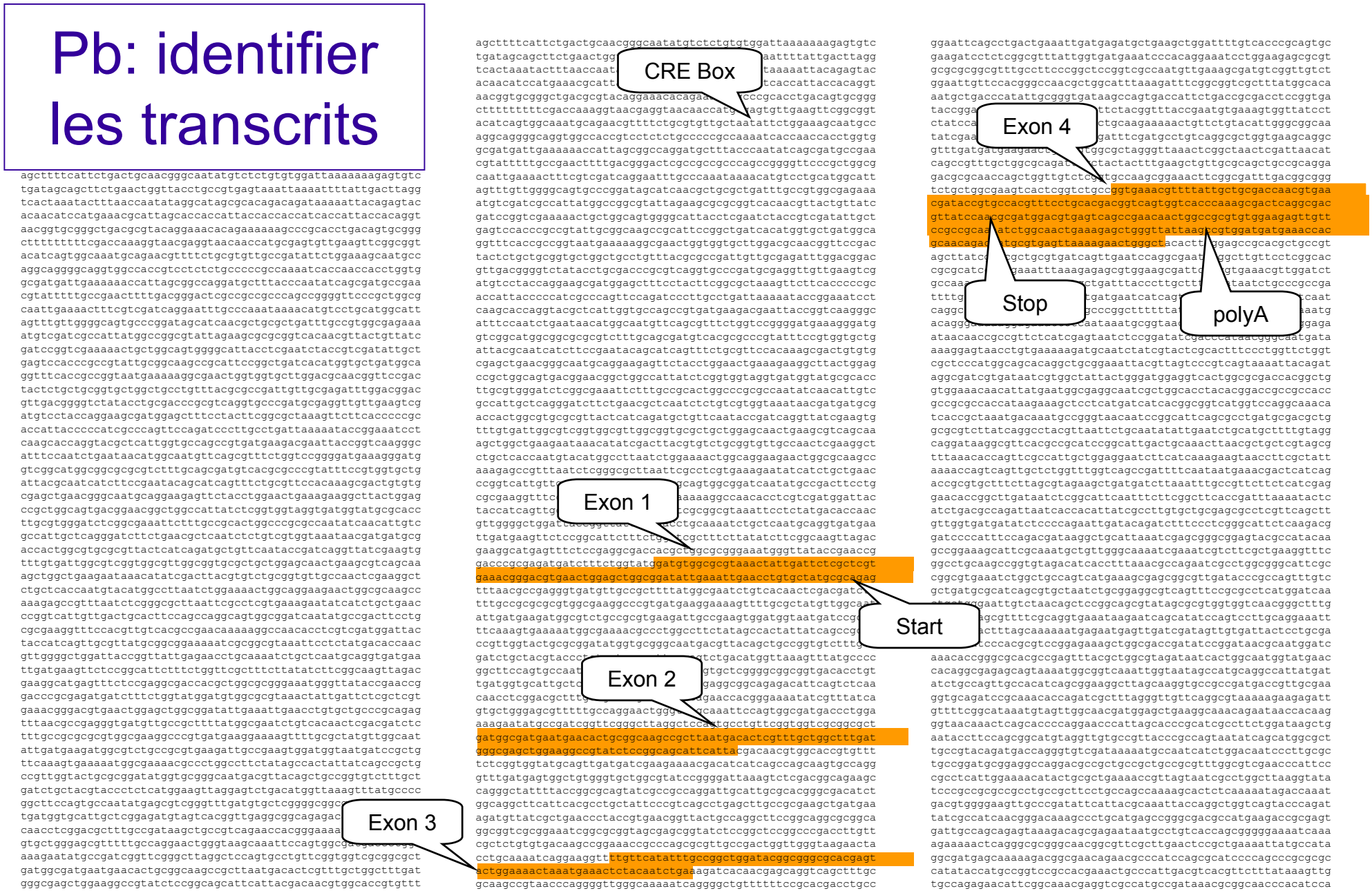
Analyse des transcrits et des transcrits alternatifs par les EST

Daniel Gautheret, 2003
ESIL, Université de la Méditerranée

V.2003.2

Pb: identifier les transcrits



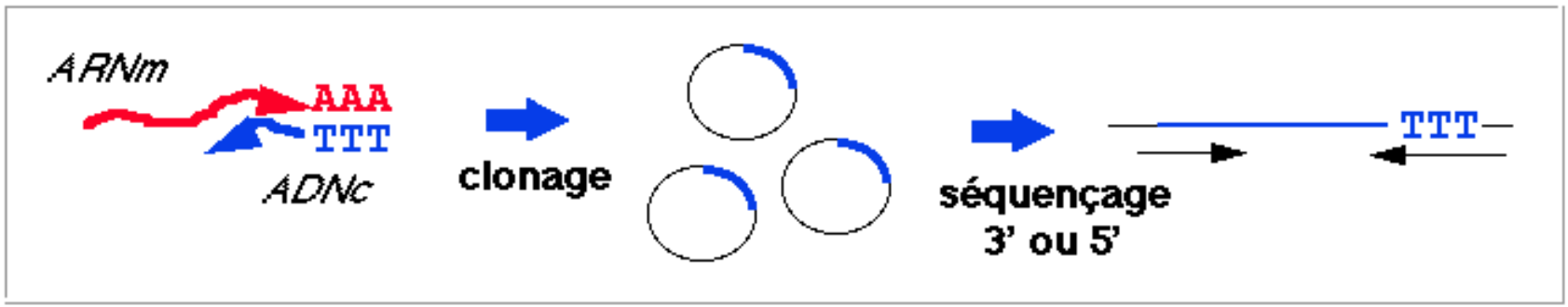
Signaux des gènes eucaryotes

- ★ **Boite TATA et autres signaux promoteur**
- ★ **Ilot CpG**
- ★ **Jonctions introns-exons**
- ★ **Signal polyA**
- ★ **Cadre de lecture**
- ★ **Composition en codons/hexamères**
- ★ **…**

Les signaux sont insuffisants pour la prédiction des transcrits

★ Même avec les progrès statistiques (consensus->profils->HMM), le recours aux données expérimentales reste un must

EST (expressed sequence tags)



