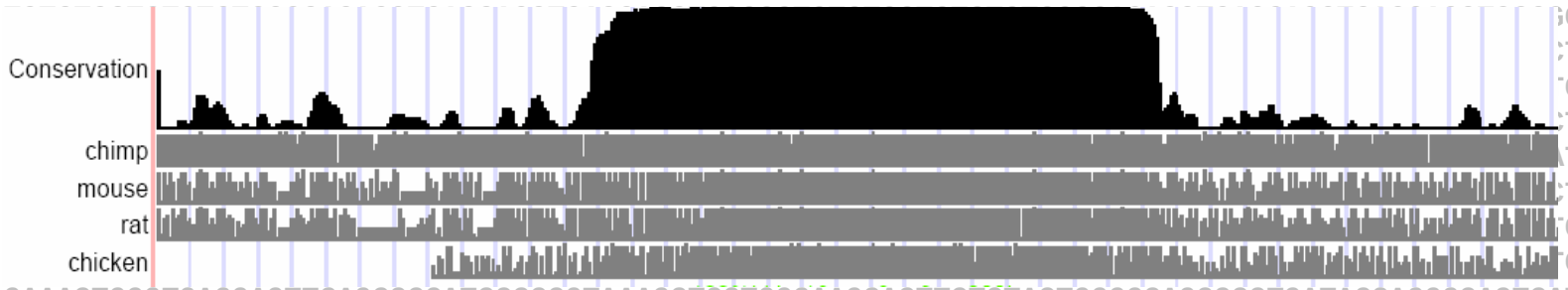
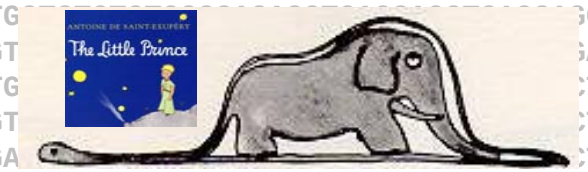
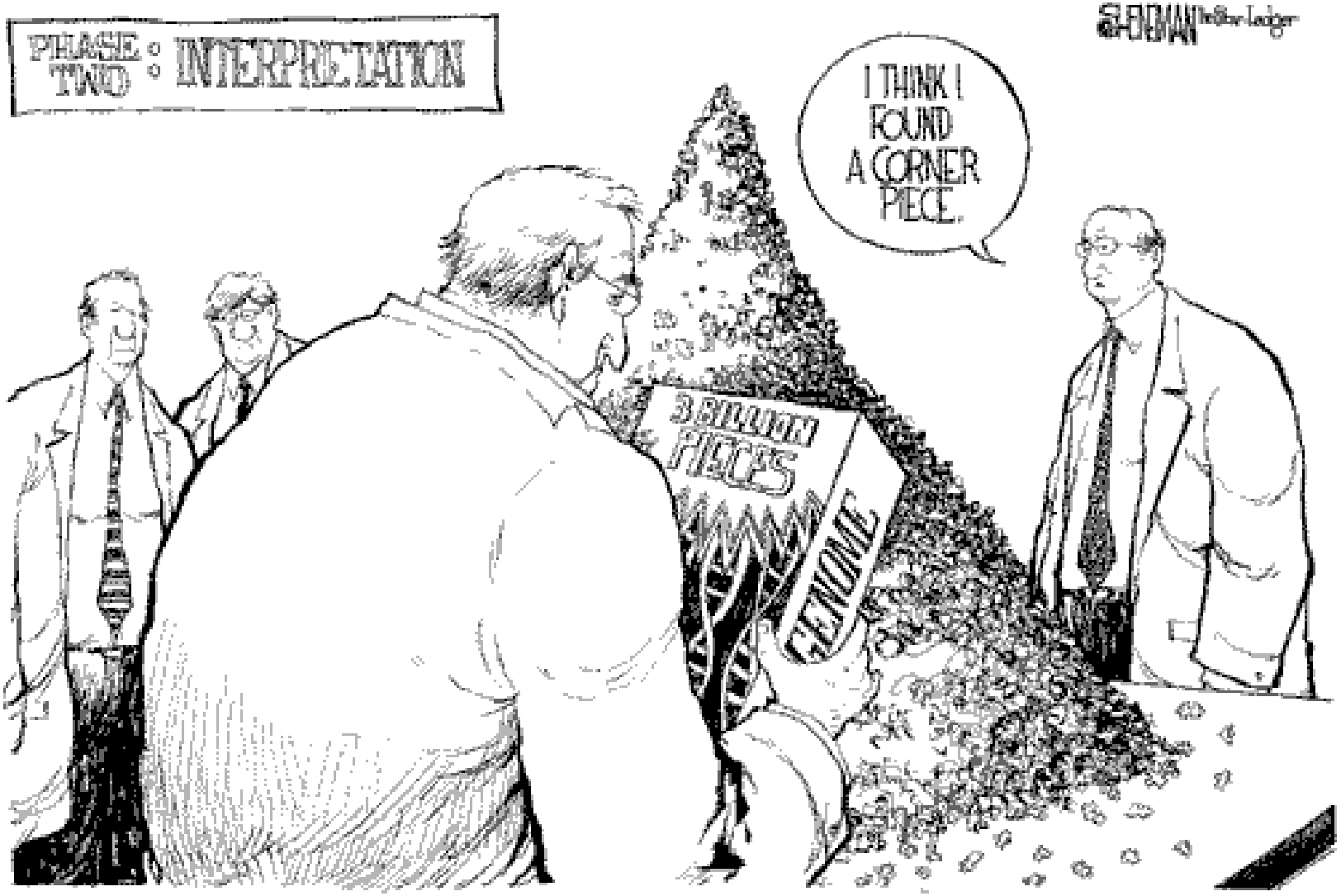
# Ultraconserved Elements in Eukaryote Genome



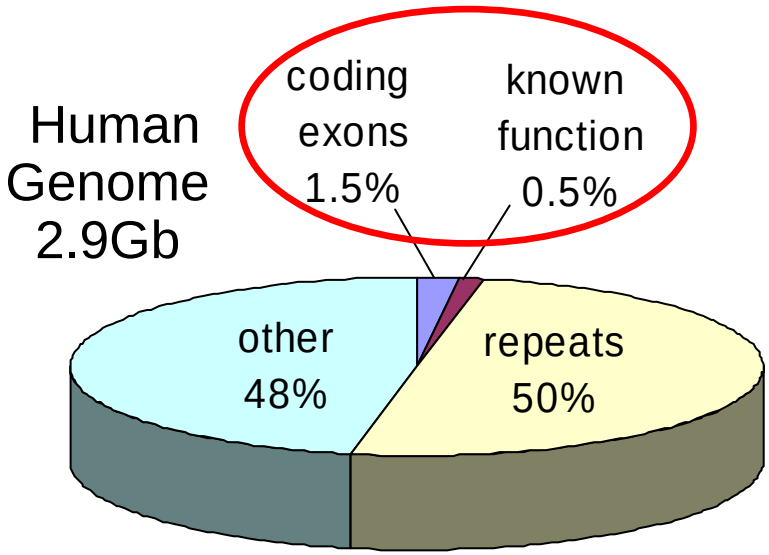



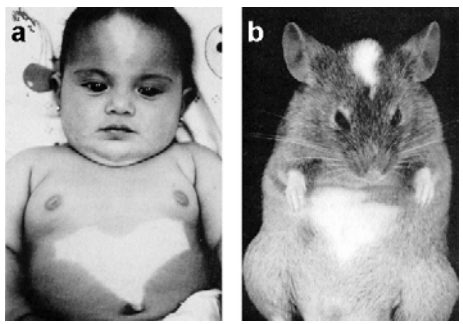

**Gill Bejerano**
**UC Santa Cruz**

# Genomics


PHASE TWO: INTERPRETATION
I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME

# Conserved Non Coding Regions




40% DNA alignable
95% coding genes shared

[*Science* 2004 Breakthrough of the Year, 5th runner up]

of the 5%
1/3 coding
**2/3 non coding**

# What To Do?

Assay! Which Assay? Which elements?

UCSC Genome Browser Home - Mozilla Firefox

**http://genome.ucsc.edu**

UCSC Genome Bioinformatics

Genomes - Blat - Tables - Gene Sorter - PCR - Proteome - FAQ - Help

Genome Browser
ENCODE
Blat
Table Browser
Gene Sorter
In Silico PCR
Proteome Browser
Utilities
Downloads
Release Log
Custom Tracks
Mirrors
Archives
Credits
Publications
Training

PROTOCOL

PUBLISHED IN ASSOCIATION WITH
COLD SPRING HARBOR LABORATORY PRESS

**one *million* elements genome wide...**

http://www.nature.com/naturemethods

## Computational screening of conserved genomic DNA in search of functional noncoding elements

Gill Bejerano[1], Adam C Siepel[1], W James Kent[1] & David Haussler[1,2]

[1]Center for Biomolecular Science and Engineering, [2]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, California 95064, USA. Correspondence should be addressed to G.B. (jill@soe.ucsc.edu).

The sequencing of the mouse genome allowed, for the first time, the large-scale estimation of the extent of sequence conservation within our own genome. In particular, it suggested that in mammals there is at least twice as much conserved genomic DNA as there is protein coding DNA[1]. The abundance of conserved noncoding regions holds even for so-called ultraconserved elements at the very tip of the mammalian

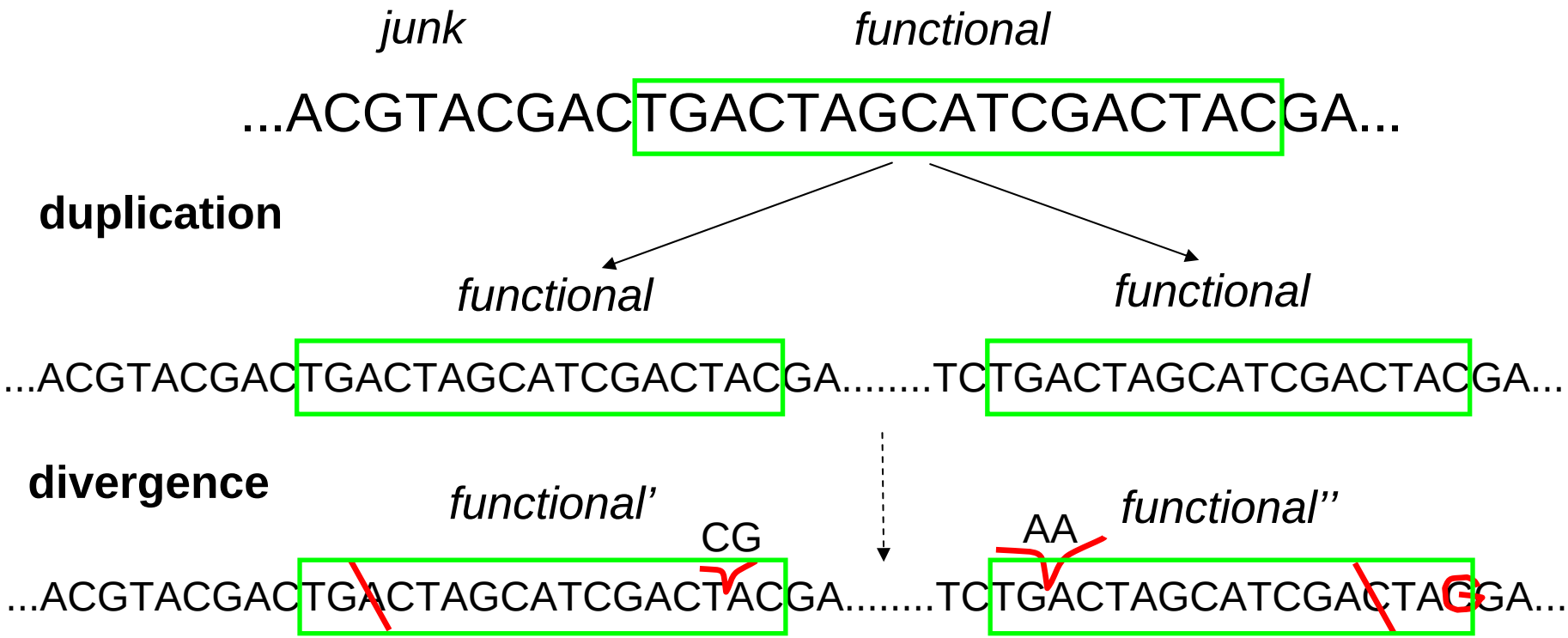
[Bejerano *et al.*, *Nature Methods* 2005]

# DNA Replication is Imperfect

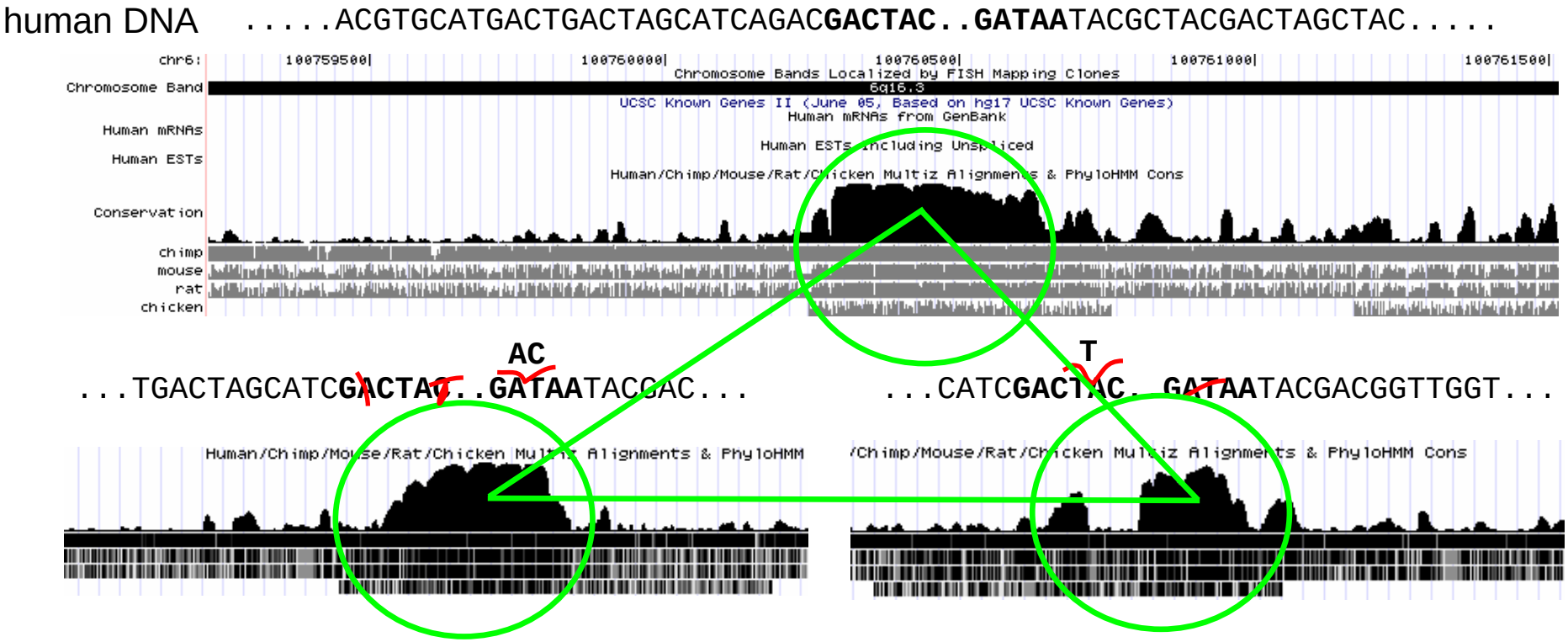
Small Scale: bases are mutated, deleted, inserted

