6.047/6.878/HST.507
Computational Biology: Genomes, Networks, Evolution

Lecture 17

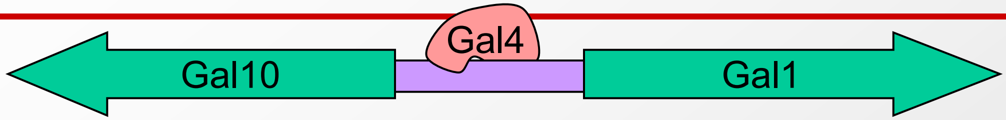
Comparative genomics I:

Genome annotation using evolutionary signatures

Module V: Comparative genomics and evolution

- Today: Whole-genome comparative genomics
  - Evolutionary signatures for systematic genome annotation
- Next week: Phylogenetics and Phylogenomics
  - Distance-based and model-based phylogenetics approaches
  - Gene trees and species trees, reconciliation, coalescence
- Computational foundations:
  - Evolutionary rates and models of evolution
  - Dynamic programming on two-dimensional tree structures
  - Synteny-based alignment, genome assembly

Key goal: Evolution preserves functional elements



```
Scer   TTATATTGAATTTTCAAAAATTCTTACTTTTTTTTTGGATGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATATACA
Spar   CTATGTTGATCTTTTCAGAATTTTT-CACTATATTAAGATGGGTGCAAAGAAGTGTGATTATTATATTACATCGCTTTCCTATCATACACA
Smik   GTATATTGAATTTTTCAGTTTTTTTTCACTATCTTCAAGGTTATGTAAAAAA-TGTCAAGATAATATTACATTTCGTTACTATCATACACA
Sbay   TTTTTTTGATTTCTTTAGTTTTCTTTCTTTAACTTCAAAATTATAAAAGAAAGTGTAGTCACATCATGCTATCT-GTCACTATCACATATA
        * * ****  * *  *   ** ** *  *   **           ** ** * *    *    **   **    *  * * ** * * *

                          TBP
Scer   TATCCATATCTAATCTTACTTATATGTTGT-GGAAAT-GTAAAGAGCCCCATTATCTTAGCCTAAAAAAACC--TTCTCTTTGGAACTTTCAGTAATACG
Spar   TATCCATATCTAGTCTTACTTATATGTTGT-GAGAGT-GTTGATAACCCCAGTATCTTAACCCAAGAAAGCC--TT-TCTATGAAACTTGAACTG-TACG
Smik   TACCGATGTCTAGTCTTACTTATATGTTAC-GGGAATTGTTGGTAATCCCAGTCTCCCAGATCAAAAAAGGT--CTTTCTATGGAGCTTTG-CTA-TATG
Sbay   TAGATATTTCTGATCTTTCTTATATATTATAGAGAGATGCCAATAAACGTGCTACCTCGAACAAAAGAAGGGGATTTTCTGTAGGGCTTTCCCTATTTTG
       **   ** ***  **** ******* **   *  *   *     *  *    *  *       **  **      * *** *    ***    *  *  *

                                        GAL4               GAL4              GAL4
Scer   CTTAACTGCTCATTGC-----TATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGACTCTCCTCCGTGCGTCCTCGTCT
Spar   CTAAACTGCTCATTGC-----AATATTGAAGTACGGATCAGAAGCCGCCGAGCGGACGACAGCCCTCCGACGGAATATTCCCCTCCGTGCGTCGCCGTCT
Smik   TTTAGCTGTTCAAG--------ATATTGAAATACGGATGAGAAGCCGCCGAACGGACGACAATTCCCCGACGGAACATTCTCCTCCGCGCGGCGTCCTCT
Sbay   TCTTATTGTCCATTACTTCGCAATGTTGAAATACGGATCAGAAGCTGCCGACCGGATGACAGTACTCCGGCGGAAAACTGTCCTCCGTGCGAAGTCGTCT
             **  **          ** ***** ******* ****** ***** ***  ****   * *** ***** * *  ****** ***    * ***

                                        GAL4
Scer   TCACCGG-TCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTACAA-----TACTAGCTTTT--ATGGTTATGAA
Spar   TCGTCGGGTTGTGTCCCTTAA-CATCGATGTACCTCGCGCCGCCCTGCTCCGAACAATAAGGATTCTACAAGAAA-TACTTGTTTTTTTATGGTTATGAC
Smik   ACGTTGG-TCGCGTCCCTGAA-CATAGGTACGGCTCGCACCACCGTGGTCCGAACTATAATACTGGCATAAAGAGGTACTAATTTCT--ACGGTGATGCC
Sbay   GTG-CGGATCACGTCCCTGAT-TACTGAAGCGTCTCGCCCCGCCATACCCCGAACAATGCAAATGCAAGAACAAA-TGCCTGTAGTG--GCAGTTATGGT
            ** *   ** *** *      *      ***** ** *  *   ****** **     *   * **     * *             ** ***

                                    MIG1
Scer   GAGGA-AAAATTGGCAGTAA----CCTGGCCCCACAAACCTT-CAAATTAACGAATCAAATTAACAACCATA-GGATGATAATGCGA------TTAG--T
Spar   AGGAACAAAATAAGCAGCCC----ACTGACCCCATATACCTTTCAAACTATTGAATCAAATTGGCCAGCATA-TGGTAATAGTACAG------TTAG--G
Smik   CAACGCAAAATAAACAGTCC----CCCGGCCCCACATACCTT-CAAATCGATGCGTAAAACTGGCTAGCATA-GAATTTTGGTAGCAA-AATATTAG--G
Sbay   GAACGTGAAATGACAATTCCTTGCCCCT-CCCCAATATACTTTGTTCCGTGTACAGCACACTGGATAGAACAATGATGGGGTTGCGGTCAAGCCTACTCG
              ****    *         *   *****     ***              * * *    *  * *    *     *           **

                MIG1                                                      TBP
Scer   TTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCG--ATGATTTTT-GATCTATTAACAGATATATAAATGGAAAAGCTGCATAACCAC-----TT
Spar   GTTTT--TCTTATTCCTGAGACAATTCATCCGCAAAAAATAATGGTTTTT-GGTCTATTAGCAAACATATAAATGCAAAAGTTGCATAGCCAC-----TT
Smik   TTCTCA--CCTTTCTCTGTGATAATTCATCACCGAAATG--ATGGTTTA--GGACTATTAGCAAACATATAAATGCAAAAGTCGCAGAGATCA-----AT
Sbay   TTTTCCGTTTTACTTCTGTAGTGGCTCAT--GCAGAAAGTAATGGTTTTCTGTTCCTTTTGCAAACATATAAATATGAAAGTAAGATCGCCTCAATTGTA
        * *      *    ***       * **   *  *     *** ***   *  *  **  ** * ********   ****    *

Scer   TAACTAATACTTTCAACATTTTCAGT--TTGTATTACTT-CTTATTCAAAT----GTCATAAAAGTATCAACA-AAAAATTGTTAATATACCTCTATACT
Spar   TAAATAC-ATTTGCTCCTCCAAGATT--TTTAATTTCGT-TTTGTTTTATT----GTCATGGAAATATTAACA-ACAAGTAGTTAATATACATCTATACT
```

We can ‘read’ evolution to reveal functional elements

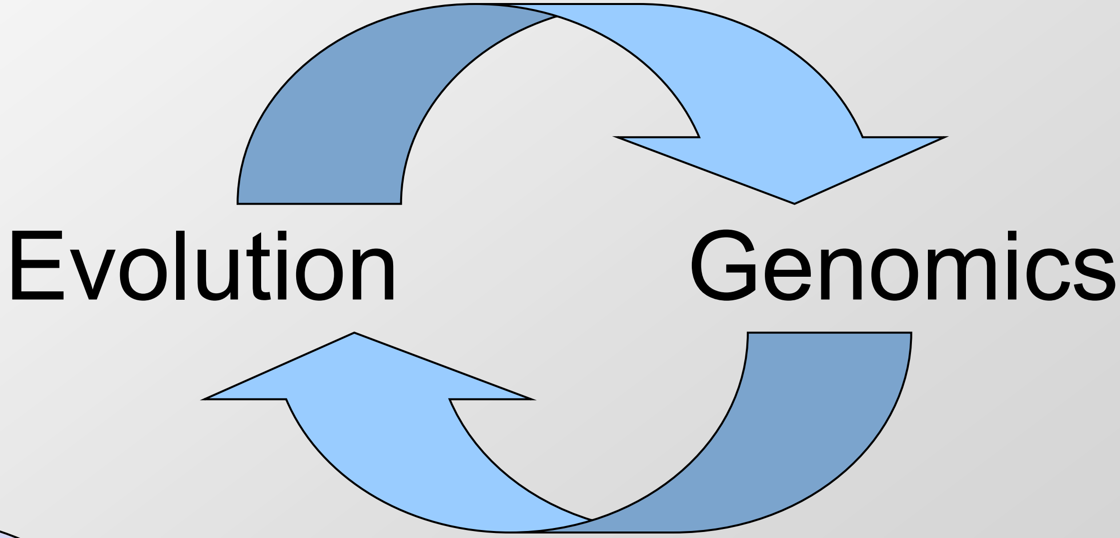
Conservation island

Yeast (Kellis et al, Nature 2003), Mammals (Xie, Nature 2005), Fly (Stark et al, Nature 07)

Comparative Genomics

Lecture 17 (Today):

Using evolution to study genomes



Evolution

Genomics

Lectures 18-19 (Thursday):

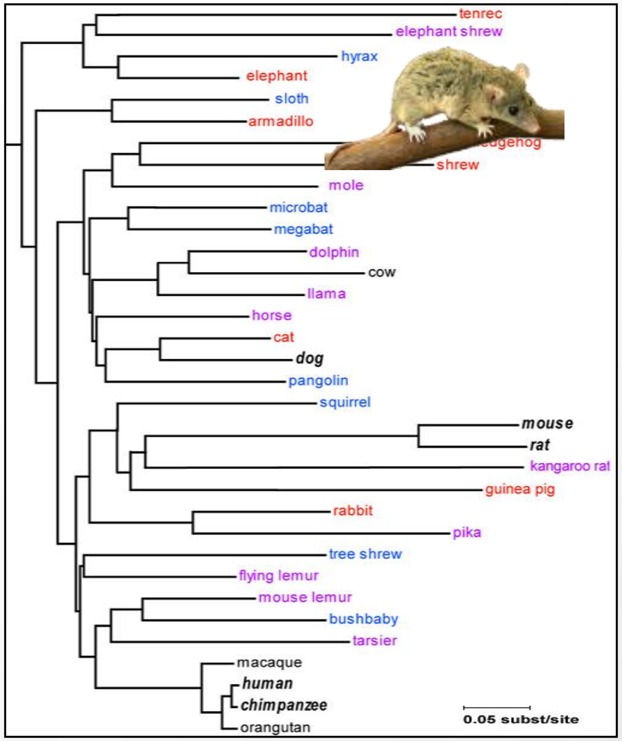
Using genomics to study evolution

Comparative genomics I: Evolutionary signatures

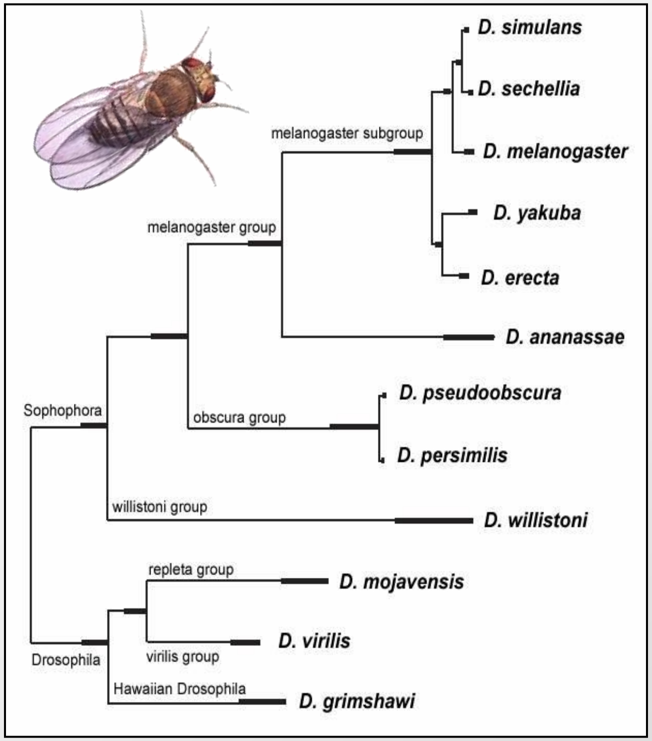
- **Nucleotide conservation: evolutionary constraint**
  - – Purifying selection, neutral branch length, discovery power
  - – Detect constrained elements: nucleotides, windows, HMM
  - – Estimate fraction constrained: signal vs. background
- **Evolutionary signatures: focus on pattern of change**
  - – Different functions ⇔ Characteristic patterns of evolution
- **Signatures of protein-coding genes**
  - – Reading-frame conservation, codon-substitution frequency
  - – Likelihood ratio framework: Estimating $Q_C Q_N$, scoring
  - – Revise genes, read-through, excess constraint regions
- **Signatures of microRNA genes**
  - – Structural and evolutionary features of microRNAs
  - – Combining features: decision trees, random forests
  - – Sense/anti-sense miRNAs, mature/star arm cooperation
- **Measuring selection within the human lineage**

Comparative genomics for genome annotation

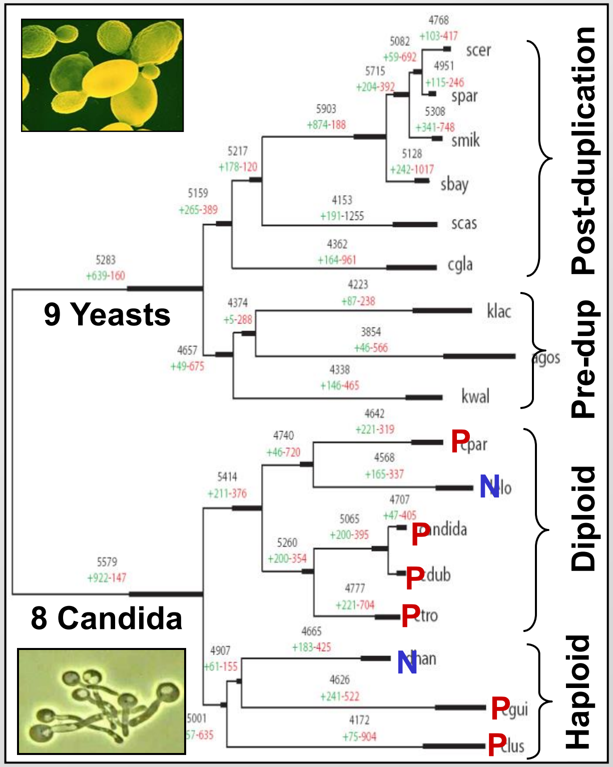
29 mammals



12 flies

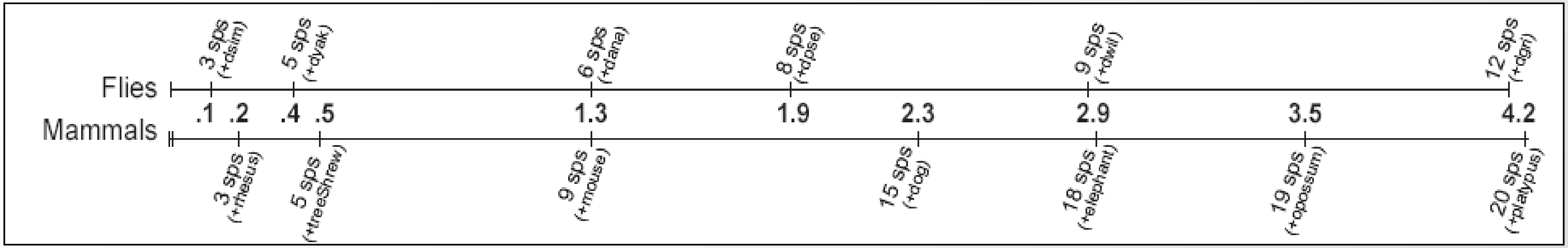


17 fungi



- **Compare related species to discover functional elmts**
- **Evolution process: random mutation, natural selection**
  - – Non-functional regions: accumulate mutations, kept
  - – Functional regions: accumulate mutations, decrease fitness
  - – Evolutionary time: less fit organisms & their genes thin out

Power of many closely related: total branch length

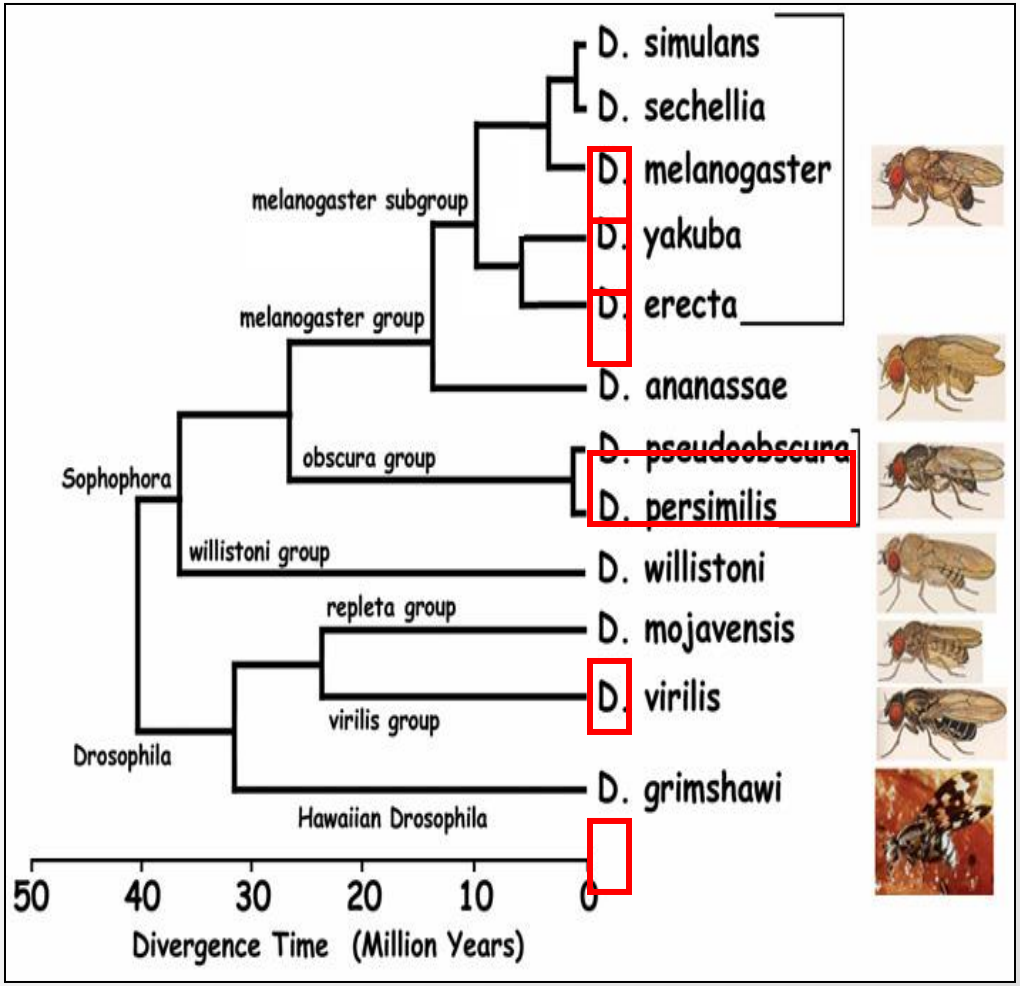


- **More branch length ➔ more events ➔ more power**
  - – Goal: functional vs. non-functional based on # of mutations
  - – Very close distances: no mutations in either region
  - – Sufficient distance: ability to distinguish increases
  - – Very far distances: functional regions no longer conserved
- **Many closely related species >> few distantly related**
  - – For same total branch length: prefer many close species
  - – Functional regions conserved for each pair of species
  - – Non-functional regions accumulate noise **independently**
  - – Analogy: recording a concert with multiple microphones

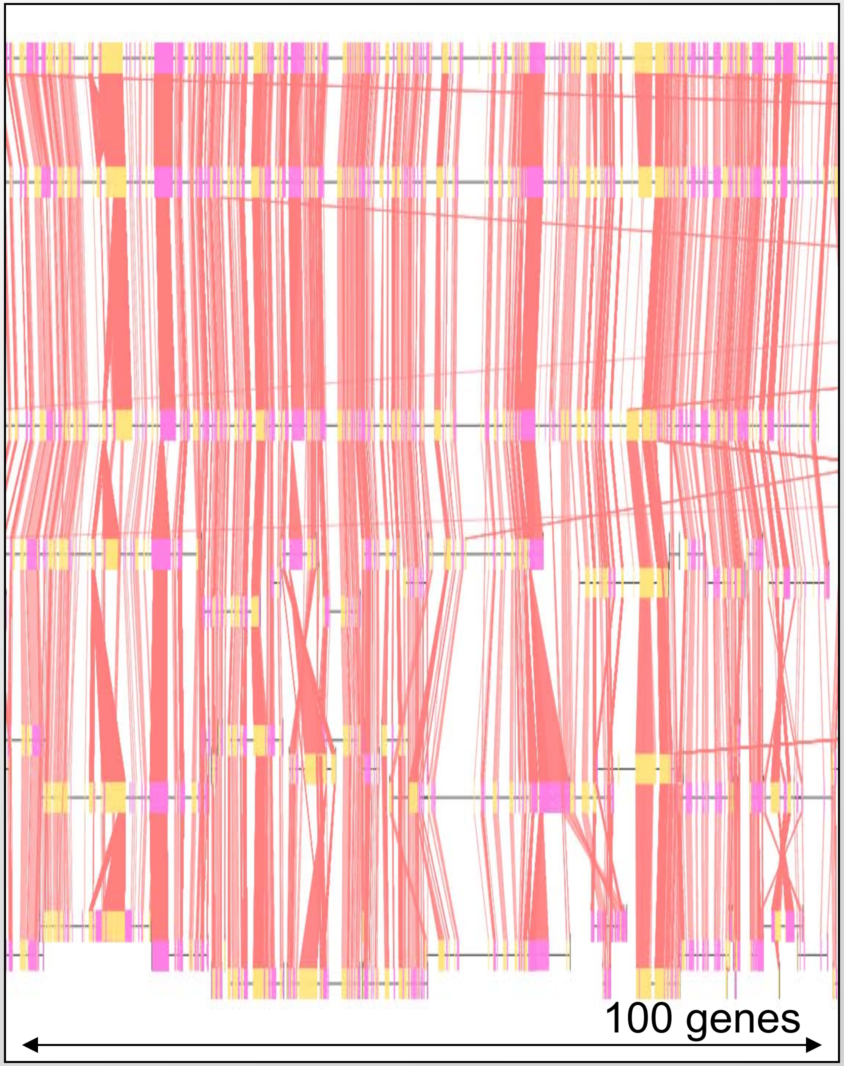
Comparative genomics I: Evolutionary signatures

- **Nucleotide conservation: evolutionary constraint**
  - Purifying selection, neutral branch length, discovery power
  - Detect constrained elements: nucleotides, windows, HMM
  - Estimate fraction constrained: signal vs. background
- **Evolutionary signatures: focus on pattern of change**
  - Different functions ⇔ Characteristic patterns of evolution
- **Signatures of protein-coding genes**
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_C Q_N$, scoring
  - Revise genes, read-through, excess constraint regions
- **Signatures of microRNA genes**
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation

Genome-wide alignments reveal orthologous segments

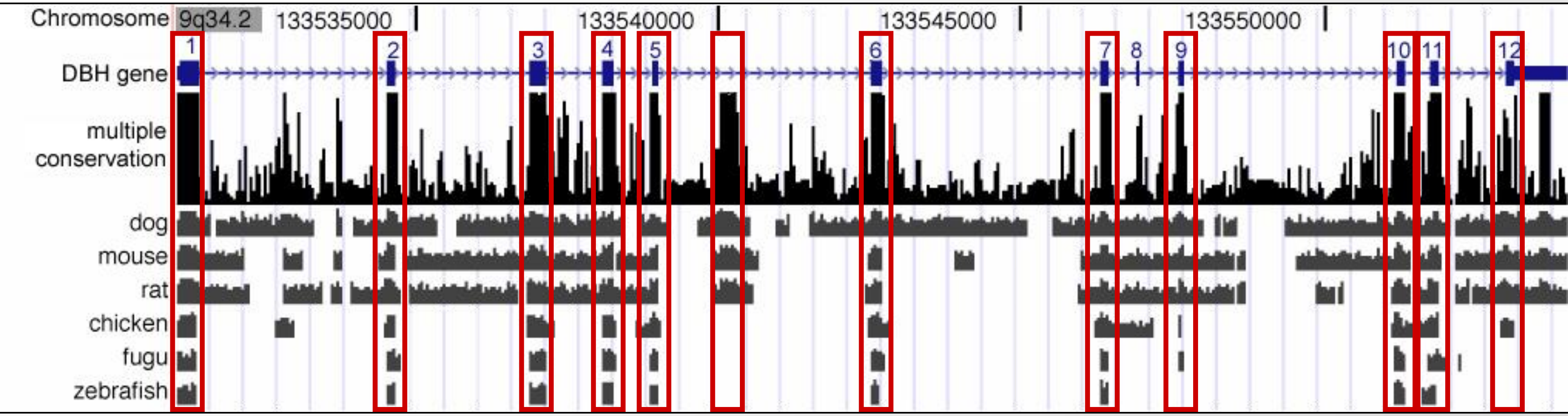


Courtesy of Don Gilbert. Used with permission.



- **Genome-wide alignments span entire genome**
- **Comparative identification of functional elements**

Comparative genomics and evolutionary signatures



- **Comparative genomics can reveal functional elements**
  - For example: exons are deeply conserved to mouse, chicken, fish
  - Many other elements are also strongly conserved: exons / regulatory?
- **Develop methods for estimating the level of constraint**
  - Count the number of edit operations, number of substitutions and gaps
  - Estimate the number of mutations (including estimate of back-mutations)
  - Incorporate information about neighborhood: conservation ‘windows’
  - Estimate the probability of a constrained ‘hidden state’: HMMs next week
  - Use phylogeny to estimate tree mutation rate, or ‘rejected substitutions’
  - Allow different portions of the tree to have different rates: phylogenetics

Detecting rates and patterns of selection (ω/π)

Estimating intensity of constraint (ω):

- Probabilistic model of substitution rate
- Maximum Likelihood (ML) estimation of ω
  - Report rate ω
  - Report log odds score that non-neutral
- Window-based vs. sitewise application

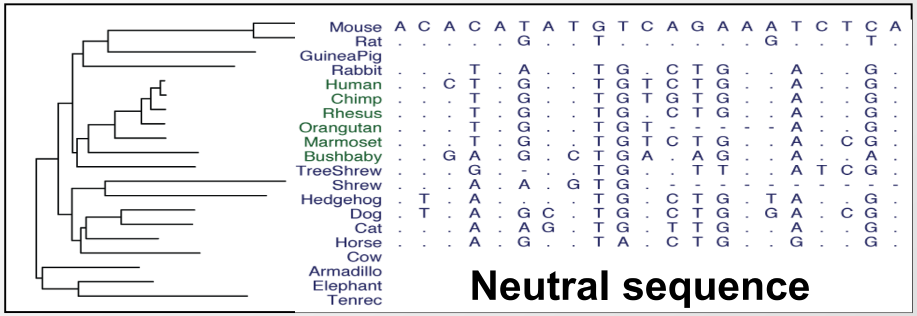
Detect unusual substitution pattern (π):

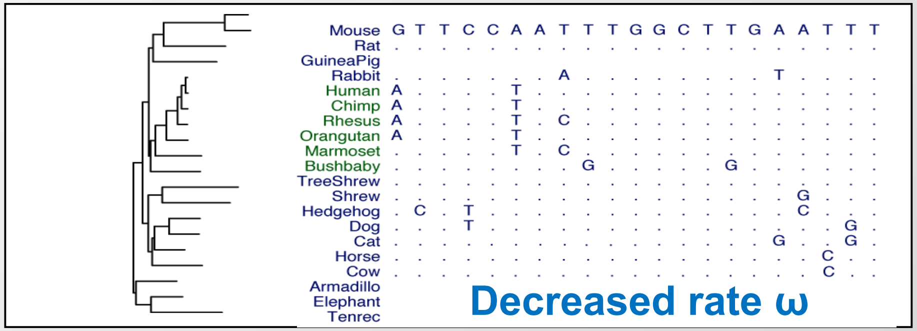
•Probabilistic model of stationary distribution that is different from background.

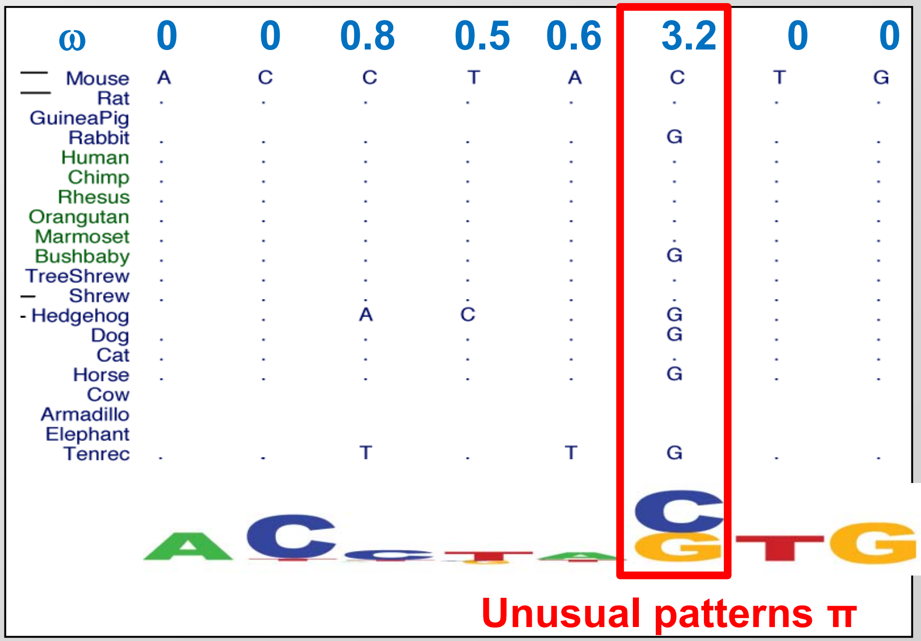
•ML estimator (π) of this vector

- Report PWM for each k-mer in genome.
- Report log odds score that non-neutral

Manuel Garber, Or Zuk, Xiaohui Xie


Neutral sequence

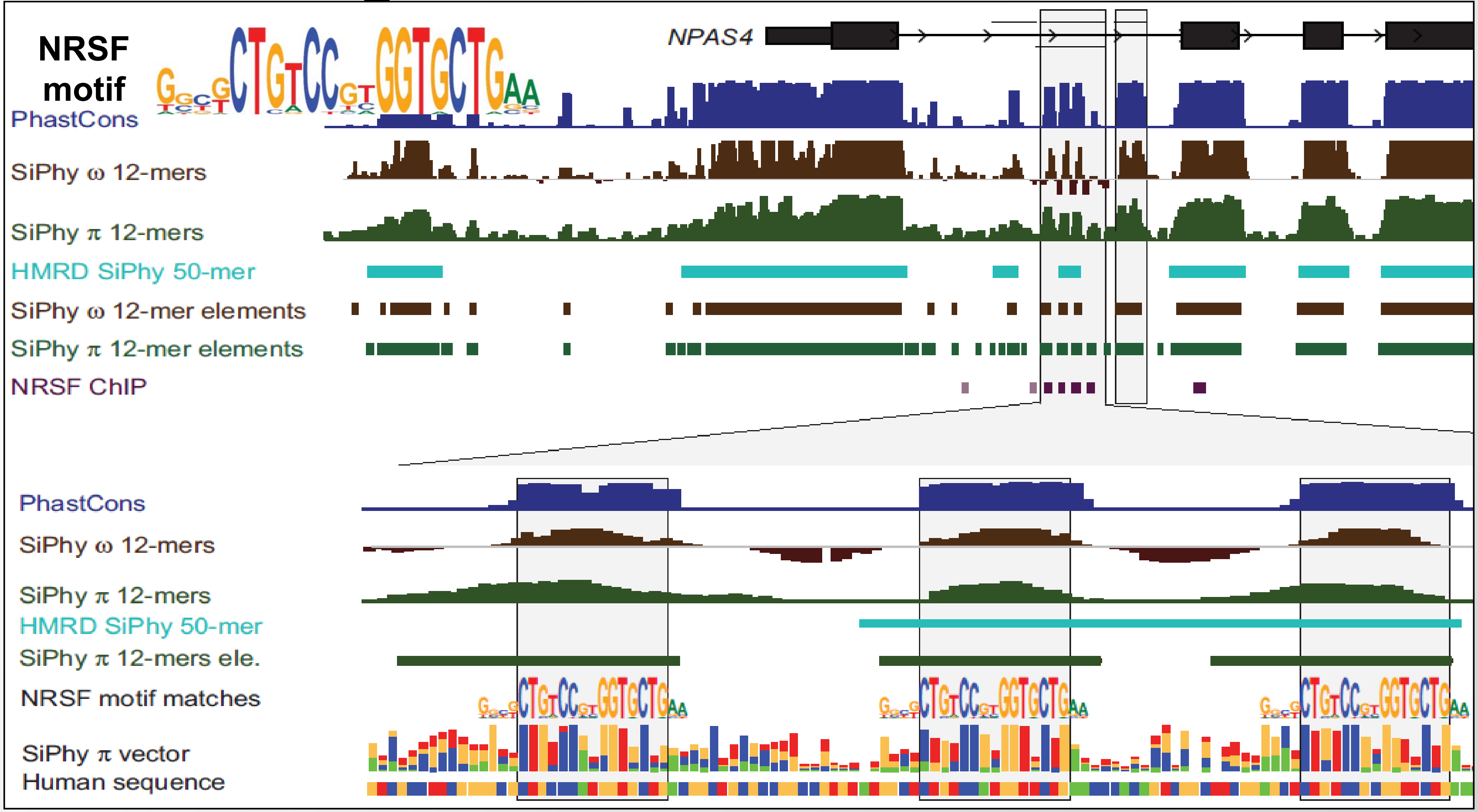

Decreased rate ω



ω	0	0	0.8	0.5	0.6	3.2	0	0
Mouse	A	C	C	T	A	C	T	G
Rat	.	.	.	.	.	.	.	.
GuineaPig								
Rabbit	.	.	.	.	.	G	.	.
Human	.	.	.	.	.	.	.	.
Chimp	.	.	.	.	.	.	.	.
Rhesus	.	.	.	.	.	.	.	.
Orangutan	.	.	.	.	.	.	.	.
Marmoset	.	.	.	.	.	.	.	.
Bushbaby	.	.	.	.	.	G	.	.
TreeShrew	.	.	.	.	.	.	.	.
Shrew	.	.	.	.	.	.	.	.
Hedgehog	.	.	A	C	.	G	.	.
Dog	.	.	.	.	.	G	.	.
Cat	.	.	.	.	.	.	.	.
Horse	.	.	.	.	.	G	.	.
Cow								
Armadillo								
Elephant								
Tenrec	.	.	T	.	T	G	.	.