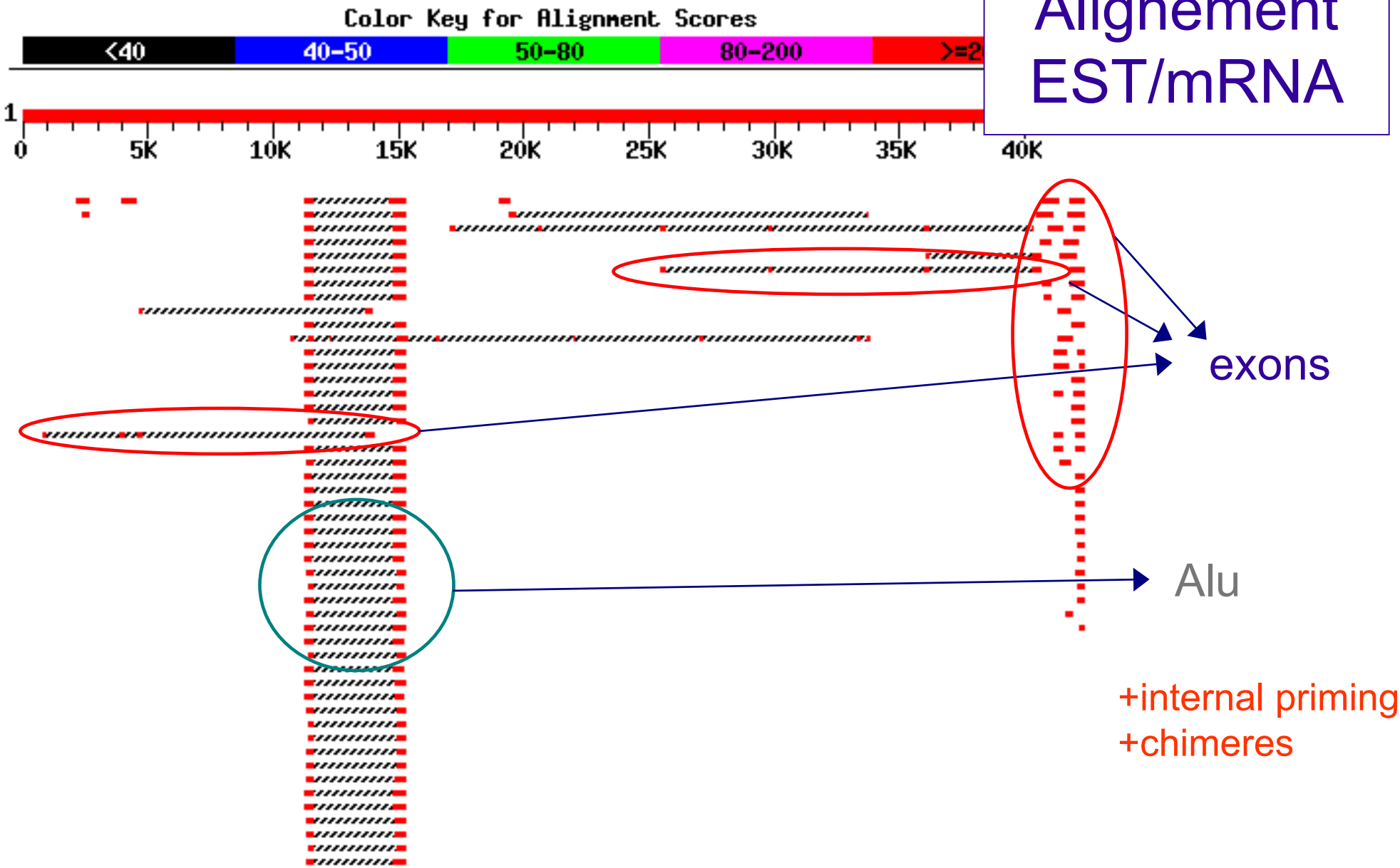
Normalisation

- Pour éviter de réamplifier sans cesse les transcrits les plus fréquents: réhybridation (normalisation) ou hybrydation contre bibliothèque de référence (soustraction).

Limitations de l'approche EST

- Chimérisme , Clônes inversés, Priming interne, Rétention d'introns, Epissage alternatif

Alignement EST/mRNA



+internal priming
+chimeres

EST data: state of the art

★ **dbEST release 100303 Summary by Organism - October 3, 2003**

★ **Number of public entries: 18,762,324**

- Homo sapiens (human) 5,426,001
- Mus musculus + domesticus (mouse) 3,881,878
- Rattus sp. (rat) 538,073
- Triticum aestivum (wheat)
- 500,898 Ciona intestinalis
- 492,488 Gallus gallus (chicken)
- 451,565 Zea mays (maize) 383,416
- Danio rerio (zebrafish) 362,362
- Hordeum vulgare + subsp. vulgare (barley) 348,233
- Xenopus laevis (African clawed frog) 344,695
- Glycine max (soybean) 341,573
- Bos taurus (cattle) 322,074
- Drosophila melanogaster (fruit fly) 261,404

- Human: 4445 libraries (623 cancer)
- Mouse: 441 libraries (23 cancer)

Alignement EST/mRNA

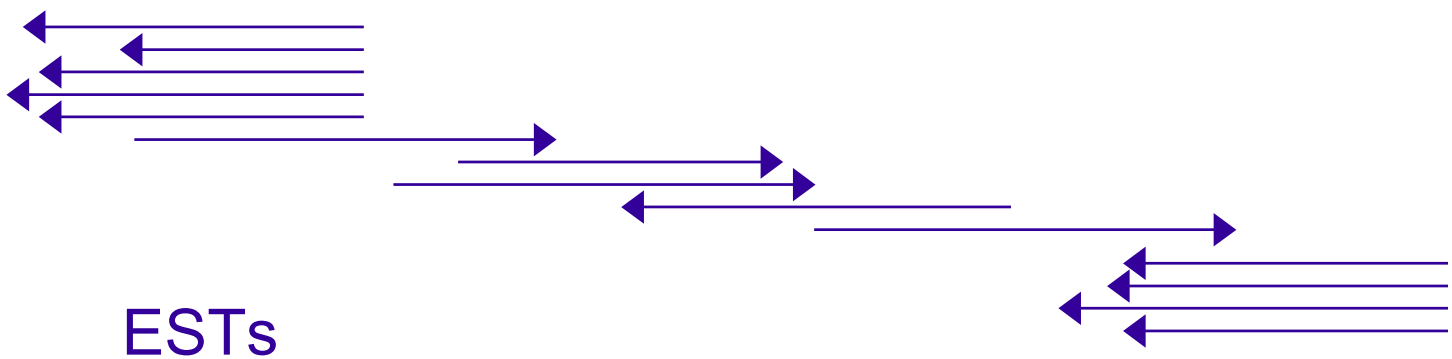
1998: analyse de clusters d'EST…

Table 2. Characteristics of the 1000 Largest Clusters

Clusters matching a GenBank primate sequence (%)[a]	Average no. of contigs per cluster	ESTs assigned to internal priming (%)[b]	Clusters exhibiting poly(A) sites (no.)		
			2	3	4
72.7	1.6	13.9	159	27	3

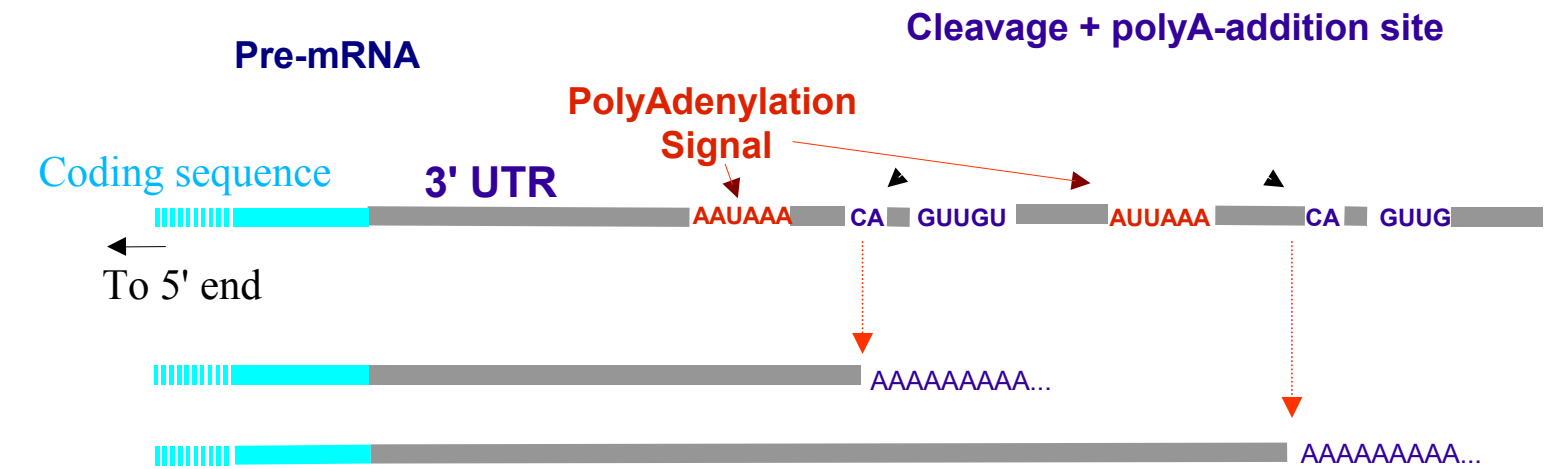
[a]With a BLAST score of 150 or higher.
[b]ESTs are considered as originating from internal priming when followed in the contig by a stretch of 6 A's or more, or by a 10 nucleotide sequence containing 7 A's or more. Only ESTs that fully align with their corresponding contig (see Methods for details) are considered.

