

New Methods for Comparative genomics

-Mark Yandell HHMI/BDGP

New Methods for Comparative genomics

-Mark Yandell HHMI/BDGP

- CGL: a software library for comparative genomics
- Explore recent history (5-70 myr years)
- Explore ancient history (70-1000 myr years)

Part I:

CGL: A software library for comparative genomics

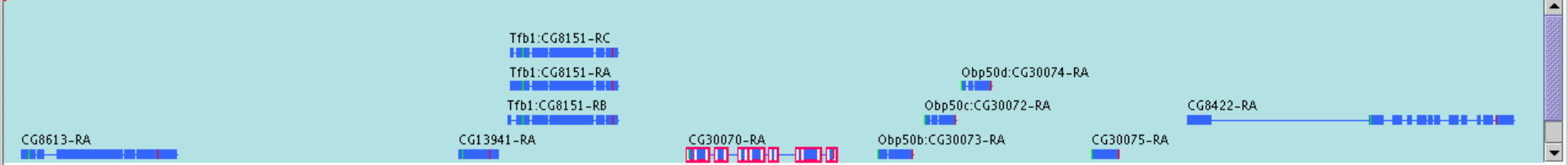
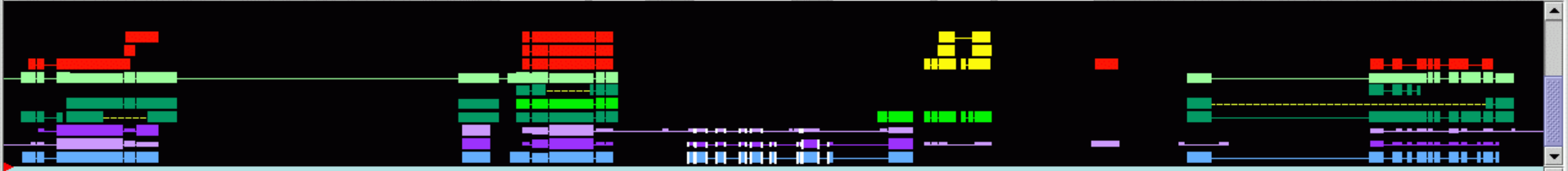
>264

MSFNNALSGVNAAQKDLNVTANNIANVNTTGFKESRAEFADVANSIFVNAKTQVGNQVA
TGAVAQQFHQALQFTNNALDLSIQNGFFVTS DGLTNLDRTFTRAGAFKLNENS YMVNN
QGNYLQGYEINTDGT PKAVSINATKPIQIPDRAGEPKMTELVEASFNLSIESKTKPTSPA
AFDPTNSATFAHSTSVTIYDSL GAPHVITKYFVRHEDPAAPGTPLTPGVKMTFTSGKLDP
TLTVPVDPIKTVALGTTAGI INNGADPTQTLEIRLGDVTQYSSPFNVTKLTQDGATVGNL
TKVEITPDGIVSATYSNATTLKVAMVALAKFANSQGLTQVGDTSWRQSLLSGDALPGTPN
SGTLGSIKSSALEQSNVDLTSQLVNLIT AQRNFAQANSRSLEVNSSLQQTILQI

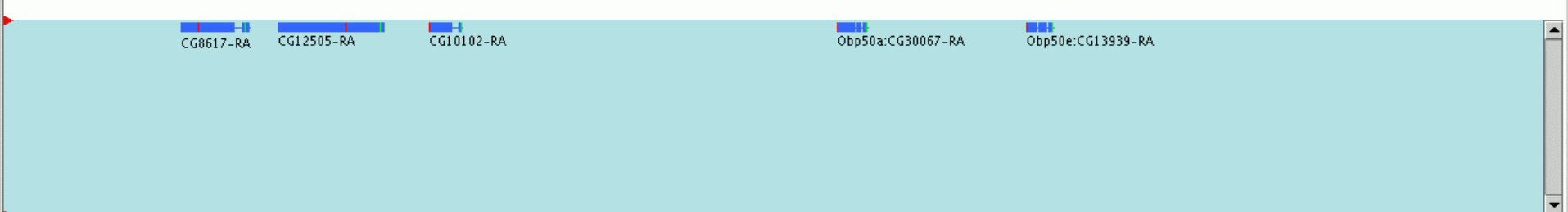
>264

ATGAAAGTTAGTTTTGAAAGAATAATTC CAAGTGAAAAAGCTCTTTCCGCACACTGCAT
AATAACTCTCCTATTTCTGAATTTAAATGGGAGTATCATTATCATCCGGAAATAGA ACTG
GTATGTGTAATTTCCGGGAAGTGGCACACGGCATGTAGGCTACCATAAAAGCAATTATACA
AACGGAGATCTTGTGTTAATAGGTTCAAACATTCCACATTCGGATTTGGACTGAATTCT
GTTGATCCGCATGAAGAAATAGTACTTCAGTTCAGGGAAGAGATTTTGCATTTTCCACAA
CAGGAAGTTGAAACAAGAGCCGTGAAAGATCTACTGGAACGCTCTAAATATGGTATTCTG
TATAGTACAGCTACAAAAAAGCTGCTCATGCCGAAACTAAAAAAGCTTCTGGAATCCGAA
GGCTACAAAAGATACTTACTACTTCTGGAGATTCTCTTCGAACTTTCTTTGTGCGAGGAA
TATGAATTGTTGAACAAAGAAATTATGCCTTATAACCATAATCTCTAAAAATAAAACAAGA
CTGGAAAATATCTTTACCTATGTGGAACATCATTACGATAAGGAAATAAATATAGAGGAT
GTTGCAAAGCTGGCTAATCTTACTCTTCCTGCATTTTGTAAATTTTTTTAAAAAAGCAACA
CAGATTACCTTTACAGAATTTGTCAACCGTTACCGTATTAATAAAGCCTGCCTTCTGATG
ACTCAGGATAAAACAATATCCGAATGCAGCTACAGTTGTGGCTTTAACAATGTTACTTAT
TTCAACAGAATGTTTTAAAAAATATAACCAATAAAACGCCATCAGAATTT

Chromosome 2R Start 9405773 End 9459007 Expand Load



9.42Mb 9.425Mb 9.43Mb 9.435Mb 9.44Mb 9.445Mb 9.45Mb

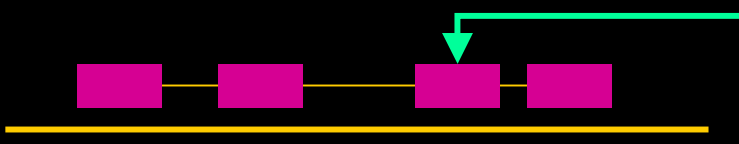


Position

Zoom x10 x2 x5 x1 Reset Zoom factor = 1.5662

Type	Name	Range	Gene: CG30070	
Gene	CG30070-RA	9430773-94...	Gene Transcript: CG30070-RA	
			Genomic Range	Id
			9430773-9430883	CG30070:1
			9430952-9431172	CG30070:2
			9431402-9431558	CG30070:3
			9431916-9432029	CG30070:4
			9432166-9432386	CG30070:5
			9432604-9432695	CG30070:6

Position 9415681 Feature Action

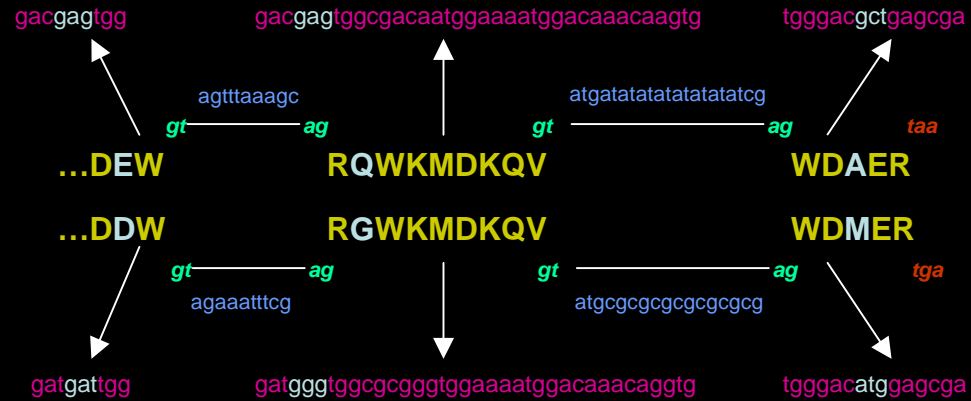


...DEW RQWKMDKQV WDA ER
...DDW RGWKMDKQV WDMER

Gene structure

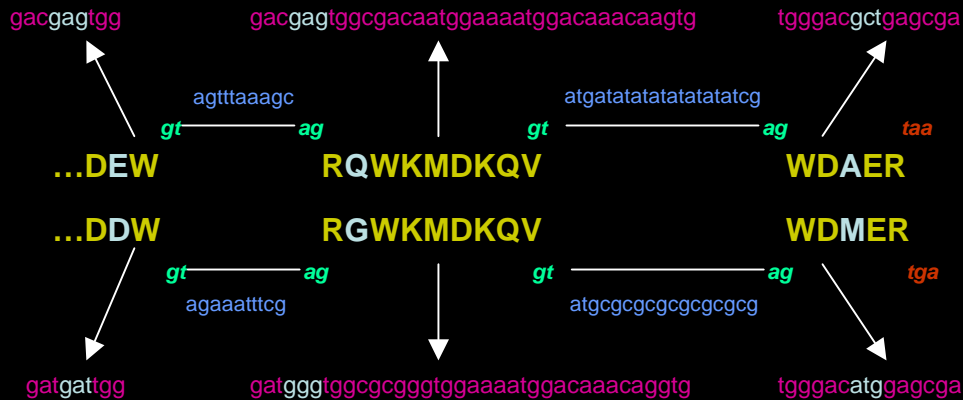
Sequence similarity

Aligning two sequences implicitly aligns two genes



Using BLASTP alignments to make sense of genomic annotations

Annotation A



Annotation B

Using genomic annotations to make sense of BLASTP alignments

Score = 42.0 bits (97), Expect(2) = $2e-25$

Identities = 23/64 (35%), Positives = 40/64 (61%), Gaps = 2/64 (3%)

Frame = -3

```
Query: 1 MFQNDVSSPRELQLMAAKVEKELGPVDILVNNASLMPMTSTP-SLKSDEIDTILQLNL-G 59
      + + DVS+ E++ M KV + GP+DIL+NNA ++ T P + +E D ++ +NL G
Sbjct: 29 VVKADVSNREEVREMVKKVIDKFGPIDILINNAGILGKTKDPLEVTDEEWDRVISVNLKG 88
```

Score = 37.7 bits (86), Expect(2) = $2e-25$

Identities = 17/43 (39%), Positives = 29/43 (66%)

Frame = -1

```
Query: 60 VTGAGHGLGRAISLELAKKGCHIAVVDINVSGAEDTVKQIQDI 92
      +TGA G+GRAI++ELAK+G ++ IN SGAE+ K+ +++
Sbjct: 89 ITGASRGIGRAIAIELAKRGVNVV---INYSGAEEEEAKKTEEL 128
```

Using genomic annotations to make sense of BLASTP alignments

Score = 42.0 bits (97), Expect(2) = $2e-25$

Identities = 23/64 (35%), Positives = 40/64 (61%), Gaps = 2/64 (3%)

