La diversité des transcrits

Daniel Gautheret, 2006

1. Identification des transcrits par les EST
2. Diversité du transcriptome: travaux récents
3. Diversité des extrémités 3’

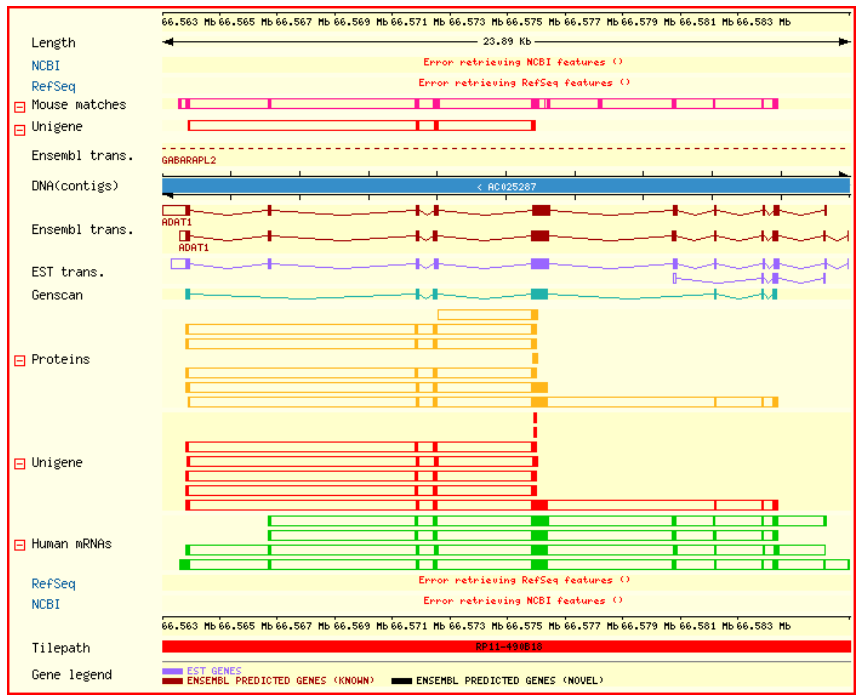
I. Principes de l’identification des transcrits par les EST

Combien de gènes dans le génome de mammifère?

Humain/souris 2005: 22000 gènes annotés

Est-on proche de la fin?

Si peu de différence avec la mouche (15000 gènes) ou C. elegans (18000) ?

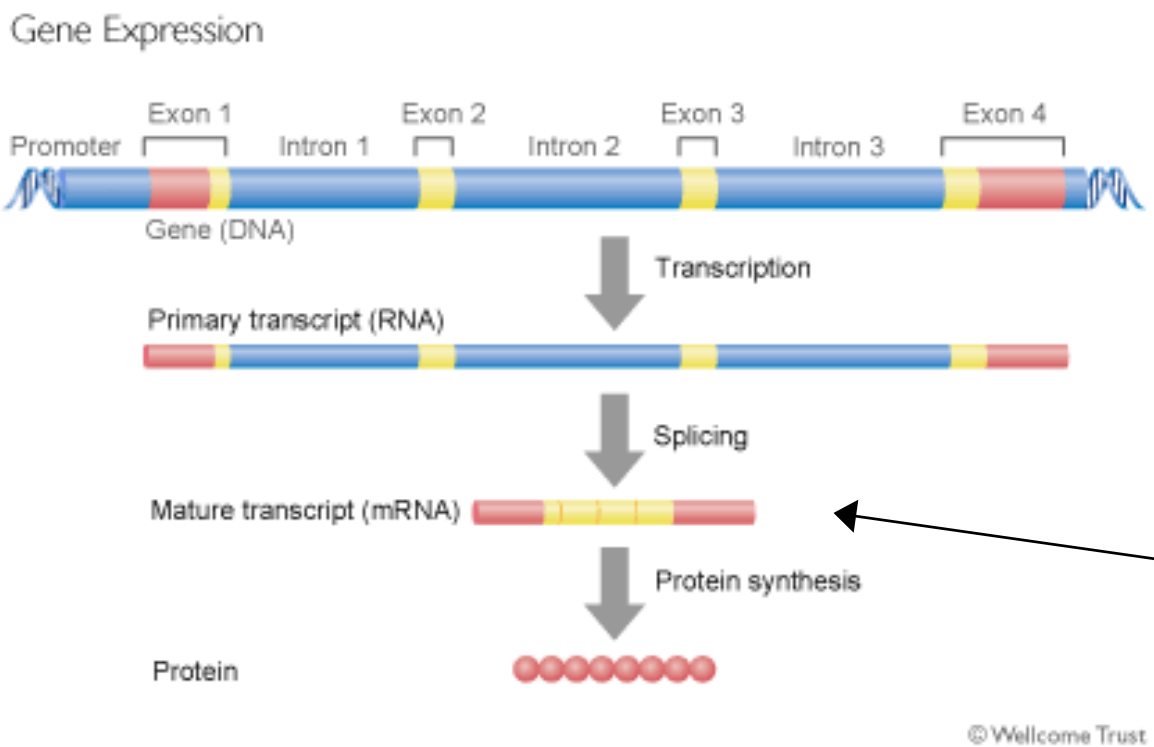


Une page Ensembl

Prédiction de gènes

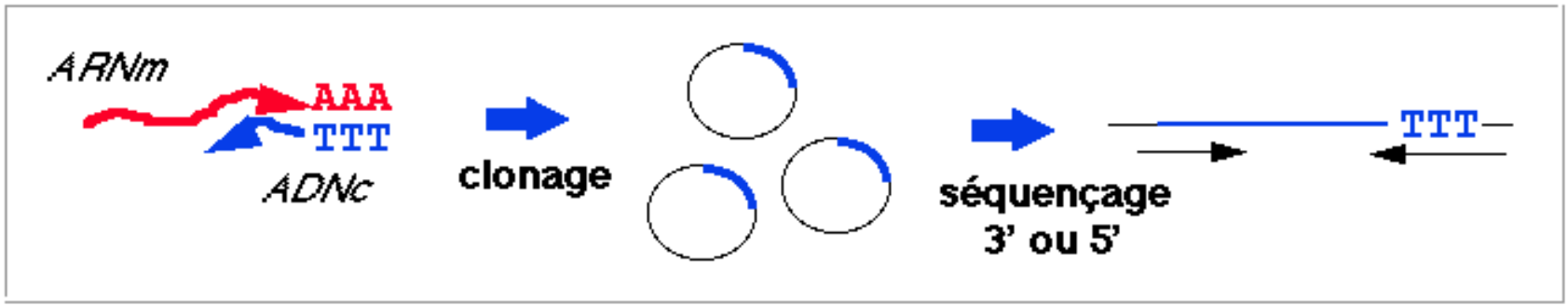
- Modèles purement bioinformatiques (HMM): <50% des gènes correctement prédits (difficulté de prédire les gènes complets)
- Très nombreux exons prédits en dehors des gènes reconnus: faux positifs?
- Conservation inter-espèce: 200.000 séquences conservées <u>en plus</u> des 200.000 exons connus

Réponse: Identifier les transcrits



+ facile à isoler et séquencer

Les EST (expressed sequence tags): le premier outil d'annotation des génomes



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, Formes alternatives

EST data: state of the art

NCBI

dbEST: database of "Expressed Sequence Tags"

dbEST release 091605

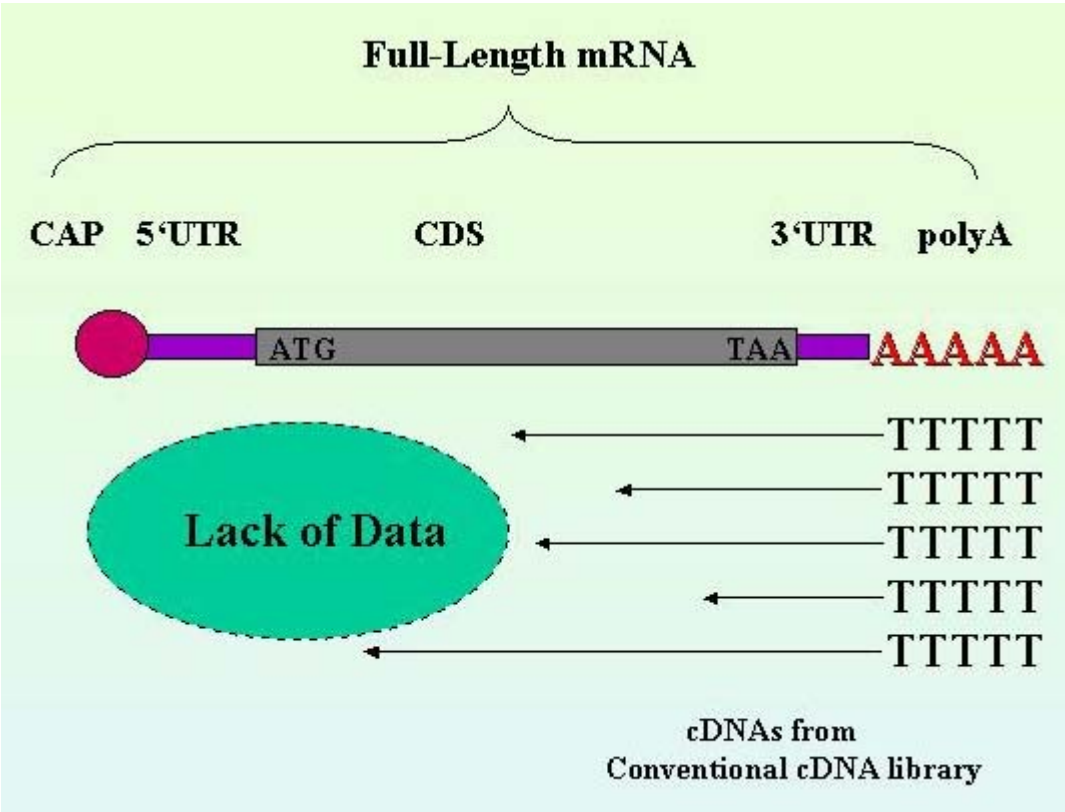
Summary by Organism - September 16, 2005

Number of public entries: 29,222,672

Homo sapiens (human)	6,133,870
Mus musculus + domesticus (mouse)	4,334,152
Xenopus tropicalis	1,038,272
Rattus sp. (rat)	701,072
Bos taurus (cattle)	684,354
Ciona intestinalis	684,319
Danio rerio (zebrafish)	651,991
Triticum aestivum (wheat)	599,989
Zea mays (maize)	566,404
Gallus gallus (chicken)	565,272
Sus scrofa (pig)	495,926
Xenopus laevis (African clawed frog)	473,289
Arabidopsis thaliana (thale cress)	420,789
Oryza sativa (rice)	406,651
Hordeum vulgare + subsp. vulgare (barley)	394,996
Drosophila melanogaster (fruit fly)	383,407

- Human: >5000 libraries (>600 cancer)
- Mouse: >500 libraries (>30 cancer)

Pourquoi des cDNA pleine longueur? (full length cDNAs)



Solutions:

- Long-distance RT enzymes
- Cap trapping

Clustering d'EST pour reconstruire les transcrits (désuet)

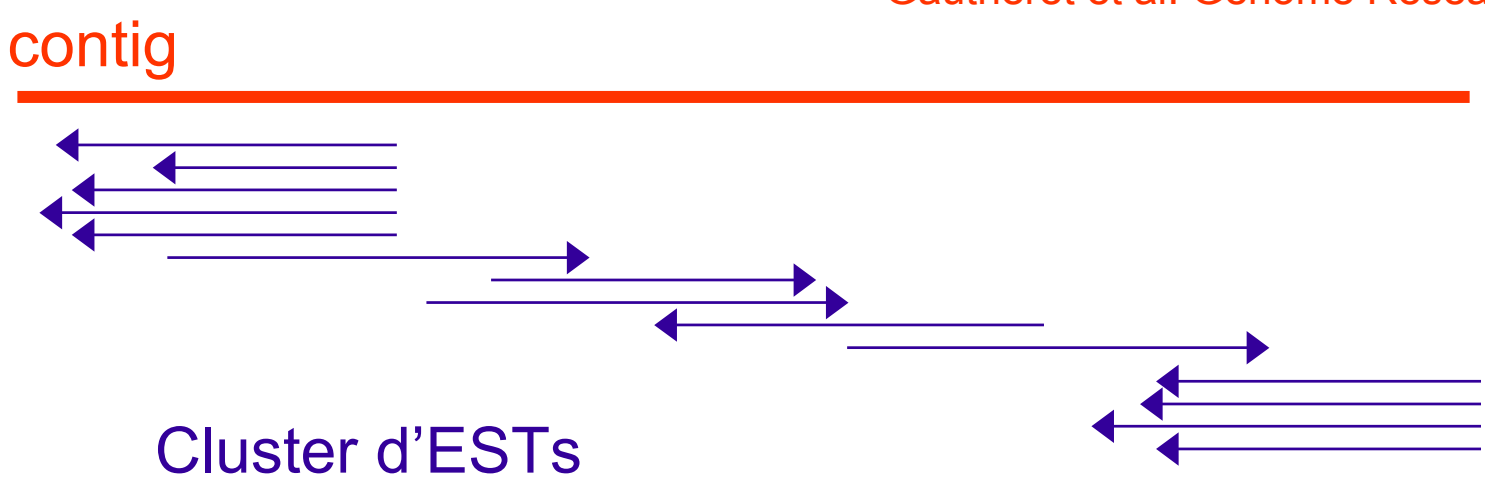
Réalisation de clusters d'EST pour reconstruire les transcrits complets (génomes non disponibles)