contig



ESTs

La polyadénylation alternative



mRNA alternative transcripts

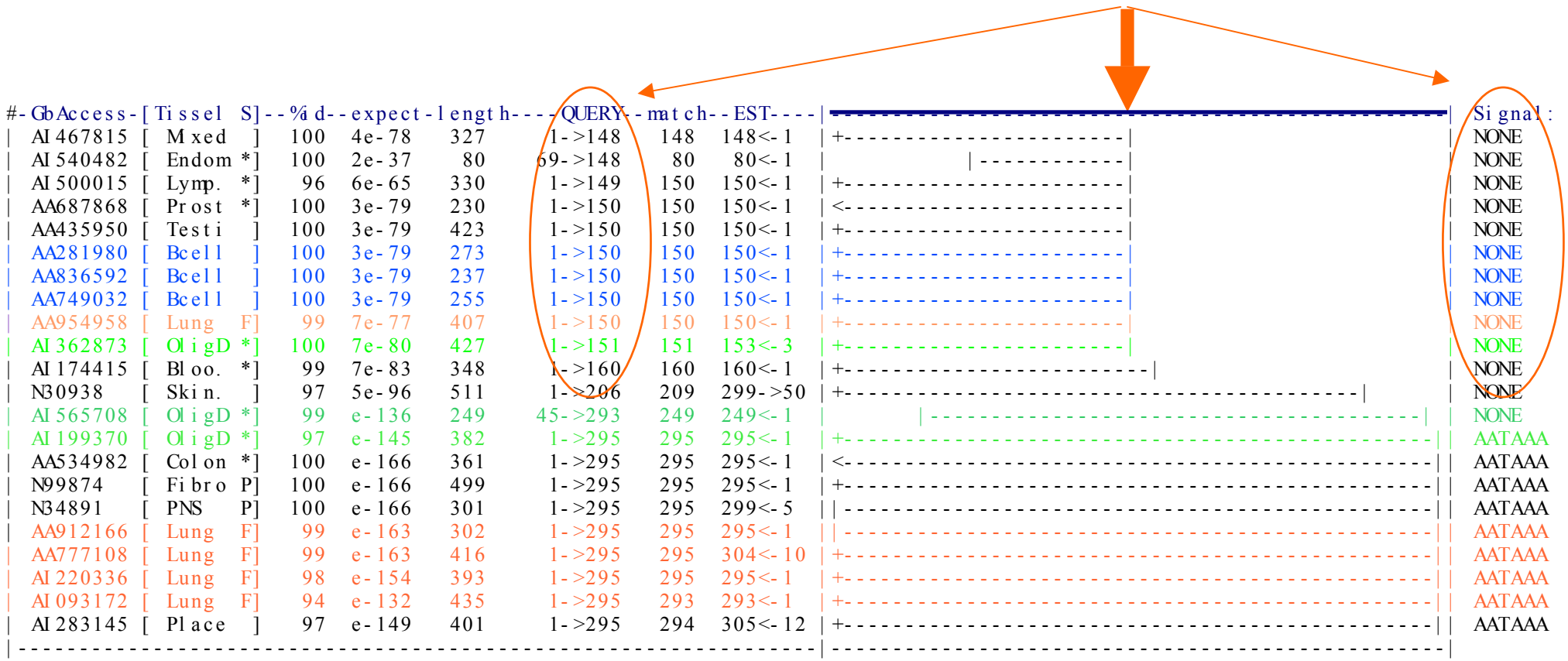
Des signaux poly(A) insuffisants pour constituer une signature

AAUAAA: court et non obligatoire, variants

Région riche en GU: plusieurs motifs possibles, non obligatoire (<60%)

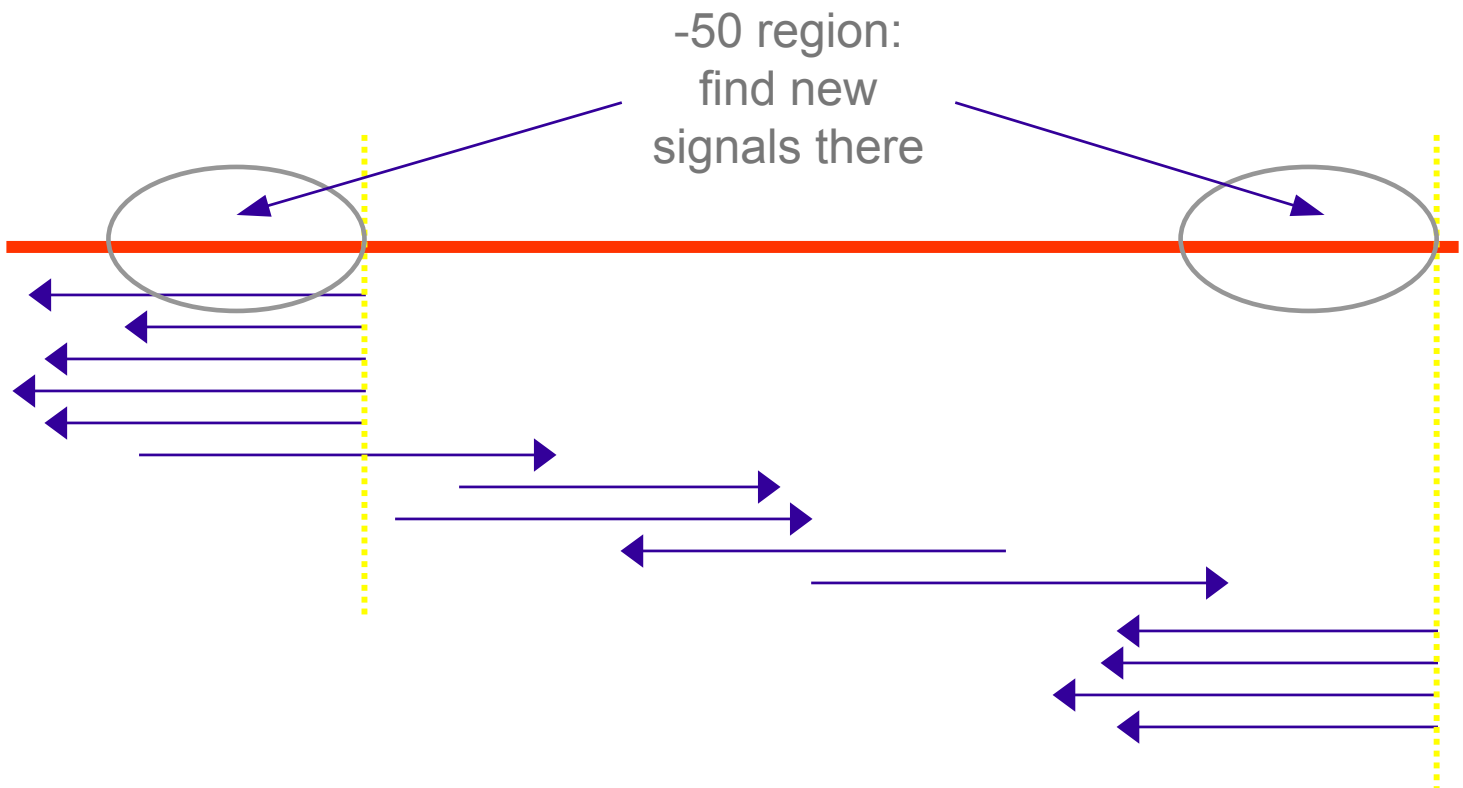
La meilleure technique de prédiction demeure le recours aux données expérimentales

Identification des variants de PA par les EST



#-GbAccess-	[Tissel	S]-	%id-	expect-	length-	QUERY-	match-	EST-	Signal:
AI467815	[Mixed	]	100	4e-78	327	1->148	148	148<-1	NONE
AI540482	[Endom	*]	100	2e-37	80	69->148	80	80<-1	NONE
AI500015	[Lymp.	*]	96	6e-65	330	1->149	150	150<-1	NONE
AA687868	[Prost	*]	100	3e-79	230	1->150	150	150<-1	NONE
AA435950	[Testi	]	100	3e-79	423	1->150	150	150<-1	NONE
AA281980	[Bcell	]	100	3e-79	273	1->150	150	150<-1	NONE
AA836592	[Bcell	]	100	3e-79	237	1->150	150	150<-1	NONE
AA749032	[Bcell	]	100	3e-79	255	1->150	150	150<-1	NONE
AA954958	[Lung	F]	99	7e-77	407	1->150	150	150<-1	NONE
AI362873	[OligD	*]	100	7e-80	427	1->151	151	153<-3	NONE
AI174415	[Bloo.	*]	99	7e-83	348	1->160	160	160<-1	NONE
N30938	[Skin.	]	97	5e-96	511	1->206	209	299->50	NONE
AI565708	[OligD	*]	99	e-136	249	45->293	249	249<-1	NONE
AI199370	[OligD	*]	97	e-145	382	1->295	295	295<-1	AATAAA
AA534982	[Colon	*]	100	e-166	361	1->295	295	295<-1	AATAAA
N99874	[Fibro	P]	100	e-166	499	1->295	295	295<-1	AATAAA
N34891	[PNS	P]	100	e-166	301	1->295	295	299<-5	AATAAA
AA912166	[Lung	F]	99	e-163	302	1->295	295	295<-1	AATAAA
AA777108	[Lung	F]	99	e-163	416	1->295	295	304<-10	AATAAA
AI220336	[Lung	F]	98	e-154	393	1->295	295	295<-1	AATAAA
AI093172	[Lung	F]	94	e-132	435	1->295	293	293<-1	AATAAA
AI283145	[Place	]	97	e-149	401	1->295	294	305<-12	AATAAA

Analyse des signaux associés aux sites PA



Hexamères les plus fréquents

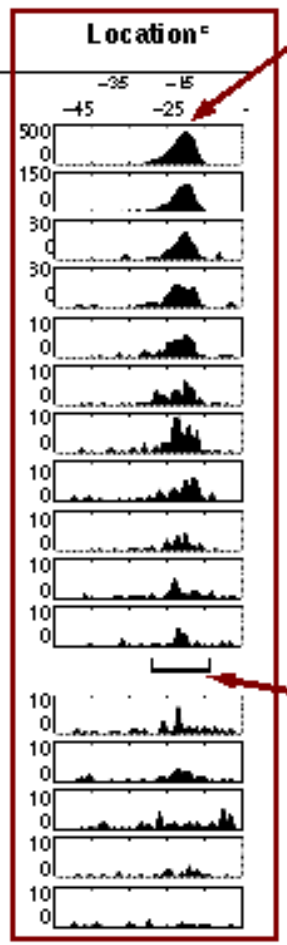
The AAUAAA and AUUAAA hexamers are the most frequently found.