Frame = -3

Query: 1 MFQNDVSSPRELQLMAAKVEKELGPVDILVNNASLMPMTSTP-SLKSDEIDTILQLNL-G 59
+ + DVS+ E++ M KV + GP+DIL+NNA ++ T P + +E D ++ +NL G
Sbjct: 29 VVKADVSNREEVREMVKKVIDKFGPIDILINNAGILGKTKDPLEVTDEEWDRVISVNLKG 88

splice junction

splice junction

splice junction

splice junction

Score = 37.7 bits (86), Expect(2) = $2e-25$

Identities = 17/43 (39%), Positives = 29/43 (66%)

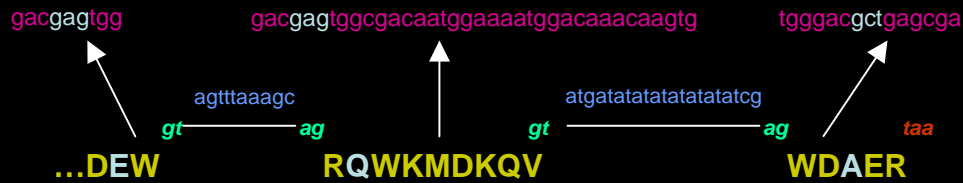
Frame = -1

Query: 60 VTGAGHGLGRAISLELAKKGGCHIAVVDINVSGAEDTVKQIQDI 92
+TGA G+GRAI++ELAK+G ++ IN SGAE+ K+ +++
Sbjct: 89 ITGASRGIGRAIAIELAKRGNVNVV---INYSGAEEEEAKKTEEL 128

splice junction

splice junction

Annotation A



Some un-annotated genome

Using genomic annotations to make sense of TBLASTX alignments

Score = 112 bits (240), Expect(2) = 2e-25
Identities = 58/99 (58%), Positives = 61/99 (61%)
Frame = -3 / +1

Query: 461 VTSATLPLTSFSLSPSANR*SNRCLRKMASRTRGRRSRKTMTFRLKMARVVSSLRCA
+TSATLPLTSFSL P ANR*S RC MASR RG SRK + FRLK ARVVSSLR AC A
Sbjct: 45613 LTSATLPLTSFSLRPRANR*SRRCFSPMASRIRGSSSRKVRIFRLKRARVVSSLRRACEA

Query: 281 ACVAVGVAEAVGPVTVT*PAGTAVVVRVMDSRKPAVSTT 165
A VAVG * V+RVMDS+KPAVSTT
Sbjct: 45793 AWWAVGAVGVAEDEGTV*AGPAGAVMRVMDSMKPAVSTT 45909

Using genomic annotations to make sense of TBLASTX alignments

Score = 112 bits (240), Expect(2) = 2e-25
Identities = 58/99 (58%), Positives = 61/99 (61%)
Frame = -3 / +1

5'-UTR

1st Coding Exon

Query: 461 VTSATLPLTSFSLSPSANR*SNRCLRKMASRTGRRSRKTMTFRLKMARVVSSLRCAC
+TSATLPLTSFSL P ANR*S RC MASR RG SRK + FRLK ARVVSSLR AC A
Sbjct: 45613 LTSATLPLTSFSLRPRANR*SRRCFSPMASRIRGSSSRKVRIFRLKRARVVSSLRRACEA

1st Intron

2nd Coding Exon

Query: 281 ACVAVGVAEAVGPVTVT*PAGTAVVVRVMDSRKPAVSTT 165
A VAVG * V+RVMDS+KPAVSTT
Sbjct: 45793 AWWAVGAVGVAEDEGTV*AGPAGAVMRVMDSMKPAVSTT 45909

Using genomic annotations to make sense of TBLASTX alignments

Score = 112 bits (240), Expect(2) = 2e-25
Identities = 58/99 (58%), Positives = 61/99 (61%)
Frame = -3 / +1

5'-UTR **1st Coding Exon**

Query: 461 VTSATLPLTSFSLSPSANR*SNRCLRKMASRTRGRRSRKTMFRLKMARVVSSLRCAC
+TSATLPLTSFSL P ANR*S RC MASR RG SRK + FRLK ARVSSLR AC A
Sbjct: 45613 LTSATLPLTSFSLRPRANR*SRRCFSPMASRIRGSSSRKVRIFRLKRARVVSSLRRACEA

?

1st Intron **2nd Coding Exon**

Query: 281 ACVAVGVAEAVGPVTVT*PAGTAVVVRVMDSRKPAVSTT* 165
A VAVG * V+RVMDS+KPAVSTT*
Sbjct: 45793 AWVAVGAVGVAEDEGTV*AGPAGAVMRVMDSMKPAVSTT* 45909

?

Using TBLASTN alignments to identify orthologous exons and introns in an un-annotated genome

Score = 53.5 bits (127), Expect(2) = 3e-10
Identities = 29/58 (50%), Positives = 30/58 (51%)
Frame = -2

1st coding exon

splice junction



Query: 1 MSEVVRNTLVKLQAIIALIGLAAITPLLILVALLGRLIAKLCWCSAPKSIAGEV+V++ 58
MSEVVRNTLVKLQAIIALIGLAAITPLLILVALLGRLIAKLCWCSAPKSIAGEV+V++
Sbjct: 2097 MSEVVRNTLVKLQAIIALIGLAAITPLLILVALLGRLIAKLCWCSAPKSIAGEVAVVS 22



orthologous intron begins here (2252)

Score = 105 bits (261), Expect(2) = 3e-10
Identities = 52/52 (100%), Positives = 52/52 (100%)
Frame = -1

splice junction

2nd coding exon

Query: 49 SIAGEV+V++ EVMVITGAGHGLGRAISLELAKKGGCHIAVVDINVSGAEDTVKQIQDIYKVRAC 99
SIAGEV+V++ EVMVITGAGHGLGRAISLELAKKGGCHIAVVDINVSGAEDTVKQIQDIYKVRAC
Sbjct: 3991 NVMPEVMSVITGAGHGLGRAISLELAKKGGCHIAVVDINVSGAEDTVKQIQDIYKVRAC 4003



orthologous intron ends here (4003)

Using TBLASTN alignments to identify orthologous exons and introns in an un-annotated genome

Score = 53.5 bits (127), Expect(2) = 3e-10
Identities = 29/58 (50%), Positives = 30/58 (51%)
Frame = -2

1st coding exon

splice junction



Query: 1 MSEVVRNTLVKLQAIIALIGLAAITPLLILVALLGRLIAKLCWCSAPKSIAGEV**VMVIT** 58
MSEVVRNTLVKLQAIIALIGLAAITPLLILVALLGRLIAKLCWCSAPKSIAGEV+V++
Sbjct: 2097 MSEVVRNTLVKLQAIIALIGLAAITPLLILVALLGRLIAKLCWCSAPKSIAGEVAVVS 22

'1st coding exon'



orthologous intron begins here (2252)

Score = 105 bits (261), Expect(2) = 3e-10
Identities = 52/52 (100%), Positives = 52/52 (100%)
Frame = -1

splice junction

2nd coding exon

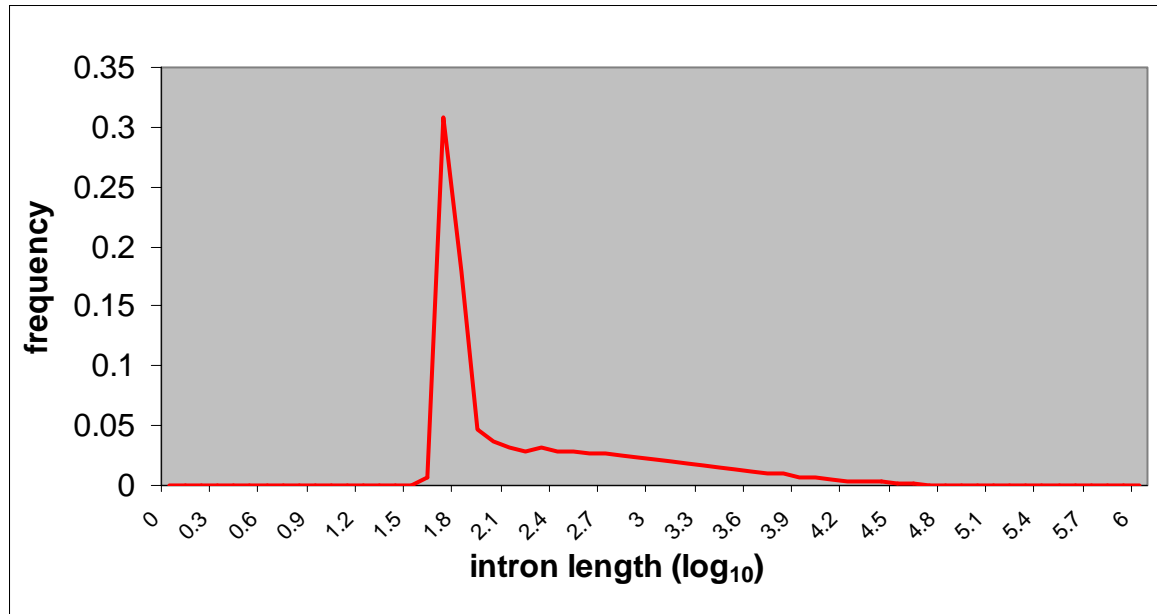
Query: 49 SIAGEV**VMVIT**GAGHGLGRAISLELAKKGGCHIAVVDINVSGAEDTVKQIQDIYKVR**AK**
+++ EVMVITGAGHGLGRAISLELAKKGGCHIAVVDINVSGAEDTVKQIQDIYKVR**AK**
Sbjct: 3991 NVMPE**VMVIT**GAGHGLGRAISLELAKKGGCHIAVVDINVSGAEDTVKQIQDIYKVR**AK**



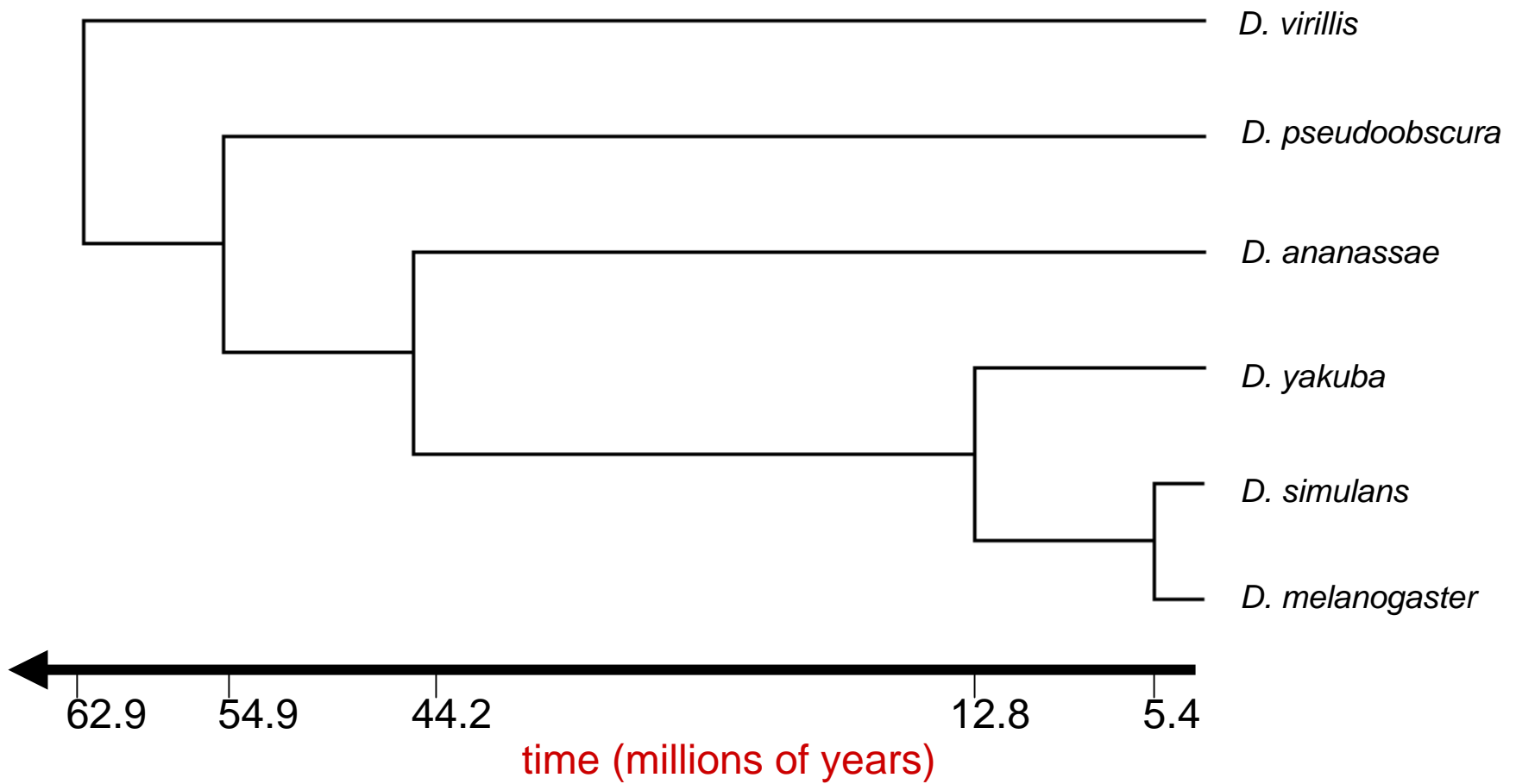
'2nd coding exon'

