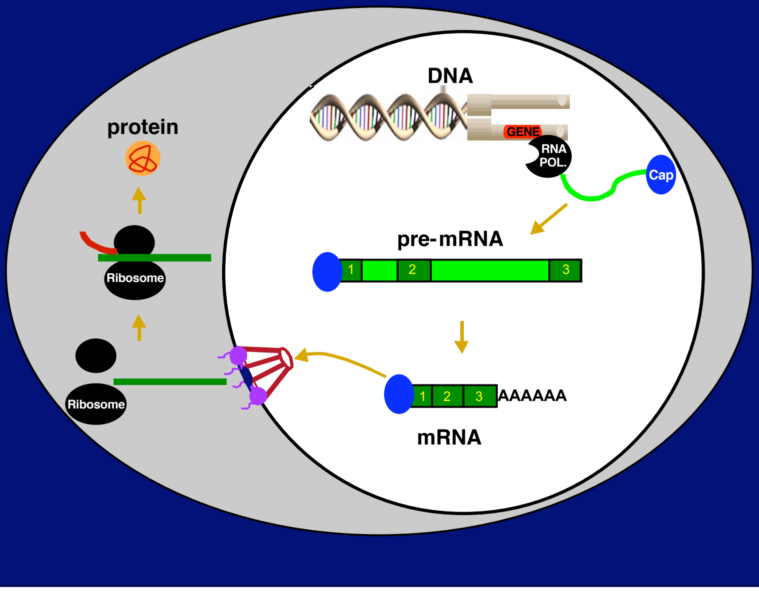
Computational Molecular Biology of Genome Expression and Regulation

Michael Q. Zhang
Cold Spring Harbor Labratory

- An ESE (enhancer-sequence-element) SNP can alter RNA splicing (Brac1:exon18)
- Classification of 5'UTRs by *CART*
- Promoter prediction and analysis
- *CSEdb* and human gene number

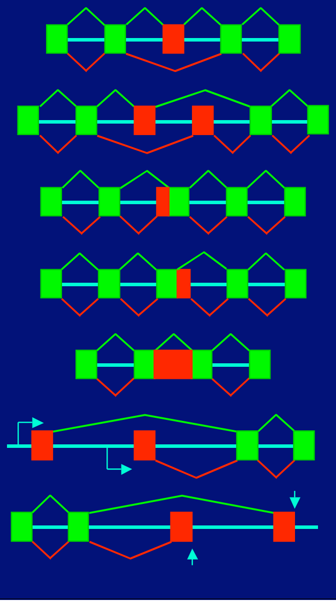
Classical view of eukaryotic gene expression



An ESE (enhancer-sequence-element) SNP can alter RNA splicing

(Collaborated with Krainer lab at CSHL: *Nature Genet*. 2001)