26.8% of the 3'' fragments do not contain a usual polyadenylation signal.

Ten variant motifs are found accounting for 14.9% of the putative mRNA 3' ends

Most significant hexamers in 3' fragments

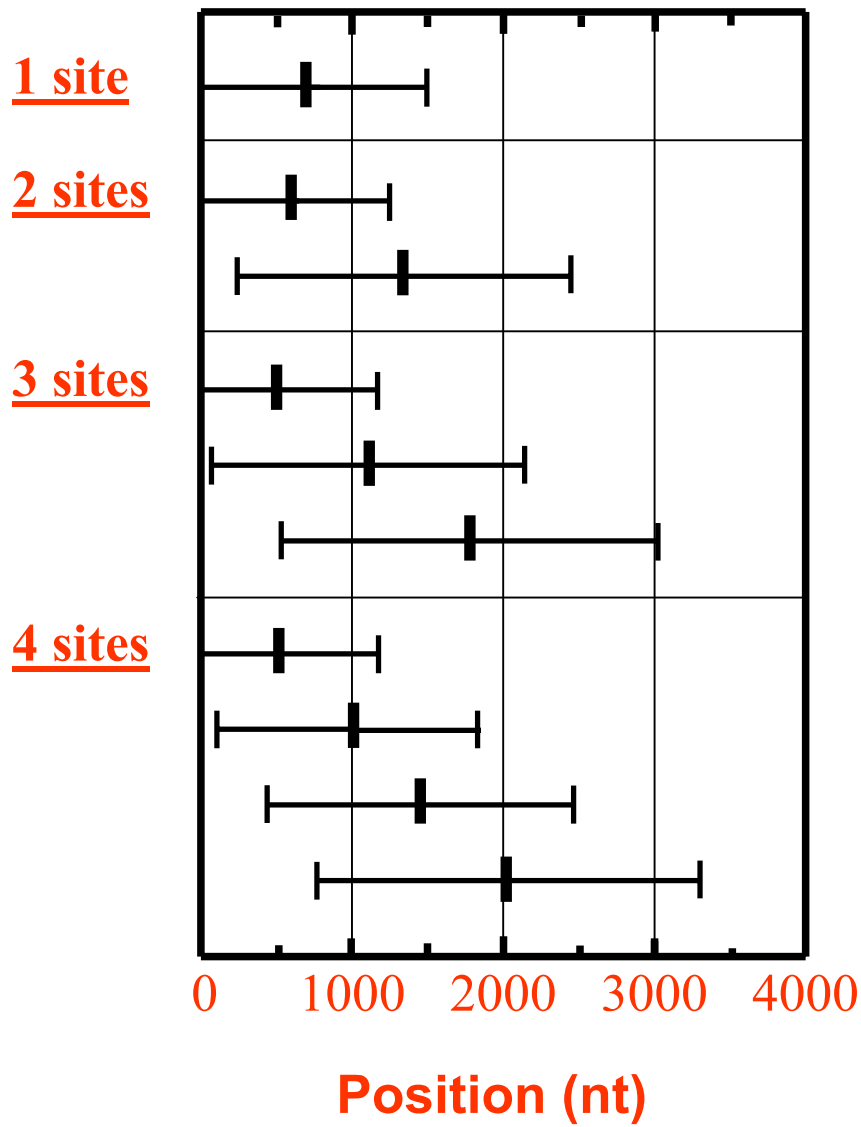
Hexamer	Observed (expected)[a]	% sites	P[b]	Position ave +\- SD	Location[c]
From Clustered Hexamers					
AAUAAA	3286 (317)	58.2	0	−16 +/− 4.7	
AUUAAA	843 (112)	14.9	0	−17 +/− 5.3	
AGUAAA	156 (32)	2.7	6×10^{-57}	−16 +/− 5.9	
UAUAAA	180 (53)	3.2	4×10^{-48}	−18 +/− 7.8	
CAUAAA	76 (23)	1.3	1×10^{-25}	−17 +/− 5.9	
GAUAAA	72 (21)	1.3	2×10^{-25}	−18 +/− 6.9	
AAUAUA	96 (33)	1.7	2×10^{-19}	−18 +/− 6.9	
AAUACA	70 (16)	1.2	5×10^{-23}	−18 +/− 8.7	
AAUAGA	43 (14)	0.7	1×10^{-09}	−18 +/− 6.3	
AAAAAG	49 (11)	0.8	5×10^{-17}	−18 +/− 8.9	
ACUAAA	36 (11)	0.6	1×10^{-05}	−17 +/− 8.1	
From Scattered Hexamers					
AAGAAA	62 (10)	1.1	9×10^{-25}	−19 +/− 11	
AAUGAA	49 (10)	0.8	4×10^{-15}	−20 +/− 10	
UUUAAA	69 (20)	1.2	3×10^{-15}	−17 +/− 12	
AAAACA	29 (5)	0.5	8×10^{-12}	−20 +/− 10	
GGGGCU	22 (3)	0.3	9×10^{-12}	−24 +/− 13	



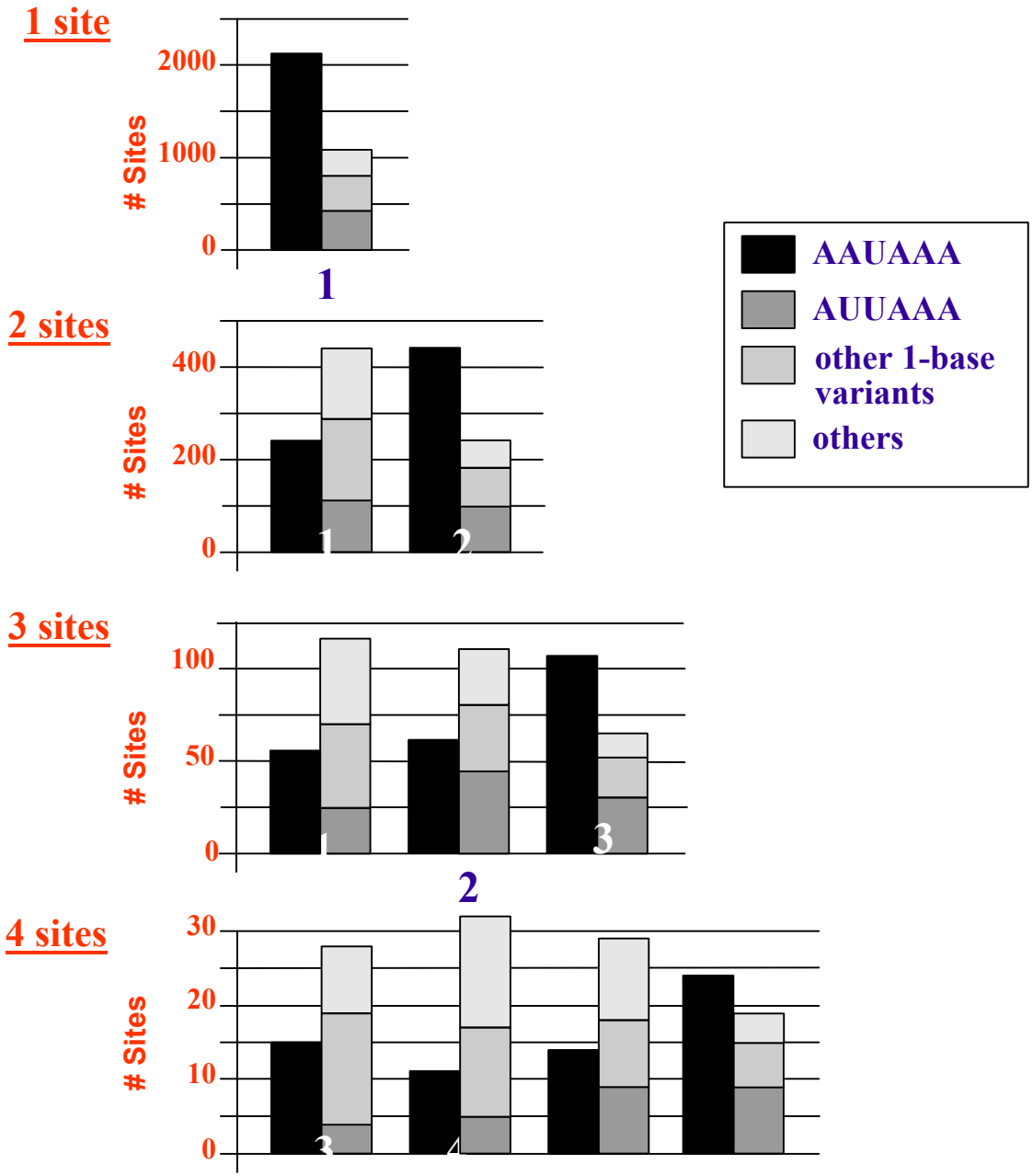
As expected from experimentally validated signals, AAUAAA and AUUAAA hexamers are clearly clustered around −15/−16 nt upstream of the putative poly(A) site.