Medium Scale: regions are duplicated, deleted, inverted

Large Scale: whole chromosomes are duplicated, merged



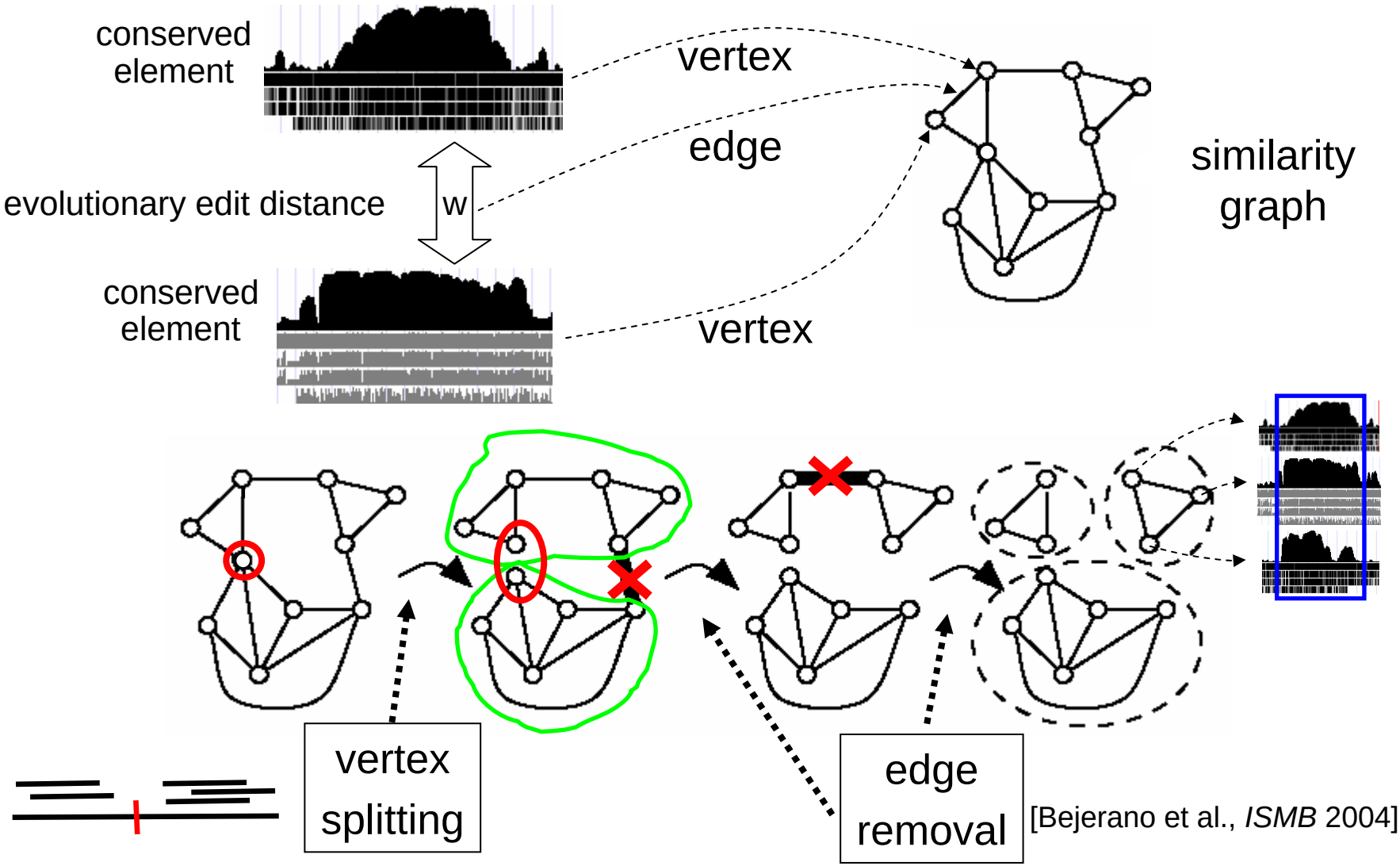
# Computational Approach




Group them into ***paralog families*** of human functional regions of common origins:

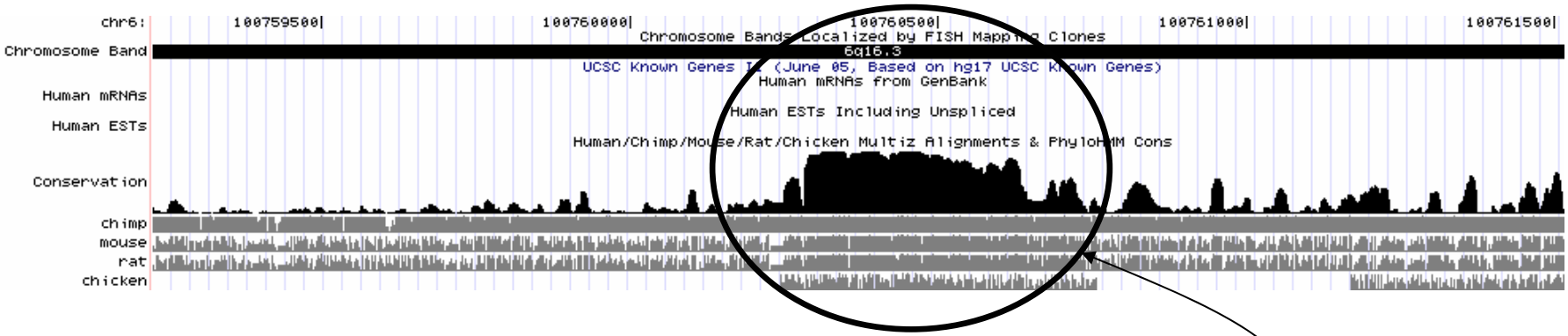
- Annotated members induce function on all.
- Examine core, substitutions in family.
- Test for "*guilt by association".* [Bejerano et al., *ISMB* 2004]

# Conserved Element Clustering



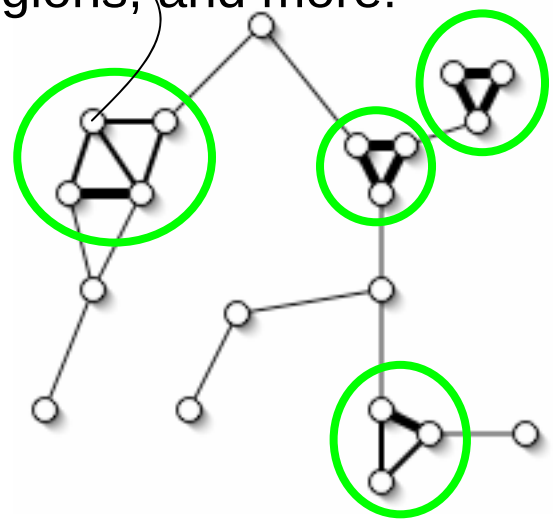

[Bejerano et al., *ISMB* 2004]

# Functional Annotation by Families




After removing from top 5% Human *all* annotated regions, and more:

700,000 elements, covering 3.5% Human Genome



metrics

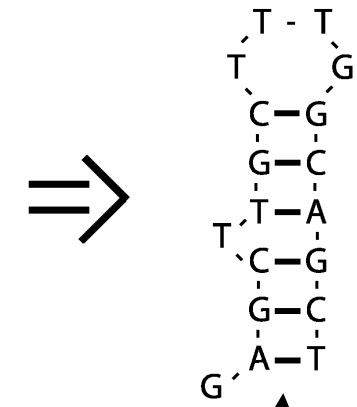
**Puzzling News:**
96% of the 700,000 are unique(!)

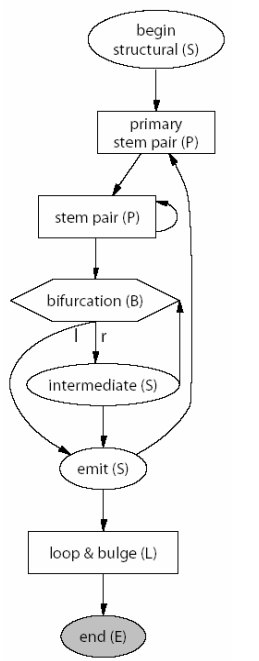
**Good News:**
We still find
**12,027 families**

**novel** putative ncRNAs, cis-regulatory elements, etc.

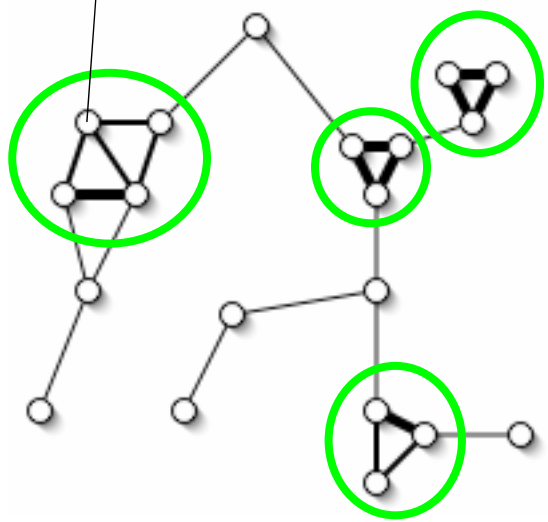
[Bejerano et al., *ISMB* 2004]

# Families of Novel ncRNAs

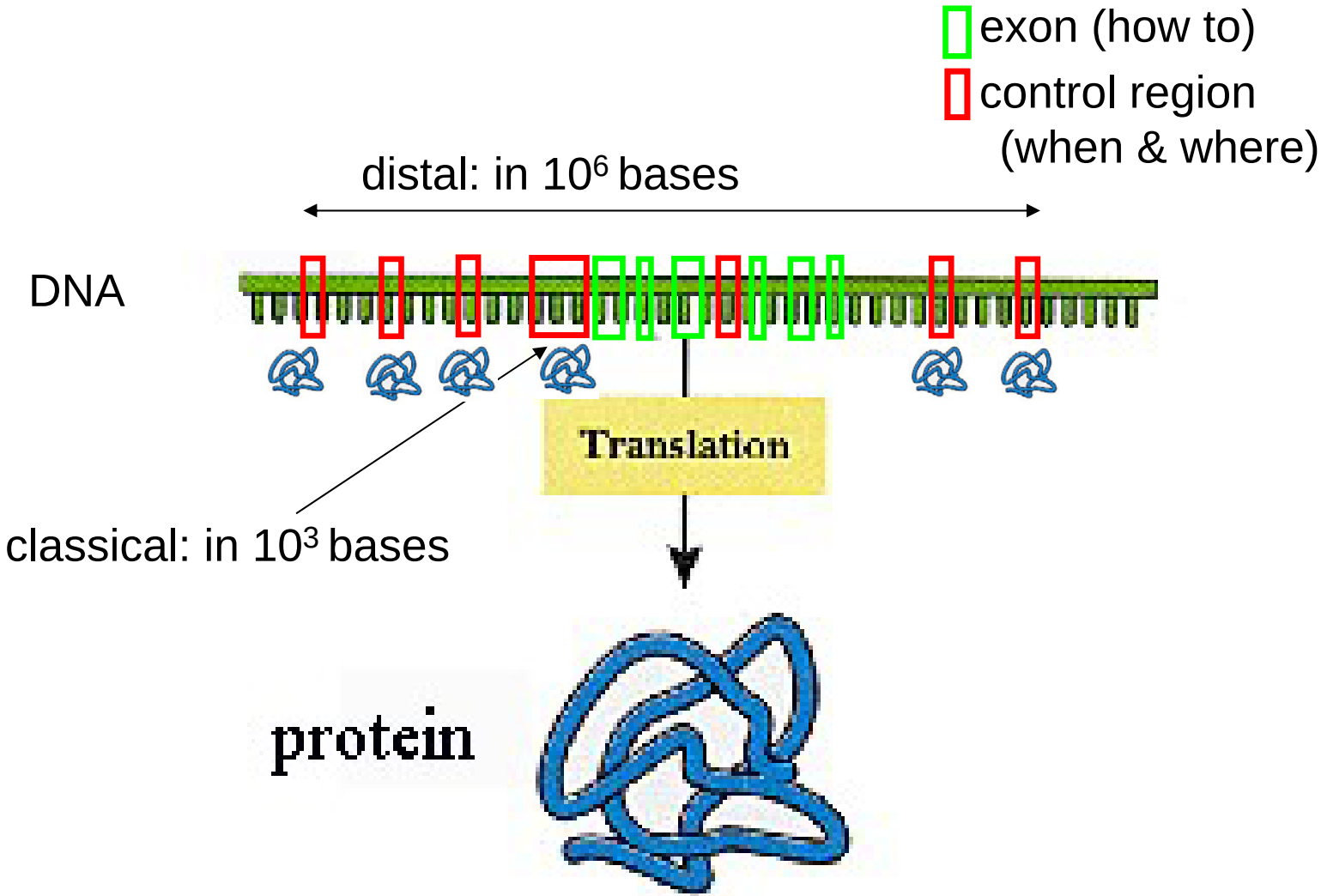




stochastic context free grammar model



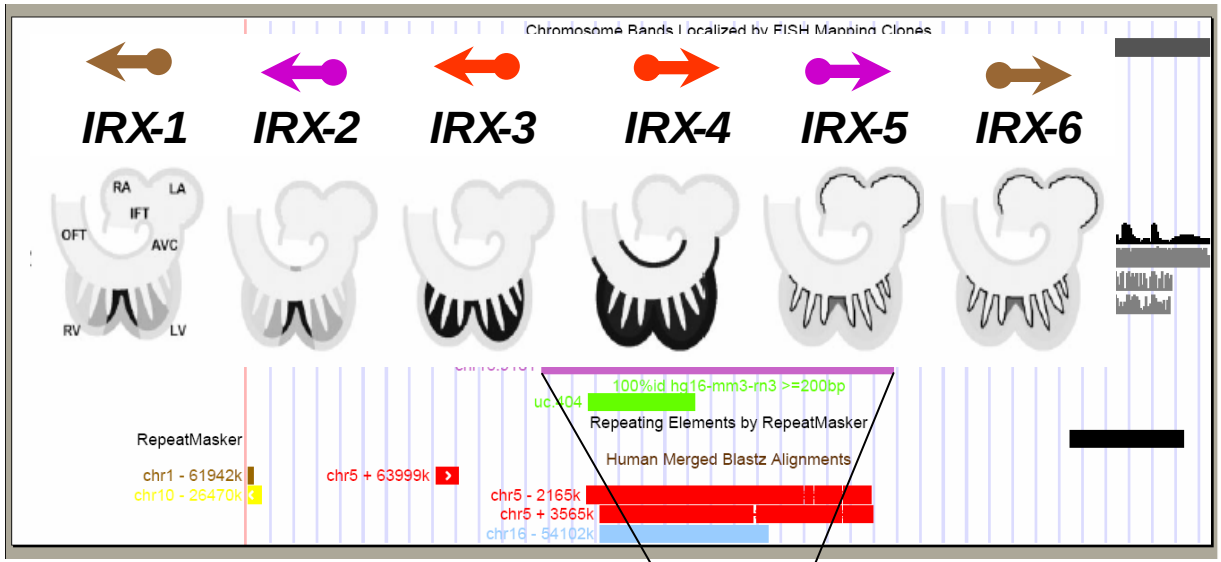
[Pedersen, Bejerano et al., *PLoS CompBio*, 2006]

