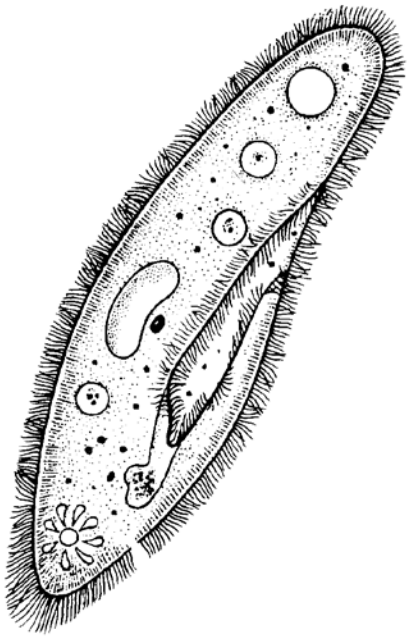
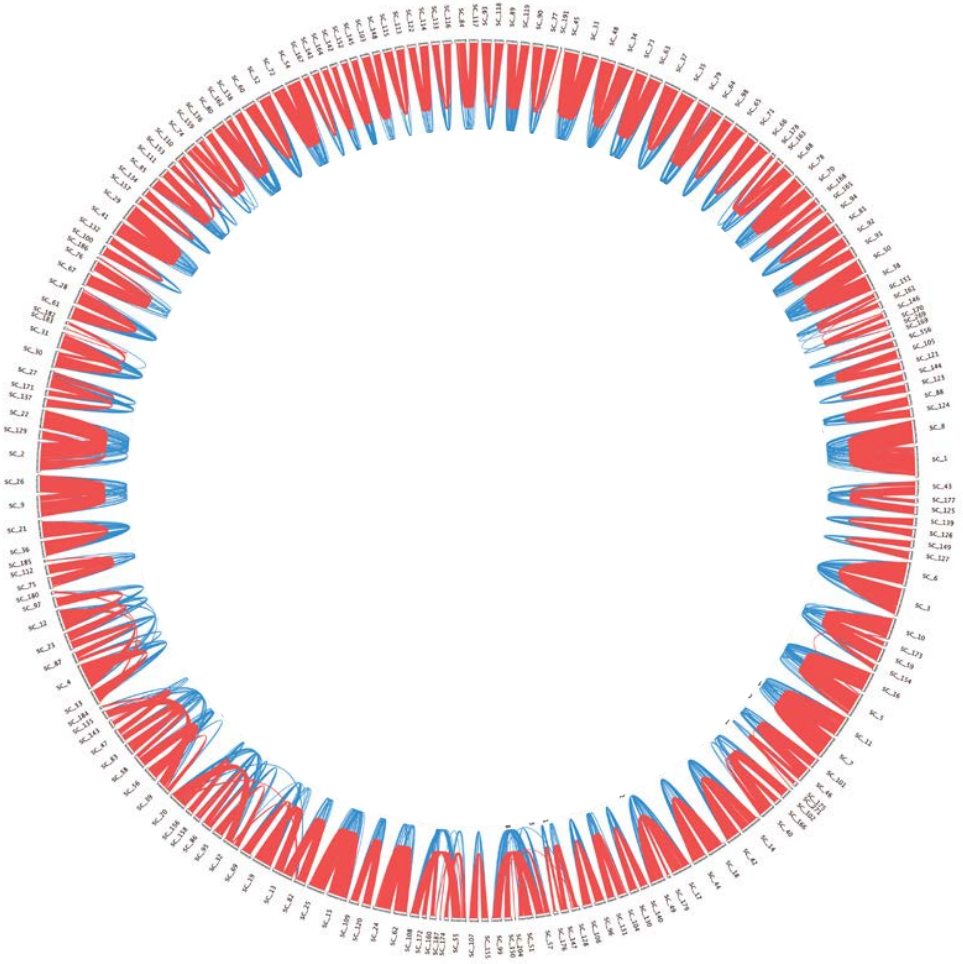
An Optimal System for Evolutionary Cell Biology: the genus *Paramecium*