Patterns of alternative pre-mRNA splicing



Exon inclusion/skipping

Mutually exclusive exons

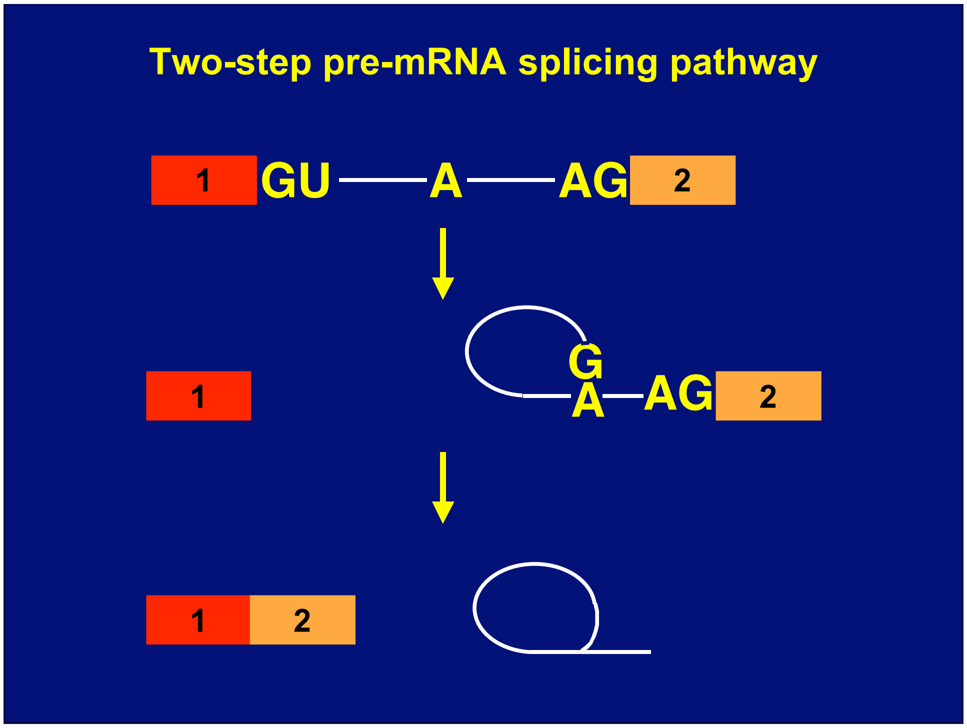
Alternative 3' splice sites

Alternative 5' splice sites

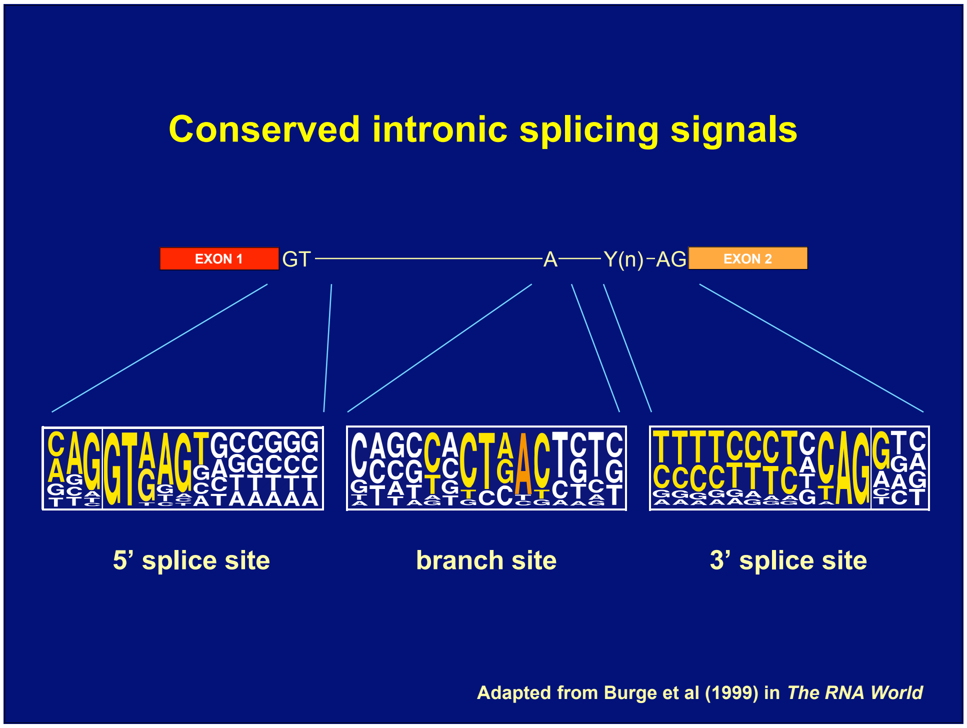
Retained intron

Alternative promoters

Alternative polyadenylation sites



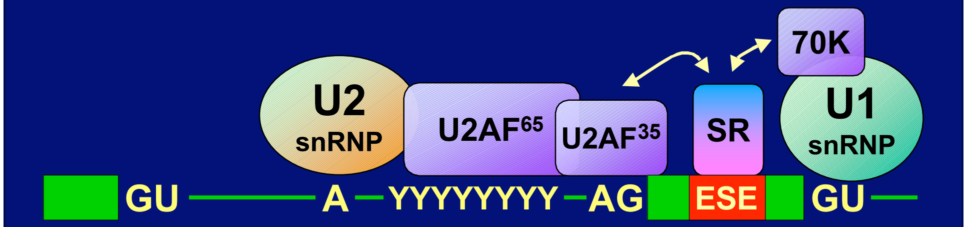
Two-step pre-mRNA splicing pathway

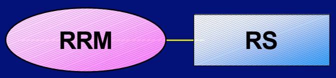


Conserved intronic splicing signals

Adapted from Burge et al (1999) in *The RNA World*

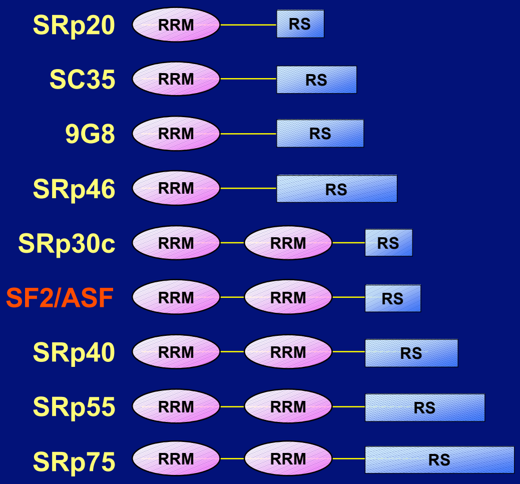
ESE-dependent splicing

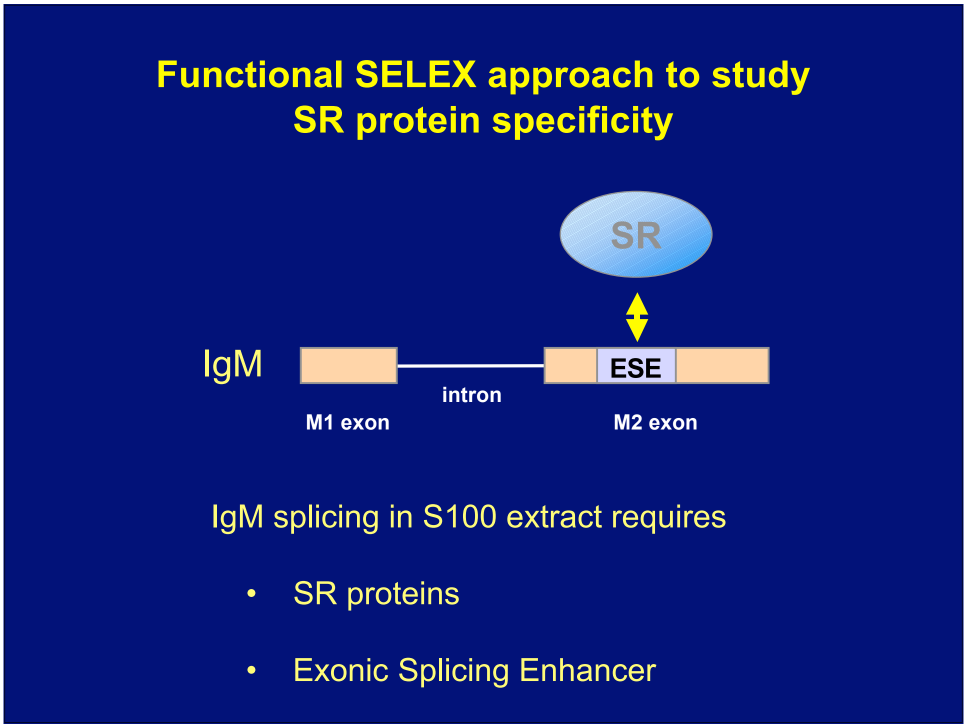




SR protein structure

SR proteins

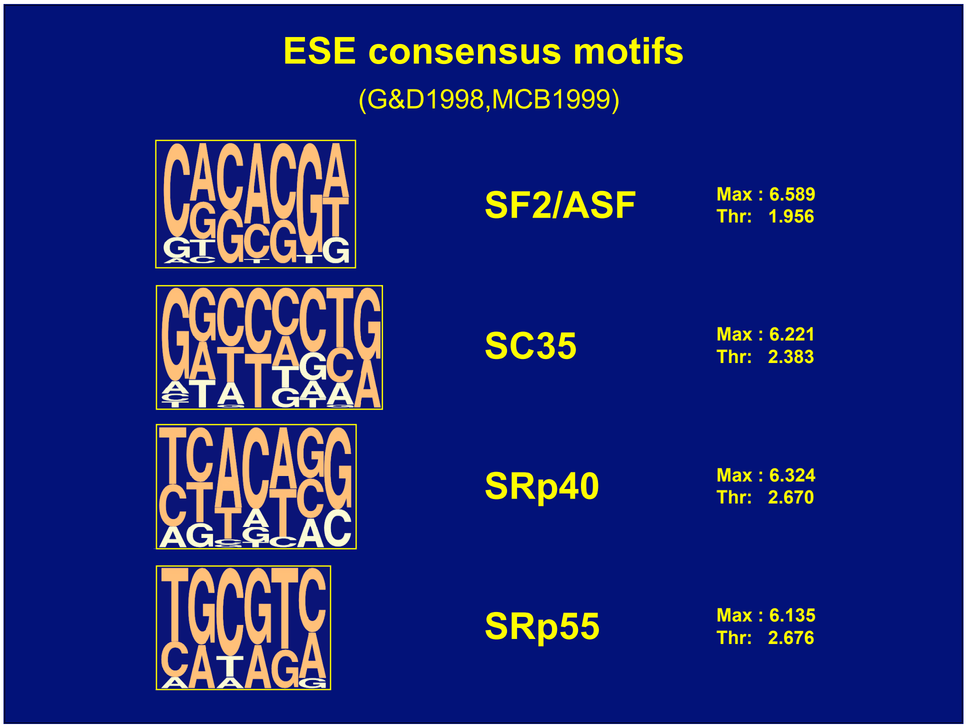




Functional SELEX approach to study SR protein specificity

IgM splicing in S100 extract requires

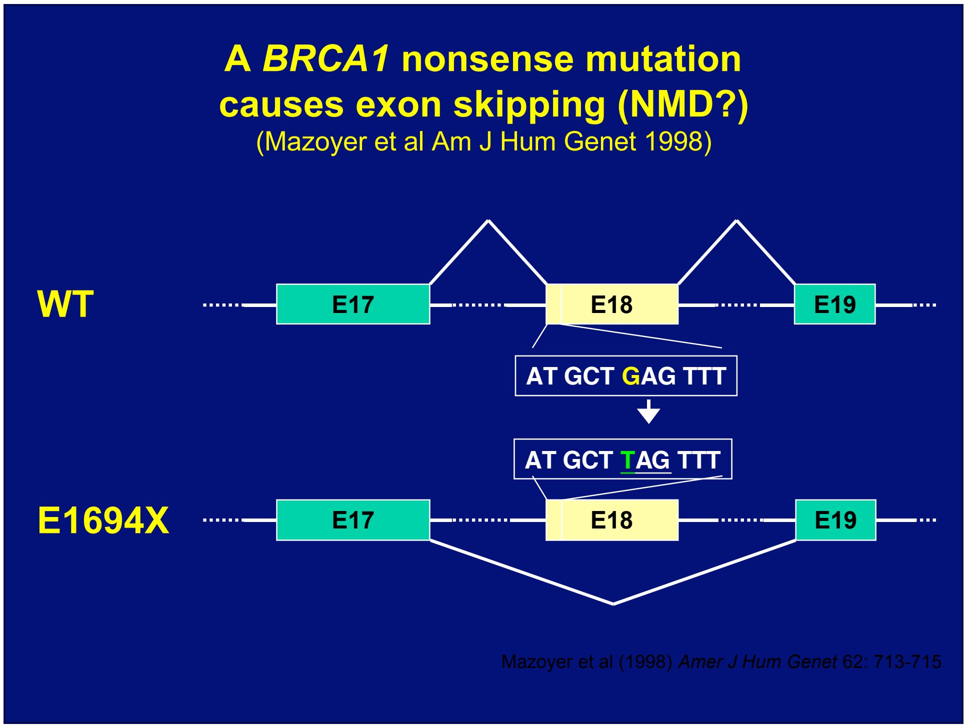
- SR proteins
- Exonic Splicing Enhancer



ESE consensus motifs

(G&D1998,MCB1999)

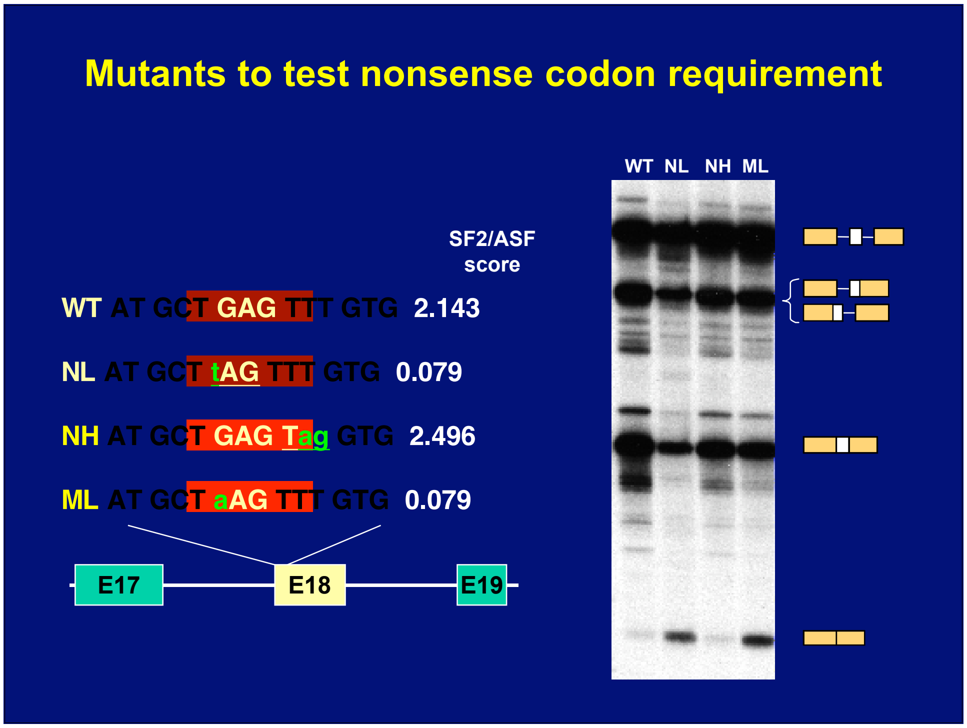
	SF2/ASF	Max : 6.589 Thr: 1.956
	SC35	Max : 6.221 Thr: 2.383
	SRp40	Max : 6.324 Thr: 2.670
	SRp55	Max : 6.135 Thr: 2.676



A *BRCA1* nonsense mutation causes exon skipping (NMD?)

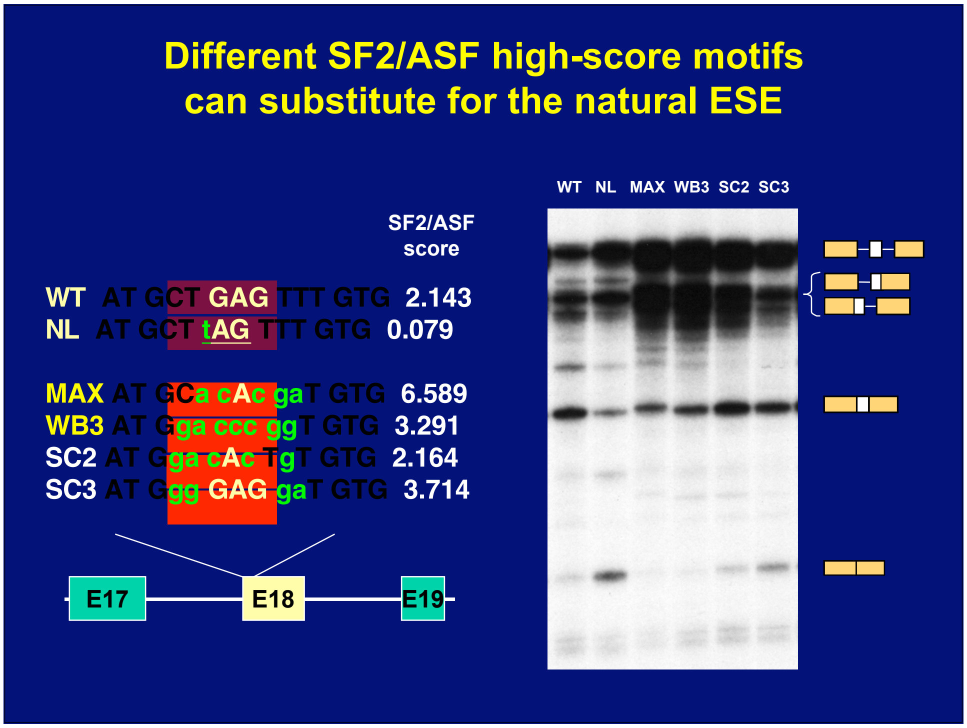
(Mazoyer et al Am J Hum Genet 1998)

Mazoyer et al (1998) *Amer J Hum Genet* 62: 713-715



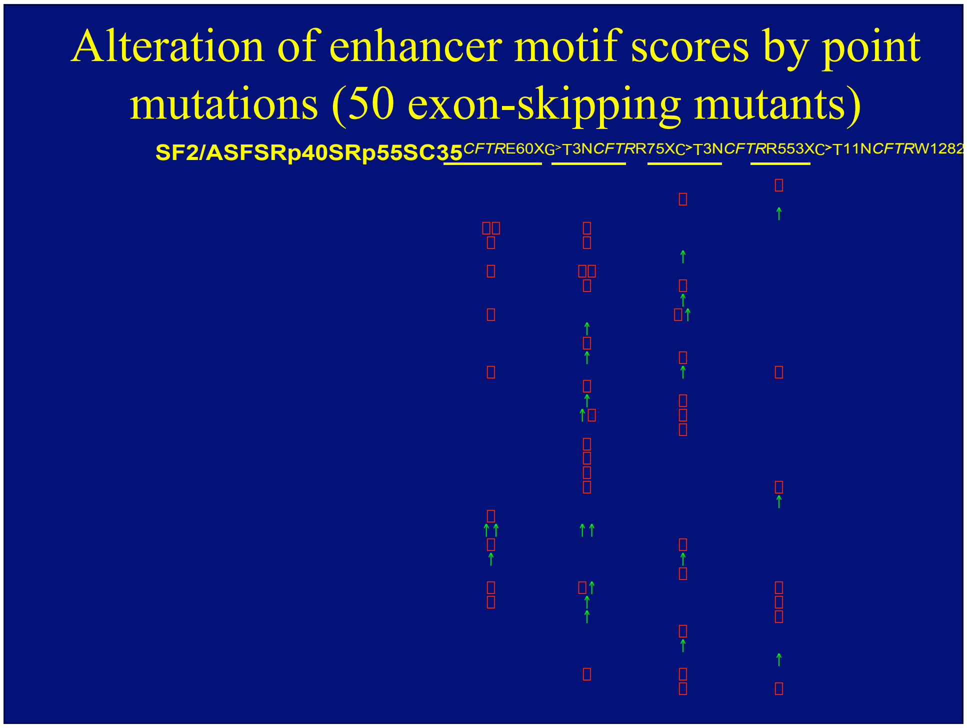
Mutants to test nonsense codon requirement

		SF2/ASF score
WT	AT GCT GAG TTT GTG	2.143
NL	AT GCT tAG TTT GTG	0.079
NH	AT GCT GAG Tag GTG	2.496
ML	AT GCT aAG TTT GTG	0.079



Different SF2/ASF high-score motifs can substitute for the natural ESE

		SF2/ASF score
WT	AT GCT GAG TTT GTG	2.143
NL	AT GCT tAG TTT GTG	0.079
MAX	AT GCa cAc gaT GTG	6.589
WB3	AT Gga ccc ggT GTG	3.291
SC2	AT Gga cAc TgT GTG	2.164
SC3	AT Ggg GAG gaT GTG	3.714

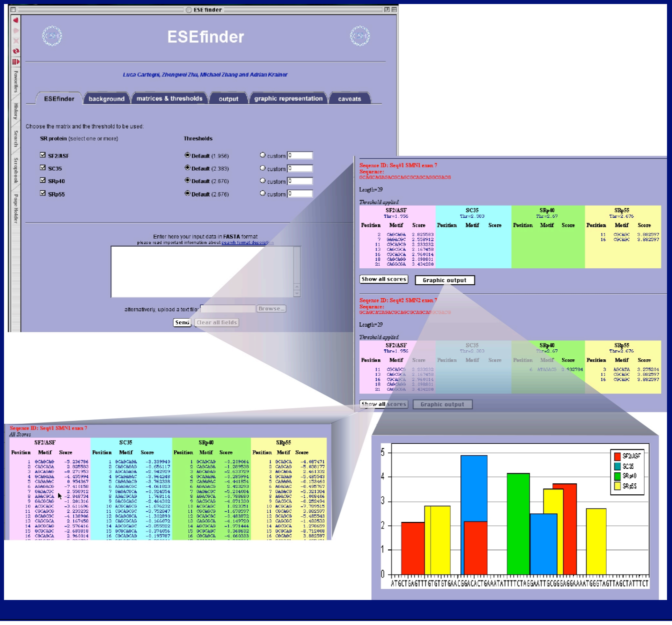


Alteration of enhancer motif scores by point mutations (50 exon-skipping mutants)

SF2/ASFSRp40SRp55SC35*CFTR*E60XG>T3N*CFTR*R75XC>T3N*CFTR*R553XC>T11N*CFTR*W1282

ESEfinder: a web resource to identify ESEs

(http://exon.cshl.edu/ESE/ Cartegni et al. submitted)



ESEfinder

ESEfinder

Luca Cartegni, Zhengwei Zhu, Michael Zhang and Adrian Krainer

ESEfinder | background | matrices & thresholds | output | graphic representation | caveats

Choose the matrix and the threshold to be used:

SR protein (select one or more)	Thresholds	
☑ SF2/ASF	● Default (1.956)	○ custom 0
☑ SC35	● Default (2.383)	○ custom 0
☑ SRp40	● Default (2.670)	○ custom 0
☑ SRp55	● Default (2.676)	○ custom 0

Enter here your input data in FASTA format
please read important information about search format description

alternatively, upload a text file: Browse...