# Genes, Proteins and Gene Control

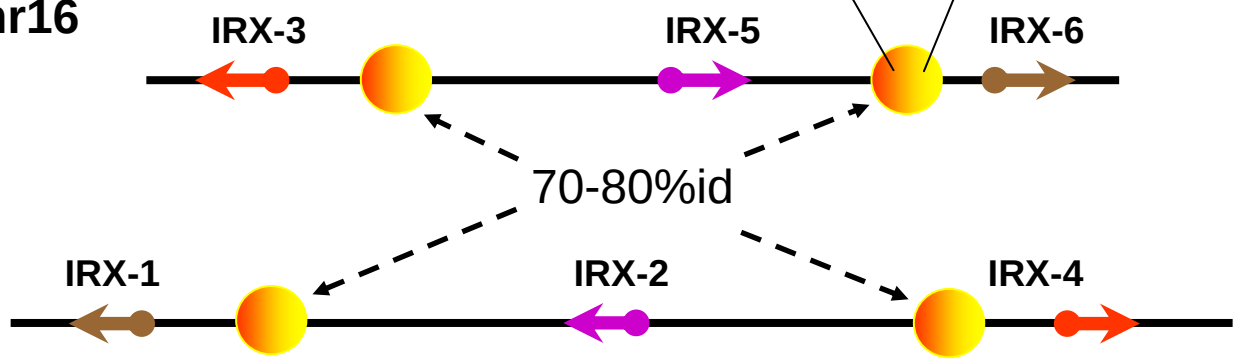


# Cis-Regulatory Families

early
body
patterning

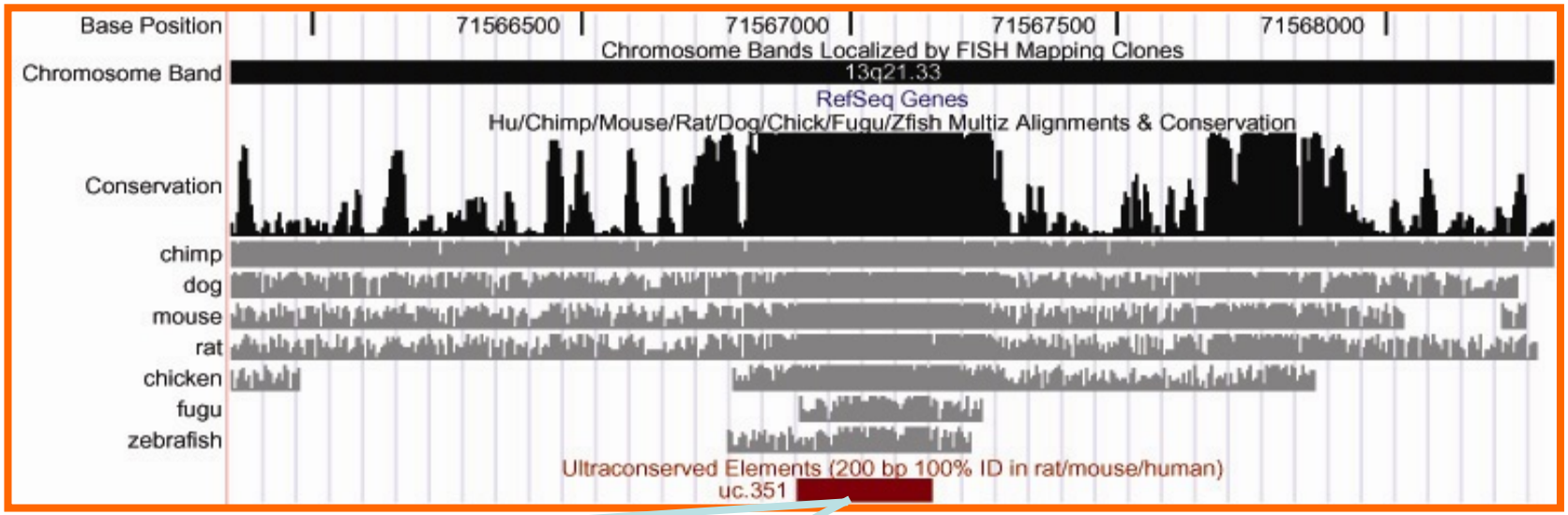


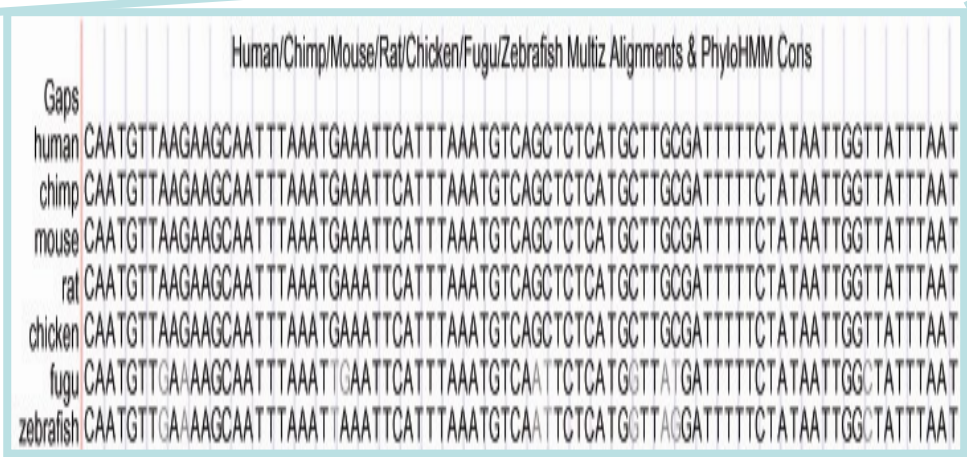

**Hs.chr16**

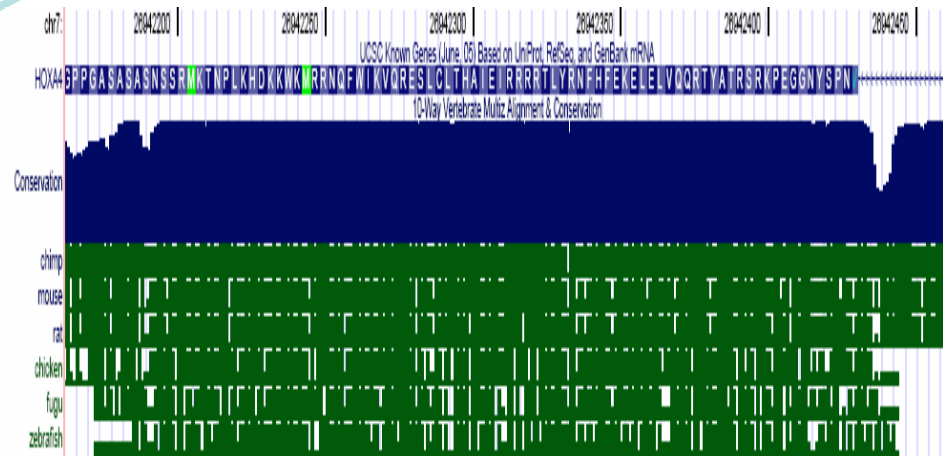



**Hs.chr5**

# Ultraconserved Elements

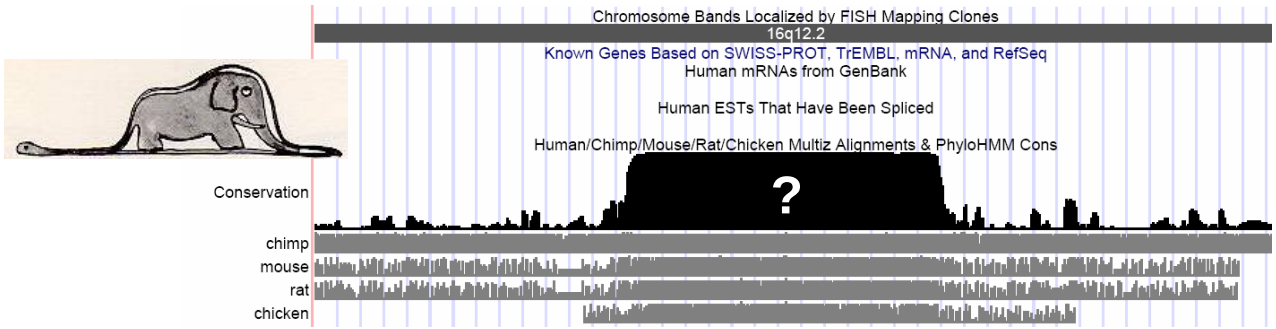


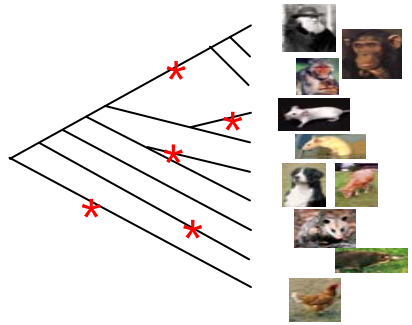




[Bejerano et al., *Science* 2004]

# No known function requires *this* much conservation



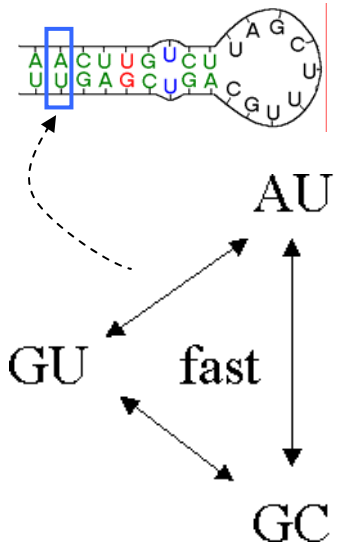



**The genetic code**

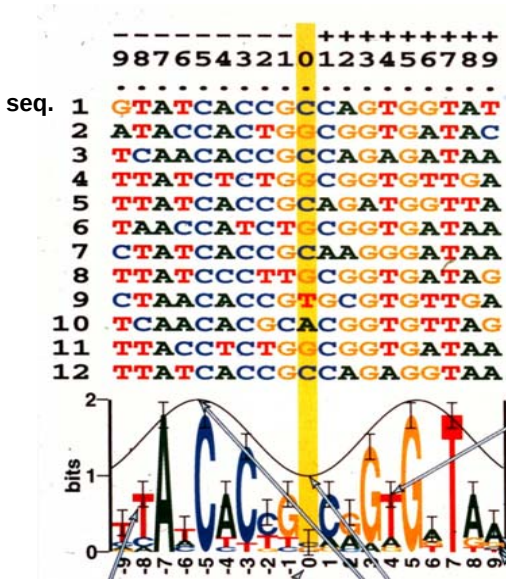
CDS

| | U | C | A | G | |
|---|---|---|---|---|---|
| U | Phe | | | | U |
| | Leu | CUU | | | C |
| C | Leu | CUC | Leucine | | A |
| | | CUA | | | G |
| A | Ile | | | | U |
| | MET | CUG | | | C |
| G | Val | | | | A |
| | | | Glu | | G |

ncRNA



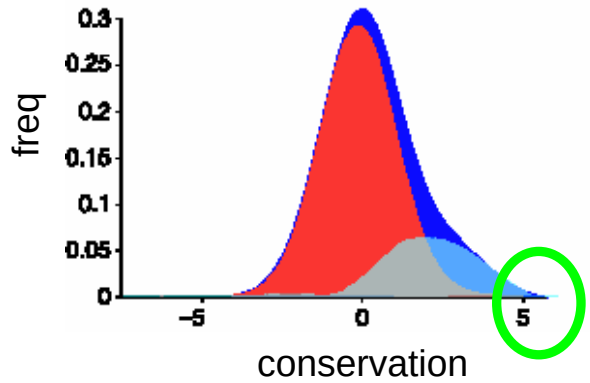

TFBS

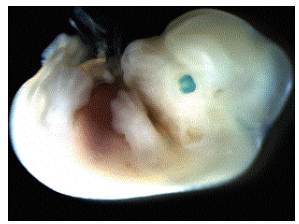


# Ultraconserved Elements

481 regions *perfectly* conserved over 200bp or more, between human, mouse and rat ($P<10^{-22}$ neutrally evolving)