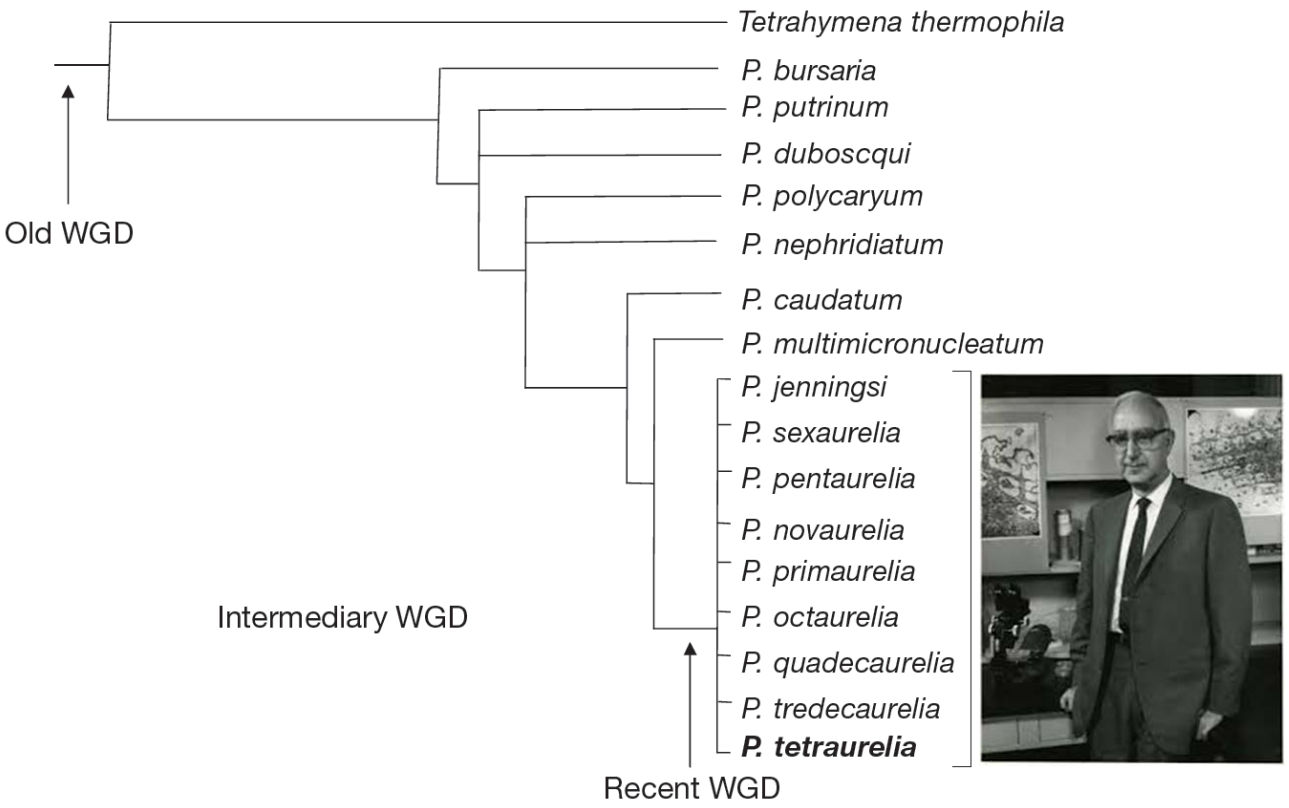
- Presence of a transcriptionally silent germline (micronucleus) and an expression-active somatic macronucleus.
- Geographically ubiquitous, large, free-living cells.
- Can be maintained as pure lines by autogamy.
- Direct effects of gene variants can be evaluated via macronuclear transformation.
- A product of two ancient whole-genome duplication events.
- Availability of pre-duplication outgroup species.
- Complete genome sequences for all species.
- One of the simplest genomic architectures in all eukaryotes.

Paramecium tetraurelia: a Descendant of Three Whole-Genome Duplications

- ~40,000 protein-coding genes in ~100 linkage groups, all grouping into sets of four syntenic groups (Aury et al., 2006, Nature).



Paramecium aurelia lineage: a cryptic species complex with at least 14 members.



Tracy Sonneborn
(1905 – 1981)

Figure 5 | Dating of genome duplication events. The phylogenetic tree of the *Paramecium* genus was adapted from ref. 47. Phylogenetic analyses indicate that the old WGD occurred before the divergence of *Paramecium* and *Tetrahymena*, and the recent WGD before the divergence of *P. tetraurelia* and *P. octaurelia*. There are currently not enough data to date the intermediary WGD.