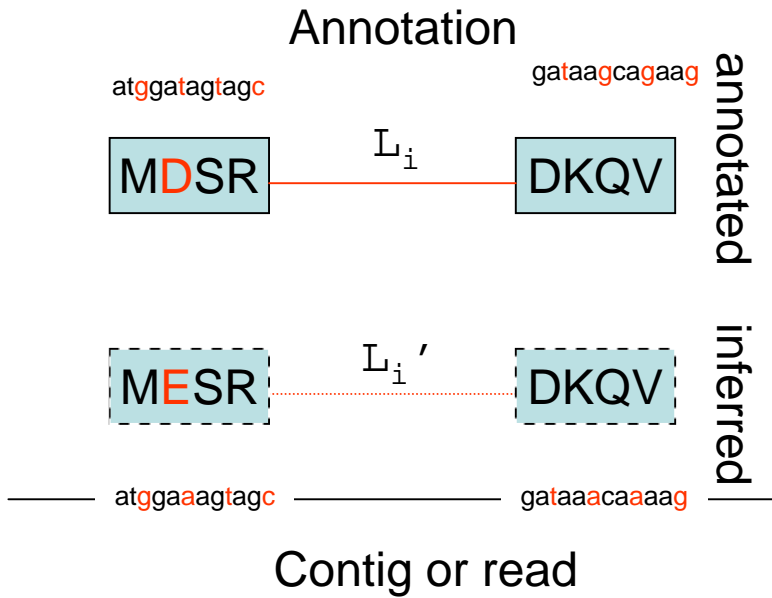
orthologous intron ends here (4003)

Part II
Exploring recent history with **CGL**

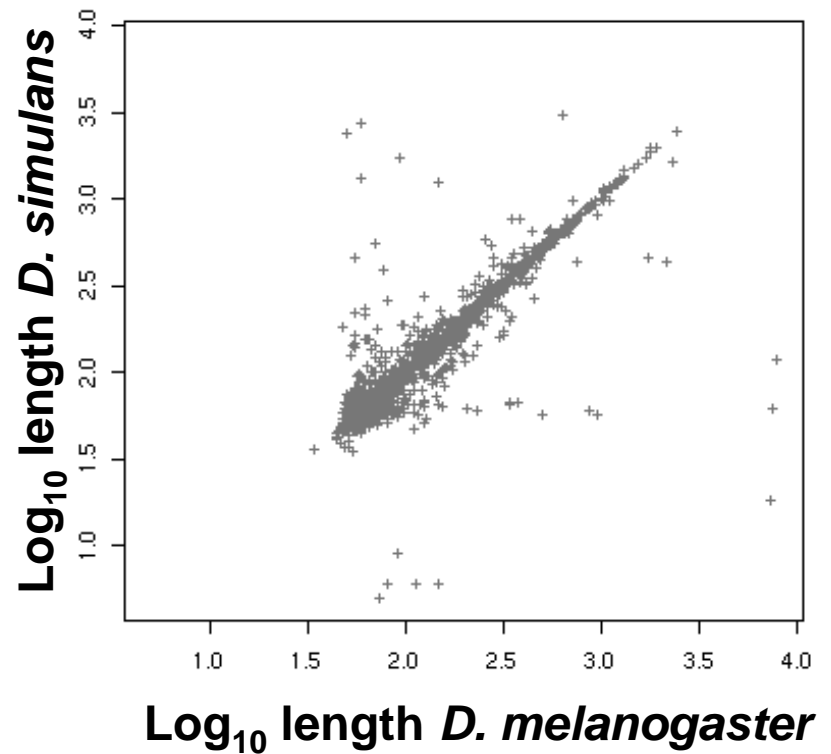


***D. melanogaster* Intron length distribution**

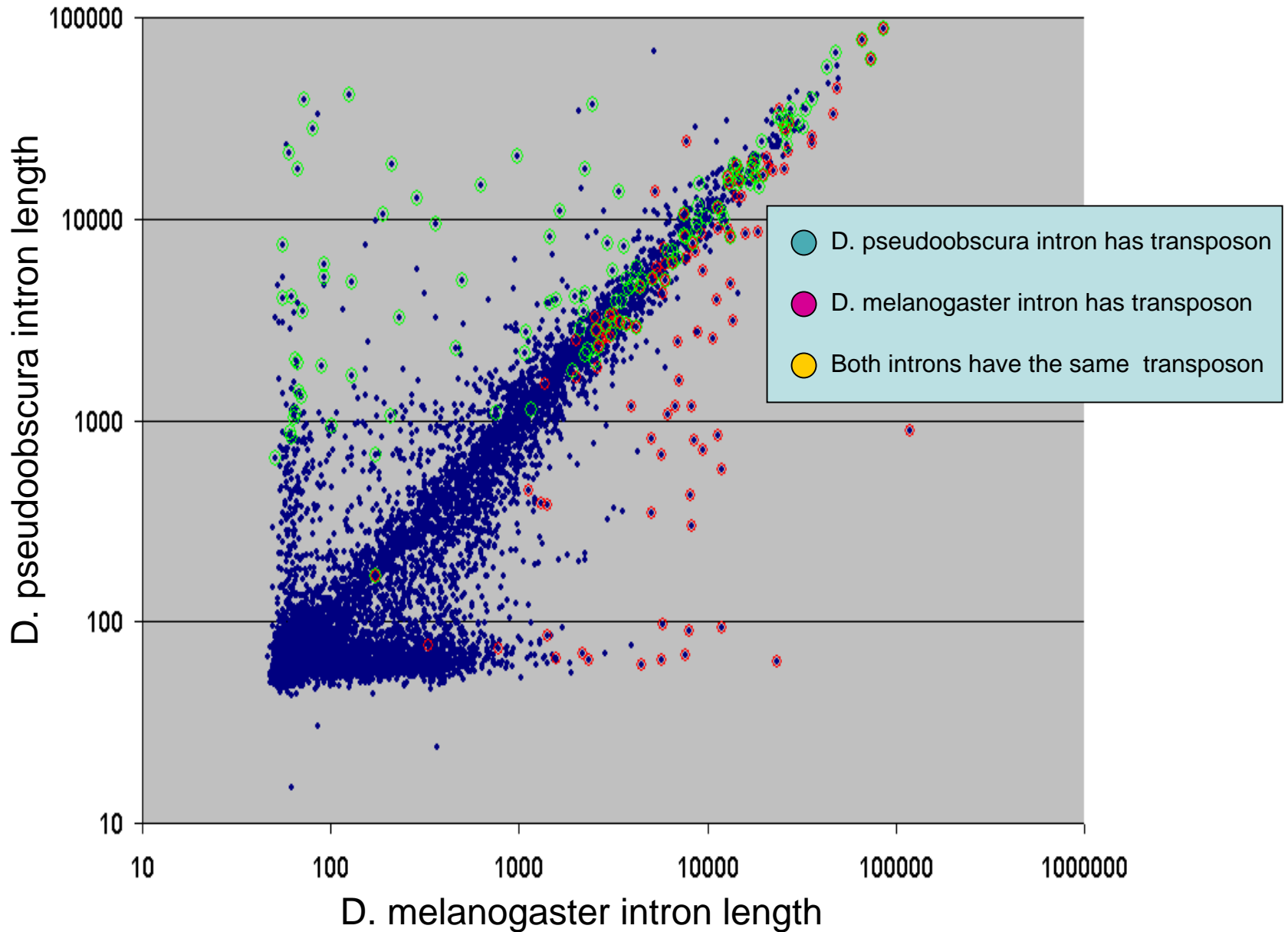




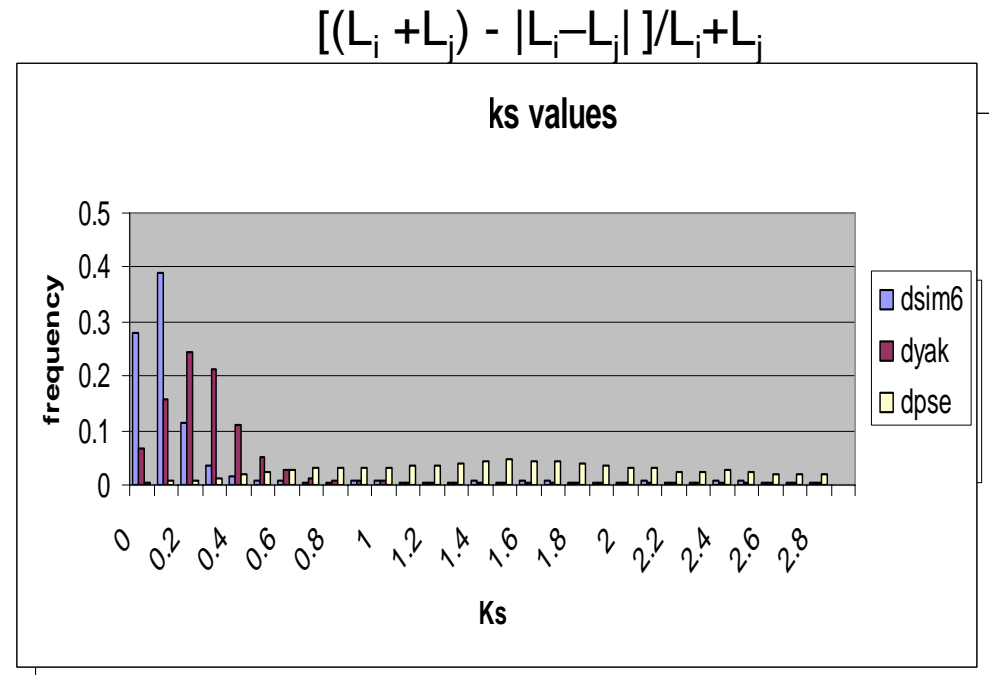
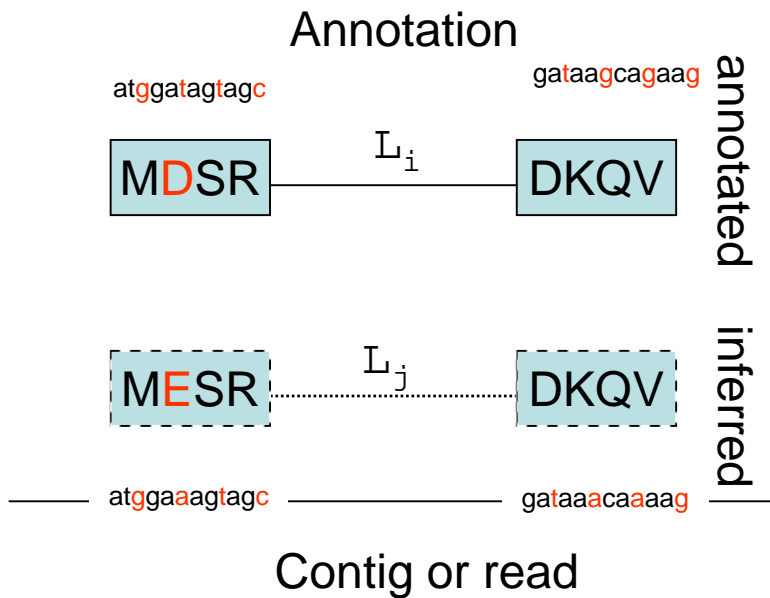
Compare intron lengths



Why do intron lengths change over time?