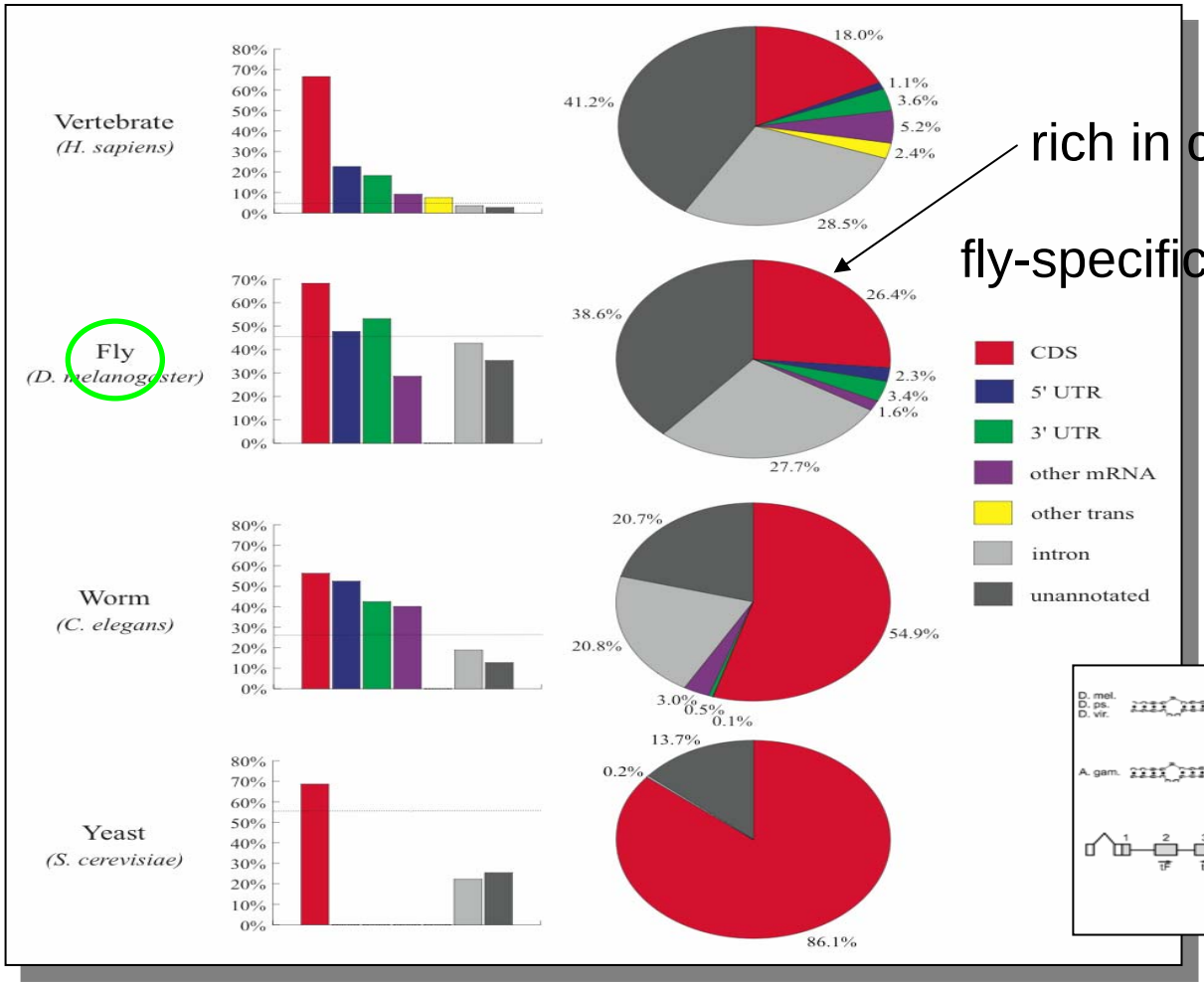
- 20-fold fewer SNPs than human average.
- Most do *not* overlap coding DNA.
- The **non-exonic** tend to cluster spatially, in or near DNA binding proteins. Dozens validated as enhancers by now.
- The **exonic** tend to overlap alternatively spliced exons, in RNA binding proteins.
- The ultras cannot be found beyond vertebrates.
- The tip of a continuum of very slowly evolving elements.



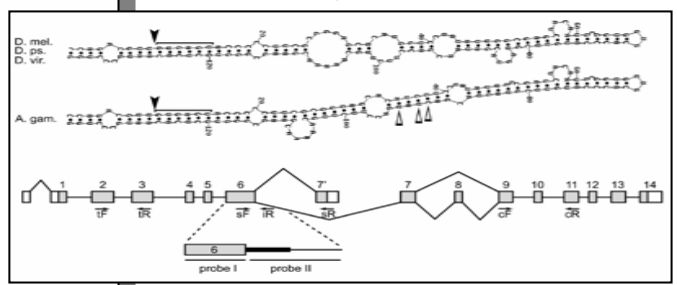
[Bejerano et al., *Science* 2004
Chicken Consortium, *Nature* 2004]

# Similar Phenomena in Flies




rich in conserved non-coding

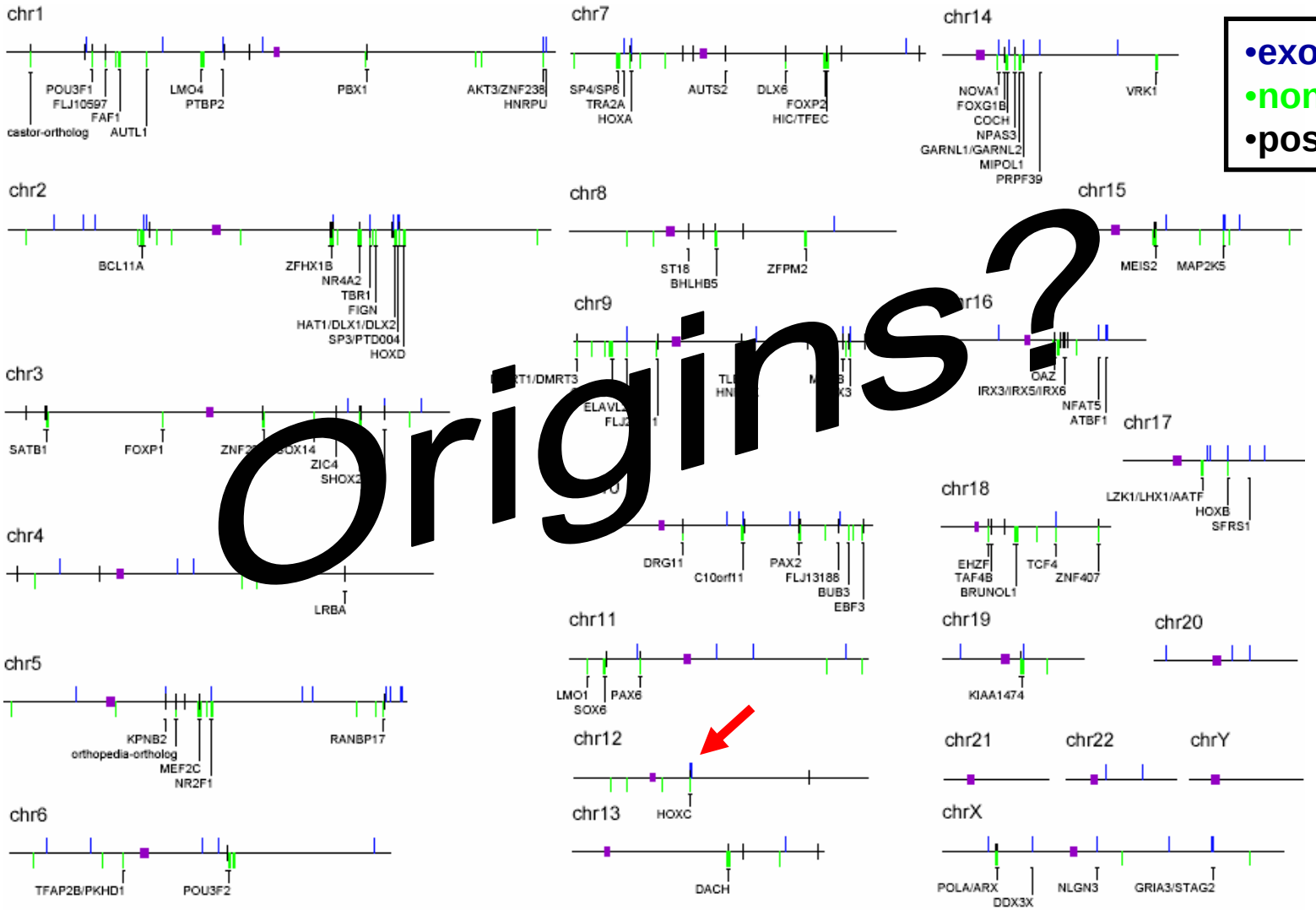
fly-specific ultraconserved elements




[Siepel, Bejerano et al., *Genome Research* 2005]

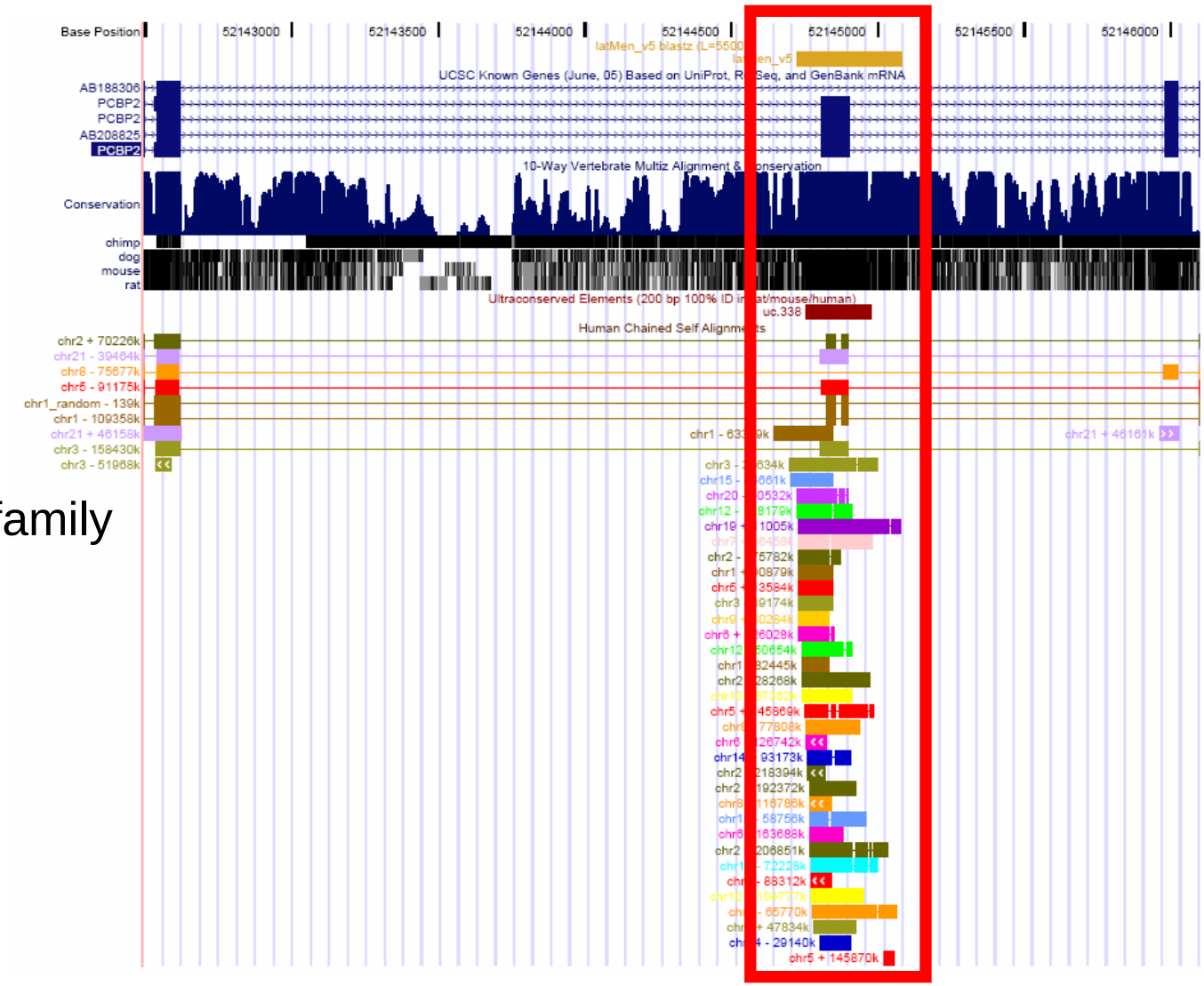
[Glazov, ..., Bejerano, Mattick, *Genome Research* 2005]

# Mammalian Ultras - Genomic Distribution



# Ultraconserved Element uc.338

uc.338 paralog family

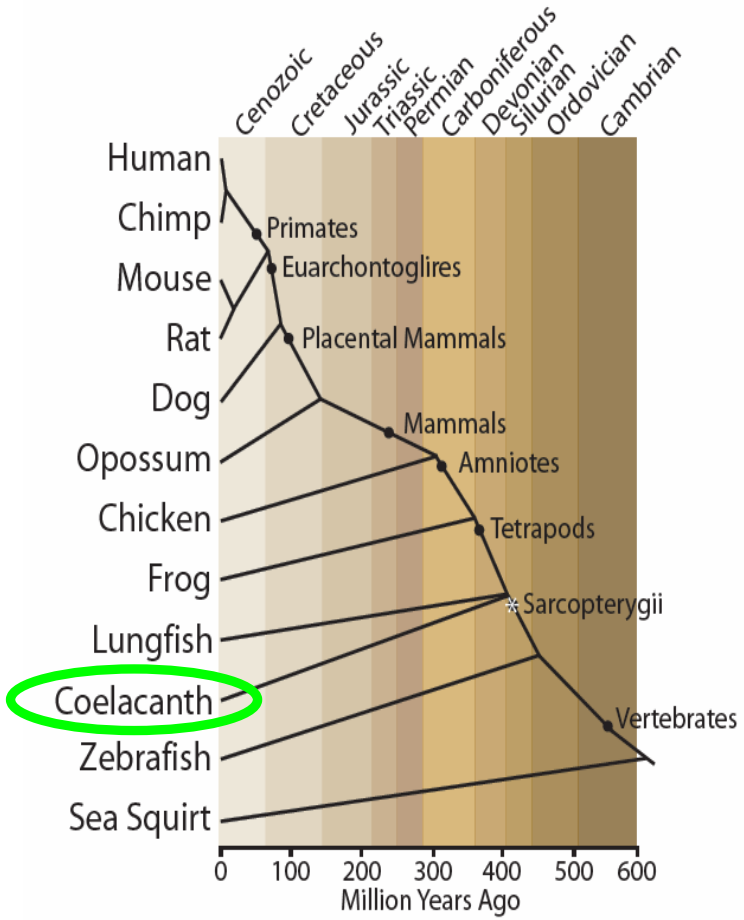


# Coelacanth Homologs Closer than Human Ones

```
Searching....10....20....30....40....50....60....70....80....90....100% done

                                                                    Smallest
                                                                      Sum
                                                             High   Probability
Sequences producing High-scoring Segment Pairs:              Score  P(N)       N

EM_HUM:AC023509 AC023509.42 Homo sapiens 12 BAC RP11-793H...  1315  3.6e-51    1
EM_MUS:AC137156 AC137156.3 Mus musculus BAC clone RP23-11...  1297  2.3e-50    1
EM_MUS:AF236844 AF236844.1 Mus musculus alpha-CP2
EM_OV:AC151571 AC151571.1 Latimeria menadoensis cl
EM_PAT:AR237520 AR237520.1 Sequence 52 from patent
EM_OV:AC150308 AC150308.1 Latimeria menadoensis cl
EM_HUM:AC006127 AC006127.1 Homo sapiens chromosome
EM_HUM:AF254822 AF254822.1 Homo sapiens SMARCA4 is
EM_OV:AC150309 AC150309.1 Latimeria menadoensis cl
EM_OV:AC150283 AC150283.1 Latimeria menadoensis cl
EM_OV:AC150284 AC150284.1 Latimeria menadoensis cl
EM_OV:AC147788 AC147788.1 Latimeria menadoensis cl
EM_HUM:AC005076 AC005076.2 Homo sapiens BAC clone
EM_OV:AF131253 AF131253.1 Latimeria chalumnae rhoc
EM_MUS:MMNUABPRO L19661.1 Mus musculus (clones pB1001, pB...   498  1.2e-14    1
EM_RO:BC078909 BC078909.1 Rattus norvegicus poly(rC) bind...   498  1.3e-14    1
```

[Bejerano, Lowe et al., *Nature*, 2006]

# Coelacanth “the Living Fossil” Fish




Appeared in Fossil Record >360Mya; Peaked 240Mya; Disappeared 80Mya; Rediscovered (by science) in 1938.

