCODE regions, our method was useful for classifying annotated introns as U12 or U2. 0.3% of introns were classifi ed as U12. Moreover, at the gene level, we found that 3.3% of the 487 known and novel genes annotated by HAVANA possess U12 introns. This signifi cant percentage underscores the importance of our effort to accurately predict and annotate these minor intron subtypes.
5. When taking the intersection of the `GeneID-U12` and `SGP2-U12` predictions on the ENCODE regions, the specifi city increases dramatically with little decrease in sensitivity. When an additional fi lter for cross-species conservation of intron junctions is applied, the results become even more conclusive. We plan to apply these same criteria to our analysis of the entire human genome in order to identify novel U12 introns not necessarily supported by transcript evidence.
6. Our method does not currently predict U12 introns in UTRs, nor does it predict more than one isoform per locus. Future work will address these limitations.

Acknowledgments

We are grateful to the GRIB academic staff, especially Enrique Blanco and Francisco Cámara, for helpful discussions regarding the `GeneID` source code and training scripts. Tyler Alioto is supported by Fundació IMIM. This work has been supported by a grant from the Ministerio de Ciencia y Tecnologia (Spain).

* To whom correspondence should be addressed:

Grup de Recerca en Informàtica Biomèdica

IMIM - Universitat Pompeu Fabra - Centre de Regulació Genòmica (CRG)

Pg. Marítim de la Barceloneta 37-49, 08003 - Barcelona (Spain).

Phone: +34 93 224 0886

e-mail: **talioto@imim.es**