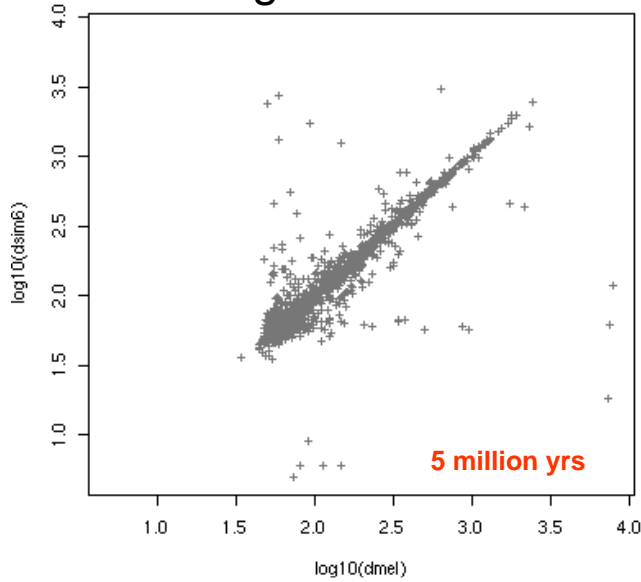


Does selection on the protein influence intron lengths ?

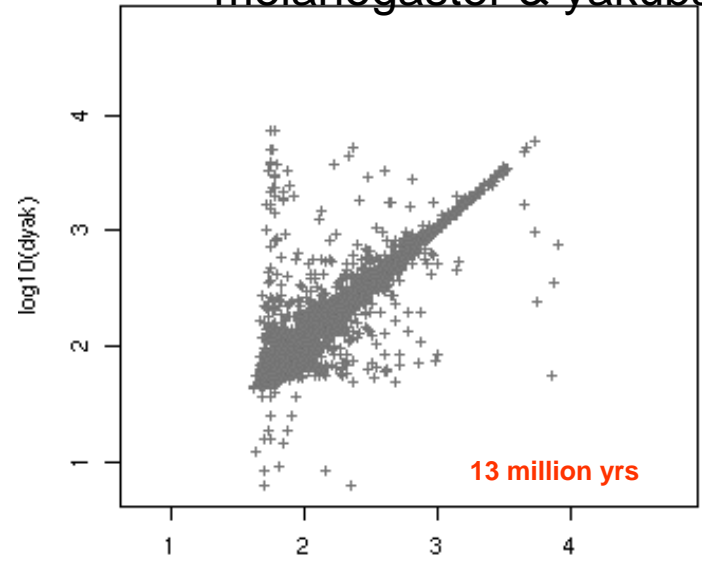


A new evolutionary clock?