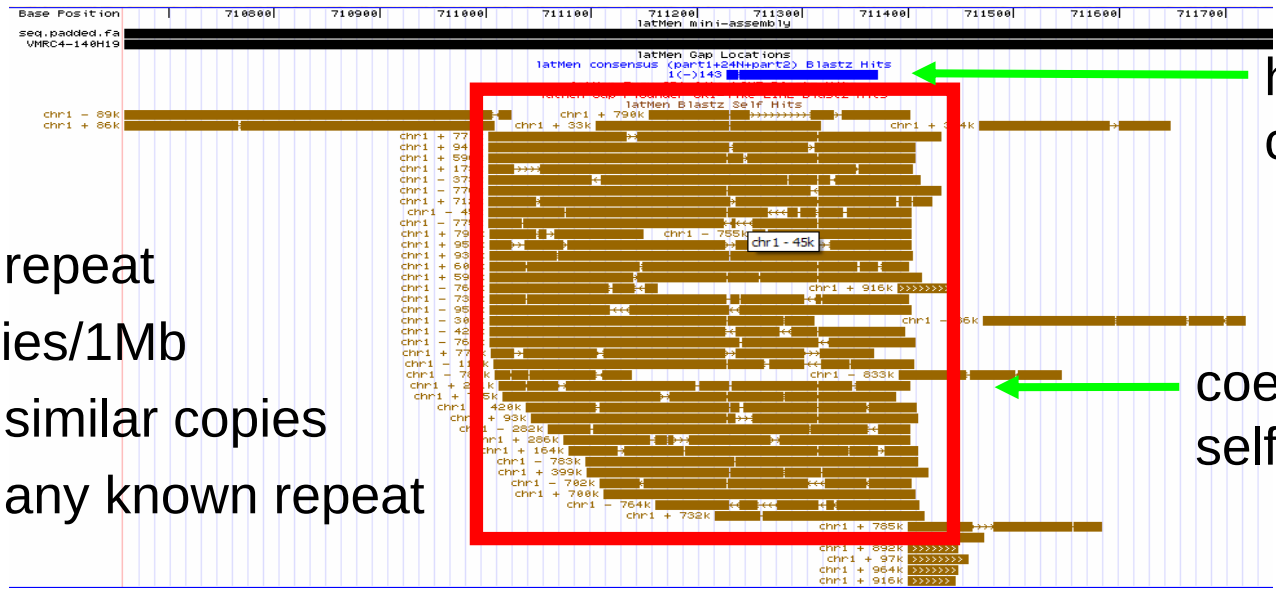
# Coelacanth Repeat

1 Mb sequenced.

4 genomic regions.

Instances in all.

| Region Of Homology | BACs | Non-overlapping Length (bp) | **uc.338 like** Instances |
|---|---|---|---|
| Protocadherin Cluster | AC150310<br>AC150283<br>AC150284<br>AC150309<br>AC150308 | 609,120 | 21 |
| Hox Cluster | AC151571 | 187,392 | 21 |
| Hox Cluster | AC147788 | 168,364 | 4 |
| Genomic | AC140159 | 92,794 | 13 |
| Total | 8 | 1,057,670 | 59 |



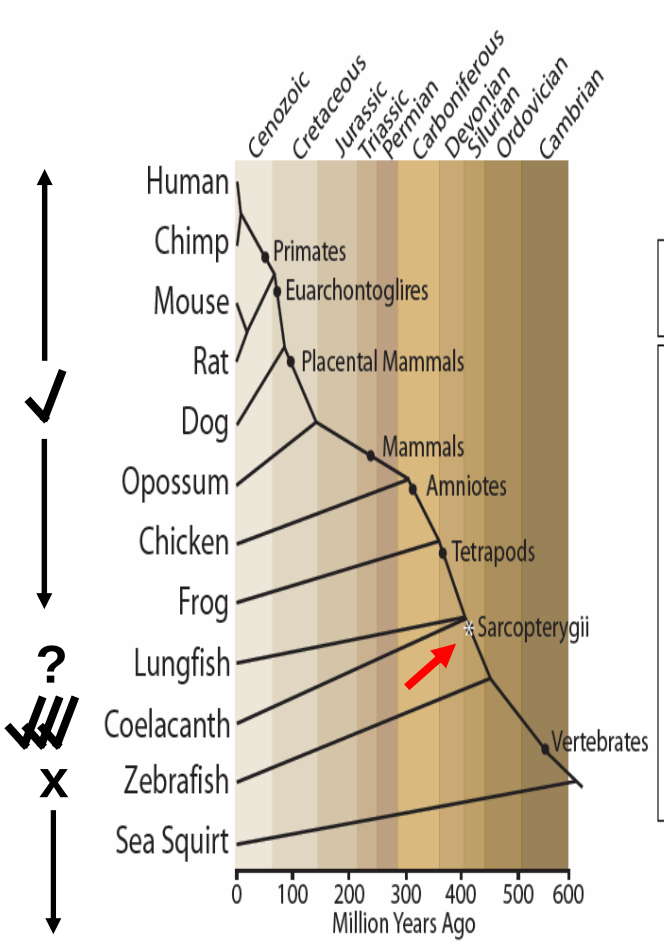

481bp repeat

59 copies/1Mb

Highly similar copies

Unlike any known repeat

# Coelacanth Repeat Distribution




Reconstruct ancestral coelacanth repeat.

Search all available genomes.

| Species | UCSC Assembly | **repeat** Detected | Species | UCSC Assembly | **repeat** Detected |
|---|---|---|---|---|---|
| *Homo sapiens* | hg17 | Yes | *Danio rerio* | danRer2 | No |
| *Pan troglodytes* | panTro1 | Yes | *Tetraodon nigroviridis* | tetNig1 | No |
| *Macaca mulatta* | rheMac1 | Yes | *Takifugu rubripes* | fr1 | No |
| *Mus musculus* | mm6 | Yes | *Ciona intestinalias* | ci1 | No |
| *Rattus norvegicus* | rn3 | Yes | *Strongylocentrotus purpuratus* | strPur1 | No |
| *Canis familiaris* | canFam1 | Yes | *Drosophila melanogaster* | dm2 | No |
| *Bos taurus* | bosTau1 | Yes | *Anopheles gambiae* | anoGam1 | No |
| *Monodelphis demestica* | monDom1 | Yes | *Caenorhabditis elegans* | ce2 | No |
| *Gallus gallus* | galGal2 | Yes | *Saccharomyces cerevisiae* | sacCer1 | No |
| *Xenopus tropicalis* | xenTro1 | Yes | | | |

Similar distribution in GenBank, Trace Archives.

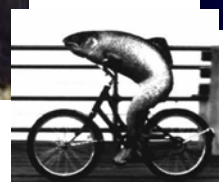
# Interim Summary


480bp recently active repeat

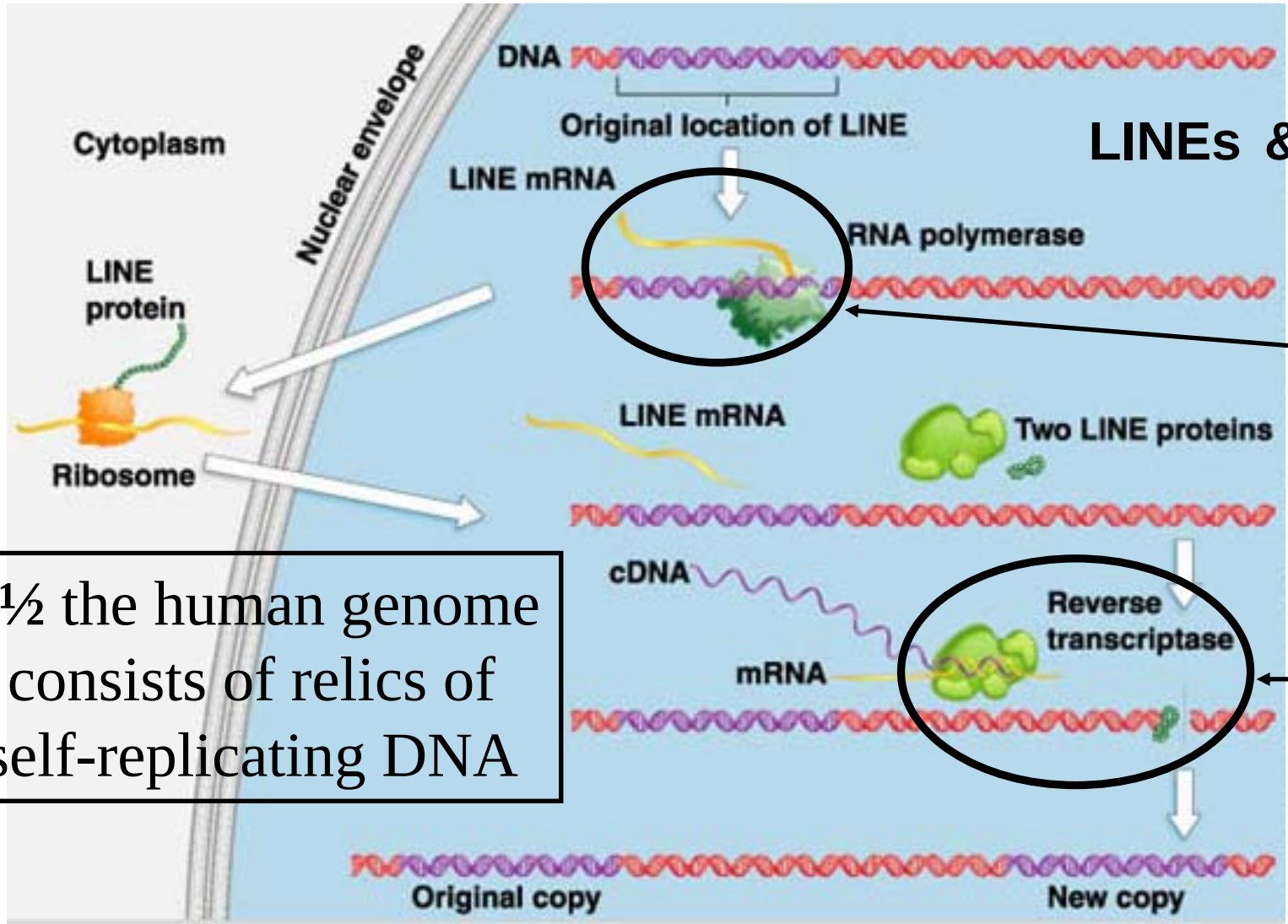
80%id to 360bp incl. uc.338

$10^5$ copies/genome

Unlike any known repeat



# Repeats ***are*** a Major Force in Vertebrate Evolution




**LINEs & SINEs**

**out**

**back**

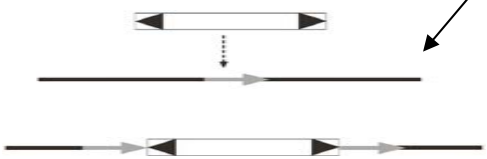
>½ the human genome consists of relics of self-replicating DNA

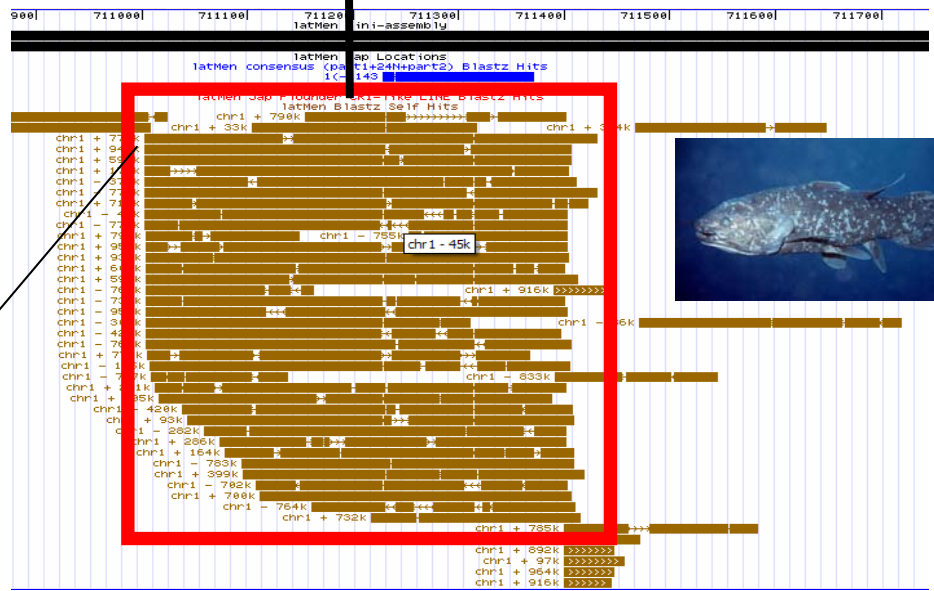
# The LF SINE (for Lobefin Fish / “Living Fossil”)


