7.91 / 7.36 / 20.490 / 20.390 /
6.874 / 6.801 / HST.506

C. Burge Lecture #6

Feb 13, 2014

Comparative Genomics

Global Alignment of Protein Sequences (NW, SW, PAM, BLOSUM)

- Global sequence alignment (Needleman-Wunch-Sellers)
- Gapped local sequence alignment (Smith-Waterman)
- Substitution matrices for protein comparison

Background: Z&B Chapters 4,5 (esp. pp. 119-125)

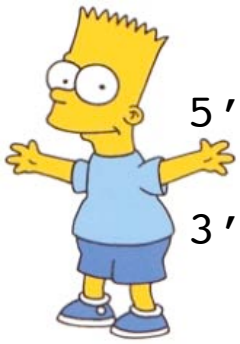
DNA Sequence Evolution

Generation ***n-1*** (grandparent)

```
5′ TGGCATGCACCCTGTAAGTCAATATAAATGGCTACGCCTAGCCCATGCGA 3′
   ||||||||||||||||||||||||||||||||||||||||||||||||||
3′ ACCGTACGTGGGACATTCAGTTATATTTACCGATGCGGATCGGGTACGCT 5′
```

Generation ***n*** (parent)

```
5′ TGGCATGCACCCTGTAAGTCAATATAAATGGCTATGCCTAGCCCATGCGA 3′
   ||||||||||||||||||||||||||||||||||||||||||||||||||
3′ ACCGTACGTGGGACATTCAGTTATATTTACCGATACGGATCGGGTACGCT 5′
```



Generation ***n+1*** (child)

```
5′ TGGCATGCACCCTGTAAGTCAATATAAATGGCTATGCCTAGCCCGTGCGA 3′
   ||||||||||||||||||||||||||||||||||||||||||||||||||
3′ ACCGTACGTGGGACATTCAGTTATATTTACCGATACGGATCGGGCACGCT 5′
```

Markov Model (aka Markov Chain)

Stochastic Process:
- a random process or
- a sequence of Random Variables

Classical Definition

A discrete stochastic process $X_1, X_2, X_3, \ldots$ which has the Markov property:

$$P(X_{n+1} = j \mid X_1=x_1, X_2=x_2, \ldots X_n=x_n) = P(X_{n+1} = j \mid X_n=x_n)$$

(for all x_i, all j, all n)

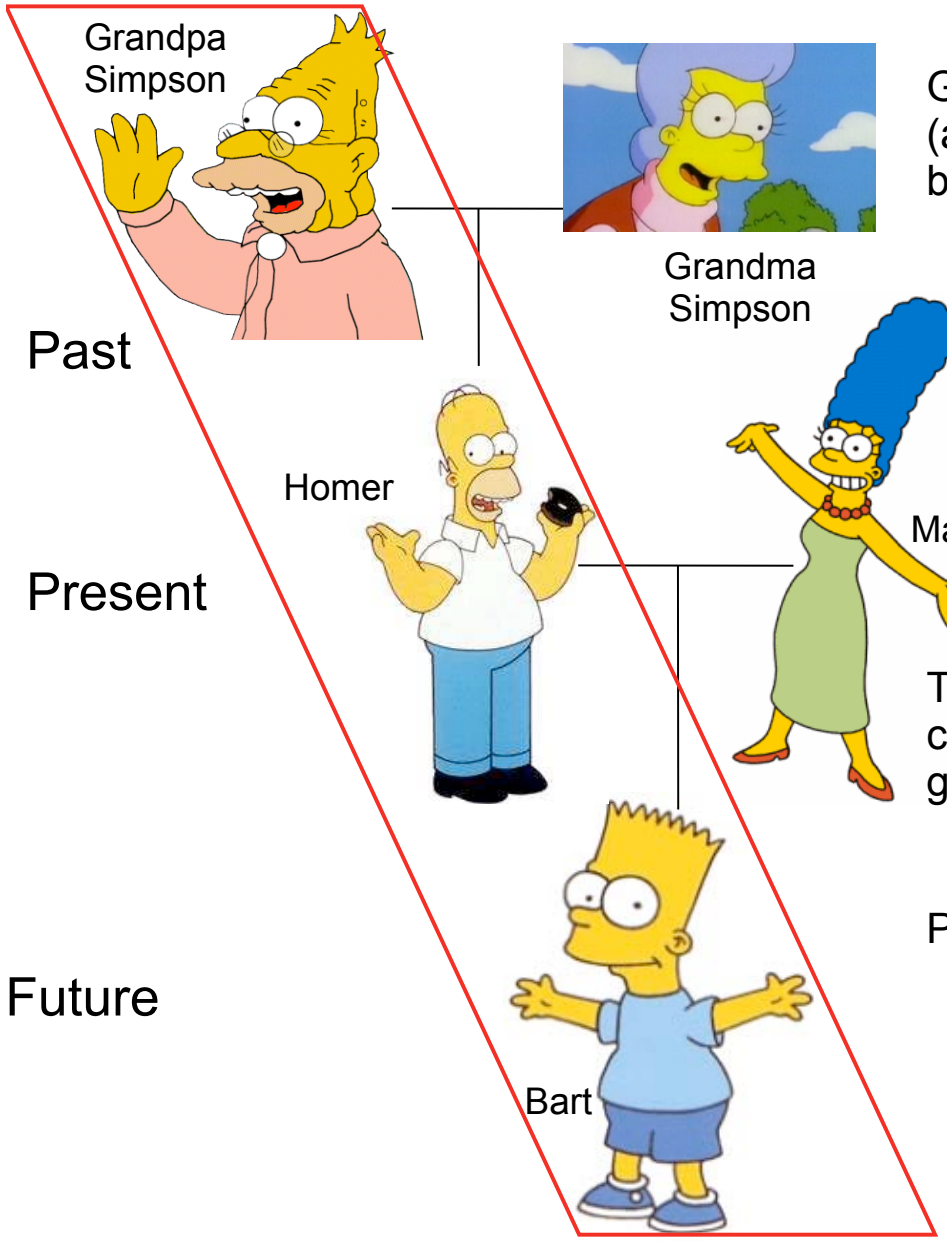
Image is in the public domain.

In words:

A random process which has the property that the future (next state) is conditionally independent of the past given the present (current state)

Andrey Markov, a Russian mathematician (1856 - 1922)

Markov Model Example



Genotype at the Apolipoprotein locus (alleles A and a) in successive generations of boxed Simpson lineage forms a Markov model

This is because, e.g., Bart's genotype is conditionally independent of Grandpa Simpson's genotype given his father Homer's genotype:

$$P(\text{Bart} = a/a \mid \text{Grandpa} = A/a \;\&\; \text{Homer} = a/a) = P(\text{Bart} = a/a \mid \text{Homer} = a/a)$$

Review:
Vector/Matrix Notation for Markov Chains

Assuming no selection

S_n = base at generation n

$$P_{ij} = P(S_{n+1} = j \mid S_n = i)$$

$$P = \begin{array}{c} \\ \text{from: A} \\ \text{C} \\ \text{G} \\ \text{T} \end{array} \begin{array}{c} \begin{array}{cccc} \text{to:} & & & \\ \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} P_{AA} & P_{AC} & P_{AG} & P_{AT} \\ P_{CA} & P_{CC} & P_{CG} & P_{CT} \\ P_{GA} & P_{GC} & P_{GG} & P_{GT} \\ P_{TA} & P_{TC} & P_{TG} & P_{TT} \end{pmatrix} \end{array}$$

$\vec{q}^{\,n} = (q_A, q_C, q_G, q_T)$ = vector of prob's of bases at gen. n

Handy relations: $\vec{q}^{\,n+1} = \vec{q}^{\,n} P$ $\qquad$ $\vec{q}^{\,n+k} = \vec{q}^{\,n} P^k$

What happens after a long time? i.e. what is $\lim_{n \to \infty} \vec{q} P^n$?

PAM matrix derivation

$$M_{a,b} = \Lambda\, m_b \frac{A_{a,b}}{\sum_i A_{i,b}}$$

Set scale factor Λ so that

$$\sum_b f_b M_{b,b} = 0.99$$

$M_{a,b}$ = mutation prob. matrix

$A_{a,b}$ = observed subs of a,b

m_b = mutability of b

f_b = frequency of b

Λ = a scaling constant

i.e. chance of mutating is ~ 1%

This gives a probability matrix for an evolutionary distance of 1 PAM. Use matrix multiplication to calculate prob. matrices for other PAM distances, e.g., 20, 40, 60, 120, 250.

substitution scores for evolutionary distance d:

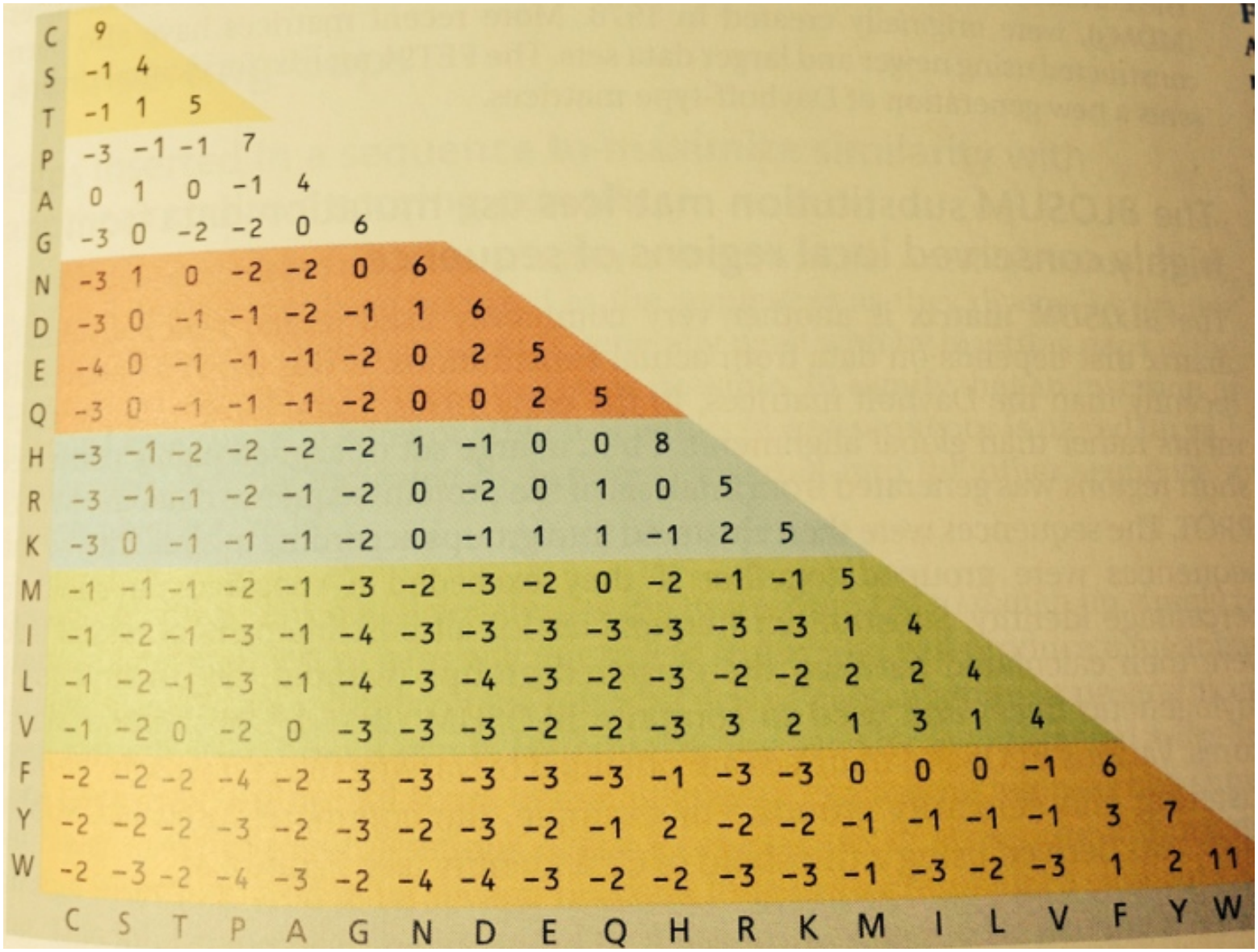
$$s_{a,b} = 2 \log_2 (M^d_{a,b} / f_b)$$

Recall: matrix multiplication

Issues with PAM Series?

- Read text about BLOSUM series.
- BLOSUM62 is the most commonly used matrix in practice.