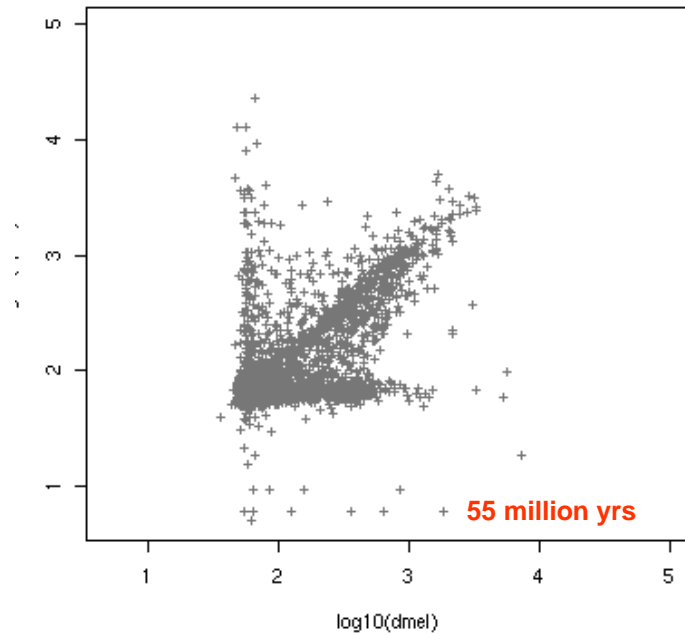
melanogaster & simulans



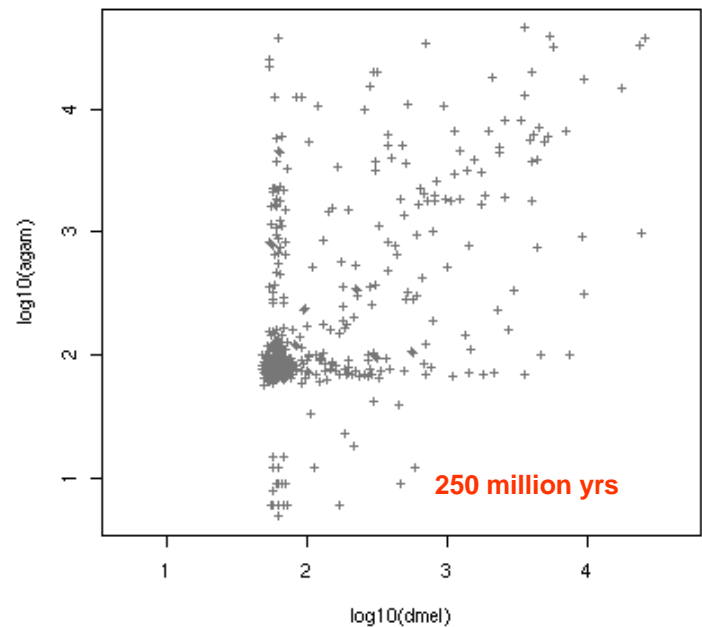
melanogaster & yakuba



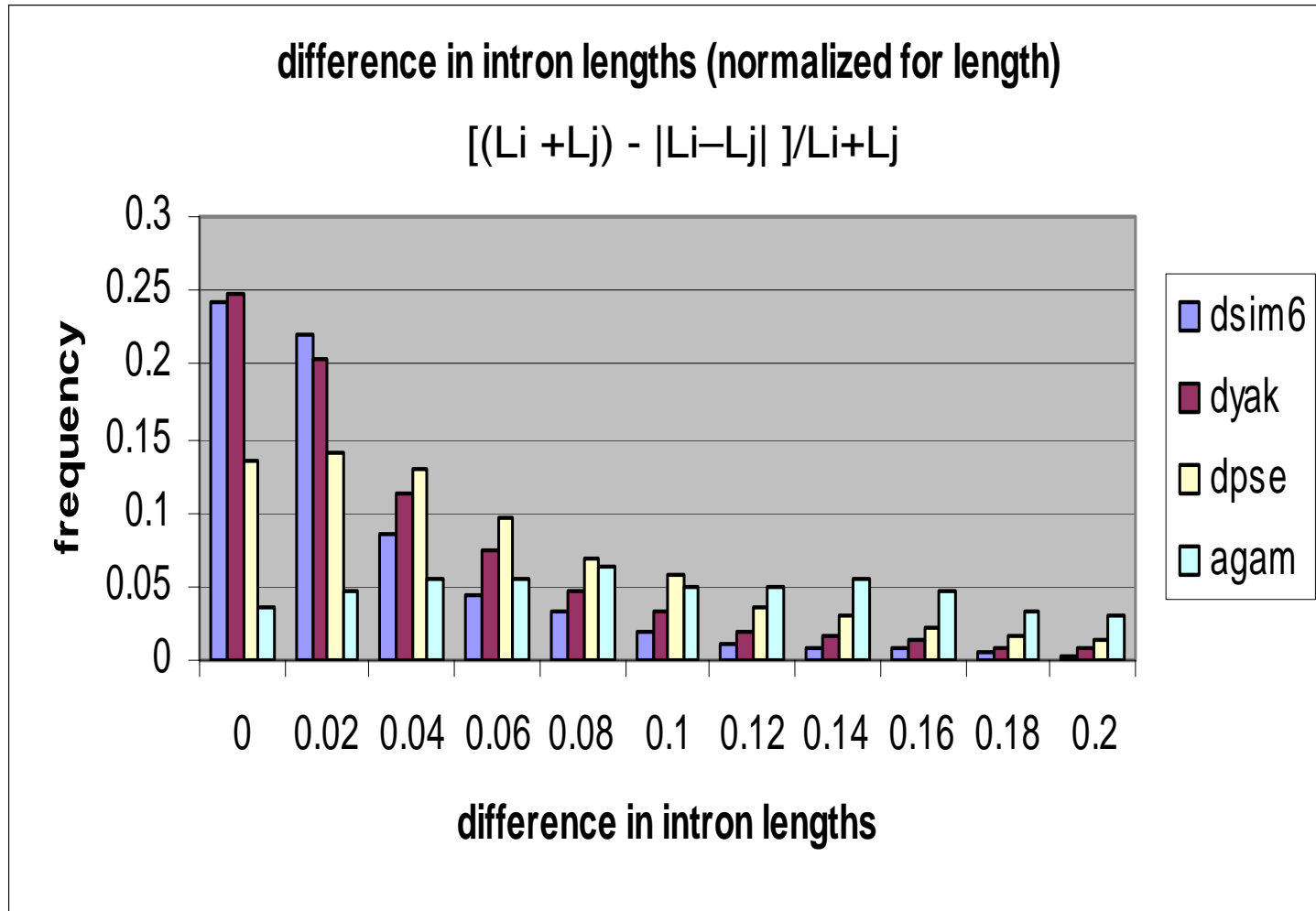
melanogaster & pseudoobscura

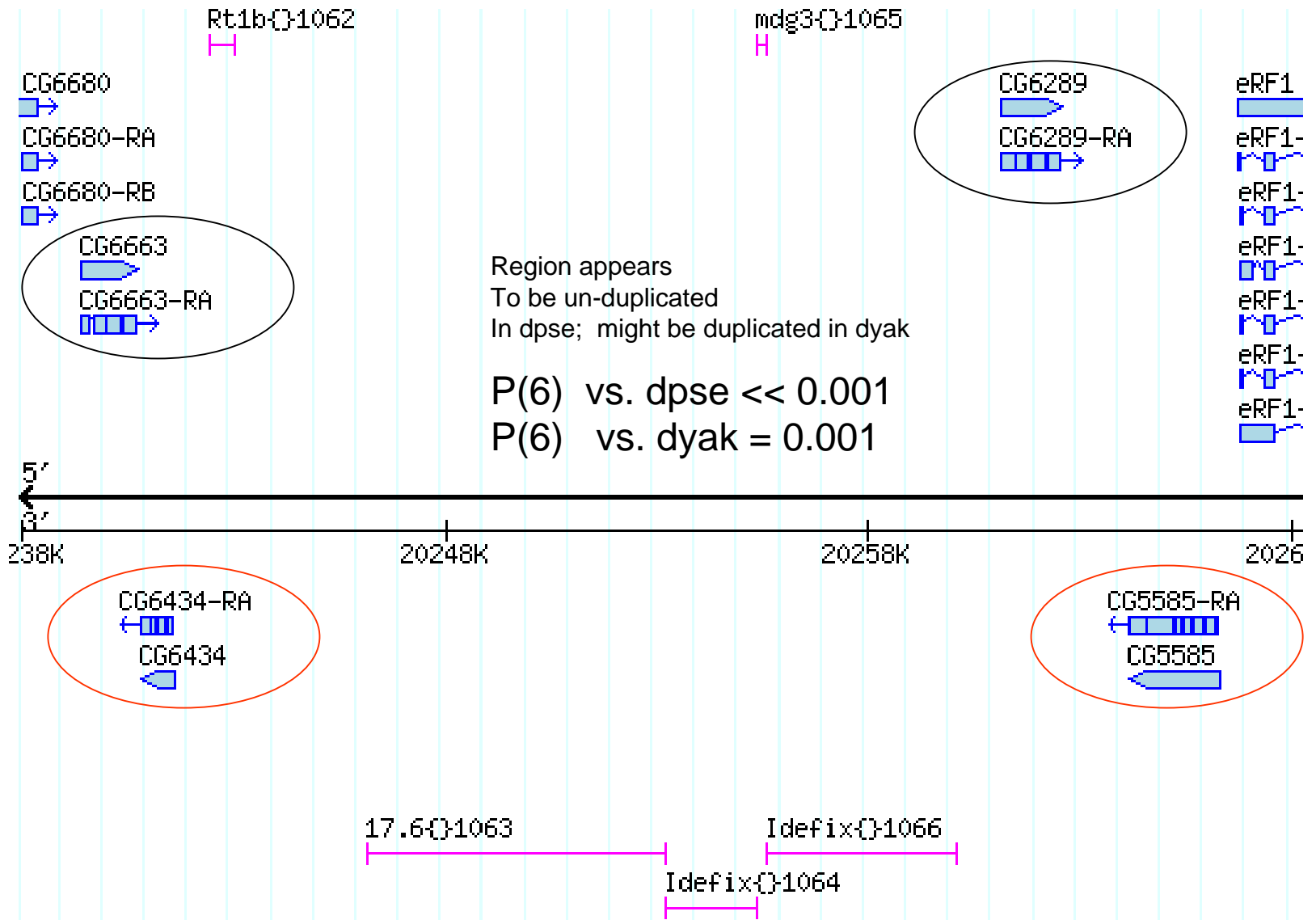


melanogaster & mosquito



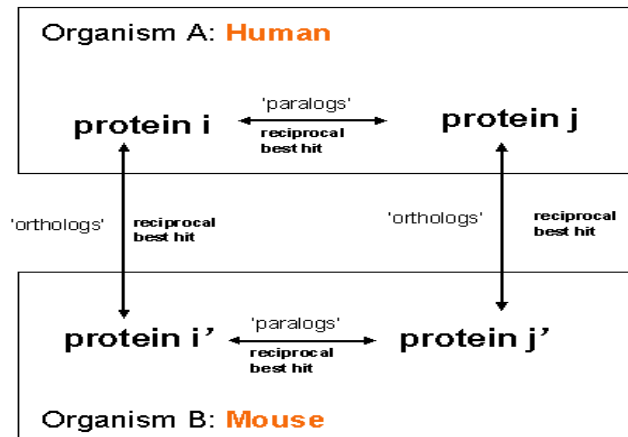
A new evolutionary clock



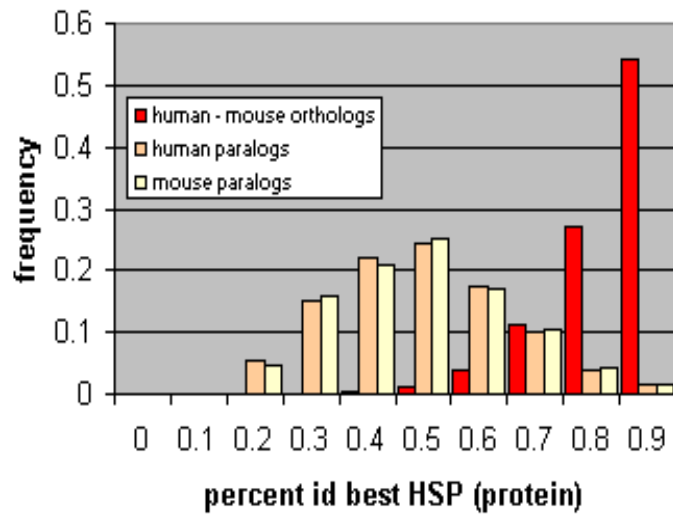
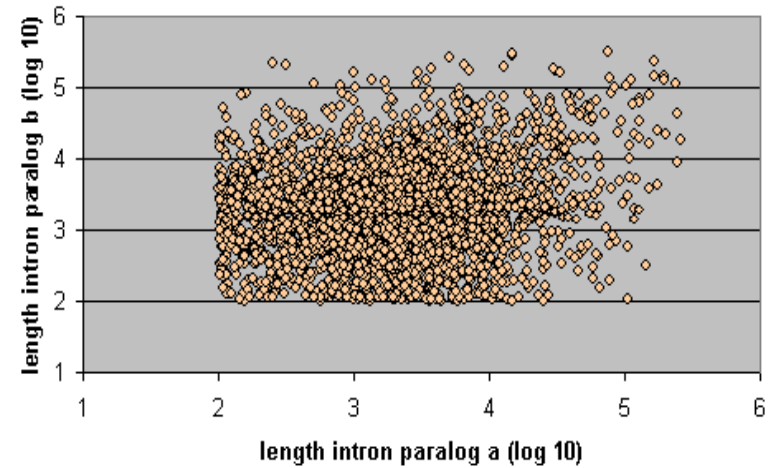


KEY:

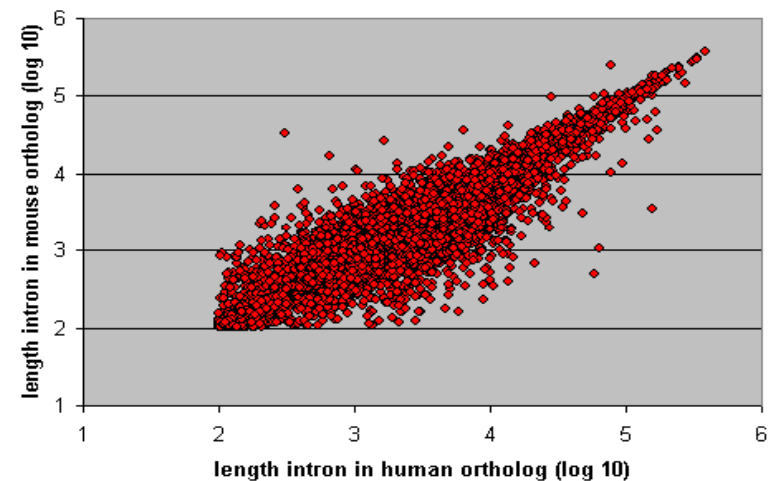
QUARTETS



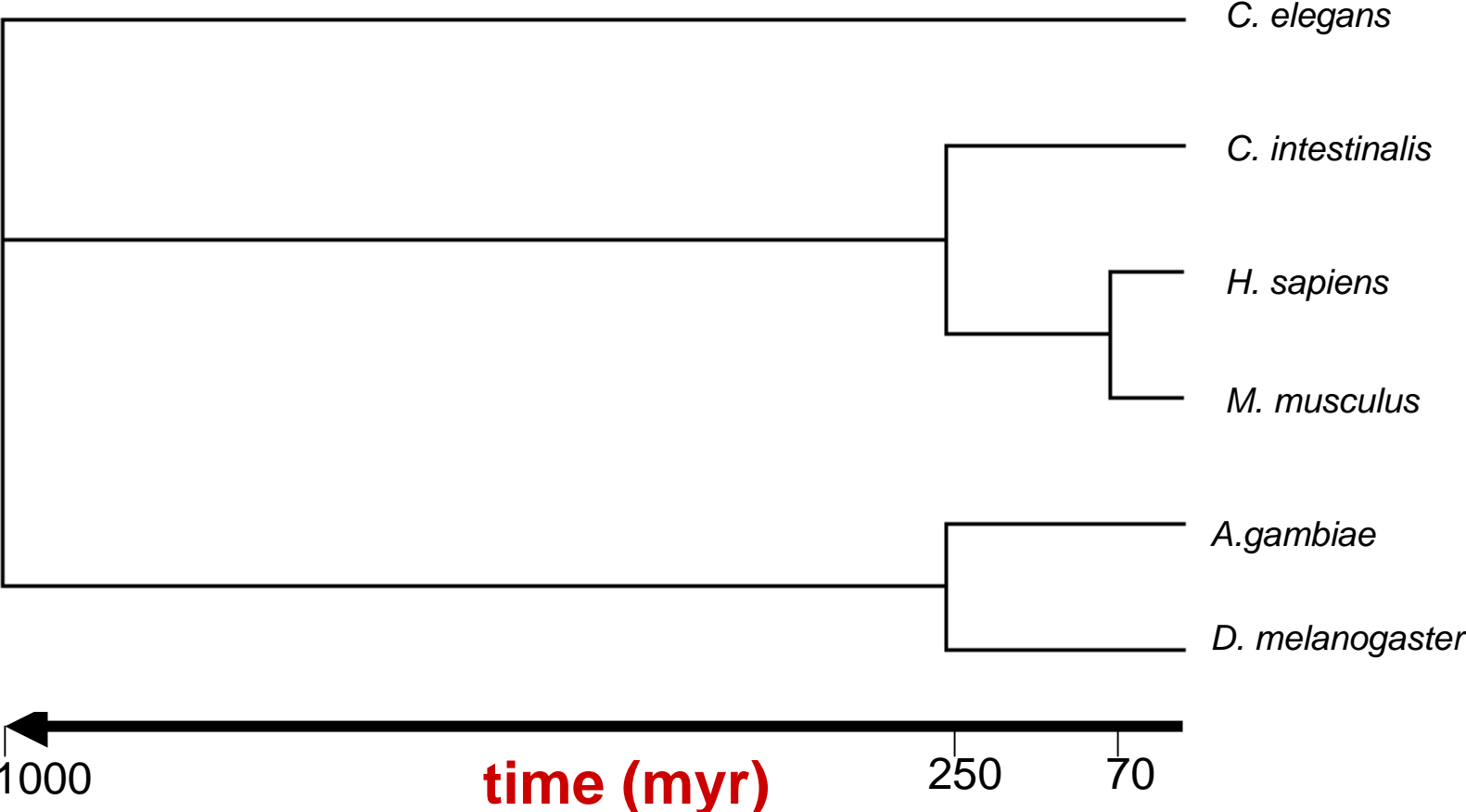
Human quartet members (paralogs)