From Aury et al. (2006)

Key Unresolved Issues:

- When did the whole-genome duplications in the *Paramecium* lineage occur?
  - Are all *Paramecium* species “octoploids”?
  - Are just the *aurelia* species or just *P. tetraurelia* polyploid?
  - Are there phylogenetically independent genome-duplication events?

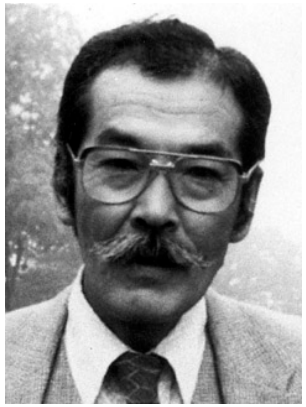
- How old are the *Paramecium* lineages and the underlying duplication events?

- What is the fundamental gene number in *Paramecium*?

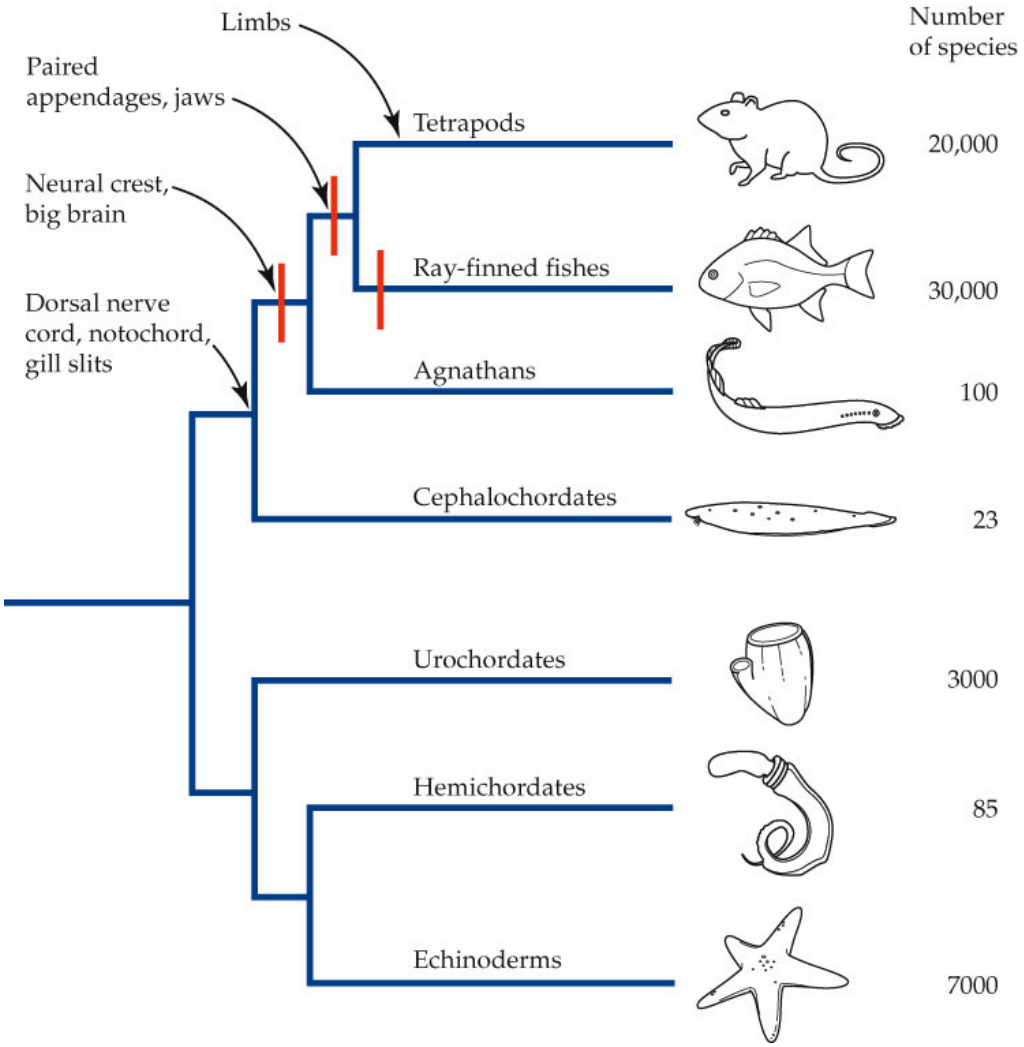
- What are the fates of the duplicate genes?