Reconstruction

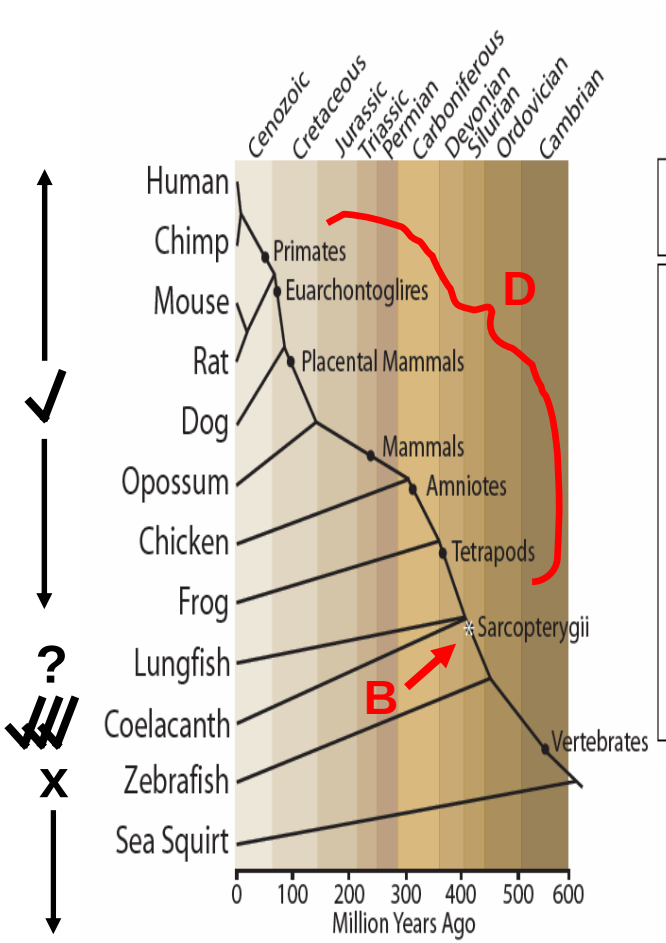
```
                100         110         120         130         140
LF-SINE         GAAAGTCA-TTACCATCTGATGGCTGTTCAGTGGCCTA---TGTGAAATGAG
                : ::::::  ::: : :::    ::   :::::  :::    ::: :::::  :
Core Domain     GCAAGTCACTTAACCTCTCTGAGCC--TCAGTTTCCTCATCTGTAAAATGGG
                  100         110         120         130         140

                       90          100           110         120
V-SINE(shark)   CTGTGTGGAGTTTGCACATTCTC----CCCGTGTCTGCGTGGGTTTCCTC
                :::::::    :: :   : :: :    ::   : :  :: :  ::::::::
Core Domain     CTGTGTGACCTTGGGCAAGTCACTTAACCTCTCTGAGCCTCAGTTTCCTC
                       90          100         110         120        130
```

target site duplications





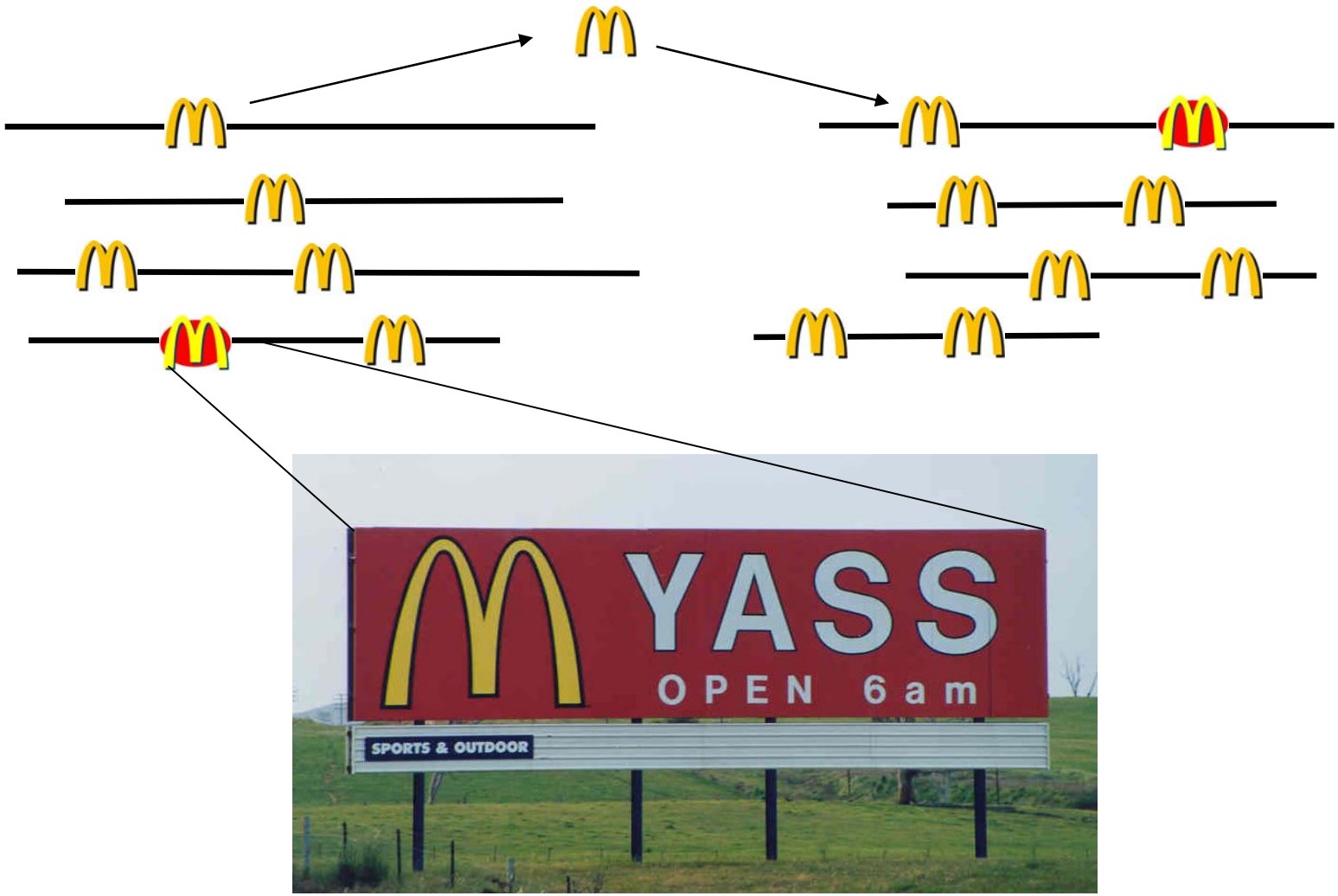
# >360My Old and Going Strong




Upto 80%id between Coelacanth SINE and some human instances, inc uc.338.

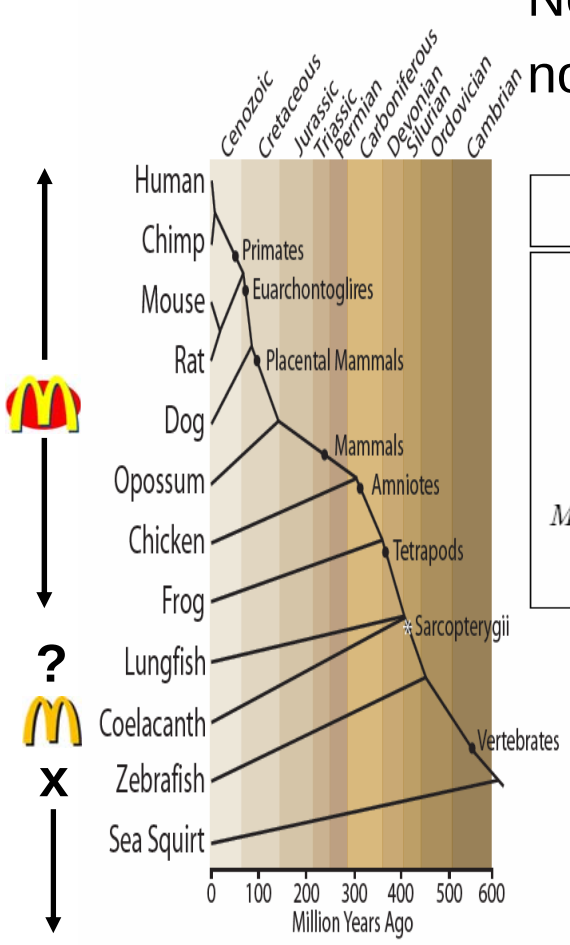
| Species | UCSC Assembly | LF-SINE Detected | Species | UCSC Assembly | LF-SINE Detected |
|---|---|---|---|---|---|
| *Homo sapiens* | hg17 | Yes | *Danio rerio* | danRer2 | No |
| *Pan troglodytes* | panTro1 | Yes | *Tetraodon nigroviridis* | tetNig1 | No |
| *Macaca mulatta* | rheMac1 | Yes | *Takifugu rubripes* | fr1 | No |
| *Mus musculus* | mm6 | Yes | *Ciona intestinalias* | ci1 | No |
| *Rattus norvegicus* | rn3 | Yes | *Strongylocentrotus purpuratus* | strPur1 | No |
| *Canis familiaris* | canFam1 | Yes | *Drosophila melanogaster* | dm2 | No |
| *Bos taurus* | bosTau1 | Yes | *Anopheles gambiae* | anoGam1 | No |
| *Monodelphis demestica* | monDom1 | Yes | *Caenorhabditis elegans* | ce2 | No |
| *Gallus gallus* | galGal2 | Yes | *Saccharomyces cerevisiae* | sacCer1 | No |
| *Xenopus tropicalis* | xenTro1 | Yes | | | |

# Co-Option (Exaptation) of Mobile Elements

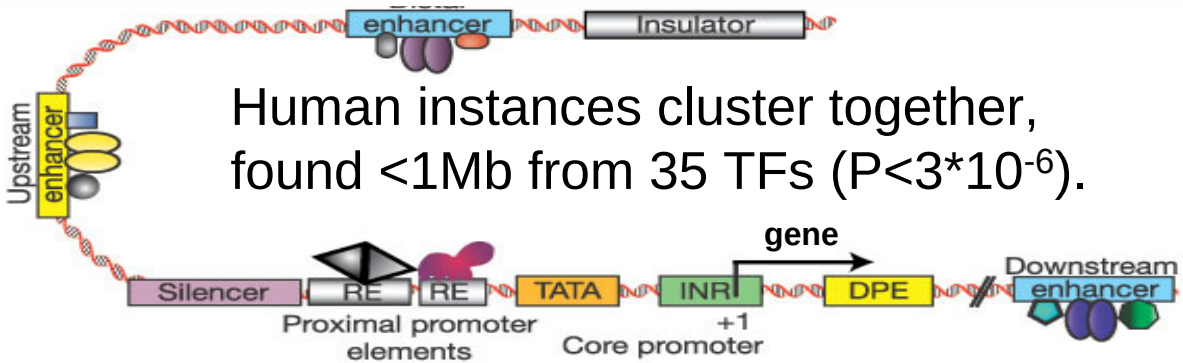


# Exapted Into Which Cellular Roles?

No evidence for Transcription (Tx) as small RNAs,
no orientation preference in introns, not in antisense Tx.



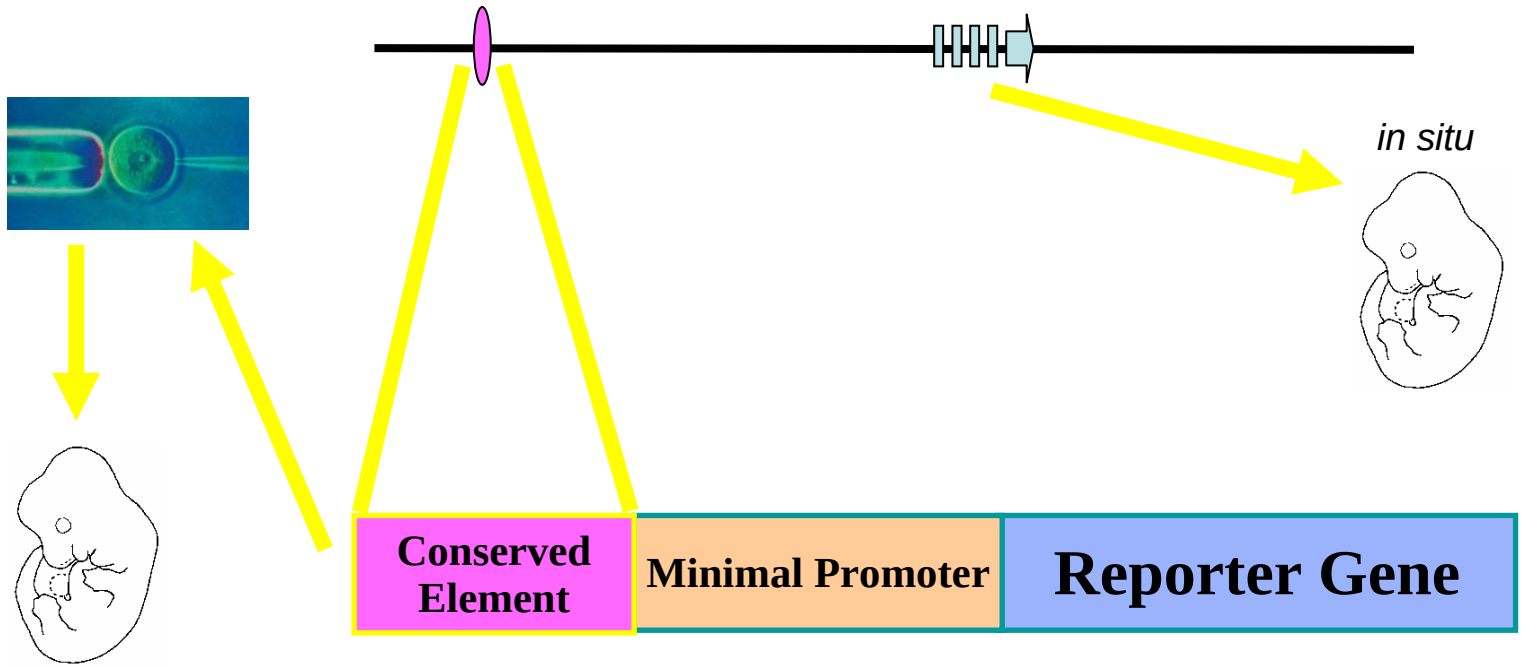

| Organism | 5' UTR | 3' UTR | Exonic | | Intronic | Intergenic | Total |
|---|---|---|---|---|---|---|---|
| | | | Alt-Spliced | Total | | | |
| *Homo sapiens* | 1 | 0 | 12 | 13 | 68 | 163 | 245 |
| *Pan troglodytes* | - | - | - | - | - | - | 210 |
| *Macaca mulatta* | - | - | - | - | - | - | 229 |
| *Canis familiaris* | - | - | - | - | - | - | 235 |
| *Bos taurus* | - | - | - | - | - | - | 169 |
| *Mus musculus* | 0 | 1 | 7 | 8 | 25 | 57 | 91 |
| *Rattus norvegicus* | - | - | - | - | - | - | 87 |
| *Monodelphis domestica* | - | - | - | - | - | - | 394 |
| *Gallus gallus* | 0 | 1 | 2 | 3 | 244 | 451 | 699 |
| *Xenopus tropicalis* | 0 | 0 | 1 | 2 | 10 | 14 | 26 |




Human instances cluster together,
found <1Mb from 35 TFs ($P<3*10^{-6}$).