Table 2. Characteristics of the 1000 Largest Clusters

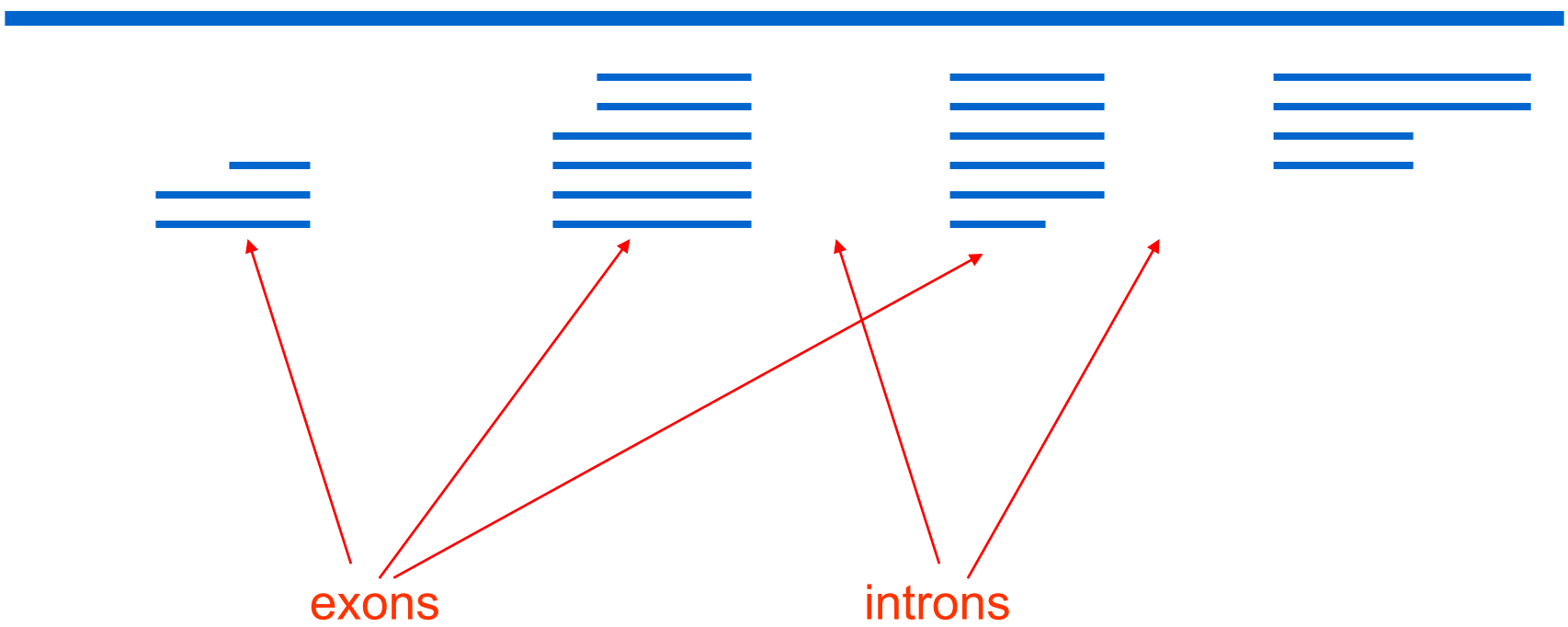
Clusters matching a GenBank primate sequence (%)[a]	Average no. of contigs per cluster	ESTs assigned to internal priming (%)[b]	Clusters exhibiting poly(A) sites (no.)		
			2	3	4
72.7	1.6	13.9	159	27	3

[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.

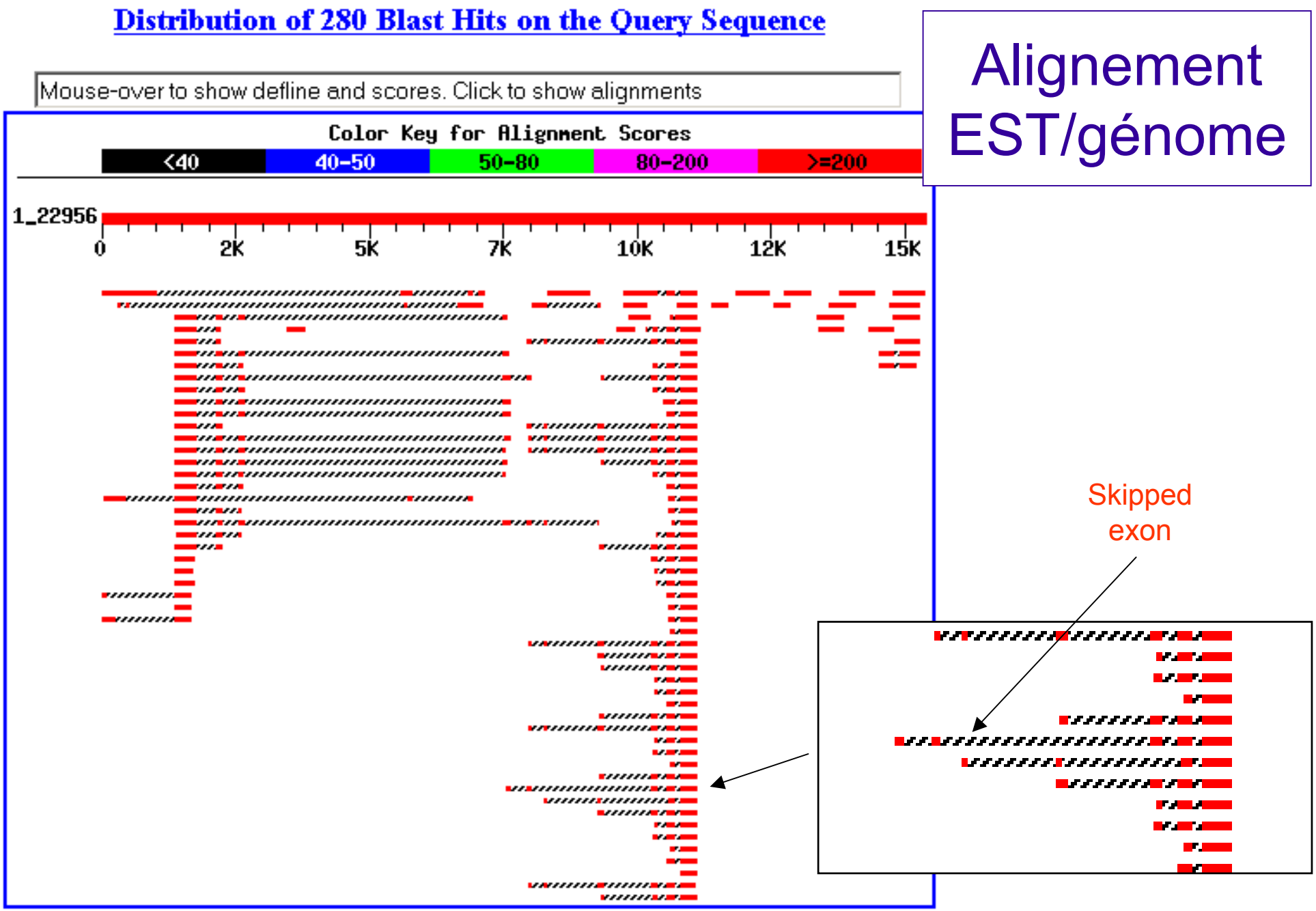
Gautheret et al. Genome Research, 1998



Alignements EST/cDNA contre génome

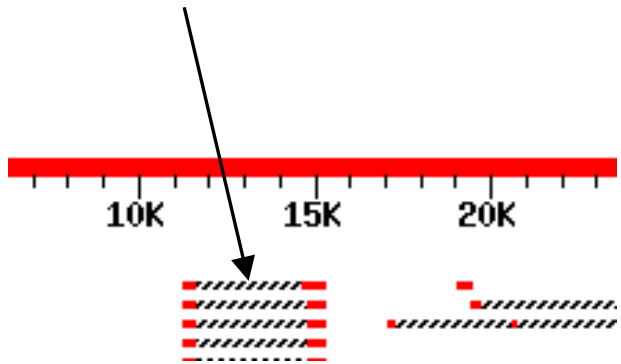


Alignement EST/génome



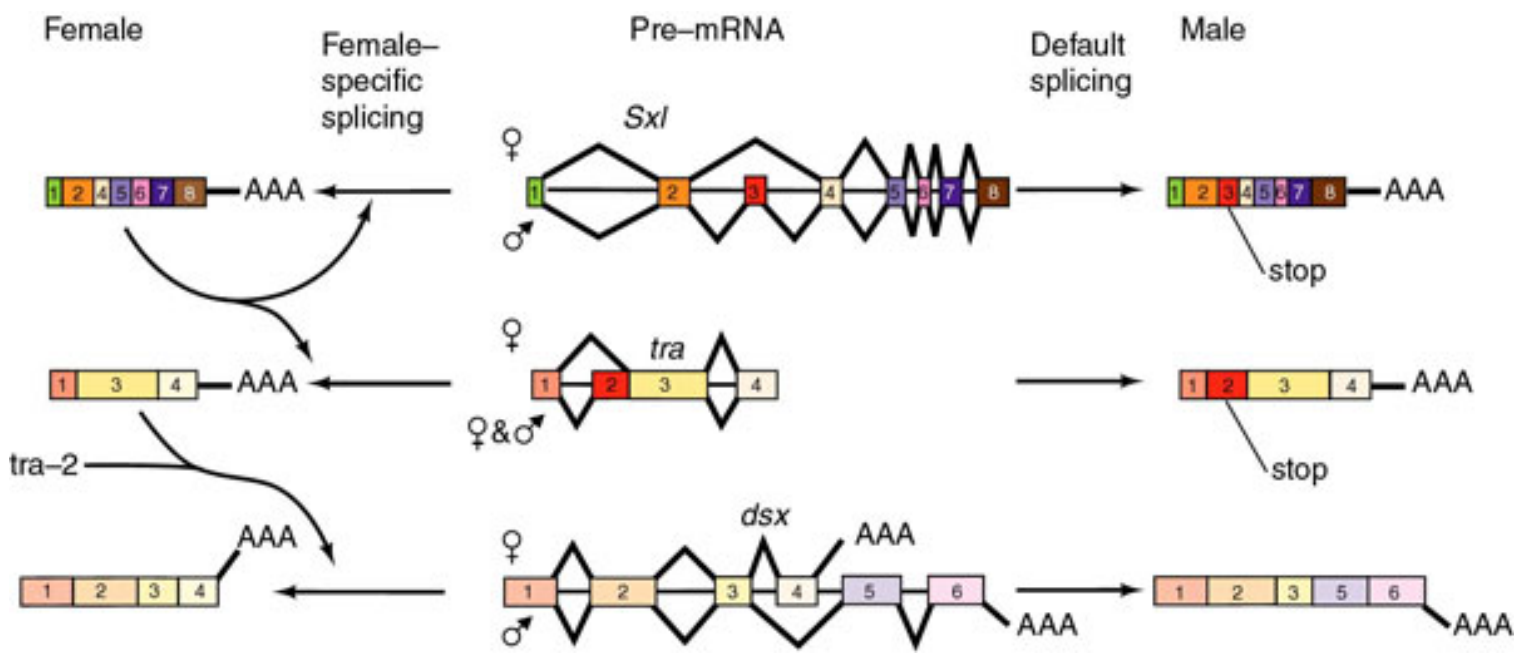
Ecueils à la prédiction des transcrits par les EST

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



Epissage alternatif

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



Epissage alternatif via EST/cDNA

- ✓ Nombreux travaux réalisés
- ✓ Sociétés créées exclusivement sur le thème (par ex. Compugen http://www.cgen.com/)
- ✓ Ex: 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.

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).