Variant hexamers are also clustered around positions −15/−20.

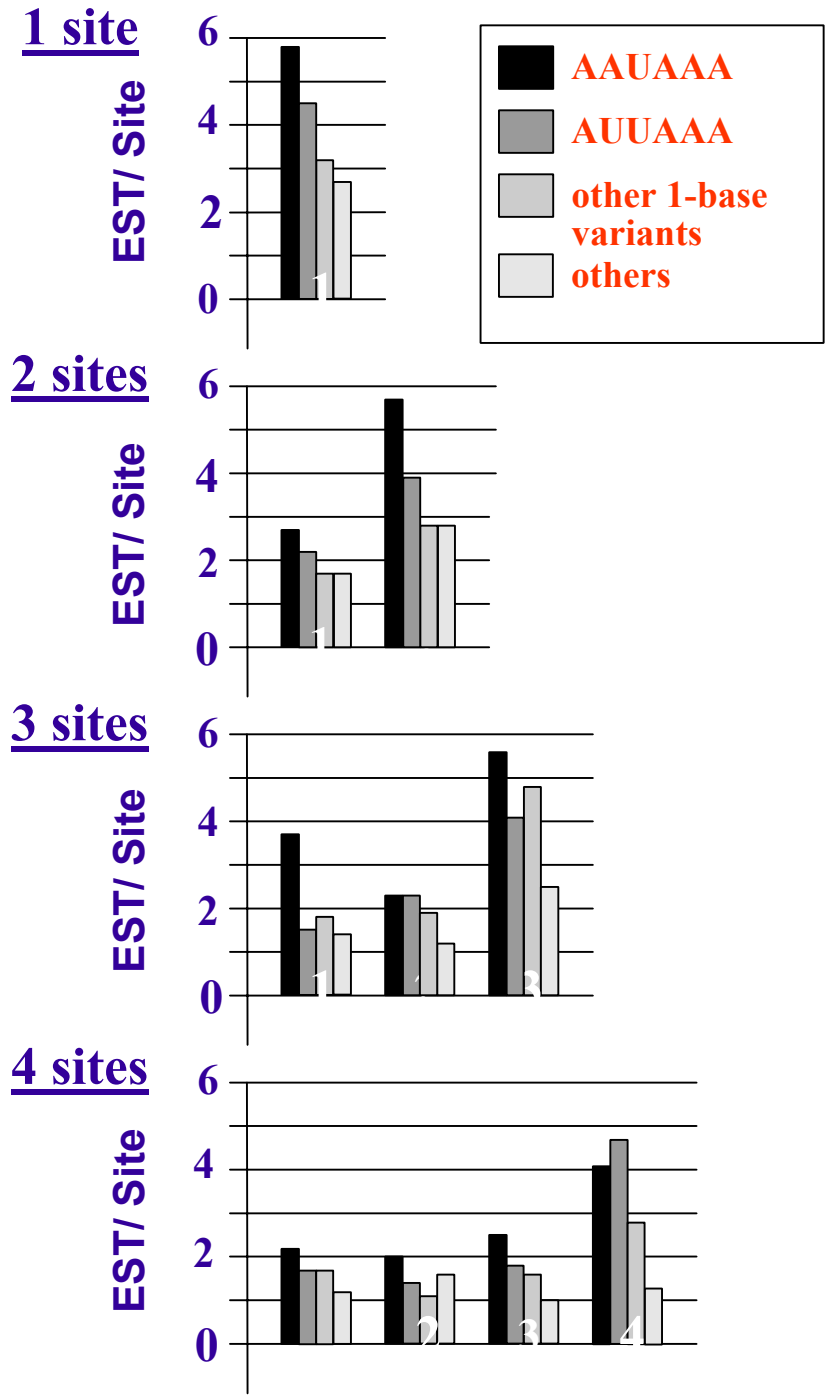
Position des sites de polyadenylation



Nombre de sites

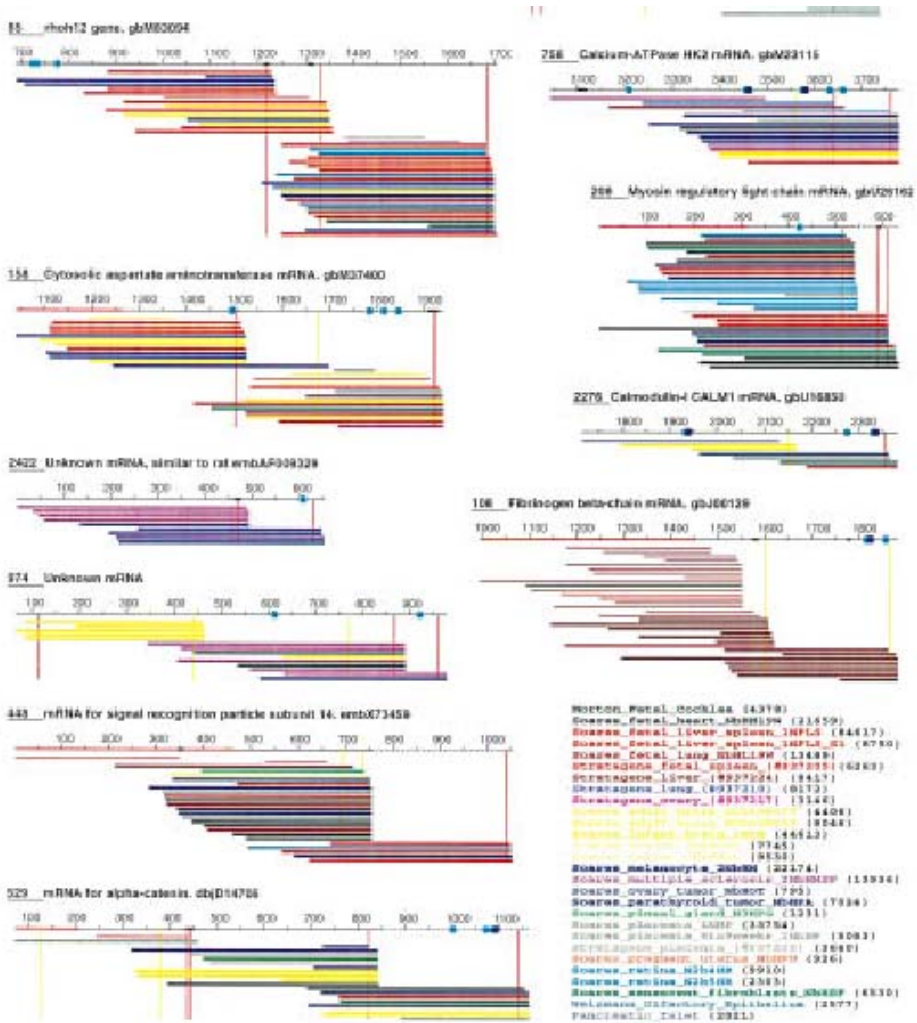


Efficacité des signaux



Tissue-specific polyadenylation

Article de 98:



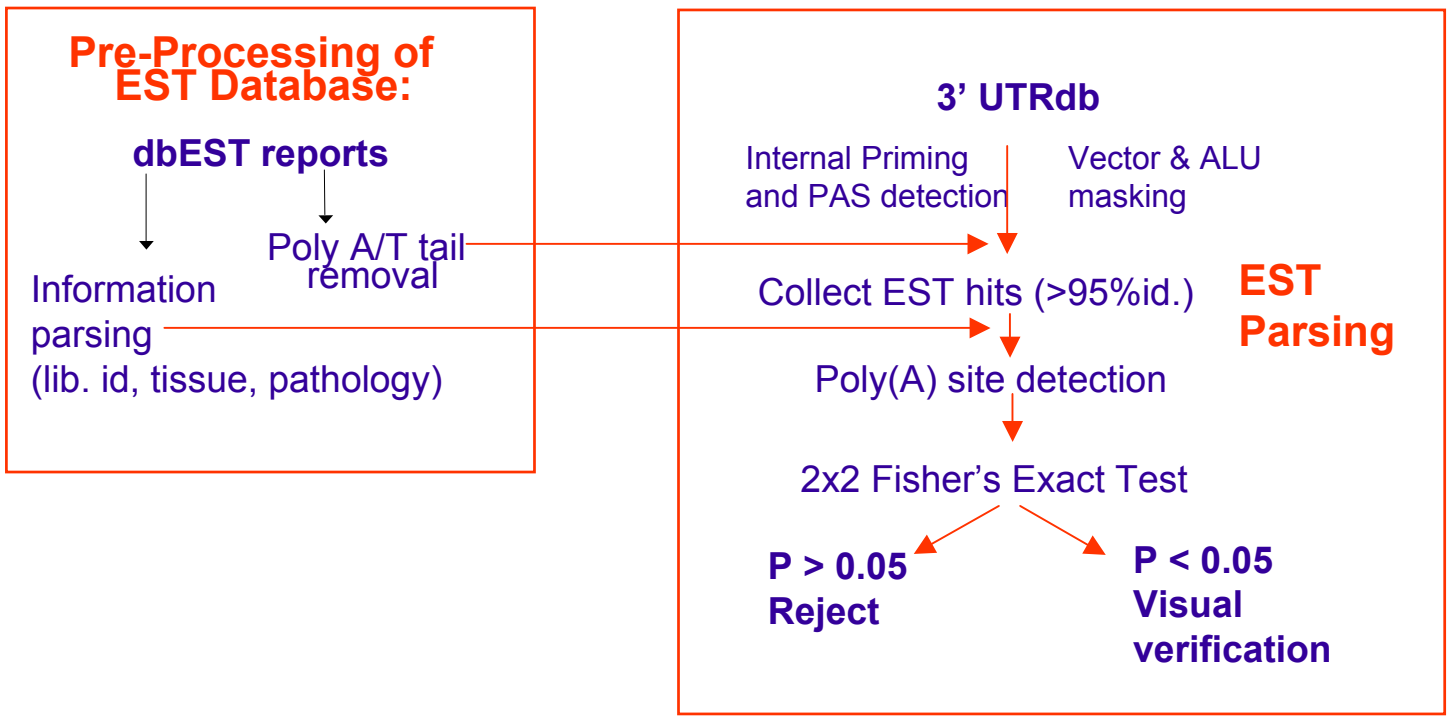
Clusters of 38 ESTs aligned with their respective contigs *(top* line of each cluster). Contigs annotated with a GenBank entry name can be considered as identical to the corresponding mRNA (BLAST2 score ù 2000, 97%– 99% identity over highest scoring segment). (Unknown mRNA) Contigs do not show any significant resemblance (BLAST score ù150 or *P* ù 0.02) to a human non-EST sequence in GenBank release 104. mRNA extensions are shown with broken lines. Contigs that do not extend corresponding mRNAs are numbered from the mRNA 58 end; other contigs are numbered from position 1. Thicker segments in contigs indicate possible internal priming sites (see Methods). Potential destabilization signals are shown with blue and dark blue dots, corresponding to sequences AUUUA and UUAUUUA(U/A)(U/A), respectively. ESTs are colored according to their source library, as indicated at *bottom right.* Numbers in parentheses indicate the total number of ESTs in each library. Red lines on contigs indicate coding sequences. Vertical red and yellow lines give the positions of all AAUAAA and AUUAAA sequences, among which are the actual polyadenylation signals (see text). Only ESTs that fully match their respective contig are shown.Clusters are numbered according to the number of ESTs they contain (1 is largest).

4500 human EST libraries

```
>431
Lib Name:        Soares mouse p3NMF19.5
Organism:        Mus musculus
Develop. stage:  19.5 dpc total fetus
Lab host:        DH10B (ampicillin resistant)
Vector:          pT7T3D (Pharmacia) with a modified polylinker
R. Site 1:       Not I
R. Site 2:       Eco RI
Description:     1st strand cDNA was primed with a Not I - oligo(dT) primer [5'
                 TGTTACCAATCTGAAGTGGGAGCGGCCGCATTTTTTTTTTTTTTTTTTT 3'],
                 double-stranded cDNA was size selected, ligated to Eco RI
                 adapters (Pharmacia), digested with Not I and cloned into the
                 Not I and Eco RI sites of a modified pT7T3 vector (Pharmacia).
                 Library went through one round of normalization to a Cot = 5.
                 Library constructed by Bento Soares and M.Fatima Bonaldo. RNA
                 was kindly provided by Dr. Minoru Ko (Wayne State University).


>432
Lib Name:        Human retina cDNA Tsp509I-cleaved sublibrary
Organism:        Homo sapiens
Sex:             mixed (males and females)
Organ:           eye
Tissue type:     retina
Develop. stage:  adult
Lab host:        E. coli strain K802
Vector:          lambda gt10
R. Site 1:       EcoRI
R. Site 2:       EcoRI
Description:     The library used for sequencing was a sublibrary derived from
                 a human retina cDNA library. Inserts from retina cDNA library
                 DNA were isolated, cleaved with Tsp509I, size selected, and
                 cloned into lamda gt10. Individual plaques were arrayed and
                 used as templates for PCR amplification and these PCR products
                 were used for sequencing.
```