- Is whole-genome duplication a causal factor in the emergence of the cryptic *aurelia* species complex?

Ohno's 2R Hypothesis and the Origin of Vertebrate Innovations

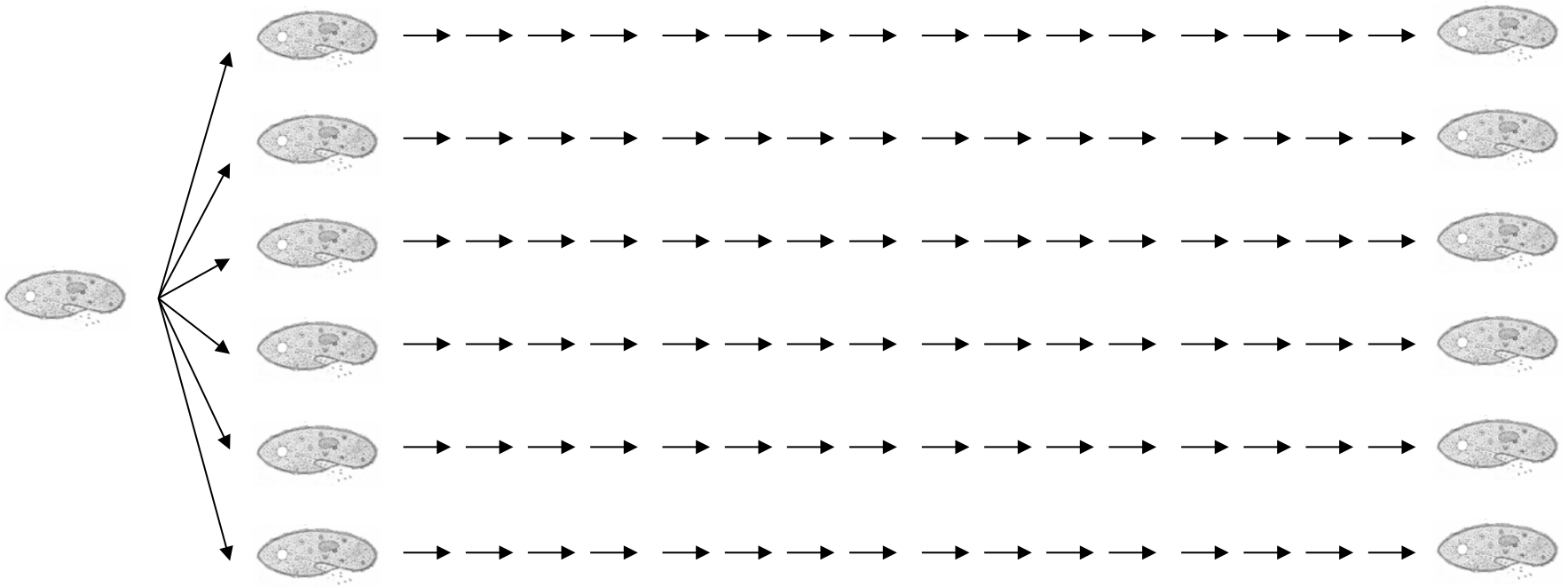


Susumu Ohno
(1928-2000)