Send | Clear all fields

Seqence ID: Seq#1 SMN1 exon 7
Sequence:
GCAGCAGAGACGCAGCGCAGCAGGCGACG

Length=29

Threshold applied

SF2/ASF Thr=1.956			SC35 Thr=2.383			SRp40 Thr=2.67			SRp55 Thr=2.676		
Position	Motif	Score	Position	Motif	Score	Position	Motif	Score	Position	Motif	Score
2	CAGCAGA	2.825583							11	CGCAGC	3.882597
7	GAGACGC	2.558912							16	CGCAGC	3.882597
11	CGCAGCG	2.233232									
13	CAGCGCA	2.167458									
16	CGCAGCA	2.960014									
18	CAGCAGG	2.098801									
21	CAGGCGA	3.434280									

Show all scores | Graphic output

Seqence ID: Seq#2 SMN2 exon 7
Sequence:
GCAGCATAGACGCAGCGCAGCAGGCGACG

Length=29

Threshold applied

SF2/ASF Thr=1.956			SC35 Thr=2.383			SRp40 Thr=2.67			SRp55 Thr=2.676		
Position	Motif	Score	Position	Motif	Score	Position	Motif	Score	Position	Motif	Score
11	CGCAGCG	2.233232				6	ATAGACG	2.932784	3	AGCATA	3.275204
13	CAGCGCA	2.167458							11	CGCAGC	3.882597
16	CGCAGCA	2.960014							16	CGCAGC	3.882597
18	CAGCAGG	2.098801									
21	CAGGCGA	3.434280									

Show all scores | Graphic output

Seqence ID: Seq#1 SMN1 exon 7
All Scores

SF2/ASF			SC35			SRp40			SRp55		
Position	Motif	Score	Position	Motif	Score	Position	Motif	Score	Position	Motif	Score
1	GCAGCAG	-5.236786	1	GCAGCAGA	-3.339940	1	GCAGCAG	-0.209064	1	GCAGCA	-4.087471
2	CAGCAGA	2.825583	2	CAGCAGAG	-0.656117	2	CAGCAGA	-1.289538	2	CAGCAG	-5.038177
3	AGCAGAG	-0.271953	3	AGCAGAGA	-2.942929	3	AGCAGAG	-2.633729	3	AGCAGA	2.461332
4	GCAGAGA	-4.455994	4	GCAGAGAC	-3.964248	4	GCAGAGA	-0.285994	4	GCAGAG	-5.485543
5	CAGAGAC	0.954367	5	CAGAGACG	-3.762338	5	CAGAGAC	-4.441854	5	CAGAGA	-0.153460
6	AGAGACG	-7.410158	6	AGAGACGC	-4.061823	6	AGAGACG	2.423293	6	AGAGAC	-0.495767
7	GAGACGC	2.558912	7	GAGACGCA	-0.924554	7	GAGACGC	-5.204084	7	GAGACG	-5.321384
8	AGACGCA	-2.848734	8	AGACGCAG	1.768114	8	AGACGCA	-0.788680	8	AGACGC	-1.008406
9	GACGCAG	-1.201316	9	GACGCAGC	-2.464302	9	GACGCAG	-4.871330	9	GACGCA	-0.252494
10	ACGCAGC	-3.611696	10	ACGCAGCG	-1.676232	10	ACGCAGC	1.823351	10	ACGCAG	-7.789515
11	CGCAGCG	2.233232	11	CGCAGCGC	-3.752847	11	CGCAGCG	-1.678597	11	CGCAGC	3.882597
12	GCAGCGC	-4.138906	12	GCAGCGCA	-1.302890	12	GCAGCGC	-0.483872	12	GCAGCG	-5.485543
13	CAGCGCA	2.167458	13	CAGCGCAG	-0.166072	13	CAGCGCA	-4.149720	13	CAGCGC	-1.483533
14	AGCGCAG	-2.576416	14	AGCGCAGC	-3.855522	14	AGCGCAG	-1.971444	14	AGCGCA	1.278659
15	GCGCAGC	-2.683818	15	GCGCAGCA	-0.374056	15	GCGCAGC	0.368632	15	GCGCAG	-8.712008
16	CGCAGCA	2.960014	16	CGCAGCAG	-0.195787	16	CGCAGCA	-4.060333	16	CGCAGC	3.882597

- ***ASDB*** **(with Stamm/Max-Plank, Nakai/Kyoto, 2001)**
- ***mATDB*** **(Wang & Zhang, to be submitted)**
- **The RNA-mediated Annealing, Selection and Ligation Assay (RASL): Application in Alternative Splicing** (with Fu&Gribskov/UCSD and Fan/Illumina, NCI funded)

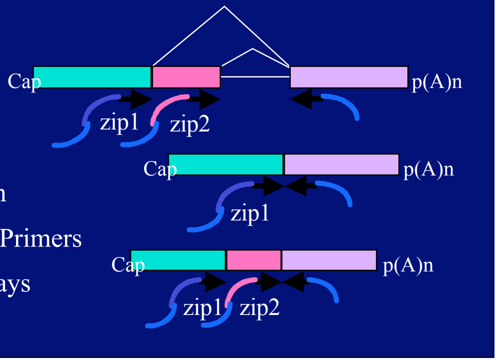
Step 1: Oligo Annealing

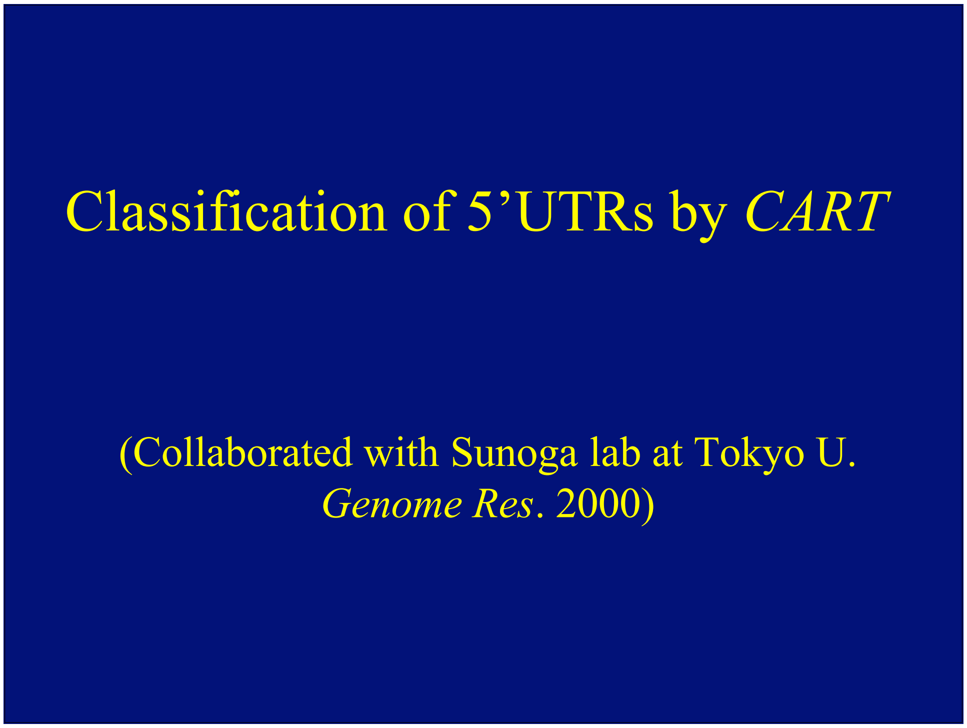
Step 2: Poly A^+ Selection

Step 3: RNA-dependent Oligo Ligation

Step 4: Amplification using Universal Primers

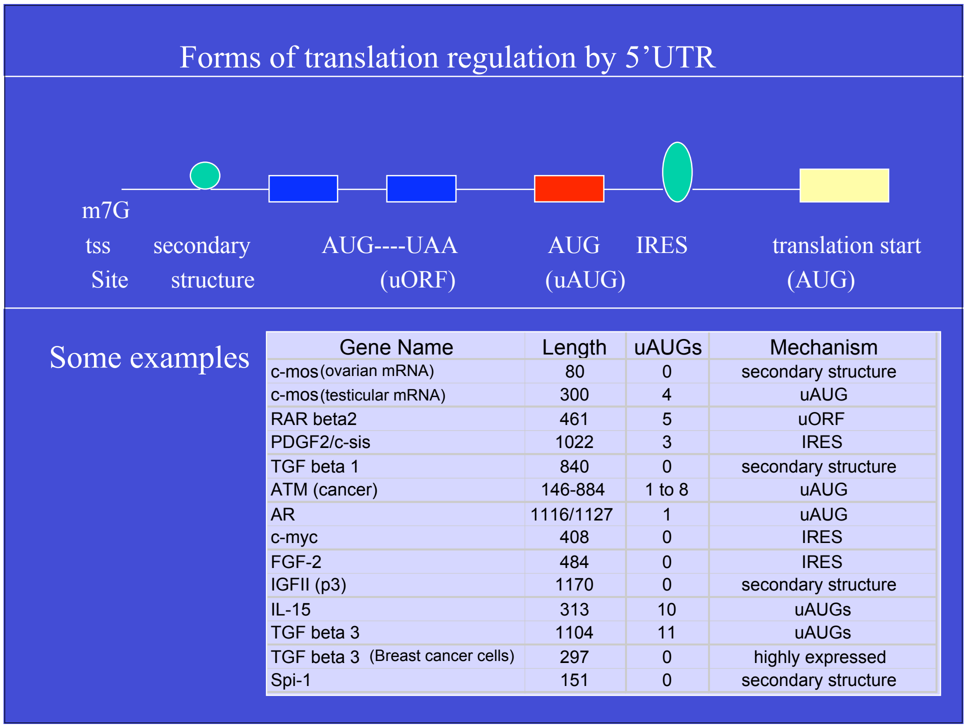
Step 5: Hybridization to Zip-Code Arrays





Classification of 5’UTRs by *CART*

(Collaborated with Sunoga lab at Tokyo U.
Genome Res. 2000)



Forms of translation regulation by 5’UTR

m7G
tss Site

secondary structure

AUG----UAA (uORF)

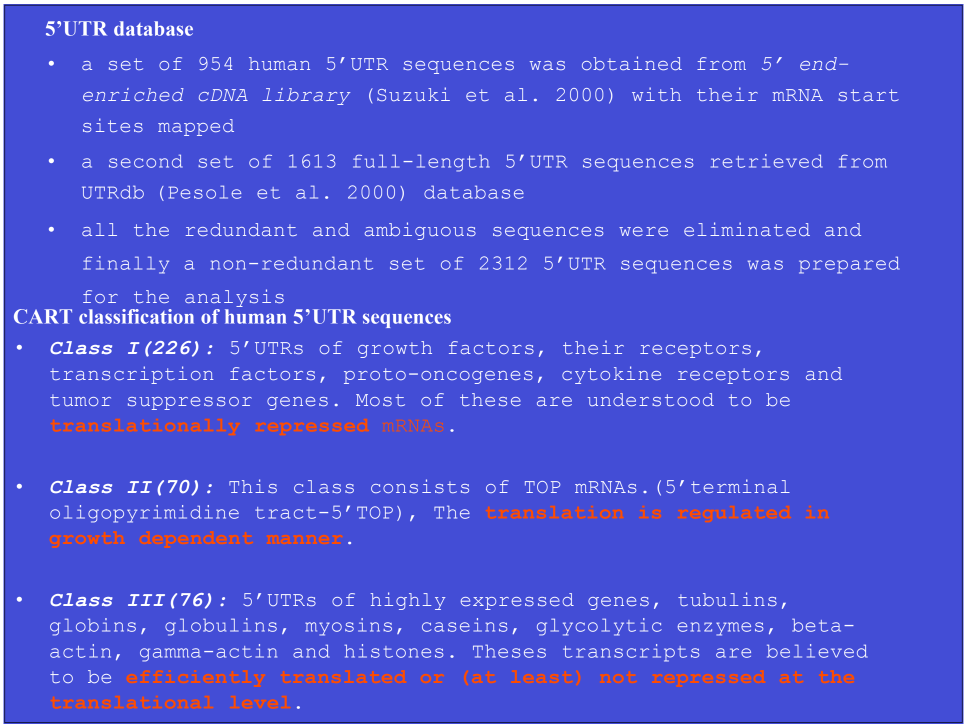
AUG (uAUG)