II. Travaux récents sur la diversité du transcriptome

- ★ Full-length cDNA
- ★ CAGE
- ★ Tiling array

The Transcriptional Landscape of the Mammalian Genome

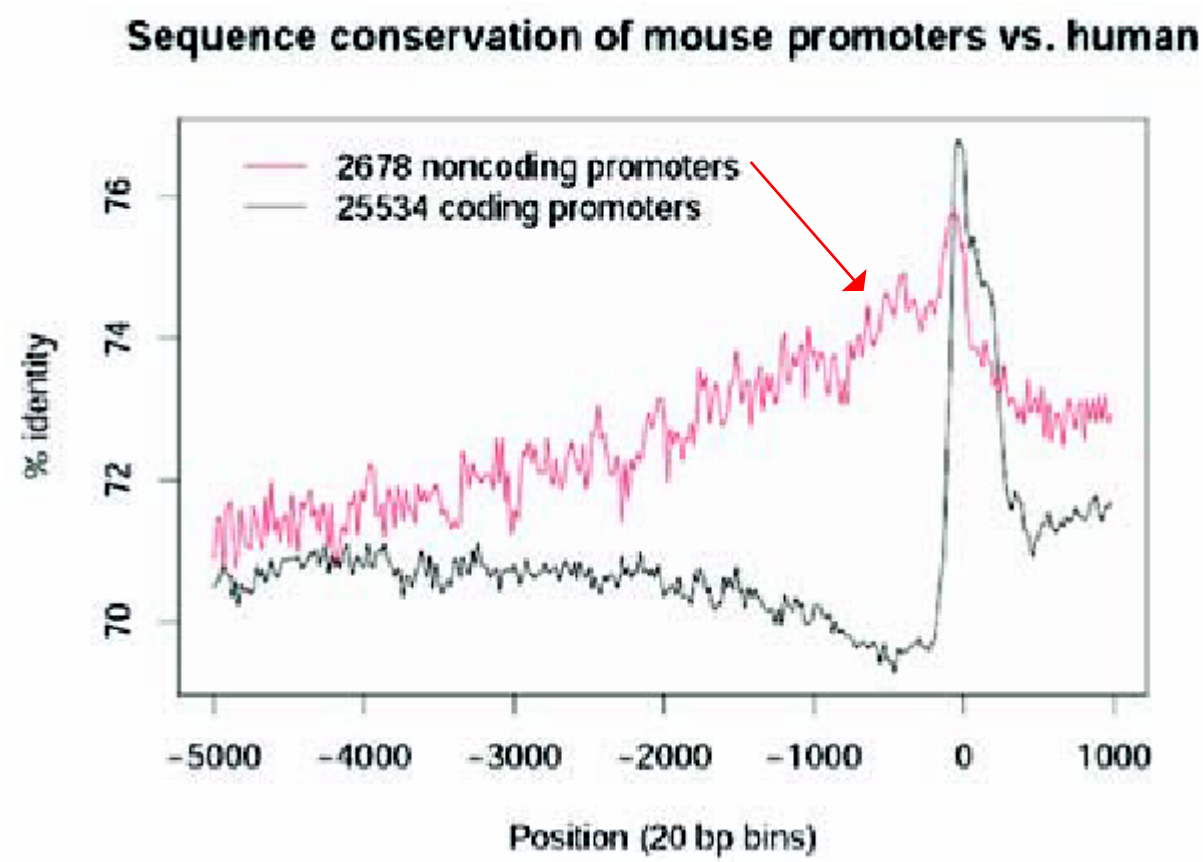
The FANTOM Consortium* and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group)*

Science, 2005

- ★ 100,000 Full length cDNAs
- ★ + 1M « CAGE » (sortes de SAGE en 5')

Utilité des full-lengths

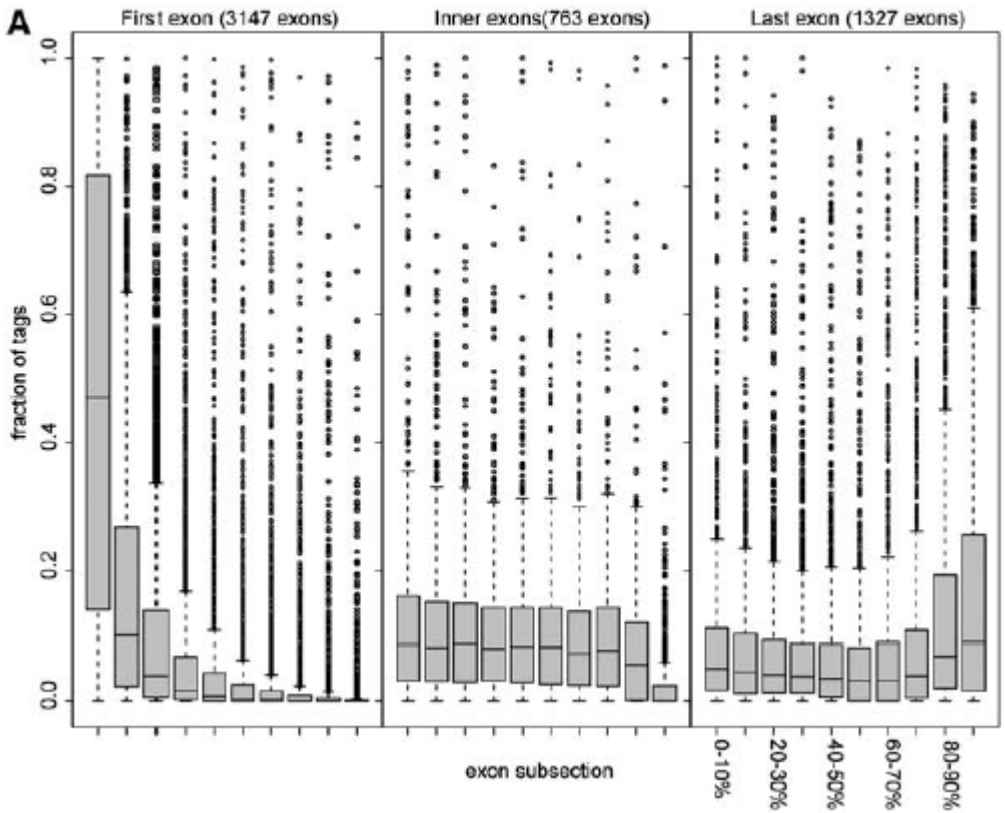
- ★ Etude des promoteurs (impossible avec les EST)
- ★ TSS alternatifs, exons 5' alternatifs



Principaux résultats de « FANTOM3 »

- ★ 32000 transcrits non-codants
- ★ 16000 nouveaux transcripts codants
- ★ 5000 nouvelles protéines
- ★ La majorité du génome est transcrit sur les deux brins
- ★ La grande diversité des transcrits dans chaque Unité de Transcription soulève le problème de la précision des puces ADN conventionnelles, dans lesquelles chaque sonde hybride différents transcrits

CAGE tags: démarrages de transcription dans les gènes et en 3’



Transcrits et Unités de Transcription

- ★ TU: Transcription Unit. mRNAs sharing at least 1 nt and with same location and orientation

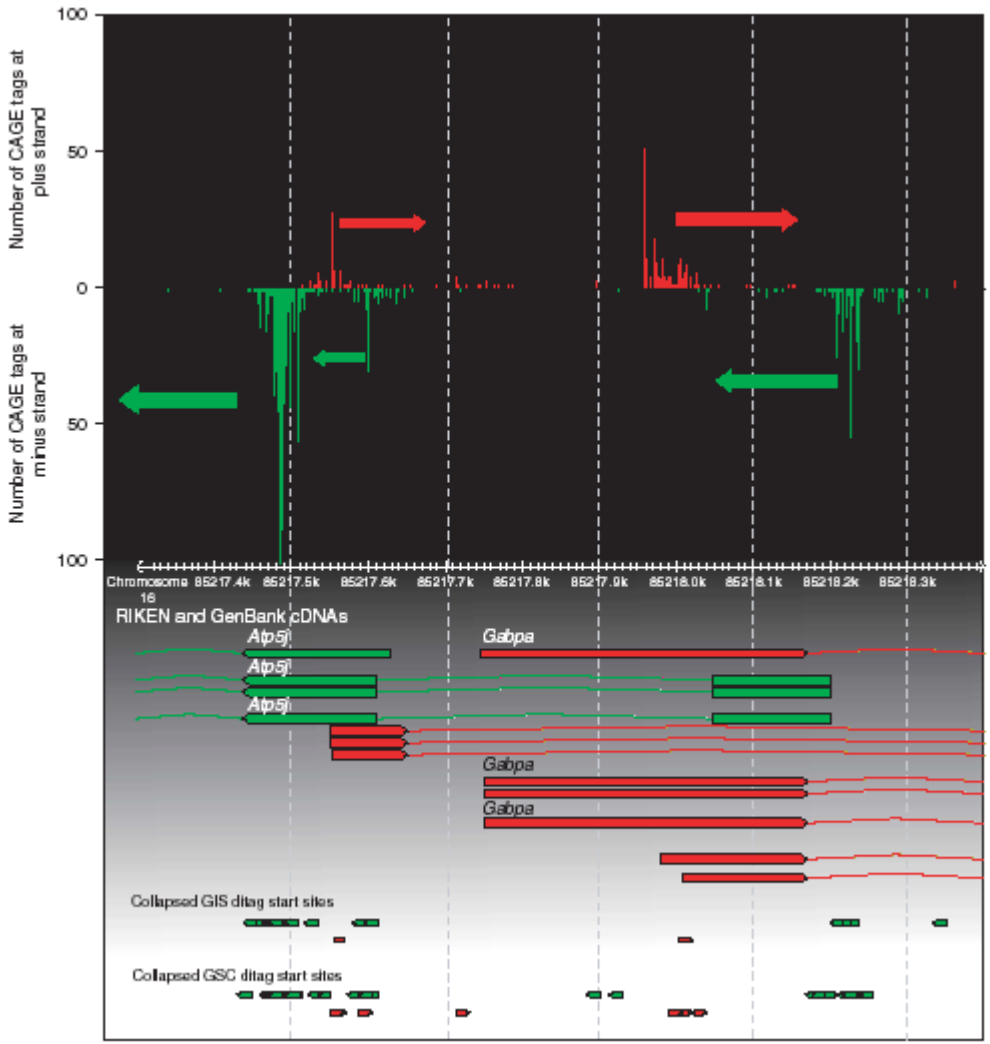
Table 2. Transcript grouping and classification. The extent of splice variation was calculated by excluding T-cell receptor and immunoglobulin genes from the transcripts. The remaining 144,351 transcripts were grouped in 43,539 TUs, of which 18,627 (42.8%) consist of single-exon transcripts, 8110 (18.6%) contain a single multiexon transcript, and the remaining 16,802 TUs (38.6%) contain at least two spliced transcripts. Among these TUs, 5862 (34.9%) show no evidence of splice variation, whereas 10,940 (65.1%) contain multiple splice forms.

	Total	Average per TU cluster
Total number of transcripts	158,807	7.59
RIKEN full-length	102,801	
Public (non-RIKEN) mRNAs	56,006	
GFs	25,027	1.20
Framework clusters	31,992	1.53
TUs	44,147	2.11
With proteins	20,929	1.00
Without proteins	23,218	1.11

Exemple d'annotation CAGE + cDNA

Genome Browser du laboratoire Riken

Transcrits chevauchants et antisens



Le Tiling Array

Review TRENDS in Genetics Vol.21 No.2 February 2005

Full text provided by www.sciencedirect.com

Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments

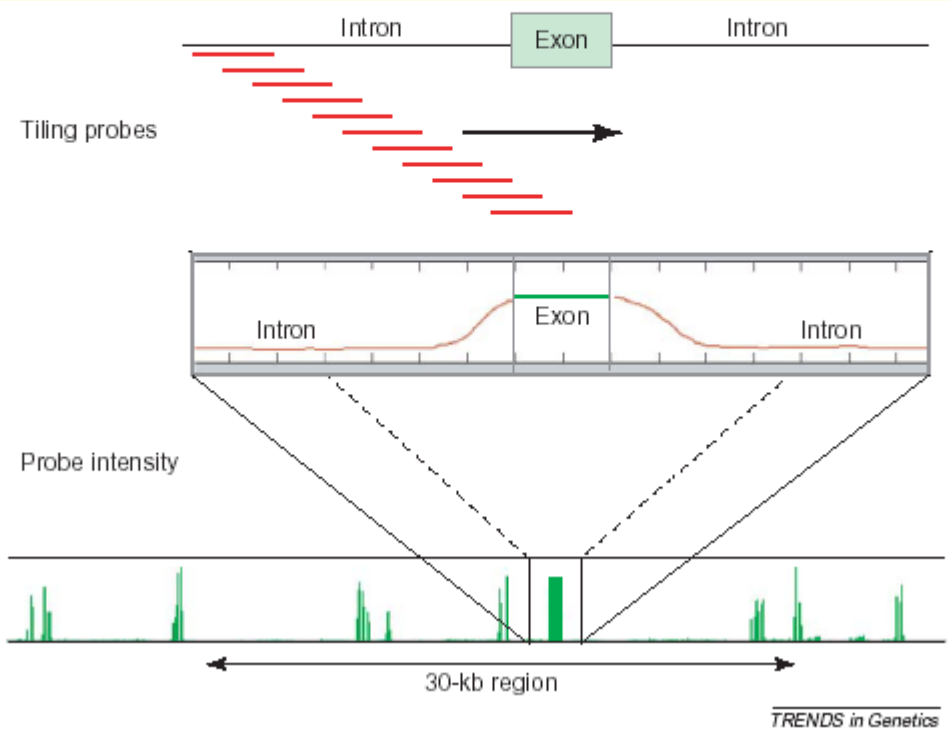
Jason M. Johnson[1], Stephen Edwards[1], Daniel Shoemaker[2] and Eric E. Schadt[1]

[1]Rosetta Inpharmatics LLC*, 401 Terry Avenue North, Seattle, WA 98109, USA
[2]GHC Technologies, 505 Coast Boulevard South, Suite 309, La Jolla, CA 92037, USA

★ Rosetta: technologie de spottage d'oligonuléotides par jet d'encre (inkjet)

Box 1. Tiling microarray experiments

Tiling microarrays are designed to assay transcription at regular intervals of the genome using regularly spaced probes (horizontal red lines) that can be overlapping (Figure I) or separated. The distance between the centers of successive probes is the 'step' size and probes can be selected to be complementary to one strand (as shown) or both strands. Probes can be synthesized directly onto or spotted onto glass slides, and can be synthesized oligonucleotides or PCR products. They are hybridized with fluorescently labeled cRNA or cDNA prepared from cell samples. Regions of greater fluorescent intensity (green peaks in lower panel) can reveal transcription within a large genomic region. In addition, the correlation of probe intensities in several different tissues (co-expression analysis) can be used to identify probes that are detecting exons of the same transcript. The lower panel shows the extent of a hypothetical transcript within the genome. The middle panel is a schematic, magnified view of the hybridization of a genomic region containing an exon.