BLOSUM 62



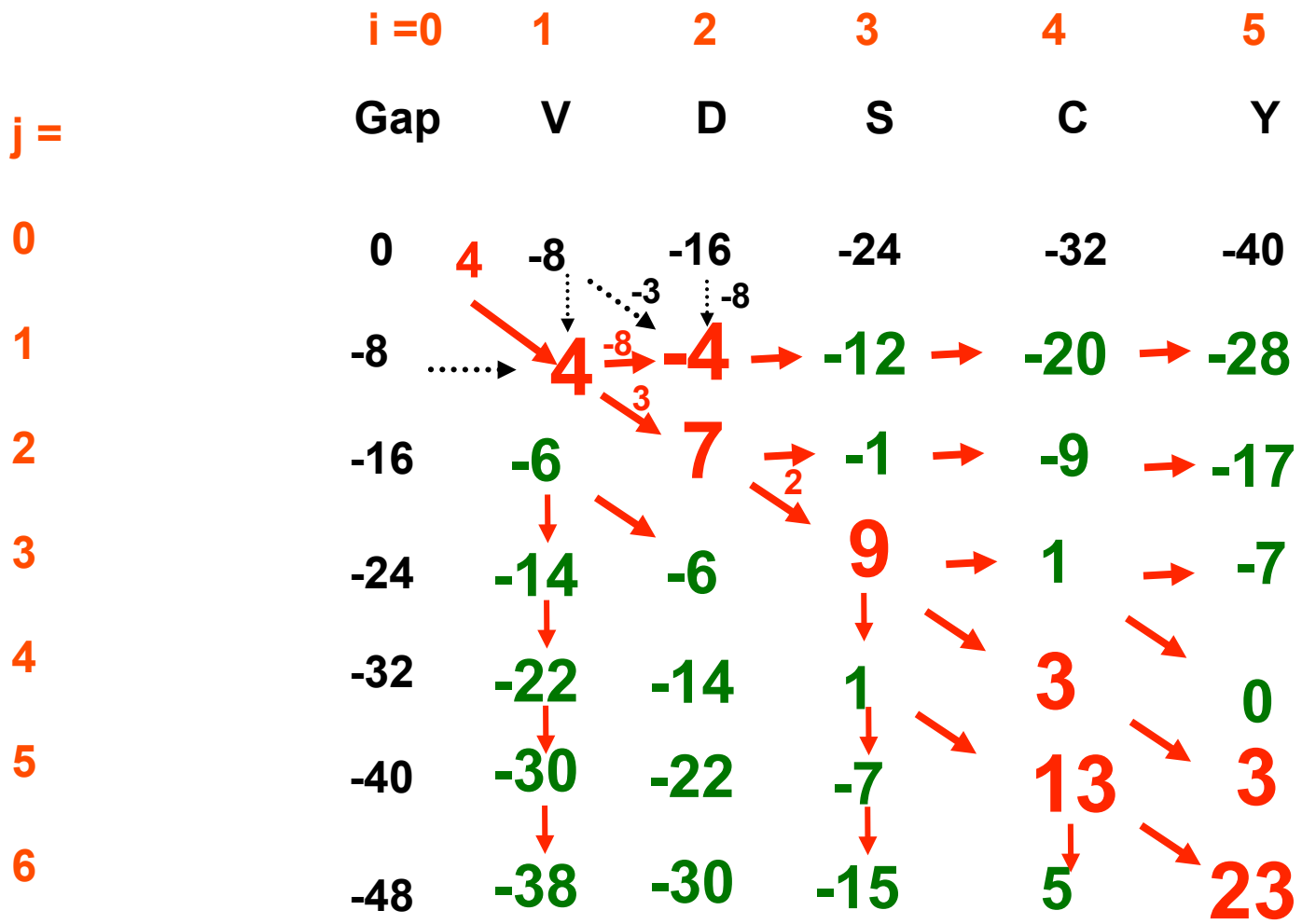
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	3	2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

Multiple Sequence Alignments

- Sequences are aligned so as to bring the greatest number of single characters into register, and maximize a score that rewards matches and penalizes mismatches, gaps

2 sequence alignment

Comp. complexity? O(mn) or O(n^2) if both have length n



For 3 sequences….

Length

```
ARDFSHGLLENKLLGCDSMRWE  m
GRDYKMALLEQWILGCD-MRWD  n
SRDW--ALIEDCMV-CNFFRWD  p
```

An O(mnp) problem

Consider sequences each 300 amino acids

2 sequences – $(300)^2$

3 sequences – $(300)^3$

but for *k* sequences – $(300)^k$

=> Need a more efficient algorithm (e.g., CLUSTALW - see Z&B Ch.6)

Comparative Genomics

- Markov models
- Jukes-Cantor, Kimura models
- Types of Selection: neutral, negative, positive

- Comparative genomics to understand gene regulation
  - a dozen examples

Readings:

12 papers posted under Comparative Genomics (optional)

Sabeti review (first 3 pages recommended)

Limit Theorem for Markov Chains

S_n = base at generation *n* $P_{ij} = P(S_{n+1} = j \mid S_n = i)$

What happens after a long time? i.e. what is $\lim_{n\to\infty} \vec{q} P^n$

If $P_{ij} > 0$ for all *i,j* (and $\sum_j P_{ij} = 1$ for all *i*)

then there is a unique vector $\vec{r}$ such that

$\vec{r} = \vec{r}P$ and $\lim_{n\to\infty} \vec{q} P^n = \vec{r}$ (for any probability vector $\vec{q}$)

$\vec{r}$ is called the "stationary" or "limiting" distribution of P

See Ch. 4, Taylor & Karlin, An Introduction to Stochastic Modeling, 1984 for details

Stationary Distribution Examples

2-letter alphabet: R = purine, Y = pyrimidine

Stationary distributions for:

$$P = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$$

(1/2, 1/2)

$0 < p < 1$

$$P' = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$

(q/(p+q), p/(p+q))

$0 < p < 1,\ 0 < q < 1$

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

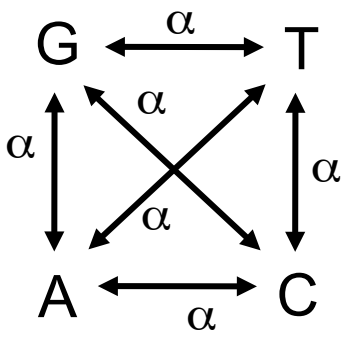
any vector

$$Q = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

(1/2, 1/2)*

* Stationary but not unique limiting distribution

Jukes-Cantor Model



Assume each nucleotide equally likely
to change into any other nt,
with rate of change=α.
Overall rate of substitution = 3α
…so if G at t=0, at t=1, $P_{G(1)}=1-3\alpha$

and $P_{G(2)}=(1-3\alpha)P_{G(1)}+\alpha\,[1-P_{G(1)}]$

Solving recursion gives $P_{G(t)}=1/4+(3/4)e^{-4\alpha t}$

Can show that this gives $K=-3/4\ \ln[1-(4/3)d]$

K = true number of substitutions that have occurred,
d = fraction of nt that differ by a simple count ($d \le 3/4$)

Captures general behavior…

More realistic models of DNA evolution

4-letter alphabet: A, C, G, T

Kimura model(1)

q = transition rate

p = transversion rate

(q = ~2p)

$$P = \begin{pmatrix} 1-2p-q & p & q & p \\ p & 1-2p-q & p & q \\ q & p & 1-2p-q & p \\ p & q & p & 1-2p-q \end{pmatrix}$$

$0 < p < q < 1$

Dinucleotide => Dinucleotide models(2)

AA => AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, …
AC => AA, AC, AG, AT, CA, …

Strand-specific models(3)

(1) Kimura J Mol Evol 1980

(2) Zhang & Gerstein Nucl Acids Res 2003

(3) Green et al. Nature Genet 2003

Detecting Positive/Negative Selection:

Calculation of Ka/Ks ratio (aka dN/dS ratio)

```
          M   Q   R   P   F   G   K   A   R   G   V   S
human   ATG CAA CGG CCT TTG GGA AAG GCC AGA GGA GTC TCC

          M   Q   R   P   V   G   K   A   R   A   L   S
baboon  ATG CAG CGG CCT GTG GGG AAG GCC AGA GCA CTC TCC
```

```
          P   T   A   P   G   V   T   G   V   T
human   CCT ACA GCC CCA GGG GTA ACC GGC GTT ACA

          P   A   A   A   G   V   P   G   V   P
baboon  CCT GCA GCT GCA GGG GTA CCC GGC GTT CCA
```

dN = **Ka** = nonsynonymous substitutions / nonsynonymous sites
dS = **Ks** = synonymous substitutions / synonymous sites

Detecting Negative and Positive Selection in Coding Regions

K_a/K_s or d_N/d_S ratio:

K_a or d_N = normalized rate of **nonsynonymous** changes per position

K_s or d_S = normalized rate of **synonymous** changes per position

Corrected version: $d_S = 3/4 \ln(1 - 4/3\ p_S)$, etc. (see Z&B pp. 240-241)

Common applications:

Identify genes or regions with Ka/Ks significantly less than one
- These regions are likely to be under selection to conserve amino acid sequence

What kinds of genes or regions would you expect to have Ka/Ks ~ 1?

Identify genes or regions with Ka/Ks significantly greater than one
- These regions are likely to be under selection to change amino acid sequence

More sophisticated tests for positive selection: McDonald-Kreitman, etc.

A dozen comparative genomics papers

To illustrate some of the types of things we can learn about gene regulation by comparing genomes, often using fairly simple methods

To provide examples of successful computational biology research projects

To gain experience in reading the literature in regulatory genomics

Types of comparative genomic analyses

Identification of regulatory elements of unknown function
Bejerano et al. 2002
...characterization of their functions
Pennacchio et al 2006, Visel et al 2008, Lareau et al 2007
...exploration of their origins
Bejerano et al 2006

Inference of the targeting rules for a class of trans-acting factors
Lewis et al 2003, 2005

Identification of regulatory targets of a class of trans-acting factors
Kheradpour et al 2008, Friedman et al 2009

Identification of new intra-genic interacting regulatory elements
Graveley 2005

Identification of a new class of trans-acting factors
Jansen et al 2002

Identification of trans-genomic interacting regulatory elements
Bolotin et al 2005

Bejerano et al. 2004 “Ultraconserved elements”

Defined “ultraconserved elements” (UCEs) as unusually long segments that are 100% identical between human, mouse and rat using whole-genome alignments of the 3 species and studied their properties

From the SOM:

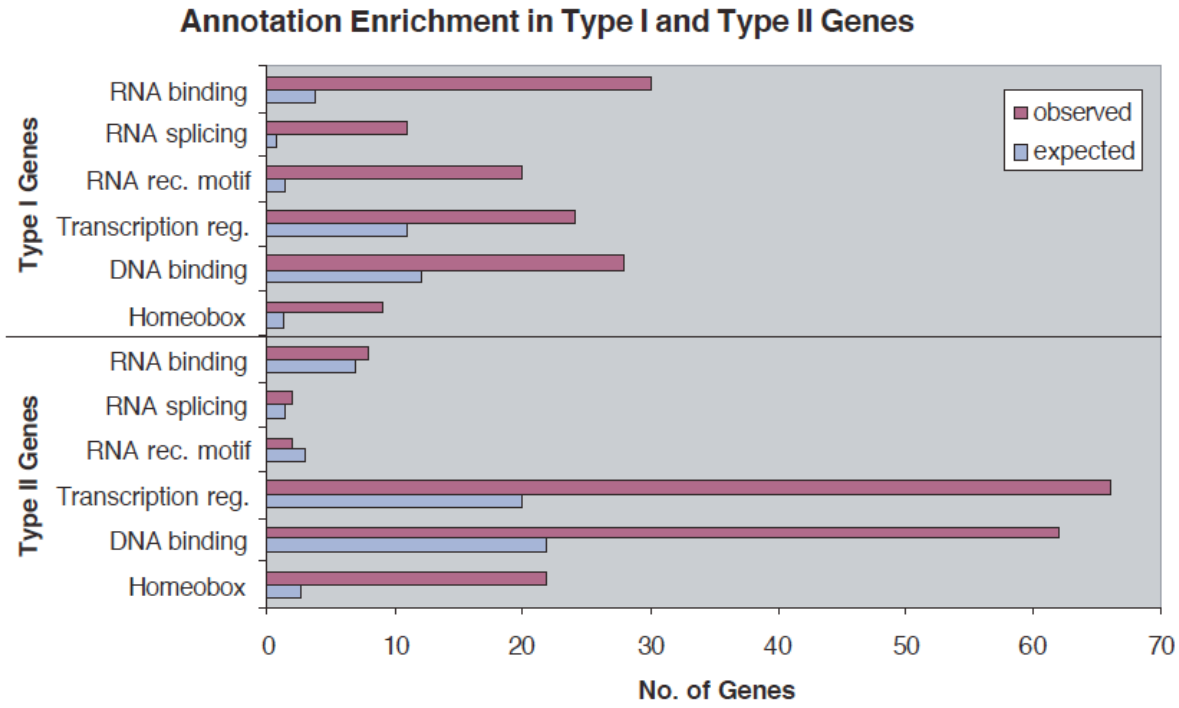
Each column in the orthologous multiple alignment is considered to be an independent observation of a Bernoulli random variable that is 1 (“heads”) if the bases are completely conserved between the three species (a “3-way identity”) and 0 (“tails”) otherwise. ...The largest percent identity among ancestral repeat sites we obtained for any 1 Mb window with enough ancestral repeat sites to get a good estimate, i.e. at least 1000 sites, was actually 0.68. The distribution of the number of runs of at least 200 heads in a series of 2.9 billion tosses of a biased coin with probability p = 0.7 of heads can be approximated quite well using a Poisson distribution with mean $(1-p)\cdot p^{200}$, and the probability of one or more such runs is very close to the mean of the Poisson distribution in this case, which is at most 10^{-22}