EST parser algorithm

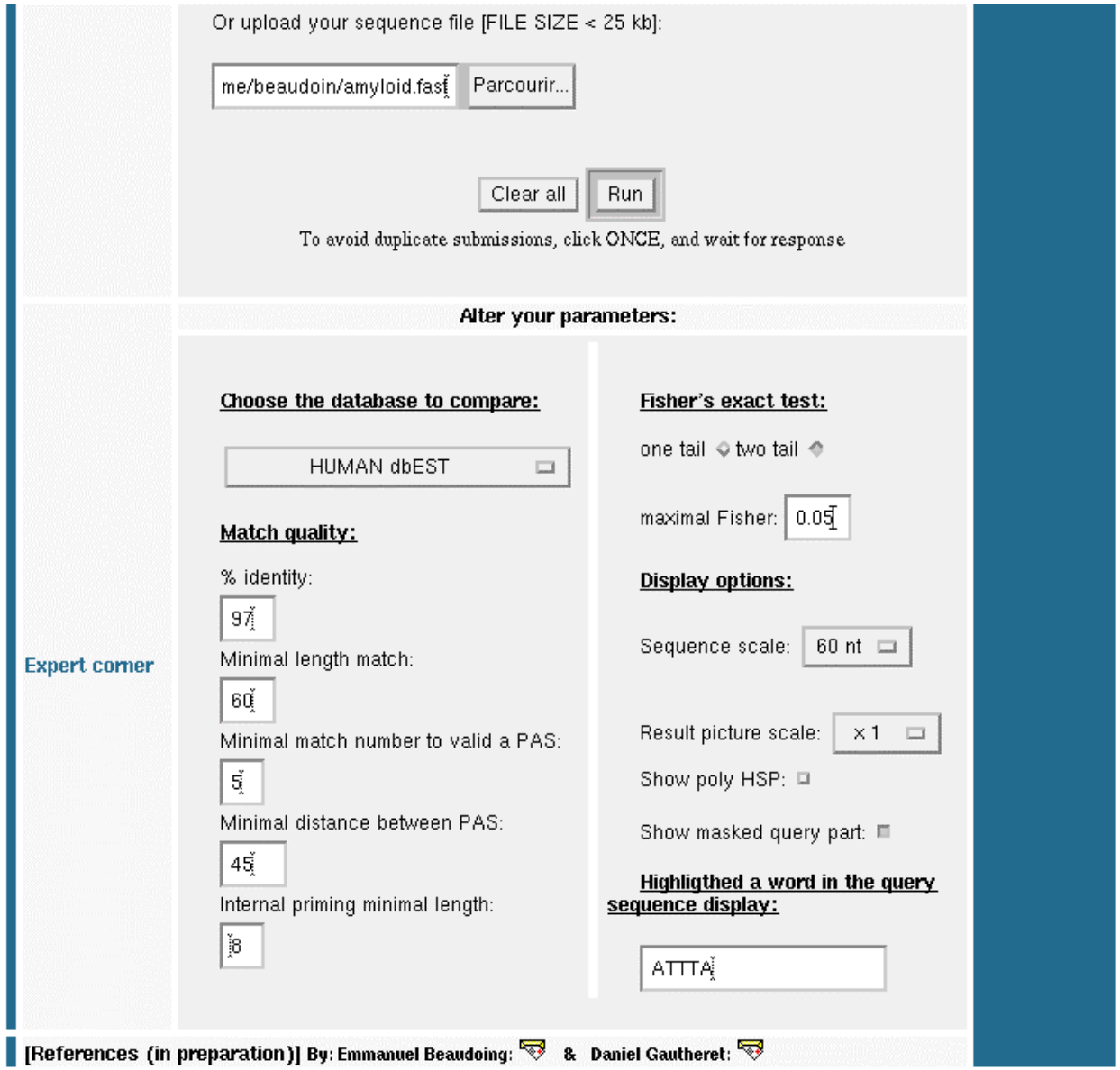


Human and mouse dbEST usage repartition on UTRdb

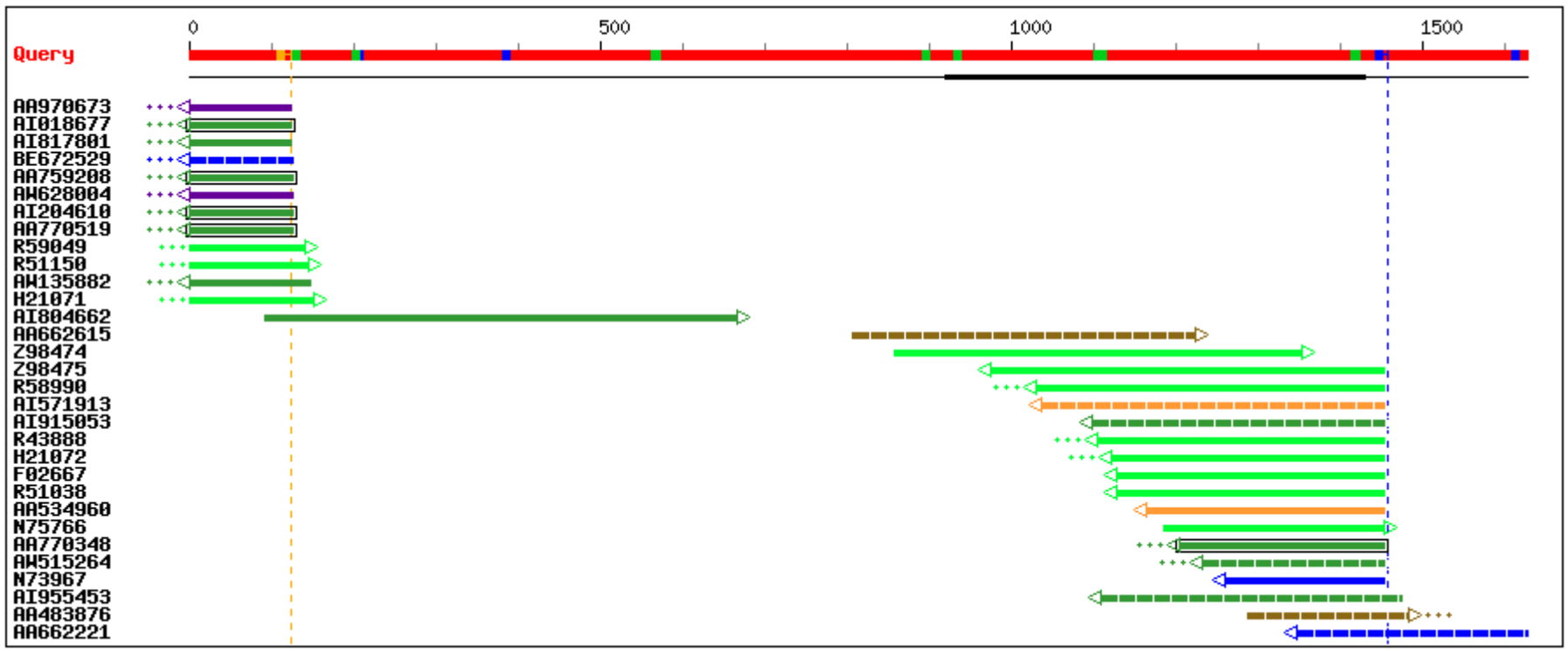
ORGAN SYSTEM	MAIN TISSUE	LIB REPARTITION			dbEST REPARTITION			EST matching known UTR		
		Homo	Mus	Total	Homo	Mus	Total	Homo	Mus	Total
Central Nervous	ear, eye, retina, RPE, olfact, any part of brain, etc.	271	72	343	304,661	248,959	553,620	127,1476	76,047	127,476
Peripheral Nervous	nervous, oblongata, spinal, etc.	285	9	294	43,141	60,429	103,570	15,695	18,355	34,050
Digestive	colon, intestin, liver, pancreas, gallbladder, stomach, esophagus, mouth, etc.	794	37	831	413,753	103,705	517,458	171,318	32,372	203,690
Respiratory	bronchi, larynx, lung, pharynx, throat, etc.	147	12	159	192,624	41,182	233,806	86,673	12,382	99,055
Uro-genital	endometrial, epididym, kidney, ovary, oviduct, placenta, prostate, testis, uterus, etc.	767	61	828	687,306	245,465	932,771	314,882	75,893	390,765
Adipose tissue	adipo	6	3	9	10,884	6,869	10,884	1,638	1,854	3,492
Hematopoietic	bone-marrow	103	10	113	21,879	1,701	23,580	10,656	436	11,092
Vascular	aort, blood, platelet, etc.	31	16	47	19,635	16,140	35,775	8,231	5,205	13,436
						...				

ESTparser: web inteface

ESTparser

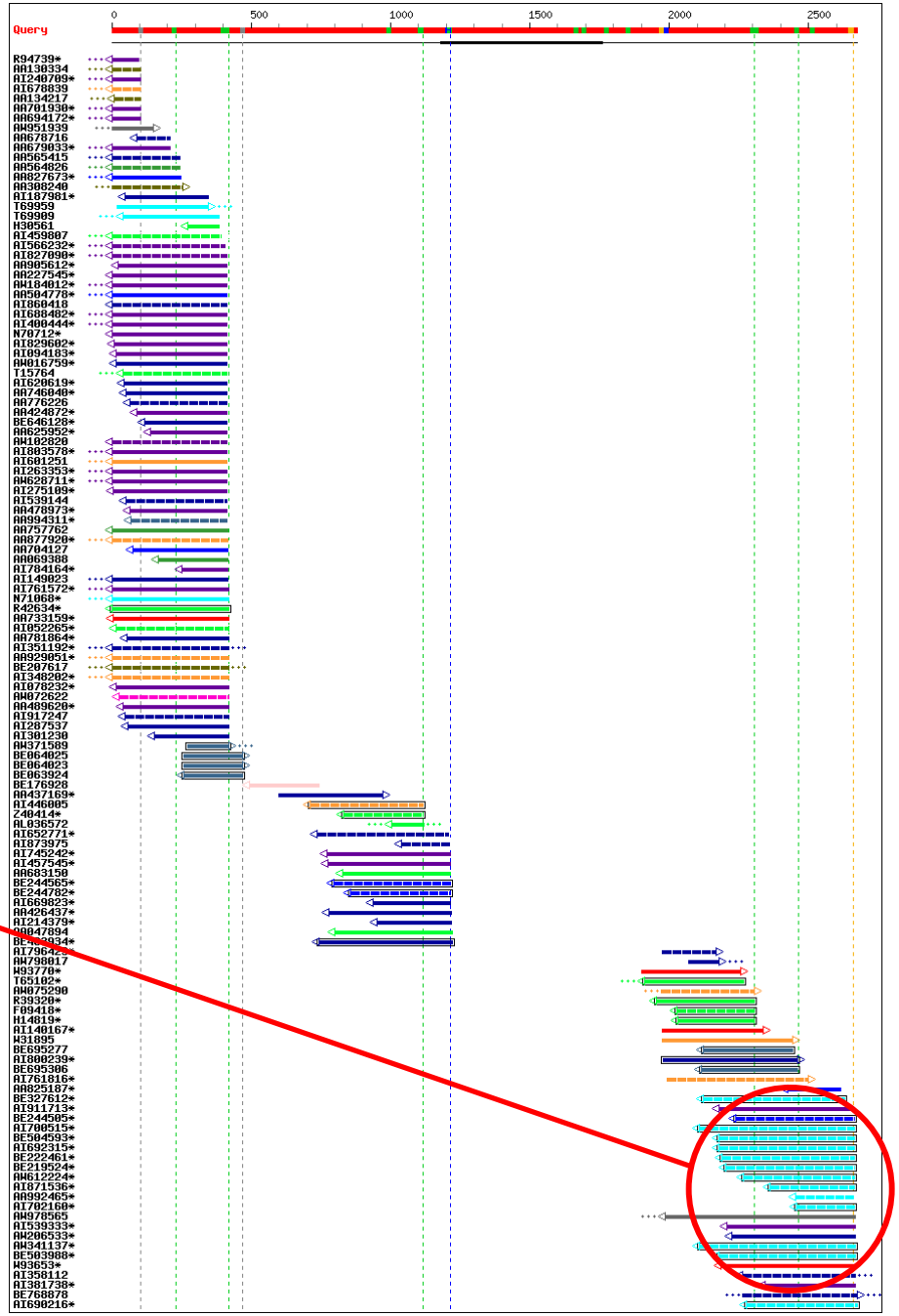

Or upload your sequence file [FILE SIZE < 25 kb]:
me/beaudoin/amyloid.fas
Parcourir...
Clear all
Run
To avoid duplicate submissions, click ONCE, and wait for response
Alter your parameters:
Expert corner
Choose the database to compare:
HUMAN dbEST
Match quality:
% identity:
97
Minimal length match:
60
Minimal match number to valid a PAS:
5
Minimal distance between PAS:
45
Internal priming minimal length:
8
Fisher's exact test:
one tail
two tail
maximal Fisher:
0.05
Display options:
Sequence scale:
60 nt
Result picture scale:
x 1
Show poly HSP:
Show masked query part:
Highligthed a word in the query sequence display:
ATTTA
[References (in preparation)] By: Emmanuel Beaudoing: & Daniel Gautheret:

ESTparser output



EST profiling with EST Parser

Tissue specific form?