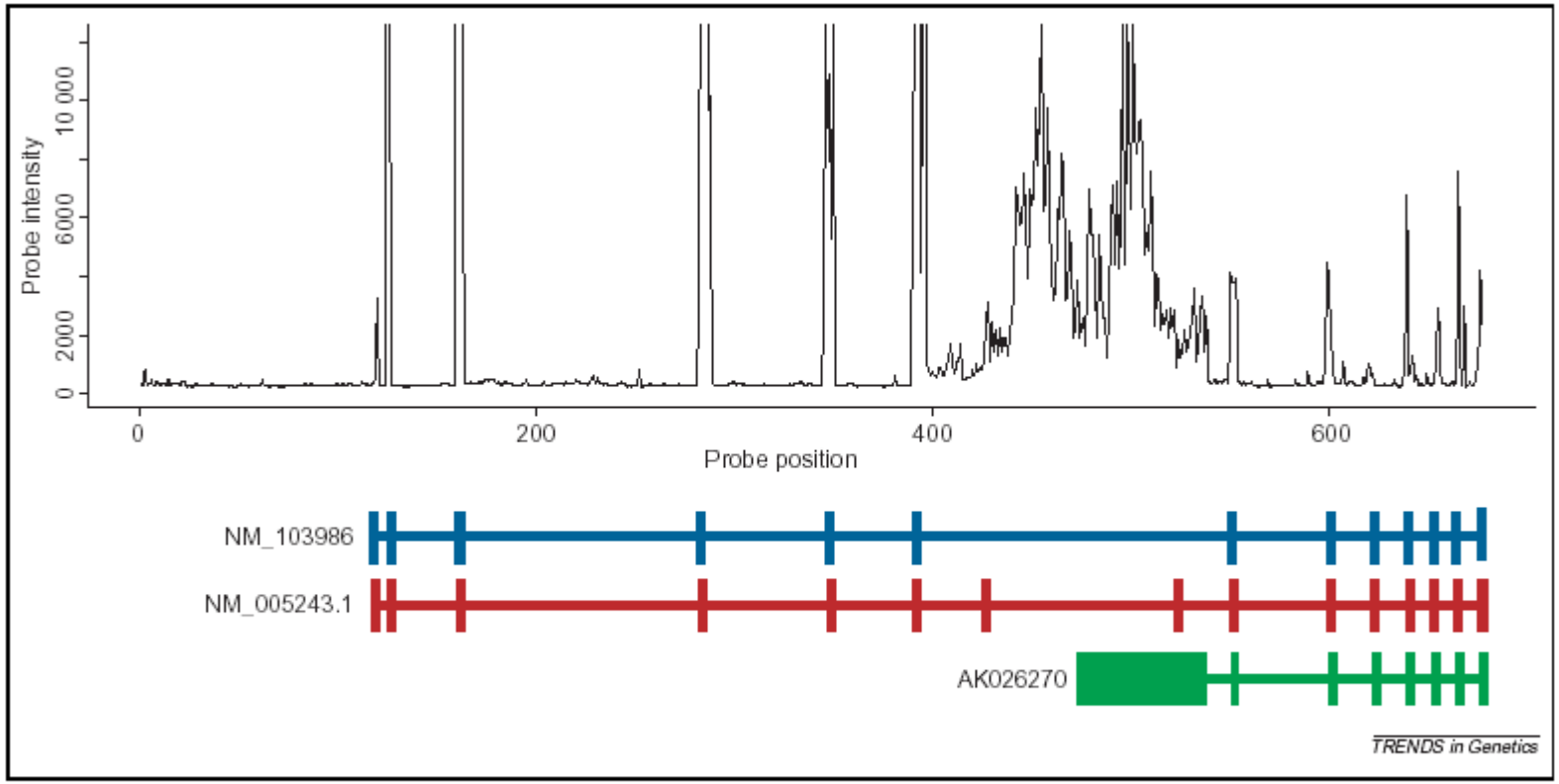


Figure 1. Microarray intensity profile for human thymus poly(A)$^+$ cDNA profiled on a 60mer ink-jet tiling array representing the genomic locus of the Ewing sarcoma breakpoint region 1 gene (*EWSR1*) in 30-nt steps, with probe index as the *x*-axis. No probes are shown for repeat-masked regions. The tiling data for this locus are shown in relation to the exon positions (below the plot) of three EWSR1 cDNAs (Genbank accession numbers: NM_013986, NM_005243.1 and AK026270). The 5′-most exon lies in a repeat-masked region and is not shown. A few peaks with the highest intensity have been truncated in Figures 1–3.

★ Gène bien caractérisé: confirmation des exons « refseq » et apparition de nouvelles régions transcrites (ici correspondant à un cDNA déjà observé)

The first example shows expression activity for a well-characterized gene, Ewing sarcoma breakpoint region 1 (*EWSR1*; Figure 1). The tiling data clearly detect exons corresponding to the RefSeq transcript of this gene but there also appears to be additional signal in at least one intron. The transcript represented by the cDNA sequence AK026270, shown in Figure 1, might correspond to a portion of this activity (as might others such as AL833489 and BX648769, not shown). In this case, the unexplained transcription suggests there are alternatively spliced isoforms or other uncharacterized RNAs transcribed from this locus.

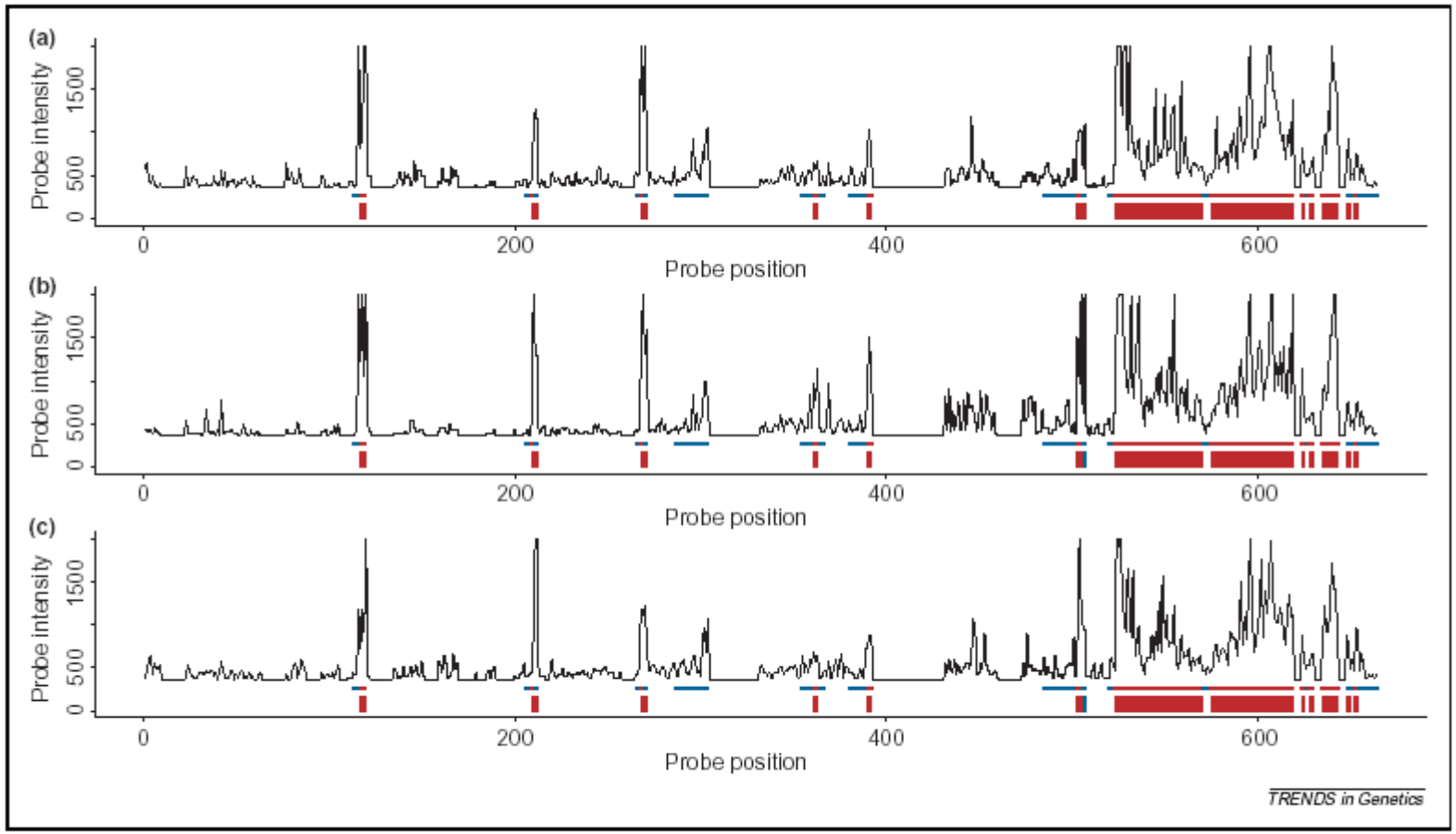


Figure 2. Microarray tiling confirmation of a predicted gene. Microarray tiling probe intensities for a region of human chromosome 20 that contains an *ab initio* gene prediction (made by the program GENSCAN [48]). Predicted exons for this gene are shown with blue lines. The transcription detected by microarrays has been grouped into a single transcriptional unit (dark red) by the correlated behavior of these probes across different human samples [11]. The conditions displayed are **(a)** thalamus, **(b)** testes and **(c)** uterus.

★ Gène prédit: confirmation ou non des exons prédit par Genscan ou autre.

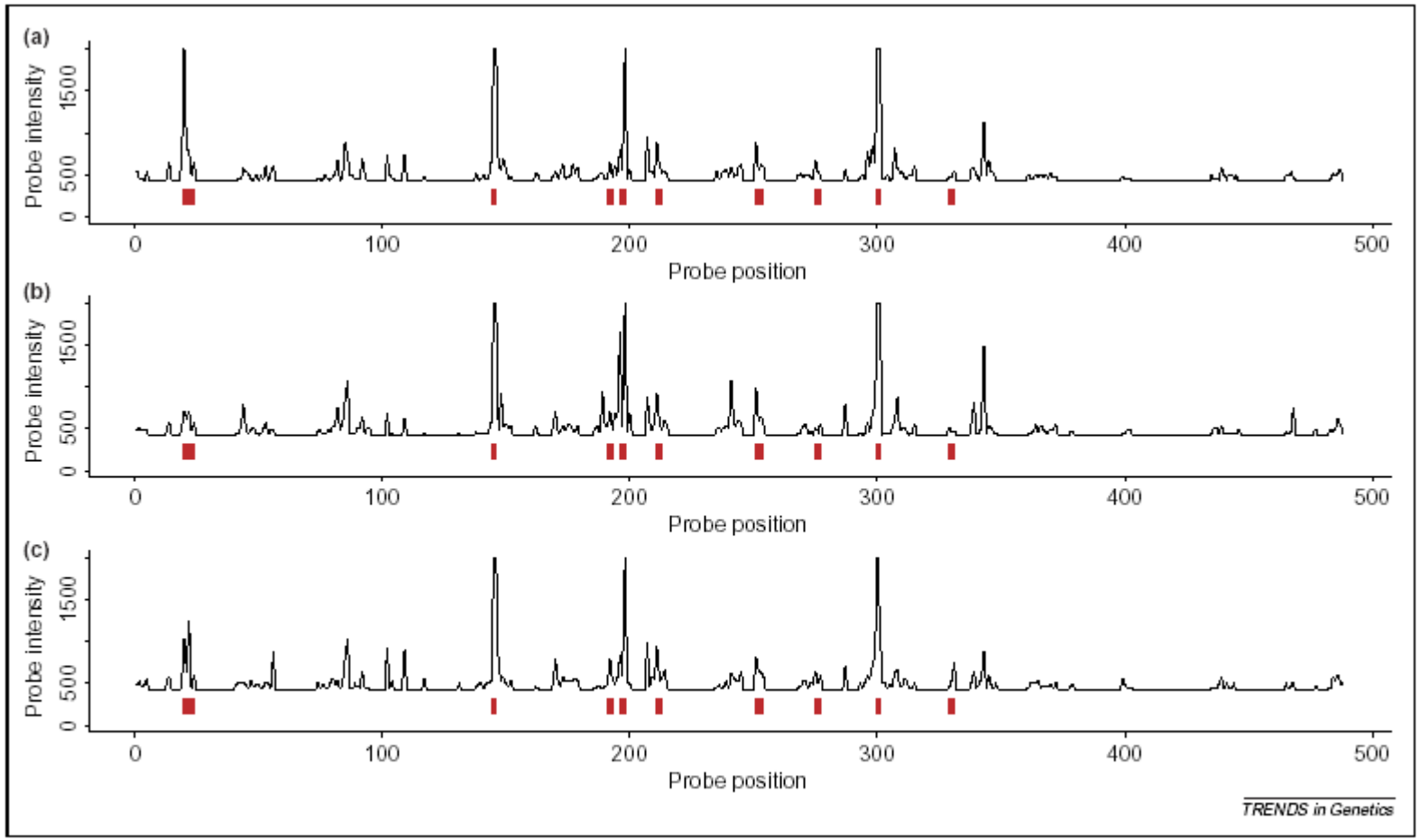


Figure 3. Tiling intensity profile for poly(A)$^+$ RNA samples from **(a)** thalamus, **(b)** testes and **(c)** uterus from an intergenic region of human chromosome 20 that contains no mapped ESTs or computational gene predictions. Dark red bars indicate transcription activity that has been grouped into a single transcriptional unit given the relative activity of these probes with respect to one another across different samples as described in Schadt *et al.* [11].

★ Apparition de nouveaux gènes dans régions intergéniques sans EST ni prédiction bioinformatique.

Que trouve-t-on?