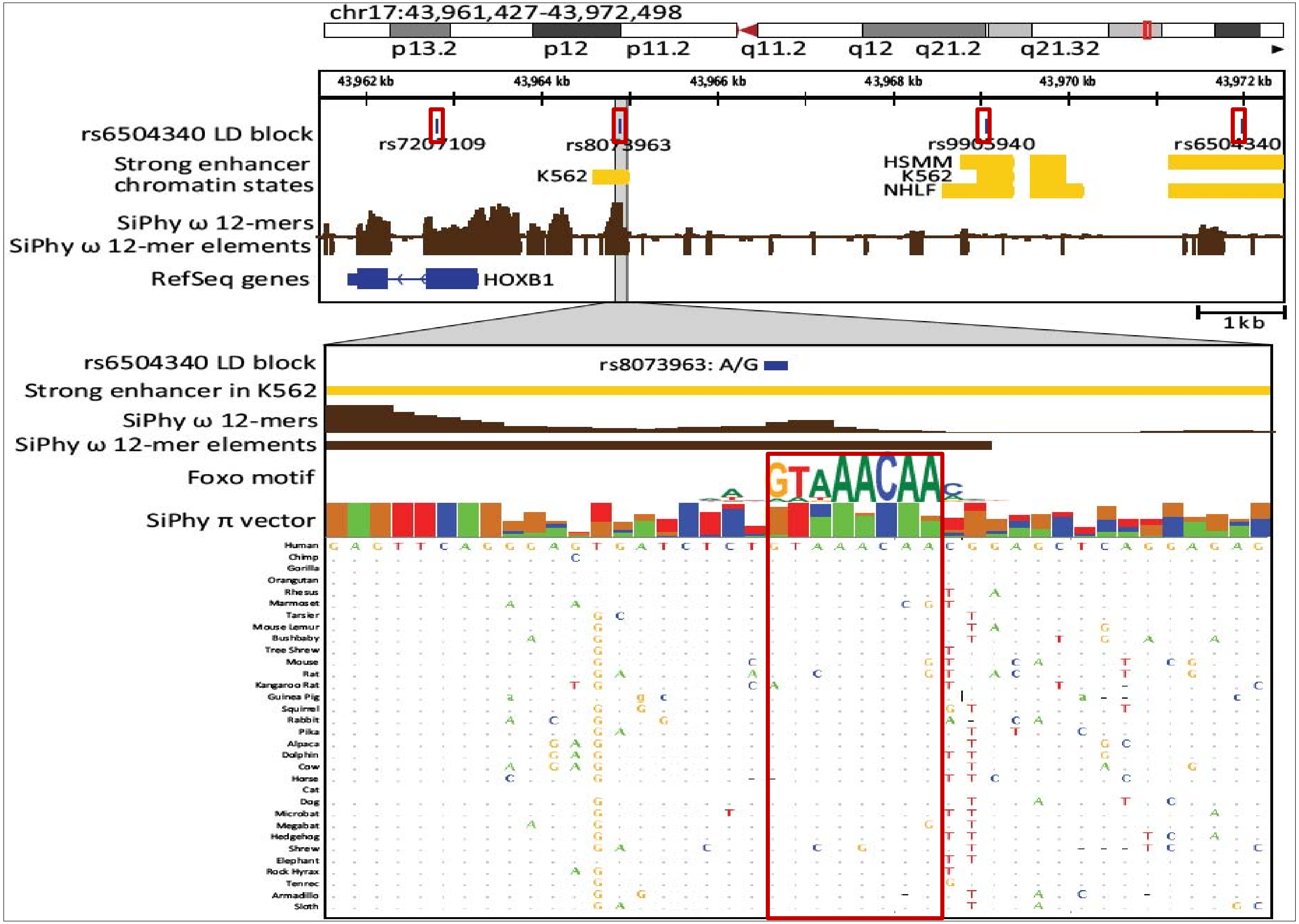
Unusual patterns π

Measuring constraint at individual nucleotides



- **Reveal individual transcription factor binding sites**
- **Within motif instances reveal position-specific bias**
- **More species: motif consensus directly revealed**

Detect SNPs that disrupt conserved regulatory motifs



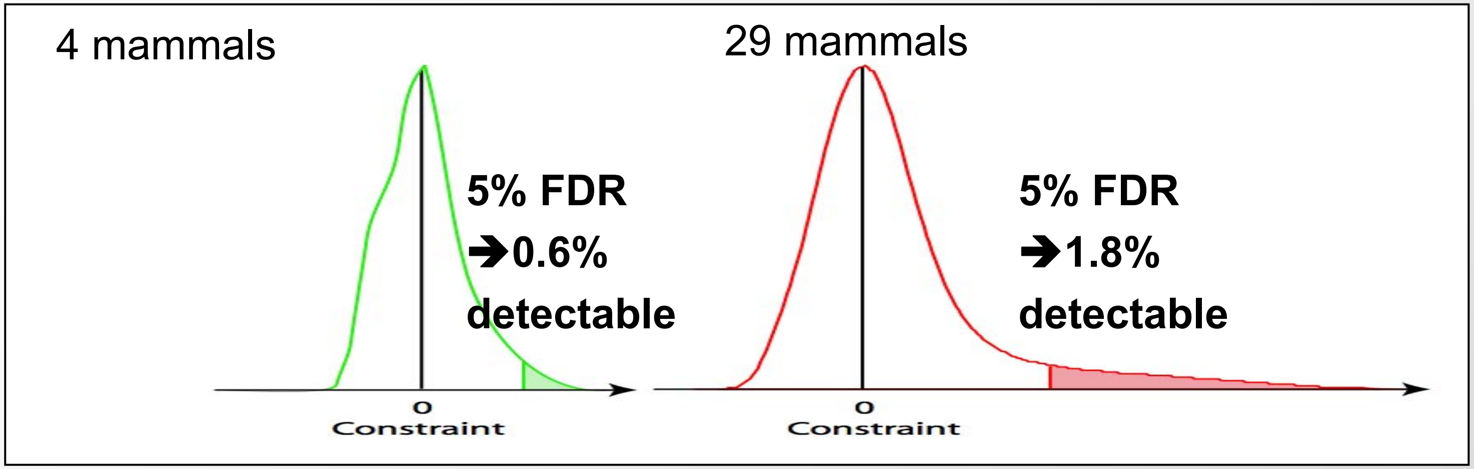
- Functionally-associated SNPs enriched in states, constraint
- Prioritize candidates, increase resolution, disrupted motifs

Comparative genomics I: Evolutionary signatures

- **Nucleotide conservation: evolutionary constraint**
  - Purifying selection, neutral branch length, discovery power
  - Detect constrained elements: nucleotides, windows, HMM
  - Estimate fraction constrained: signal vs. background
- **Evolutionary signatures: focus on pattern of change**
  - Different functions ⬄ Characteristic patterns of evolution
- **Signatures of protein-coding genes**
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_C Q_N$, scoring
  - Revise genes, read-through, excess constraint regions
- **Signatures of microRNA genes**
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation

Estimating portion of the genome under constraint

Constraint calculated over a **50mer**



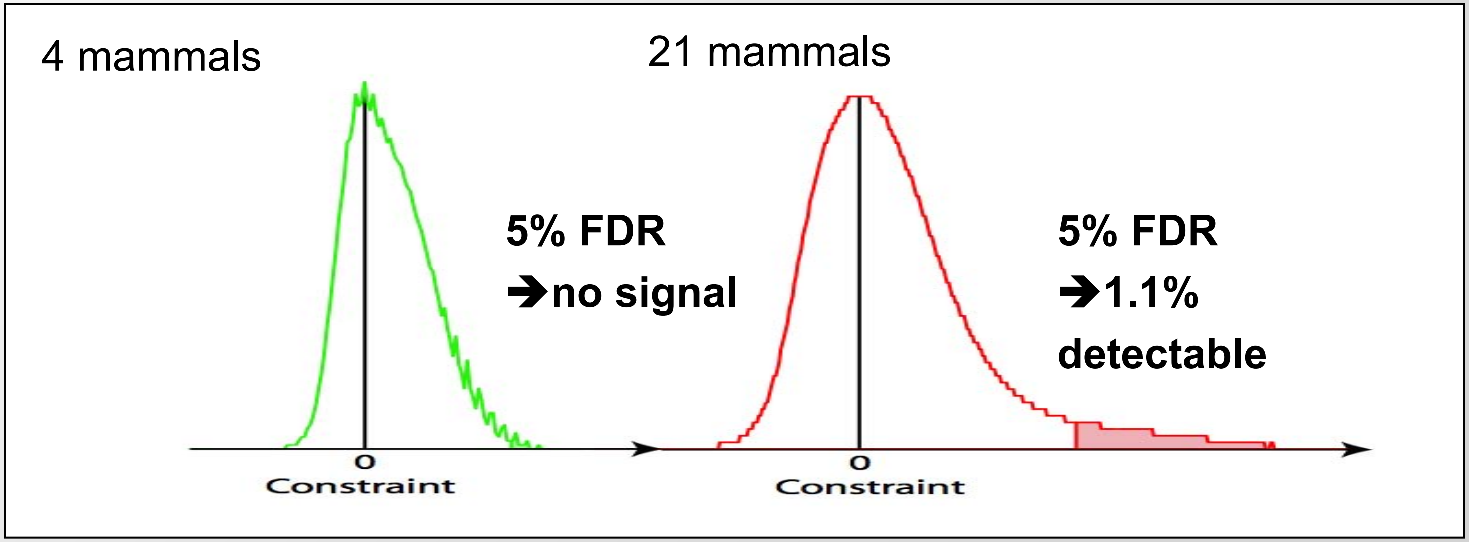
4 mammals

5% FDR
➔0.6%
detectable

29 mammals

5% FDR
➔1.8%
detectable

Constraint calculated over a **12mer**



4 mammals

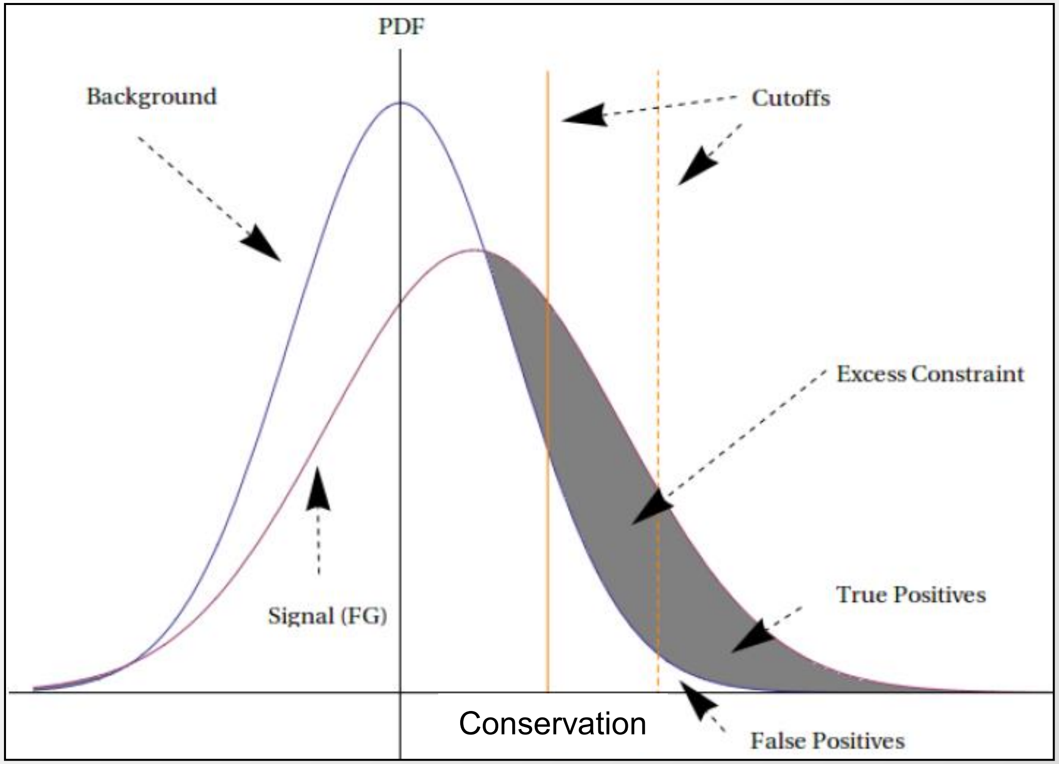
5% FDR
➔no signal

21 mammals

5% FDR
➔1.1%
detectable

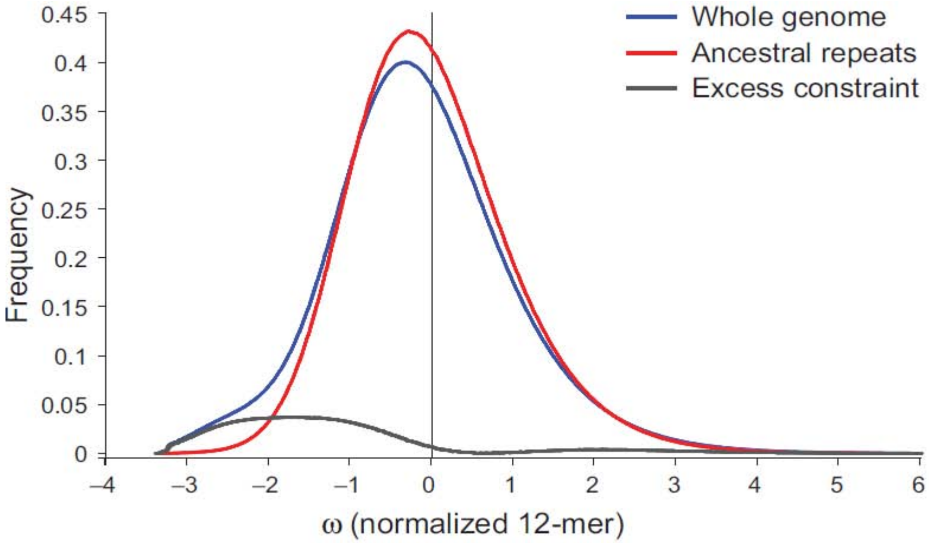
Or Zuk, Manuel Garber

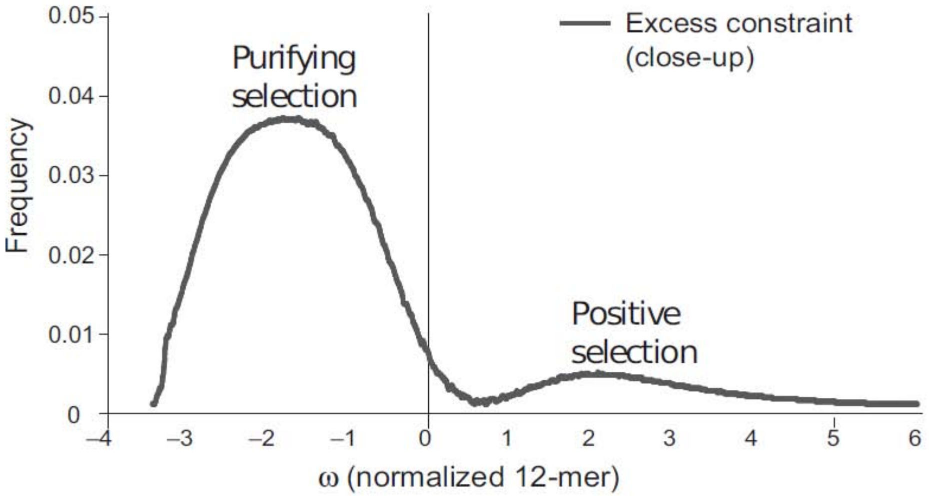
Estimating total fraction under constraint



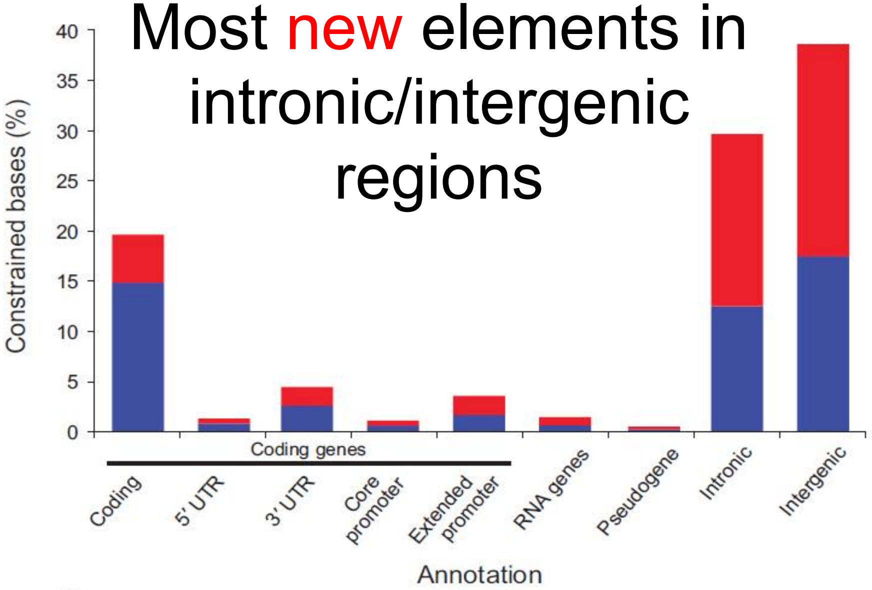
- Actual distribution of conservation scores (Signal) vs. expected distribution if no constraint (Background).
- At any cutoff: true positives (TP) and false predictions (FP)
- Can't **detect** all constrained elements since curves overlap
- But we can **estimate** the total amount of excess constraint by integrating over entire area between the two curves

Detection of evolutionarily constrained elements

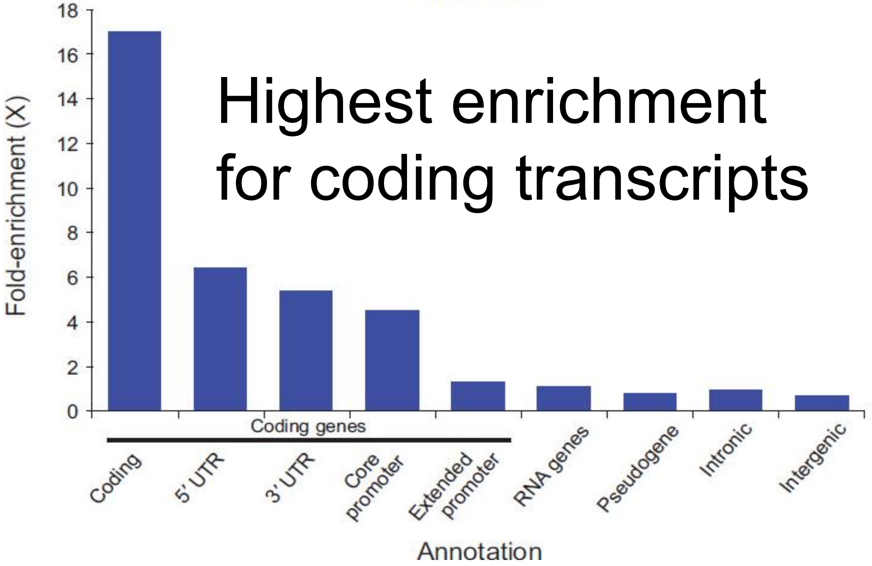




Excess positive/purifying selection



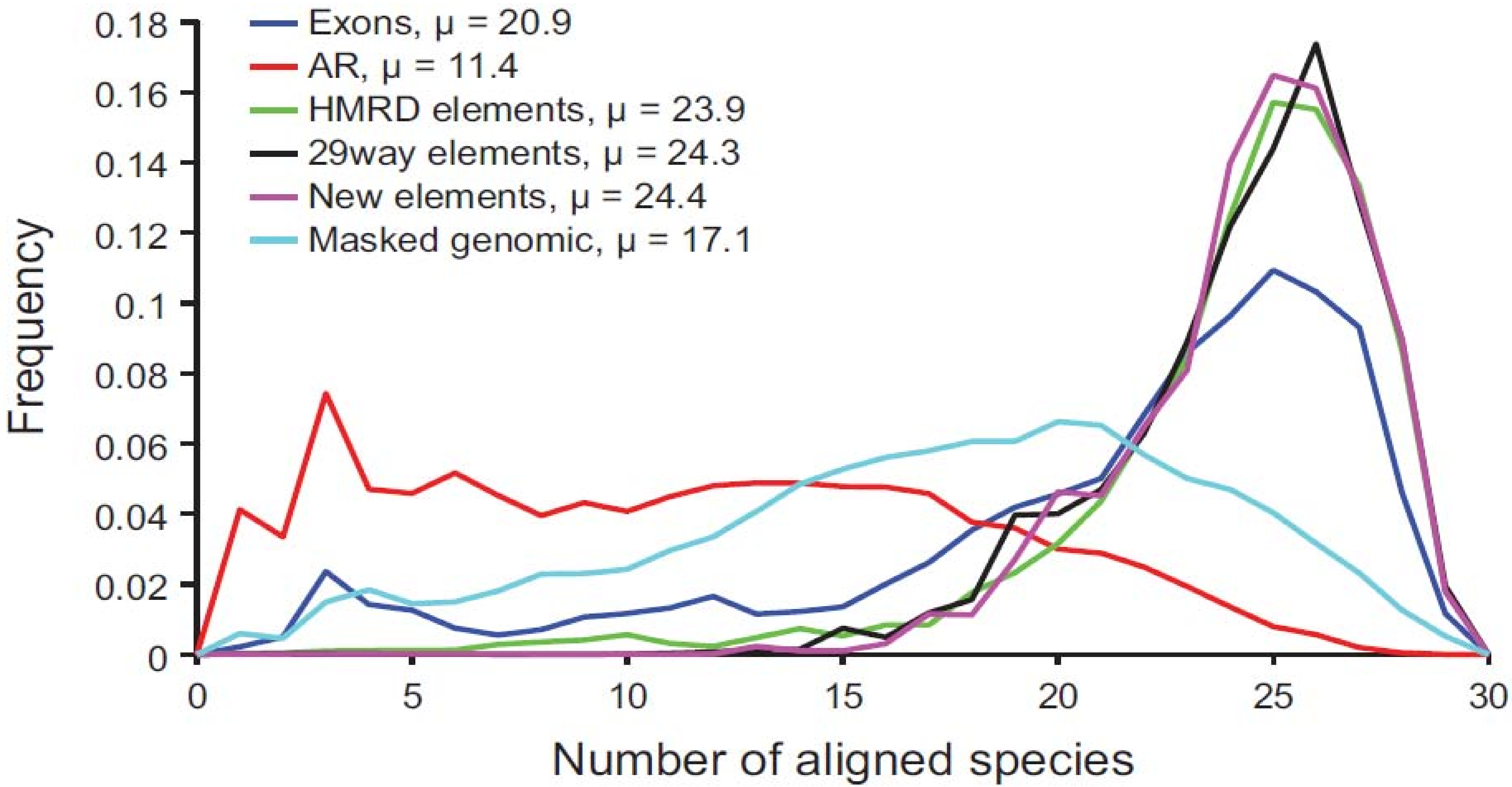
Most new elements in intronic/intergenic regions



Highest enrichment for coding transcripts

Distribution of constraint

Coverage depth higher in functional regions



Challenges of low-coverage genomes: varying aligment depth
Evidence of selection against deletions in functional regions

Increase in power from HMRD to 29 mammals

	π log-odds (12mers)	π log-odds (50mers)	ω (12mers)	ω (50mers)
29 mammals	7.1/1.5/4.6	6.8/1.8/4.1	5.7/ 1.1/3.8	5.7/1.8/3.0
(HMRD) Human Mouse Rat Dog	4.2/0.0/0.0	5.3/0.1/0.3	4.5/0.0/0.0	5.1/0.6/1.7

Estimated / kmers detectable at 5% FDR / base pairs detectable at 5% FDR

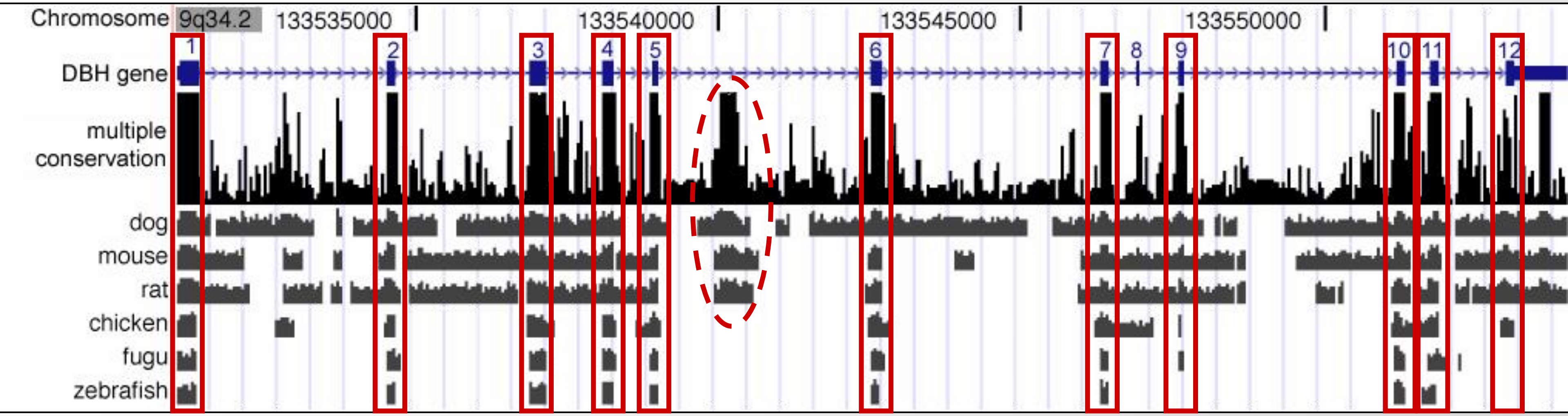
Small increase in estimate of genome percentage under constraint
Dramatic increase in power to detect small constrained elements

Manuel Garber, Or Zuk

Comparative genomics I: Evolutionary signatures

- **Nucleotide conservation: evolutionary constraint**
  - Purifying selection, neutral branch length, discovery power
  - Detect constrained elements: nucleotides, windows, HMM
  - Estimate fraction constrained: signal vs. background
- **Evolutionary signatures: focus on pattern of change**
  - Different functions ⇔ Characteristic patterns of evolution
- **Signatures of protein-coding genes**
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_C Q_N$, scoring
  - Revise genes, read-through, excess constraint regions
- **Signatures of microRNA genes**
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation

Comparative genomics and evolutionary signatures



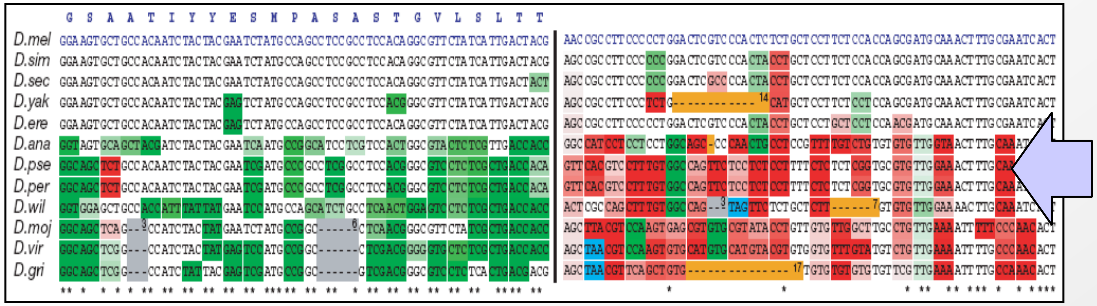
- **Comparative genomics can reveal functional elements**
  - For example: exons are deeply conserved to mouse, chicken, fish
  - Many other elements are also strongly conserved: exons / regulatory?