Bejerano et al. Science 2004

Features of UCEs

481 UCEs (≥ 200 bp):
~100 overlap exons of known protein-coding genes
~100 located in introns of known genes
~300 intergenic

93 **type I genes** overlap with exonic ultraconserved elements
225 genes that are near the non-exonic elements are called **type II genes**



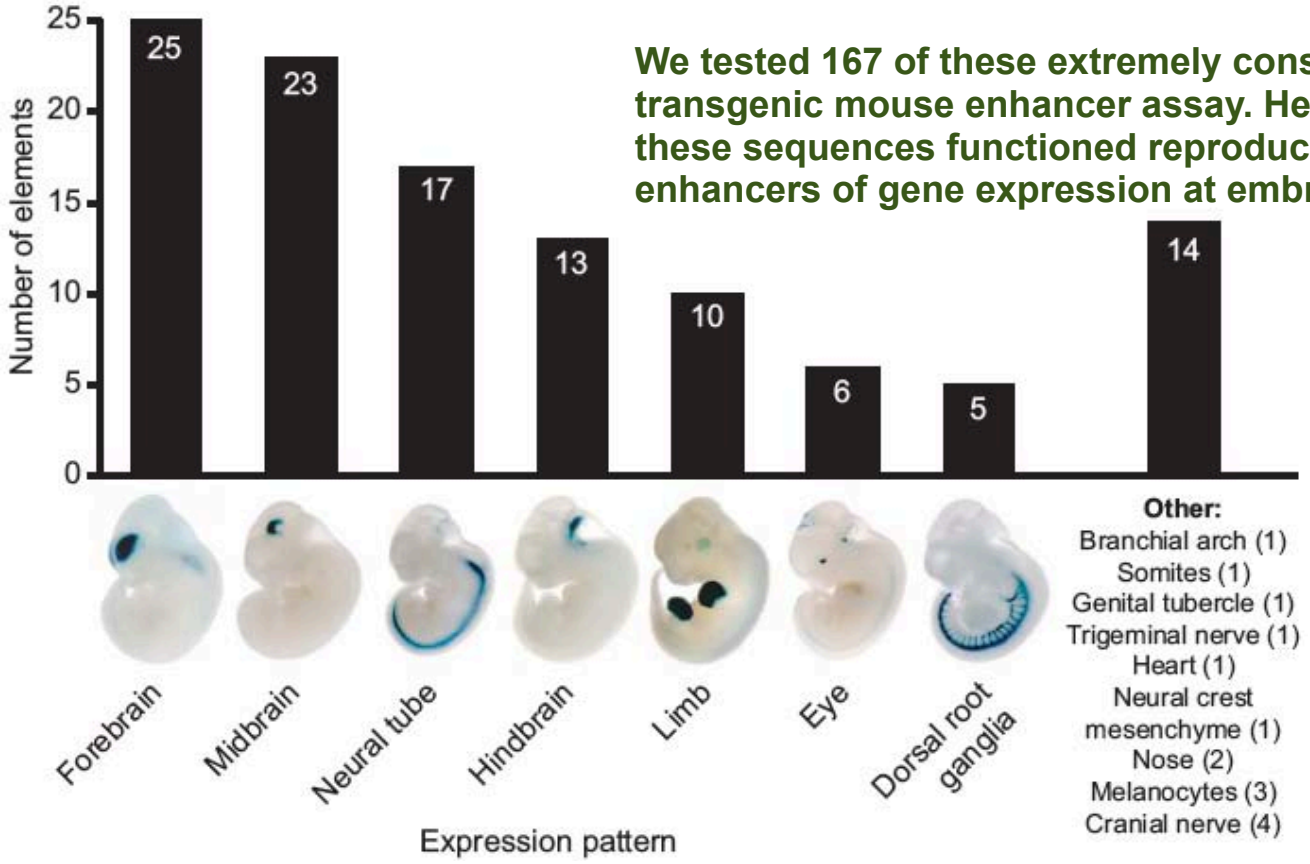
Source: Bejerano, Gill, Michael Pheasant, et al. "Ultraconserved Elements in the Human Genome." *Science* 304, no. 5675 (2004): 1321-5.

Bejerano et al. Science 2004

What do intergenic (ultra)conserved elements do?

well established transgenic mouse enhancer assay that links the human conserved fragment to a minimal mouse heat shock promoter fused to a lacZ reporter gene... determined tissue-specific reporter gene expression at embryonic day 11.5 (e11.5), as this developmental stage allows for whole-mount staining and whole-embryo visualization. Moreover, at this time-point many of the major tissues and organs have been specified. We also expected this stage to be particularly informative because 'extreme' conserved non-coding elements tend to be enriched and clustered near genes expressed during embryonic development.

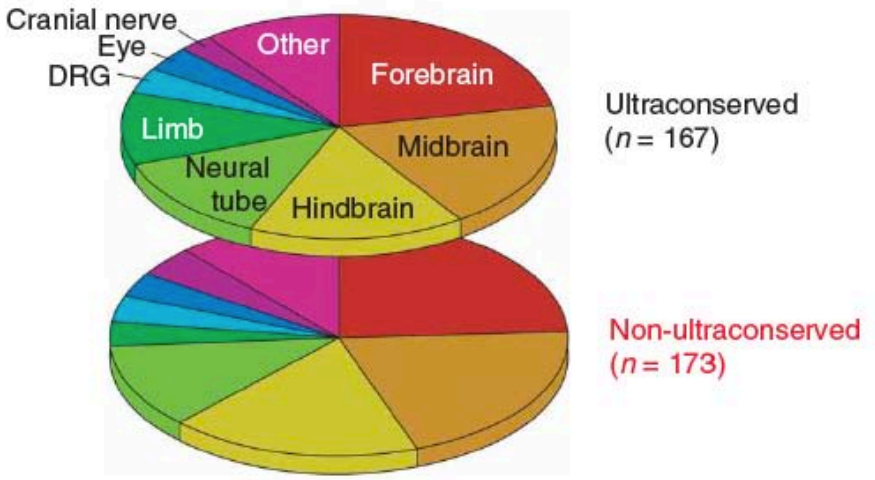
We tested 167 of these extremely conserved sequences in a transgenic mouse enhancer assay. Here we report that 45% of these sequences functioned reproducibly as tissue-specific enhancers of gene expression at embryonic day 11.5.

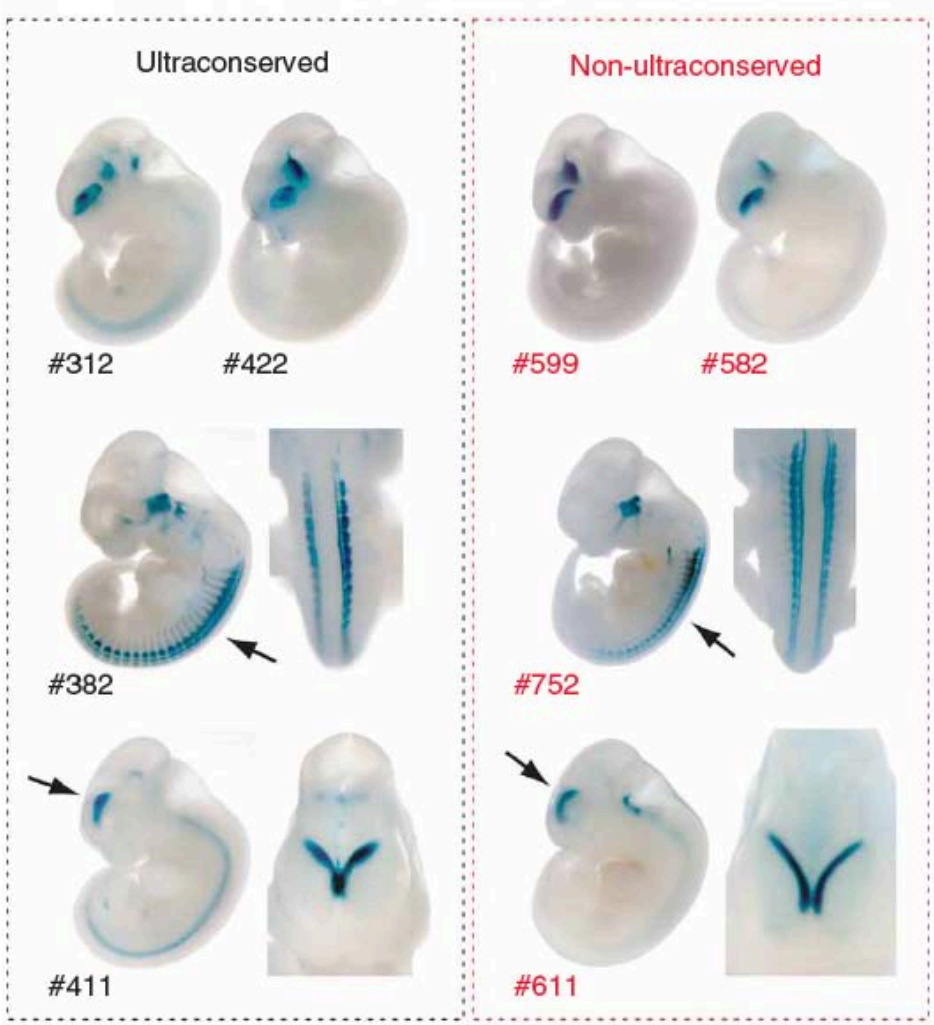


Source: Pennacchio, Len A., Nadav Ahituv, et al. "In Vivo Enhancer Analysis of Human Conserved Non-coding Sequences." *Nature* 444, no. 7118 (2006): 499-502.

Pennacchio et al. Nature 2006

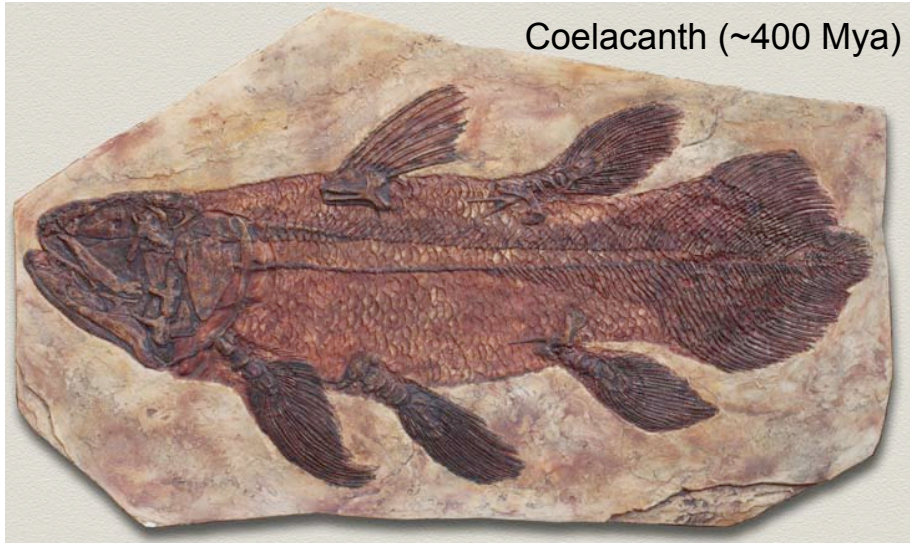
Do **ultraconserved** differ from **highly conserved** enhancers?