IRES

translation start (AUG)

Some examples

Gene Name	Length	uAUGs	Mechanism
c-mos (ovarian mRNA)	80	0	secondary structure
c-mos (testicular mRNA)	300	4	uAUG
RAR beta2	461	5	uORF
PDGF2/c-sis	1022	3	IRES
TGF beta 1	840	0	secondary structure
ATM (cancer)	146-884	1 to 8	uAUG
AR	1116/1127	1	uAUG
c-myc	408	0	IRES
FGF-2	484	0	IRES
IGFII (p3)	1170	0	secondary structure
IL-15	313	10	uAUGs
TGF beta 3	1104	11	uAUGs
TGF beta 3 (Breast cancer cells)	297	0	highly expressed
Spi-1	151	0	secondary structure

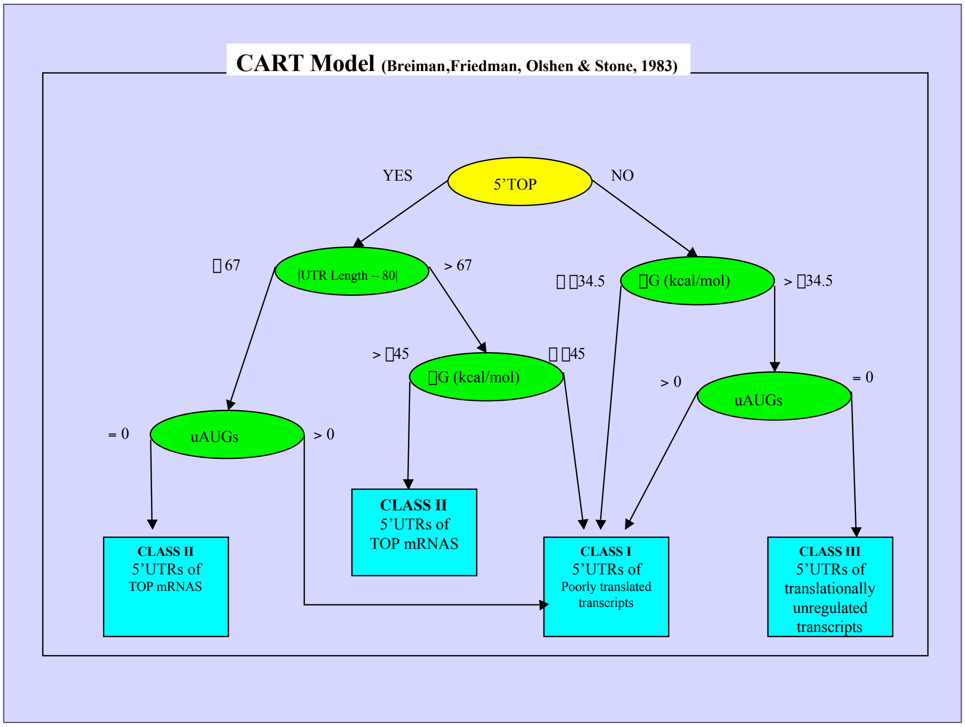


5’UTR database

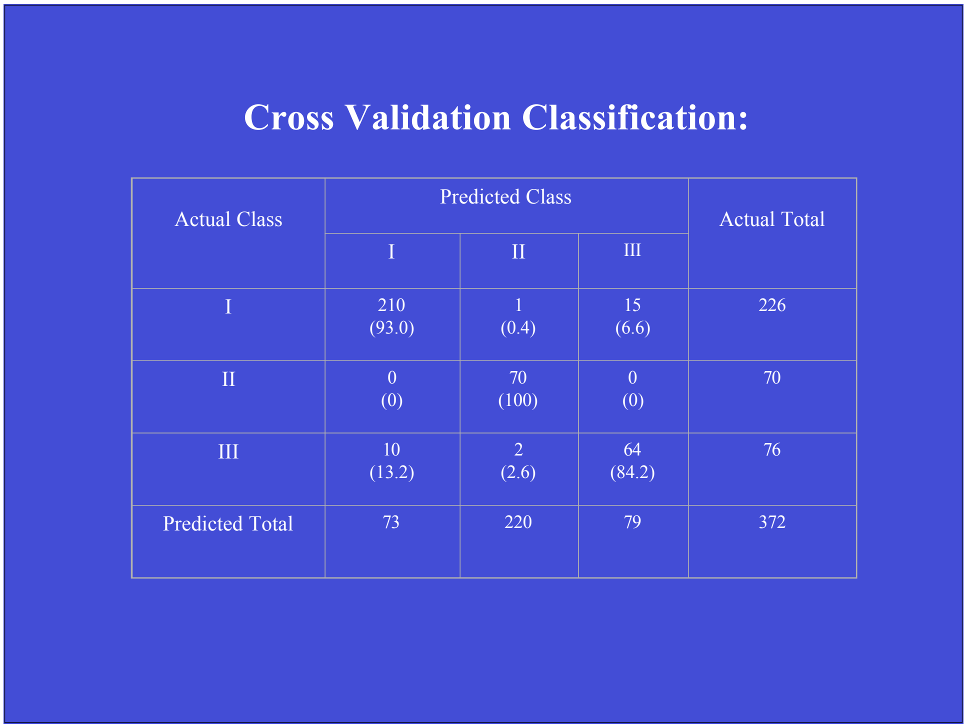
- a set of 954 human 5’UTR sequences was obtained from *5’ end-enriched cDNA library* (Suzuki et al. 2000) with their mRNA start sites mapped
- a second set of 1613 full-length 5’UTR sequences retrieved from UTRdb (Pesole et al. 2000) database
- all the redundant and ambiguous sequences were eliminated and finally a non-redundant set of 2312 5’UTR sequences was prepared for the analysis

CART classification of human 5’UTR sequences

- ***Class I(226):*** 5’UTRs of growth factors, their receptors, transcription factors, proto-oncogenes, cytokine receptors and tumor suppressor genes. Most of these are understood to be **translationally repressed** mRNAs.
- ***Class II(70):*** This class consists of TOP mRNAs.(5’terminal oligopyrimidine tract-5’TOP), The **translation is regulated in growth dependent manner**.
- ***Class III(76):*** 5’UTRs of highly expressed genes, tubulins, globins, globulins, myosins, caseins, glycolytic enzymes, beta-actin, gamma-actin and histones. Theses transcripts are believed to be **efficiently translated or (at least) not repressed at the translational level**.

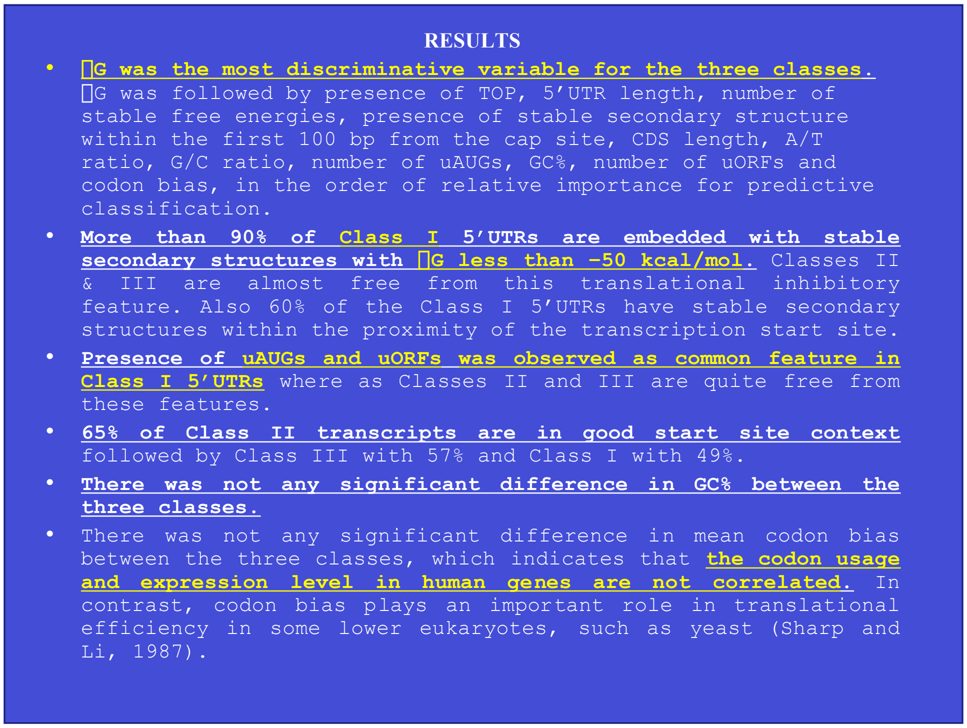


CART Model (Breiman,Friedman, Olshen & Stone, 1983)



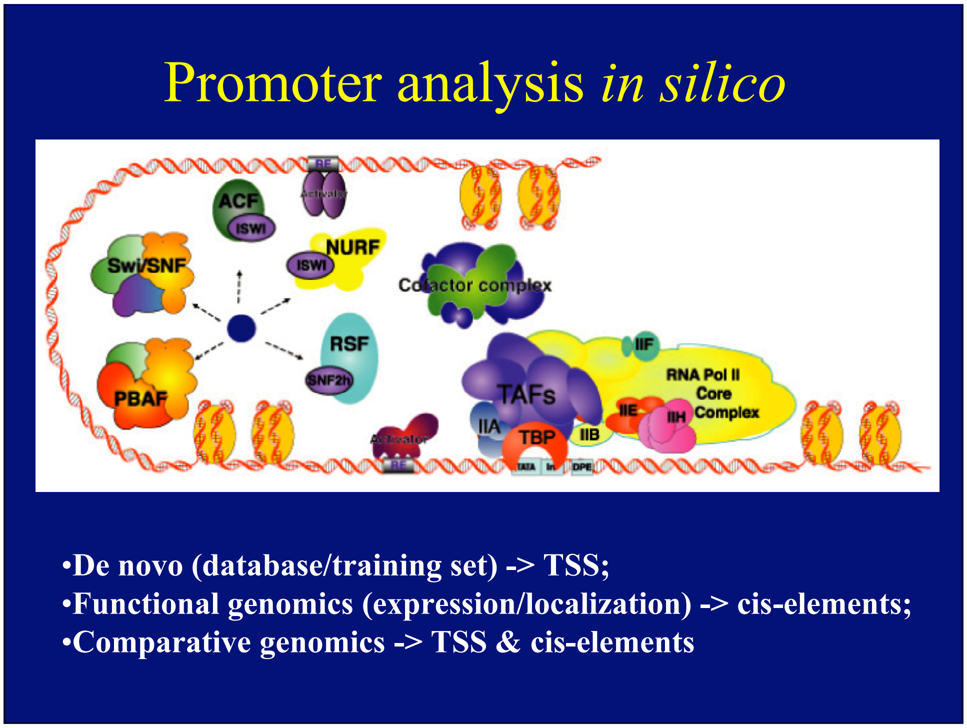
Cross Validation Classification:

Actual Class	Predicted Class			Actual Total
	I	II	III	
I	210 (93.0)	1 (0.4)	15 (6.6)	226
II	0 (0)	70 (100)	0 (0)	70
III	10 (13.2)	2 (2.6)	64 (84.2)	76
Predicted Total	73	220	79	372



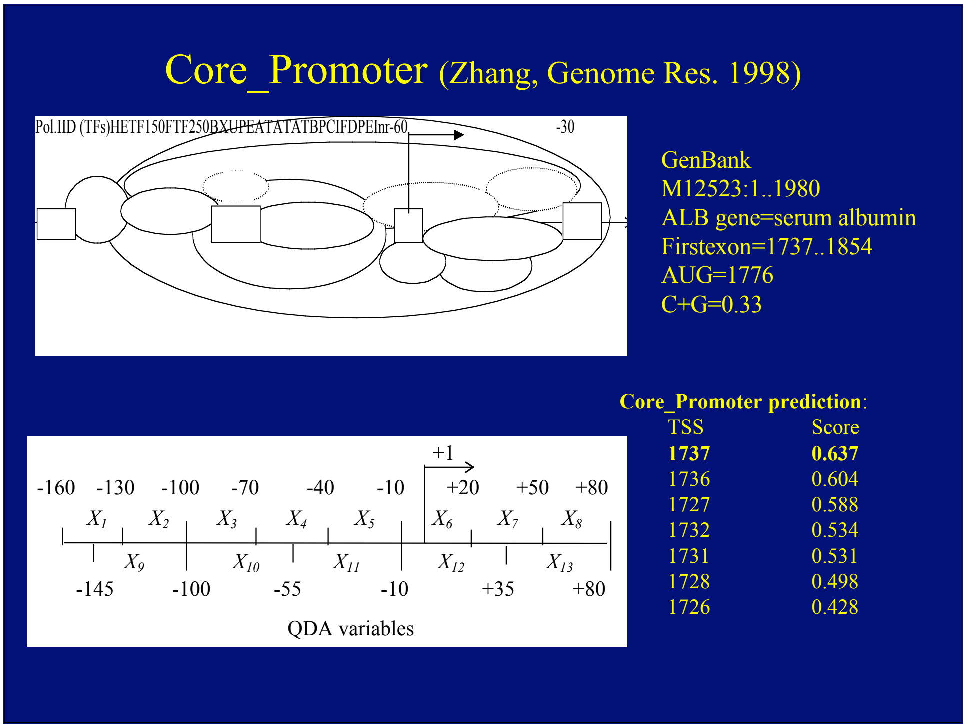
RESULTS

- **ΔG was the most discriminative variable for the three classes.** ΔG was followed by presence of TOP, 5′UTR length, number of stable free energies, presence of stable secondary structure within the first 100 bp from the cap site, CDS length, A/T ratio, G/C ratio, number of uAUGs, GC%, number of uORFs and codon bias, in the order of relative importance for predictive classification.
- **More than 90% of Class I 5′UTRs are embedded with stable secondary structures with ΔG less than -50 kcal/mol.** Classes II & III are almost free from this translational inhibitory feature. Also 60% of the Class I 5′UTRs have stable secondary structures within the proximity of the transcription start site.
- **Presence of uAUGs and uORFs was observed as common feature in Class I 5′UTRs** where as Classes II and III are quite free from these features.
- **65% of Class II transcripts are in good start site context** followed by Class III with 57% and Class I with 49%.
- **There was not any significant difference in GC% between the three classes.**
- There was not any significant difference in mean codon bias between the three classes, which indicates that **the codon usage and expression level in human genes are not correlated.** In contrast, codon bias plays an important role in translational efficiency in some lower eukaryotes, such as yeast (Sharp and Li, 1987).