Beaudoing & Gautheret,
Genome Res. 2001

Fisher's Exact Test

An example of such data (not real) are 100 persons classified after sex and e-mail address:

	female	male
e-mail address	3	15
no e-mail address	37	45

I.e. 3 of 40 women have e-mail address and 15 of 60 men have e-mail address.

The data could have been collected in different ways:

1. We have asked 60 men and 40 women. I.e. the total number of men and women is fixed.
2. We have asked 100 persons about their sex and whether they have e-mail address. I.e. only the total number of persons is fixed in advance.
3. We have asked all persons in Norway born 24/11-68. I.e.: The total number of persons is not fixed in advance.

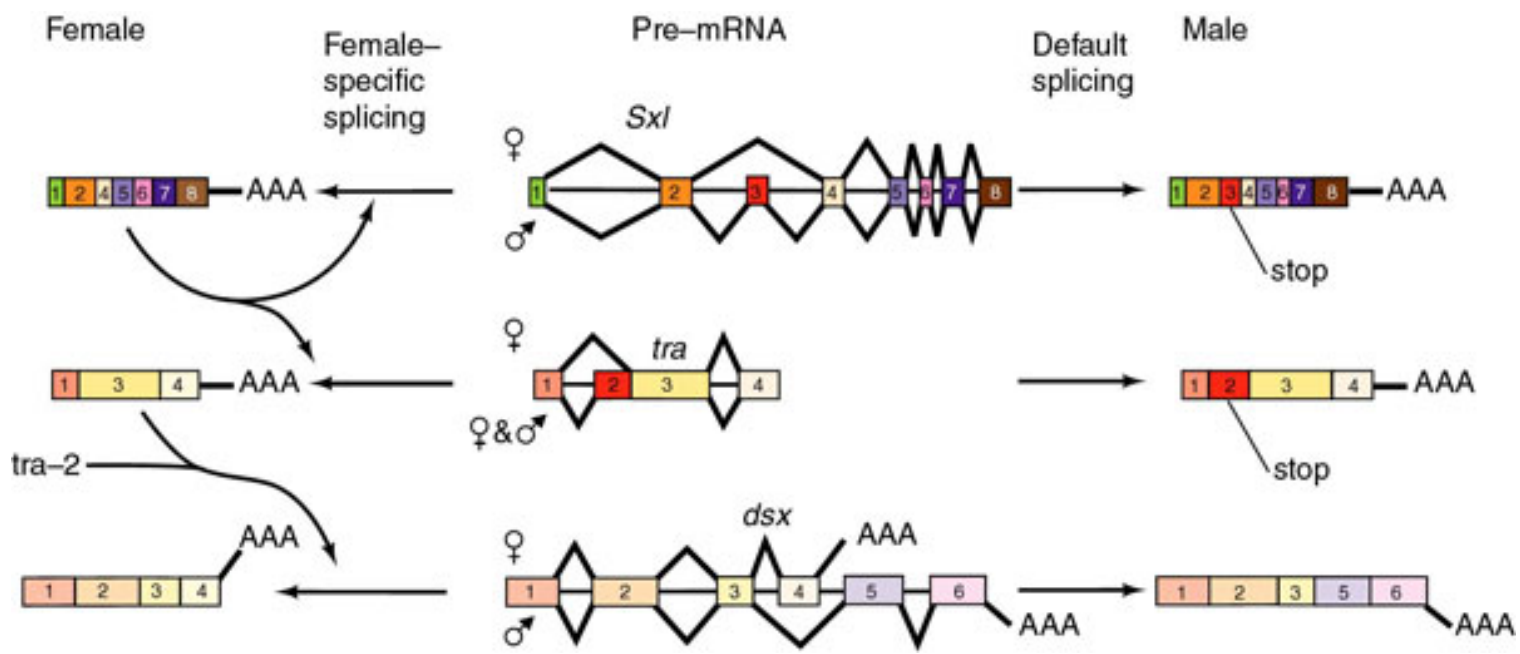
Pris sur le site de oyvind.langsrud@matforsk.no

1942 biases in 951 different human 3'UTR

Tissue/Organ	Nb. biases	Nb. UTRs	Nb. ESTs cumulate
skin	251	81	39,96
colon	180	70	59,3
nervous	169	48	6,21
breast	167	63	22,98
brain	151	77	168,24
liver	99	57	141,19
germinal	88	54	44,57
uterus	84	46	52,48
lung	78	49	113,57
Bcell	78	41	70,1
stomach	74	31	27,75
mixed	72	55	110,76
testis	53	40	46,27
heart	52	32	45,2
ovary	52	23	38,67
kidney	50	36	85,82
		...	

Epissage alternatif

3 gènes de détermination du sexe chez la drosophile, épissés différemment selon le sexe de l'individu:



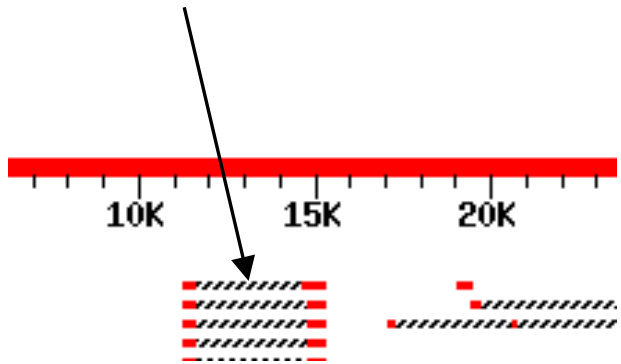
Epissage alternatif via EST

- ✓ Nombreux travaux réalisés à partir des EST
- ✓ Sociétés créées exclusivement sur le thème (par ex. compugen http://www.cgen.com/)
- ✓ Voir. Modrek B, Resch A, Grasso C, Lee C. Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res 2001 Jul 1;29(13):2850-9 :

We have identified **6201 alternative splice relationships in human** genes, through a genome-wide analysis of expressed sequence tags (ESTs). Starting with approximately 2.1 million human mRNA and EST sequences, we mapped expressed sequences onto the draft human genome sequence and only accepted splices that obeyed the standard splice site consensus. A large fraction (47%) of these were observed multiple times, indicating that they comprise a substantial fraction of the mRNA species. **The vast majority of the detected alternative forms appear to be novel, and produce highly specific, biologically meaningful control of function in both known and novel human genes,** e.g. specific removal of the lysosomal targeting signal from HLA-DM beta chain, replacement of the C-terminal transmembrane domain and cytoplasmic tail in an FC receptor beta chain homolog with a different transmembrane domain and cytoplasmic tail, likely modulating its signal transduction activity. **Our data indicate that a large proportion of human genes, probably 42% or more, are alternatively spliced,** but that this appears to be observed mainly in certain types of molecules (e.g. cell surface receptors) and systemic functions, particularly the immune system and nervous system. These results provide a comprehensive dataset for understanding the role of alternative splicing in the human genome, accessible at http://www.bioinformatics.ucla.edu/HASDB.

Ecueils à la prédiction des transcrits

- ★ Priming interne
- ★ Rétention d'introns => notion de « spliced EST »
- ★ Gènes chevauchants
- ★ Mauvaise couverture 5'
  - (cf banques RIKEN)



Faible conservation des transcrits alternatifs

(comparaison homme-souris. Modrek et al. 2003)

- ★98% des exons « constitutifs » sont conservés
- ★25% des exons « mineurs » sont conservés
- ★Exons mineurs: souvent des exons « récents » et d'expression tissu-spécifique

ASAP, AltExtron, ASD, SpliceNest…

ASAP interface
(Lee et al. 2003)

Figure 1. Screenshot of ASAP's gene view for transcobalamin I, and sequence view for its second exon. Each view has a title bar (tag #1); left menu (tag #2) that switches between various views for the current object; right menu (tag #3) that controls what view is shown when the user clicks on a feature; and maximize/split button (tag #4) that toggles between maximizing the view to fill the whole browser window, or splitting it into two views so the user can click on a feature in the upper view and see its detailed results in the lower view (tag #5). New searches, help, and additional information are available from the navigation bar (tag #6).

Les besoins actuels

- ★ Intégration initiation+épissage+transcription
- ★ Etude fonctionnelle (domaines, etc.)
- ★ Conservation
- ★ Validation expérimentale
- ★ Tissu-spécificité

The Alternative Transcript Diversity Project (ATD), 6e PCRD