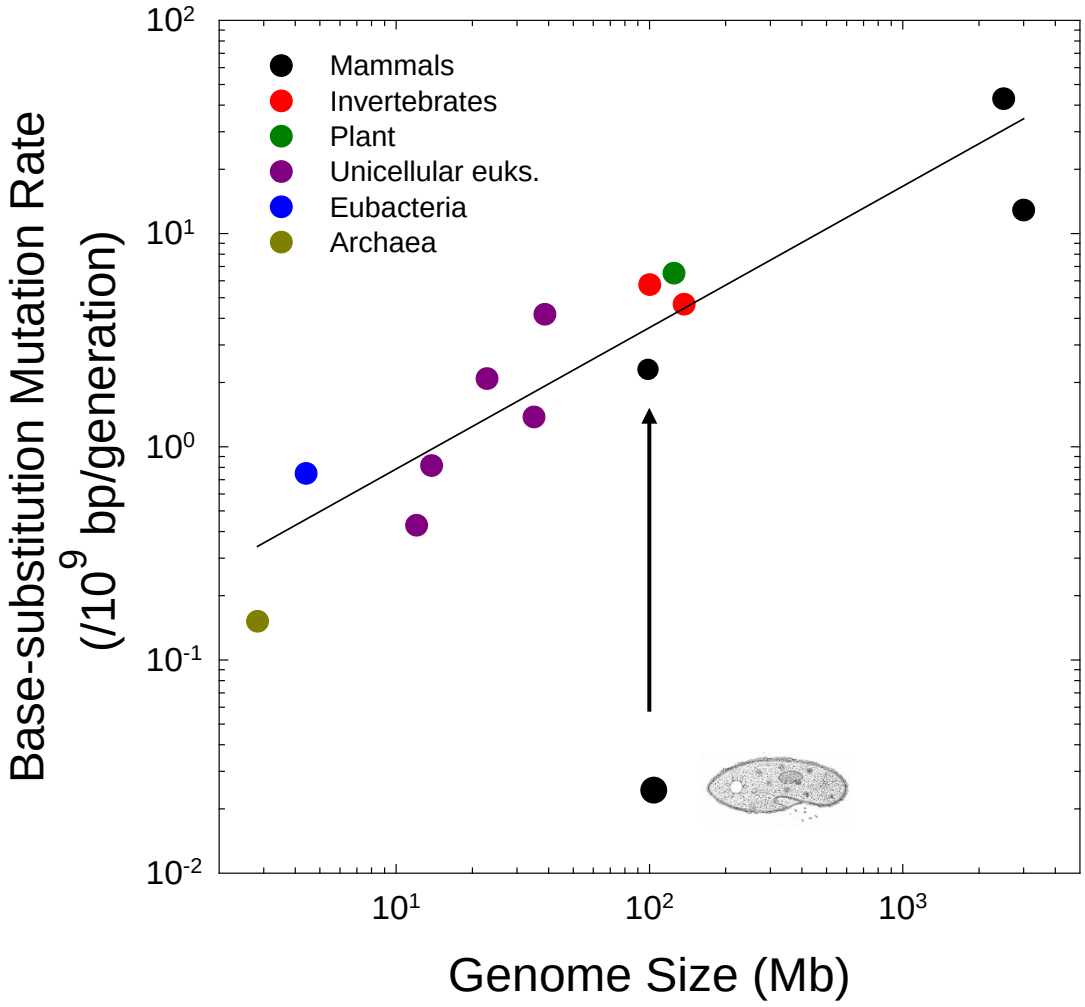


Analysis of Genome Stability with a Mutation-accumulation Experiment

- Starting with a single stem mother, sublines are maintained by single-progeny descent, preventing selection from removing spontaneous mutations.
- Mutations in the transcriptionally silent germline are invisible to selection.
- This protocol is continued for thousands of cell divisions.



Although *P. tetraurelia* Has the Lowest Known Mutation Rate Per Cell Division, Its Rate Per Sexual Episode is Compatible With Other Species



Further work underway:

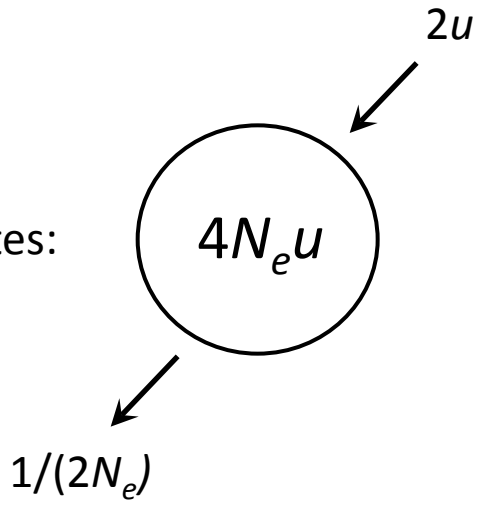
Paramecium biaurelia
Paramecium sexaurelia
Tetrahymena thermophila

Sung et al. (2012, PNAS)