Promoter analysis *in silico*

- •De novo (database/training set) -> TSS;
- •Functional genomics (expression/localization) -> cis-elements;
- •Comparative genomics -> TSS & cis-elements



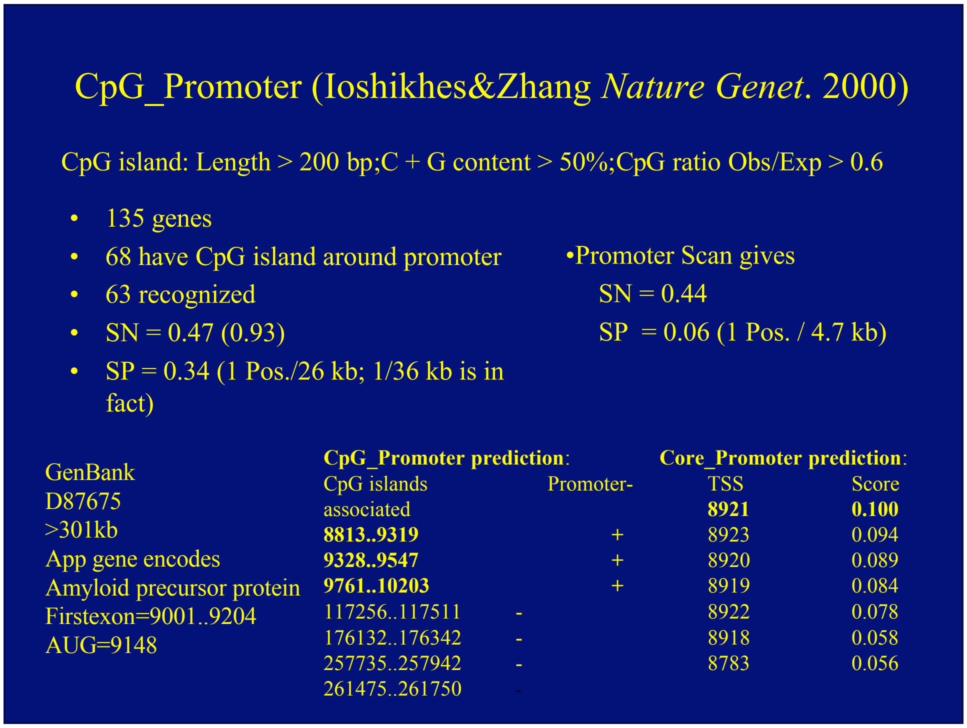
Core_Promoter (Zhang, Genome Res. 1998)

GenBank
M12523:1..1980
ALB gene=serum albumin
Firstexon=1737..1854
AUG=1776
C+G=0.33

QDA variables

Core_Promoter prediction:

TSS	Score
1737	**0.637**
1736	0.604
1727	0.588
1732	0.534
1731	0.531
1728	0.498
1726	0.428



CpG_Promoter (Ioshikhes&Zhang *Nature Genet*. 2000)

CpG island: Length > 200 bp;C + G content > 50%;CpG ratio Obs/Exp > 0.6

- 135 genes
- 68 have CpG island around promoter
- 63 recognized
- SN = 0.47 (0.93)
- SP = 0.34 (1 Pos./26 kb; 1/36 kb is in fact)

•Promoter Scan gives
SN = 0.44
SP = 0.06 (1 Pos. / 4.7 kb)

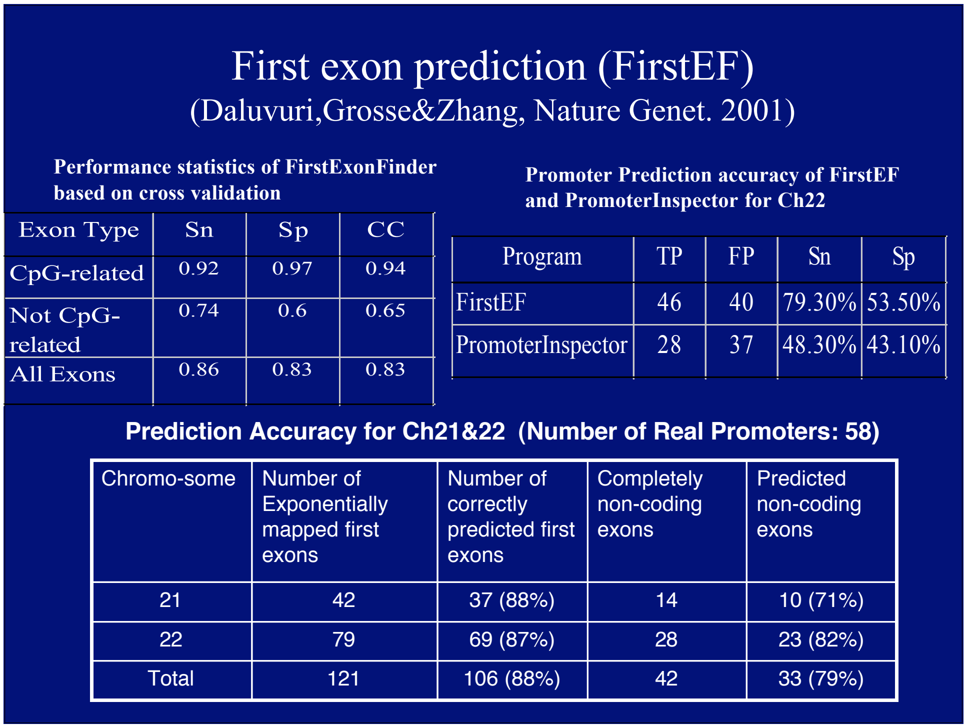
GenBank
D87675
>301kb
App gene encodes
Amyloid precursor protein
Firstexon=9001..9204
AUG=9148

CpG_Promoter prediction:

CpG islands	Promoter-associated
8813..9319	+
9328..9547	+
9761..10203	+
117256..117511	-
176132..176342	-
257735..257942	-
261475..261750	-

Core_Promoter prediction:

TSS	Score
8921	**0.100**
8923	0.094
8920	0.089
8919	0.084
8922	0.078
8918	0.058
8783	0.056



First exon prediction (FirstEF)

(Daluvuri,Grosse&Zhang, Nature Genet. 2001)

Performance statistics of FirstExonFinder based on cross validation

Exon Type	Sn	Sp	CC
CpG-related	0.92	0.97	0.94
Not CpG-related	0.74	0.6	0.65
All Exons	0.86	0.83	0.83

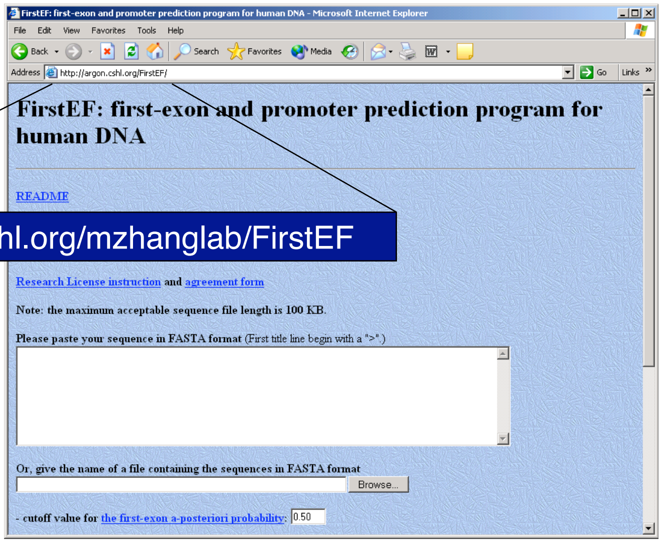
Promoter Prediction accuracy of FirstEF and PromoterInspector for Ch22

Program	TP	FP	Sn	Sp
FirstEF	46	40	79.30%	53.50%
PromoterInspector	28	37	48.30%	43.10%

Prediction Accuracy for Ch21&22 (Number of Real Promoters: 58)

Chromo-some	Number of Exponentially mapped first exons	Number of correctly predicted first exons	Completely non-coding exons	Predicted non-coding exons
21	42	37 (88%)	14	10 (71%)
22	79	69 (87%)	28	23 (82%)
Total	121	106 (88%)	42	33 (79%)

FirstEF program


FirstEF: first-exon and promoter prediction program for human DNA - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address
Go
Links
Browse...

http://argon.cshl.org/FirstEF/

FirstEF: first-exon and promoter prediction program for human DNA

README

http://www.cshl.org/mzhanglab/FirstEF

Research License instruction and agreement form

Note: the maximum acceptable sequence file length is 100 KB.

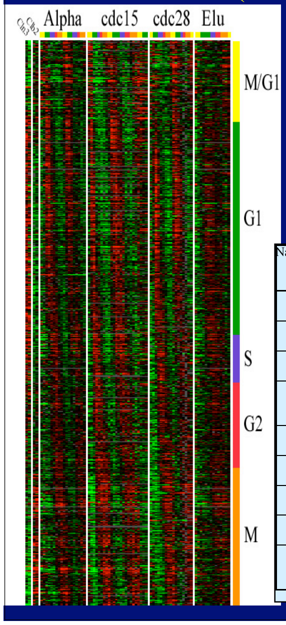
Please paste your sequence in FASTA format (First title line begin with a ">".)

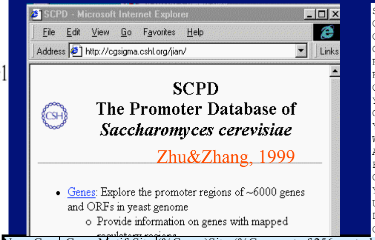
Or, give the name of a file containing the sequences in FASTA format

- cutoff value for the first-exon a-posteriori probability: 0.50

Cell Cycle Regulation (Spellman *et al*. MBC,1998)

(Brown&Botstein labs – Futcher&Zhang labs)




SCPD - Microsoft Internet Explorer
File Edit View Go Favorites Help
Address
Links

http://cgsigma.cshl.org/jian/

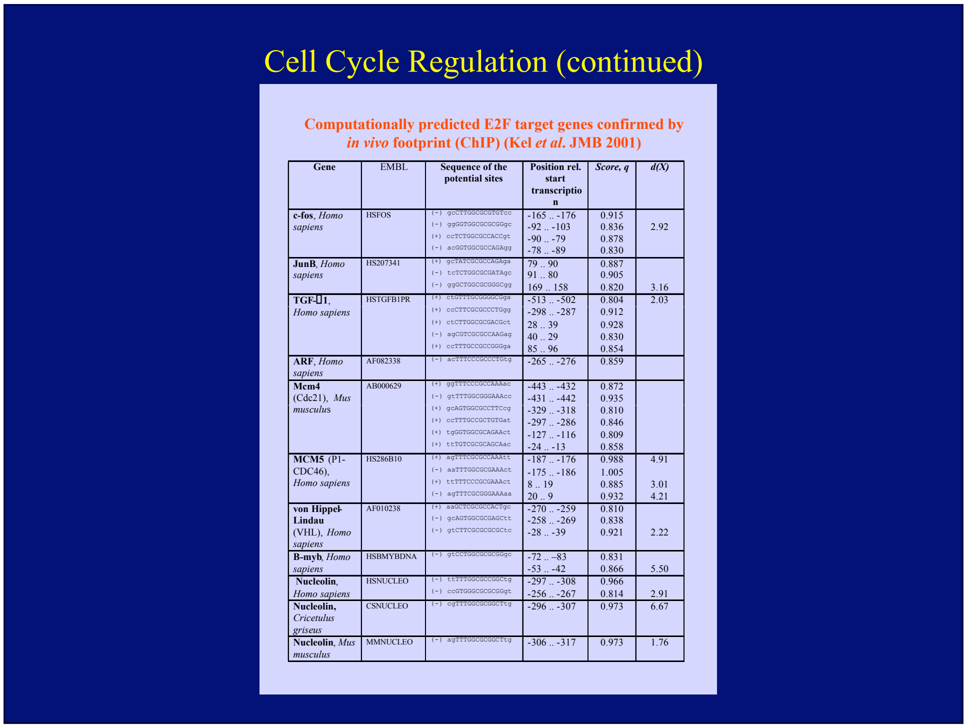
SCPD
The Promoter Database of *Saccharomyces cerevisiae*