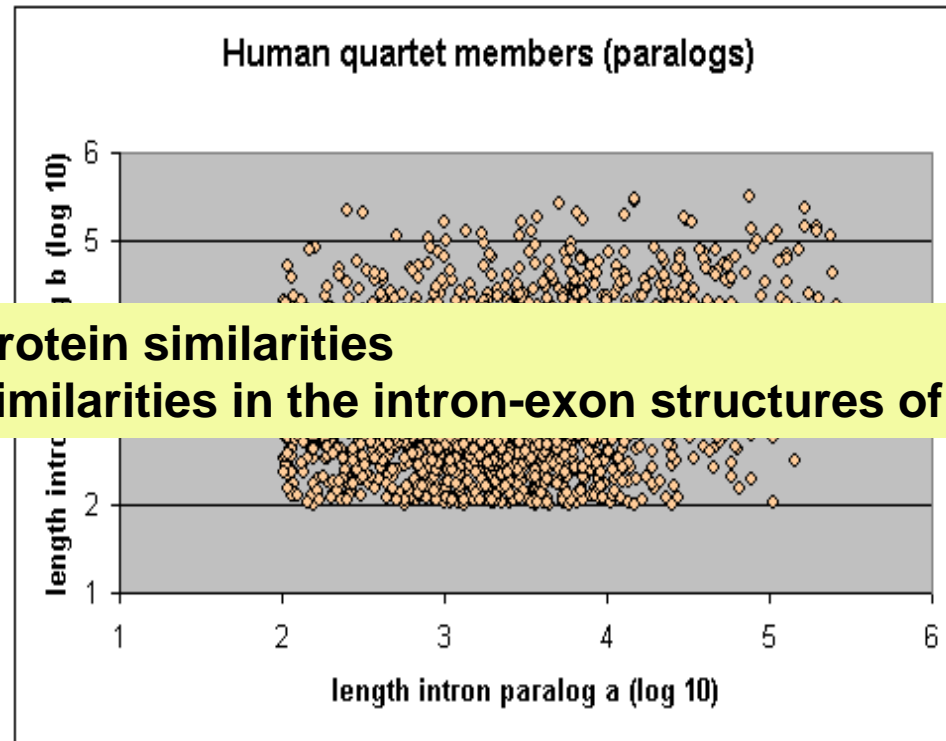


Human quartet members (orthologs)

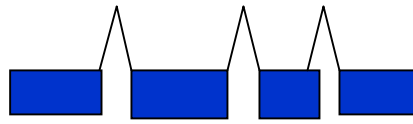


Part III
Exploring ancient history with **CGL**

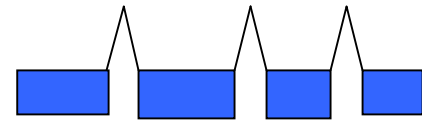




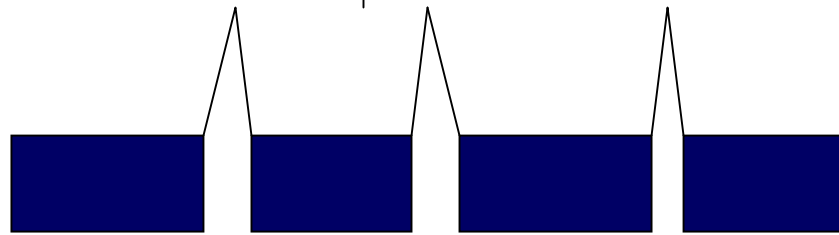
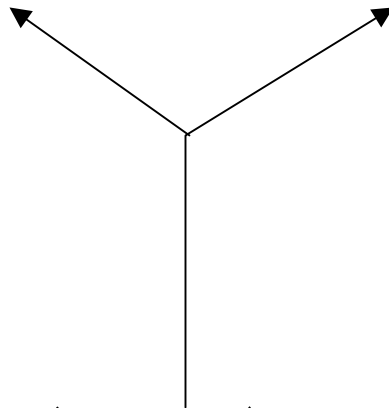
- protein similarities
- similarities in the intron-exon structures of genes