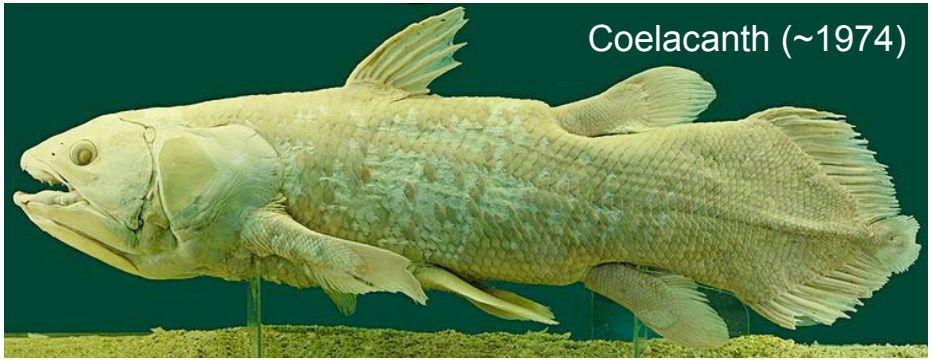


Visel et al. Nature Genet 2008

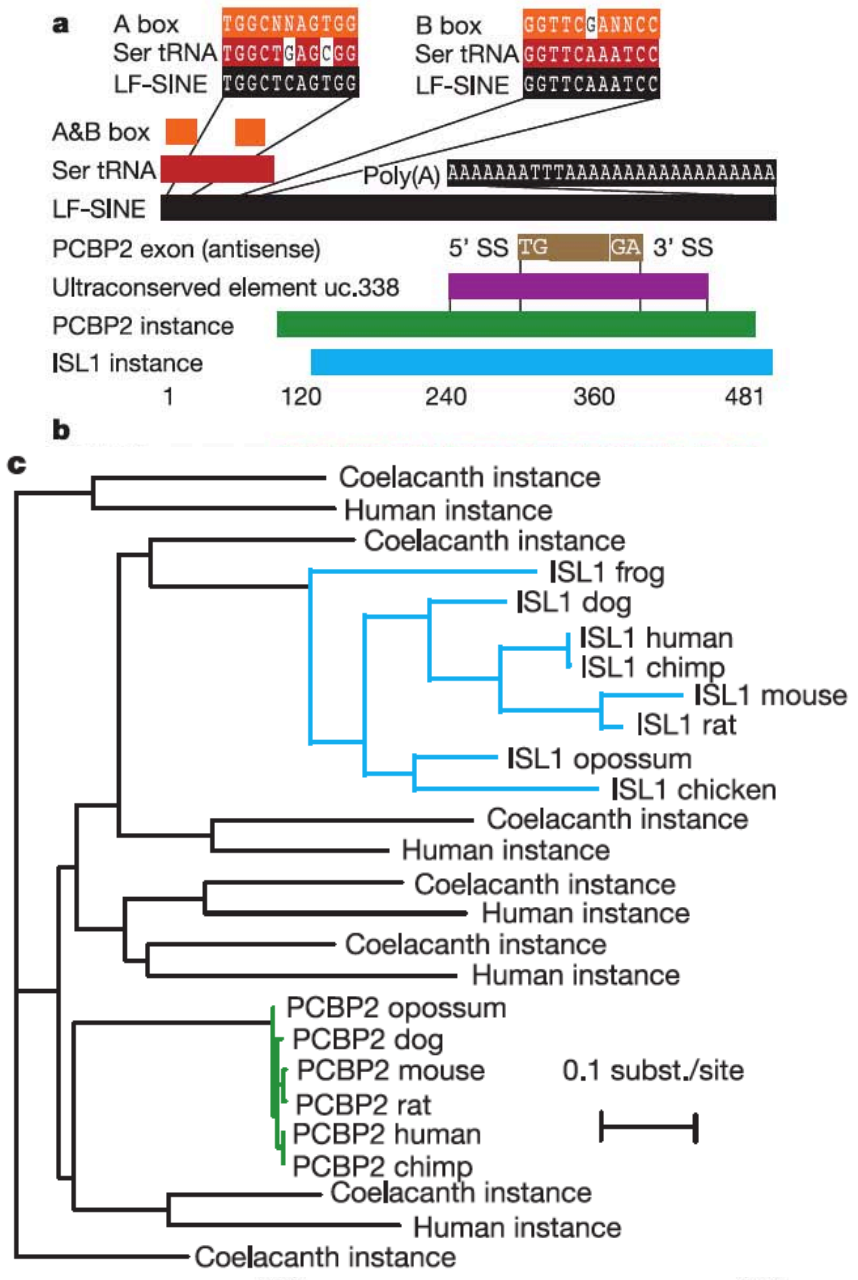
Where do UCEs come from?



Courtesy of Stuart Gold. Used with permission.



Bejerano et al. Nature 2006



Courtesy of Macmillan Publishers Limited. Used with permission. Source: Bejerano, Gill, Craig B. Lowe, et al. "A Distal Enhancer and An Ultraconserved Exon are Derived from a Novel Retroposon." *Nature* 441, no. 7089 (2006): 87-90.

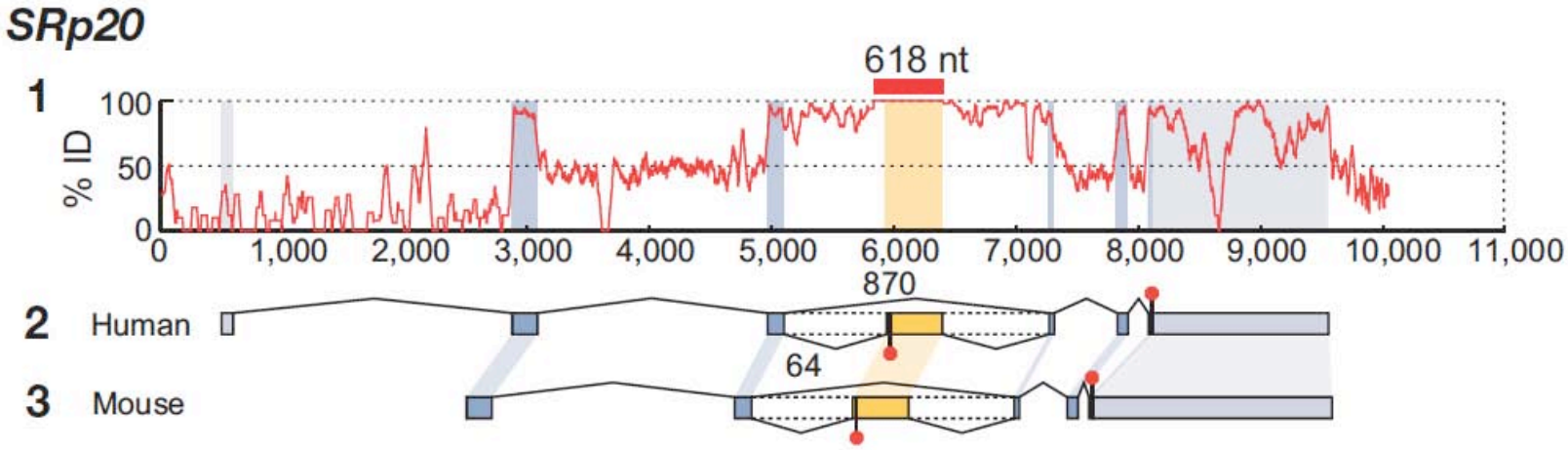
Interpretation

After discovering mobile DNA elements, Barbara McClintock suggested that they were fundamentally involved in gene regulation, an idea further developed by Britten and Davidson, who speculated on the benefit of obtaining similar control regions for a ‘battery’ of co-regulated genes through **exaptation**.

At least 50% of our genome originates from characterized transposon-derived DNA ... it seems possible that, because these elements optimize their interaction with the host machinery under strong, virus-like evolutionary pressures, they are a particularly fecund source of evolutionary innovations, including new gene regulatory elements, and these are at times **exapted** by the host to improve its own fitness. If so, it is possible that many more of the one million conserved vertebrate genomic elements originated from ancient retroposon families.

Bejerano et al. Nature 2006

What about exonic UCEs?

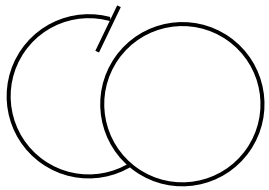


Source: Lareau, Liana F., Maki Inada, et al. "Unproductive Splicing of SR Genes Associated with Highly Conserved and Ultraconserved DNA Elements." *Nature* 446, no. 7138 (2007): 926-29.

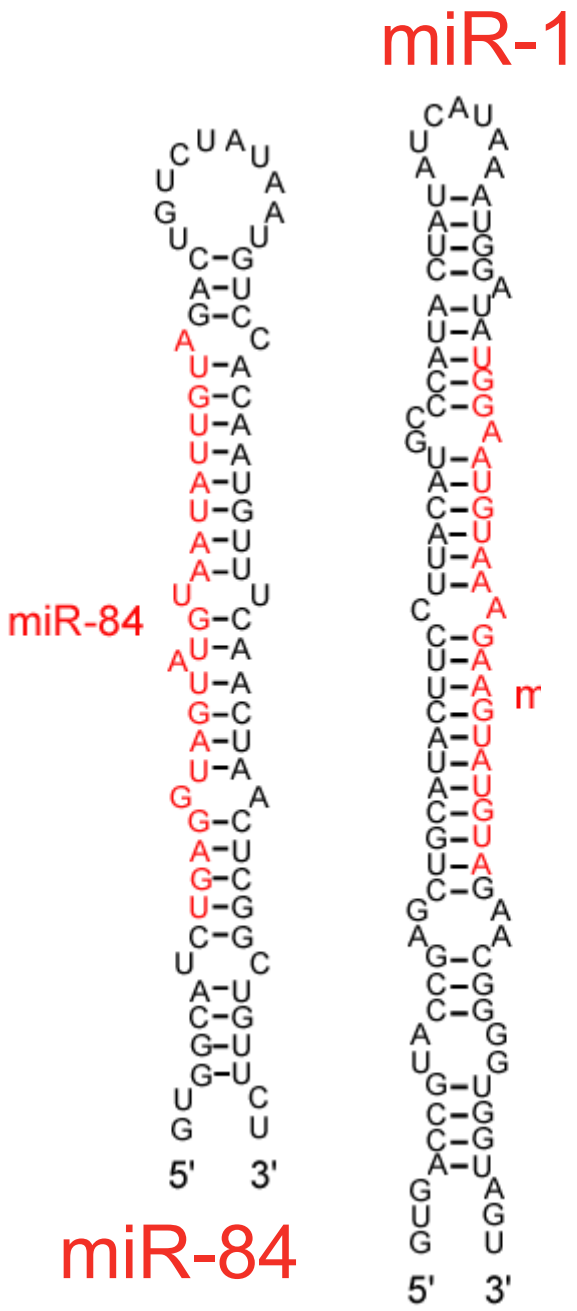
SRp20 - a splicing factor involved in constitutive and alternative splicing - contains one of the longest UCEs, overlapping a “poison cassette exon”

• mRNAs containing “premature” termination codons (PTCs) are commonly degraded by the nonsense-mediated mRNA decay pathway

• many splicing factor genes express PTC-containing mRNA isoforms - and in some cases the SF is known to promote splicing of PTC isoforms from its own locus - may ensure reduced variability in levels between cells



Lareau et al. Nature 2007



microRNAs

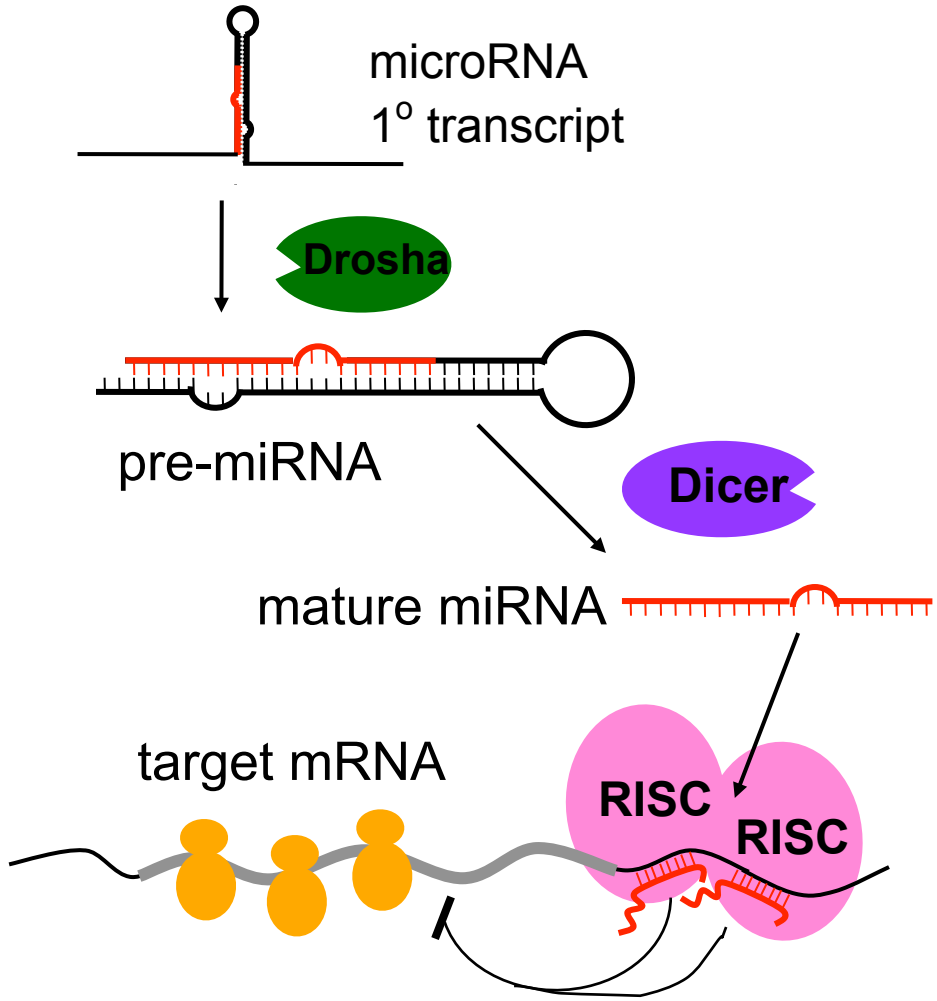
defining features:

- RNA 2° structure of precursor
- Dicer processing
- Expressed product is ~20-23 nt
- Sequence conservation

Called microRNAs or miRNAs

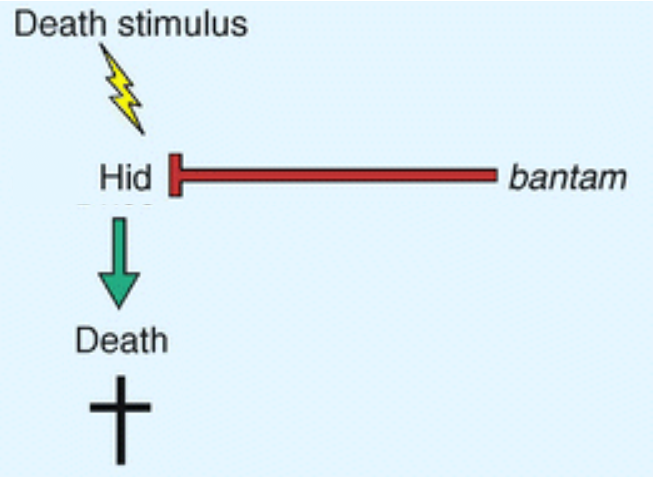
Named: *mir-X* (gene), miR-X (RNA)

microRNA biogenesis/function



translational repression
and/or mRNA degradation

MicroRNAs and apoptosis in *Drosophila*

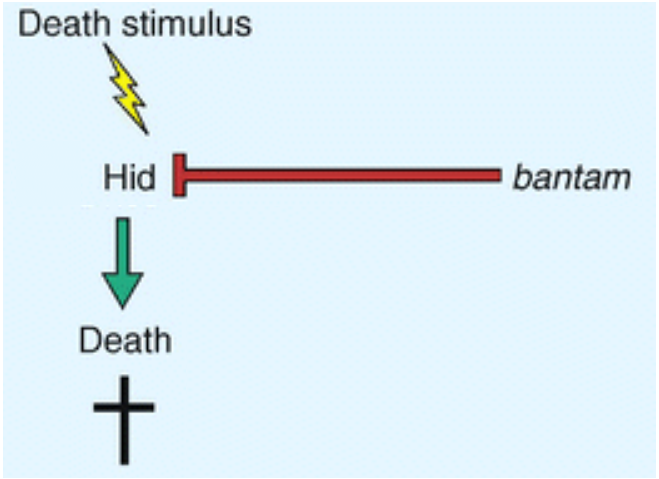


Courtesy of Elsevier. Used with permission.

Brennecke *Curr. Biol.* 2003

Phenotype of the *bantam* knockout fly pupa

bantam knockout pupa



Courtesy of Elsevier. Used with permission.

Pupa expressing *bantam* microRNA

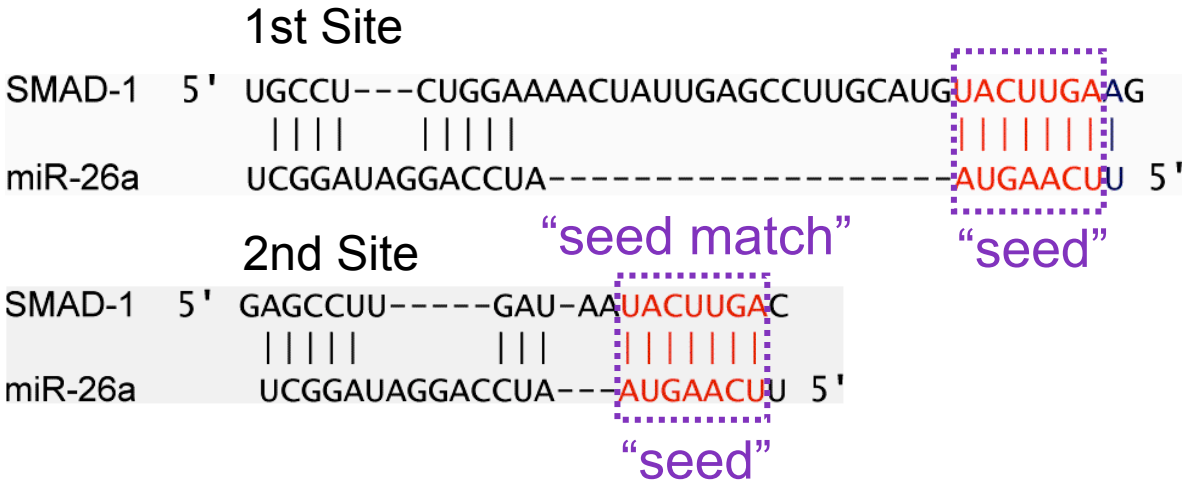
Brennecke et al. *Cell* 2003

Courtesy of Elsevier. Used with permission.
Source: Brennecke, Julius, David R. Hipfner, et al. "*bantam* Encodes a Developmentally Regulated MicroRNA that Controls Cell Proliferation and Regulates the Proapoptotic Gene *hid* in *Drosophila*." *Elsevier journal* 113, no. 1 (2003): 25-36.

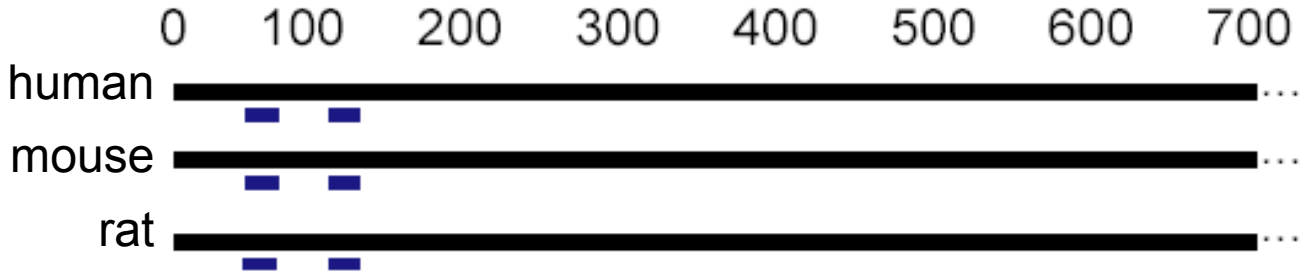
Original TargetScan Algorithm

Example: miR-26a / SMAD-1

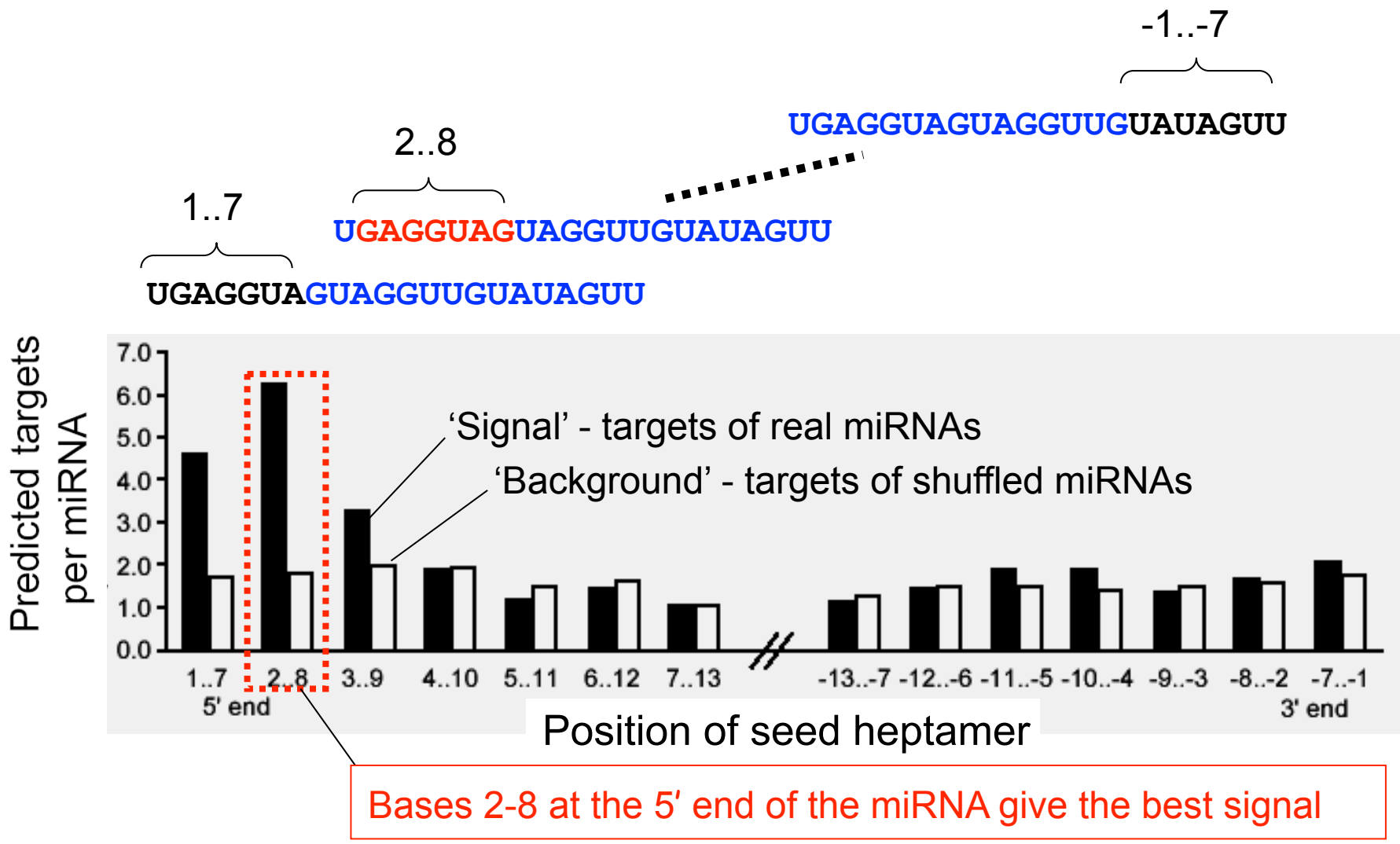
Require Watson-Crick pairing to miRNA bases 2-8



SMAD-1 3' UTR



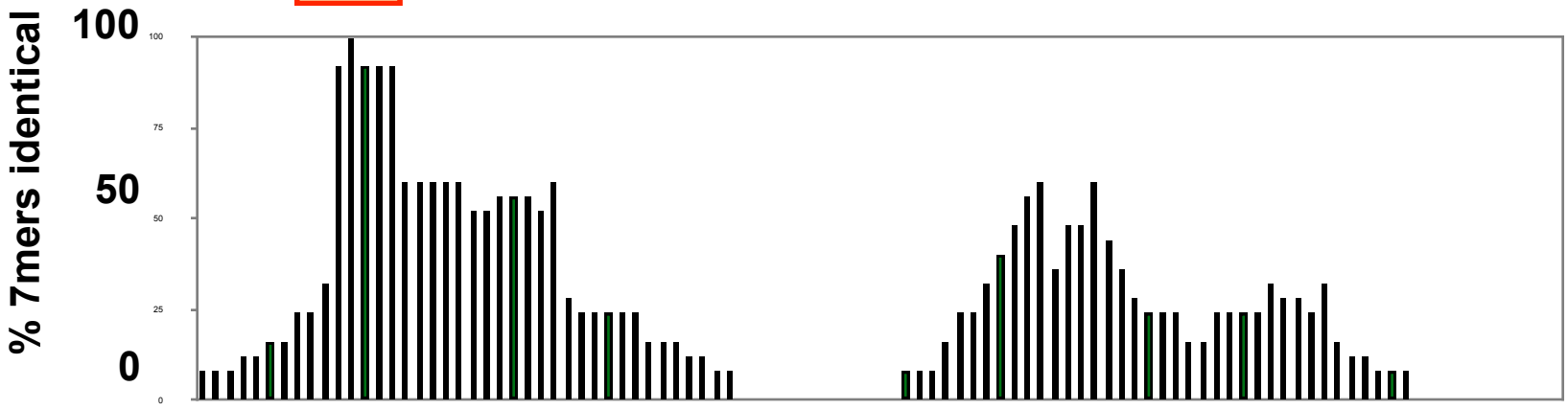
Perturbing the Model: “Sliding Seed” Experiment