- **Can we also pinpoint specific functions of each region? Yes!**
  - Patterns of change distinguish different types of functional elements
  - Specific function ⇔ Selective pressures ⇔ Patterns of mutation/inse/del

- **Develop evolutionary signatures characteristic of each function**

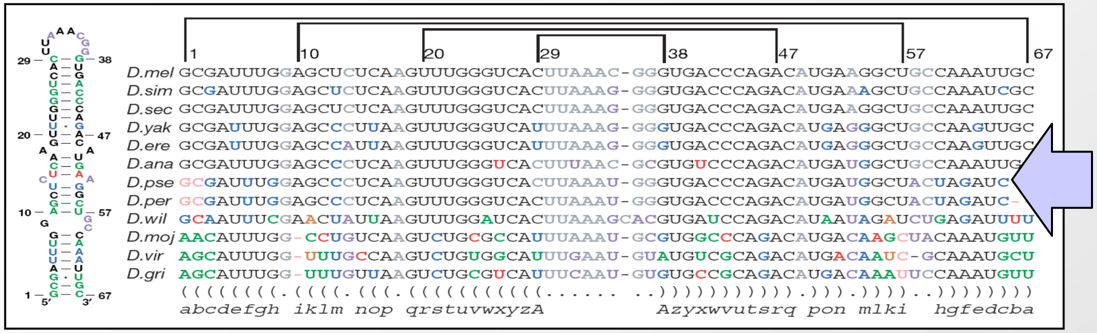
Stark *et al*, Nature 2007

Evolutionary signatures for diverse functions



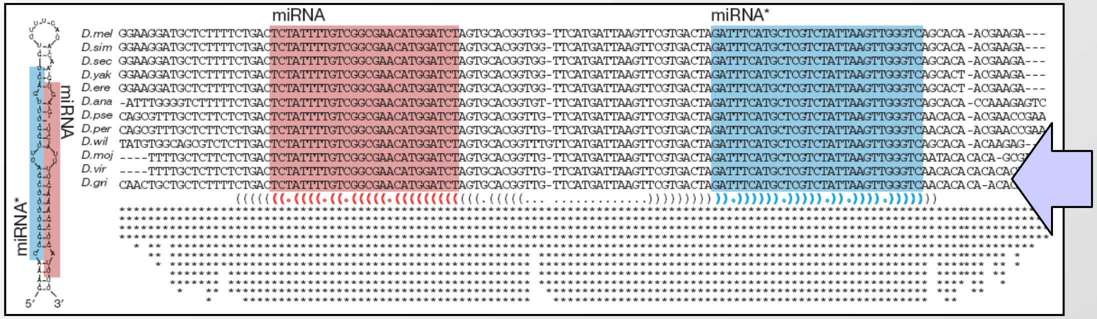
Protein-coding genes

- Codon Substitution Frequencies
- Reading Frame Conservation



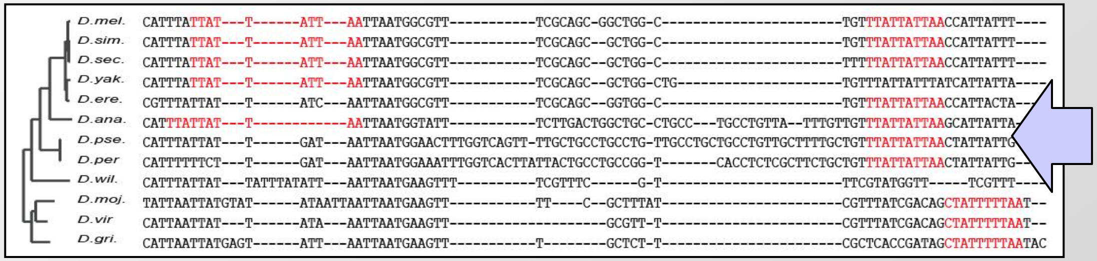
RNA structures

- Compensatory changes
- Silent G-U substitutions



microRNAs

- Shape of conservation profile
- Structural features: loops, pairs
- Relationship with 3'UTR motifs



Regulatory motifs

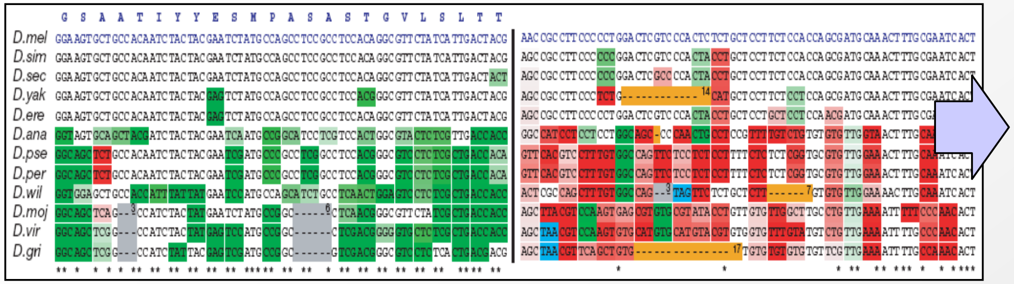
- Mutations preserve consensus
- Increased Branch Length Score
- Genome-wide conservation

Courtesy of Macmillan Publishers Limited. Used with permission.

Source: Stark, Alexander et al. "Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures." Nature 450, no. 7167 (2007): 219-232.

Stark et al, Nature 2007

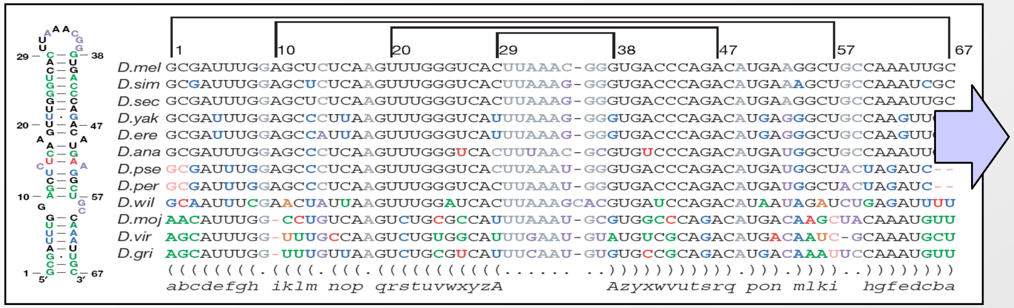
Implications for genome annotation / regulation



Novel protein-coding genes

Revised gene annotations

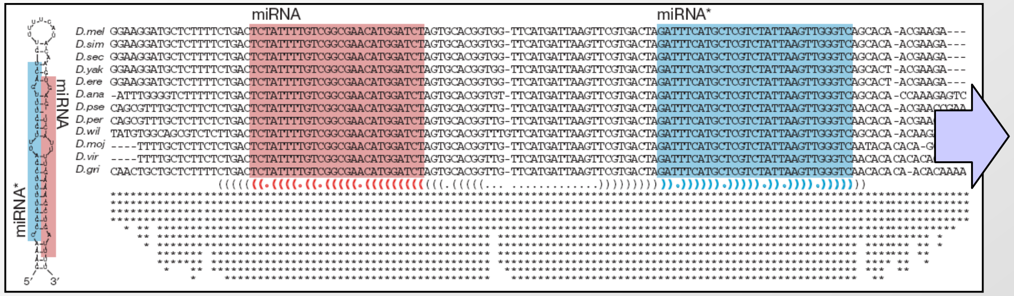
Unusual gene structures



Novel structural families

Targeting, editing, stability

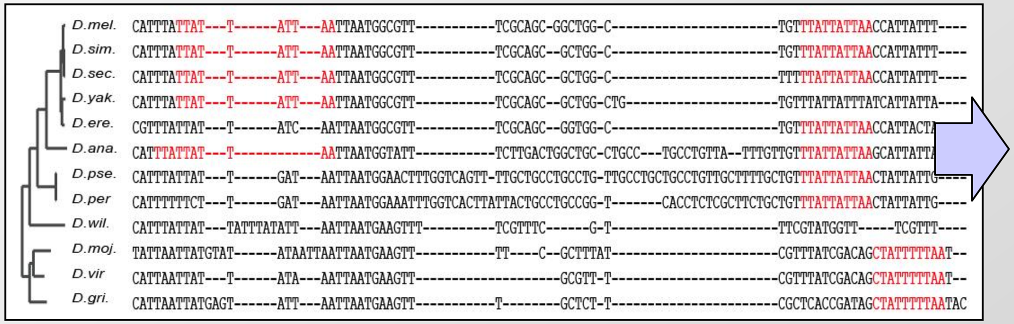
Riboswitches in mammals



Novel/expanded miR families

miR/miR* arm cooperation

Sense/anti-sense miR switches



Novel regulatory motifs

Regulatory motif instances

TF/miRNA regulatory networks

Single binding site resolution

Courtesy of Macmillan Publishers Limited. Used with permission.

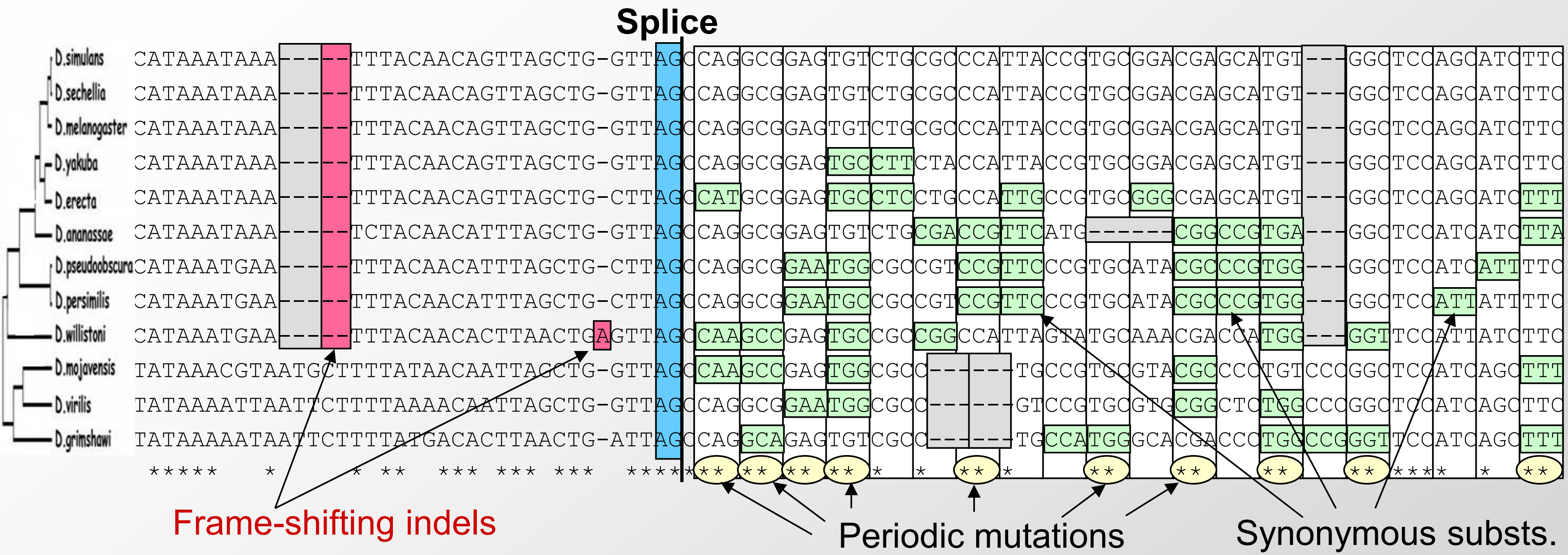
Source: Stark, Alexander et al. "Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures." Nature 450, no. 7167 (2007): 219-232.

Stark et al, Nature 2007

Comparative genomics I: Evolutionary signatures

- **Nucleotide conservation: evolutionary constraint**
  - Purifying selection, neutral branch length, discovery power
  - Detect constrained elements: nucleotides, windows, HMM
  - Estimate fraction constrained: signal vs. background
- **Evolutionary signatures: focus on pattern of change**
  - Different functions ⇔ Characteristic patterns of evolution
- **Signatures of protein-coding genes**
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_C Q_N$, scoring
  - Revise genes, read-through, excess constraint regions
- **Signatures of microRNA genes**
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation

Evolutionary signatures for protein-coding genes



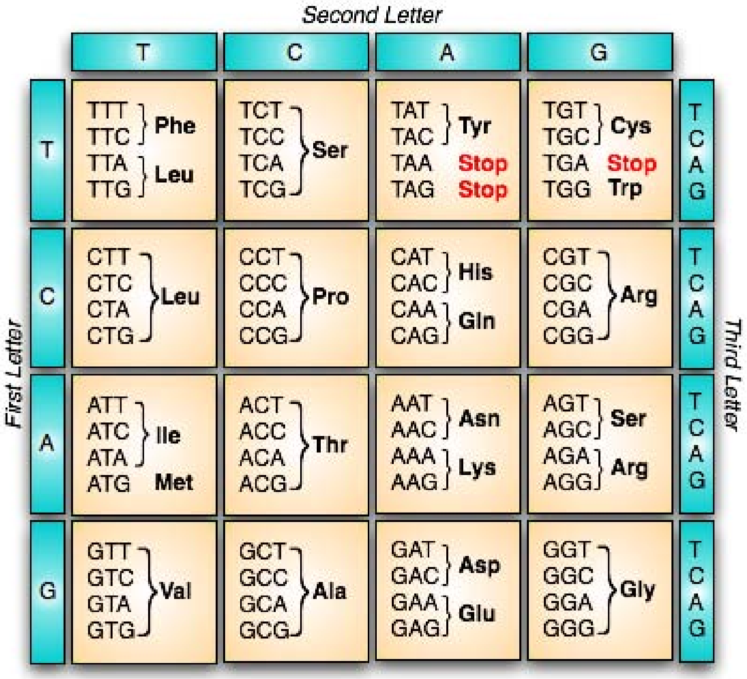
```
D.simulans      CATAAATAAA-----TTTACAACAGTTAGCTG-GTTAGCCAGGCGGAGTGTCTGCGCCCATTACCGTGCGGACGAGCATGT---GGCTCCAGCATCTTC
D.sechellia     CATAAATAAA-----TTTACAACAGTTAGCTG-GTTAGCCAGGCGGAGTGTCTGCGCCCATTACCGTGCGGACGAGCATGT---GGCTCCAGCATCTTC
D.melanogaster  CATAAATAAA-----TTTACAACAGTTAGCTG-GTTAGCCAGGCGGAGTGTCTGCGCCCATTACCGTGCGGACGAGCATGT---GGCTCCAGCATCTTC
D.yakuba        CATAAATAAA-----TTTACAACAGTTAGCTG-GTTAGCCAGGCGGAGTGCCTTCTACCATTACCGTGCGGACGAGCATGT---GGCTCCAGCATCTTC
D.erecta        CATAAATAAA-----TTTACAACAGTTAGCTG-CTTAGCCATGCGGAGTGCCTCCTGCCATTGCCGTGCGGGCGAGCATGT---GGCTCCAGCATCTTT
D.ananassae     CATAAATAAA-----TCTACAACATTTAGCTG-GTTAGCCAGGCGGAGTGTCTGCGACCGTTCATG------CGGCCGTGA---GGCTCCATCATCTTA
D.pseudoobscura CATAAATGAA-----TTTACAACATTTAGCTG-CTTAGCCAGGCGGAATGGCGCCGTCCGTTCCCGTGCATACGCCCGTGG---GGCTCCATCATTTTC
D.persimilis    CATAAATGAA-----TTTACAACATTTAGCTG-CTTAGCCAGGCGGAATGCCGCCGTCCGTTCCCGTGCATACGCCCGTGG---GGCTCCATTATTTTC
D.willistoni    CATAAATGAA-----TTTACAACACTTAACTGAGTTAGCCAAGCCGAGTGCCGCCGGCCATTAGTATGCAAACGACCATGG---GGTTCCATTATCTTC
D.mojavensis    TATAAACGTAATGCTTTTATAACAATTAGCTG-GTTAGCCAAGCCGAGTGGCGCC------TGCCGTGCGTACGCCCCTGTCCCGGCTCCATCAGCTTT
D.virilis       TATAAAATTAATTCTTTTAAAACAATTAGCTG-GTTAGCCAGGCGGAATGGCGCC------GTCCGTGCGTGCGGCTCTGGCCCGGCTCCATCAGCTTC
D.grimshawi     TATAAAAATAATTCTTTTATGACACTTAACTG-ATTAGCCAGGCAGAGTGTCGCC------TGCCATGGGCACGACCCTGGCCGGGTTCCATCAGCTTT
                 *****   *     * **  *** *** ***  ******* ** ** ** *  *  ** *     **    **    **    ** ****  *  **
```

- **Same conservation levels, distinct patterns of divergence**
  - Gaps are multiples of three (preserve amino acid translation)
  - Mutations are largely 3-periodic (silent codon substitutions)
  - Specific triplets exchanged more frequently (conservative substs.)
  - Conservation boundaries are sharp (pinpoint individual splicing signals)

➔ Evolutionary signatures of protein-coding selection

Evolutionary signatures of protein-coding genes

	the	fat	cat	sat
Δ1	the	atc	ats	at
Δ2	the	tca	tsa	t
Δ3	the	cat	sat	



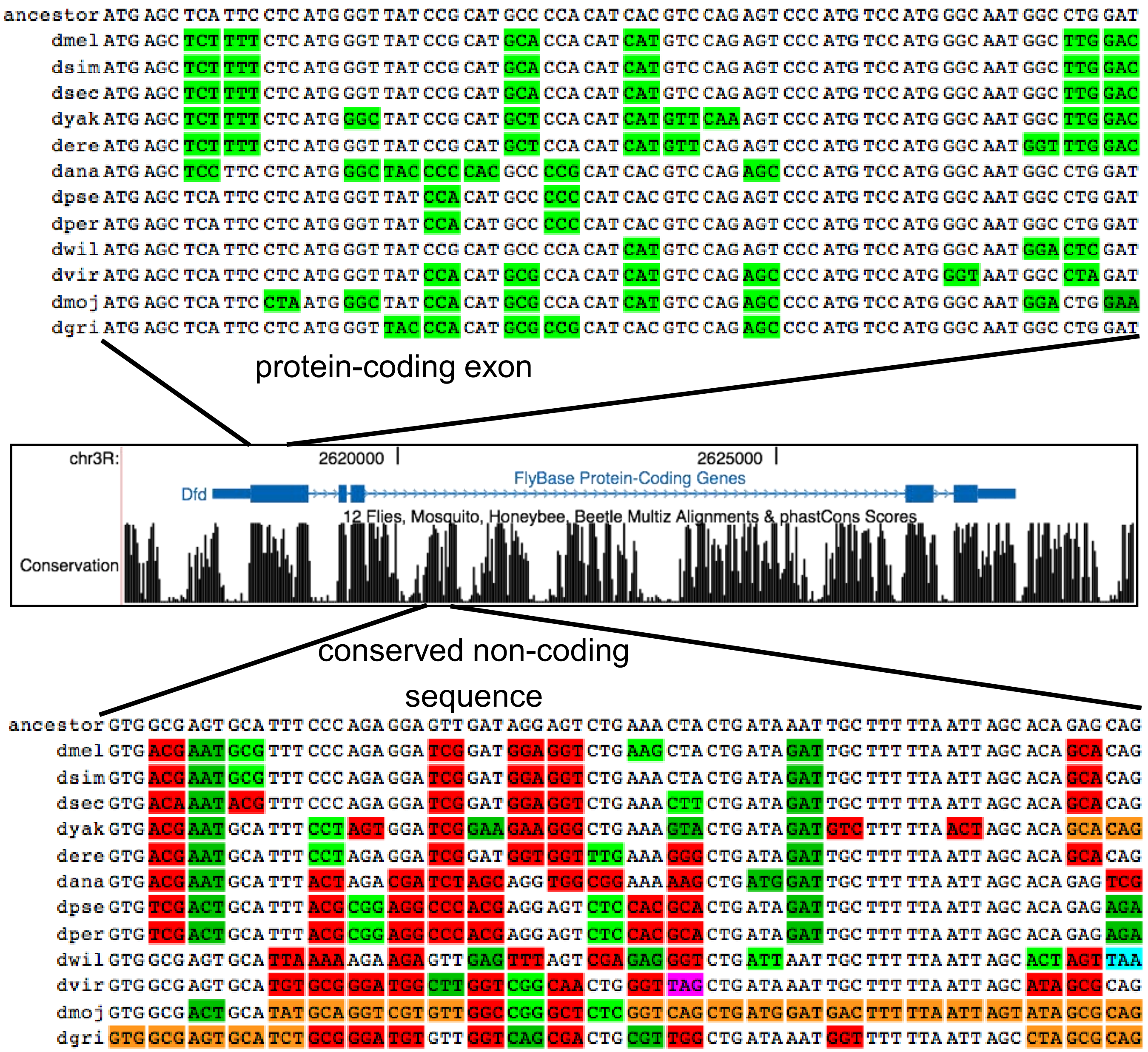
	Second Letter				
First Letter	T	C	A	G	Third Letter
T	TTT TTC Phe TTA TTG Leu	TCT TCC TCA TCG Ser	TAT TAC Tyr TAA Stop TAG Stop	TGT TGC Cys TGA Stop TGG Trp	T C A G
C	CTT CTC CTA CTG Leu	CCT CCC CCA CCG Pro	CAT CAC His CAA CAG Gln	CGT CGC CGA CGG Arg	T C A G
A	ATT ATC ATA Ile ATG Met	ACT ACC ACA ACG Thr	AAT AAC Asn AAA AAG Lys	AGT AGC Ser AGA AGG Arg	T C A G
G	GTT GTC GTA GTG Val	GCT GCC GCA GCG Ala	GAT GAC Asp GAA GAG Glu	GGT GGC GGA GGG Gly	T C A G

DNA insertions and deletions can either insert/remove AAs, or totally mangle the remainder of the protein (frameshift).

Some point mutations to the DNA sequence do not change its protein translation at all.

Natural selection tends to tolerate mutations with little/no effect on the protein.

Protein-coding sequences tolerate distinctive types of change



```
ancestor ATG AGC TCA TTC CTC ATG GGT TAT CCG CAT GCC CCA CAT CAC GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dmel ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCA CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dsim ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCA CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dsec ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCA CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dyak ATG AGC TCT TTT CTC ATG GGC TAT CCG CAT GCT CCA CAT CAT GTT CAA AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dere ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCT CCA CAT CAT GTT CAG AGT CCC ATG TCC ATG GGC AAT GGT TTG GAC
    dana ATG AGC TCC TTC CTC ATG GGC TAC CCC CAC GCC CCG CAT CAC GTC CAG AGC CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dpse ATG AGC TCA TTC CTC ATG GGT TAT CCA CAT GCC CCC CAT CAC GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dper ATG AGC TCA TTC CTC ATG GGT TAT CCA CAT GCC CCC CAT CAC GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dwil ATG AGC TCA TTC CTC ATG GGT TAT CCG CAT GCC CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGA CTC GAT
    dvir ATG AGC TCA TTC CTC ATG GGT TAT CCA CAT GCG CCA CAT CAT GTC CAG AGC CCC ATG TCC ATG GGT AAT GGC CTA GAT
    dmoj ATG AGC TCA TTC CTA ATG GGC TAT CCA CAT GCG CCA CAT CAT GTC CAG AGC CCC ATG TCC ATG GGC AAT GGA CTG GAA
    dgri ATG AGC TCA TTC CTC ATG GGT TAC CCA CAT GCG CCG CAT CAC GTC CAG AGC CCC ATG TCC ATG GGC AAT GGC CTG GAT
```

protein-coding exon

chr3R:
2620000
2625000
FlyBase Protein-Coding Genes
Dfd
12 Flies, Mosquito, Honeybee, Beetle Multiz Alignments & phastCons Scores
Conservation

conserved non-coding sequence

```
ancestor GTG GCG AGT GCA TTT CCC AGA GGA GTT GAT AGG AGT CTG AAA CTA CTG ATA AAT TGC TTT TTA ATT AGC ACA GAG CAG
    dmel GTG ACG AAT GCG TTT CCC AGA GGA TCG GAT GGA GGT CTG AAG CTA CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dsim GTG ACG AAT GCG TTT CCC AGA GGA TCG GAT GGA GGT CTG AAA CTA CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dsec GTG ACA AAT ACG TTT CCC AGA GGA TCG GAT GGA GGT CTG AAA CTT CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dyak GTG ACG AAT GCA TTT CCT AGT GGA TCG GAA GAA GGG CTG AAA GTA CTG ATA GAT GTC TTT TTA ACT AGC ACA GCA CAG
    dere GTG ACG AAT GCA TTT CCT AGA GGA TCG GAT GGT GGT TTG AAA GGG CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dana GTG ACG AAT GCA TTT ACT AGA CGA TCT AGC AGG TGG CGG AAA AAG CTG ATG GAT TGC TTT TTA ATT AGC ACA GAG TCG
    dpse GTG TCG ACT GCA TTT ACG CGG AGG CCC ACG AGG AGT CTC CAC GCA CTG ATA GAT TGC TTT TTA ATT AGC ACA GAG AGA
    dper GTG TCG ACT GCA TTT ACG CGG AGG CCC ACG AGG AGT CTC CAC GCA CTG ATA GAT TGC TTT TTA ATT AGC ACA GAG AGA
    dwil GTG GCG AGT GCA TTA AAA AGA AGA GTT GAG TTT AGT CGA GAG GGT CTG ATT AAT TGC TTT TTA ATT AGC ACT AGT TAA
    dvir GTG GCG AGT GCA TGT GCG GGA TGG CTT GGT CGG CAA CTG GGT TAG CTG ATA AAT TGC TTT TTA ATT AGC ATA GCG CAG
    dmoj GTG GCG ACT GCA TAT GCA GGT CGT GTT GGC CGG GCT CTC GGT CAG CTG ATG GAT GAC TTT TTA ATT AGT ATA GCG CAG
    dgri GTG GCG AGT GCA TCT GCG GGA TGT GTT GGT CAG CGA CTG CGT TGG CTG ATA AAT GGT TTT TTA ATT AGC CTA GCG CAG
```

synonymous

conservative

non-conservative

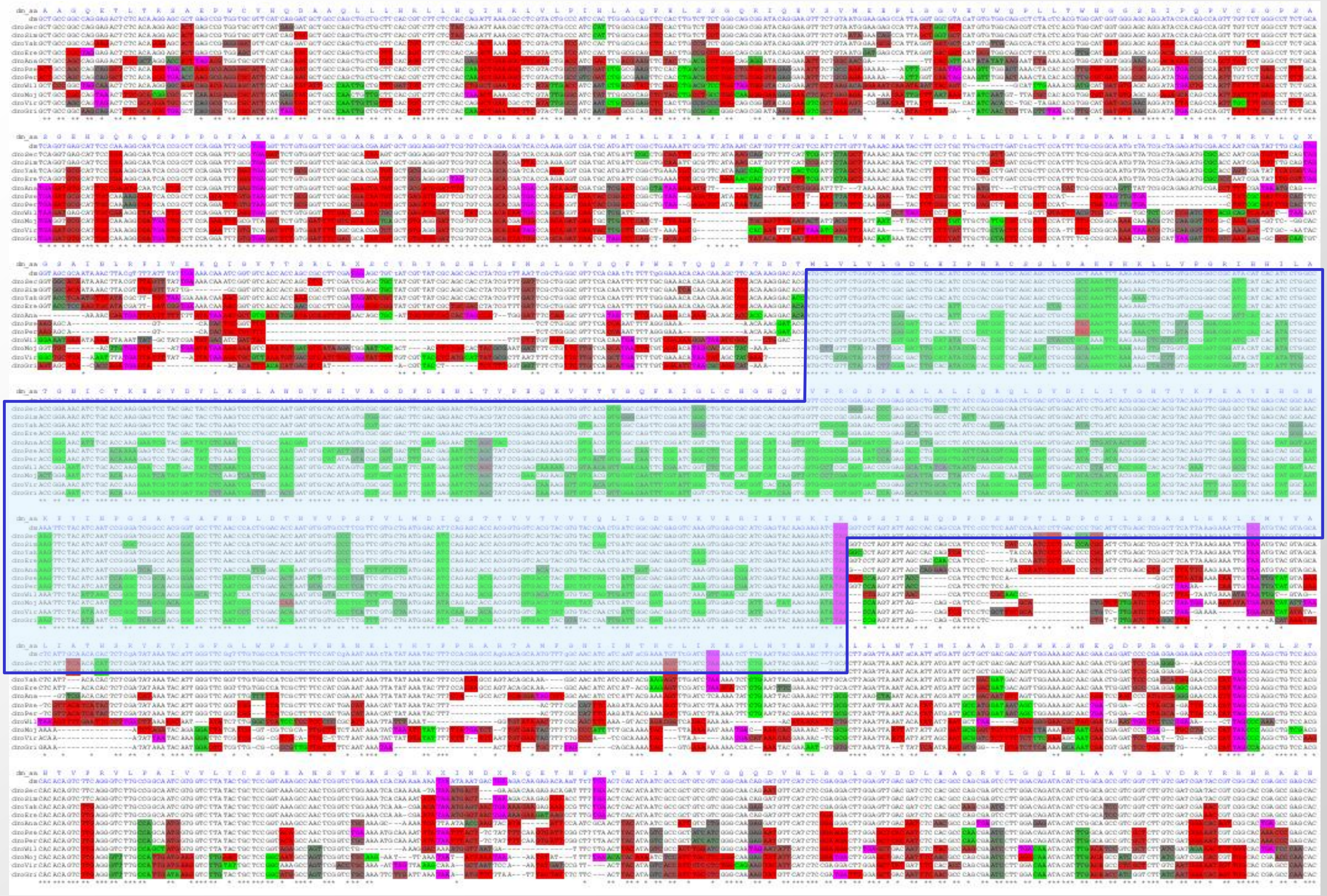
frame-shifted

three stop codons

Known genes stand out

🟩 Substitution typical of protein-coding regions

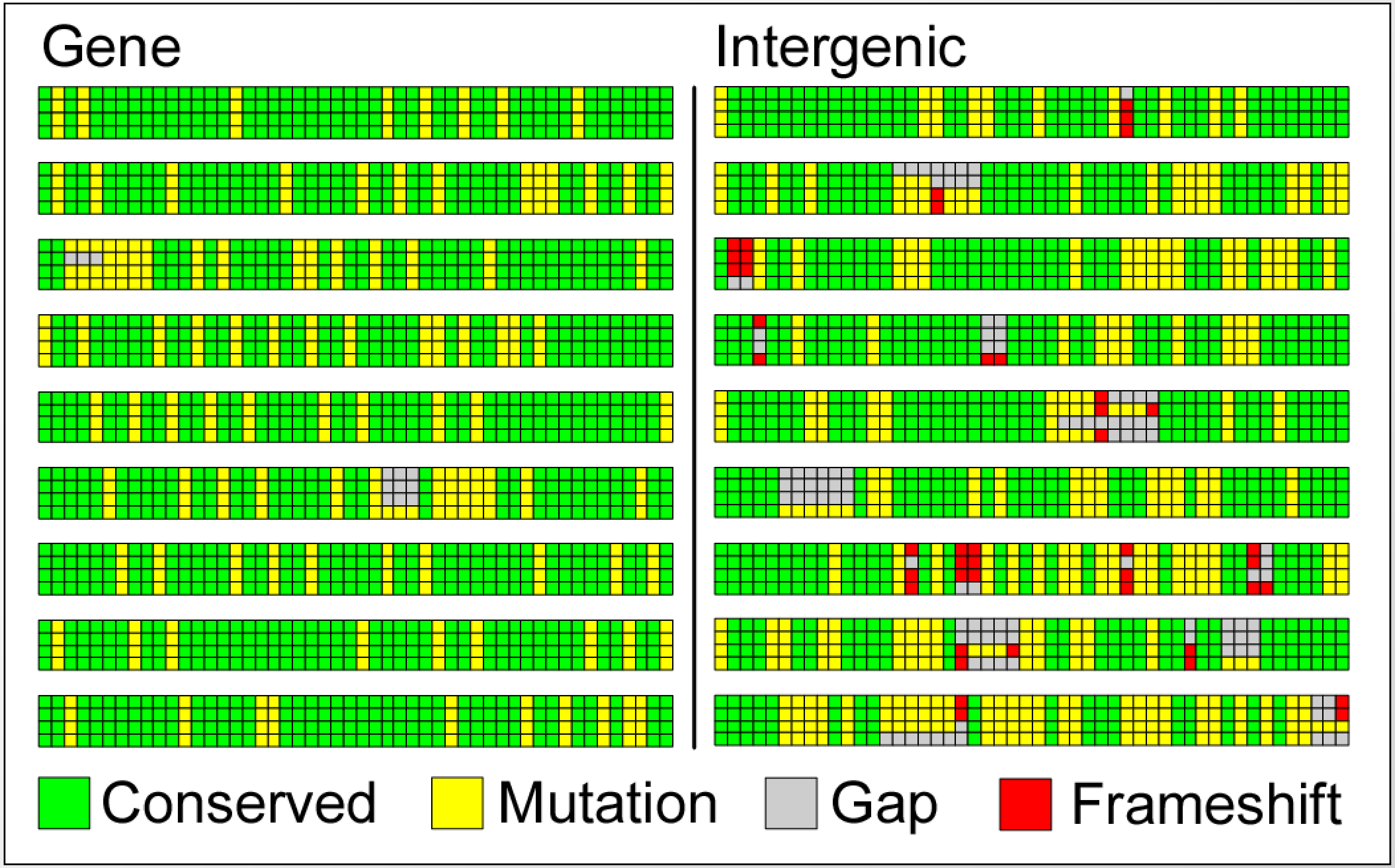
🟥 Substitution typical of intergenic regions



Comparative genomics I: Evolutionary signatures

- **Nucleotide conservation: evolutionary constraint**
  - Purifying selection, neutral branch length, discovery power
  - Detect constrained elements: nucleotides, windows, HMM
  - Estimate fraction constrained: signal vs. background
- **Evolutionary signatures: focus on pattern of change**
  - Different functions ⬄ Characteristic patterns of evolution
- **Signatures of protein-coding genes**
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_C Q_N$, scoring
  - Revise genes, read-through, excess constraint regions
- **Signatures of microRNA genes**
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation

