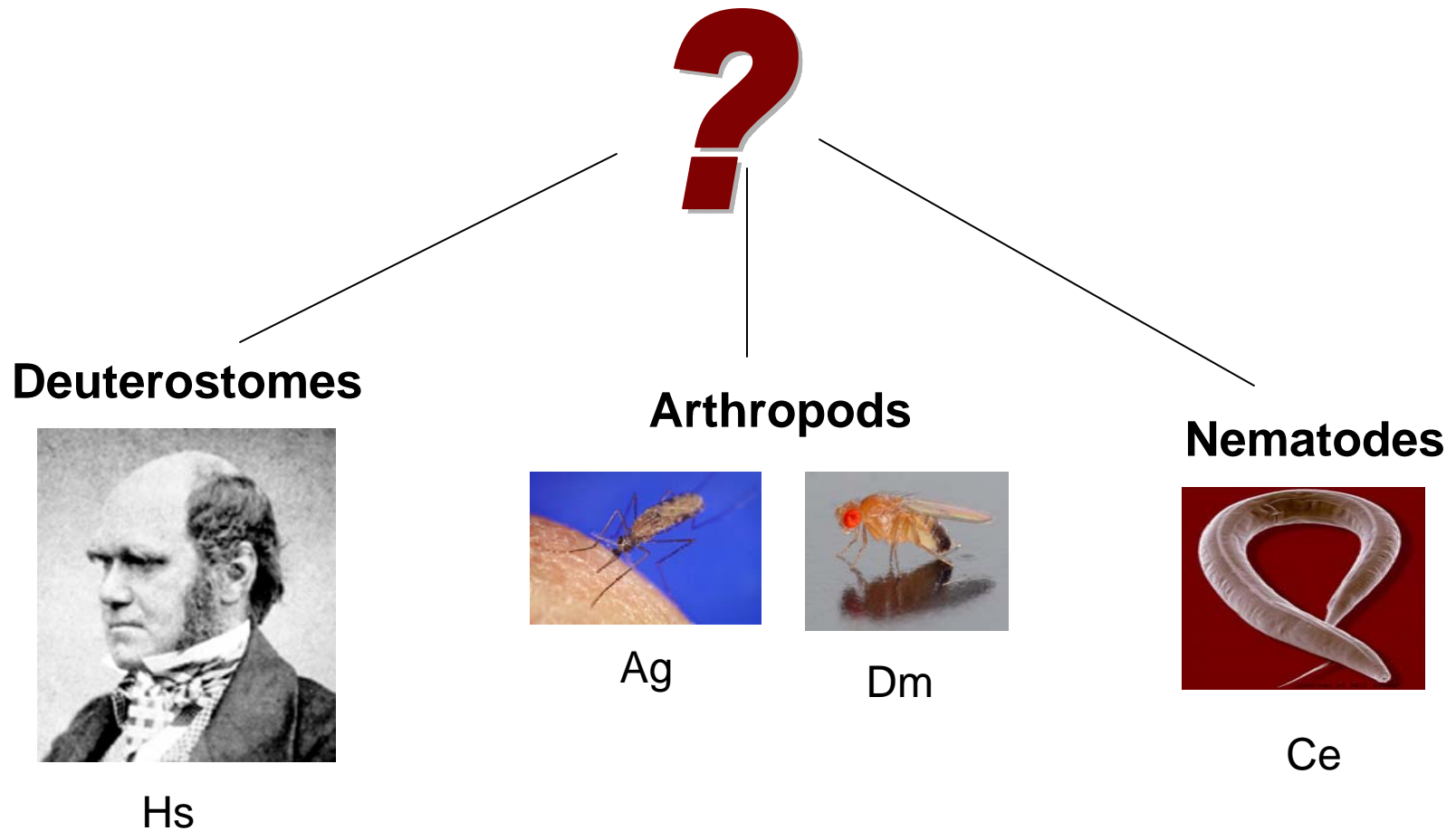


Intron Pattern Analysis Supports the Coelomata Clade of Animals

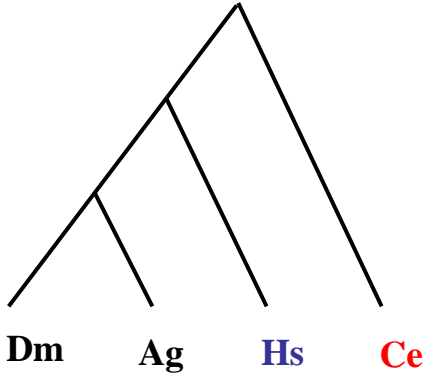
Jie Zheng
NCBI,NLM,NIH
with
Igor B. Rogozin
Eugene V. Koonin
Teresa M. Przytycka



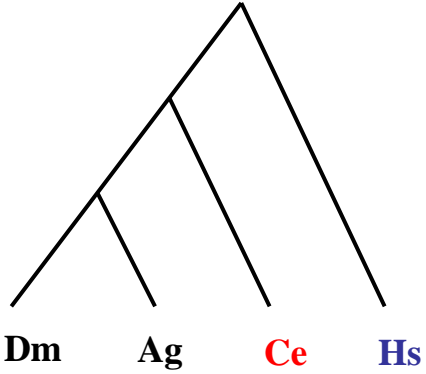
3 groups of animals in Tree of Life



Hot, persistent debate



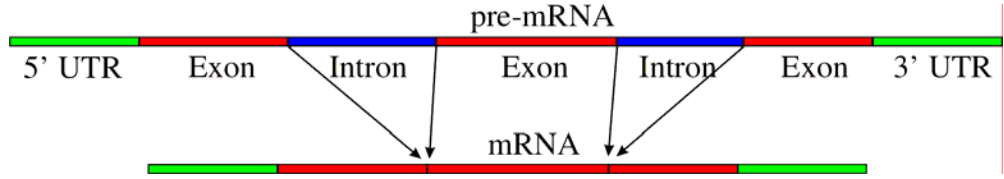
Coelomata



Ecdysozoa

Outline

- Intron data
- Retention dependence between branches
- Conserved introns support Coelomata
- Tests support Coelomata



Intron positions 33 55 144 169

Pf MSRRTKKVGLTGKYGTRYGSSLRKQIKKIELMQHAKYLCTFCGKTATKRKTCVGIWKCK-