MIKEVFRPDKFGMDL

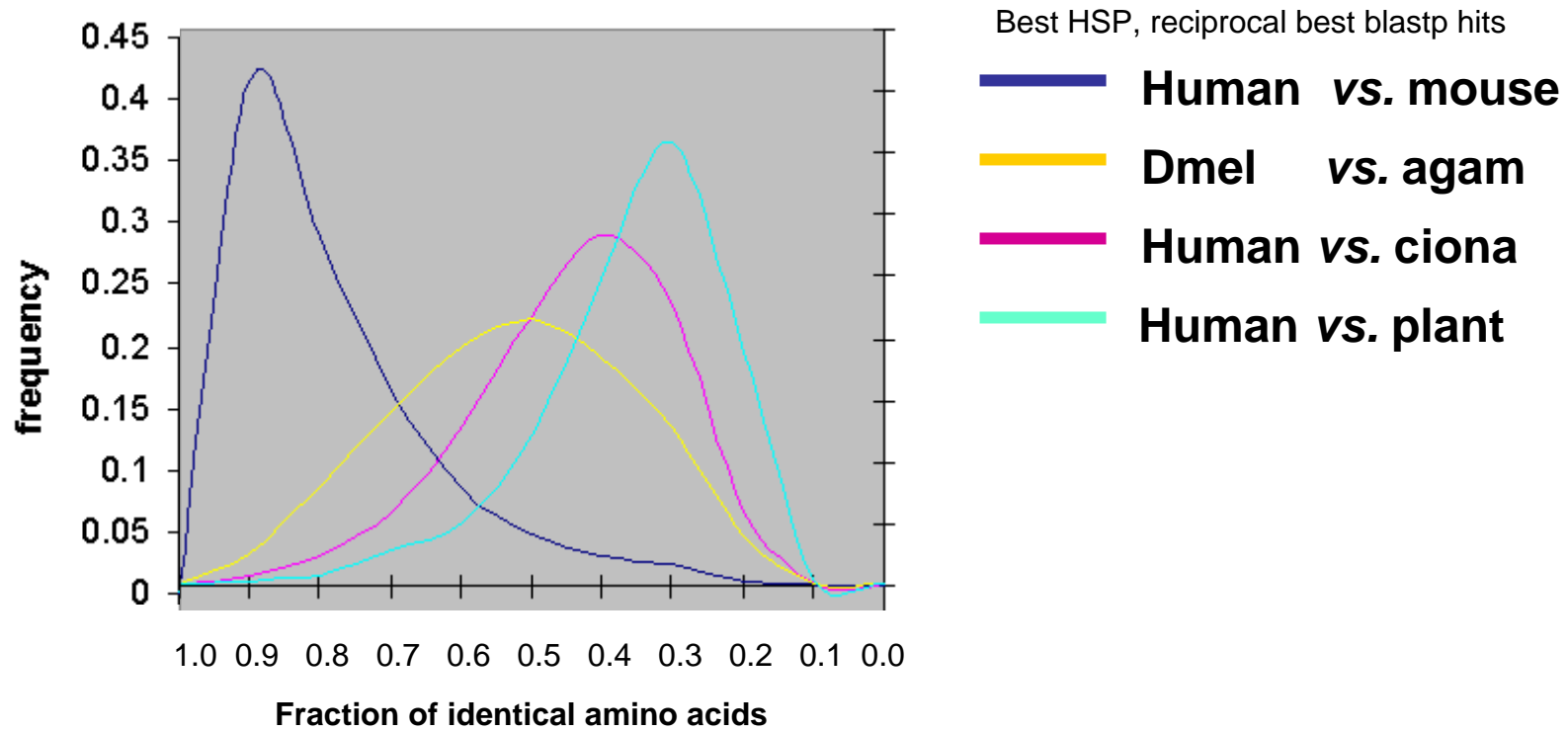


MILEVFRPDKFGMDL

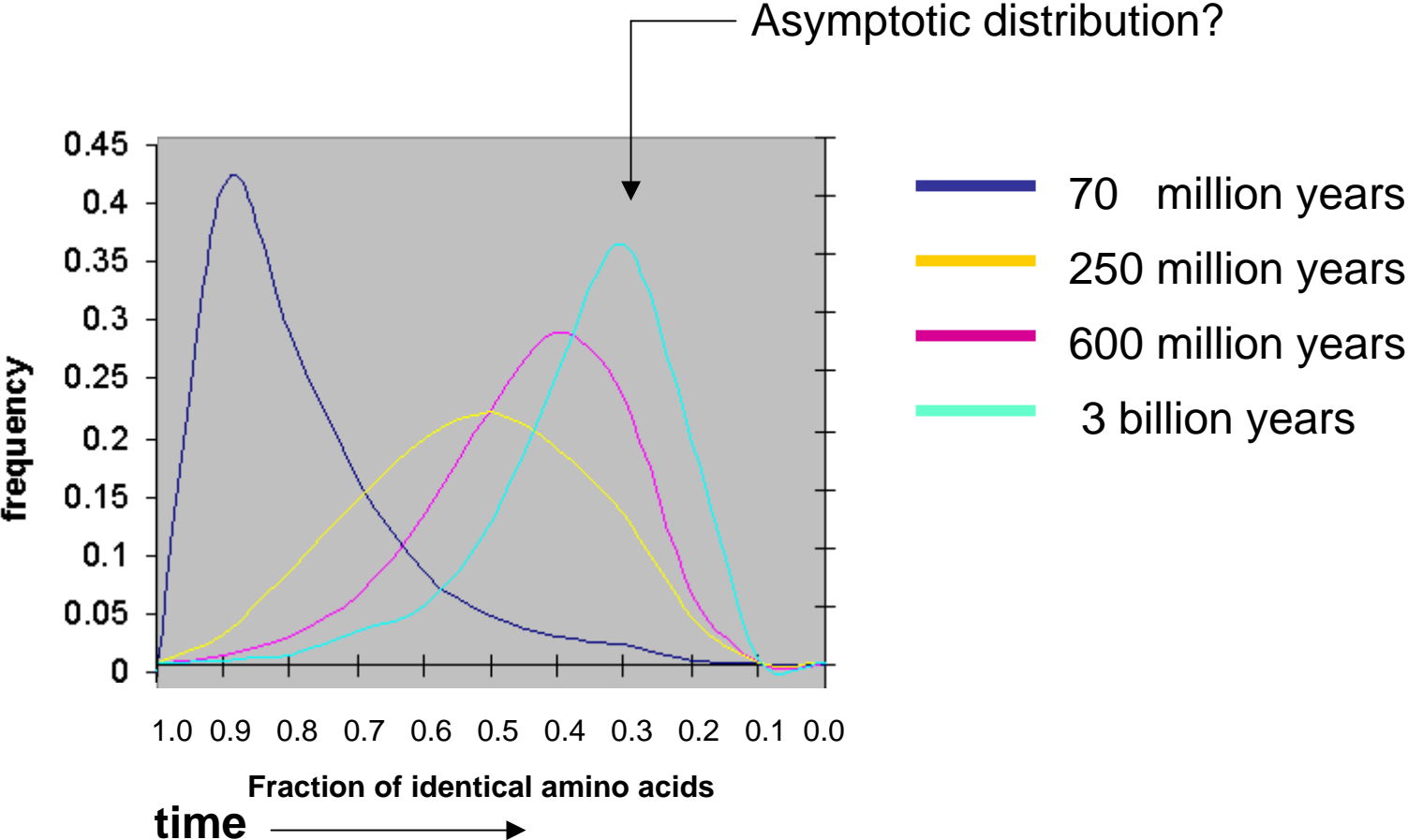


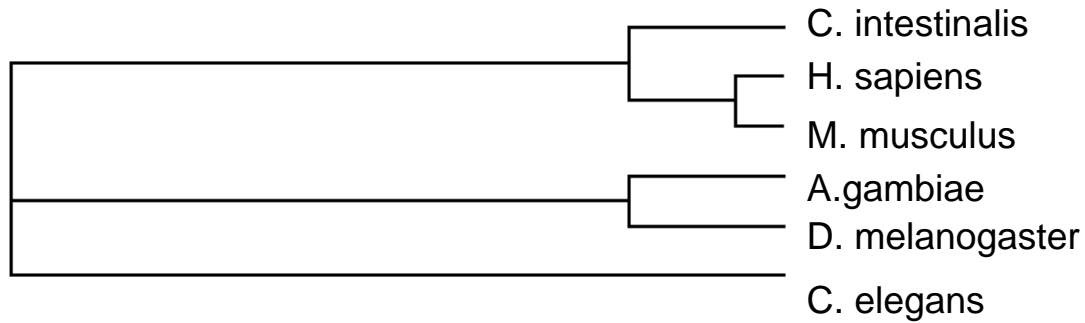
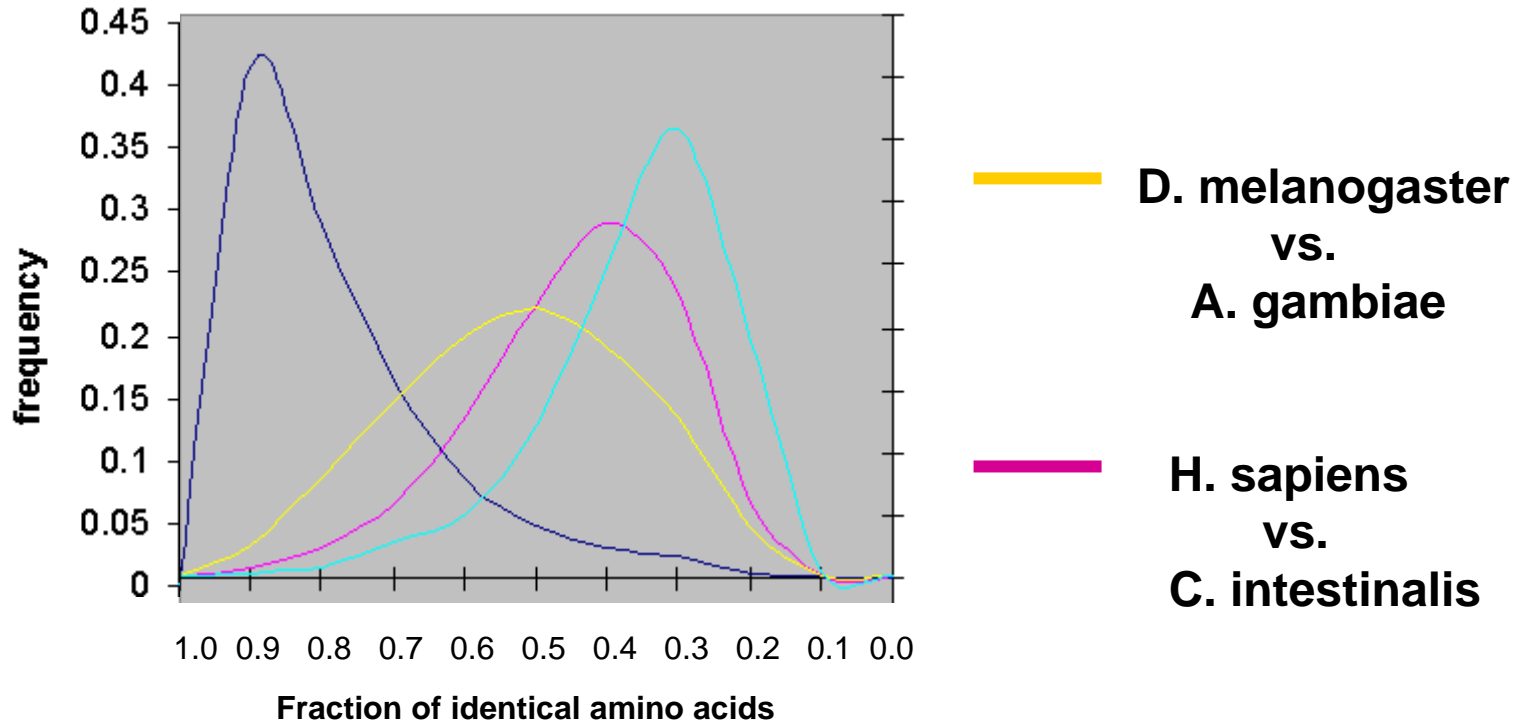
MIKDVFRPDKFGIDL

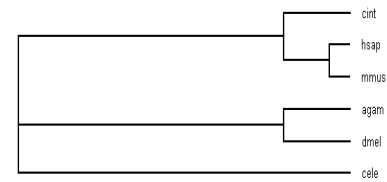
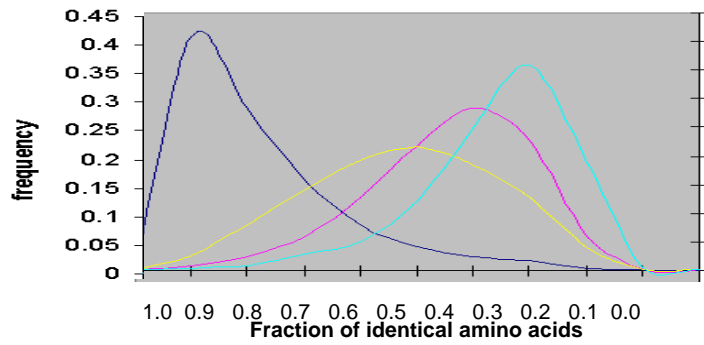
The evolutionary trajectory of protein similarities



The evolutionary trajectory of protein similarities









Every reciprocal best hit

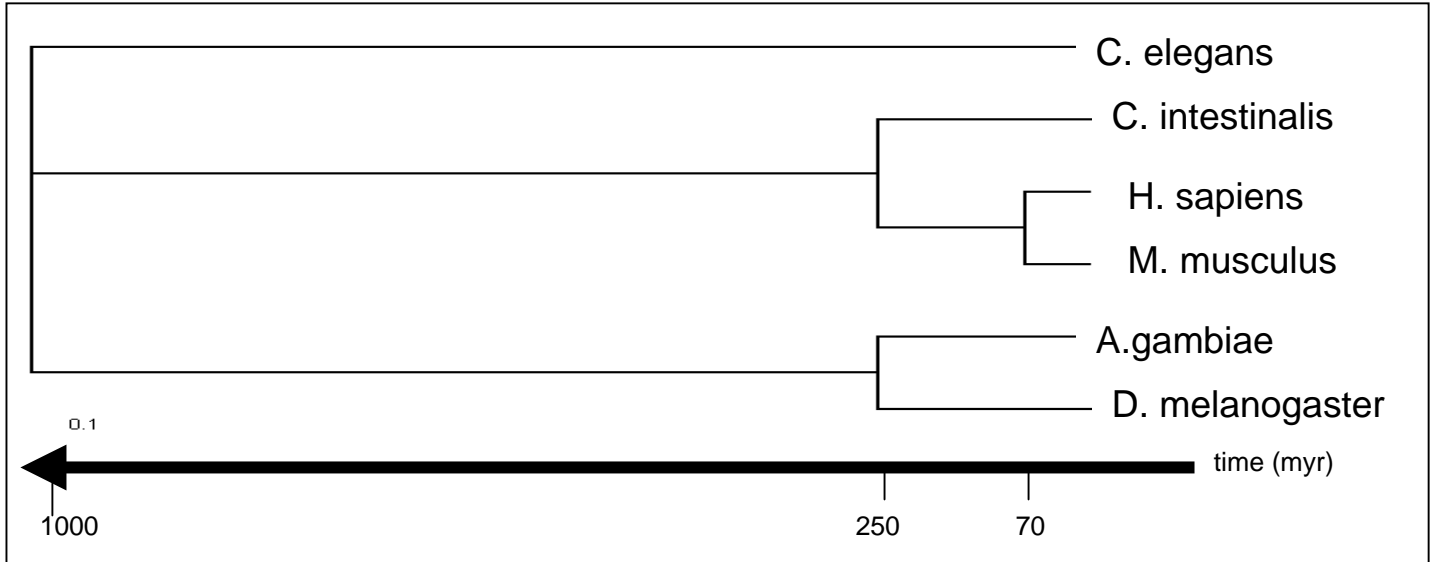
Query: 1 MDEWRQWKNDKQVWDAER 18
 MD+WR WKNDKQVWD ER
Sbjct: 1 MDDWRGQWKNDKQVWDMER 18

For every HSP

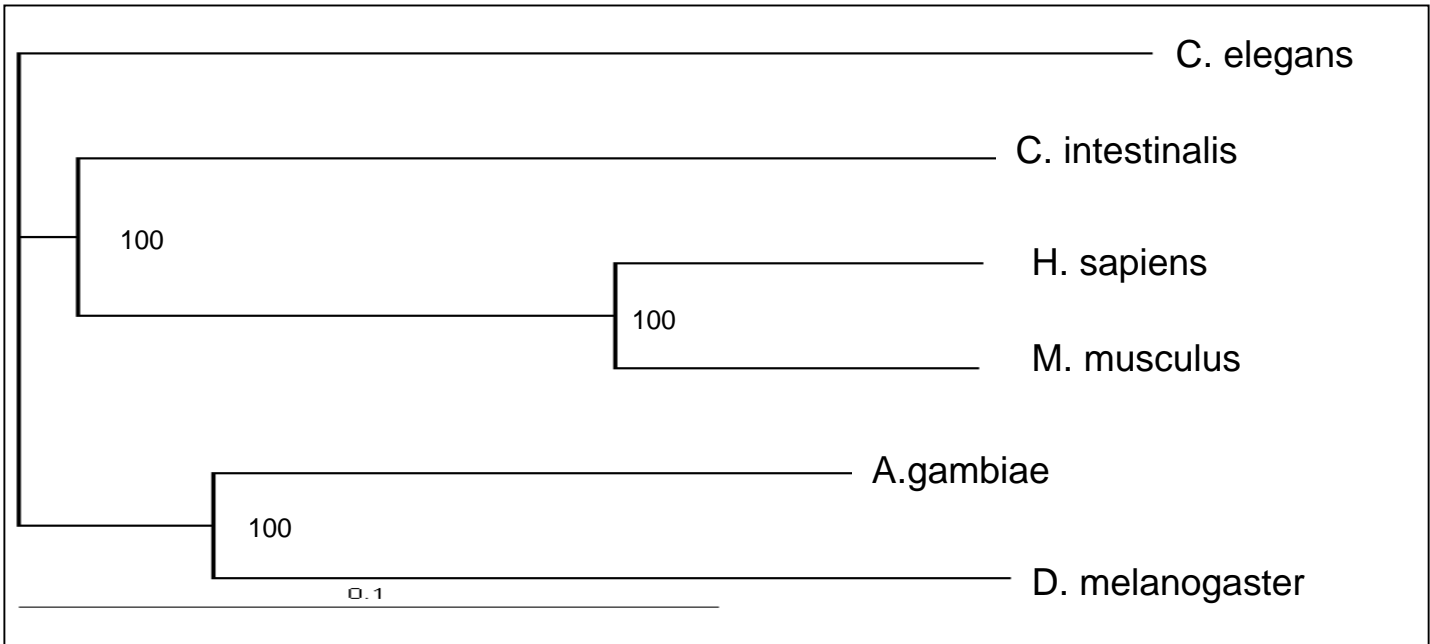
$$H = \sum_{i=1}^{20} \sum_{j=1}^i q_{ij} \log_2(q_{ij} / p_i p_j)$$