Signature 1: Reading frame conservation

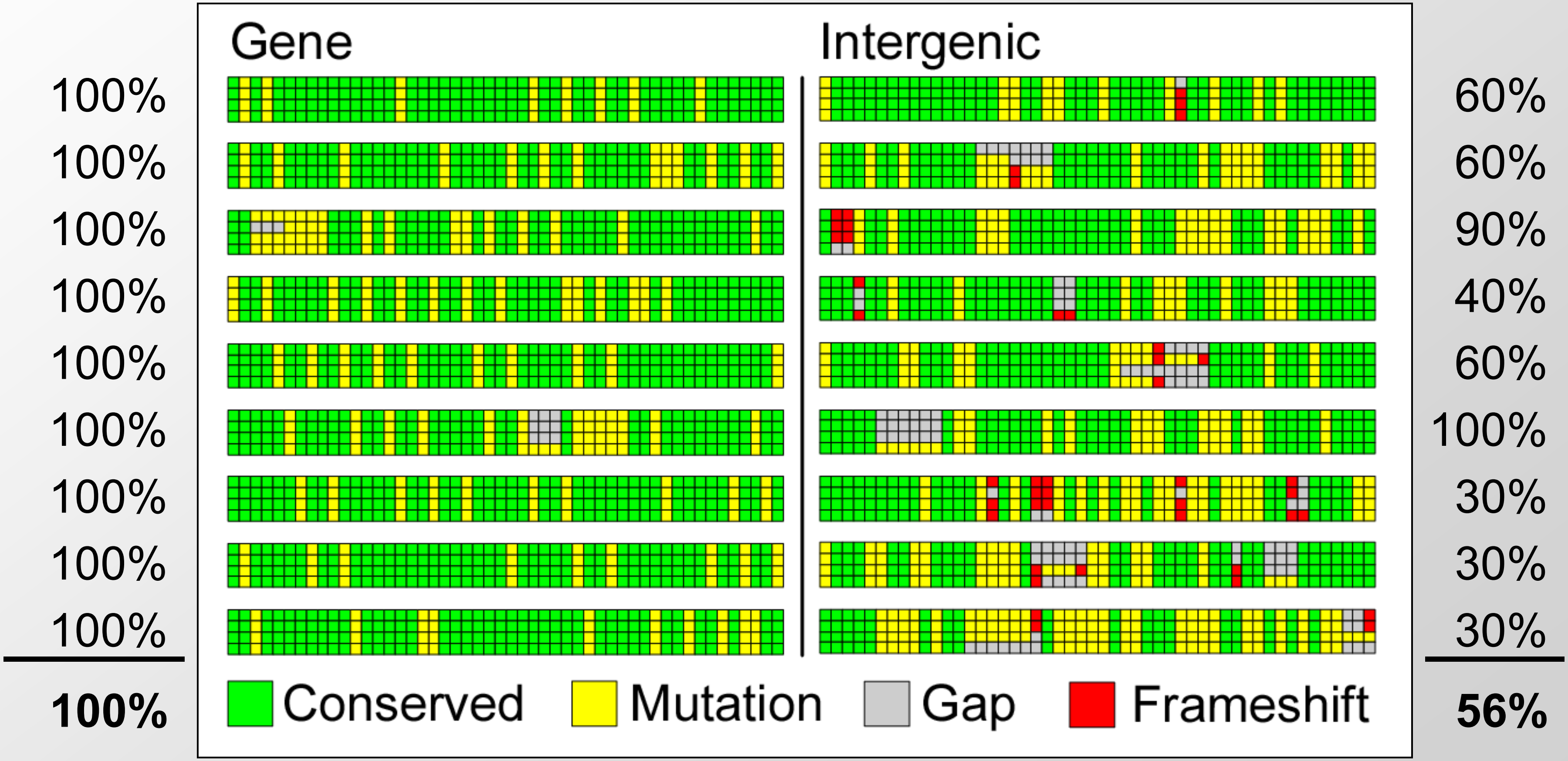


RFC	RFC
100%	60%
100%	55%
100%	90%
100%	40%
100%	60%
100%	100%
100%	20%
100%	30%
100%	40%
→100%	→60%

	Genes	Intergenic		Separation
🟨 Mutations	30%	58%	→	2-fold
⬜ Gaps	1.3%	14%	→	10-fold
🟥 Frameshifts	0.14%	10.2%	→	**75-fold**

Reading Frame Conservation Test

```
Scer     CTTCTAGATTTTCATCTT-GTCGATGTTCAAACAACGTGTTA-----TCAGAGAAACAGCTCTATGAGAAATCAGCTGATG
Scer_f1  123123123123123123-12312312312312312312312-----3123123123123123123123123123123123

Spar     TATTCATA-TCTCATCTTCATCAATGTTCAAACAGCGTGTTACAGACACAGAGAAACAGCTTC-TGAGAAGTCAGCCGGTG    RFC
Spar_f1  12312312-312312312312312312312312312312312312312312312312312312-31231231231231231  →  43%
Spar_f2  23123123-123123123123123123123123123123123123123123123123123123-12312312312312312  →  34%
Spar_f3  31231231-231231231231231231231231231231231231231231231231231231-23123123123123123  →  23%

         ←F1→     ←F2→     ←————————F1————————→      ←————F2————→     ←————F3————→
```

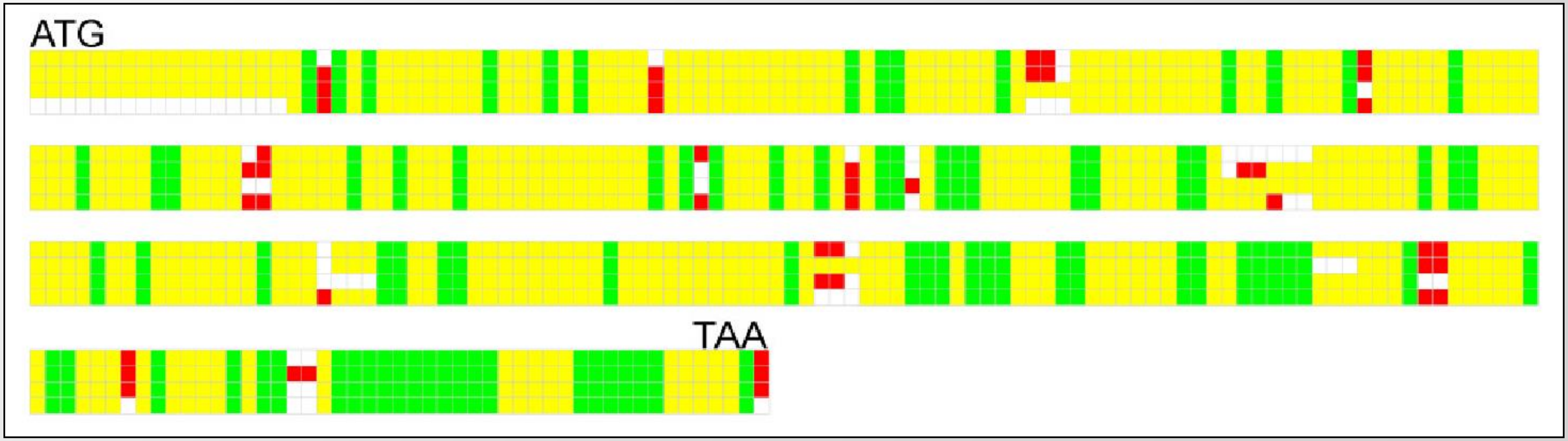


Revisiting gene content with RFC test

	Accept	Reject
~4000 named genes	99.9%	0.1%
~300 intergenic regions	1%	99%
2000 Hypothetical ORFs	1500	500

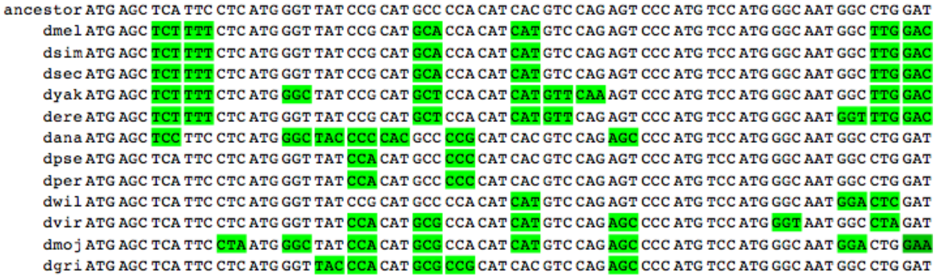
High sensitivity and specificity

Example of a rejected ORF



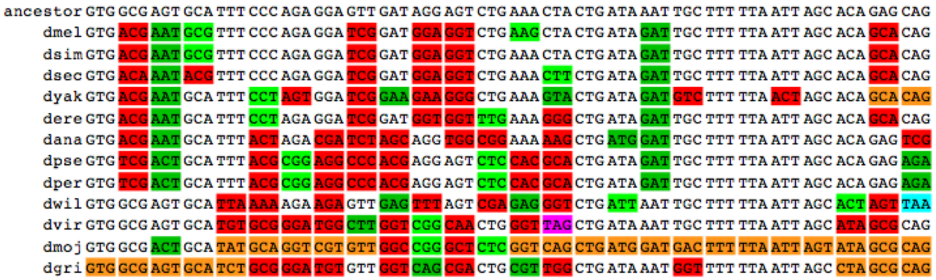
Comparative genomics I: Evolutionary signatures

- **Nucleotide conservation: evolutionary constraint**
  - Purifying selection, neutral branch length, discovery power
  - Detect constrained elements: nucleotides, windows, HMM
  - Estimate fraction constrained: signal vs. background
- **Evolutionary signatures: focus on pattern of change**
  - Different functions ⬄ Characteristic patterns of evolution
- **Signatures of protein-coding genes**
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_C Q_N$, scoring
  - Revise genes, read-through, excess constraint regions
- **Signatures of microRNA genes**
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation



```
ancestor ATG AGC TCA TTC CTC ATG GGT TAT CCG CAT GCC CCA CAT CAC GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dmel ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCA CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dsim ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCA CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dsec ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCA CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dyak ATG AGC TCT TTT CTC ATG GGC TAT CCG CAT GCT CCA CAT CAT GTT CAA AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dere ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCT CCA CAT CAT GTT CAG AGT CCC ATG TCC ATG GGC AAT GGT TTG GAC
    dana ATG AGC TCC TTC CTC ATG GGC TAC CCC CAC GCC CCG CAT CAC GTC CAG AGC CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dpse ATG AGC TCA TTC CTC ATG GGT TAT CCA CAT GCC CCC CAT CAC GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dper ATG AGC TCA TTC CTC ATG GGT TAT CCA CAT GCC CCC CAT CAC GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dwil ATG AGC TCA TTC CTC ATG GGT TAT CCG CAT GCC CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGA CTC GAT
    dvir ATG AGC TCA TTC CTC ATG GGT TAT CCA CAT GCG CCA CAT CAT GTC CAG AGC CCC ATG TCC ATG GGT AAT GGC CTA GAT
    dmoj ATG AGC TCA TTC CTA ATG GGC TAT CCA CAT GCG CCA CAT CAT GTC CAG AGC CCC ATG TCC ATG GGC AAT GGA CTG GAA
    dgri ATG AGC TCA TTC CTC ATG GGT TAC CCA CAT GCG CCG CAT CAC GTC CAG AGC CCC ATG TCC ATG GGC AAT GGC CTG GAT
```

protein-coding exon



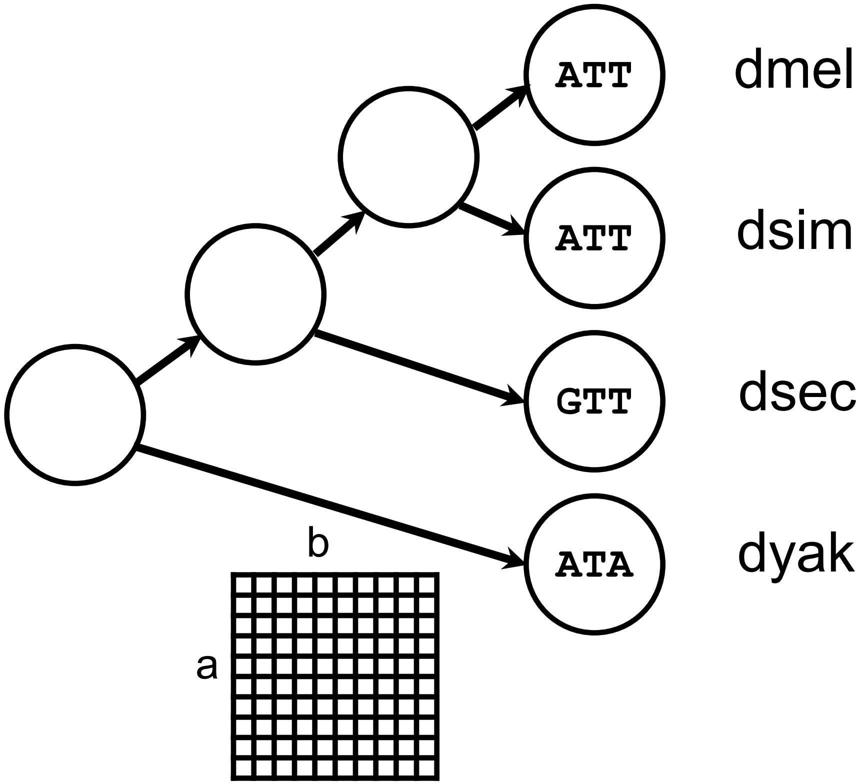
```
ancestor GTG GCG AGT GCA TTT CCC AGA GGA GTT GAT AGG AGT CTG AAA CTA CTG ATA AAT TGC TTT TTA ATT AGC ACA GAG CAG
    dmel GTG ACG AAT GCG TTT CCC AGA GGA TCG GAT GGA GGT CTG AAG CTA CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dsim GTG ACG AAT GCG TTT CCC AGA GGA TCG GAT GGA GGT CTG AAA CTA CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dsec GTG ACA AAT ACG TTT CCC AGA GGA TCG GAT GGA GGT CTG AAA CTT CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dyak GTG ACG AAT GCA TTT CCT AGT GGA TCG GAA GAA GGG CTG AAA GTA CTG ATA GAT GTC TTT TTA ACT AGC ACA GCA CAG
    dere GTG ACG AAT GCA TTT CCT AGA GGA TCG GAT GGT GGT TTG AAA GGG CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dana GTG ACG AAT GCA TTT ACT AGA CGA TCT AGC AGG TGG CGG AAA AAG CTG ATG GAT TGC TTT TTA ATT AGC ACA GAG TCG
    dpse GTG TCG ACT GCA TTT ACG CGG AGG CCC ACG AGG AGT CTC CAC GCA CTG ATA GAT TGC TTT TTA ATT AGC ACA GAG AGA
    dper GTG TCG ACT GCA TTT ACG CGG AGG CCC ACG AGG AGT CTC CAC GCA CTG ATA GAT TGC TTT TTA ATT AGC ACA GAG AGA
    dwil GTG GCG AGT GCA TTA AAA AGA AGA GTT GAG TTT AGT CGA GAG GGT CTG ATT AAT TGC TTT TTA ATT AGC ACT AGT TAA
    dvir GTG GCG AGT GCA TGT GCG GGA TGG CTT GGT CGG CAA CTG GGT TAG CTG ATA AAT TGC TTT TTA ATT AGC ATA GCG CAG
    dmoj GTG GCG ACT GCA TAT GCA GGT CGT GTT GGC CGG GCT CTC GGT CAG CTG ATG GAT GAC TTT TTA ATT AGT ATA GCG CAG
    dgri GTG GCG AGT GCA TCT GCG GGA TGT GTT GGT CAG CGA CTG CGT TGG CTG ATA AAT GGT TTT TTA ATT AGC CTA GCG CAG
```

conserved non-coding sequence

A method to distinguish these evolutionary signatures should:

- **Quantify the distinctiveness of all 64^2 possible codon substitutions**
  - Synonymous: very frequent in protein-coding sequences
  - Nonsense: much more frequent in non-coding than coding regions
- **Model the phylogenetic relationship among the species**
  - Multiple apparent substitutions may be explained by one evolutionary event
- **Tolerate uncertainty in the input**
  - Unknown ancestral sequences
  - Alignment gaps, missing data
- **Report the [un]certainty of the result**
  - Quantify confidence that given alignment is protein-coding
  - Units: p-value, bits, decibans, etc.

Codon evolution can be modeled as a Bayesian network



Conditional probability distribution (CPD) giving,
for all codons a & b, $\Pr(\mathrm{dyak} = b | \mathrm{Ancestor} = a)$

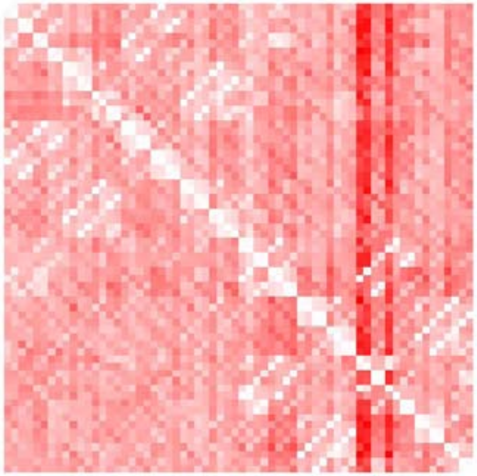
Each site (codon alignment column) is treated independently.

Given the topology and CPDs, we can simulate evolution of an ancestral sequence.

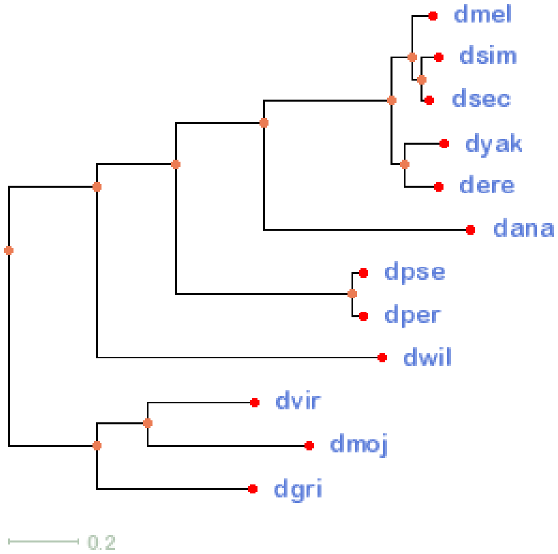
Additionally given extant (leaf) sequences, the ancestral sequences can be inferred.

For *L* leaves, CPDs total about $(2L-2) \cdot 64^2$ parameters.

The Bayes net is parameterized as a continuous-time Markov process



Rate matrix (***Q***)



Branch lengths t

Each CPD is determined by a rate matrix shared throughout the tree and a branch-specific 'time' (branch length):

$$\Pr(\text{child} = b | \text{parent} = a; t) = [e^{\mathbf{Q}t}]_{a,b}$$

Intuition: The branch lengths specify how much 'time' passed between any two nodes. The rate matrix describes the relative frequencies of codon substitutions *per unit branch length*. Synonymous substitutions have high rates and nonsense substitutions have low rates.

We can obtain maximum likelihood estimates of $(2L-2)+64^2$ parameters using EM in training data.

Example nucleotide (4x4) rate & substitution matrices

$$\mathbf{Q} = \begin{array}{cc} \begin{pmatrix} -4 & 2 & 1 & 1 \\ 2 & -4 & 1 & 1 \\ 1 & 1 & -4 & 2 \\ 1 & 1 & 2 & -4 \end{pmatrix} & \begin{array}{c} \text{A} \\ \text{G} \\ \text{C} \\ \text{T} \end{array} \\ \begin{array}{cccc} \text{A} & \text{G} & \text{C} & \text{T} \end{array} & \end{array}$$

$$\Pr(\text{child} = b | \text{parent} = a; t) = [e^{\mathbf{Q}t}]_{a,b}$$

$e^{\mathbf{Q}t} = \sum_{n=0}^{\infty} \frac{t^n}{n!}\mathbf{Q}^n$ is the solution to the system of differential equations describing the Markov process model of evolution.

```
MatrixExp[Q * 0] // MatrixForm
```

```
1  0  0  0
0  1  0  0
0  0  1  0
0  0  0  1
```

```
MatrixExp[Q * 0.001] // MatrixForm // NumberForm[#, 4] &
```

```
  0.996     0.001993  0.000998  0.000998
0.001993     0.996    0.000998  0.000998
0.000998  0.000998     0.996    0.001993
0.000998  0.000998  0.001993     0.996
```

```
MatrixExp[Q * 0.01] // MatrixForm // NumberForm[#, 4] &
```

```
 0.9611     0.01932   0.009803  0.009803
 0.01932    0.9611    0.009803  0.009803
0.009803   0.009803    0.9611    0.01932
0.009803   0.009803   0.01932    0.9611
```

```
MatrixExp[Q * 0.1] // MatrixForm // NumberForm[#, 4] &
```

```
  0.692    0.1432   0.08242  0.08242
 0.1432     0.692   0.08242  0.08242
0.08242   0.08242    0.692    0.1432
0.08242   0.08242   0.1432     0.692
```

```
MatrixExp[Q * 1.0 ] // MatrixForm // NumberForm[#, 4] &
```

```
0.2558  0.2533  0.2454  0.2454
0.2533  0.2558  0.2454  0.2454
0.2454  0.2454  0.2558  0.2533
0.2454  0.2454  0.2533  0.2558
```

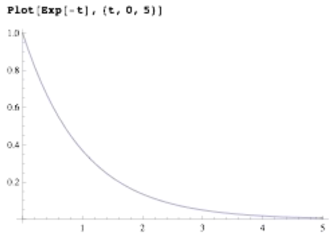
```
MatrixExp[Q * 10.0] // MatrixForm // NumberForm[#, 4] &
```

```
0.25  0.25  0.25  0.25
0.25  0.25  0.25  0.25
0.25  0.25  0.25  0.25
0.25  0.25  0.25  0.25
```

Analogy: $y(t) = e^{qt}$

solves the differential equation

$$\frac{dy}{dt} = qy$$


Plot[Exp[-t], {t, 0, 5}]

Side note: Jukes-Cantor and Kimura models are set up so that the entries of e^{Qt} have closed-form solutions.

The hairy math: how do we estimate *Q*?

- Collect many alignments of known protein-coding sequences (training data)
- Consider the probability of the training data as a function of ***Q***

$$\text{Likelihood}(\mathbf{Q}) = \Pr(\text{Training Data}; \mathbf{Q}, \underline{t})$$

Still computed using Felsenstein algorithm

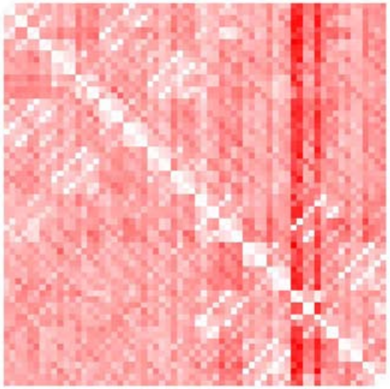
- Choose the ***Q*** that maximizes that probability:

$$\hat{\mathbf{Q}} = \underset{\mathbf{Q}}{\text{argmax}}\left(\text{Likelihood}(\mathbf{Q})\right)$$

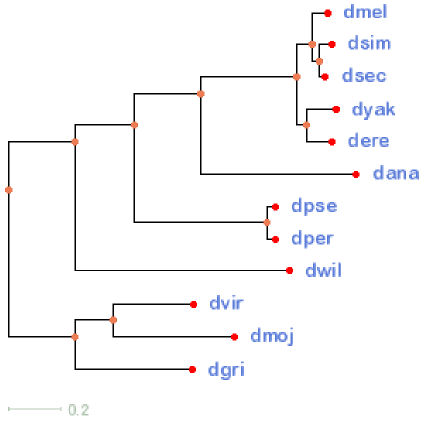
Note: ***Q*** represents thousands of parameters

- Maximization strategies: expectation-maximization; gradient ascent; simulated annealing; spectral decomposition; others
- Branch lengths can also be optimized in the same way (simultaneously)
- Non-coding model estimated similarly, with random non-coding regions as training data.

Given this generative model of codon evolution:

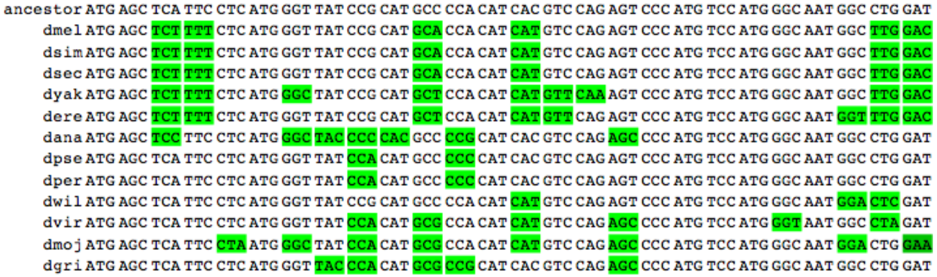


Rate matrix (***Q***)



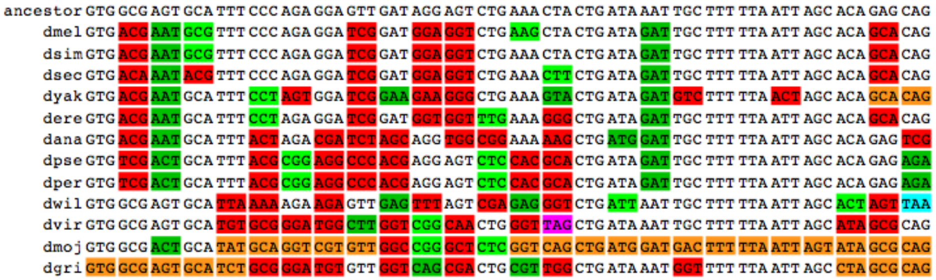
Branch lengths $\underline{t}$

We can compute the probability of any given alignment, marginalizing over all possible ancestral sequences, using Felsenstein's pruning algorithm.



```
ancestor ATG AGC TCA TTC CTC ATG GGT TAT CCG CAT GCC CCA CAT CAC GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dmel ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCA CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dsim ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCA CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dsec ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCA CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dyak ATG AGC TCT TTT CTC ATG GGC TAT CCG CAT GCT CCA CAT CAT GTT CAA AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dere ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCT CCA CAT CAT GTT CAG AGT CCC ATG TCC ATG GGC AAT GGT TTG GAC
    dana ATG AGC TCC TTC CTC ATG GGC TAC CCC CAC GCC CCG CAT CAC GTC CAG AGC CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dpse ATG AGC TCA TTC CTC ATG GGT TAT CCA CAT GCC CCC CAT CAC GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dper ATG AGC TCA TTC CTC ATG GGT TAT CCA CAT GCC CCC CAT CAC GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dwil ATG AGC TCA TTC CTC ATG GGT TAT CCG CAT GCC CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGA CTC GAT
    dvir ATG AGC TCA TTC CTC ATG GGT TAT CCA CAT GCG CCA CAT CAT GTC CAG AGC CCC ATG TCC ATG GGT AAT GGC CTA GAT
    dmoj ATG AGC TCA TTC CTA ATG GGC TAT CCA CAT GCG CCA CAT CAT GTC CAG AGC CCC ATG TCC ATG GGC AAT GGA CTG GAA
    dgri ATG AGC TCA TTC CTC ATG GGT TAC CCA CAT GCG CCG CAT CAC GTC CAG AGC CCC ATG TCC ATG GGC AAT GGC CTG GAT
```

protein-coding exon



```
ancestor GTG GCG AGT GCA TTT CCC AGA GGA GTT GAT AGG AGT CTG AAA CTA CTG ATA AAT TGC TTT TTA ATT AGC ACA GAG CAG
    dmel GTG ACG AAT GCG TTT CCC AGA GGA TCG GAT GGA GGT CTG AAG CTA CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dsim GTG ACG AAT GCG TTT CCC AGA GGA TCG GAT GGA GGT CTG AAA CTA CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dsec GTG ACA AAT ACG TTT CCC AGA GGA TCG GAT GGA GGT CTG AAA CTT CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dyak GTG ACG AAT GCA TTT CCT AGT GGA TCG GAA GAA GGG CTG AAA GTA CTG ATA GAT GTC TTT TTA ACT AGC ACA GCA CAG
    dere GTG ACG AAT GCA TTT CCT AGA GGA TCG GAT GGT GGT TTG AAA GGG CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dana GTG ACG AAT GCA TTT ACT AGA CGA TCT AGC AGG TGG CGG AAA AAG CTG ATG GAT TGC TTT TTA ATT AGC ACA GAG TCG
    dpse GTG TCG ACT GCA TTT ACG CGG AGG CCC ACG AGG AGT CTC CAC GCA CTG ATA GAT TGC TTT TTA ATT AGC ACA GAG AGA
    dper GTG TCG ACT GCA TTT ACG CGG AGG CCC ACG AGG AGT CTC CAC GCA CTG ATA GAT TGC TTT TTA ATT AGC ACA GAG AGA
    dwil GTG GCG AGT GCA TTA AAA AGA AGA GTT GAG TTT AGT CGA GAG GGT CTG ATT AAT TGC TTT TTA ATT AGC ACT AGT TAA
    dvir GTG GCG AGT GCA TGT GCG GGA TGG CTT GGT CGG CAA CTG GGT TAG CTG ATA AAT TGC TTT TTA ATT AGC ATA GCG CAG
    dmoj GTG GCG ACT GCA TAT GCA GGT CGT GTT GGC CGG GCT CTC GGT CAG CTG ATG GAT GAC TTT TTA ATT AGT ATA GCG CAG
    dgri GTG GCG AGT GCA TCT GCG GGA TGT GTT GGT CAG CGA CTG CGT TGG CTG ATA AAT GGT TTT TTA ATT AGC CTA GCG CAG
```

conserved non-coding sequence

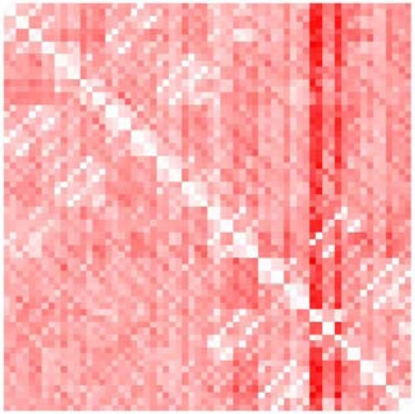
$$\Pr(\text{Leaves}; \mathbf{Q}, \underline{t}) = \frac{1}{10^{117}}$$

$$\Pr(\text{Leaves}; \mathbf{Q}, \underline{t}) = \frac{1}{10^{275}}$$

↑

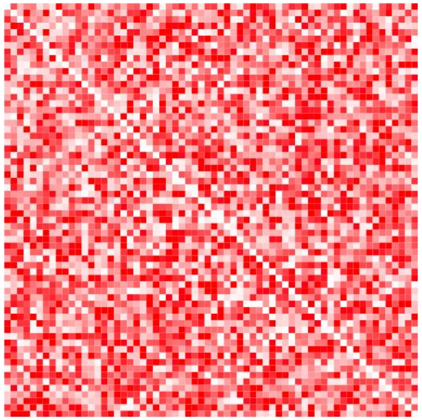
If I simulate alignments randomly according to the model, I'll get this exact alignment once every 10^{117} samples

Now suppose we've estimated two rate matrices:



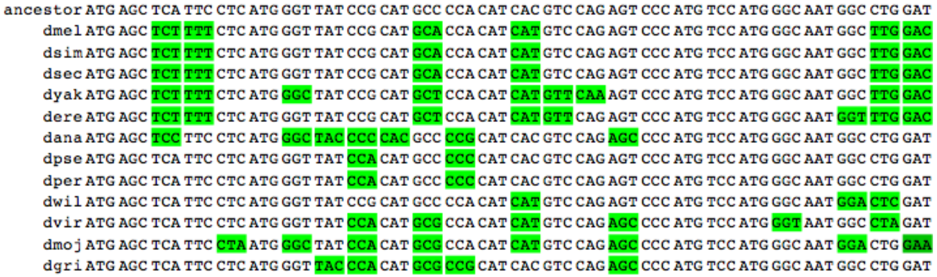
Q

$\mathbf{Q}_C$ estimated from known coding regions

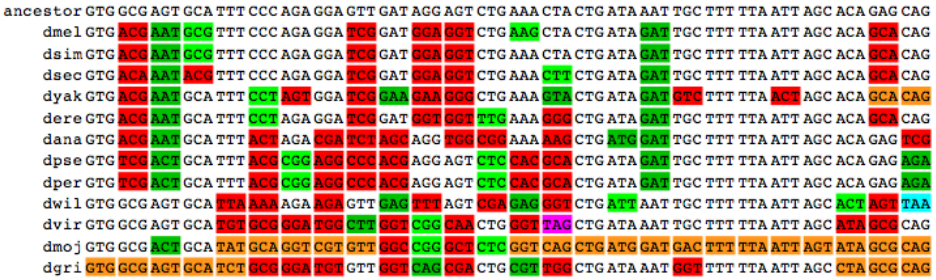


$\mathbf{Q}_N$ estimated from non-coding regions

These specify different rates of codon substitution, which in turn lead to different probabilities of any given alignment:



```
ancestor ATG AGC TCA TTC CTC ATG GGT TAT CCG CAT GCC CCA CAT CAC GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dmel ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCA CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dsim ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCA CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dsec ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCA CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dyak ATG AGC TCT TTT CTC ATG GGC TAT CCG CAT GCT CCA CAT CAT GTT CAA AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dere ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCT CCA CAT CAT GTT CAG AGT CCC ATG TCC ATG GGC AAT GGT TTG GAC
    dana ATG AGC TCC TTC CTC ATG GGC TAC CCC CAC GCC CCG CAT CAC GTC CAG AGC CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dpse ATG AGC TCA TTC CTC ATG GGT TAT CCA CAT GCC CCC CAT CAC GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dper ATG AGC TCA TTC CTC ATG GGT TAT CCA CAT GCC CCC CAT CAC GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dwil ATG AGC TCA TTC CTC ATG GGT TAT CCG CAT GCC CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGA CTC GAT
    dvir ATG AGC TCA TTC CTC ATG GGT TAT CCA CAT GCG CCA CAT CAT GTC CAG AGC CCC ATG TCC ATG GGT AAT GGC CTA GAT
    dmoj ATG AGC TCA TTC CTA ATG GGC TAT CCA CAT GCG CCA CAT CAT GTC CAG AGC CCC ATG TCC ATG GGC AAT GGA CTG GAA
    dgri ATG AGC TCA TTC CTC ATG GGT TAC CCA CAT GCG CCG CAT CAC GTC CAG AGC CCC ATG TCC ATG GGC AAT GGC CTG GAT
```



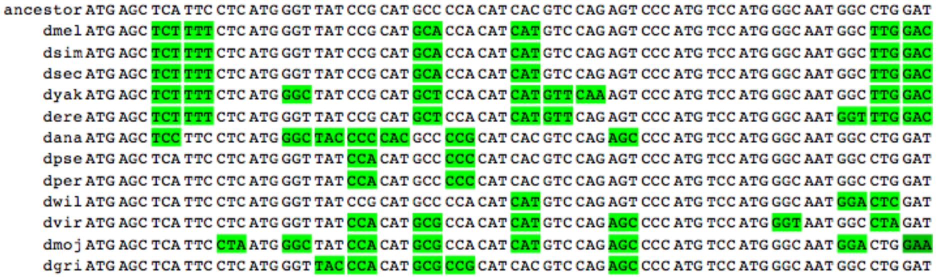
```
ancestor GTG GCG AGT GCA TTT CCC AGA GGA GTT GAT AGG AGT CTG AAA CTA CTG ATA AAT TGC TTT TTA ATT AGC ACA GAG CAG
    dmel GTG ACG AAT GCG TTT CCC AGA GGA TCG GAT GGA GGT CTG AAG CTA CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dsim GTG ACG AAT GCG TTT CCC AGA GGA TCG GAT GGA GGT CTG AAA CTA CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dsec GTG ACA AAT ACG TTT CCC AGA GGA TCG GAT GGA GGT CTG AAA CTT CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dyak GTG ACG AAT GCA TTT CCT AGT GGA TCG GAA GAA GGG CTG AAA GTA CTG ATA GAT GTC TTT TTA ACT AGC ACA GCA CAG
    dere GTG ACG AAT GCA TTT CCT AGA GGA TCG GAT GGT GGT TTG AAA GGG CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dana GTG ACG AAT GCA TTT ACT AGA CGA TCT AGC AGG TGG CGG AAA AAG CTG ATG GAT TGC TTT TTA ATT AGC ACA GAG TCG
    dpse GTG TCG ACT GCA TTT ACG CGG AGG CCC ACG AGG AGT CTC CAC GCA CTG ATA GAT TGC TTT TTA ATT AGC ACA GAG AGA
    dper GTG TCG ACT GCA TTT ACG CGG AGG CCC ACG AGG AGT CTC CAC GCA CTG ATA GAT TGC TTT TTA ATT AGC ACA GAG AGA
    dwil GTG GCG AGT GCA TTA AAA AGA AGA GTT GAG TTT AGT CGA GAG GGT CTG ATT AAT TGC TTT TTA ATT AGC ACT AGT TAA
    dvir GTG GCG AGT GCA TGT GCG GGA TGG CTT GGT CGG CAA CTG GGT TAG CTG ATA AAT TGC TTT TTA ATT AGC ATA GCG CAG
    dmoj GTG GCG ACT GCA TAT GCA GGT CGT GTT GGC CGG GCT CTC GGT CAG CTG ATG GAT GAC TTT TTA ATT AGT ATA GCG CAG
    dgri GTG GCG AGT GCA TCT GCG GGA TGT GTT GGT CAG CGA CTG CGT TGG CTG ATA AAT GGT TTT TTA ATT AGC CTA GCG CAG
```

$$\Pr(\text{Leaves}; \mathbf{Q}_C, \underline{t}) = \frac{1}{10^{117}}$$

$$\Pr(\text{Leaves}; \mathbf{Q}_N, \underline{t}) = \frac{1}{10^{152}}$$

$$\Pr(\text{Leaves}; \mathbf{Q}_C, \underline{t}) = \frac{1}{10^{275}}$$

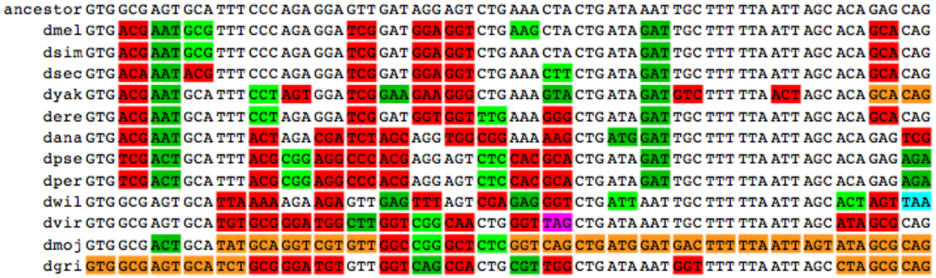
$$\Pr(\text{Leaves}; \mathbf{Q}_N, \underline{t}) = \frac{1}{10^{254}}$$



```
ancestor ATG AGC TCA TTC CTC ATG GGT TAT CCG CAT GCC CCA CAT CAC GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dmel ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCA CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dsim ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCA CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dsec ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCA CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dyak ATG AGC TCT TTT CTC ATG GGC TAT CCG CAT GCT CCA CAT CAT GTT CAA AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dere ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCT CCA CAT CAT GTT CAG AGT CCC ATG TCC ATG GGC AAT GGT TTG GAC
    dana ATG AGC TCC TTC CTC ATG GGC TAC CCC CAC GCC CCG CAT CAC GTC CAG AGC CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dpse ATG AGC TCA TTC CTC ATG GGT TAT CCA CAT GCC CCC CAT CAC GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dper ATG AGC TCA TTC CTC ATG GGT TAT CCA CAT GCC CCC CAT CAC GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dwil ATG AGC TCA TTC CTC ATG GGT TAT CCG CAT GCC CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGA CTC GAT
    dvir ATG AGC TCA TTC CTC ATG GGT TAT CCA CAT GCG CCA CAT CAT GTC CAG AGC CCC ATG TCC ATG GGT AAT GGC CTA GAT
    dmoj ATG AGC TCA TTC CTA ATG GGC TAT CCA CAT GCG CCA CAT CAT GTC CAG AGC CCC ATG TCC ATG GGC AAT GGA CTG GAA
    dgri ATG AGC TCA TTC CTC ATG GGT TAC CCA CAT GCG CCG CAT CAC GTC CAG AGC CCC ATG TCC ATG GGC AAT GGC CTG GAT
```

$$\frac{\Pr(\text{Leaves}; \mathbf{Q}_C, \underline{t}) = \dfrac{1}{10^{117}}}{\Pr(\text{Leaves}; \mathbf{Q}_N, \underline{t}) = \dfrac{1}{10^{152}}} = 10^{35}$$

This alignment is 10^{35} times more probable under the coding model than the non-coding model.



```
ancestor GTG GCG AGT GCA TTT CCC AGA GGA GTT GAT AGG AGT CTG AAA CTA CTG ATA AAT TGC TTT TTA ATT AGC ACA GAG CAG
    dmel GTG ACG AAT GCG TTT CCC AGA GGA TCG GAT GGA GGT CTG AAG CTA CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dsim GTG ACG AAT GCG TTT CCC AGA GGA TCG GAT GGA GGT CTG AAA CTA CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dsec GTG ACA AAT ACG TTT CCC AGA GGA TCG GAT GGA GGT CTG AAA CTT CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dyak GTG ACG AAT GCA TTT CCT AGT GGA TCG GAA GAA GGG CTG AAA GTA CTG ATA GAT GTC TTT TTA ACT AGC ACA GCA CAG
    dere GTG ACG AAT GCA TTT CCT AGA GGA TCG GAT GGT GGT TTG AAA GGG CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dana GTG ACG AAT GCA TTT ACT AGA CGA TCT AGC AGG TGG CGG AAA AAG CTG ATG GAT TGC TTT TTA ATT AGC ACA GAG TCG
    dpse GTG TCG ACT GCA TTT ACG CGG AGG CCC ACG AGG AGT CTC CAC GCA CTG ATA GAT TGC TTT TTA ATT AGC ACA GAG AGA
    dper GTG TCG ACT GCA TTT ACG CGG AGG CCC ACG AGG AGT CTC CAC GCA CTG ATA GAT TGC TTT TTA ATT AGC ACA GAG AGA
    dwil GTG GCG AGT GCA TTA AAA AGA AGA GTT GAG TTT AGT CGA GAG GGT CTG ATT AAT TGC TTT TTA ATT AGC ACT AGT TAA
    dvir GTG GCG AGT GCA TGT GCG GGA TGG CTT GGT CGG CAA CTG GGT TAG CTG ATA AAT TGC TTT TTA ATT AGC ATA GCG CAG
    dmoj GTG GCG ACT GCA TAT GCA GGT CGT GTT GGC CGG GCT CTC GGT CAG CTG ATG GAT GAC TTT TTA ATT AGT ATA GCG CAG
    dgri GTG GCG AGT GCA TCT GCG GGA TGT GTT GGT CAG CGA CTG CGT TGG CTG ATA AAT GGT TTT TTA ATT AGC CTA GCG CAG
```

$$\frac{\Pr(\text{Leaves}; \mathbf{Q}_C, \underline{t}) = \dfrac{1}{10^{275}}}{\Pr(\text{Leaves}; \mathbf{Q}_N, \underline{t}) = \dfrac{1}{10^{254}}} = 10^{-21}$$

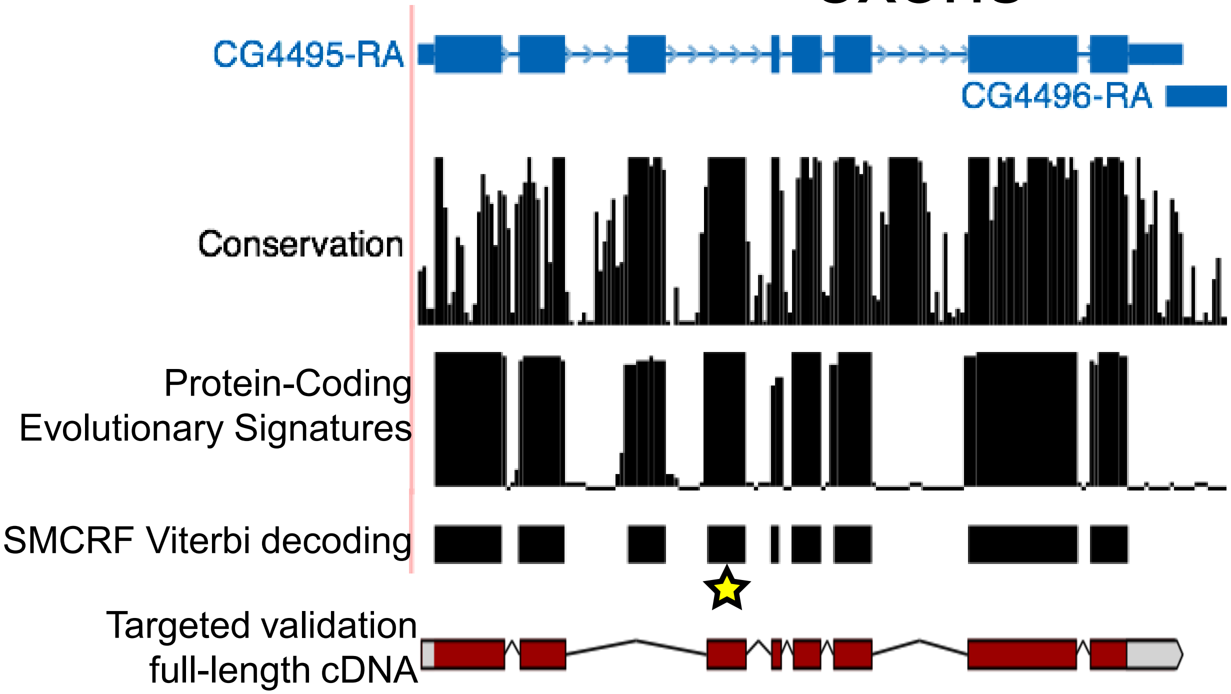
This alignment is 10^{21} times less probable under the coding model than the non-coding model.

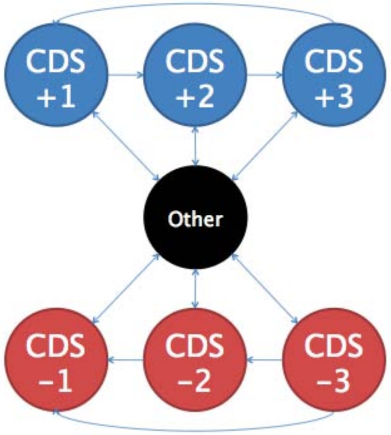
This **likelihood ratio** $\dfrac{\Pr(\text{Leaves}; \mathbf{Q}_C, \underline{t})}{\Pr(\text{Leaves}; \mathbf{Q}_N, \underline{t})}$ is our measure of confidence that the alignment is protein-coding.

Comparative genomics I: Evolutionary signatures

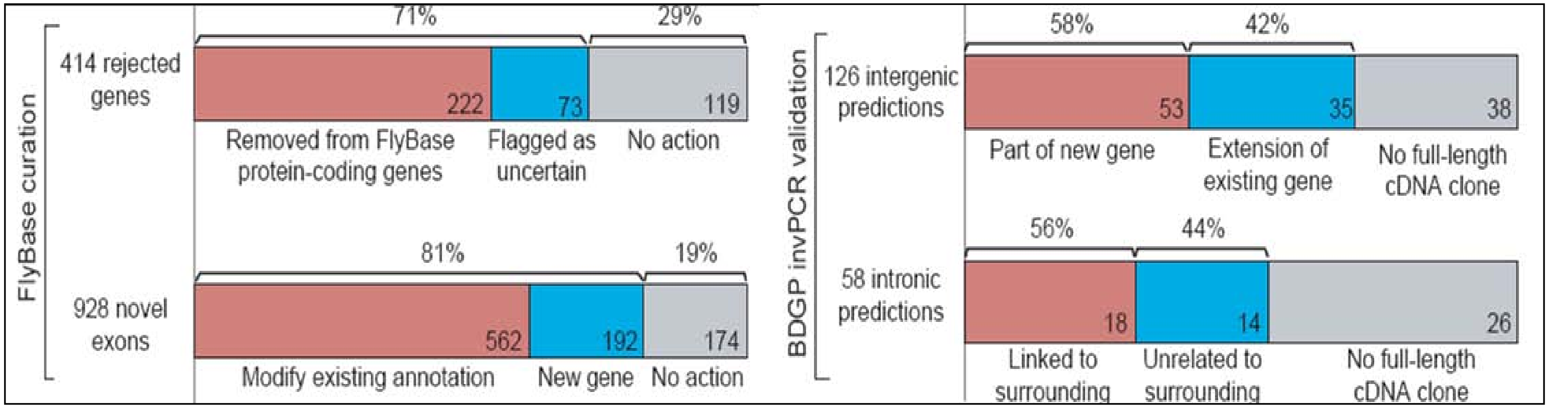
- **Nucleotide conservation: evolutionary constraint**
  - Purifying selection, neutral branch length, discovery power
  - Detect constrained elements: nucleotides, windows, HMM
  - Estimate fraction constrained: signal vs. background
- **Evolutionary signatures: focus on pattern of change**
  - Different functions ⇔ Characteristic patterns of evolution
- **Signatures of protein-coding genes**
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_C Q_N$, scoring
  - Revise genes, read-through, excess constraint regions
- **Signatures of microRNA genes**
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation

Evolutionary signatures can predict new genes and exons





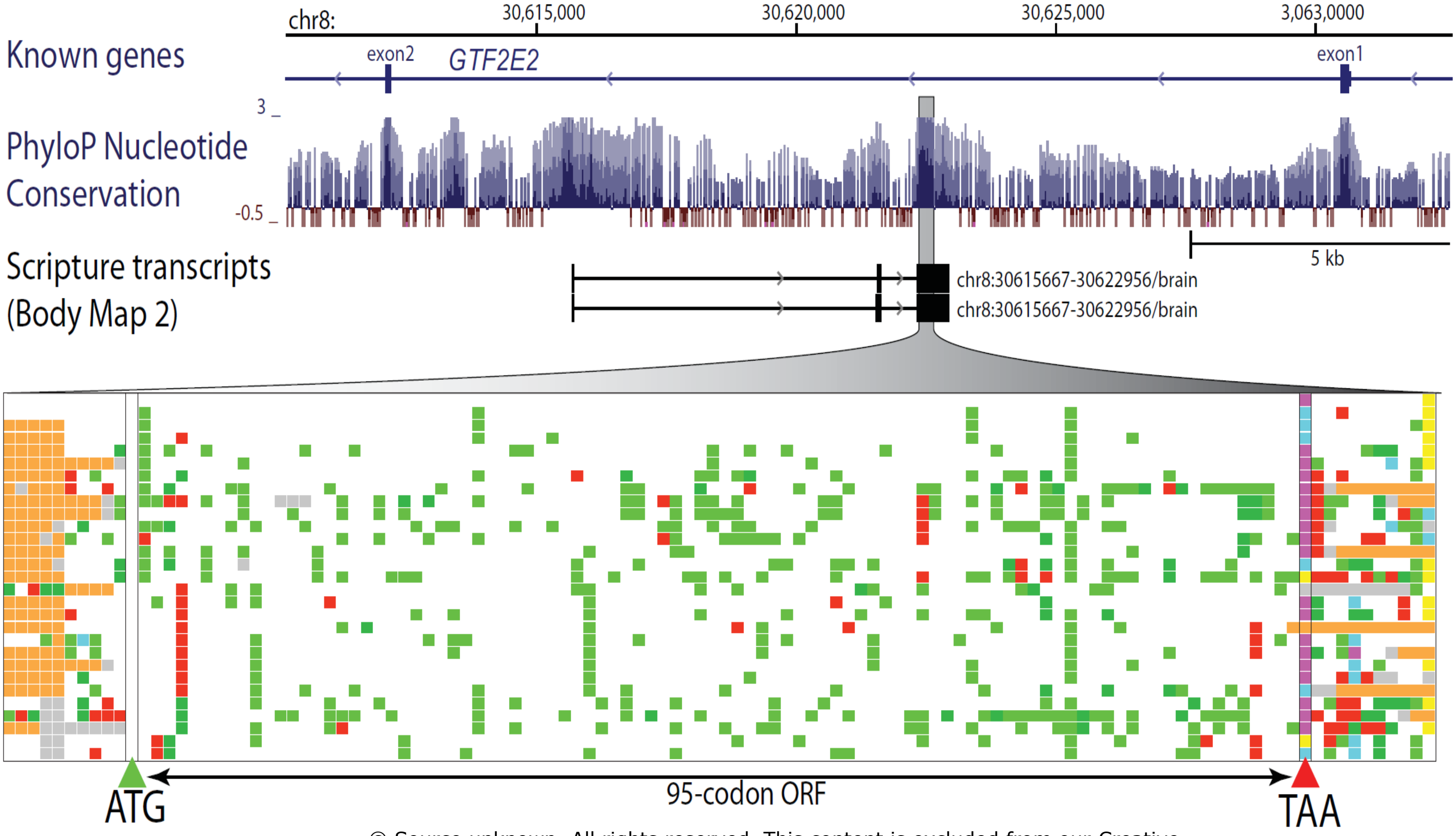
Evolutionary signatures built into a semi-Markov conditional random field to predict protein-coding exons



Courtesy of Macmillan Publishers Limited. Used with permission.

Source: Stark, Alexander et al. "Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures." Nature 450, no. 7167 (2007): 219-232.

New protein-coding genes

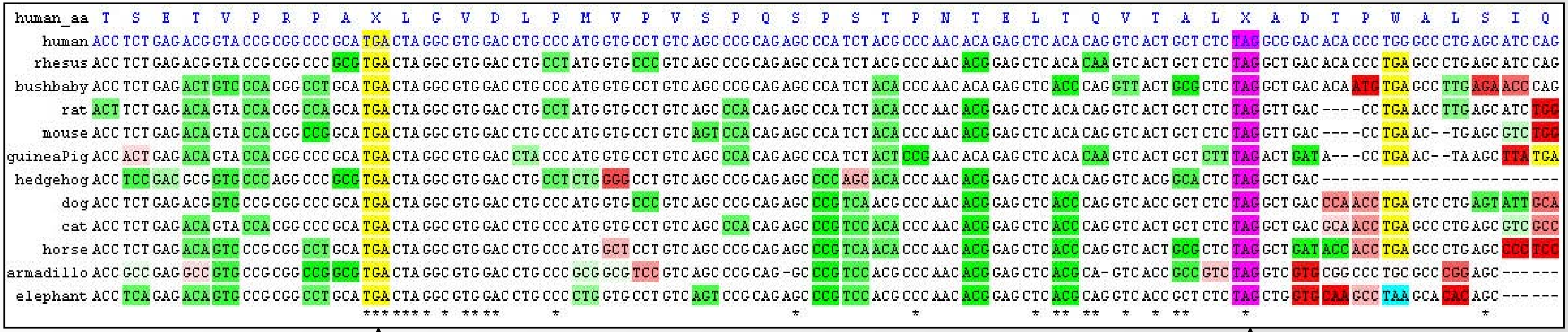

chr8:
30,615,000
30,620,000
30,625,000
3,063,0000
Known genes
exon2
GTF2E2
exon1
3
PhyloP Nucleotide Conservation
-0.5
5 kb
Scripture transcripts (Body Map 2)
chr8:30615667-30622956/brain
chr8:30615667-30622956/brain
ATG
95-codon ORF
TAA

New genes supported by Illumina BodyAtlas transcripts

Submitted to GENCODE for validation / manual curation

Translational read-through in flies and mammals

One of four novel candidates in the human genome: OPRL1 neurotransmitter



```
 human_aa  T   S   E   T   V   P   R   P   A   X   L   G   V   D   L   P   M   V   P   V   S   P   Q   S   P   S   T   P   N   T   E   L   T   Q   V   T   A   L   X   A   D   T   P   W   A   L   S   I   Q
    human ACC TCT GAG ACG GTA CCG CGG CCC GCA TGA CTA GGC GTG GAC CTG CCC ATG GTG CCT GTC AGC CCG CAG AGC CCA TCT ACG CCC AAC ACA GAG CTC ACA CAG GTC ACT GCT CTC TAG GCG GAC ACA CCC TGG GCC CTG AGC ATC CAG
   rhesus ACC TCT GAG ACG GTA CCG CGG CCC GCG TGA CTA GGC GTG GAC CTG CCT ATG GTG CCC GTC AGC CCG CAG AGC CCA TCT ACG CCC AAC ACG GAG CTC ACA CAA GTC ACT GCT CTC TAG GCT GAC ACA CCC TGA GCC CTG AGC ATC CAG
 bushbaby ACC TCT GAG ACT GTC CCA CGG CCT GCA TGA CTA GGC GTG GAC CTG CCC ATG GTG CCT GTC AGC CCG CAG AGC CCA TCT ACA CCC AAC ACA GAG CTC ACC CAG GTT ACT GCG CTC TAG GCT GAC ACA ATG TGA GCC TTG AGA ACC CAG
      rat ACT TCT GAG ACA GTA CCA CGG CCA GCA TGA CTA GGC GTG GAC CTG CCT ATG GTG CCT GTC AGC CCA CAG AGC CCA TCT ACA CCC AAC ACG GAG CTC ACA CAG GTC ACT GCT CTC TAG GTT GAC --- -CC TGA ACC TTG AGC ATC TGG
    mouse ACC TCT GAG ACA GTA CCA CGG CCG GCA TGA CTA GGC GTG GAC CTG CCC ATG GTG CCT GTC AGT CCA CAG AGC CCA TCT ACA CCC AAC ACG GAG CTC ACA CAG GTC ACT GCT CTC TAG GTT GAC --- -CC TGA AC- -TG AGC GTC TGG
guineaPig ACC ACT GAG ACA GTA CCA CGG CCC GCA TGA CTA GGC GTG GAC CTA CCC ATG GTG CCT GTC AGC CCA CAG AGC CCA TCT ACT CCG AAC ACA GAG CTC ACA CAA GTC ACT GCT CTT TAG ACT GAT A-- -CC TGA AC- -TA AGC TTA TGA
 hedgehog ACC TCC GAC GCG GTG CCC AGG CCC GCG TGA CTA GGC GTG GAC CTG CCT CTG GGG CCT GTC AGC CCG CAG AGC CCC AGC ACA CCC AAC ACG GAG CTC ACA CAG GTC ACG GCA CTC TAG GCT GAC --- --- --- --- --- --- --- ---
      dog ACC TCT GAG ACG GTG CCG CGG CCC GCA TGA CTA GGC GTG GAC CTG CCC ATG GTG CCC GTC AGC CCG CAG AGC CCG TCA ACG CCC AAC ACG GAG CTC ACC CAG GTC ACC GCT CTC TAG GCT GAC CCA ACC TGA GTC CTG AGT ATT GCA
      cat ACC TCT GAG ACA GTA CCA CGG CCC GCA TGA CTA GGC GTG GAC CTG CCC ATG GTG CCT GTC AGC CCA CAG AGC CCG TCC ACA CCC AAC ACG GAG CTC ACC CAG GTC ACT GCT CTC TAG GCT GAC GCA ACC TGA GCC CTG AGC GTC GCC
    horse ACC TCT GAG ACA GTC CCG CGG CCT GCA TGA CTA GGC GTG GAC CTG CCC ATG GCT CCT GTC AGC CCG CAG AGC CCG TCA ACA CCC AAC ACG GAG CTC ACC CAG GTC ACT GCG CTC TAG GCT GAT ACC ACC TGA GCC CTG AGC CCC TCC
armadillo ACC GCC GAG GCC GTG CCG CGG CCG GCG TGA CTA GGC GTG GAC CTG CCC GCG GCG TCC GTC AGC CCG CAG -GC CCG TCC ACG CCC AAC ACG GAG CTC ACG CA- GTC ACC GCC GTC TAG GTC GTG CGG CCC TGC GCC CGG AGC --- ---
 elephant ACC TCA GAG ACA GTG CCG CGG CCT GCA TGA CTA GGC GTG GAC CTG CCC CTG GTG CCT GTC AGT CCG CAG AGC CCG TCC ACG CCC AAC ACG GAG CTC ACG CAG GTC ACC GCT CTC TAG CTG GTG CAA GCC TAA GCA CAC AGC --- ---
                                                ***  *** * *  **  **      *                               *               *               *  **  **   *   *  **       *                               *
```

Protein-coding conservation

Stop codon read through

Continued protein-coding conservation

2nd stop codon

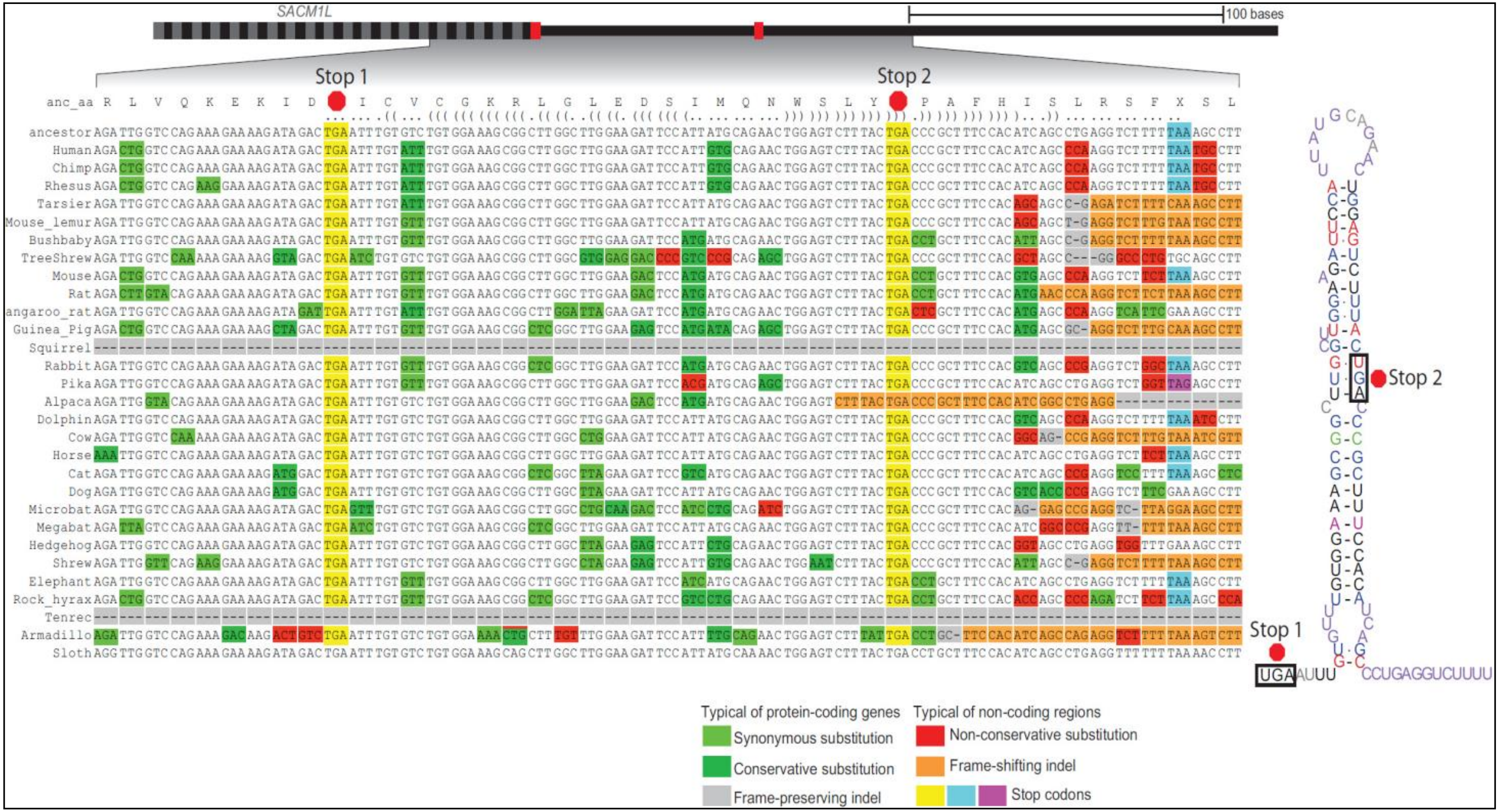
No more conservation

- **New mechanism of post-transcriptional regulation?**
  - Conserved in both mammals (4 candidates) and flies (350 candidates)
  - Strongly enriched for neurotransmitters, brain-expressed proteins, TF regulators
  - After correcting for gene length: TF enrichment remains
- **Evidence suggestive of regulatory control**
  - Read-through stop codon perfectly conserved in 93% of cases (24% at bkgrnd)
  - Upstream bases show increased conservation. Downstream is TGAC.
  - GCA triplet repeats
  - Increased RNA secondary structure