At MTKRTRKKARIVGKYGTRYGASLRKQIKKMEVVSQHNYFCEFCGKYSVVKRQVGIWCK-

Sc MAKRTRKKVGITGKYGVRYGSSLRQVKKLEIQQHARYDCSFCGKKTVKRGAAGIWTCS-

Sp MTKRTRKKVGVTVGKYGVRYGASLRDVRKIEVQQHSRYQCPFCGRLTVKRRTAAGIWKCSG

Ce MAKRTRKKVGIIVGKYGTRYGASLRKMAKKLEVAQHSRYTCSFCGKEAMKRKATGIWNCA-

Dm MAKRTRKKVGIIVGKYGTRYGASLRKRVKMEITQHSKYTCSFCGKDSMKRAVGIWCK-

Ag YLPKMAKRTRKVGIVGKYGTRYGASLRKRVKMEITQHAKYTCTFCGKDAMKRSCVGIWCK-

Hs MAKRTRKKVGIIVGKYGTRYGASLRKRVKKEIISQHAKYTCSFCGKTMMKRRAVGIWHCG-



	33	55	144	169	233
Pf	1	0	1	0	0
At	0	1	1	0	0
Sc	0	0	0	0	0
Sp	0	0	0	1	0
Ce	0	0	0	0	0
Dm	0	0	1	0	0
Ag	0	0	1	0	0
Hs	0	0	1	0	1