Zhu&Zhang, 1999

- Genes: Explore the promoter regions of ~6000 genes and ORFs in yeast genome
  - Provide information on genes with mapped regulatory regions

```
SWI5     TAACCTGTTTAGGAAAAAGGTAAACAATAAC
CLB1     CCGCCCAAAGAGGAAAAACATCAACAATCAA
CLB2-1   CGACCGAATCAGGAAAAGGTCAACAACGAAG
CLB2-2   TTTCCTAAACGGGCTCAATATGTAAACATAT
BUD4     TGACCCGATTTGGAAAAAGGTAAACAACAAT
KIN3     TTTCCTAATTAGGTTAAACGTAAATAAAAGC
CDC20    TTGCCGAAAGAGGCAAAACGTAAATAGGTTG
YOR315W  TCGCCCAATTAGGATAAAAGTAAATAACATA
CHS2     ATACTCAAATAGGAAATATATAAACAAATAG
YRO2     CAACCCGATGAGGAATCATCCCGATCTAACA
WSC4     TCGCCCAAAGTGGAAATACCGTAATAGTACC
APC11    TGACCGGTAGTGGAACTGTTCCAAACCTTTT
FAR1     ATACCTAAAAAGGAAATAGGCAATAAAATGA
CDC5     AAACCCAATAAAGAAAATCCAAAATATAGAA
YML119W  TTTCCAGATTAGGAAAGAACATAAATAAAAA
UBA2     TTTTCTGTATTGGTAAGAACTTAAATGAGAA
DBF2     TTTTCTTTTTGGGTTGGTCCTCCGGAATAGG
CDC16    TTTCTTTATTGGGTCACACAAACCGTATTAA
```

NameGenesGroupMotifsSites(%Genes)Sites(%Genesout of 256 controls)Choet al.Cln258G1MCB:ACGCGTSCB:CRCG



Cell Cycle Regulation (continued)

Computationally predicted E2F target genes confirmed by *in vivo* footprint (ChIP) (Kel *et al*. JMB 2001)

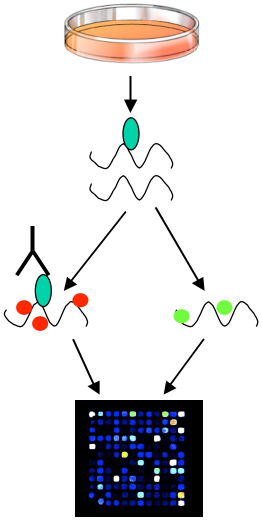
Gene	EMBL	**Sequence of the potential sites**	**Position rel. start transcription**	***Score, q***	***d(X)***
c-fos, *Homo sapiens*	HSFOS	(-) gcCTTGGCGCGTGTcc (-) ggGGTGGCGCGCGGgc (+) ccTCTGGCGCCACCgt (-) acGGTGGCGCCAGAgg	-165 .. -176 -92 .. -103 -90 .. -79 -78 .. -89	0.915 0.836 0.878 0.830	 2.92
JunB, *Homo sapiens*	HS207341	(+) gcTATCGCGCCAGAga (-) tcTCTGGCGCGATAgc (-) ggGCTGGCGCGGGCgg	79 .. 90 91 .. 80 169 .. 158	0.887 0.905 0.820	 3.16
TGF-□1, *Homo sapiens*	HSTGFB1PR	(+) ctGTTTGCGGGGCGga (+) ccCTTCGCGCCCTGgg (+) ctCTTGGCGCGACGct (-) agCGTCGCGCCAAGag (+) ccTTTGCCGCCGGGga	-513 .. -502 -298 .. -287 28 .. 39 40 .. 29 85 .. 96	0.804 0.912 0.928 0.830 0.854	2.03
ARF, *Homo sapiens*	AF082338	(-) acTTTCCCGCCCTGtg	-265 .. -276	0.859	
Mcm4 (Cdc21), *Mus musculus*	AB000629	(+) ggTTTCCCGCCAAAac (-) gtTTTGGCGGGAAAcc (+) gcAGTGGCGCCTTCcg (+) ccTTTGCCGCTGTGat (+) tgGGTGGCGCAGAAct (+) ttTGTCGCGCAGCAac	-443 .. -432 -431 .. -442 -329 .. -318 -297 .. -286 -127 .. -116 -24 .. -13	0.872 0.935 0.810 0.846 0.809 0.858	
MCM5 (P1-CDC46), *Homo sapiens*	HS286B10	(+) agTTTCGCGCCAAAtt (-) aaTTTGGCGCGAAAct (+) ttTTTCCCGCGAAAct (-) agTTTCGCGGGAAAaa	-187 .. -176 -175 .. -186 8 .. 19 20 .. 9	0.988 1.005 0.885 0.932	4.91 3.01 4.21
von Hippel-Lindau (VHL), *Homo sapiens*	AF010238	(+) aaGCTCGCGCCACTgc (-) gcAGTGGCGCGAGCtt (-) gtCTTCGCGCGCGCtc	-270 .. -259 -258 .. -269 -28 .. -39	0.810 0.838 0.921	 2.22
B-myb, *Homo sapiens*	HSBMYBDNA	(-) gtCCTGGCGCGCGGgc	-72 .. –83 -53 .. -42	0.831 0.866	 5.50
Nucleolin, *Homo sapiens*	HSNUCLEO	(-) ttTTTGGCGCCGGCtg (-) ccGTGGGCGCGCGGgt	-297 .. -308 -256 .. -267	0.966 0.814	 2.91
Nucleolin, *Cricetulus griseus*	CSNUCLEO	(-) cgTTTGGCGCGGCTtg	-296 .. -307	0.973	6.67
Nucleolin, *Mus musculus*	MMNUCLEO	(-) agTTTGGCGCGGCTtg	-306 .. -317	0.973	1.76

Does genome location analysis work with human cells?

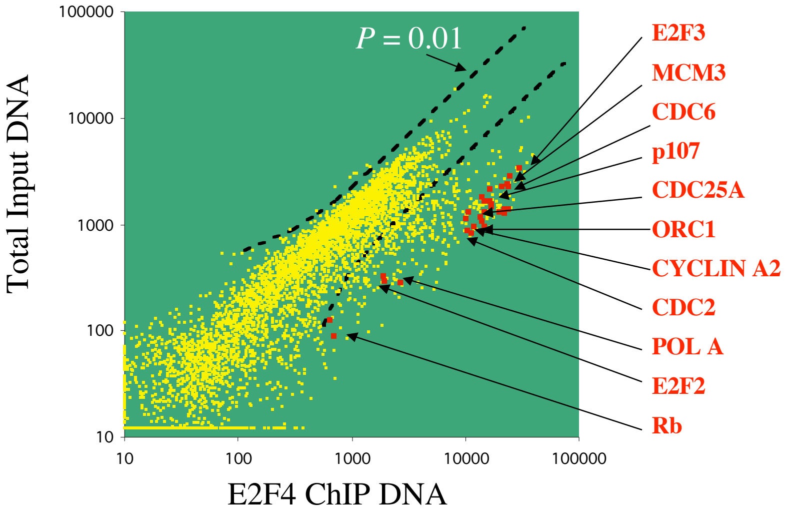
—— Challenges (In collaboration with Ren Lab UCSD)

ChIP-chip:

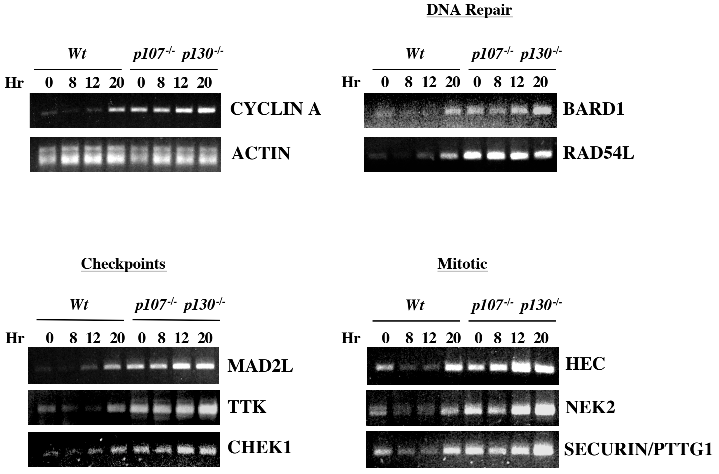
- Genome is 200 times bigger than yeast
- Abundant repetitive sequences
- Annotation of gene structure and function is much less complete
- Many different cell types
- Quality of antibodies



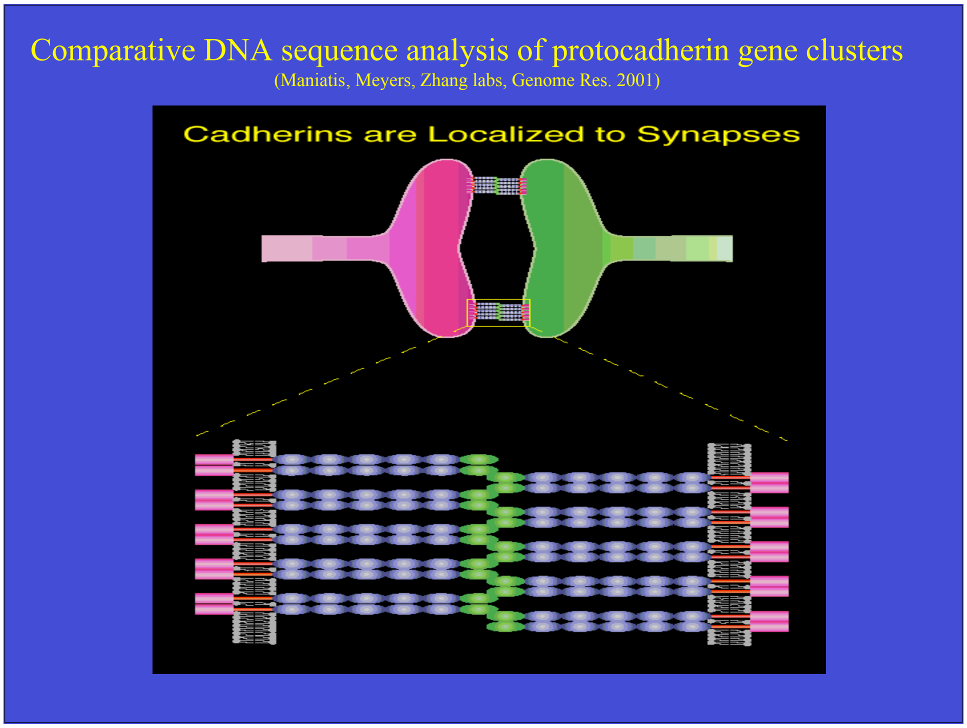
Location Analysis in Human cells: E2F4 Binding to Human Promoters (Ren et al. 2002)



RT-PRC Confirmation (Cont.)