- ★ Nouveaux gènes codants
  - >2000 transcrits avec ORF non existant dans Ensembl
- ★ Nouveaux gènes non-codants
  - >10000 transcrits
- ★ Transcrits antisens
  - >5000 gènes ont transcription antisens
- ★ Isoformes de gènes connus
  - Au moins 40% à 60% des gènes selon auteurs
  - Comprend également extensions (parfaois très longues) en 3'
- ★ Artefacts
  - biologiques: « bruit » transcriptionnel
  - expérimentaux: hybridation non-spécifique, contamination génomique…

Transcriptional Maps of 10 Human Chromosomes at 5-Nucleotide Resolution

Jill Cheng,[1*] Philipp Kapranov,[1*] Jorg Drenkow,[1] Sujit Dike,[1] Shane Brubaker,[1] Sandeep Patel,[1] Jeffrey Long,[1] David Stern,[1] Hari Tammana,[1] Gregg Helt,[1] Victor Sementchenko,[1] Antonio Piccolboni,[1] Stefan Bekiranov,[1] Dione K. Bailey,[1] Madhavan Ganesh,[1] Srinka Ghosh,[1] Ian Bell,[1] Daniela S. Gerhard,[2] Thomas R. Gingeras[1]†

Science, 2005

- ★ Technologie: Affymetrix
- ★ 25-mères espacés de 5bp
- ★ polyA+ et polyA-
- ★ Noyau et cytoplasme
- ★ 8 lignées cellulaires différentes

Etat de polyadenylation

- ★ Parmi tous les transcrits observés:
  - 19% polyA+
  - 44% polyA-
  - 37% polyA+ et polyA-
  - Donc la moitié du transcriptome humain est polyA-
  - Important, car les transcrits qu'on regardait jusqu'à présent n'étaient que les polyA+!
- ★ La plupart des polyA- nucléaires sont introniques (normal)
- ★ Mais >50% des transcrits cytoplasmiques sont polyA-

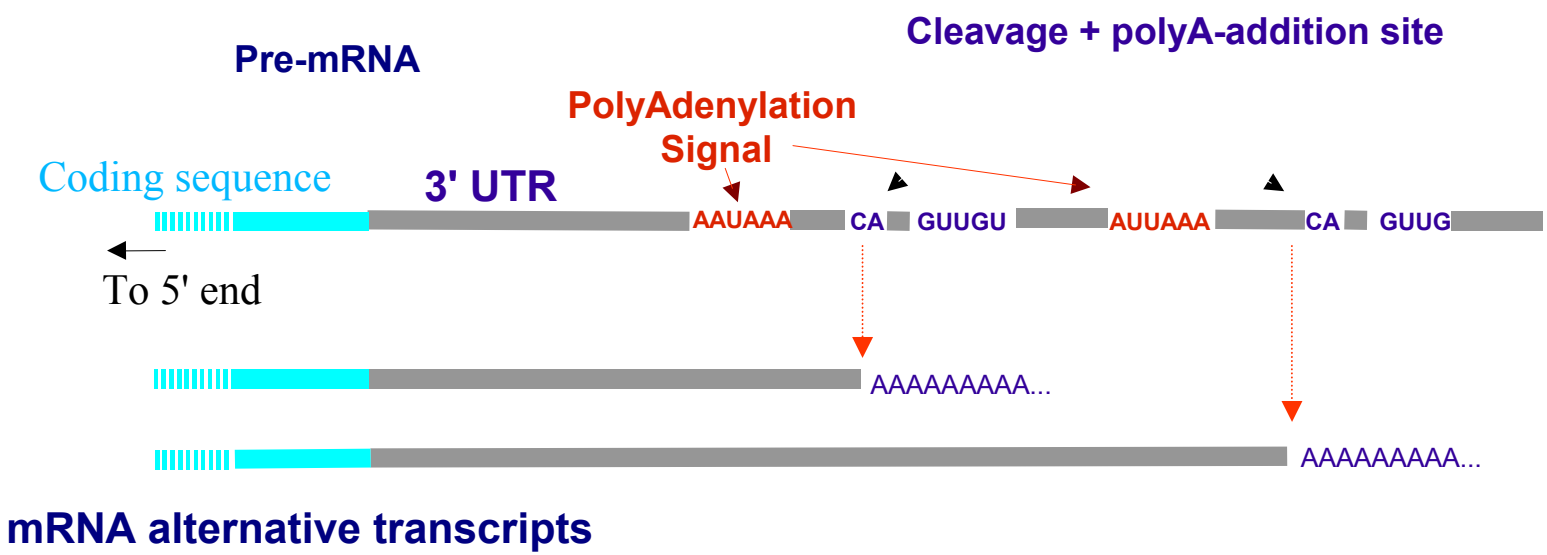
Position des transcrits

- ★ 60% des loci exprimés présentent des évidences de transcription sur 2 brins
- ★ Beaucoup de transcription dans l'intergénique:
  - 50% des polyA- cytoplasmiques et 25% des polyA- nucléaires sont intergéniques
  - 41% des polyA+ nucléaires sont intergéniques

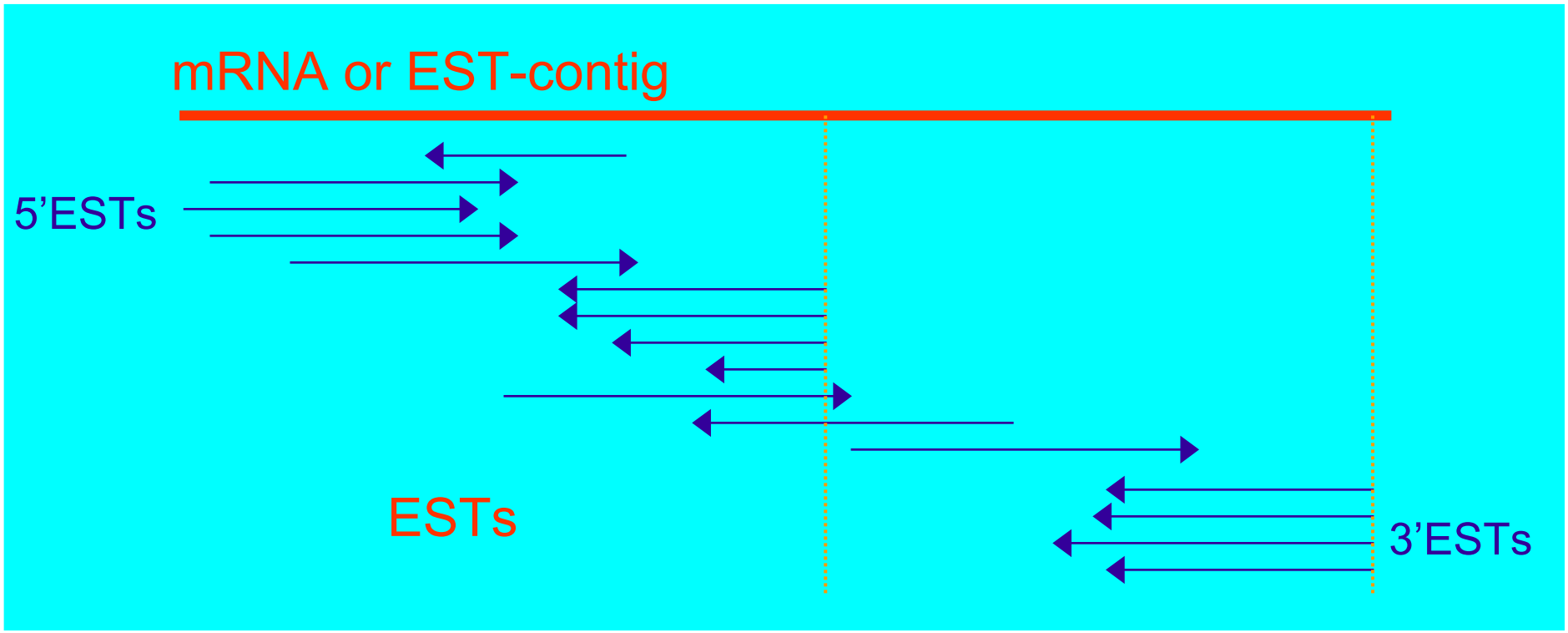
Derniers résultats approfondis sur régions ENCODE, multi-tissu:
93% du génome serait transcrit!!

III. La polyadenylation alternative

La polyadénylation alternative



PAS Discovery through EST/mRNA Alignment

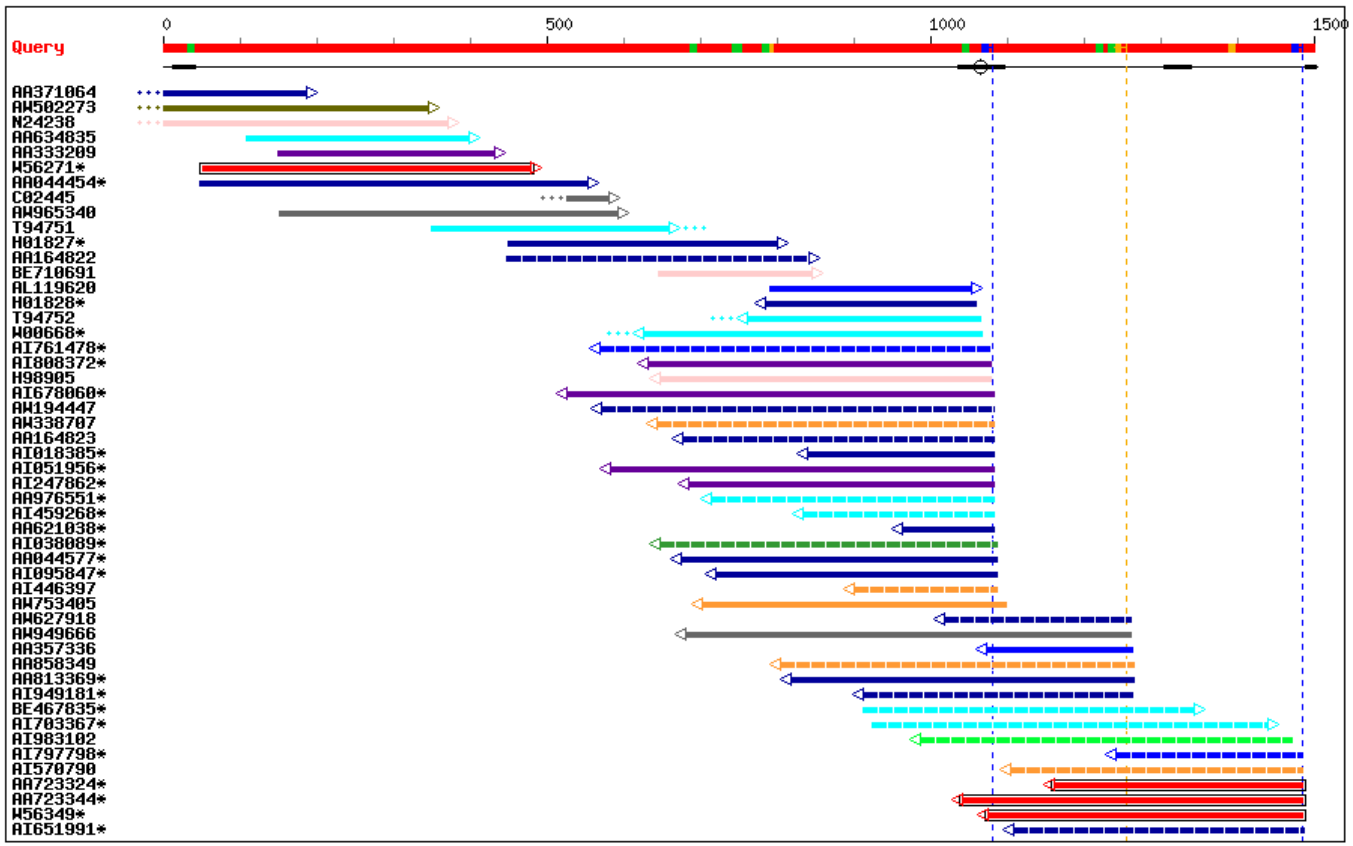


First observation in 1998: <u>189 cases</u> of alternative polyadenylation

Gautheret et al. (1998)
Genome Res. 8, 524

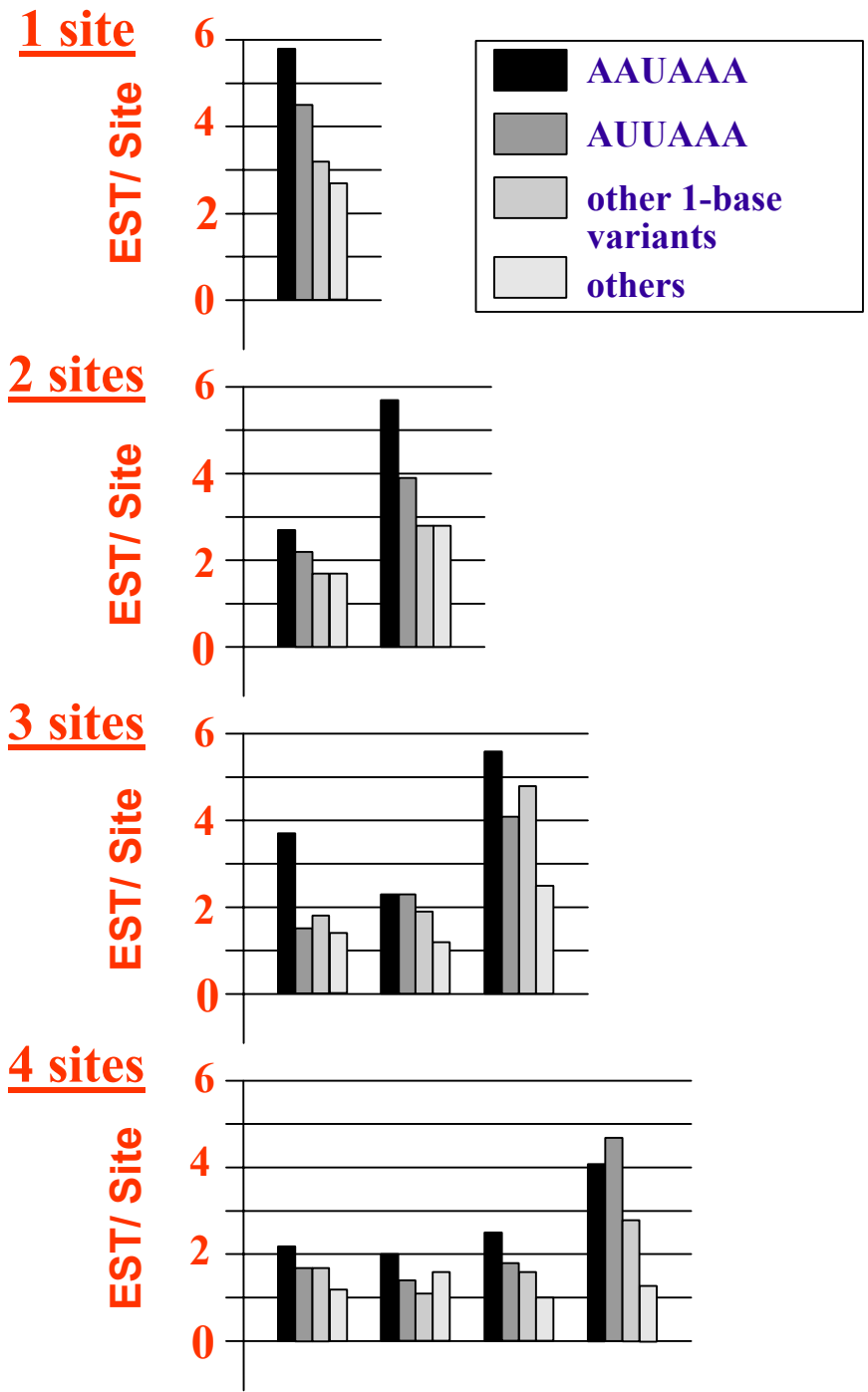
2000: ~1000 Genes Observed w/ Alt PAS

(estimation: at least 22% of genes)