- Alignment of orthologous proteins (KOG) to genomes
- 585 orthologous genes

13 Species

Ag



Dm



Honey Bee



Sea Urchin



Hs

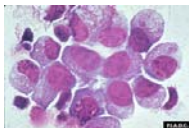


Ce

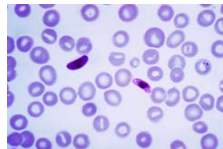
At



Tp

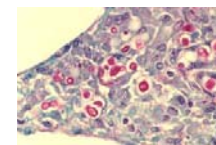


Pf

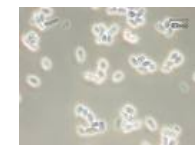


Fungi

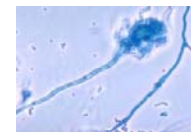
Cn



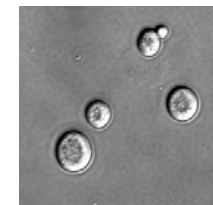
Sp



Af



Sc

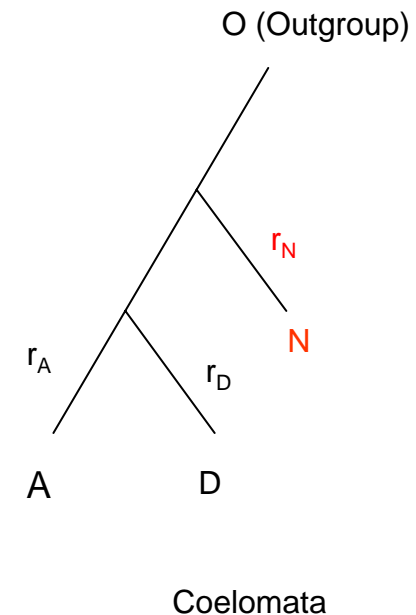


Related work

- Rogozin et al, *Curr. Biol.* 2003
- Roy and Gilbert, *PNAS* 2005
- Przytycka, *RECOMB* 2006

Notation

- Based on Roy & Gilbert, *PNAS* 2005
- ADO is # of introns present in A, D, and O, but not in N
- $ADO/DO = r_A / (1 - r_A)$



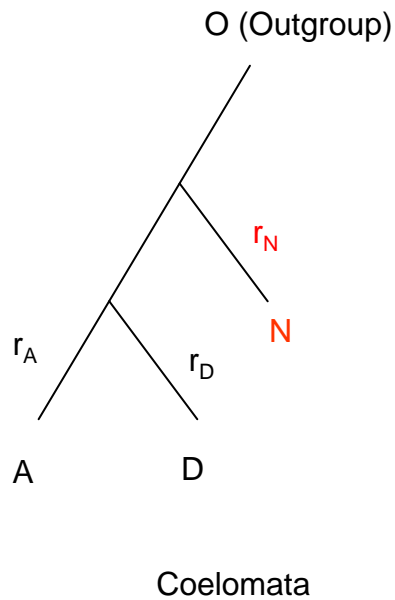
A: Arthropods (Dm, Ag)

D: Deterostomes (Hs)

N: Nematodes (Ce)

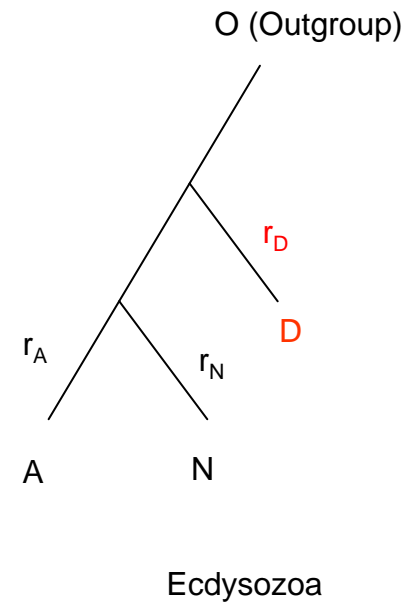
If retentions are independent between every two branches

...