Lin *et al*, Genome Research 2007

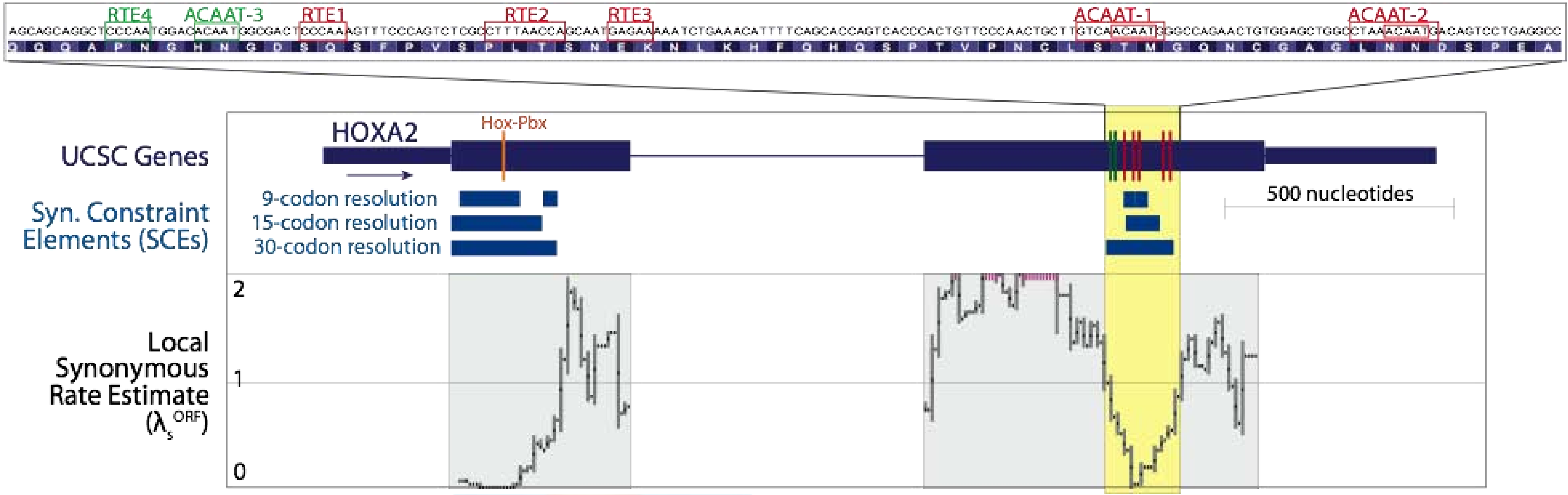
Jungreis *et al*, in preparation

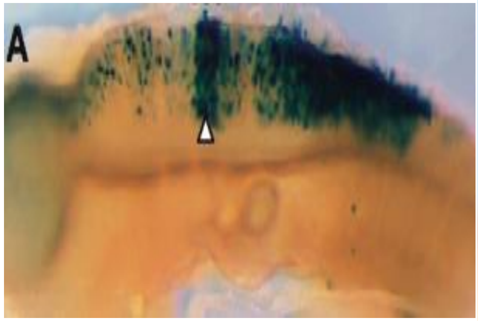
Discover of translational readthrough genes



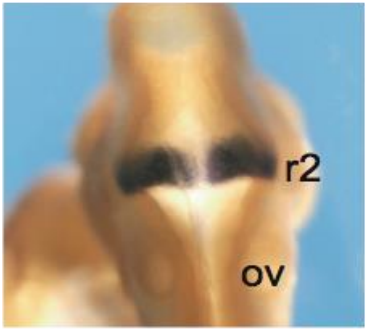
Discovery of 4 readthrough genes, abundant in many animal genomes

Overlapping selection in protein-coding exons

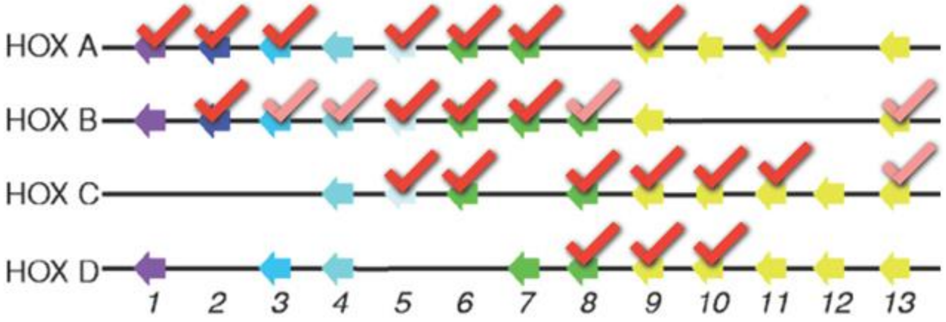




rhombomere 4 expression
(Lampe *et al*., NAR 2008)



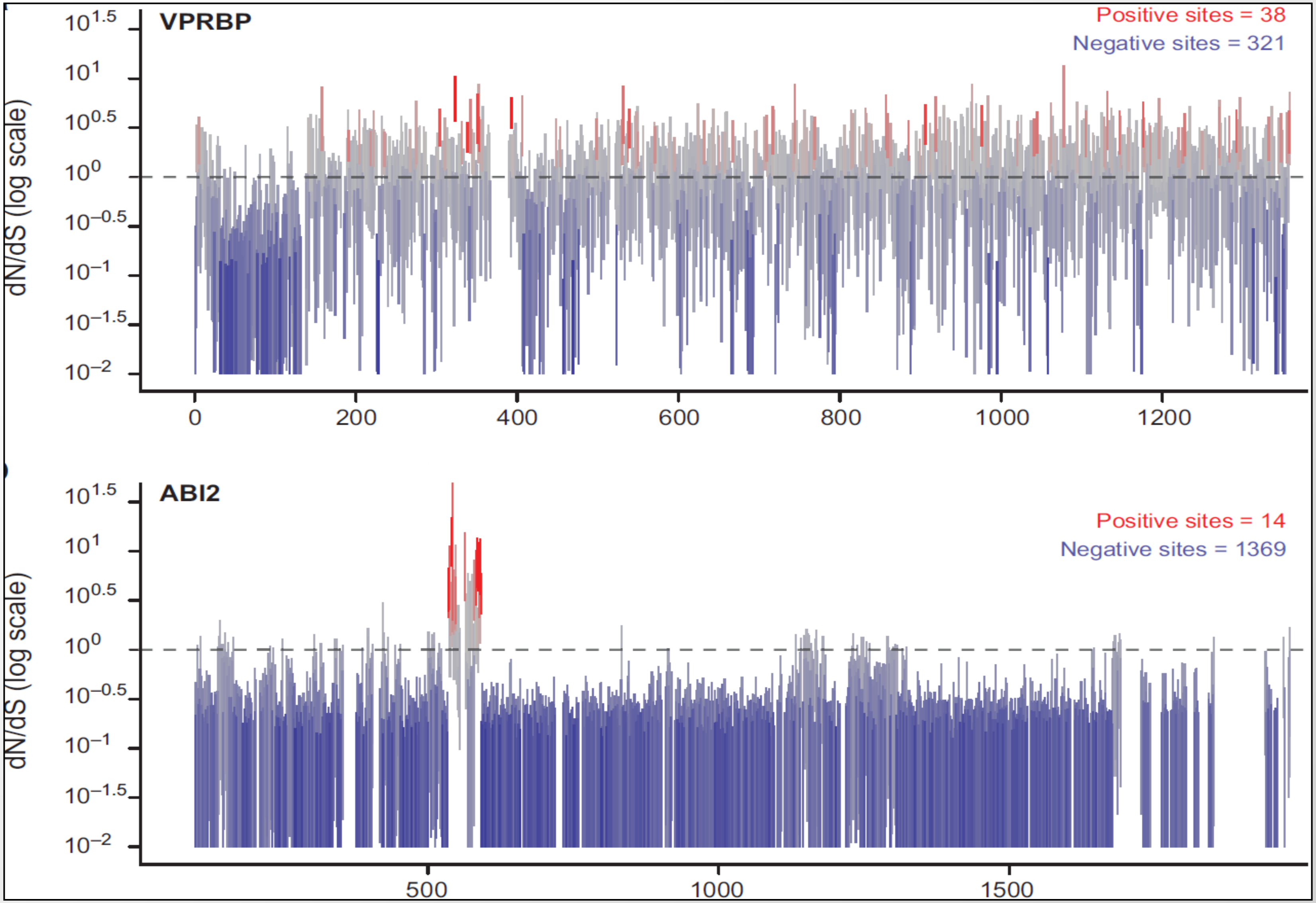
rhombomere 2 expr.
(Tümpel PNAS 2008)



10,000 overlapping synonymous constrained elements

Roles in splicing, translation, regulation

Codon-specific measures of positive selection



Gene-wide vs. punctate regions of exons positive selection

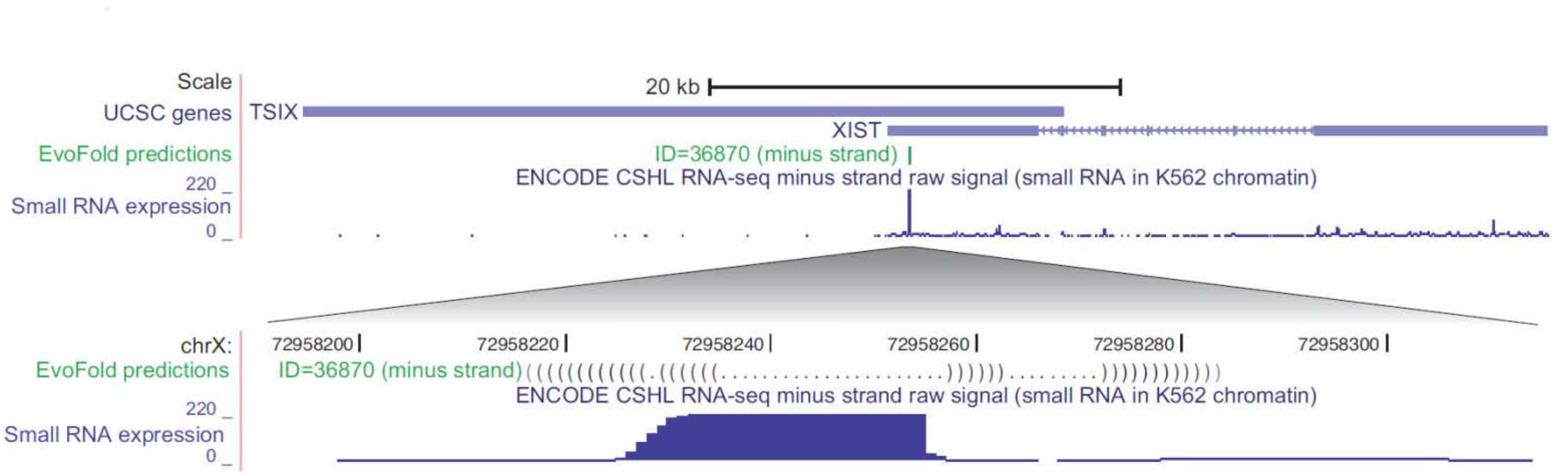
Comparative genomics I: Evolutionary signatures

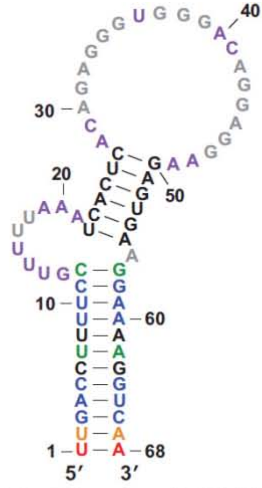
- **Nucleotide conservation: evolutionary constraint**
  - Purifying selection, neutral branch length, discovery power
  - Detect constrained elements: nucleotides, windows, HMM
  - Estimate fraction constrained: signal vs. background
- **Evolutionary signatures: focus on pattern of change**
  - Different functions ⇔ Characteristic patterns of evolution
- **Signatures of protein-coding genes**
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_C Q_N$, scoring
  - Revise genes, read-through, excess constraint regions
- **Signatures of microRNA genes**
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation

New RNA structures and families

	No. of structures	No. of novel structures	No. of families	No. of novel families	EvoFold score	RNAz overlap enrichment (x)	DNAse hypersensitivity overlap (%)	Avg. correlation of tissue-specific expression within families
EvoFold all (no CDS)	27,012	26,643	n/a	n/a	14	13.5	25 ($P \leq$ 5e–3)	n/a
Unfiltered families	3293	3081	1254	1215	18	17.3	25 ($P \leq$ 7e–3)	0.14 ($P \leq$ 1e–3)
Filtered families	725	526	220	181	18	29	32 ($P \leq$ 4e–3)	0.15 ($P \leq$ 1e–3)

New structs fall in families, supported by evolut/energy

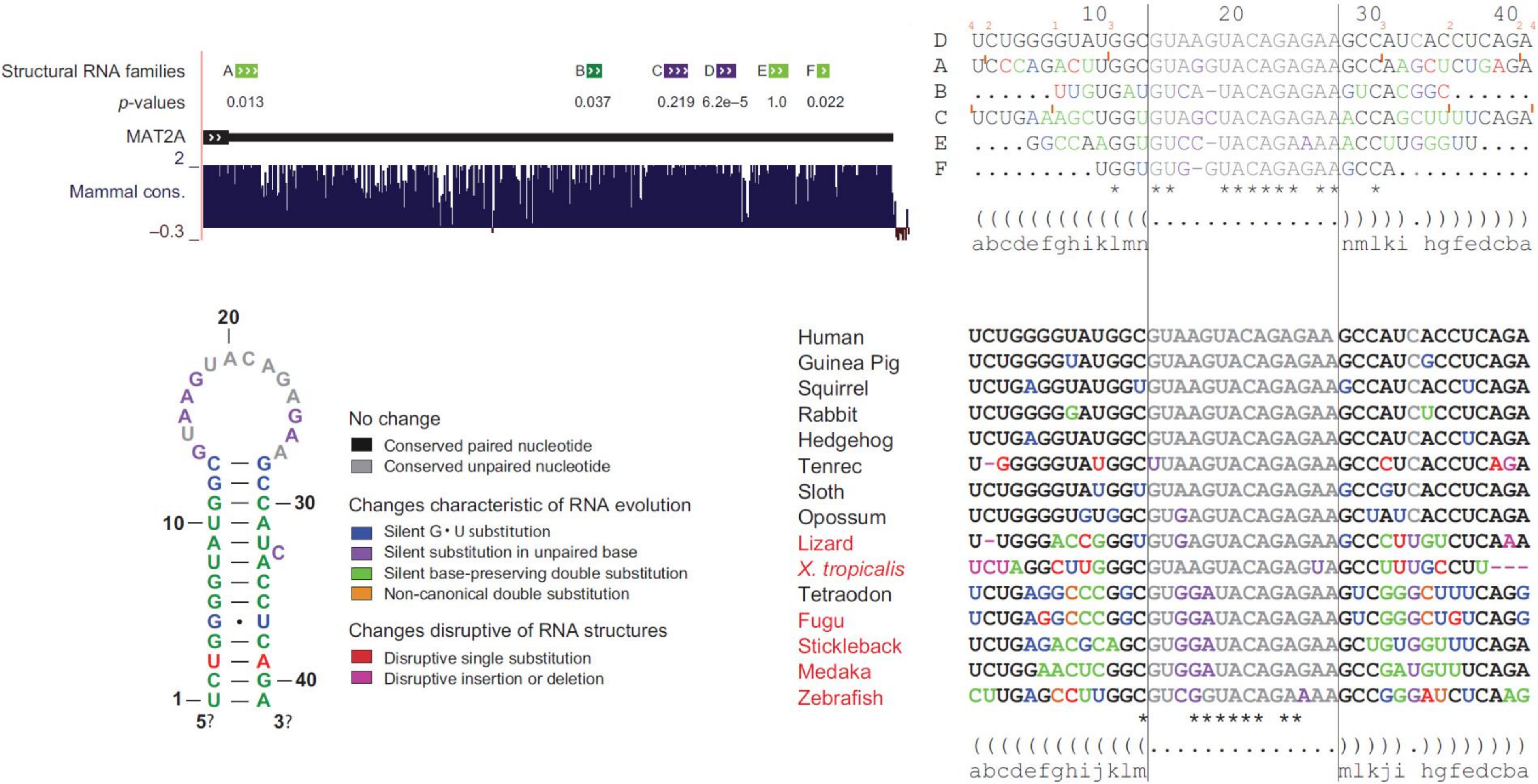

Scale
20 kb
UCSC genes
TSIX
XIST
EvoFold predictions
ID=36870 (minus strand)
ENCODE CSHL RNA-seq minus strand raw signal (small RNA in K562 chromatin)
220
Small RNA expression
0
chrX:
72958200
72958220
72958240
72958260
72958280
72958300
EvoFold predictions
ID=36870 (minus strand) ((((((((((((.((((((......................)))))).........))))))))))))
ENCODE CSHL RNA-seq minus strand raw signal (small RNA in K562 chromatin)
220
Small RNA expression
0



XIST structure (36870)

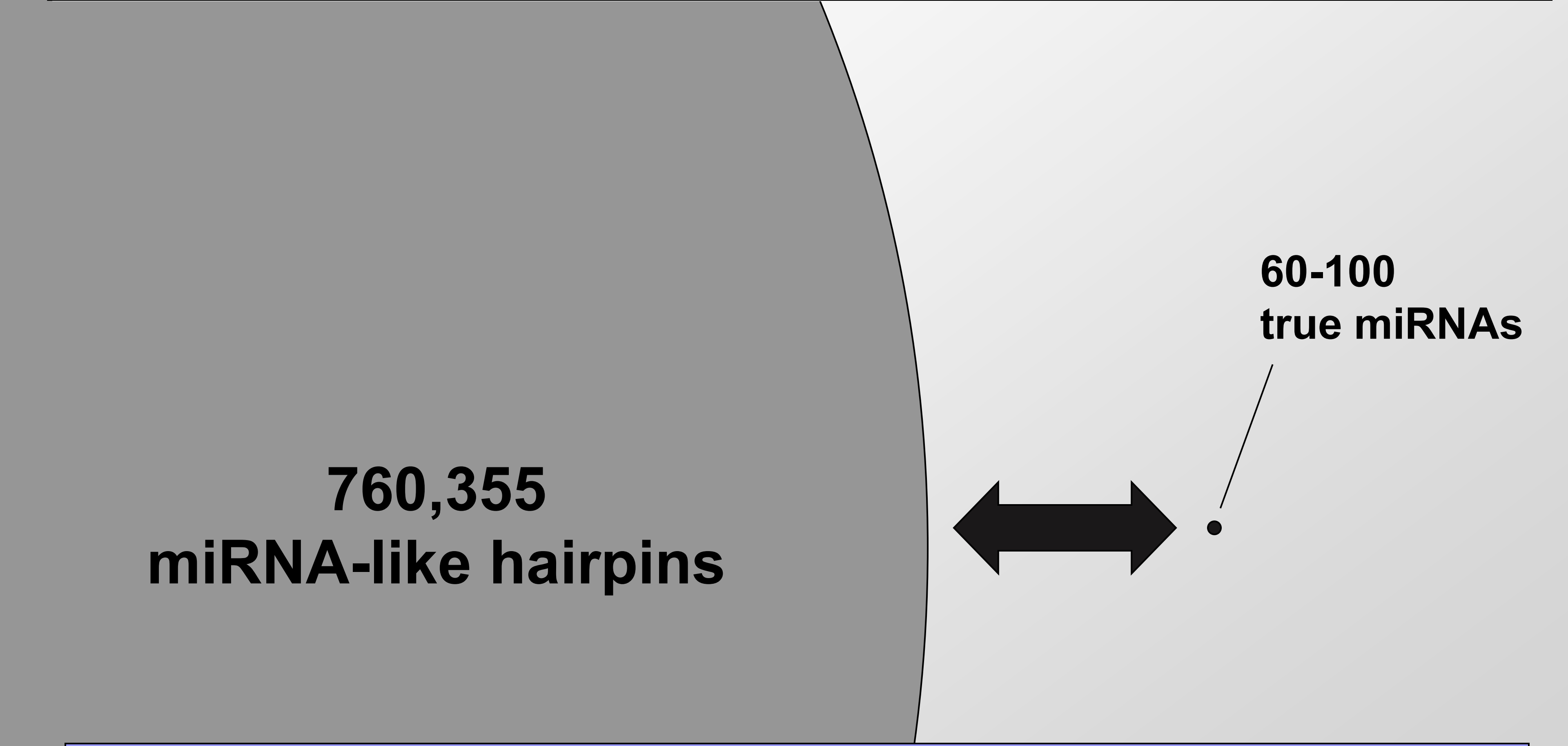
Ex: new struct in XIST long non-coding RNA
Known function in X-chromosome inactivation
Possible functional domain of XIST?

RNA families: orthologous/paralogous conservation


Structural RNA families
A B C D E F
p-values
0.013
0.037
0.219
6.2e–5
1.0
0.022
MAT2A
2
Mammal cons.
–0.3
10 20 30 40
D UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
A UCCCAGACUUGGCGUAGGUACAGAGAAGCCAAGCUCUGAGA
B ......UUGUGAUGUCA-UACAGAGAAGUCACGGC......
C UCUGAAAGCUGGUGUAGCUACAGAGAAACCAGCUUUUCAGA
E ....GGCCAAGGUGUCC-UACAGAAAAACCUUGGUU....
F ....UGGUGUG-GUACAGAGAAGCCA.........
* ** ****** ** *
((((((((((((.............))))).))))))))
abcdefghiklmn nmlki hgfedcba
20
10
30
40
1
5?
3?
No change
Conserved paired nucleotide
Conserved unpaired nucleotide
Changes characteristic of RNA evolution
Silent G • U substitution
Silent substitution in unpaired base
Silent base-preserving double substitution
Non-canonical double substitution
Changes disruptive of RNA structures
Disruptive single substitution
Disruptive insertion or deletion
Human UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Guinea Pig UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCGCCUCAGA
Squirrel UCUGAGGUAUGGUGUAAGUACAGAGAAGCCAUCACCUCAGA
Rabbit UCUGGGGGAUGGCGUAAGUACAGAGAAGCCAUCUCCUCAGA
Hedgehog UCUGAGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Tenrec U-GGGGGUAUGGCUUAAGUACAGAGAAGCCCUCACCUCAGA
Sloth UCUGGGGUAUGGUGUAAGUACAGAGAAGCCGUCACCUCAGA
Opossum UCUGGGGUGUGGCGUGAGUACAGAGAAGCUAUCACCUCAGA
Lizard U-UGGGACCGGGUGUGAGUACAGAGAAGCCCUUGUCUCAAA
X. tropicalis UCUAGGCUUGGGCGUAAGUACAGAGUAGCCUUUGCCUU---
Tetraodon UCUGAGGCCCGGCGUGGAUACAGAGAAGUCGGGCUUUCAGG
Fugu UCUGAGGCCCGGCGUGGAUACAGAGAAGUCGGGCUGUCAGG
Stickleback UCUGAGACGCAGCGUGGAUACAGAGAAGCUGUGGUUUCAGA
Medaka UCUGGAACUCGGCGUGGAUACAGAGAAGCCGAUGUUUCAGA
Zebrafish CUUGAGCCUUGGCGUCGGUACAGAAAAGCCGGGAUCUCAAG
* ****** **
((((((((((((.............))))).))))))))
abcdefghijklm mlkji hgfedcba

Example of new structural 3’UTR family in MAT2A gene likely role in detecting S-adeosyl-methionic (SAM) level

Computational challenge of miRNA discovery



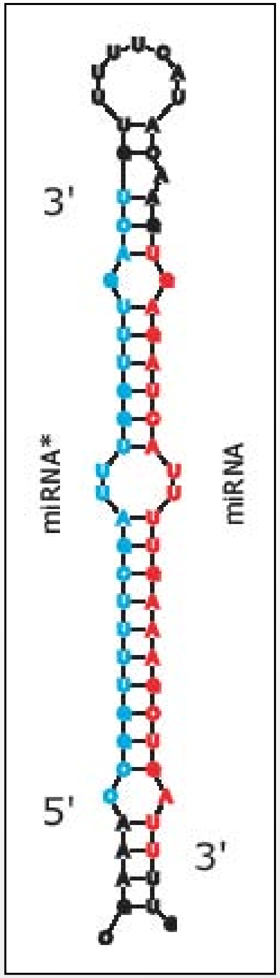
760,355
miRNA-like hairpins

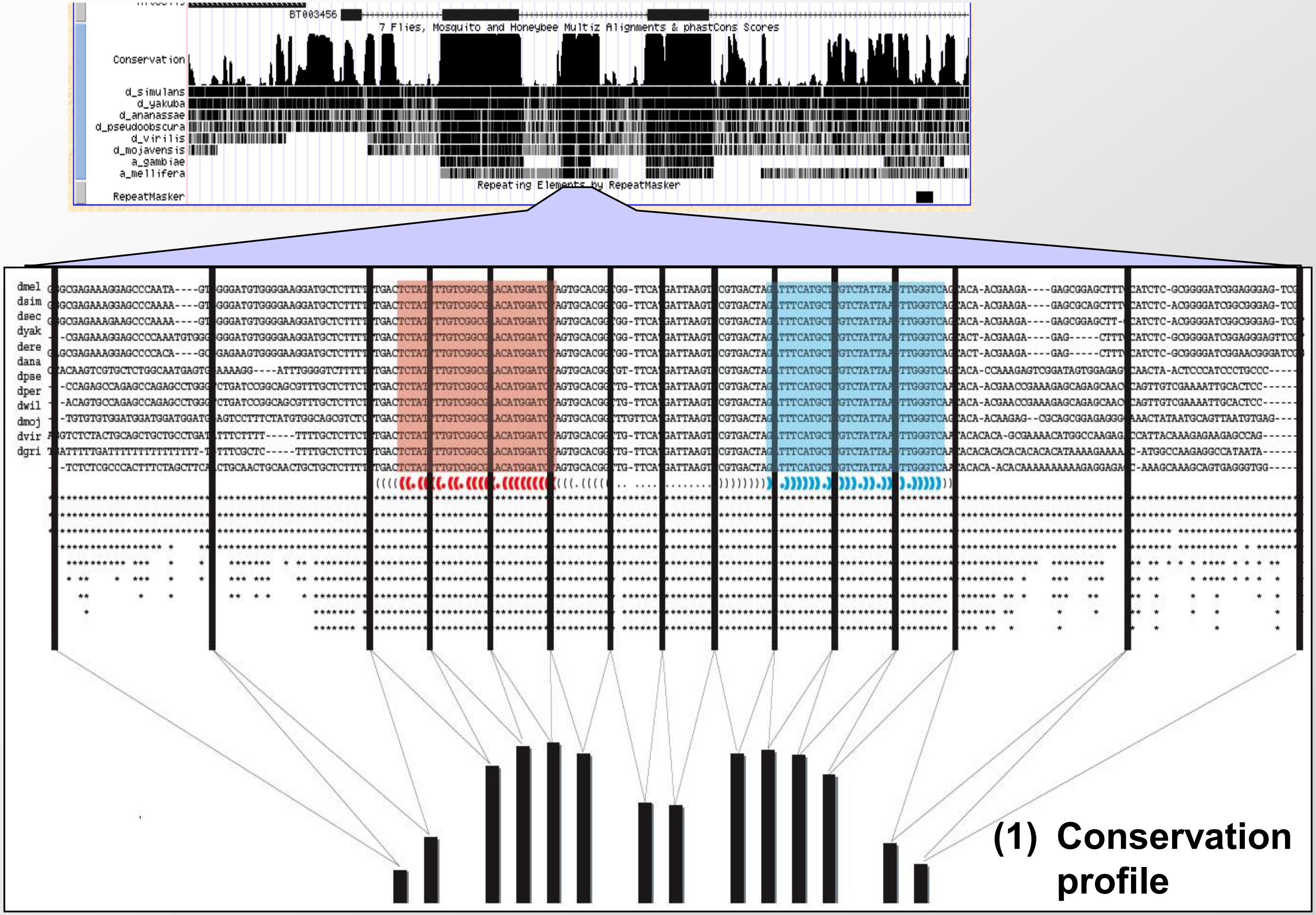
60-100
true miRNAs

A false positive rate of 0.5% ➔ 3800 spurious hairpins.