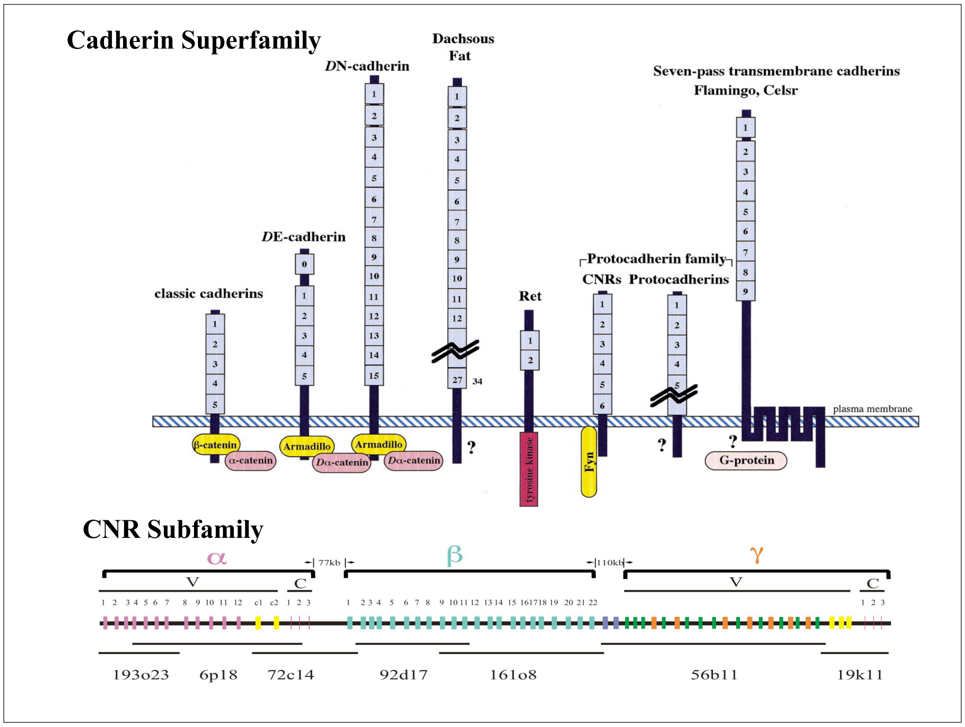


New challenge: Expression+ChIP+Promter, CompositeSites



Comparative DNA sequence analysis of protocadherin gene clusters

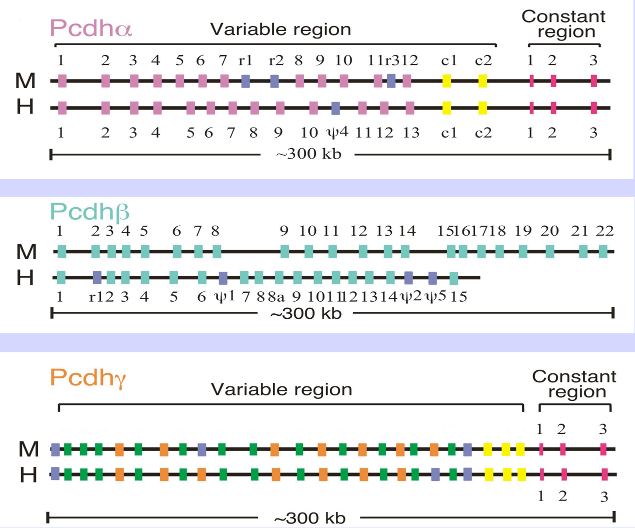
(Maniatis, Meyers, Zhang labs, Genome Res. 2001)



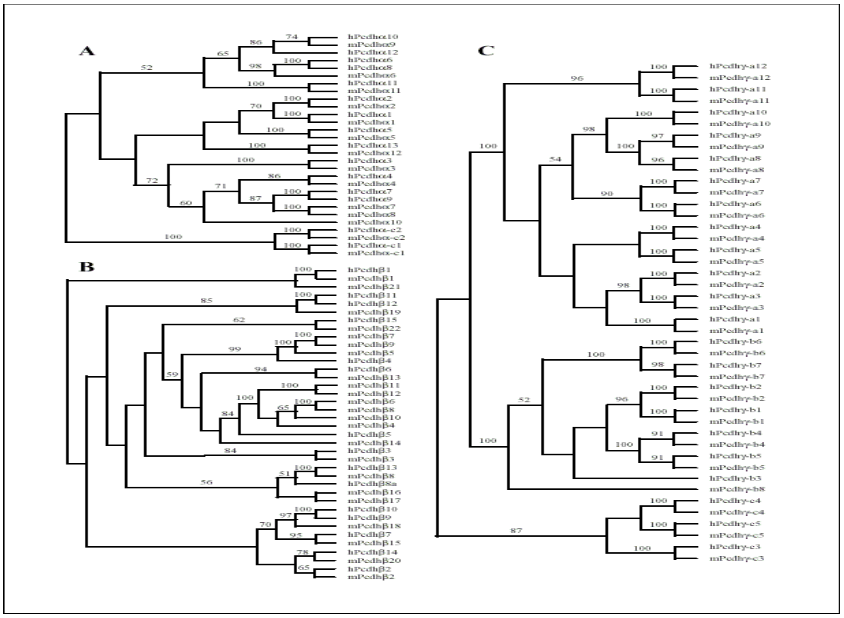
Cadherin Superfamily

CNR Subfamily

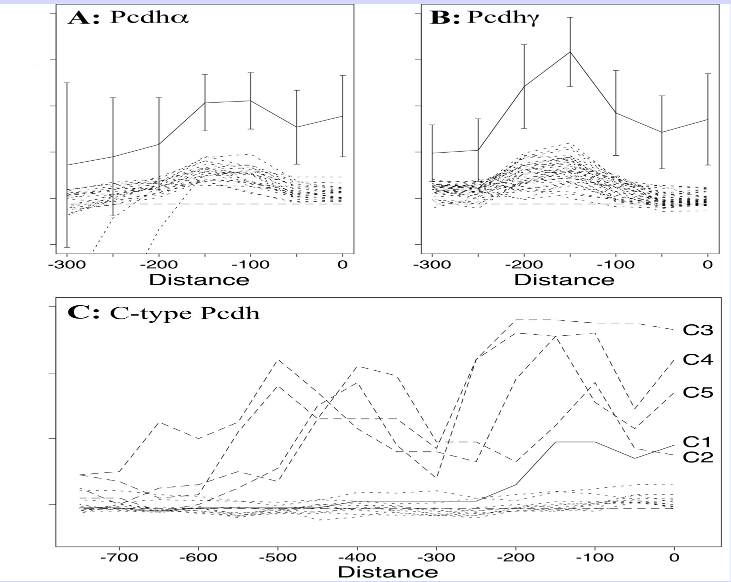
Mouse and Human CNS Clusters



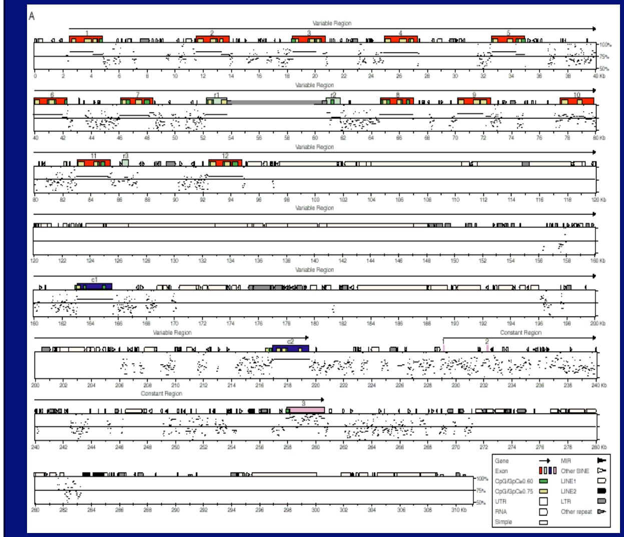
Phylogenetic relationship

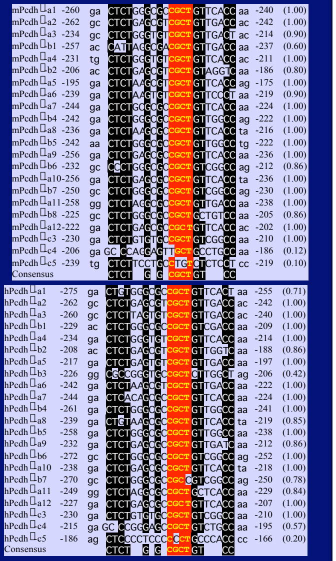


Upstream Sequence Similarity



CpG-islands, conserved promoter elements

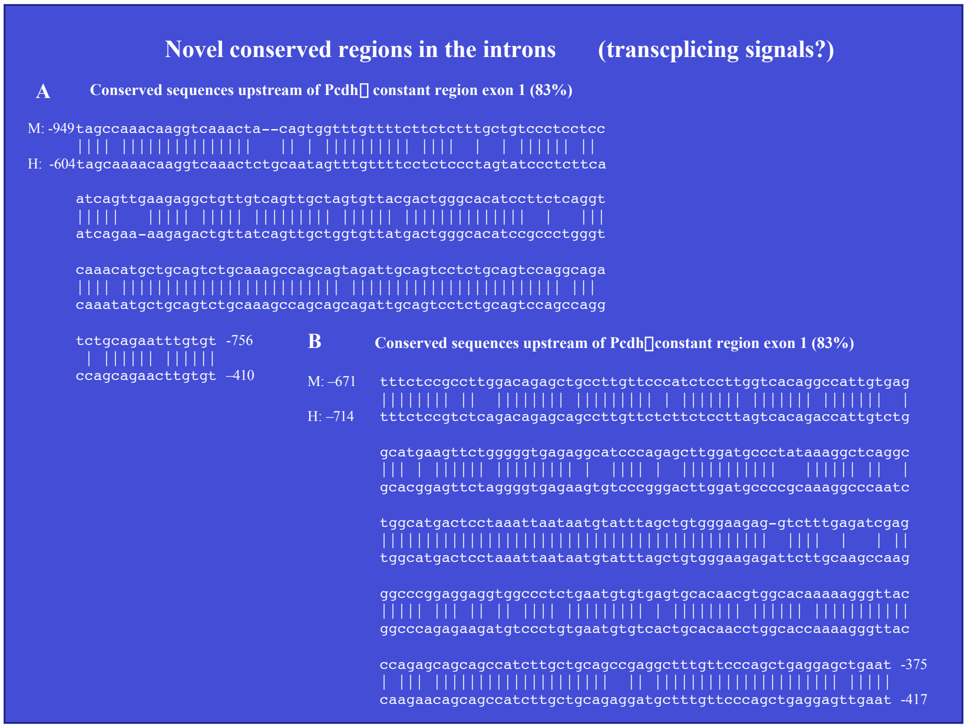




```
mPcdh□a1  -260  ga   CTCTGGGCGCCGCTGTTCACC aa -240 (1.00)
mPcdh□a2  -262  gc   CTCTGAGCGTCGCTGTTGACC ac -242 (1.00)
mPcdh□a3  -234  gc   CTCTGGGTGTCGCTGTTGACT ac -214 (0.90)
mPcdh□b1  -257  ac   CATTAGGCGACGCTGTTGACC aa -237 (0.60)
mPcdh□a4  -231  tg   CTCTGGGTGTCGCTGTTCACC ac -211 (1.00)
mPcdh□b2  -206  ac   CTCTGAGCGTCGCTGTAGGTC aa -186 (0.80)
mPcdh□a5  -195  ga   CTCTAAGCGTCGCTGTTCACC ag -175 (1.00)
mPcdh□a6  -239  ga   CTCTAAGTGTCGCTGTTCCCT aa -219 (0.90)
mPcdh□a7  -244  ga   CTCTGCGCGCCGCTGTTCACC aa -224 (1.00)
mPcdh□b4  -242  ga   CTCTGGGCGCCGCTGTTGGCC ag -222 (1.00)
mPcdh□a8  -236  ga   CTCTAAGCGCCGCTGTTCACC ta -216 (1.00)
mPcdh□b5  -242  aa   CTCTGGGCGCCGCTGTTGGCC tg -222 (1.00)
mPcdh□a9  -256  ga   CTCTGAGCGCCGCTGTTCACC aa -236 (1.00)
mPcdh□b6  -232  gc   CCCTGGGCGCCGCTGTCGGCC ag -212 (0.86)
mPcdh□a10-256   ga   CTCTGAGCGCCGCTGTTCACC ta -236 (1.00)
mPcdh□b7  -250  gc   CTCTGGGCGCCGCTGTCGGCC ag -230 (1.00)
mPcdh□a11-258   gg   CTCTAGGCGCCGCTGTTGACC aa -238 (1.00)
mPcdh□b8  -225  gc   CTCTGGGCGCCGCTGCTGTCC aa -205 (0.86)
mPcdh□a12-222   ga   CTCTGAGCGCCGCTGTTCACC ac -202 (1.00)
mPcdh□c3  -230  ga   CTCTGTGTGCCGCTGTCGGCC aa -210 (1.00)
mPcdh□c4  -206  ga GC CCAGGAGTTGCTGCCTGCC aa -186 (0.12)
mPcdh□c5  -239  tg   CTCTTCCTGCCTGTCTCTCCT cc -219 (0.10)
Consensus            CTCT   G G CGCTGT   CC

hPcdh□a1    -275  ga   CTGTGGGCGCCGCTGTTCACT aa -255 (0.71)
hPcdh□a2    -262  gc   CTCTGAGCGTCGCTGTTGACC ac -242 (1.00)
hPcdh□a3    -260  gc   CTCTTAGTGTCGCTGTTGACC ac -240 (1.00)
hPcdh□b1    -229  ac   CTCTGGGCGCCGCTGTCGACC aa -209 (1.00)
hPcdh□a4    -234  ga   CTCTGGGTGTCGCTGTTCACC aa -214 (1.00)
hPcdh□b2    -208  ac   CTCTGAGCGTCGCTGTTGGTC aa -188 (0.86)
hPcdh□a5    -217  ga   CTCTGAGTGTCGCTGTTGACC aa -197 (1.00)
hPcdh□b3    -226  gg   CGCCGGGTGCCGCTCTTGGCT ag -206 (0.42)
hPcdh□a6    -242  ga   CTCTAAGCGTCGCTGTTGACC aa -222 (1.00)
hPcdh□a7    -244  ga   CTCACAGCGCCGCTGTTCACC aa -224 (1.00)
hPcdh□b4    -261  ga   CTCTGGGCGCCGCTGTTGGCC aa -241 (1.00)
hPcdh□a8    -239  ga   CTGTAAGCGCCGCTGTTCACC ta -219 (0.85)
hPcdh□b5    -258  ga   CTCTGGGCGCCGCTGTTGGCC aa -238 (1.00)
hPcdh□a9    -232  ga   CTCTGAGCGCCGCTGTTGATC aa -212 (0.86)
hPcdh□b6    -272  gc   CTCTGGGCGCCGCTGTCGGCC ag -252 (1.00)
hPcdh□a10   -238  ga   CTCTGAGCGCCGCTGTTCACC ta -218 (1.00)
hPcdh□b7    -270  gc   CTCTGGGCGCCGCCGTCGGCC ag -250 (0.78)
hPcdh□a11   -249  gg   CTCTAGGCGCCGCTGCTCACC aa -229 (0.84)
hPcdh□a12   -227  ga   CTCTGAGCGCCGCTGTTCACC aa -207 (1.00)
hPcdh□c3    -230  ga   CTCTGTGTGCCGCTGTCGGCC aa -210 (1.00)
hPcdh□c4    -215  ga GC CCGGGAGCCGCTGTCTGCC aa -195 (0.57)
hPcdh□c5    -186  ag   CTCCCCTCCCCCCTGCCCACC cc -166 (0.20)
Consensus              CTCT   G G CGCTGT   CC
```