# Transient Transgenic Enhancer Assay

**Eddy Rubin's Lab, LBNL**

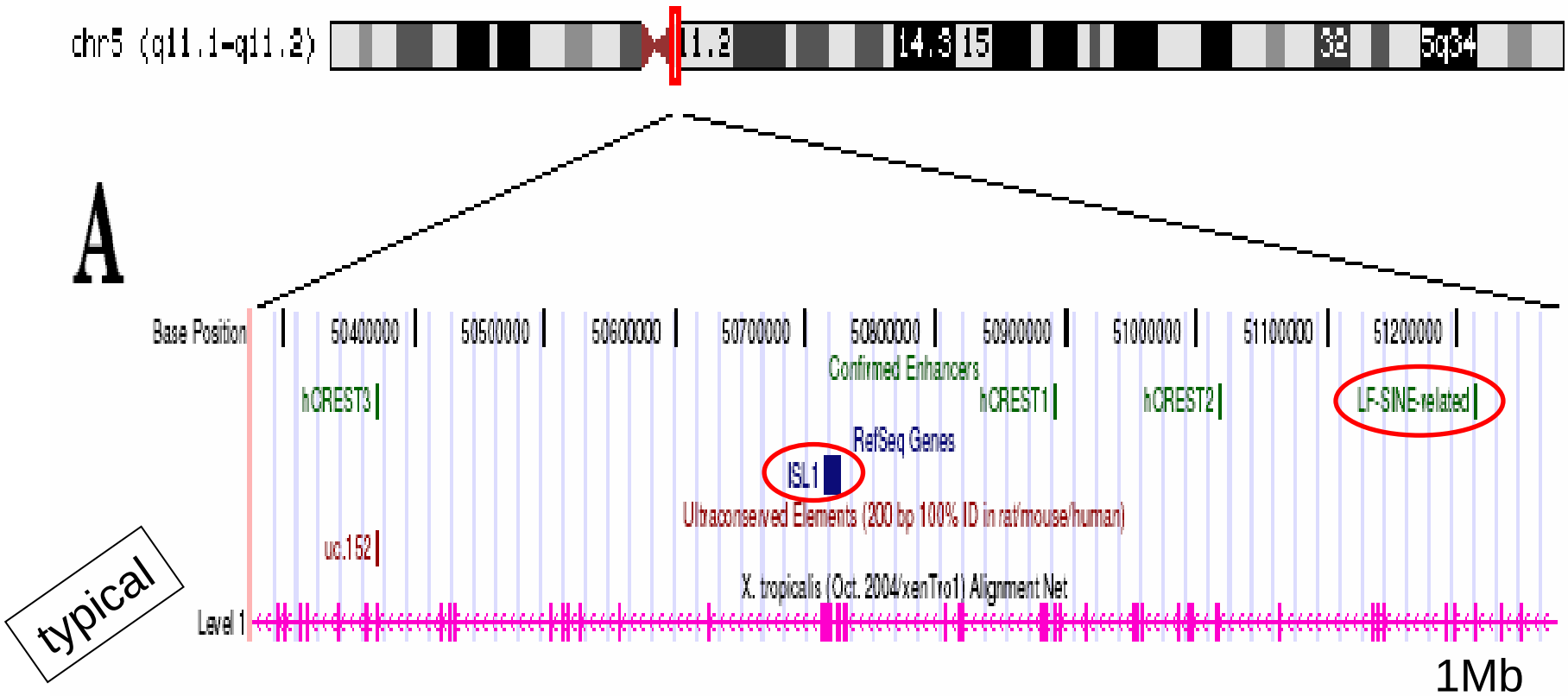



Construct is injected into 1 cell embryos

Taken out at embryonic day 10.5-14.5

Assayed for reporter gene activity

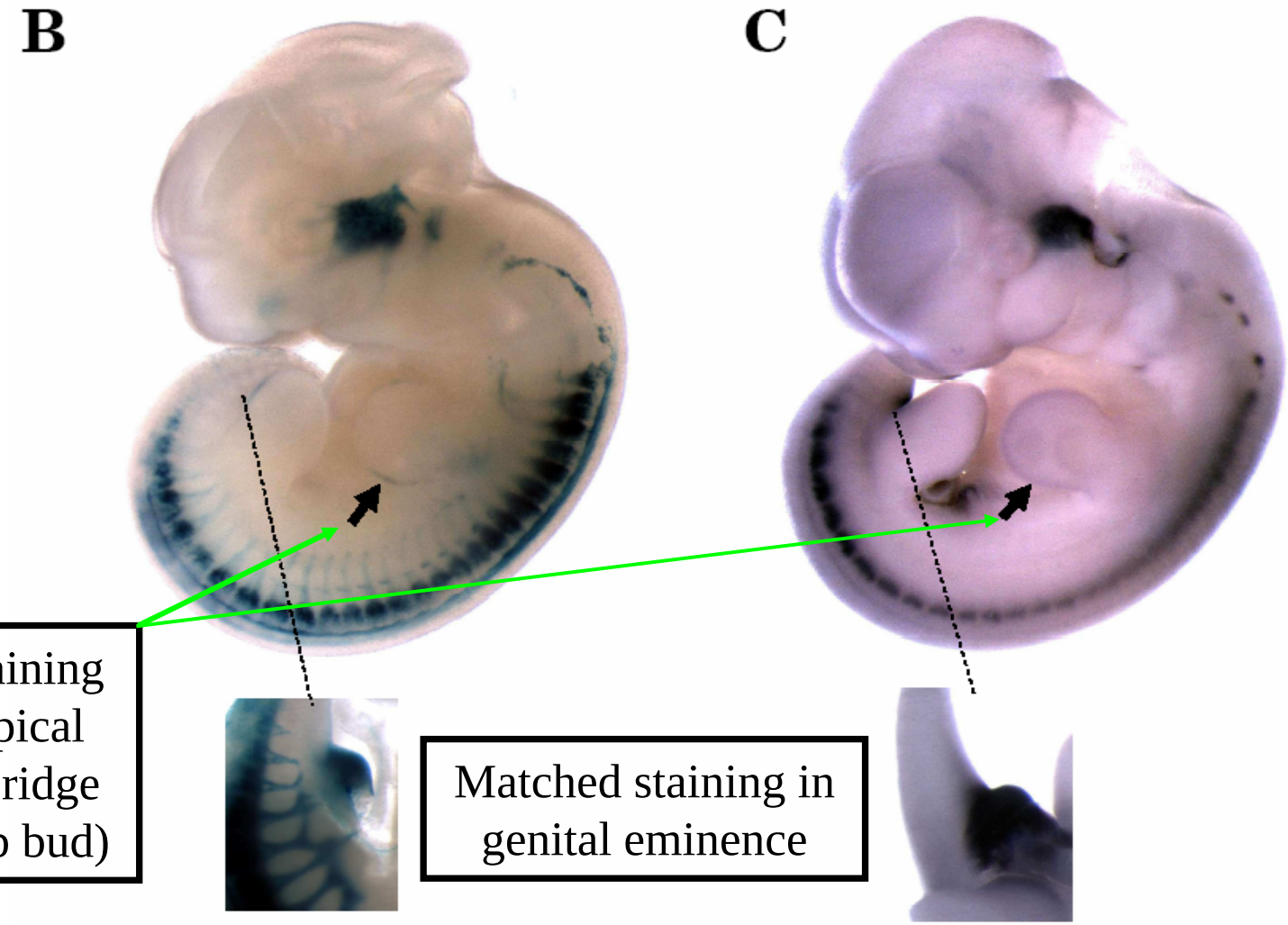
# Instance 500kb Downstream of ISL1




ISL1 is a neuro-developmental gene, also expressed in testis.

Three previously known enhancers are conserved in all vertebrates.

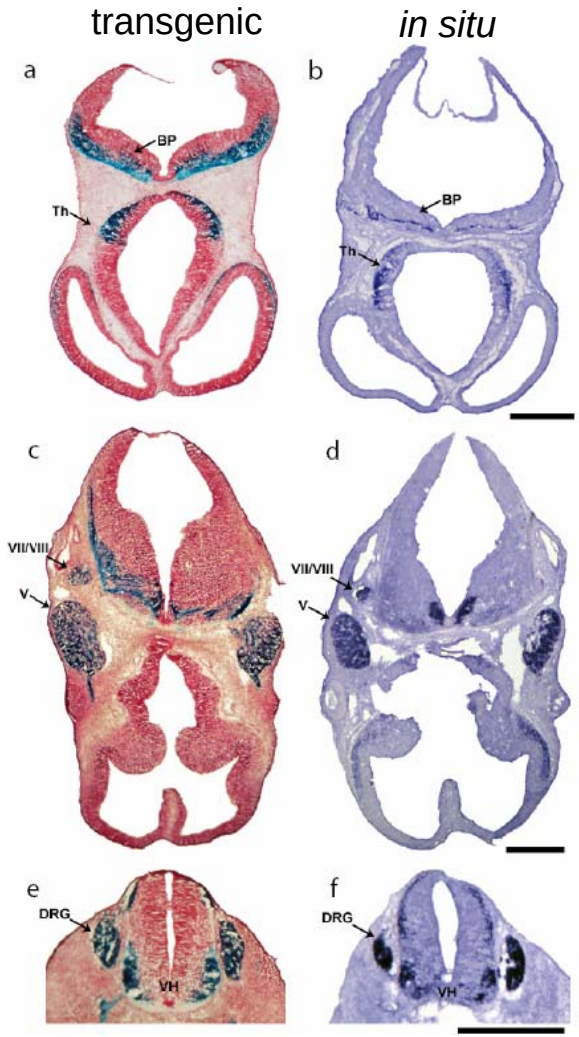
# LF SINE Transgenic (B) vs. Mouse Isl1 *in situ* (C)




Nadav Ahituv, Eddy Rubin

# Matched Level Sections

Corresponding expression patterns in:
(a, b) the developing thalamus (Th)
and basal plate (BP) in the brain.
(c, d) the trigeminal (V) ganglion and
facio-acoustic (VII/VIII) ganglia
in the head region.
(e, f) the dorsal root ganglion (DRG),
and the lateral region of the
ventral horn (VH) of the spinal cord
in thoracic sections.



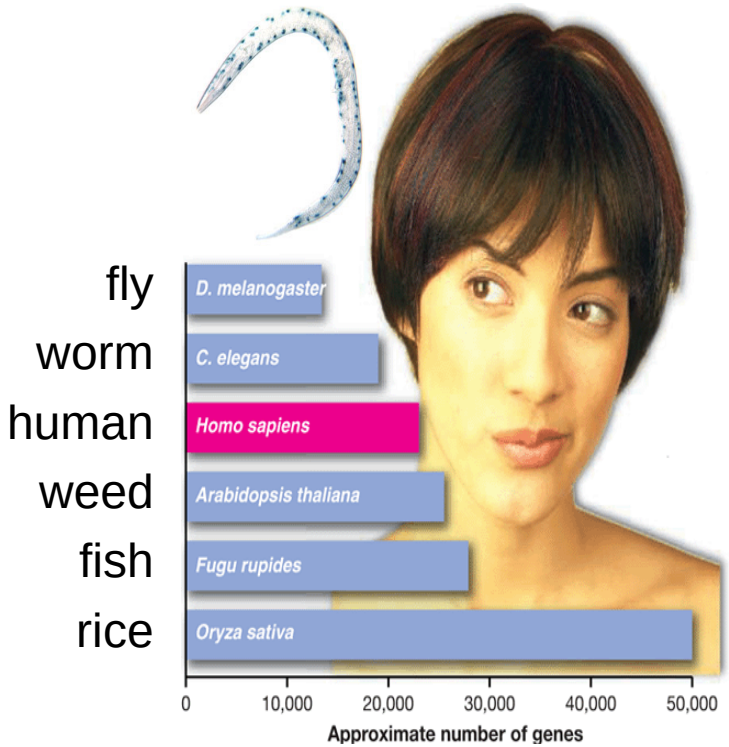

Bryan King, Sofie Salama, Nadav Ahituv, Eddy Rubin

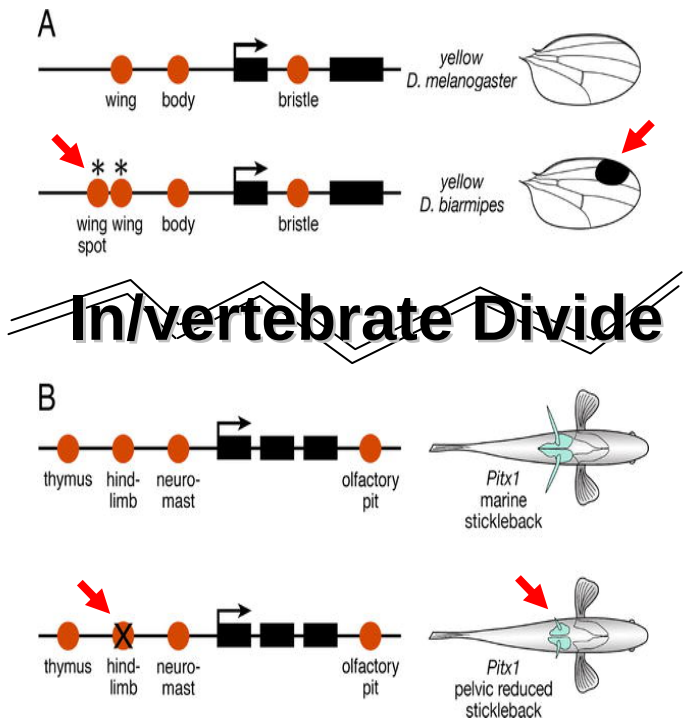
# Mobile Elements Give Birth to Distal Cis-Regulatory Elements: **Implications**

Gene numbers do not correlate with organism complexity.