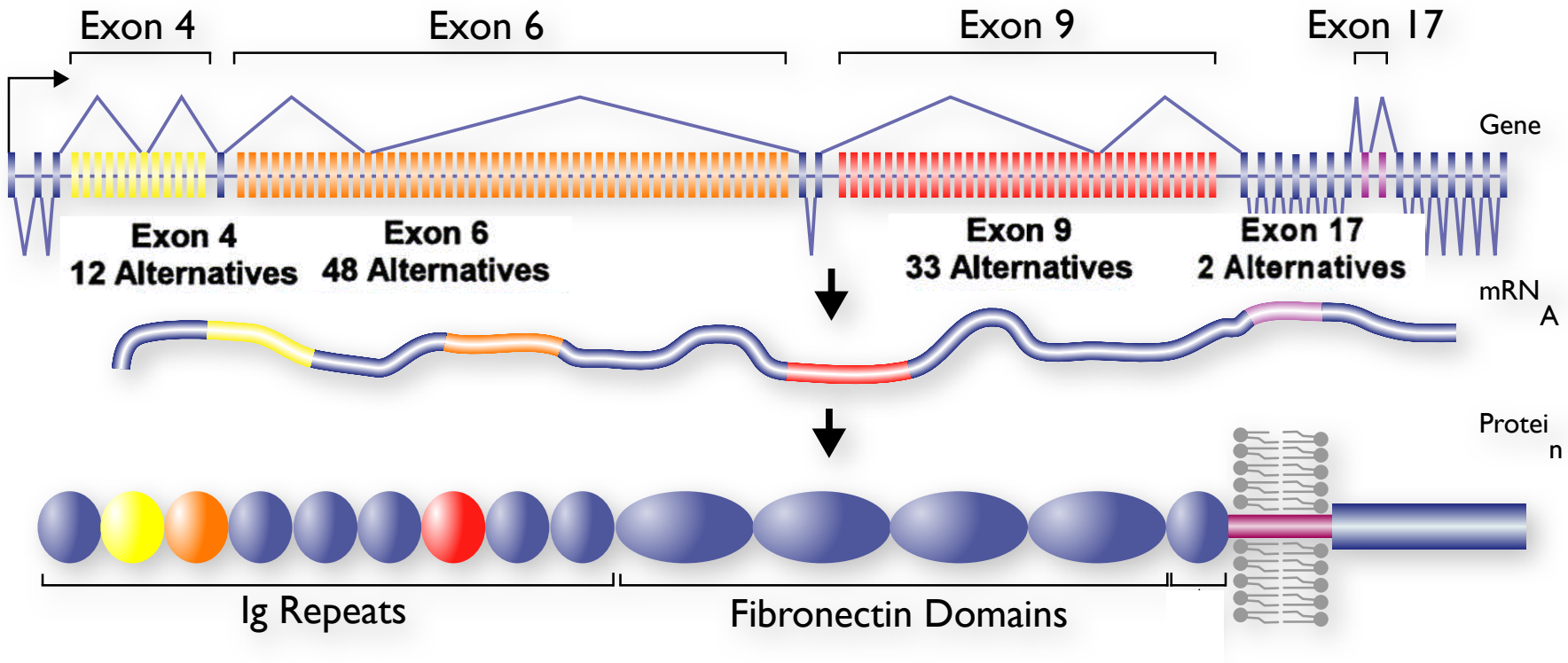
Lewis et al. *Cell* 2003

Conservation of *let-7* Foldbacks

```
mm-let-7c-1  -UGUGUGCAUCCGGGUUGAGGUAGUAGGUUGUAUGGUU--UAGAGUUACACCCUGG----------GAGUUAACUGUACAACCUUCUAGCUUUCCUUGGAGCACACU------
hs-let-7c    ------GCAUCCGGGUUGAGGUAGUAGGUUGUAUGGUU--UAGAGUUACACCCUGG----------GAGUUAACUGUACAACCUUCUAGCUUUCCUUGGAGC----------
hs-let-7a-2  ------------AGGUUGAGGUAGUAGGUUGUAUAGUU--UAGAAUUACAUCAAGG----------GAGAUAACUGUACAGCCUCCUAGCUUUCCU----------------
mm-let-7a-2  CUGCAUGUUCCCAGGUUGAGGUAGUAGGUUGUAUAGUU--UAGAGUUACAUCAAGG----------GAGAUAACUGUACAGCCUCCUAGCUUUCCUUGGGACUUGCAC-----
hs-let-7f-1  ----------UCAGAGUGAGGUAGUAGAUUGUAUAGUU-GUGGGGUAGUGAUUUUACCCUGUUCAGGAGAUAACUAUACAAUCUAUUGCCUUCCCUGA---------------
mm-let-7f-1  ---------AUCAGAGUGAGGUAGUAGAUUGUAUAGUU-GUGGGGUAGUGAUUUUACCCUGUUUAGGAGAUAACUAUACAAUCUAUUGCCUUCCCUGAG--------------
mm-let-7b    ----------GCAGGGUGAGGUAGUAGGUUGUGUGGUU-UCAGGGCAGUGAUGUUGCCCC--UCCGAAGAUAACUAUACAACCUACUGCCUUCCCUGA---------------
hs-let-7b    ----------CGGGGUGAGGUAGUAGGUUGUGUGGUU-UCAGGGCAGUGAUGUUGCCCC--UCGGAAGAUAACUAUACAACCUACUGCCUUCCCUG----------------
hs-let-7i    ----------CUGGCUGAGGUAGUAGUUUGUGCUGUUGGUCGGGUUGUGACAUUGCCCGCUGU-GGAGAUAACUGCGCAAGCUACUGCCUUGCUA----------------
mm-let-7i    ----------CUGGCUGAGGUAGUAGUUUGUGCUGUUGGUCGGGUUGUGACAUUGCCCGCUGU-GGAGAUAACUGCGCAAGCUACUGCCUUGCUAG---------------
mm-let-7g    ----------CCAGGCUGAGGUAGUAGUUUGUACAGUUUGAGGGUCUAUGAUACCACCCGGUACAGGAGAUAACUGUACAGGCCACUGCCUUGCCAGG---------------
hs-let-7g    ------------AGGCUGAGGUAGUAGUUUGUACAGUUUGAGGGUCUAUGAUACCACCCGGUACAGGAGAUAACUGUACAGGCCACUGCCUUGCCA----------------
hs-let-7a-3  -------------GGGUGAGGUAGUAGGUUGUAUAGUU--UGGGGCUCUG-CCCUGCUAU------GGGAUAACUAUACAAUCUACUGUCUUUCCU----------------
mm-let-7c-2  ---ACGGCCUUUGGGGUGAGGUAGUAGGUUGUAUGGUU--UUGGGCUCUG-CCCCGCUCU------GCGGUAACUAUACAACUCUACUGUCUUUCCUGAAGUGGCCGC------
mm-let-7d    -AAUGGGUUCCUAGGAAGAGGUAGUAGGUUGCAUAGUU-UUAGGGCAGAGAUUUUGCCCAC--AAGGAGUUAACUAUACGACCUGCUGCCUUUCUUAGGGCCUUAUU------
hs-let-7d    ---------CCUAGGAAGAGGUAGUAGGUUGCAUAGUU-UUAGGGCAGGGAUUUUGCCCAC--AAGGAGGUAACUAUACGACCUGCUGCCUUUCUUAGG-------------
hs-let-7a-1  ----------UGGGAUGAGGUAGUAGGUUGUAUAGUU-UUAGGGUCACACCCACCACUG-----GGAGAUAACUAUACAAUCUACUGUCUUUCCUA----------------
mm-let-7a-1  ----UUCACUGUGGGAUGAGGUAGUAGGUUGUAUAGUU-UUAGGGUCACACCCACCACUG-----GGAGAUAACUAUACAAUCUACUGUCUUUCCUAAGGUGAU---------
hs-let-7f-2  ---------UGUGGGAUGAGGUAGUAGAUUGUAUAGUU-UUAGGGUCAUACCC-CAUCUU-----GGAGAUAACUAUACAGUCUACUGUCUUUCCCACG--------------
mm-let-7f-2  ---------UGUGGGAUGAGGUAGUAGAUUGUAUAGUU-UUAGGGUCAUACCC-CAUCUU-----GGAGAUAACUAUACAGUCUACUGUCUUUCCCACG--------------
mm-let-7e    --CGCGCCCCCCGGGCUGAGGUAGGAGGUUGUAUAGUU---GAGGAAGACACCCGA---------GGAGAUCACUAUACGGCCUCCUAGCUUUCCCCAGGCUGCGCC------
hs-let-7e    ---------CCCGGGCUGAGGUAGGAGGUUGUAUAGUU---GAGGAGGACACCCAA---------GGAGAUCACUAUACGGCCUCCUAGCUUUCCCCAGG-------------
cb-let-7     ---ACUG-GGGUACGGUGAGGUAGUAGGUUGUAUAGUU--UAGAAUAUUACUCUCG----------GUGAACUAUGCAAGUUUCUACCUCACCGAAUACCAGG----------
ce-let-7     UACACUGUGGAUCCGGUGAGGUAGUAGGUUGUAUAGUU--UGGAAUAUUACCACCG----------GUGAACUAUGCAAUUUUCUACCUUACCGGAGACAGAACUCUUCGA
dm-let-7     ------UCUGGCAAAUUGAGGUAGUAGGUUGUAUAGU----AGUA-AUUACACAUC----------AU--ACUAUACAAUGUGCUAGCUUUCUUUGCUUGA----------
```



Dscam - an extreme case of alternative splicing

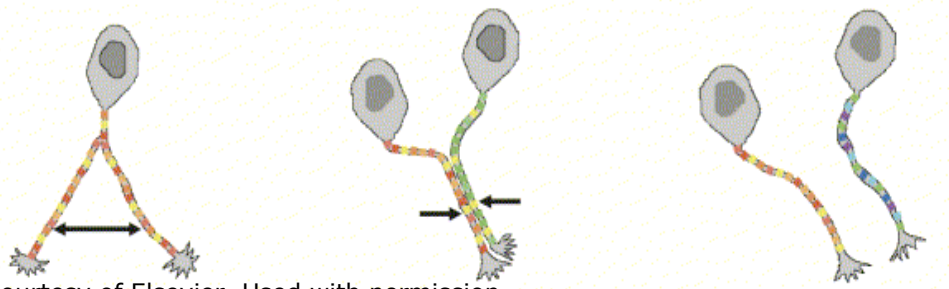


Dscam can potentially express > 38,000 isoforms, which control neuronal wiring

How is splicing regulated?