$$r_A / (1 - r_A) = ADO / DO = ADON / DON$$

$$r_D / (1 - r_D) = ADO / AO = ADON / AON$$



$$r_A / (1 - r_A) = ANO / NO = ANOD / NOD$$

$$r_N / (1 - r_N) = ANO / AO = ANOD / AOD$$

Real data reject the hypothesis of retention independence

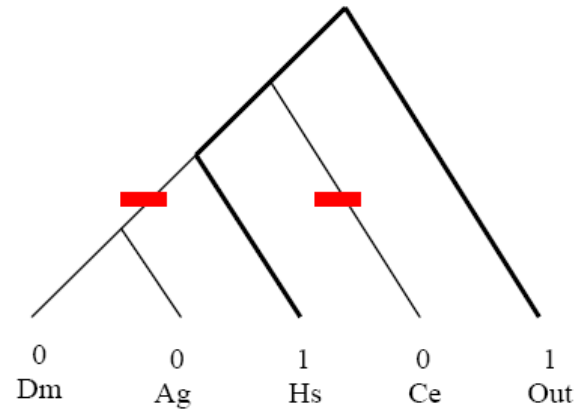
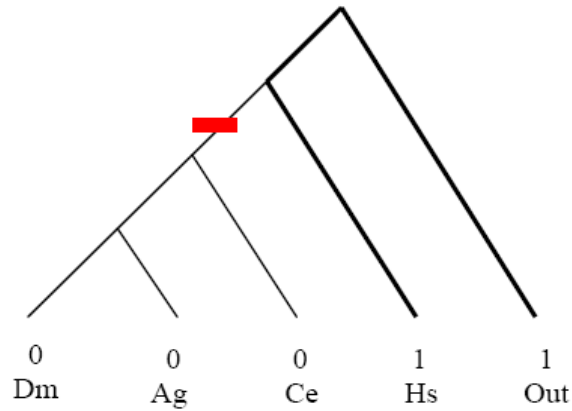
Ecdysozoa			
	Non-conserved in D	Conserved in D	p-value in Fisher test
ratio	ANO/NO	ANOD/NOD	
$r_A / (1 - r_A)$	11/85	100/146	8.01e-8
r_A	0.11	0.40	
	ANO/AO	ANOD/AOD	
$r_N / (1 - r_N)$	11/62	100/131	6.67e-6
r_N	0.15	0.43	
r_A / r_N	0.73	0.9	
Coelomata			
	Non-conserved in N	Conserved in N	p-value in Fisher test
ratio	ADO/DO	ADON/DON	
$r_A / (1 - r_A)$	131/711	100/146	1.08e-15
r_A	0.15	0.4	
ratio	ADO/AO	ADON/AON	
$r_D / (1 - r_D)$	131/62	100/11	6.67e-6
r_D	0.67	0.9	
r_A / r_D	0.22	0.44	

Definition of Conserved and Variable introns

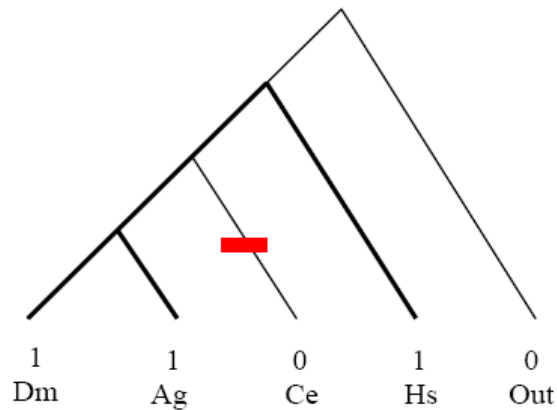
- For 5 taxa
 - A *conserved* intron is present in ≥ 3 taxa
 - A *variable* intron is present in 2 taxa

Examples of variable and conserved intron positions in 5 taxa

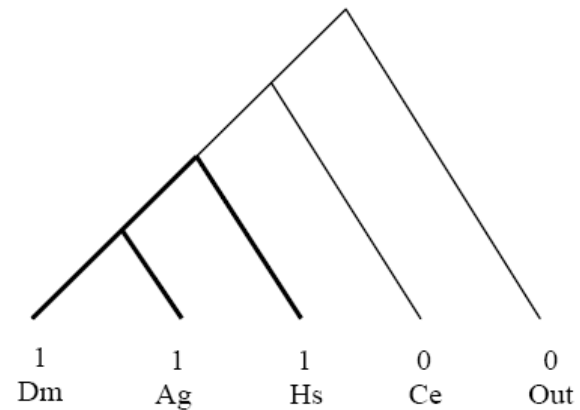
variable



conserved



Ecdysozoa



Coelomata

conserved introns are more reliable

- A intron is *informative* if it has different dollo parsimony scores for coelomata and ecdysozoa
- When informative,
 - Variable introns can have **double deletions** in both trees
 - Conserved introns can not have double deletion in either tree