Need 99.99% specificity (>5,000-fold enrichment)

Evolutionary signatures for microRNA genes



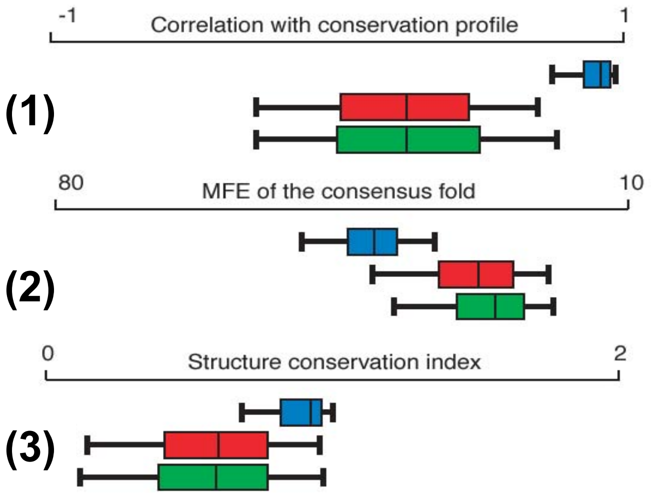


(1) Conservation profile

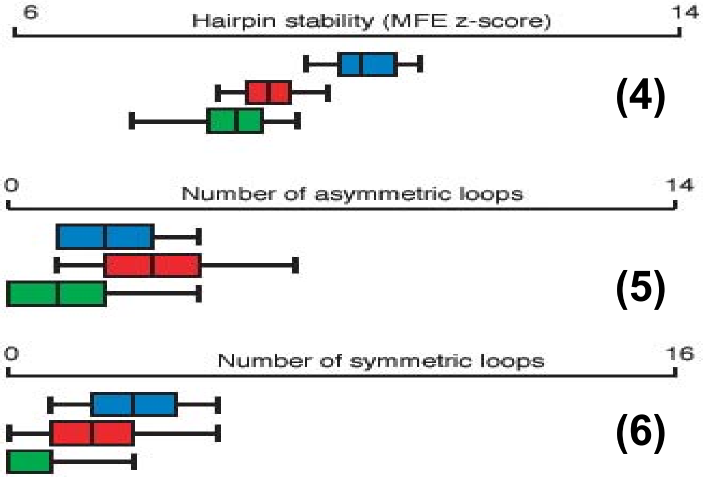
miRNAs show characteristic conservation properties

Distinguishing true miRNAs from random hairpins

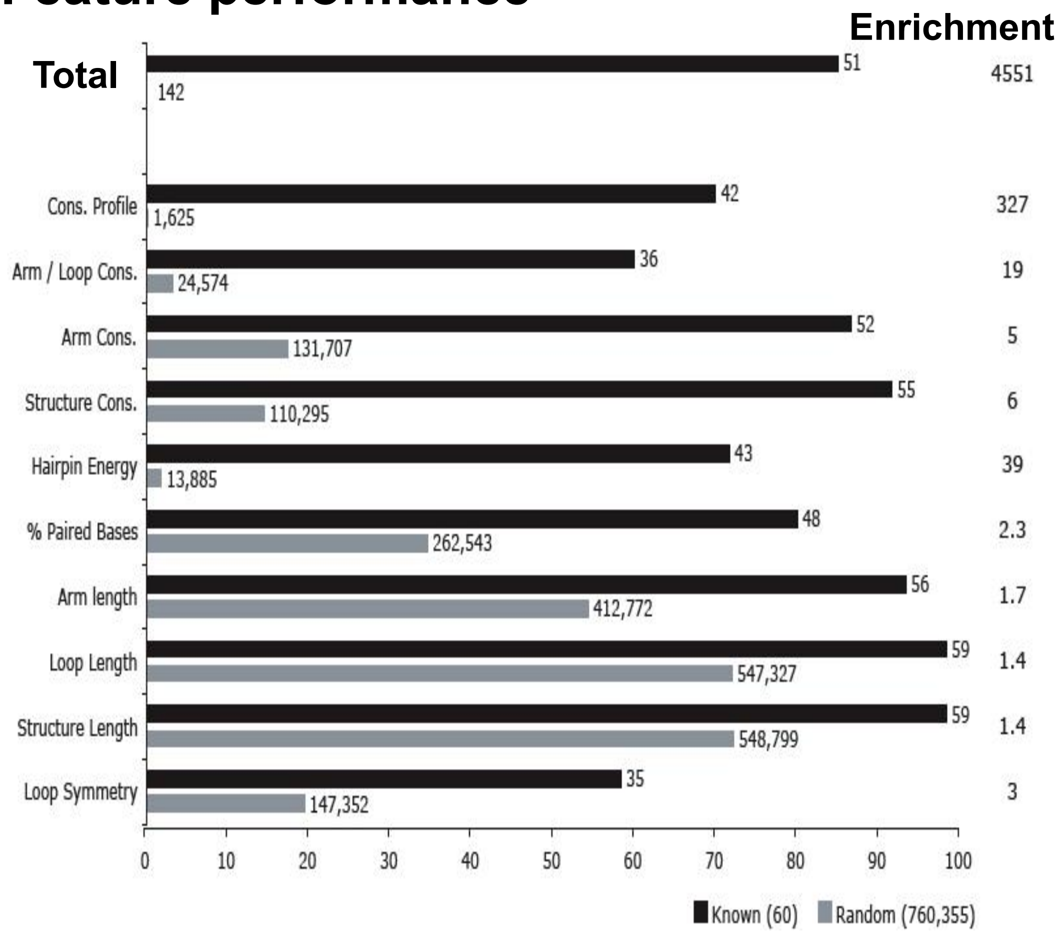
Evolutionary features



Structural features



Feature performance

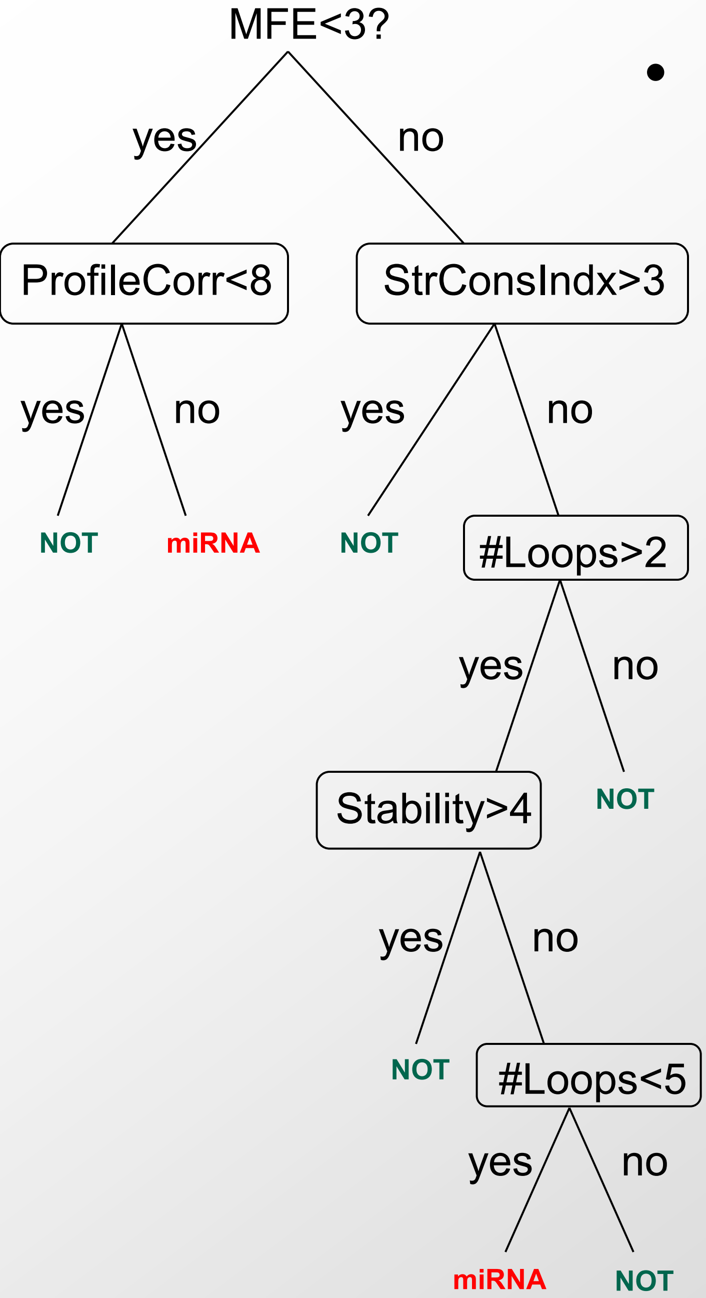


Combination of features:
> 4,500-fold enrichment

Comparative genomics I: Evolutionary signatures

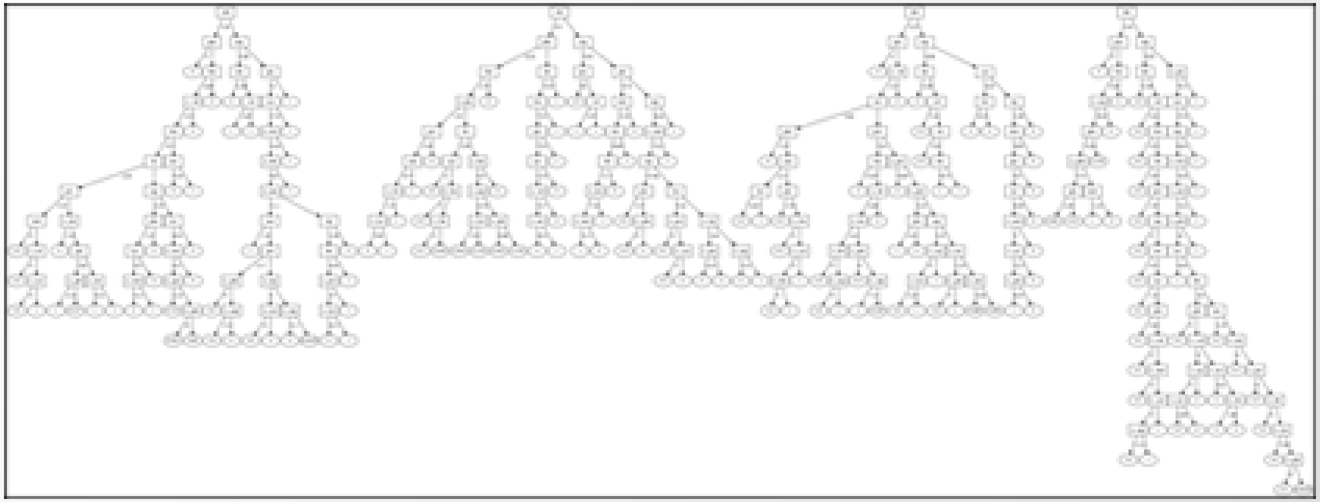
- **Nucleotide conservation: evolutionary constraint**
  - Purifying selection, neutral branch length, discovery power
  - Detect constrained elements: nucleotides, windows, HMM
  - Estimate fraction constrained: signal vs. background
- **Evolutionary signatures: focus on pattern of change**
  - Different functions ⬄ Characteristic patterns of evolution
- **Signatures of protein-coding genes**
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_C Q_N$, scoring
  - Revise genes, read-through, excess constraint regions
- **Signatures of microRNA genes**
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation

miRNA detection using many decision trees



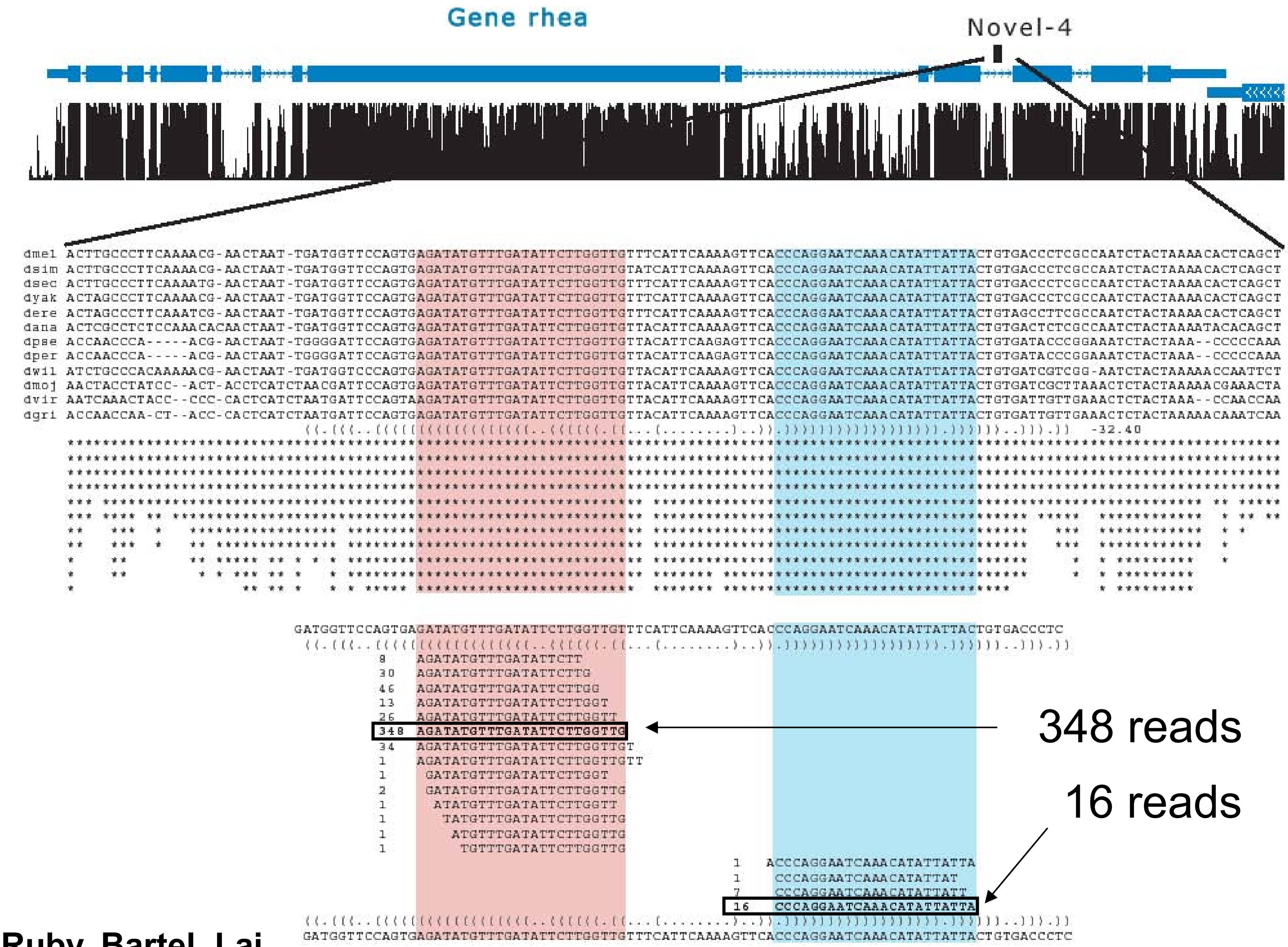
- **For each tree:**
  - Randomly select:
    - Subset of features to base classification on
    - Subset of +/- training examples
    - Remainder of testing examples
  - Use to train a decision tree classifier:
    - Select a feature and cutoff at each level
    - Continue with feature/cutoff at next level
    - (…)
  - Evaluate performance on test set:
    - Push each element down the decision tree
    - Leaf label gives classification decision
- **To combine trees:**
  - Average prediction class across trees
  - Report class with maximum # of votes

Random Forests: Combine many decision trees



- **Many decision trees:**
  - Each can select cutoffs and direction of cutoff
  - Each feature can be reused multiple times
  - Used serially (AND) and in parallel (OR)
- **Ensemble classifier**
  - Bagging: model averaging, combines predictions
  - Can take median of predictions
- **Advantages: Robustness, Feature importance**

Evidence 1: Novel miRNAs match sequencing reads



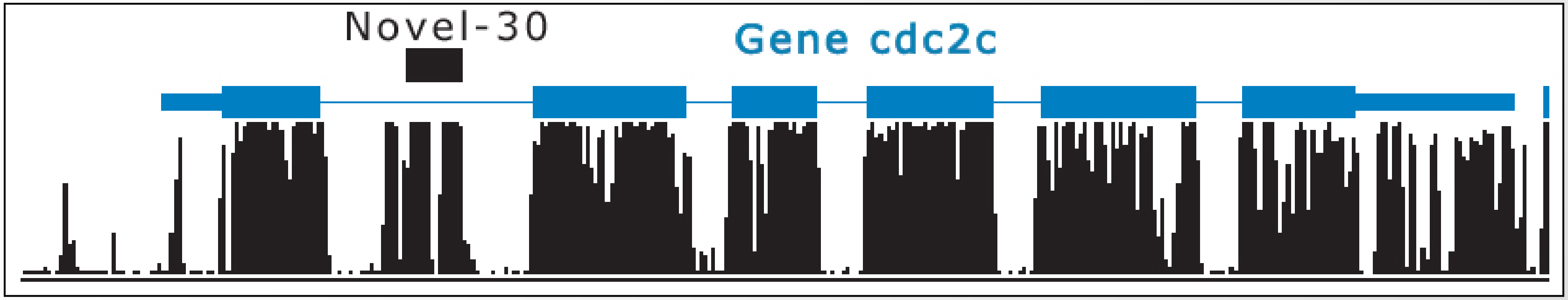
```
dmel ACTTGCCCTTCAAAACG-AACTAAT-TGATGGTTCCAGTGAGATATGTTTGATATTCTTGGTTGTTTCATTCAAAAGTTCACCCAGGAATCAAACATATTATTACTGTGACCCTCGCCAATCTACTAAAACACTCAGCT
dsim ACTTGCCCTTCAAAACG-AACTAAT-TGATGGTTCCAGTGAGATATGTTTGATATTCTTGGTTGTATCATTCAAAAGTTCACCCAGGAATCAAACATATTATTACTGTGACCCTCGCCAATCTACTAAAACACTCAGCT
dsec ACTTGCCCTTCAAAATG-AACTAAT-TGATGGTTCCAGTGAGATATGTTTGATATTCTTGGTTGTTTCATTCAAAAGTTCACCCAGGAATCAAACATATTATTACTGTGACCCTCGCCAATCTACTAAAACACTCAGCT
dyak ACTAGCCCTTCAAAACG-AACTAAT-TGATGGTTCCAGTGAGATATGTTTGATATTCTTGGTTGTTTCATTCAAAAGTTCACCCAGGAATCAAACATATTATTACTGTGACCCTCGCCAATCTACTAAAACACTCAGCT
dere ACTAGCCCTTCAAATCG-AACTAAT-TGATGGTTCCAGTGAGATATGTTTGATATTCTTGGTTGTTTCATTCAAAAGTTCACCCAGGAATCAAACATATTATTACTGTAGCCTTCGCCAATCTACTAAAACACTCAGCT
dana ACTCGCCTCTCCAAACACAACTAAT-TGATGGTTCCAGTGAGATATGTTTGATATTCTTGGTTGTTACATTCAAAAGTTCACCCAGGAATCAAACATATTATTACTGTGACTCTCGCCAATCTACTAAAATACACAGCT
dpse ACCAACCCA-----ACG-AACTAAT-TGGGGATTCCAGTGAGATATGTTTGATATTCTTGGTTGTTACATTCAAGAGTTCACCCAGGAATCAAACATATTATTACTGTGATACCCGGAAATCTACTAAA--CCCCCAAA
dper ACCAACCCA-----ACG-AACTAAT-TGGGGATTCCAGTGAGATATGTTTGATATTCTTGGTTGTTACATTCAAGAGTTCACCCAGGAATCAAACATATTATTACTGTGATACCCGGAAATCTACTAAA--CCCCCAAA
dwil ATCTGCCCACAAAAACG-AACTAAT-TGATGGTCCCAGTGAGATATGTTTGATATTCTTGGTTGTTACATTCAAAAGTTCACCCAGGAATCAAACATATTATTACTGTGATCGTCGG-AATCTACTAAAAACCAATTCT
dmoj AACTACCTATCC--ACT-ACCTCATCTAACGATTCCAGTGAGATATGTTTGATATTCTTGGTTGTTACATTCAAAAGTTCACCCAGGAATCAAACATATTATTACTGTGATCGCTTAAACTCTACTAAAAACGAAACTA
dvir AATCAAACTACC--CCC-CACTCATCTAATGATTCCAGTAAGATATGTTTGATATTCTTGGTTGTTACATTCAAAAGTTCACCCAGGAATCAAACATATTATTACTGTGATTGTTGAAACTCTACTAAA--CCAACCAA
dgri ACCAACCAA-CT--ACC-CACTCATCTAATGATTCCAGTGAGATATGTTTGATATTCTTGGTTGTTACATTCAAAAGTTCACCCAGGAATCAAACATATTATTACTGTGATTGTTGAAACTCTACTAAAAACAAATCAA
                                                                                                                          -32.40

GATGGTTCCAGTGAGATATGTTTGATATTCTTGGTTGTTTCATTCAAAAGTTCACCCAGGAATCAAACATATTATTACTGTGACCCTC
          8    AGATATGTTTGATATTCTT
          30   AGATATGTTTGATATTCTTG
          46   AGATATGTTTGATATTCTTGG
          13   AGATATGTTTGATATTCTTGGT
          26   AGATATGTTTGATATTCTTGGTT
          348  AGATATGTTTGATATTCTTGGTTG
          34   AGATATGTTTGATATTCTTGGTTGT
          1    AGATATGTTTGATATTCTTGGTTGTT
          1     GATATGTTTGATATTCTTGGT
          2     GATATGTTTGATATTCTTGGTTG
          1      ATATGTTTGATATTCTTGGTT
          1       TATGTTTGATATTCTTGGTTG
          1        ATGTTTGATATTCTTGGTTG
          1         TGTTTGATATTCTTGGTTG
                                                 1   ACCCAGGAATCAAACATATTATTA
                                                 1    CCCAGGAATCAAACATATTAT
                                                 7    CCCAGGAATCAAACATATTATT
                                                 16   CCCAGGAATCAAACATATTATTA
GATGGTTCCAGTGAGATATGTTTGATATTCTTGGTTGTTTCATTCAAAAGTTCACCCAGGAATCAAACATATTATTACTGTGACCCTC
```

348 reads

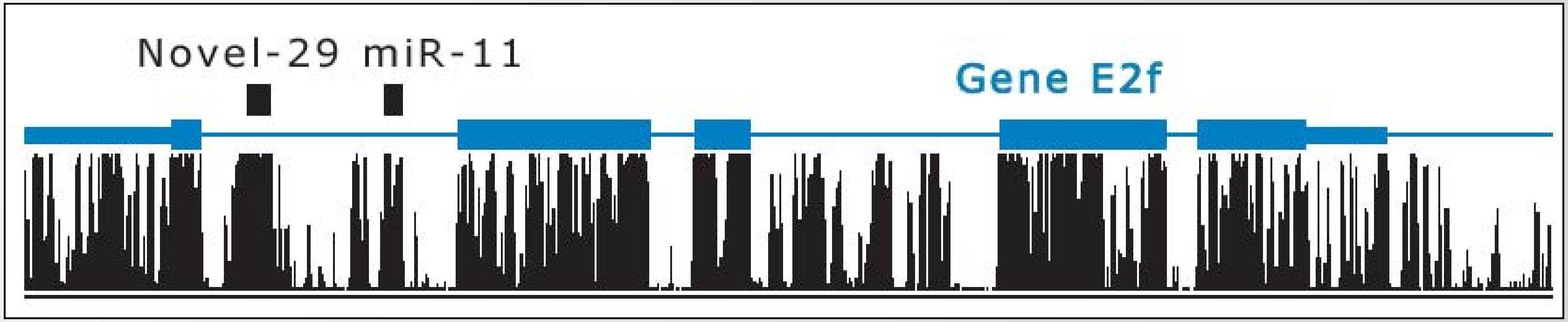
16 reads

Ruby, Bartel, Lai

Evidence 2: Genomic properties typical of miRNAs



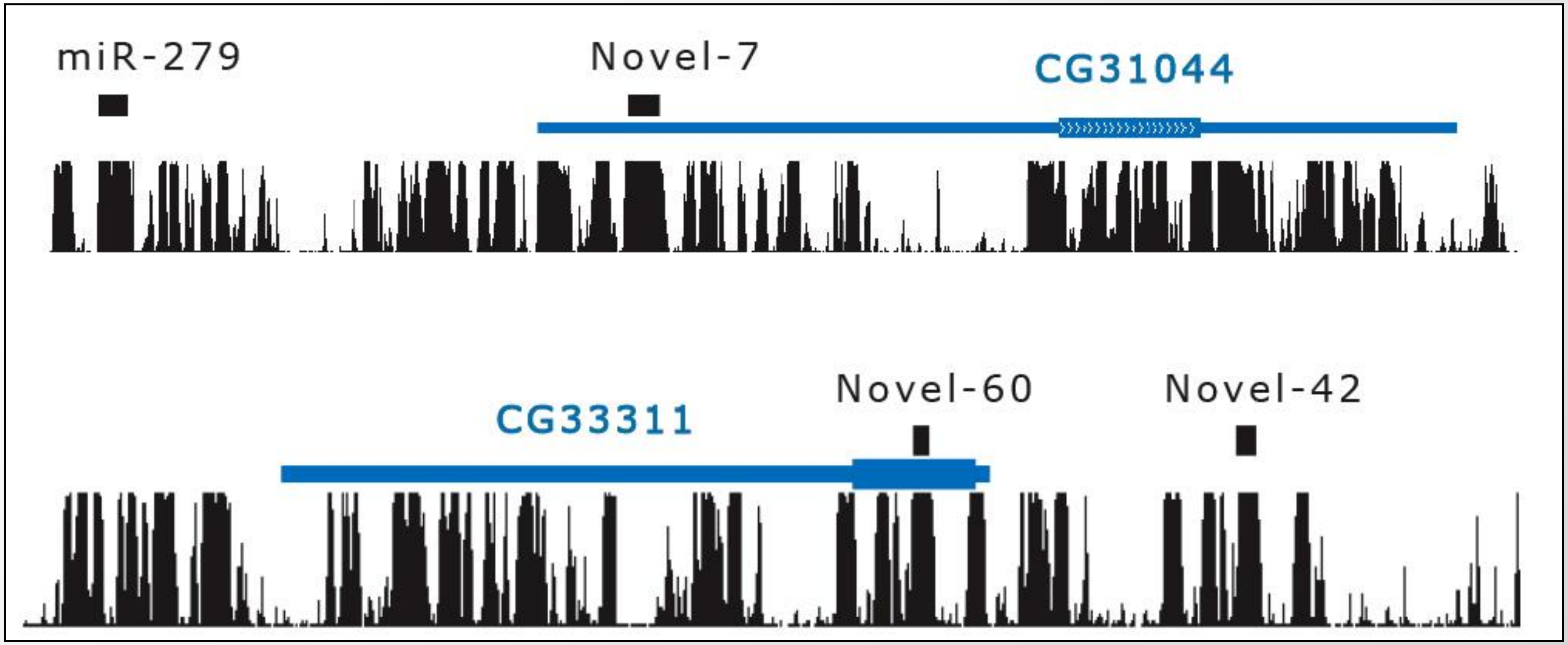
- Novel miRNAs in introns of known genes
- Preference for + strand, transcription factors



- Genomic clustering with novel / known miRNAs
- Same family, common origin / same precursor

Two ‘dubious’ protein-coding genes are in fact miRNAs

Two novel miRNAs overlap exons (5’UTR and coding!)

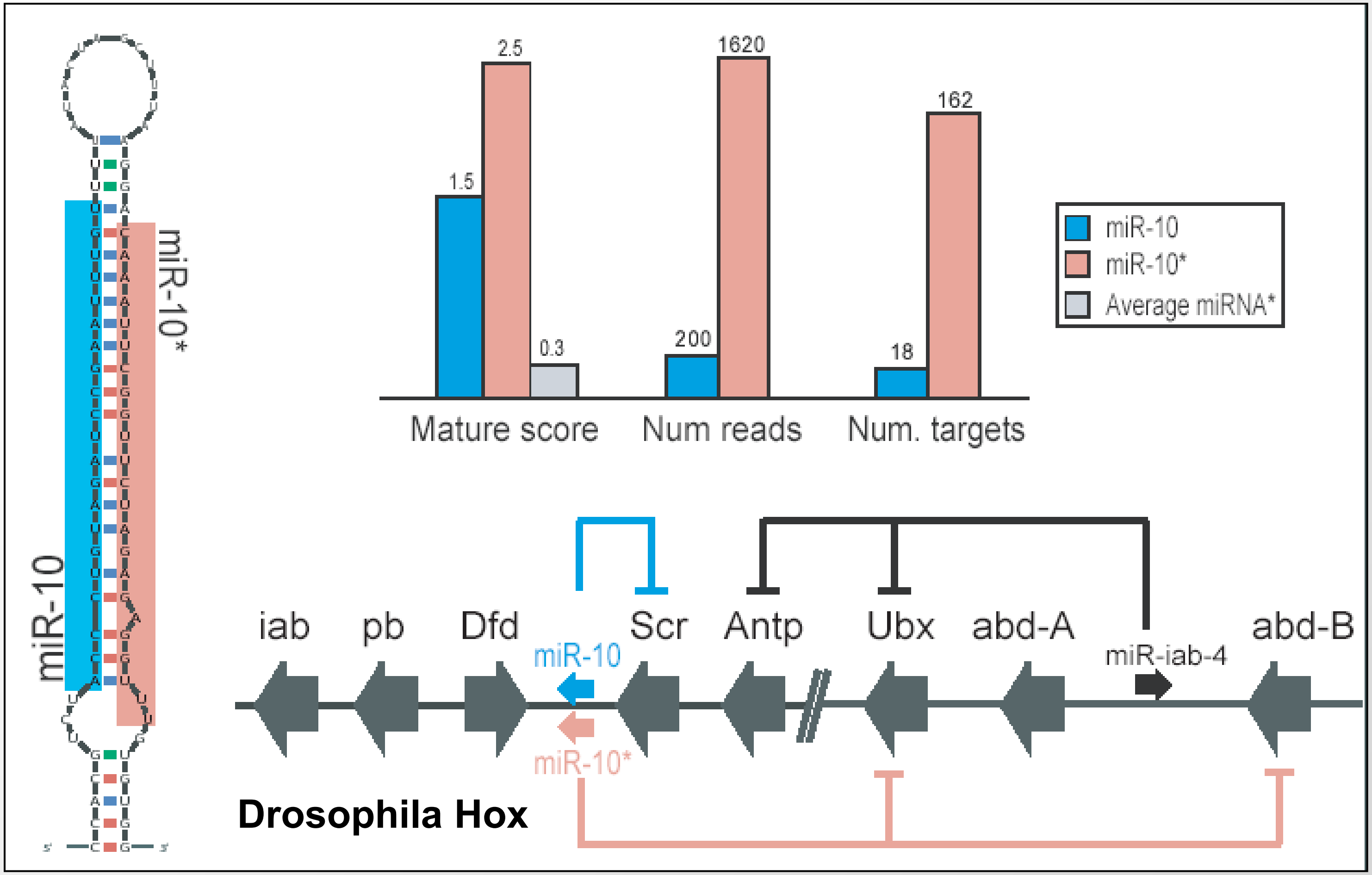


- **Both CG31044 and CG33311 were independently rejected as *dubious* based on their non-protein-coding conservation patterns (Lin *et al.*)**
- **Novel miRNA genes provide explanation for their transcripts, as their precursor miRNA**

Comparative genomics I: Evolutionary signatures

- **Nucleotide conservation: evolutionary constraint**
  - – Purifying selection, neutral branch length, discovery power
  - – Detect constrained elements: nucleotides, windows, HMM
  - – Estimate fraction constrained: signal vs. background
- **Evolutionary signatures: focus on pattern of change**
  - – Different functions ⇔ Characteristic patterns of evolution
- **Signatures of protein-coding genes**
  - – Reading-frame conservation, codon-substitution frequency
  - – Likelihood ratio framework: Estimating $Q_C Q_N$, scoring
  - – Revise genes, read-through, excess constraint regions
- **Signatures of microRNA genes**
  - – Structural and evolutionary features of microRNAs
  - – Combining features: decision trees, random forests
  - – Sense/anti-sense miRNAs, mature/star arm cooperation

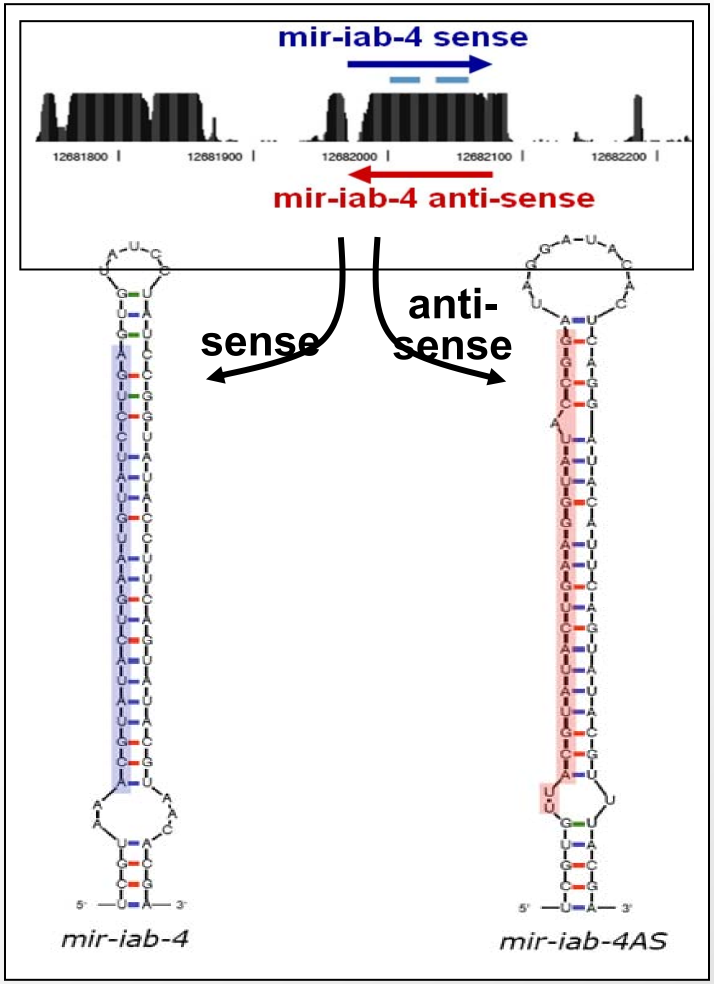
Surprise 1: microRNA & microRNA* function

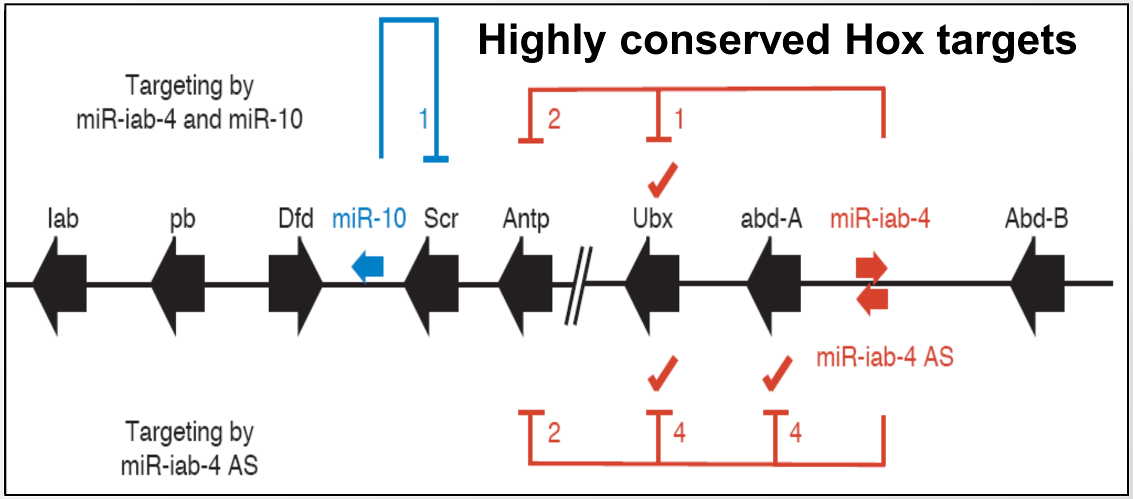


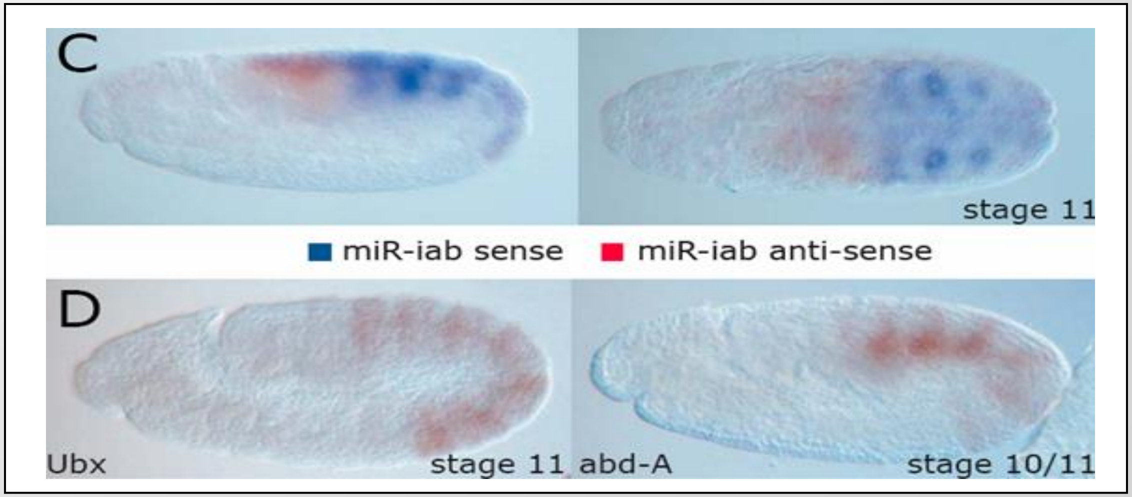
- **Both hairpin arms of a microRNA can be functional**
  - High scores, abundant processing, conserved targets
  - Hox miRNAs miR-10 and miR-iab-4 as master Hox regulators