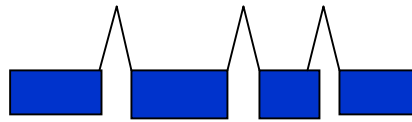
Global similarity between the two organisms' proteomes

'standard'

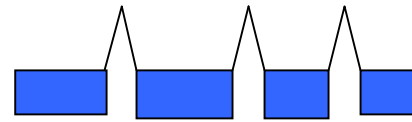


H

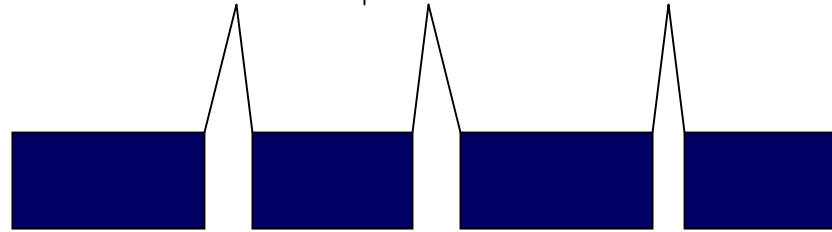
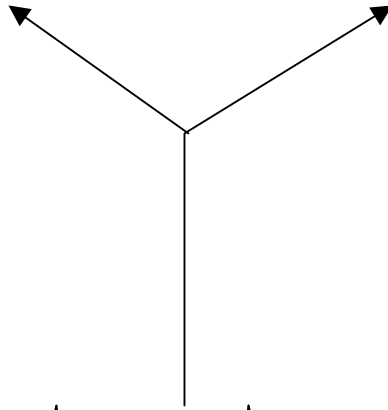




MIKEVFRPDKFGMDL



MILEVFRPDKFGMDL



MIKDVFRPDKFGIDL

Organism A

Organism B



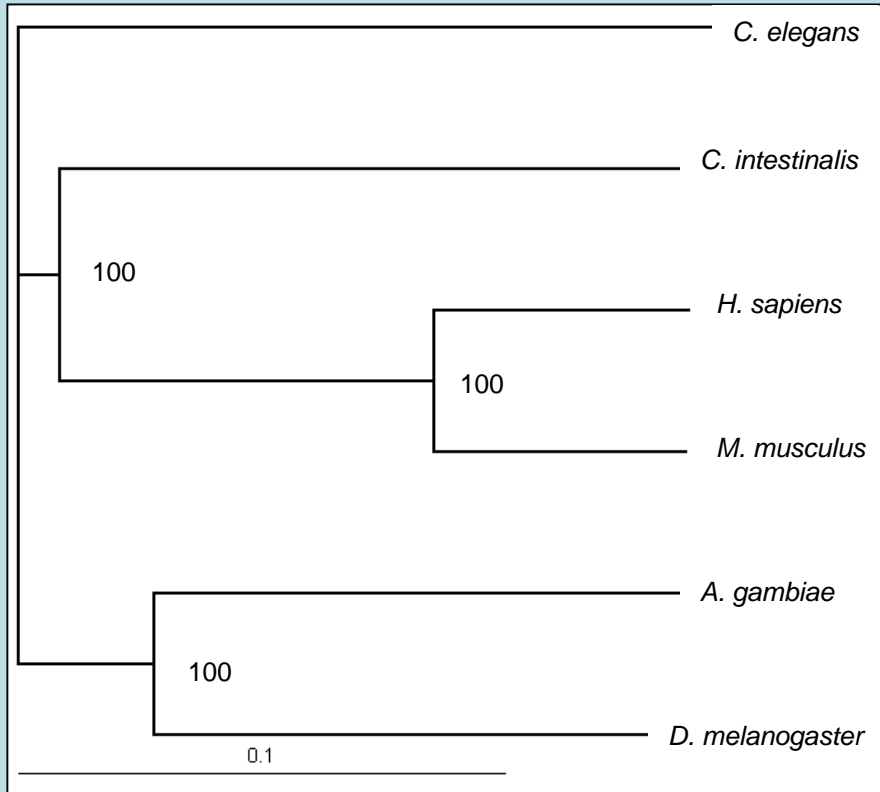
Every reciprocal best hit



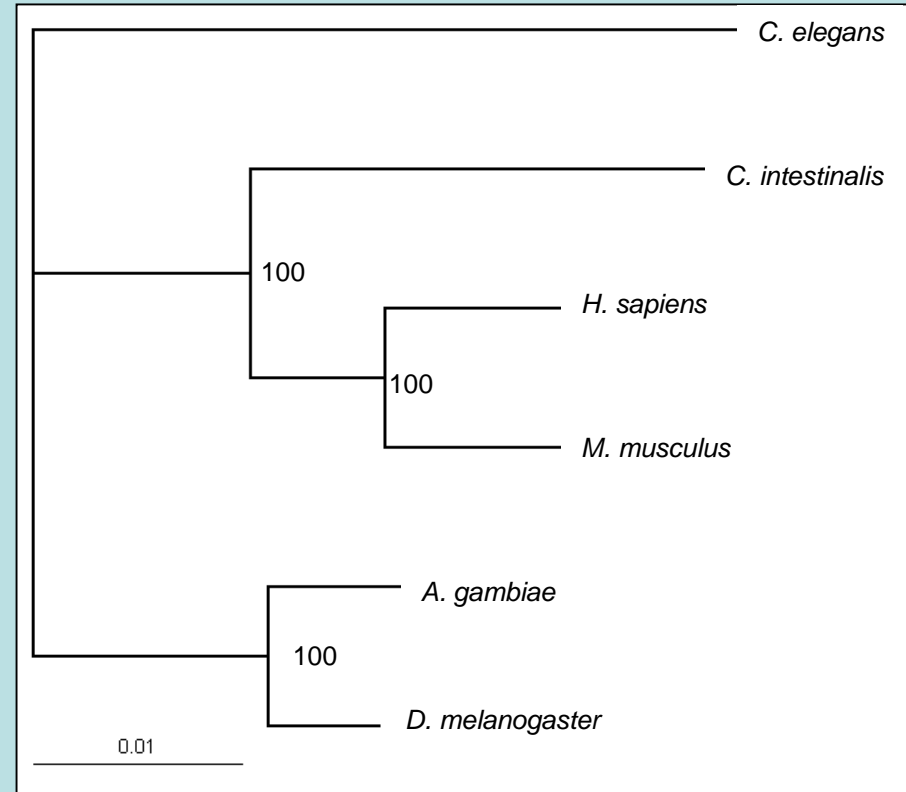
$$I = \frac{q_{\text{intron-intron}}}{p_{\text{intron}}^{\text{query}} * p_{\text{intron}}^{\text{subject}}}$$

Global similarity between the two organisms' intron-exon structures

Good trees can be made from both measures; these measures provide a rapid and robust means to calculate phylogenetic trees using the combined data from many annotated genomes simultaneously.



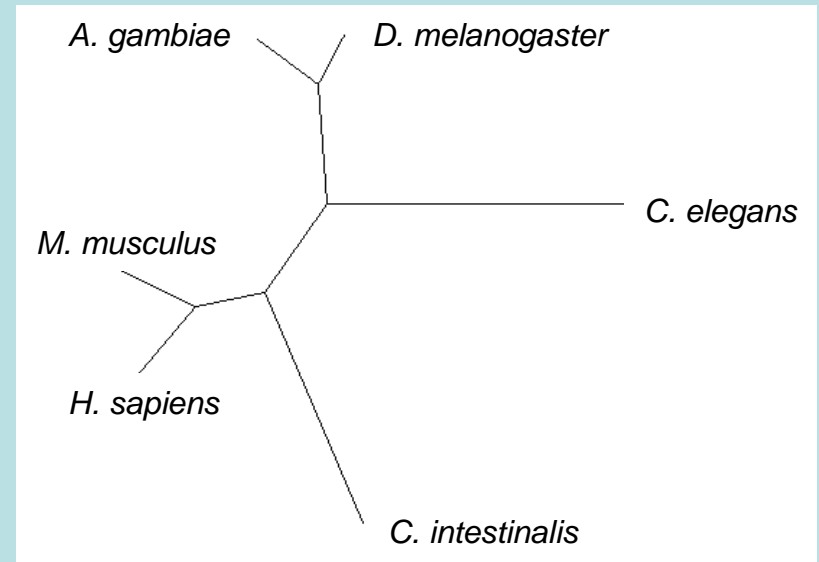
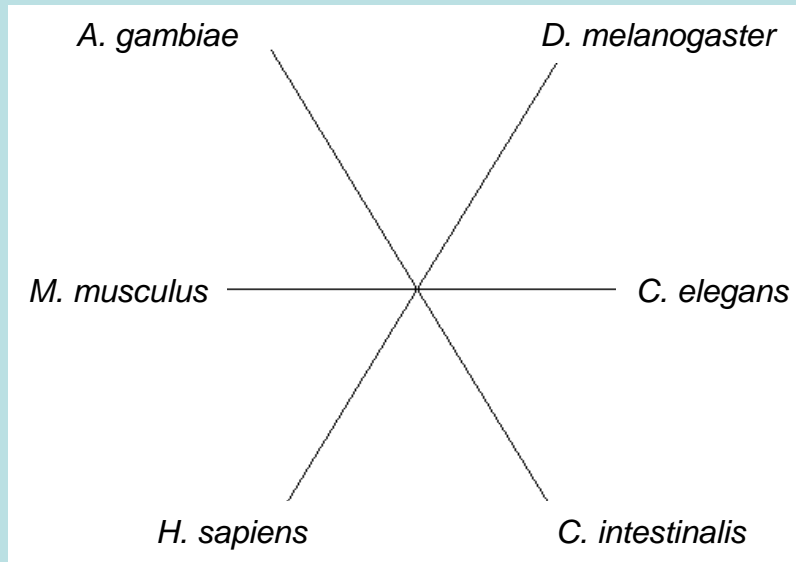
H



I

Trees based on I provide an independent check on trees made from amino-acid similarities; note that deep nodes are better resolved in the I tree.

Like intron lengths, gene-structures seem to evolve independently of protein Sequence— hold H constant and you still get the same I tree



H = 1.5
~ 50% similarity

I

New Methods for Comparative genomics

Part I: CGL: a software library for comparative genomics

- Genome annotations make possible any new kinds of genome analyses
- We have developed a (soon to be) publicly available software library called CGL; 'seagull' that greatly facilitates such analyses.

Part II: Using CGL to explore recent history (1-70 myr years)

- Intron length can be used as an evolutionary clock.
- Seems to evolve independently of protein sequence.
- Can be used to confirm and calibrate existing protein clock approaches.
- Can be used to identify recent gene duplication events.
- Can be used to strengthen and extend with many 'standard' protein based approaches, e.g. quartet analysis.

Part II: Using CGL to explore ancient history (70-1000 myr years)

- Proteins may saturate completely after ~ 1 billion years or so.
- Resolving deep evolutionary questions w/ proteins means computing 'consensus' trees; we offer one approach to this problem using 'H'.
- Like intron length, the intron-exon structures of genes contains a phylogenetic signal; the 'I' vs. 'H' trees
- This signal seems to evolve independently of protein sequence, and perhaps more slowly.

Large scale comparative genomics has much to tell us about the evolution of both genes and organisms.

Acknowledgements

Chris Mungall
Simon Prochnik
Chris Smith
Josh Kaminker
George Hartzell
Suzi Lewis
Gerald M. Rubin