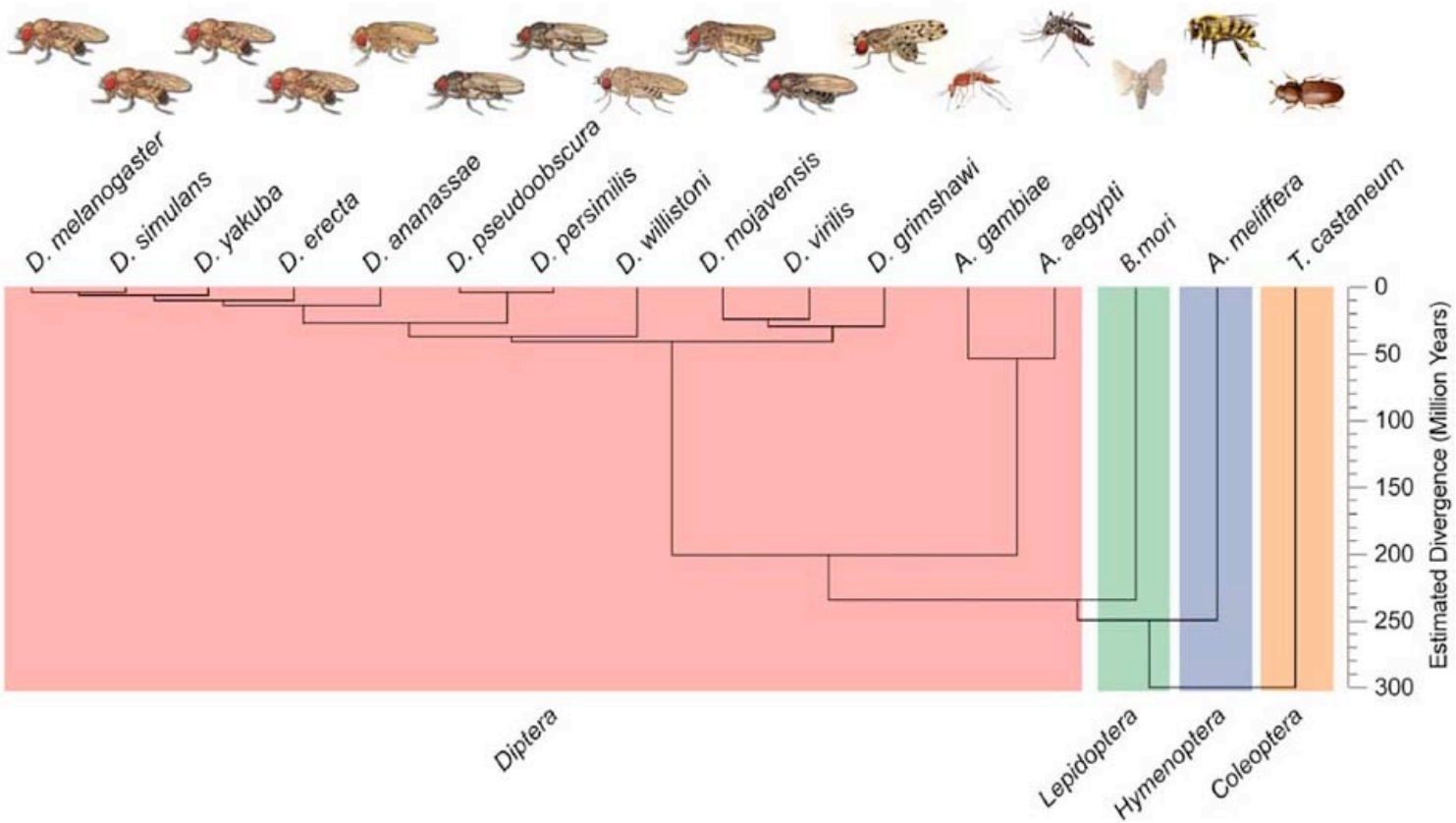
How is mutually exclusive inclusion of exons achieved?

Schmücker *et al. Cell* 2000



Courtesy of Elsevier. Used with permission.
Source: Wojtowicz, Woj M., John J. Flanagan, et al. "Alternative Splicing of *Drosophila* Dscam Generates Axon Guidance Receptors that Exhibit Isoform-Specific Homophilic Binding." *Cell* 118, no. 5 (2004): 619-33.

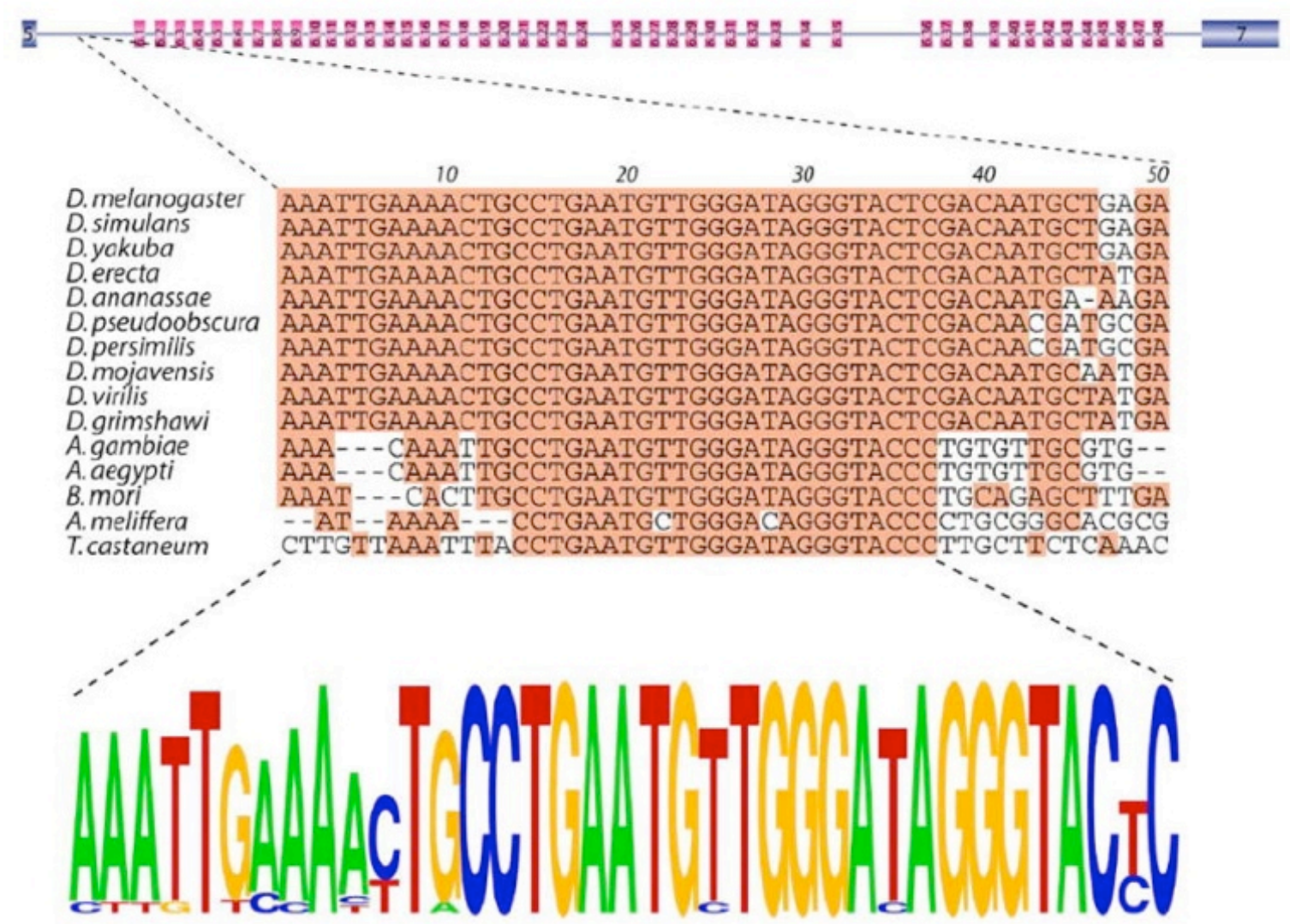
Motif downstream of DSCAM exon 5



Source: Graveley, Brenton R. "Mutually Exclusive Splicing of the Insect *Dscam* Pre-mRNA Directed by Competing Intronic RNA Secondary Structures." *Cell* 123, no. 1 (2005): 65-73.

Graveley *Cell 2005*

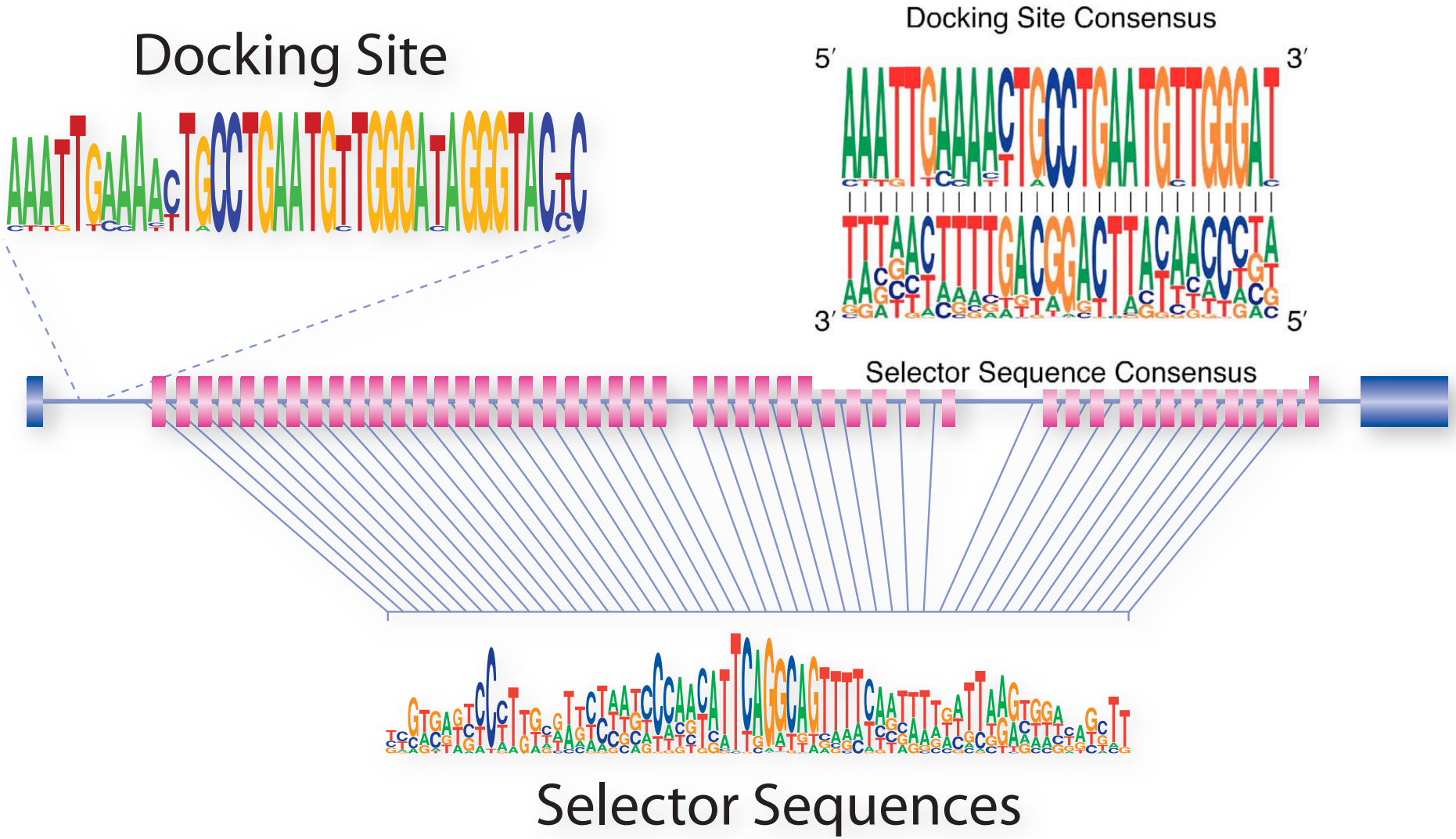
Motif downstream of DSCAM exon 5



Courtesy of Elsevier. Used with permission.
Source: Graveley, Brenton R. "Mutually Exclusive Splicing of the Insect *Dscam* Pre-mRNA Directed by Competing Intronic RNA Secondary Structures." *Cell* 123, no. 1 (2005): 65-73.

Graveley *Cell 2005*

Mutually Exclusive Splicing of the Exon 6 Cluster

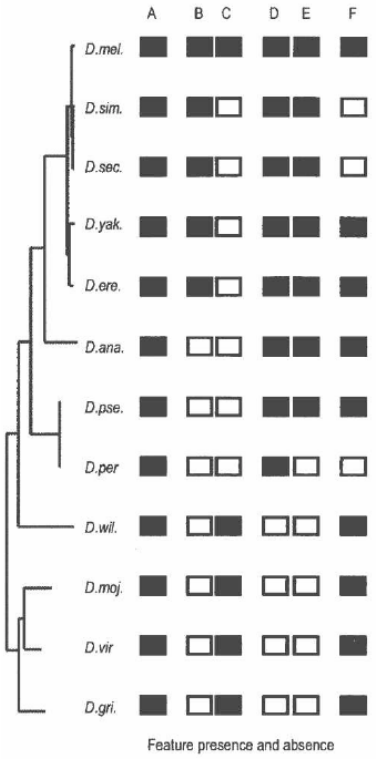


Source: Graveley, Brenton R. "Mutually Exclusive Splicing of the Insect *Dscam* Pre-mRNA Directed by Competing Intronic RNA Secondary Structures." *Cell* 123, no. 1 (2005): 65-73.