Stark *et al*, Genome Research 2007

Evidence of miR-iab-4 anti-sense (AS) function




Highly conserved Hox targets
Targeting by miR-iab-4 and miR-10
1
2
1
lab
pb
Dfd
miR-10
Scr
Antp
Ubx
abd-A
miR-iab-4
Abd-B
miR-iab-4 AS
Targeting by miR-iab-4 AS
2
4
4

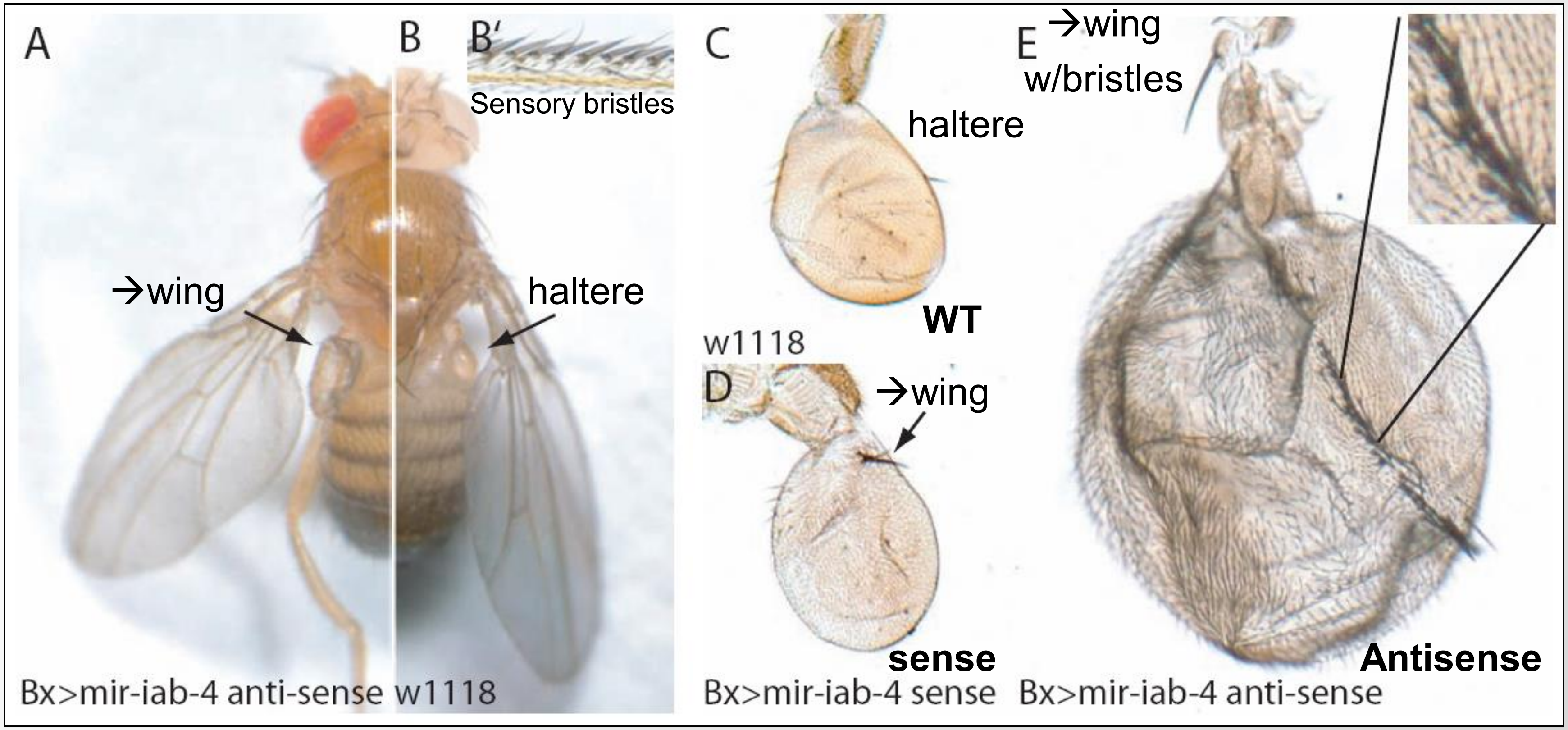


Source: Stark, Alexander et al. "A single Hox locus in Drosophila produces functional microRNAs from opposite DNA strands." Genes & development 22, no. 1 (2008): 8-13.

- **A single miRNA locus transcribed from both strands**
- **The two transcripts show distinct expression domains (mutually exclusive)**
- **Both processed to mature miRNAs: mir-iab-4, miR-iab-4AS (anti-sense)**

Stark *et al*, Genes&Development 2008

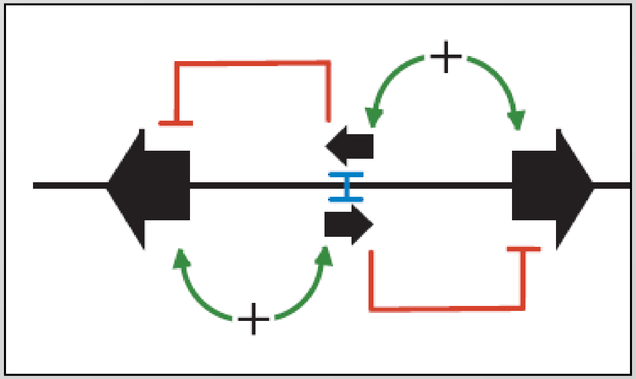
miR-iab-4AS leads to homeotic transformations



Note: C,D,E same magnification

Source: Stark, Alexander et al. "A single Hox locus in Drosophila produces functional microRNAs from opposite DNA strands." Genes & development 22, no. 1 (2008): 8-13.

- **Mis-expression of mir-iab-4S & AS: alteres→wings homeotic transform.**
- **Stronger phenotype for AS miRNA**
- **Sense/anti-sense pairs as general building blocks for miRNA regulation**
- **10 sense/anti-sense miRNAs in mouse**

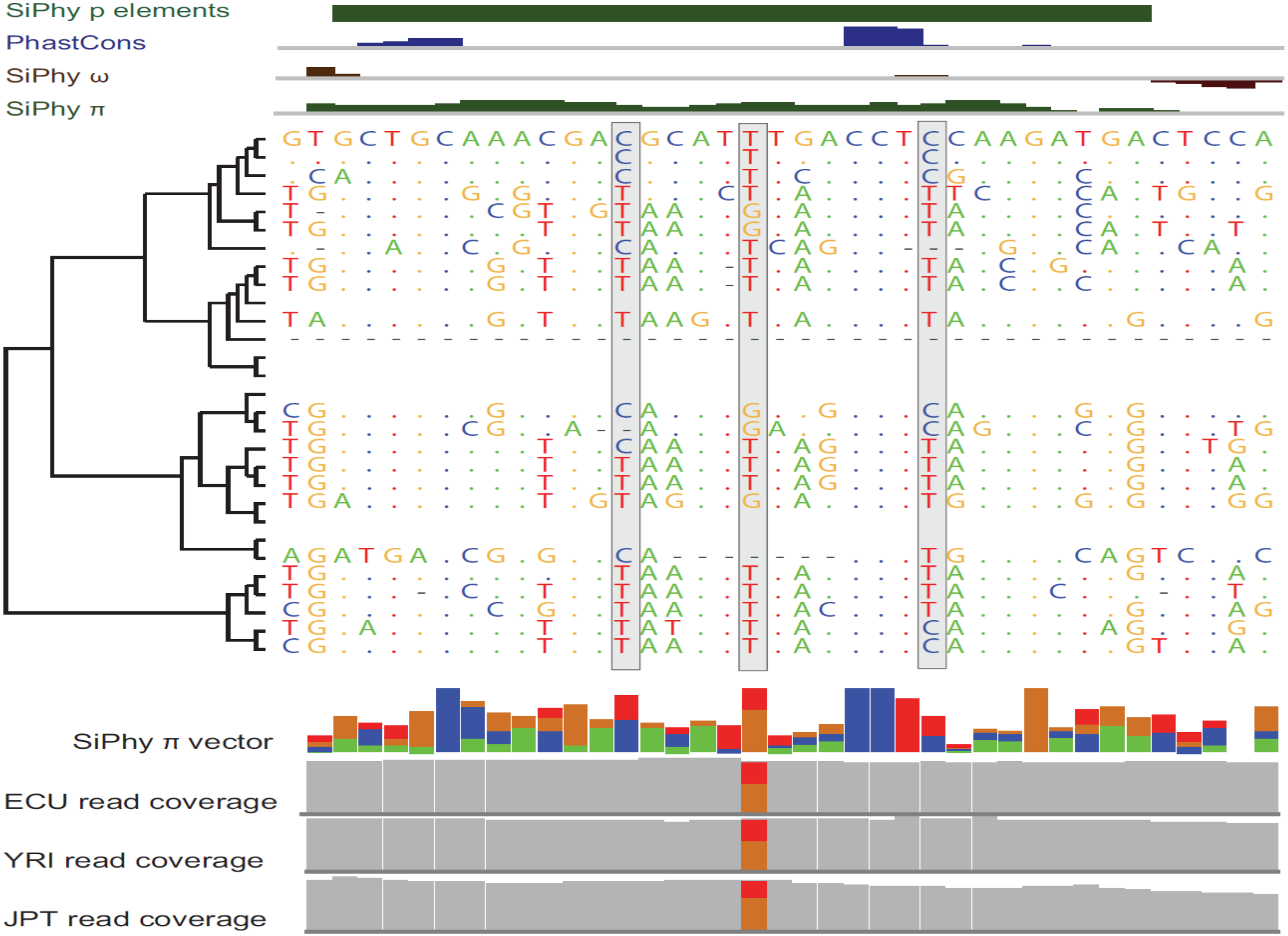


Stark *et al*, Genes&Development 2008

Comparative genomics I: Evolutionary signatures

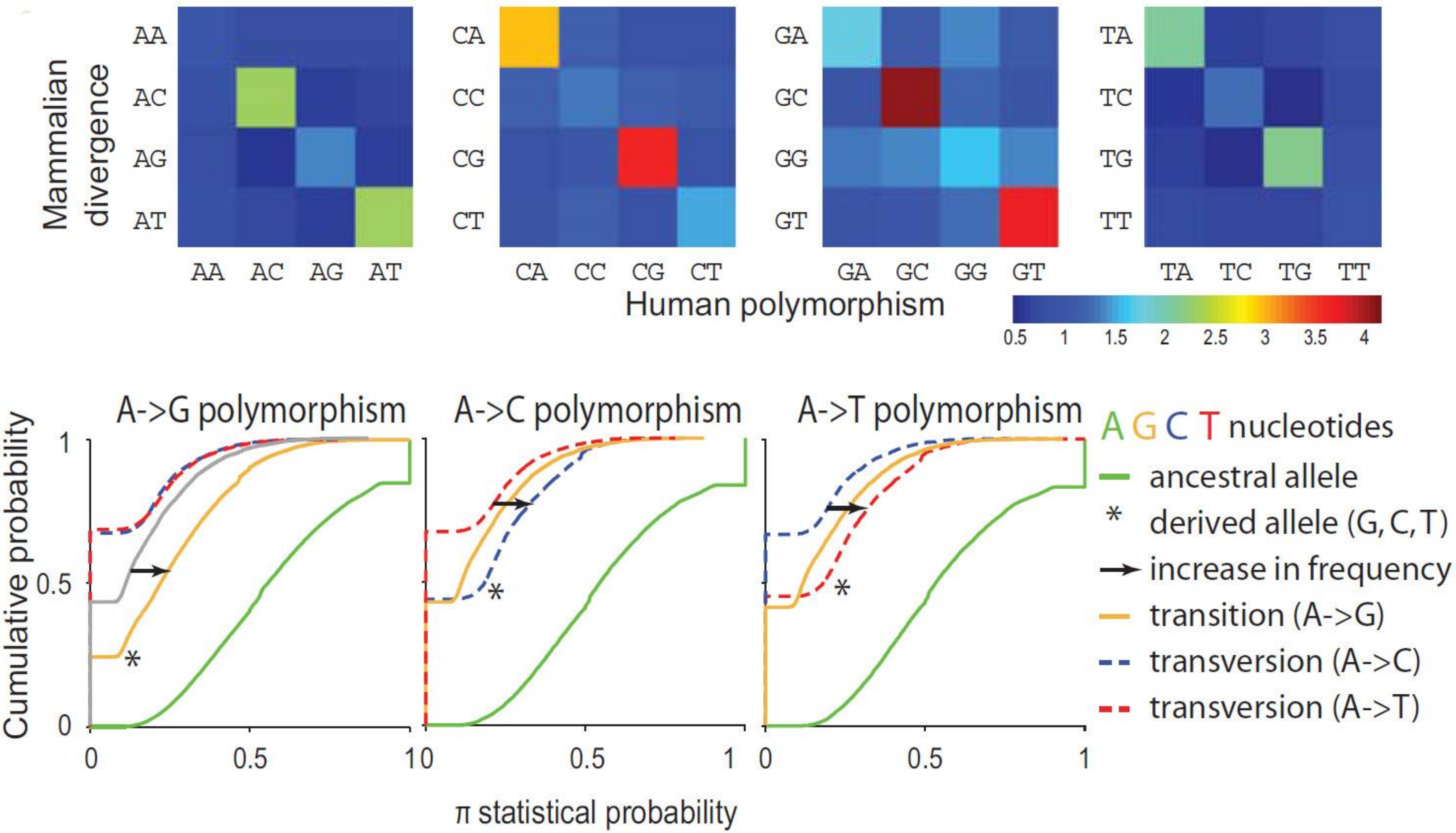
- **Nucleotide conservation: evolutionary constraint**
  - Purifying selection, neutral branch length, discovery power
  - Detect constrained elements: nucleotides, windows, HMM
  - Estimate fraction constrained: signal vs. background
- **Evolutionary signatures: focus on pattern of change**
  - Different functions ⬄ Characteristic patterns of evolution
- **Signatures of protein-coding genes**
  - Reading-frame conservation, codon-substitution frequency
  - Likelihood ratio framework: Estimating $Q_C Q_N$, scoring
  - Revise genes, read-through, excess constraint regions
- **Signatures of microRNA genes**
  - Structural and evolutionary features of microRNAs
  - Combining features: decision trees, random forests
  - Sense/anti-sense miRNAs, mature/star arm cooperation
- **Measuring selection within the human lineage**

Mammalian constraint matches Human SNPs



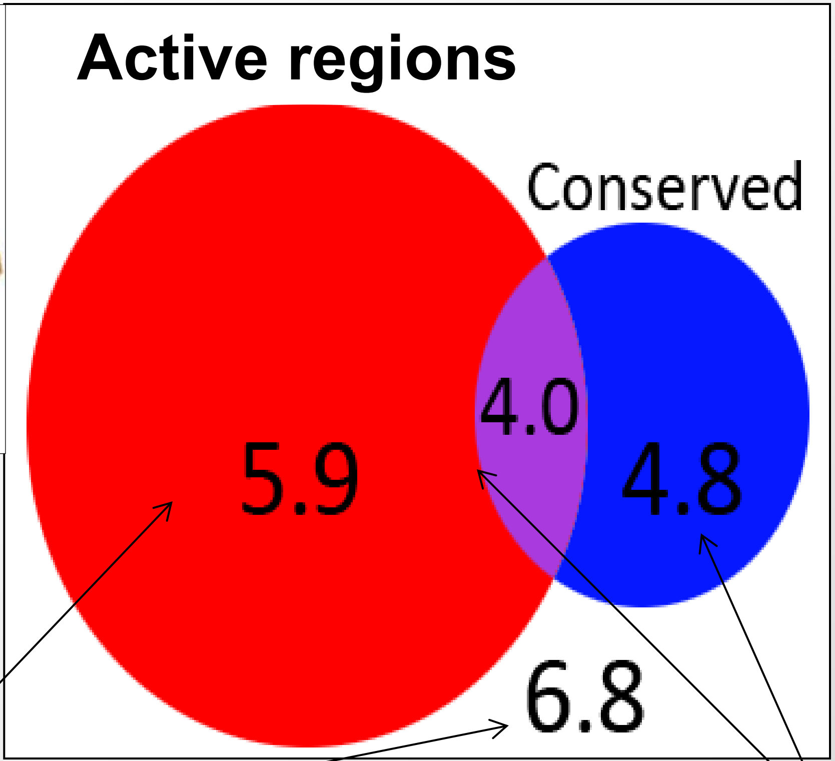
Human SNPs match mammalian-wide twofold constraint

Mammalian constraint matches Human SNPs



Genome-wide agreement of selection and polymorphisms

Human constraint outside conserved regions



Average diversity (heterozygosity)

Aggregate over the genome

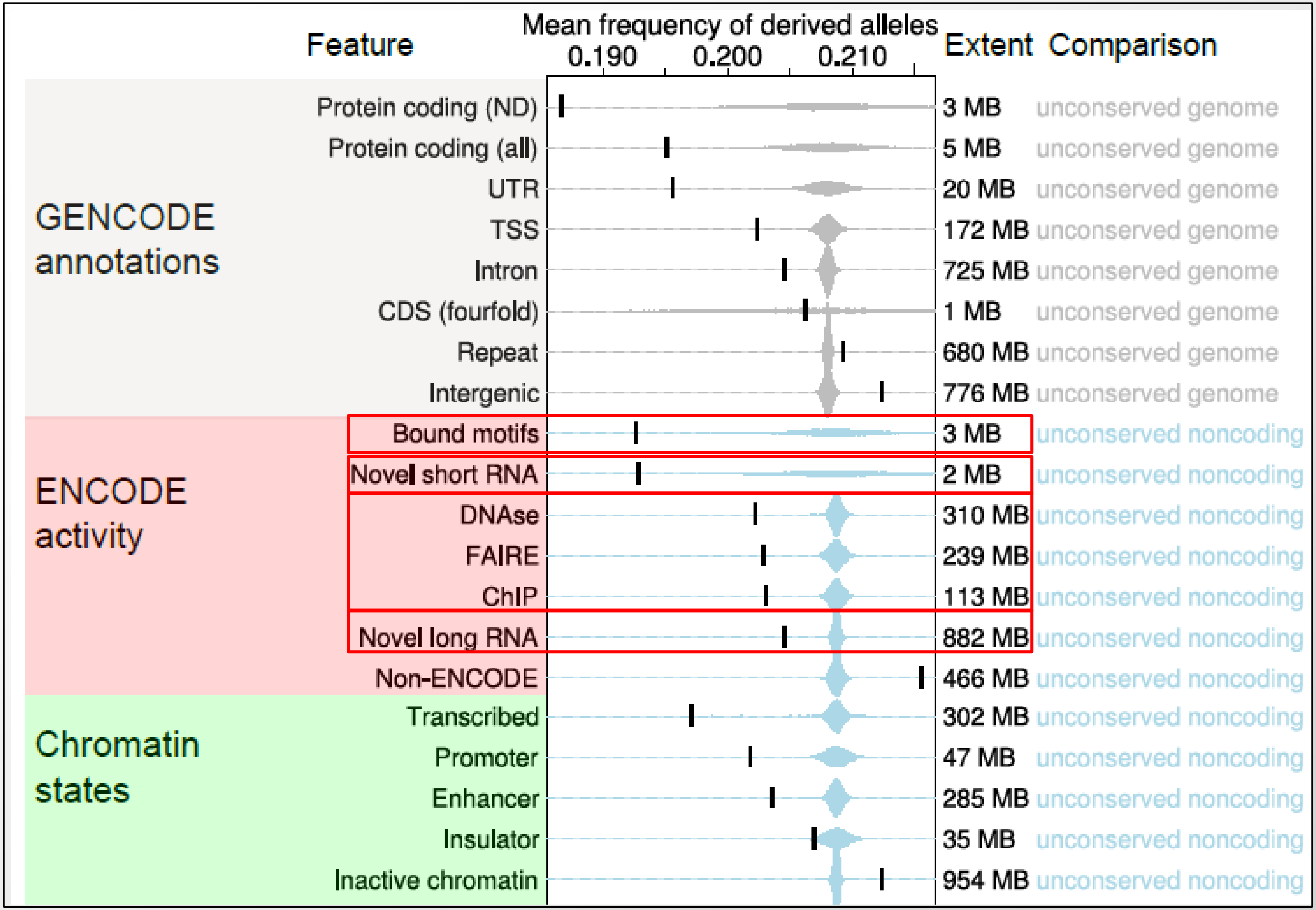
- **Non-conserved regions:**
  - – ENCODE-active regions show reduced diversity

➔ Lineage-specific constraint in biochemically-active regions

- **Conserved regions:**
  - – Non-ENCODE regions show increased diversity

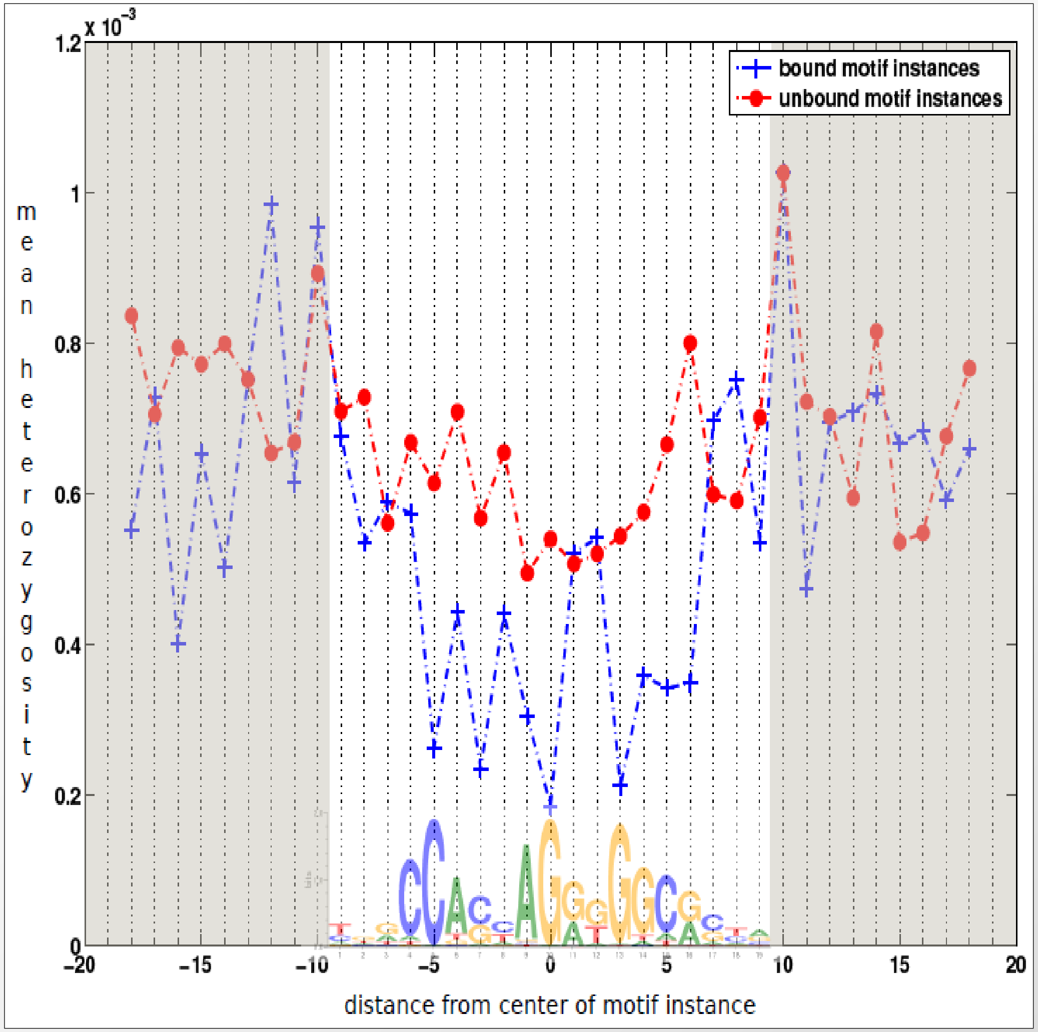
➔ Loss of constraint in human when biochemically-inactive

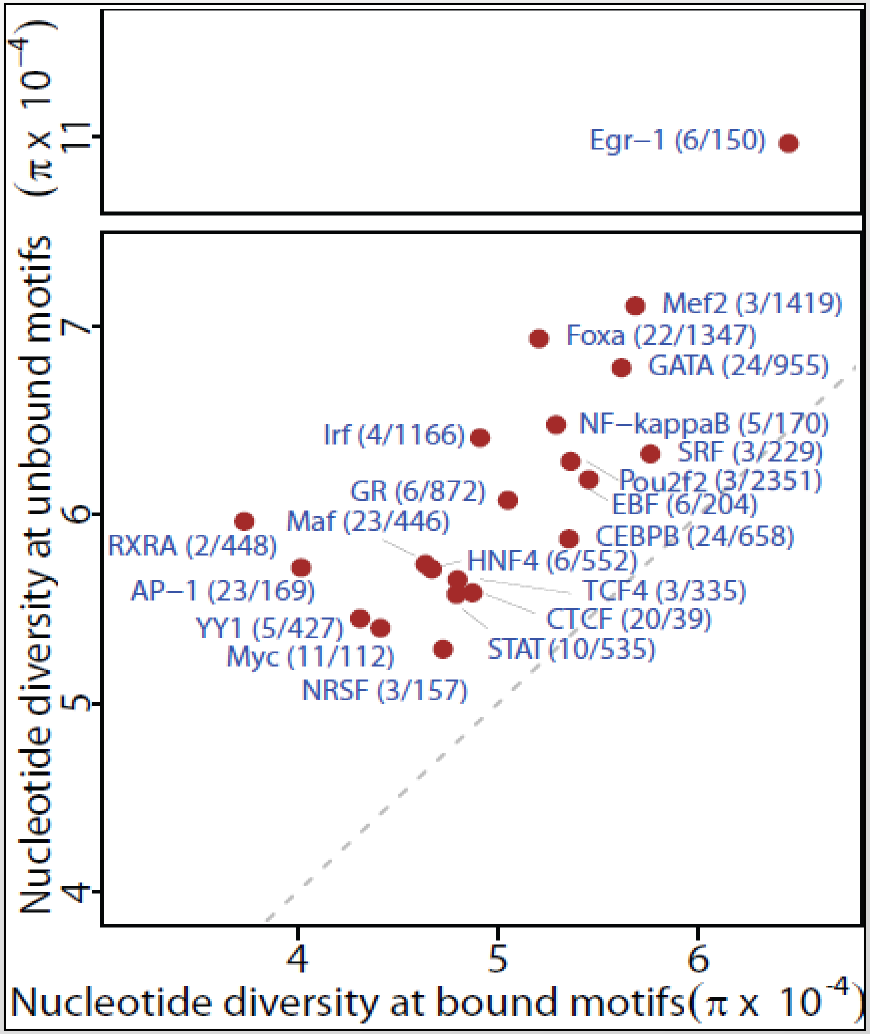
Strongest: motifs, short RNA, Dnase, ChIP, lncRNA



- **Significant derived allele depletion in active features**

Bound motifs show increased human constraint





Position-specific reduction in bound motif heterozygosity

Aggregate across thousands of CTCF motif instances

Most constrained human-specific enhancer functions

Transcription initiation from Pol2 promoter

Transcription coactivator activity

Transcription factor binding

Chromatin binding

Negative regulation of transcription, DNA-dependent

Transcription factor complex

Protein complex

Protein kinase activity

Nerve growth factor receptor signaling pathway

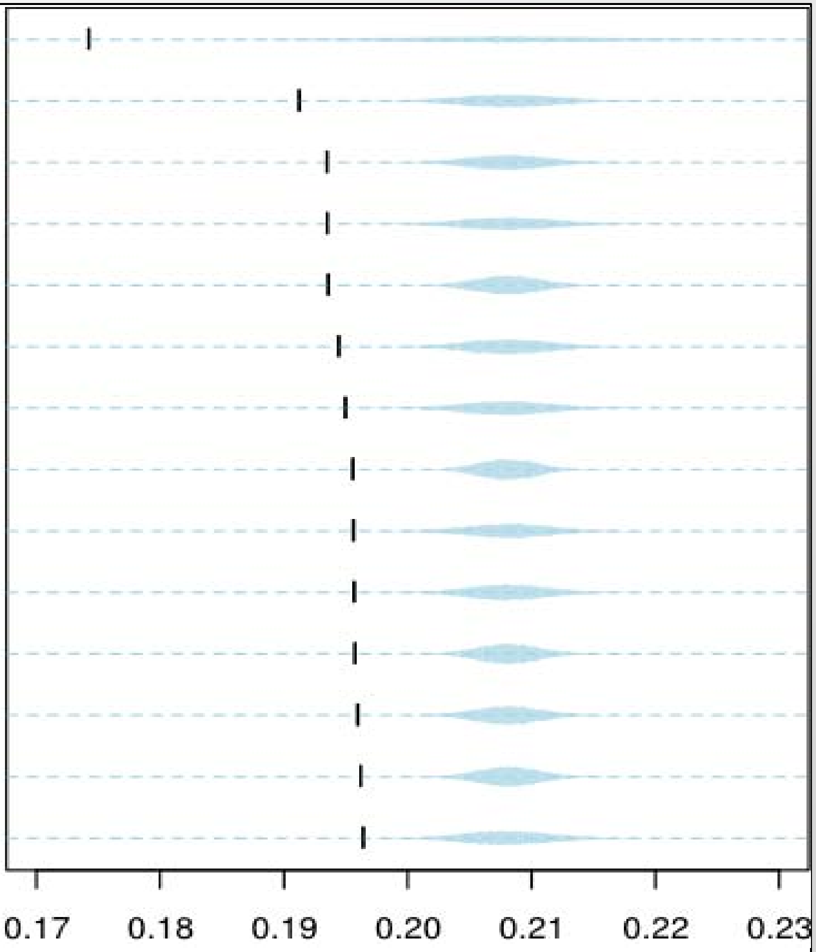
Signal transducer activity

Protein serine/threonine kinase activity

Negative regulation of transcription from Pol2 prom

Protein tyrosine kinase activity

In utero embryonic development



Regulatory genes: Transcription, Chromatin, Signaling.

Developmental enhancers: embryo, nerve growth

Comparative genomics I: Evolutionary signatures

- **Nucleotide conservation: evolutionary constraint**
  - – Purifying selection, neutral branch length, discovery power
  - – Detect constrained elements: nucleotides, windows, HMM
  - – Estimate fraction constrained: signal vs. background
- **Evolutionary signatures: focus on pattern of change**
  - – Different functions ⬄ Characteristic patterns of evolution
- **Signatures of protein-coding genes**
  - – Reading-frame conservation, codon-substitution frequency
  - – Likelihood ratio framework: Estimating $Q_C Q_N$, scoring
  - – Revise genes, read-through, excess constraint regions
- **Signatures of microRNA genes**
  - – Structural and evolutionary features of microRNAs
  - – Combining features: decision trees, random forests
  - – Sense/anti-sense miRNAs, mature/star arm cooperation
- **Measuring selection within the human lineage**