Pipmaker

Gibbs sampler



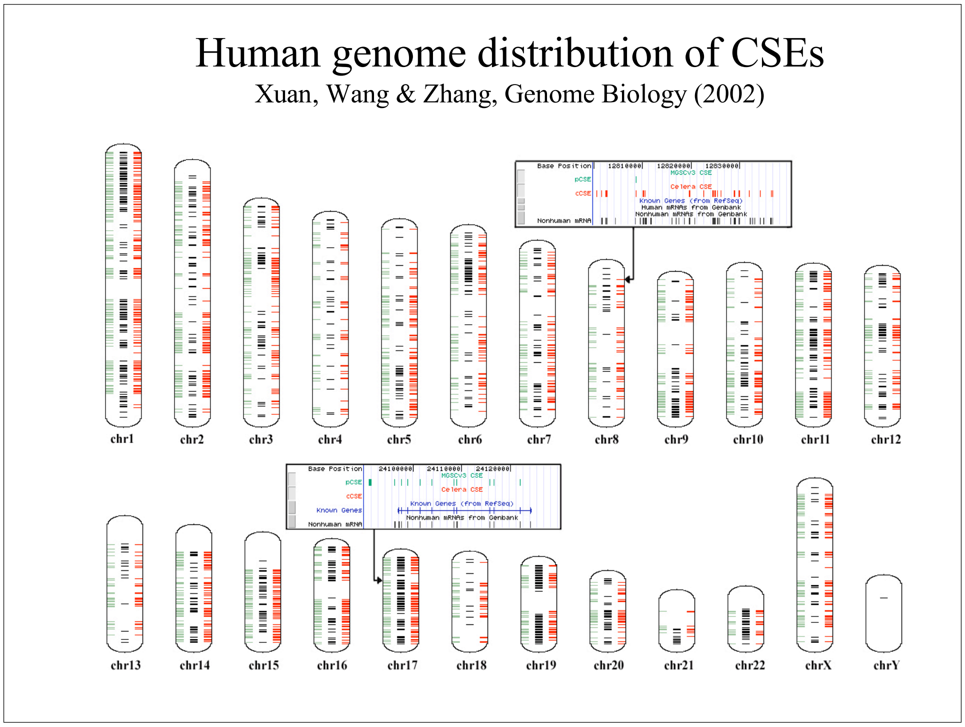
Novel conserved regions in the introns (transcplicing signals?)

A **Conserved sequences upstream of Pcdh□ constant region exon 1 (83%)**

```
M: -949 tagccaaacaaggtcaaacta--cagtggtttgttttcttctctttgctgtccctcctcc
        |||| |||||||||||||||   || | |||||||||| ||||  |  | |||||| || 
H: -604 tagcaaaacaaggtcaaactctgcaatagtttgttttcctctccctagtatccctcttca

        atcagttgaagaggctgttgtcagttgctagtgttacgactgggcacatccttctcaggt
        |||||   ||||| ||||| ||||||||| |||||| ||||||||||||||  |   |||
        atcagaa-aagagactgttatcagttgctggtgttatgactgggcacatccgccctgggt

        caaacatgctgcagtctgcaaagccagcagtagattgcagtcctctgcagtccaggcaga
        |||| ||||||||||||||||||||||||| |||||||||||||||||||||||| ||| 
        caaatatgctgcagtctgcaaagccagcagcagattgcagtcctctgcagtccagccagg

        tctgcagaatttgtgt  -756
         | |||||| ||||||
        ccagcagaacttgtgt  –410
```

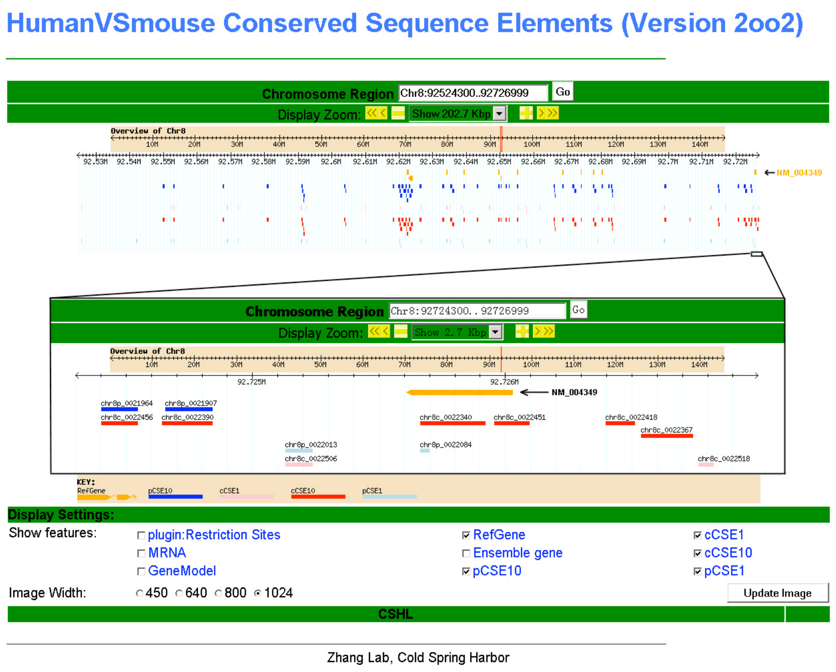
B **Conserved sequences upstream of Pcdh□constant region exon 1 (83%)**

```
M: –671   tttctccgccttggacagagctgccttgttcccatctccttggtcacaggccattgtgag
          |||||||| ||  |||||||| ||||||||| | ||||||| ||||||| |||||||  |
H: –714   tttctccgtctcagacagagcagccttgttctcttctccttagtcacagaccattgtctg

          gcatgaagttctgggggtgagaggcatcccagagcttggatgccctataaaggctcaggc
          ||| | |||||| ||||||||| |  |||| |  |||||||||||   |||||| ||  |
          gcacggagttctaggggtgagaagtgtcccgggacttggatgccccgcaaaggcccaatc

          tggcatgactcctaaattaataatgtatttagctgtgggaagag-gtctttgagatcgag
          ||||||||||||||||||||||||||||||||||||||||||||  ||||  |   | ||
          tggcatgactcctaaattaataatgtatttagctgtgggaagagattcttgcaagccaag

          ggcccggaggaggtggccctctgaatgtgtgagtgcacaacgtggcacaaaaagggttac
          ||||| ||| || || |||| ||||||||| | |||||||| |||||| |||||||||||
          ggcccagagaagatgtccctgtgaatgtgtcactgcacaacctggcaccaaaagggttac

          ccagagcagcagccatcttgctgcagccgaggctttgttcccagctgaggagctgaat  -375
          | ||| ||||||||||||||||||||  || ||||||||||||||||||||| |||||
          caagaacagcagccatcttgctgcagaggatgctttgttcccagctgaggagttgaat  -417
```



Human genome distribution of CSEs

Xuan, Wang & Zhang, Genome Biology (2002)



HumanVSmouse Conserved Sequence Elements (Version 2oo2)

Chromosome Region Chr8:92524300..92726999 Go

Display Zoom: << < − Show 202.7 Kbp ▾ + > >>

Chromosome Region Chr8:92724300..92726999 Go

Display Zoom: << < − Show 2.7 Kbp ▾ + > >>

Display Settings:

Show features:

□ plugin:Restriction Sites	☑ RefGene	☑ cCSE1
□ MRNA	□ Ensemble gene	☑ cCSE10
□ GeneModel	☑ pCSE10	☑ pCSE1

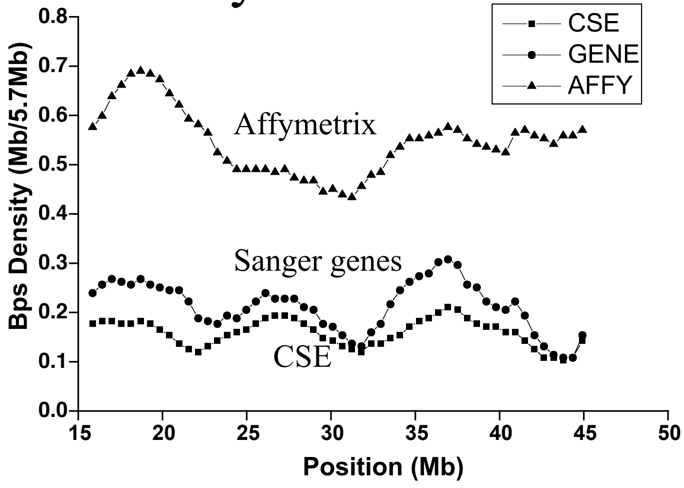
Image Width: ○ 450 ○ 640 ○ 800 ● 1024

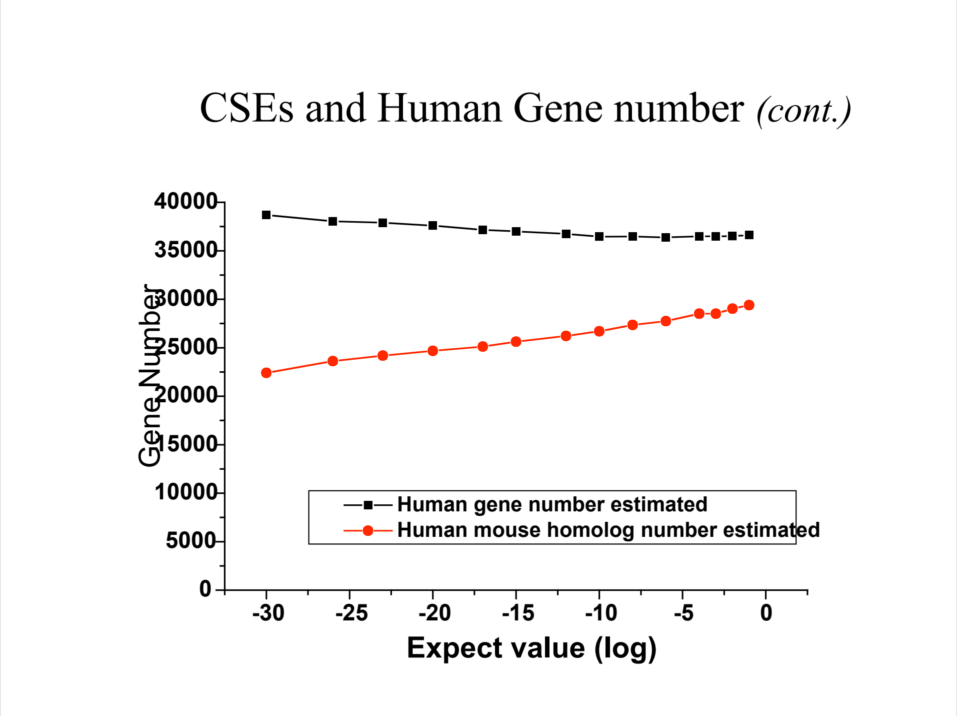
Update Image

CSHL

Zhang Lab, Cold Spring Harbor

CSE, genes and transcriptional activity in human chr22





CSEs and Human Gene number *(cont.)*

Zhang Lab Members