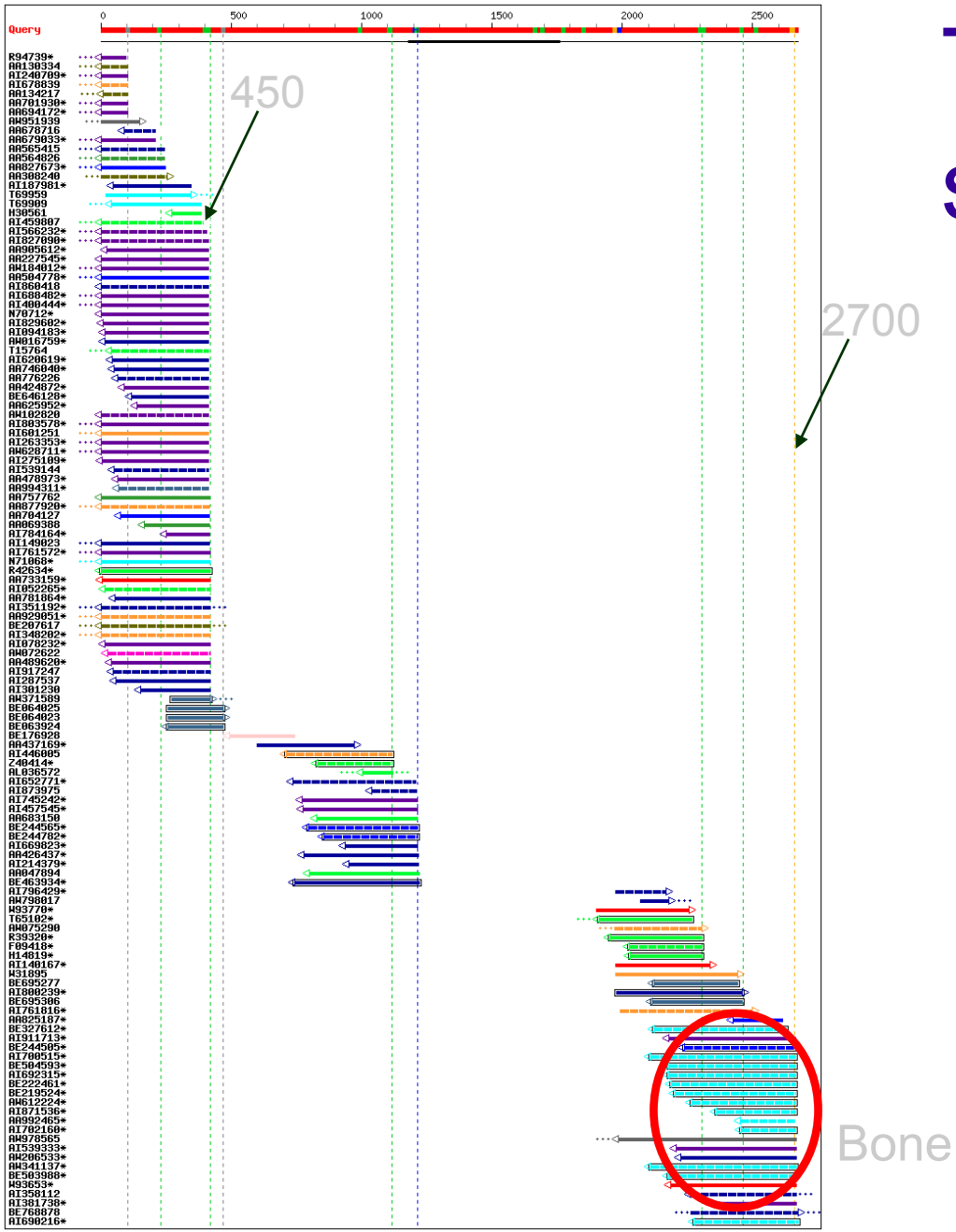
Beaudoing et al. (2000)
Genome Res. 10, 1001

EST Counts as a Measure of Signal Efficiency



- ✓Distal site more efficient than proximal
- ✓AAUAAA signal more efficient than variant signal

Tissue-specific sites



Site

	450	2700
Bone	2	10
Others	49	11

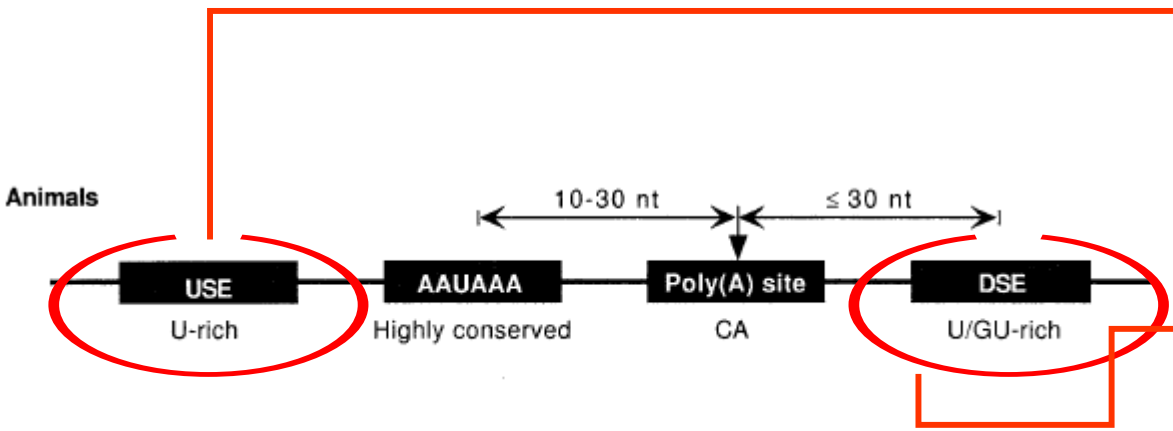
Fisher's Exact: P=0.00003

1942 biases in 951 different human 3'UTR

Beaudoing & Gautheret (2001)
Genome Res. 9, 1520

Improved Definition of PAS signals

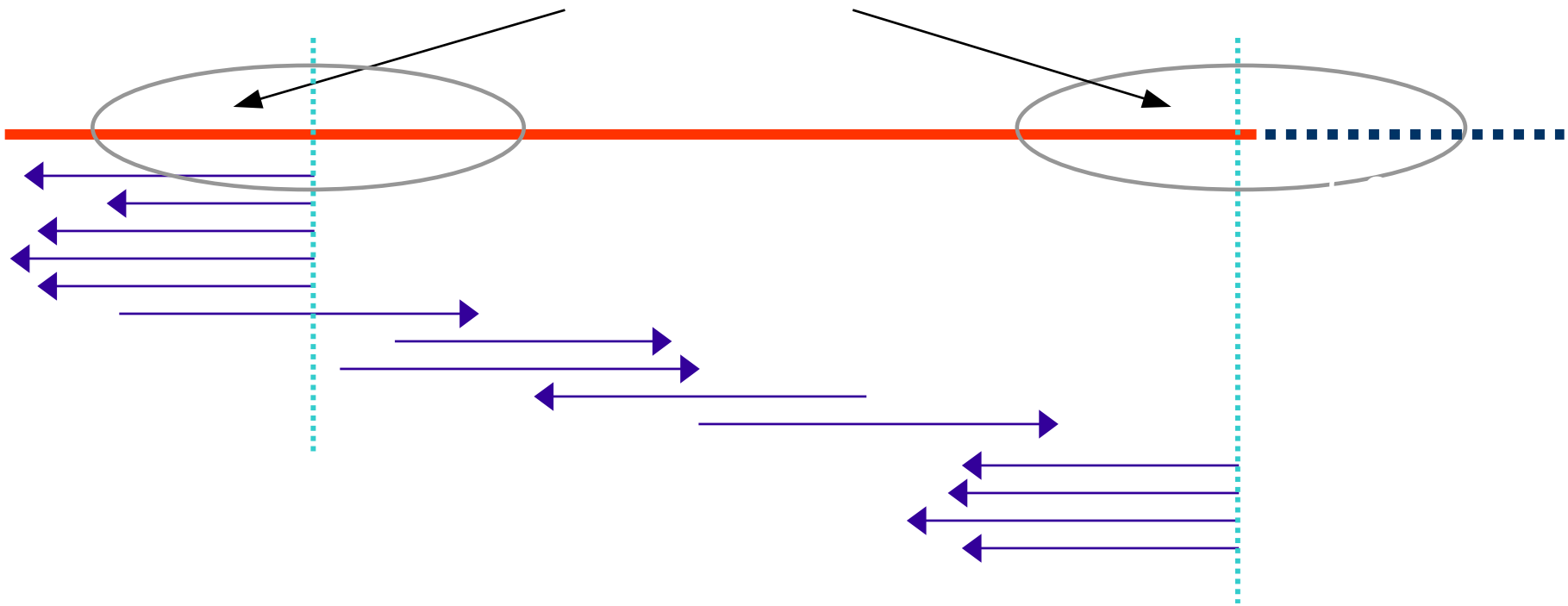
- ★ How does the polyA machinery tells a true cleavage site from a random AATAAA?
- ★ What other signals help dictate use of specific sites in certain conditions?



Upstream Seq. Elemt.
« enhancer element »
Mostly found in viral Sequences

Downstream Seq. Elemt
« constitutive »
Poorly defined
Mutations tolerated

Analysis of PAS-flanking regions



Most Frequent Hexamers

The AAUAAA and AUUAAA hexamers are the most frequently found.

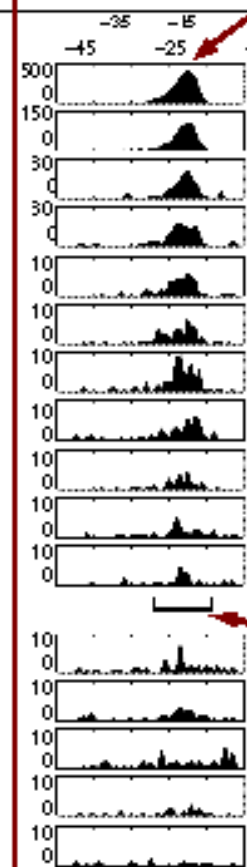
26.8% of the 3'' fragments do not contain a usual polyadenylation signal.