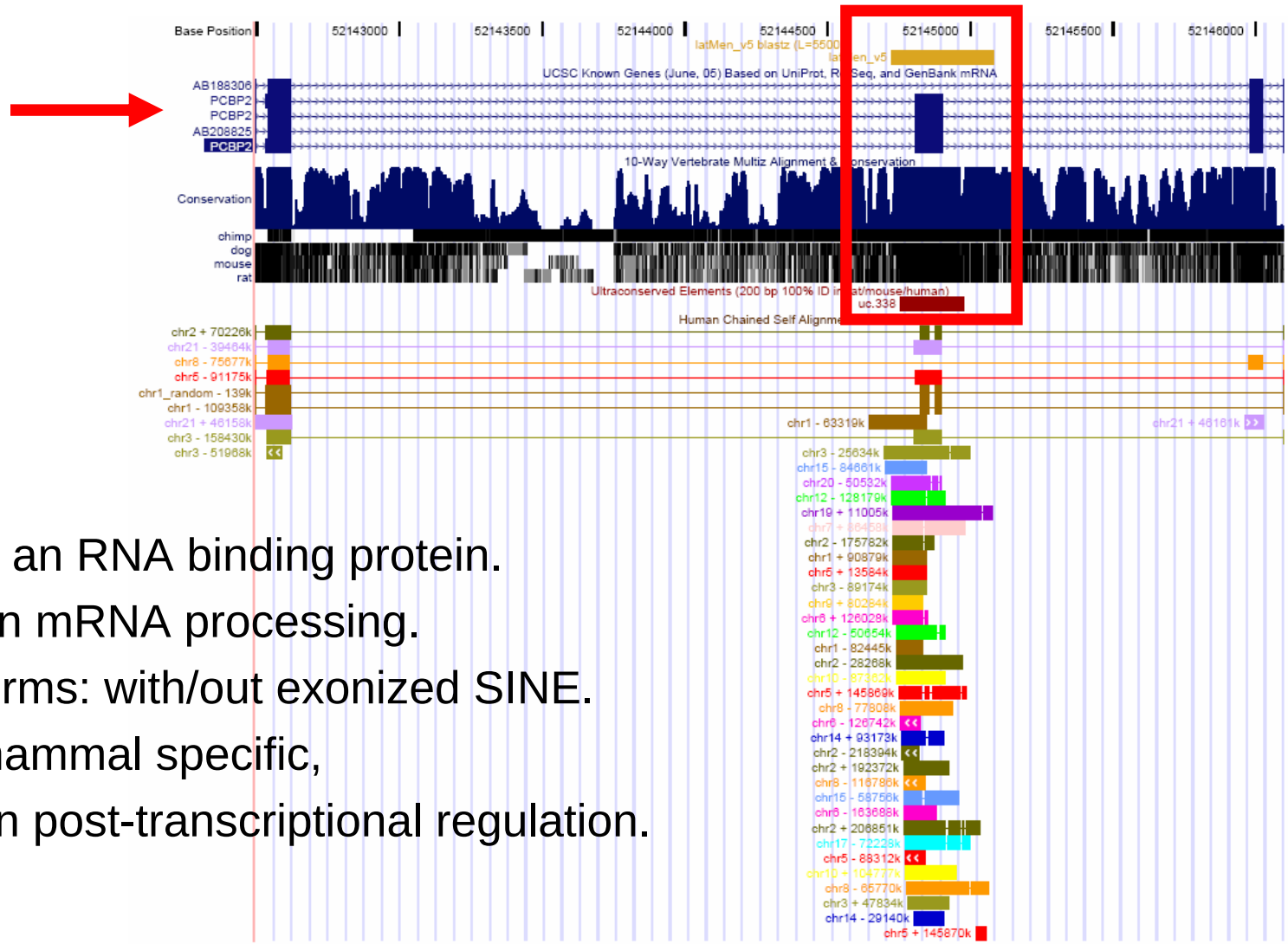
Many gene families are surprisingly old.

"Regulatory sequence evolution must be the major contribution to the evolution of form." [Sean Carroll, *PLoS Bio* 2005]



# uc.338 Has Undergone Exonization




PCBP2 is an RNA binding protein.

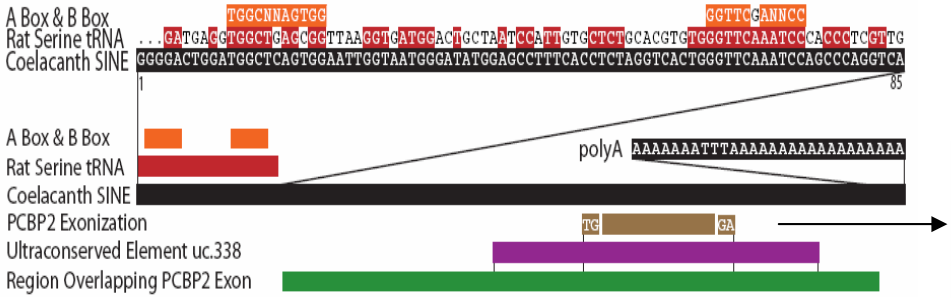
Involved in mRNA processing.

Two Isoforms: with/out exonized SINE.

Exon is mammal specific,

involved in post-transcriptional regulation.

# Exonized LF SINE Instances




19 different tetrapod exonizations

| Gene Name | Sp. | Exon No. | Alt-Spl. | CDS/UTR | Ins. F.S. | Ins. Stop | Trig. NMD | 3' S.S. ex5' | 3' S.S. Seq. | 5' S.S. ex3' | 5' S.S. Seq. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| PCBP2 | H | 9 | + | C | - | - | - | 381 | cacagGACAG | 289 | TAGAGgtgag |
| | M | | + | C | - | - | - | 381 | cacagGACAG | 289 | TAGAGgtgag |
| SMARCA4 | H | 27 | + | C | - | - | - | 381 | cacagAGCAG | 289 | TAAAGgtgag |
| | M | | + | C | - | - | - | 381 | cacagAGCAG | 289 | TAAAGgtgag |
| EEF1B2 | H | 3 | + | C | - | + | +[a] | 376 | aacagGTAGA | 101 | TCATAgtgag |
| TCERG1 | H | 22 | + | C | + | + | +[a] | 469 | tatagTTAAT | 377 | AACAGgtaca |
| | M | | + | C | + | + | +[a] | 469 | tatagTTAAT | 377 | AACAGgtaca |
| PTDSR | H | 5 | + | C | + | + | + | 376 | aacagATACA | 289 | TCGGGgtaag |
| | M | | + | C | + | + | + | 376 | aacagATACA | 289 | TCGGGgtaag |
| RORA | H | 3 | + | C | + | + | + | 366 | cgcagGGCAG | 289 | AGGTGgtaag |
| | M | | + | C | + | + | + | 366 | cgcagGGCAG | 289 | AGGTGgtaag |
| GRID1 | H | 1 | + | 5U | | | | | | 289 | GTAGGgtaag |
| ATF2 | H | 14 | + | C | + | + | +[a] | 450 | taaagTGAAT | 377 | ACCAGgtaca |
| | M | | + | 3U[b] | | | | | | | |
| FLJ22833 | H | 4 | + | C | + | + | + | 381 | cacagTCTAG | 280 | GTAAGgtaat |
| ARHGAP6 | H | 13 | + | C | - | + | + | 381 | cccagAACAA | 289 | TAAGGgtgag |
| | M | | + | C | - | + | + | 381 | cacagAACCA | 289 | TAAGGgtgag |
| KIAA1409 | H | 34 | - | C | - | - | - | 381 | cacagAACAG | 281 | GAGAGgttag |
| | M | | - | C | - | - | - | 381 | cacagAACAG | 281 | GAGAGgttag |
| | C | | - | C | - | - | - | 381 | cacagAGCAG | 281 | CAGAGgttag |
| NT5C2 | H | 9 | + | C | + | + | + | 450 | taaagTGAAT | 107 | GGATGgtaat |
| | M | | + | C | + | + | + | 450 | taaagTGAAT | 240 | GGGAGgtttg |
| LRP1B | H | 90 | + | C | - | - | - | 381 | tacagGCCAG | 289 | CTGGGgtgag |
| DHX30 | H | 4 | + | C | - | - | - | 381 | cccagATCGG | 289 | TCGAGgtaag |
| gg-DMTF1 | C | 12 | + | C | + | + | + | 450 | tttagTGAAT | 377 | AACAGgtaca |
| gg-PPP2R2C | C | 2 | + | C | + | + | + | 381 | cacagGAGAG | 289 | TGGAGgtgag |
| gg-SHFM1 | C | 3 | + | 3U[b] | | | | | | | |
| xt-MBNL1 | F | 4 | + | C | - | - | - | 381 | cacagGCCAG | 289 | TATGGgtgag |
| JGI-49280 | F | 5 | + | C | + | + | +[a] | 450 | tctagTGATT | 377 | AACAGgtaaa |

The 19 proteins are unrelated

19/19 exon antisense to SINE

17/19 novel coding exon

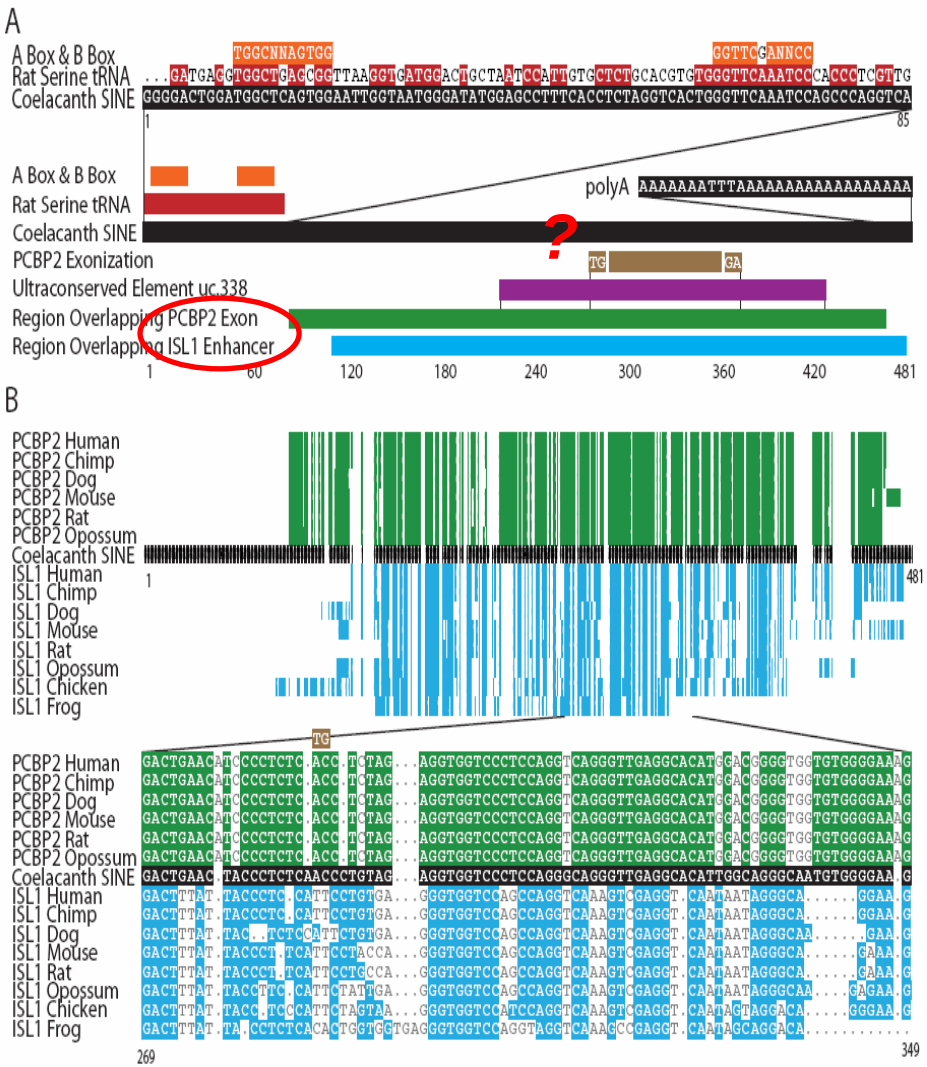
16/17 alternatively-spliced

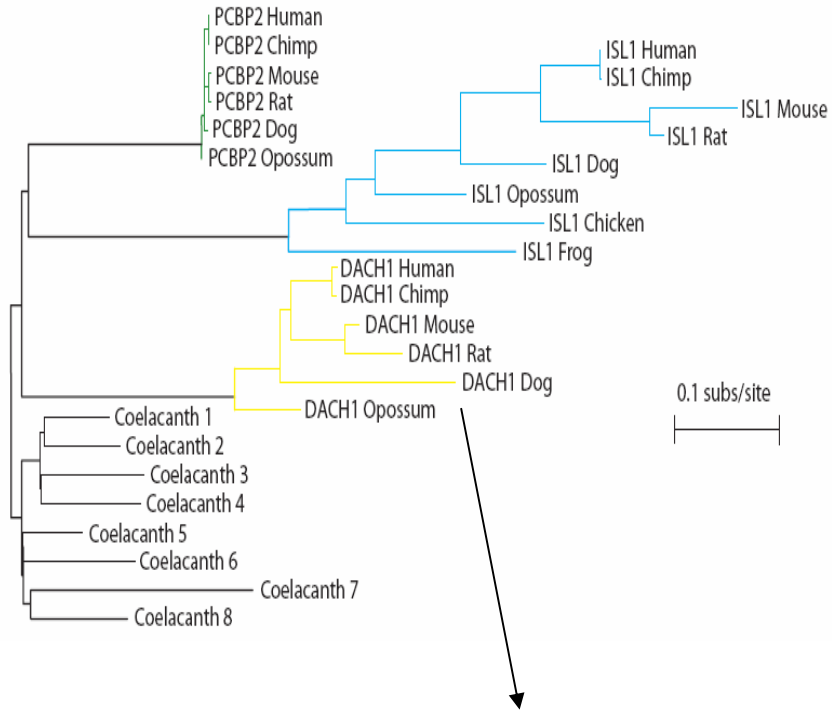
11/17 introduce early stop codon*

post-transcriptional regulation

* The 6 read-thru exons all match a 93bp internal LF-SINE region, (still!) free of stop codons in all 3 frames (p~0.002)

# Unexpected Kin: Enhancer, Exon, Repeat




one of 29 human instances spanning coelacanth SINE

# Some Challenges

- What molecular mechanisms lead to ultra conservation?
- Which ultraconserved elements (ultras) were seeded by transposons? When? Origins of the others?
- How do ultras evolve: all by purifying selection, or is there some reduced mutation or hyper-repair?
- Why the strong association with DNA binding and RNA binding genes? Transcription/splicing enhancers?
- What information do ultras encode and how?
- Are ultras part of delicate self-regulating gene networks that are critical in development?
- Are they related to human disease?
- Are vertebrate and fly ultras related?
- ...

# Kudos


UC Santa Cruz

David Haussler

Sofie Salama, Craig Lowe, Bryan King

Jim Kent, Adam Siepel, Jakob Pedersen

Katie Pollard, Courtney Onodera

Rachel Harte, Genomics/Browser Group

Lawrence Berkeley Labs

Eddy Rubin

Nadav Ahituv, Marcelo Nobrega

Northwestern U.

Jack Kessler, Alex Bassuk

McGill U.

Mathieu Blanchette

Penn State U.

Webb Miller's group

U. Queensland

John Mattick's group

Genome Sequencing Consortia

*All* GenBank contributors