Conserved introns support Coelomata

Dollo parsimony analysis: for one intron

1. $E = \text{min \# losses in Ecdysozoa}$
2. $C = \text{min \# losses in Coelomata}$
3. If $E = C$, intron is *uninformative* and discarded
4. If $E < C$, intron supports Ecdysozoa
5. If $E > C$, intron support Coelomata

Outgroup	# introns in ≥ 2 groups	# introns in exactly 2 groups	# introns in ≥ 3 groups
Non-animals	Ecdysozoa 737 / 990	Ecdysozoa 731 / 916	Coelomata 68 / 74

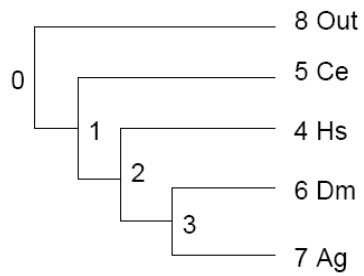
Supported tree

informative introns that support the tree

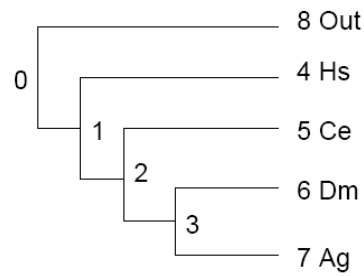
Total # informative introns

Test of retention rate

- Assumption
 - Retention rate higher for conserved introns
- Argument
 - Ecdysozoa violate above assumption with statistical significance



Coelomata



Ecdysozoa

Not significant

P < 0.02
One-sided Fisher

Coelomata Tree Topology						
branch	variable introns			strongly conserved introns		
	retained	lost	retention rate	retained	lost	retention rate
(0, 1)	841	0	1	241	0	1
(0, 8)	841	0	1	241	0	1
(1, 5)	277	756	0.268	188	91	0.674
(1, 2)	948	85	0.918	279	0	1
(2, 4)	983	65	0.938	320	27	0.922
(2, 3)	165	883	0.157	201	146	0.579
(3, 6)	175	60	0.745	149	52	0.741
(3, 7)	130	105	0.553	143	58	0.711
Ecdysozoa Tree Topology						
branch	variable introns			strongly conserved intron		
	retained	lost	retention rate	retained	lost	retention rate
(0, 1)	841	0	1	241	0	1
(0, 8)	841	0	1	241	0	1
(1, 4)	983	130	0.883	320	21	0.938
(1, 2)	402	711	0.361	341	0	1
(2, 5)	277	145	0.656	188	159	0.542
(2, 3)	165	257	0.390	146	146	0.579
(3, 6)	175	60	0.745	52	52	0.741
(3, 7)	130	106	0.553	58	58	0.711

Summary

- We used intron positions to resolve Coelomata-Ecdysozoa controversy
- Retention of introns is dependent between branches
- We divide introns into conserved and variable subsets by # taxa
- Conserved introns support Coelomata and variable support Ecdysozoa
- Our tests to support Coelomata

Discussion

- The Coelomata-Ecdysozoa dilemma requires
 - detailed analyses of more genomes
 - reconciliation of results from other characters
- Taxon sampling vs. character sampling
- Why are many introns conserved between Human and Arabidopsis?

Acknowledgements

Collaborators:

Igor B. Rogozin

Eugene V. Koonin

Teresa M. Przytycka

Helpful discussion with Yuri I. Wolf

Supported by Intramural Research program of
NCBI, NLM, NIH

Thank you