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
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^{-97}	−16 +/− 5.9
UAUAAA	180 (53)	3.2	4×10^{-48}	−18 +/− 7.8
CAUAAA	76 (23)	1.3	1×10^{-13}	−17 +/− 5.9
GAUAAA	72 (21)	1.3	2×10^{-13}	−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

Location[c]



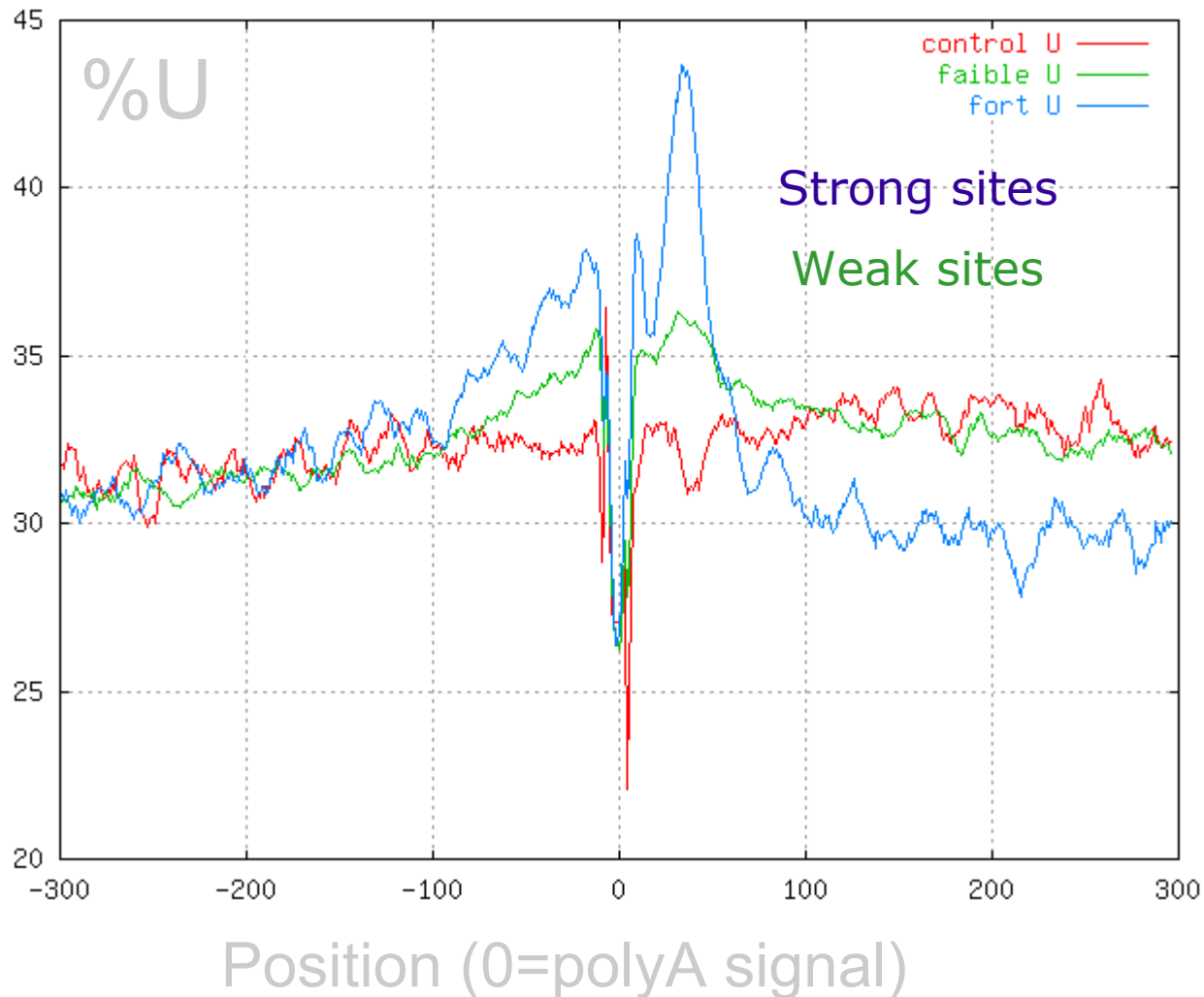
As expected from experimentally validated signals, AAUAAA and AUUAAA hexamers are clearly clustered around −15/−16 nt upstream of the putative poly(A) site.

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

USE, DSE and Polyadenylation Efficiency

USE
Paired t-test
data: USE weak and USE strong
t = -0.1826, df = 35, p-value = 0.8562

DSE
Paired t-test
data: DSE weak and DSE strong
t = -3.5876, df = 35, p-value = 0.001010



Legendre & Gautheret, BMC Genomics (2003)

EST/cDNA-based PAS Map

2006

	human	mouse	chicken
Tot PAS	89,200	72,400	2,900
PAS <3K from ENSEMBL Gene	42,500	37,900	2,400
PAS >10K from ENSEMBL Gene	43,900	33,600	300
Genes with no PAS	24%	19%	90%
Genes with 2 or more PAS*	57%	53%	6%

*relative to all genes with 1+ PAS

A large-scale analysis of mRNA polyadenylation of human and mouse genes

Bin Tian*, Jun Hu, Haibo Zhang[1] and Carol S. Lutz

Department of Biochemistry and Molecular Biology, New Jersey Medical School, UMDNJ, Newark, NJ 07101, USA and [1]Center for Computational Biology and Bioengineering, New Jersey Institute of Technology, Newark, NJ 07102, USA

-> 54% human, 32% mouse (accept multiple sites for 1 signal, but refseq UTR only)

>40,000 polyadenylated sites between Ensembl annotated genes !

What are they?

Background transcription ?

Background polyadenylation ?

What is the actual reach of the 3' UTR?

Textbook « Human Molecular Genetics 2 » (1999):

- ★ 3′ UTR Average of about 0.6 kb (see Zhang, 1998) but this is likely to be an underestimate because of underreporting of genes with long 3′ UTRs

Untranslated Regions of mRNA (Mignone et al. 2003) :

	5′ UTR				3′ UTR			
	Number of sequences	Average length	Maximum length	Minimum length	Number of sequences	Average length	Maximum length	Minimum length
Humans	1,203	210.2	2,803	18	1,247	1,027.7	8,555	21
Other mammals	142	141.3	936	20	148	441.1	3,324	37
Rodents	638	186.3	1,786	16	457	607.3	3,354	19
Aves	59	126.4	620	17	56	651.9	3,990	21
Other vertebrates	105	164.0	1,154	15	111	446.5	2,858	31
Invertebrates	5,464	221.9	4,498	14	3,736	444.5	9,142	15
Liliopsidae	144	129.8	715	17	127	273.3	1,605	22
Other Viridiplantae	1,471	103.0	1,355	12	1,699	207.7	1,911	13
Fungi	388	134.0	1,088	16	326	237.1	1,142	25

!

Several recent papers mentionning distal PAS

Long-Range Heterogeneity at the 3′ Ends of Human mRNAs

Computational analysis of 3′-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat

Jun Yan[1,3] and Thomas G. Marr[1,2,3]

A large-scale analysis of mRNA polyadenylation of human and mouse genes

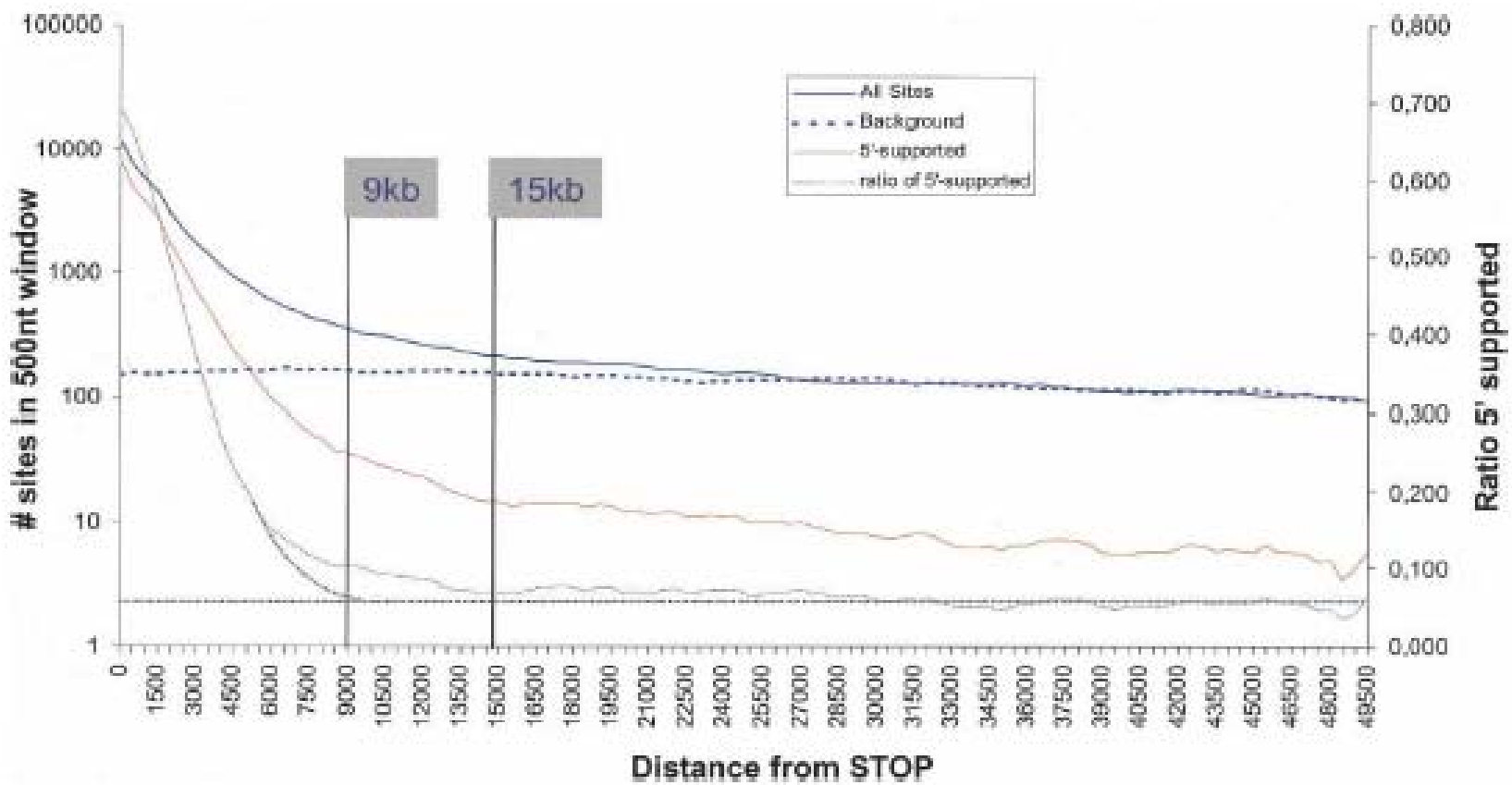
Bin Tian*, Jun Hu, Haibo Zhang[1] and Carol S. Lutz

★ No systemic study of intergenic sites yet

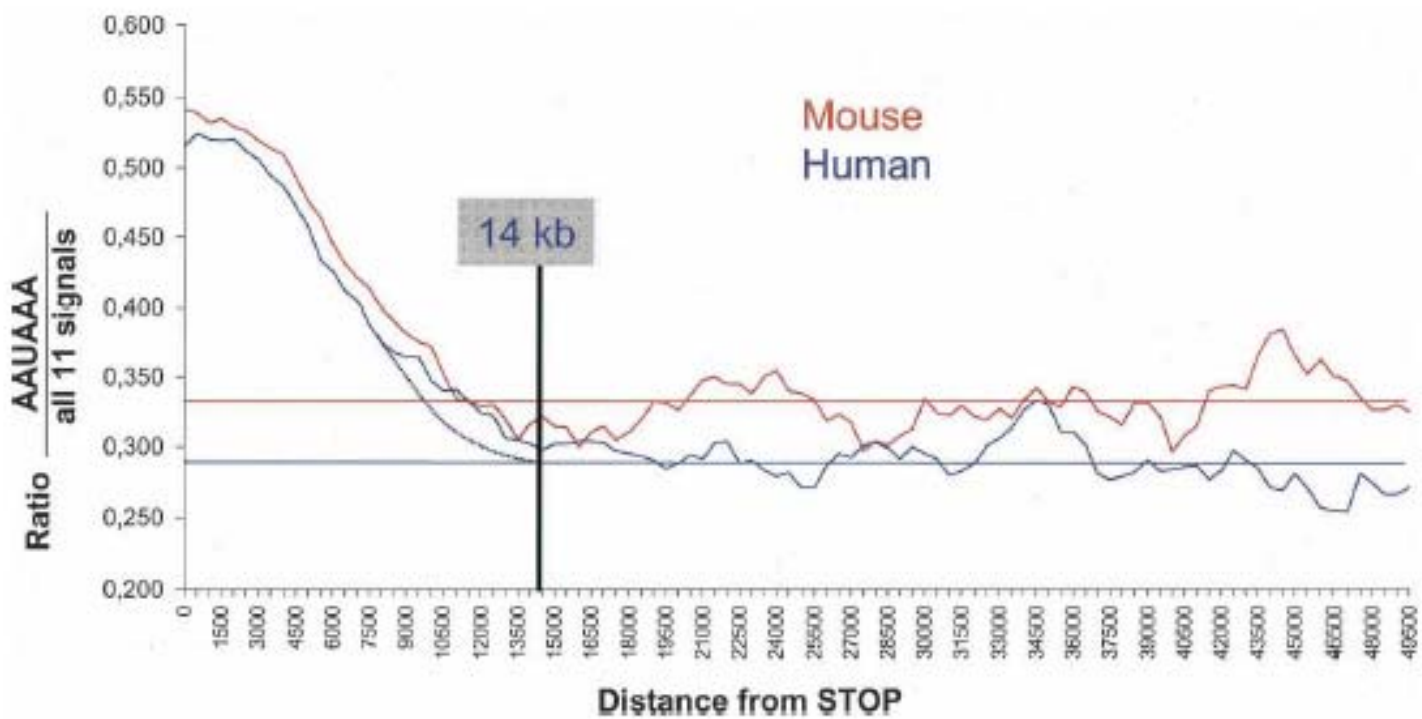
How can you make sure a PAS is true and pertains to the nearest 5’ gene ?