Graveley *Cell 2005*

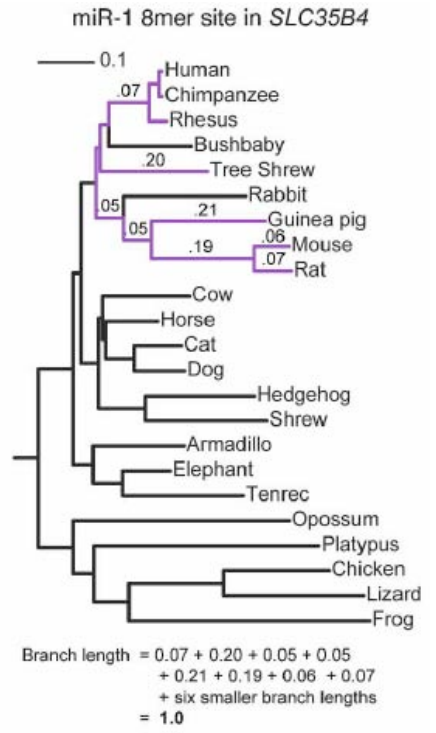
Defining a Branch Length Score to assess conservation

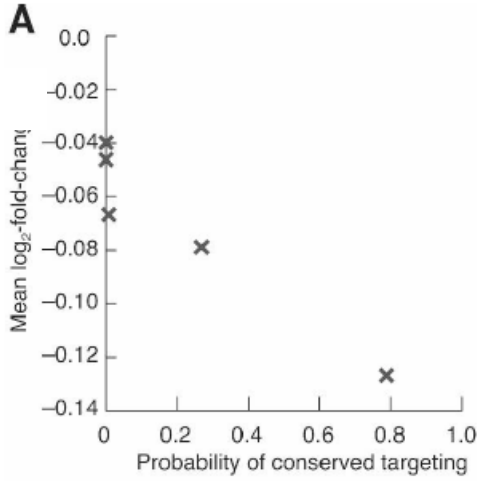
Species Set	Total BL	Total rel. BL
A	5.3	100%
B	0.4	7%
C	3.6	68%
D	2.5	48%
E	2.2	43%
F	4.9	94%

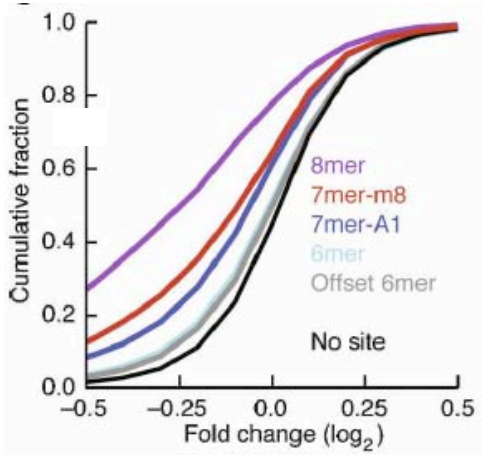


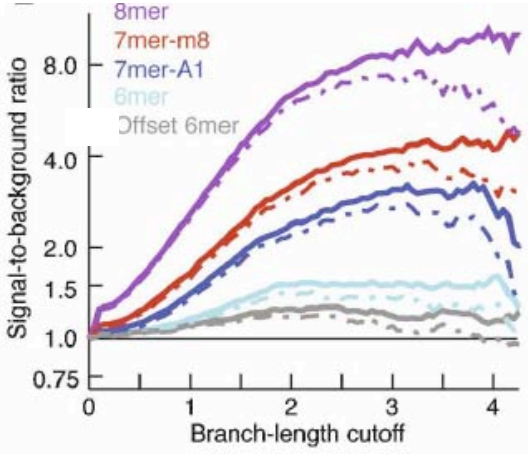
Kheradpour et al Genome Res 2008

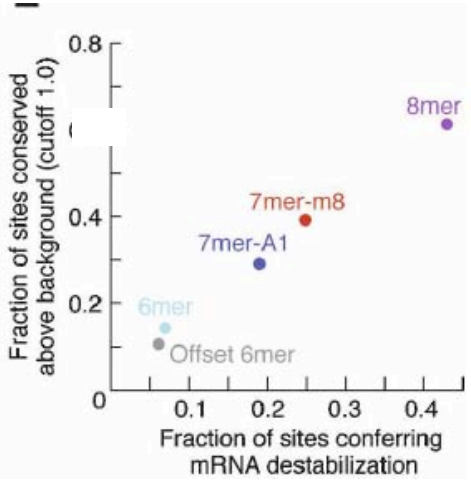
Using a similar branch length conservation measure to assess and classify mammalian miRNA target sites





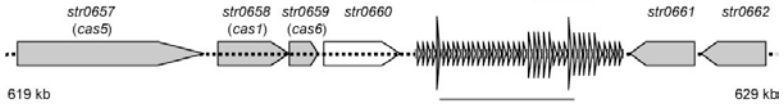






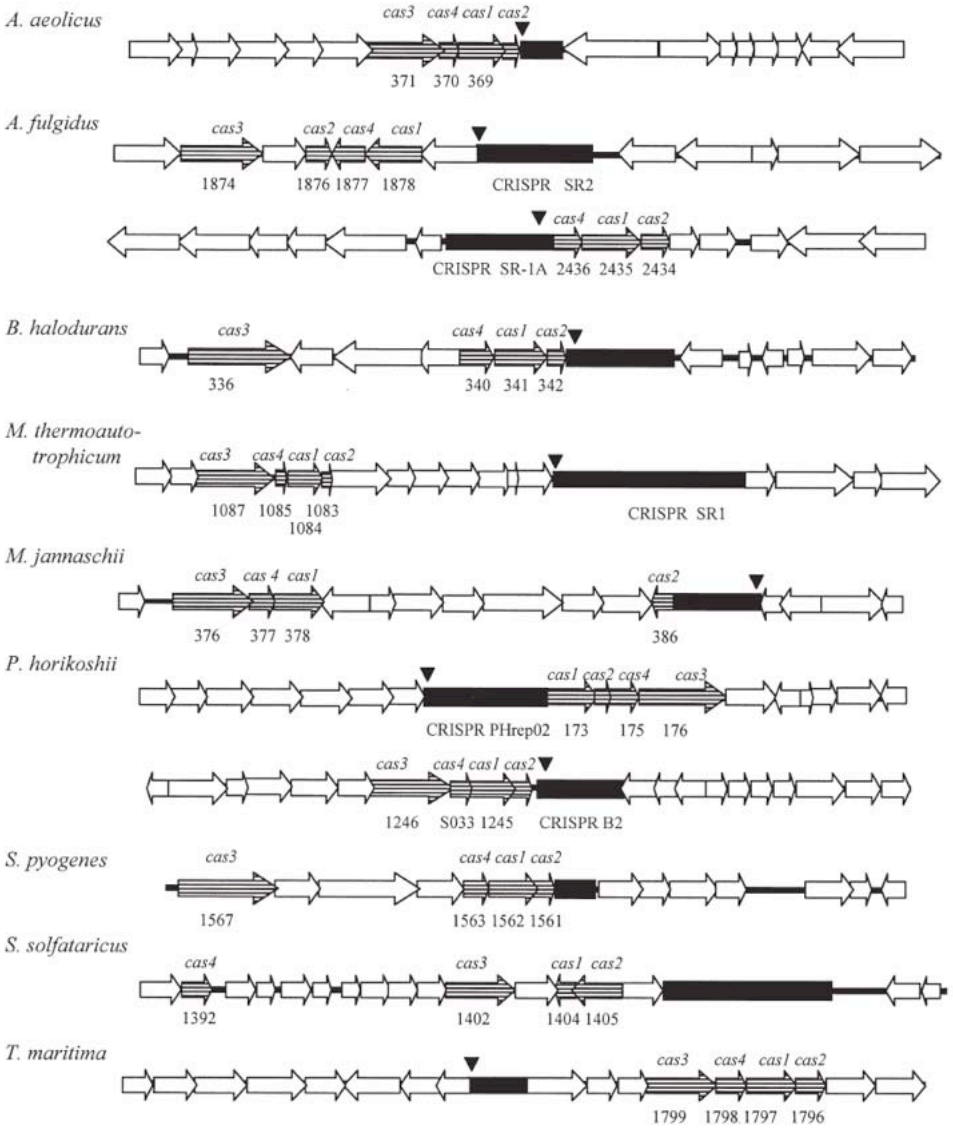
Friedman et al Genome Res 2009

Identifying a family of genes (cas) associated with a bacterial repeat structure (CRISPR)



Source: Bolotin, Alexander, Benoit Quinquis, et al. "Clustered Regularly Interspaced Short Palindrome Repeats(CRISPRs) have Spacers of Extra-chromosomal Origin." *Microbiology* 151, no. 8 (2005): 2551-61.

Jansen et al Mol Microbiol 2002



Source: Jansen, Ruud, Jan Embden, et al. "Identification of Genes that are Associated with DNA Repeats in Prokaryotes." *Molecular Microbiology* 43, no. 6 (2002): 1565-75.

CRISPR spacers match phage genomes

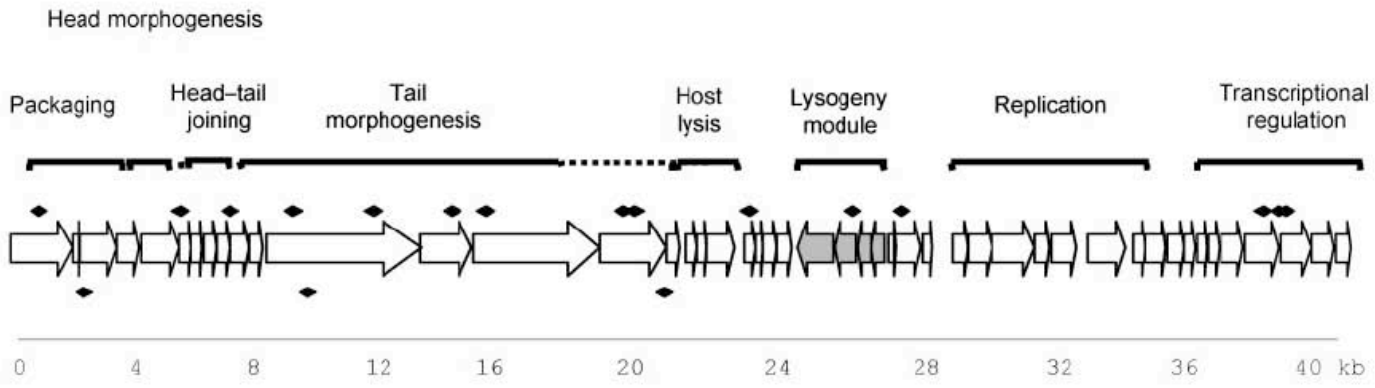


Fig. 4. Localization of spacer-matching sequences along the phage Sfi21 genome. The phage genetic map is drawn after GenBank entry NC_000872 (ORFs are shown as arrows), the regions involved in different stages of phage development, identified by comparative analysis (Desiere *et al.*, 2002), are indicated above the map, and the scale (in kb) below it. Phage regions having a BLAST E score <0·001 with the CRISPR spacers are indicated by the diamonds placed above or below the map, denoting homology with the top or the bottom DNA strand, respectively.

Number of spacers is correlated with resistance to phage

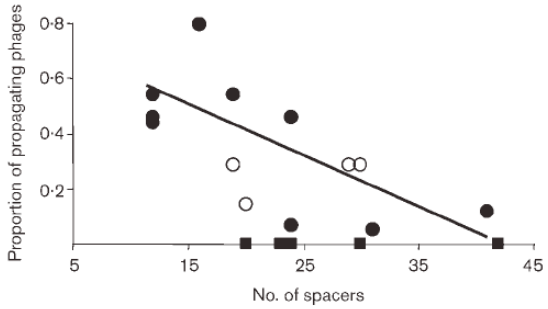


Fig. 6. Correlation of *S. thermophilus* phage resistance and the number of spacers in a CRISPR locus. Filled symbols correspond to data obtained from strains tested with the panel of 59 phages. The line of best-fit refers to strains that were not fully phage resistant (●), and for which $y = -0{\cdot}02x + 0{\cdot}77$ and $R^2 = 0{\cdot}51$. Fully phage-resistant strains (■), were not taken into account for the correlation shown. ○, Strains tested with the panel of seven phages.

Source: Bolotin, Alexander, Benoit Quinquis, et al. "Clustered Regularly Interspaced Short Palindrome Repeats(CRISPRs) have Spacers of Extrachromosomal Origin." *Microbiology* 151, no. 8 (2005): 2551-61.

Bolotin et al Microbiol 2005

7.91J / 20.490J / 20.390J / 7.36J / 6.802J / 6.874J / HST.506J Foundations of Computational and Systems Biology
Spring 2014