- ★ PAS could be just noise (>30000 PAS at >10kb from any Ensembl gene)
- ★ There could be another gene in the interval

Direct UTR counts



AAUAAA signal usage

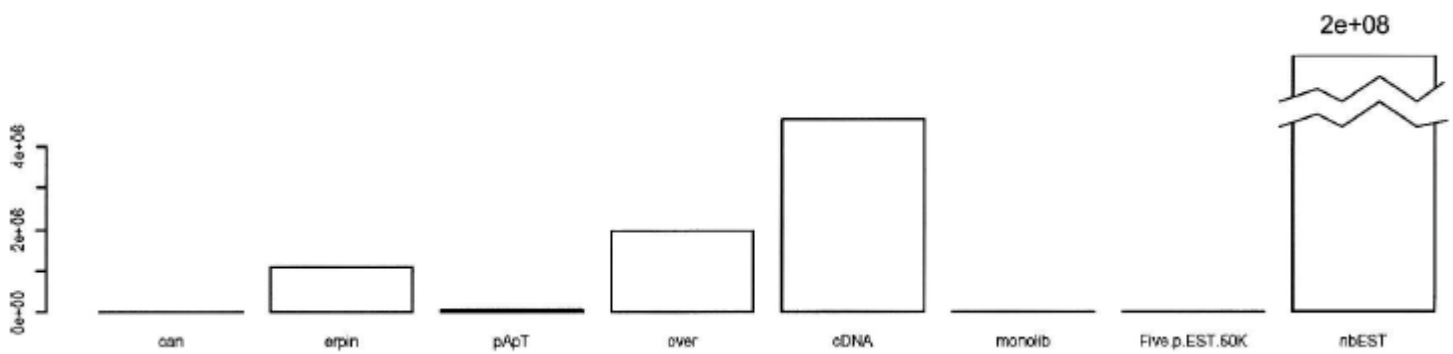


- ★ Noisy PAS are expected to use random polyA signals
- ★ Independent of EST coverage
- ★ True PAS appear dominant up to 15kb!

Evaluating PAS quality

Classification tree:

- #ESTs, fl cDNA, overlap of transcripts, profile score, canonical signal, polyA tail, 5'EST, multiple libraries.



Selected PAS or transcriptional leakage?

- 3’ UTR sizes from orthologues (Ensembl)….

UTR size in mouse

UTR size in human

bp	<100	<1k	<10k
<100	286	195	99
<1k	256	4396	1334
<10k	131	1527	3004

- Chi2 Probability ~ 0!
- Long UTR in human => long UTR in mouse

High quality PASs:

- 6000 UTR extensions
- 3500 intergenic transcripts
- 16000 signal-free sites

- Incidence on: microarray probe design, miRNA target search…

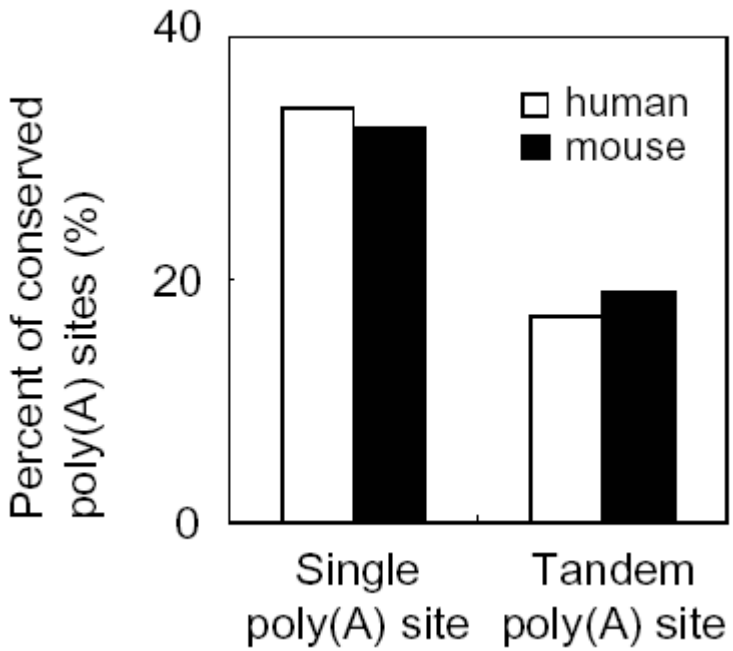
Conserved PAS

Define conserved PAS as PAS that is:

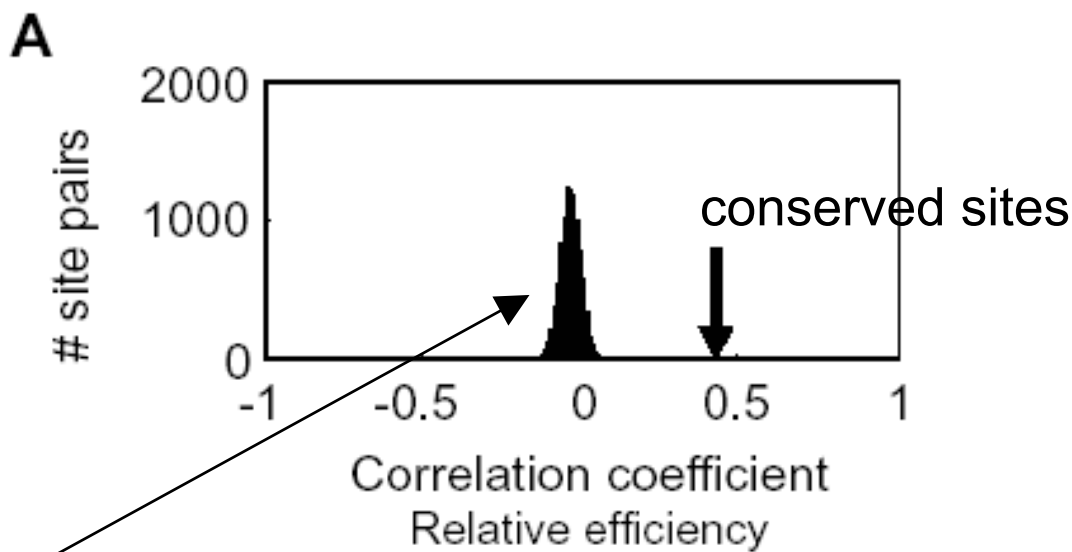
- ★ Aligned in orthologous UTR regions from human and mouse
- ★ EST-supported in both species

Tandem sites are generally less conserved than single sites

About 500 genes with 2 or more conserved tandem sites
-> most interesting functionally

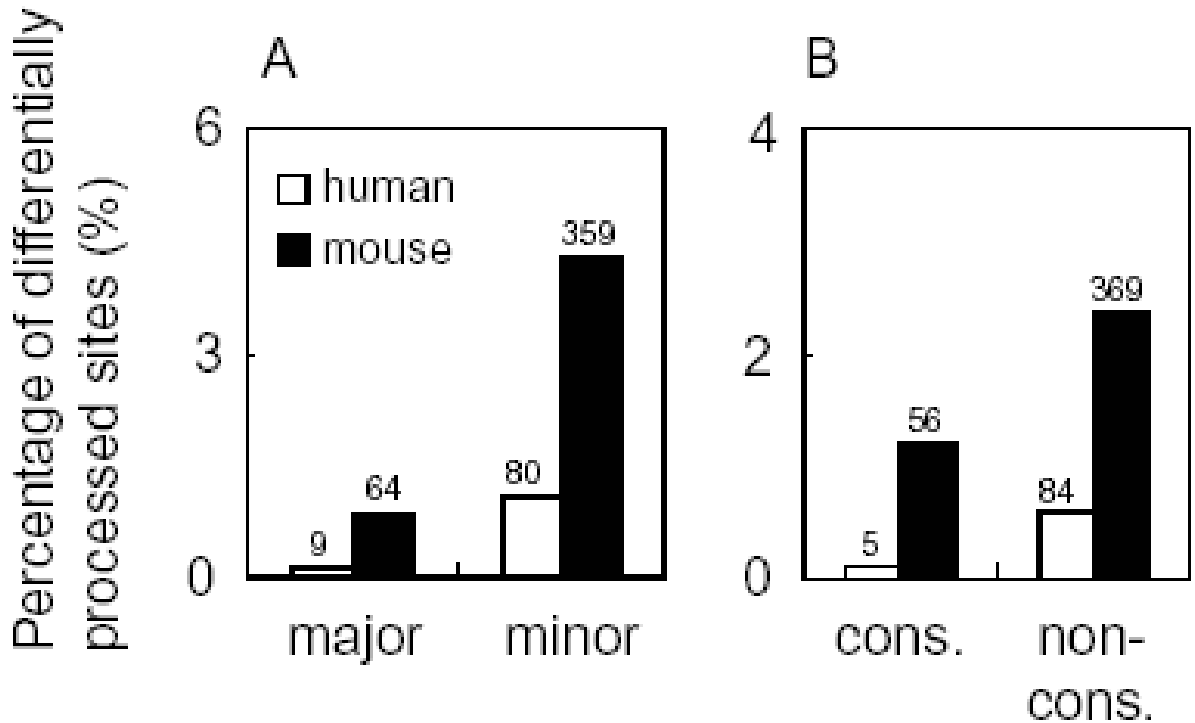


Conserved PAS in human and mouse have correlated processing efficiencies



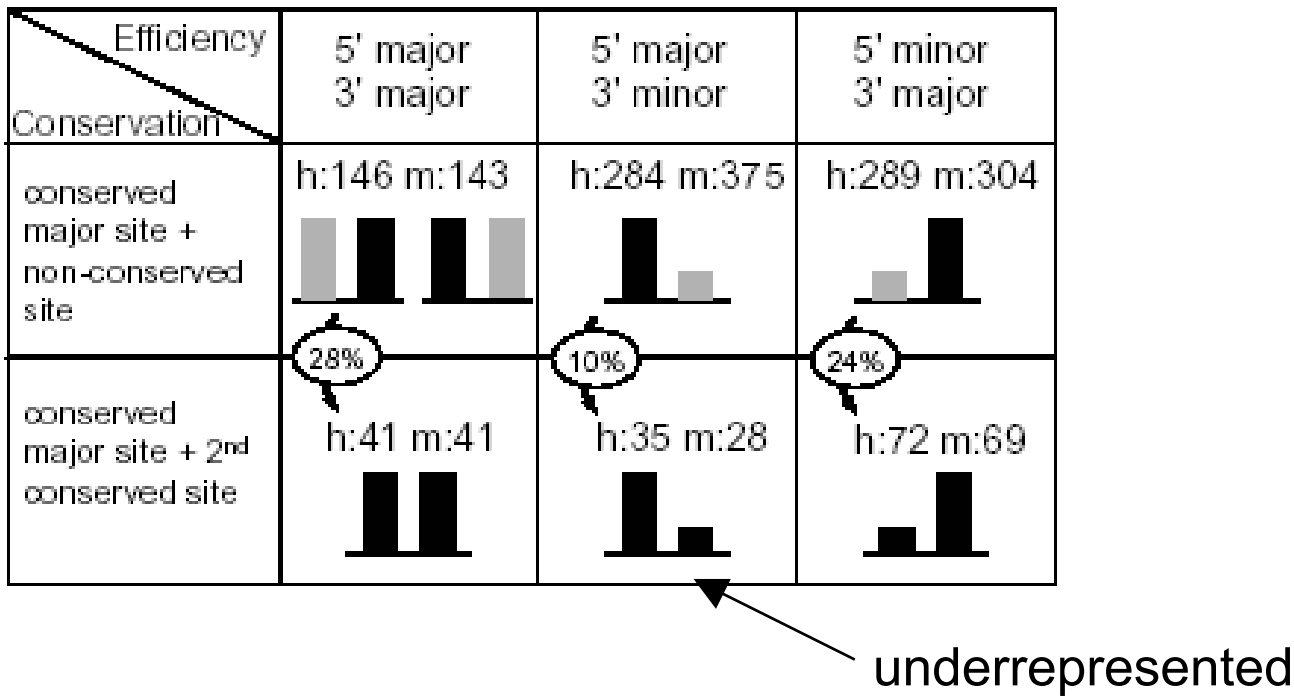
1000 pairs of non-conserved sites

Differentially processed sites are most often non-conserved



Alternative splicing is generally species-specific too (Modrek & Lee, 2003)

Selection favors downstream major sites



Consistent with a model where novel poly(A) sites arising 3' to existing sites tend to be lost more quickly, unless stronger than existing 5' sites.

Les besoins actuels

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

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

The ATD Project

Partners - it-atdp

File Edit View G

Home | Local | Institutions | Journaux | Mot/Annu | Cours/Guides | MolBio 1 | RNA | FP6-ATD | trad

ATD

Alternate Transcript Diversity

SIXTH FRAMEWORK PROGRAMME

- Literature
- News
- Contact Us
- Index

Tools

site Map

News: Daily | Weekly

Print | Tutorial |
Help | Sign In |

	ATD partner	Principal investigator
Inserm	**INSERM** *French Institute of Health and Medical Research* Marseille, FRANCE	**Daniel Gautheret**
	EMBL- EBI	
	European Molecular Biology Laboratory	**Peer Bork**
	Heidelberg, GERMANY	

	European Bioinformatics Institute	**Alphonse Thanaraj**
	Hinxton, Cambridge, UK	
UNIVERSITY of the WESTERN CAPE	**UWC** *University of Western Cape* Cape Town, SOUTH AFRICA	**Winston Hide**
Fundació IMIM	**FIMIM** *Fundacio IMIM* Barcelona, SPAIN	**Roderic Guigo**
ebc Estonian Biocentre	**EBC** *Estonian Biocentre* Tartu, ESTONIA	**Jaak Vilo**
Universitätsklinikum und Medizinische Fakultät Heidelberg	**UH** *Universitätklinikum Heidelberg* Heidelberg, GERMANY	**Magnus von Knebel Döeberitz**
MDC MAX-DELBRÜCK-CENTRUM FÜR MOLEKULARE MEDIZIN BERLIN-BUCH	**MDC** *Max-Delbrück-Centrum für Molekulare Medizin* Berlin-Buch, GERMANY	**Jens Reich**
InsermTransfert	**INSERM-TRANSFERT** Paris, FRANCE	**Christiane Dascher-Nadel**

- ✓ Integrate Splice+polyA+Init variants
- ✓ Quality control
- ✓ Tissue-specific Isoforms
- ✓ Regulatory motifs
- ✓ Isoform specific oligos
- ✓ RT-PCR validation of selected isoforms

Done

The ATD project is funded by the European Commission within its FP6 Programme, under the thematic area "Life sciences, genomics and biotechnology for health", contract number LHSG-CT-2003-503329