Nucleotide Diversity at Silent Sites in Protein-coding Genes Estimates $4N_eu$

u = base-substitution mutation rate
N_e = genetic effective population size

Expected nucleotide heterozygosity at neutral sites:



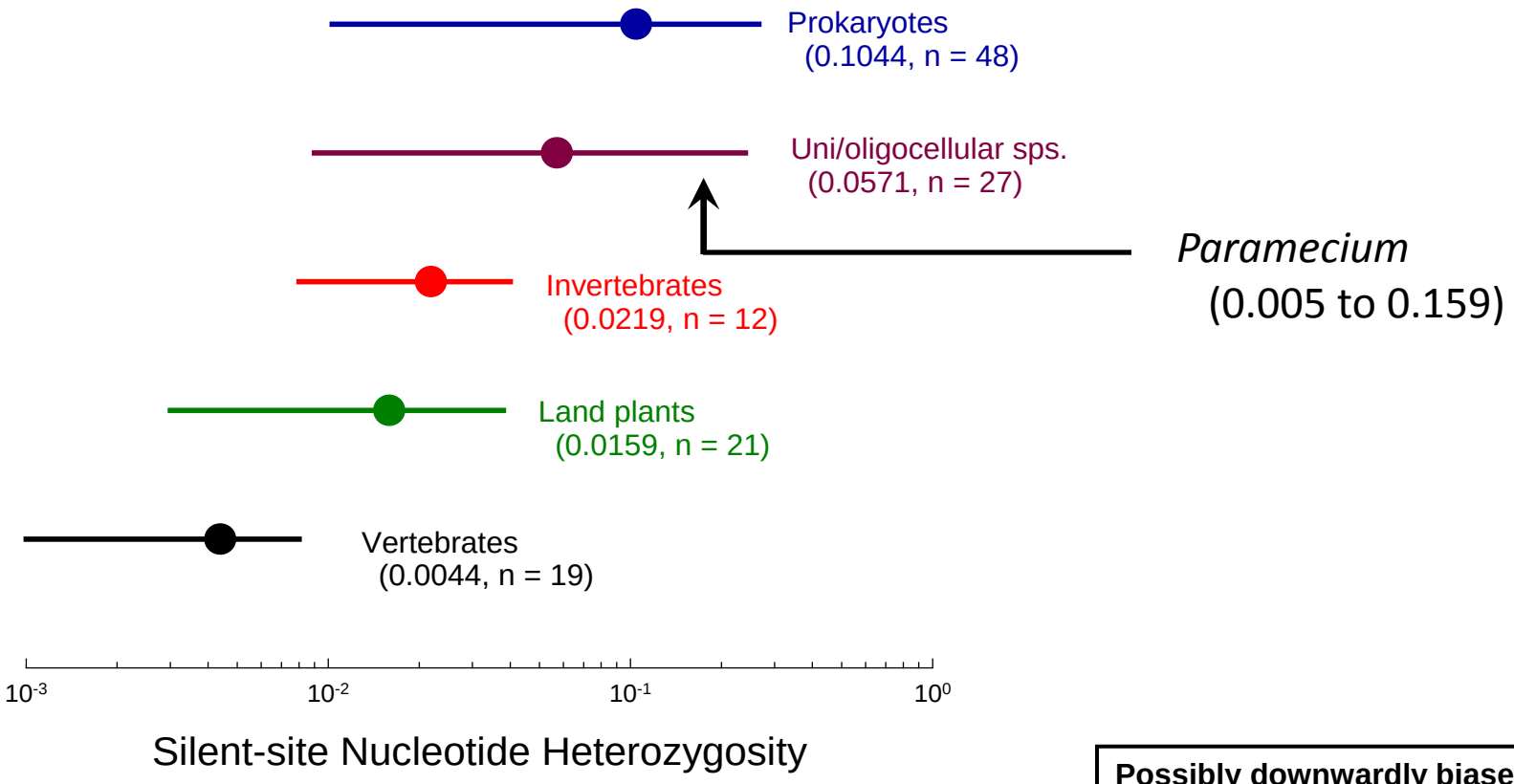
ACA
ACC
ACG
ACT

← Silent site

The Four Threonine Codons

$4N_e u$ for *Paramecium* Nuclear Genes is Exceptionally High

- Implies $N_e \approx 10^8$, the highest estimated value for a eukaryote.



Possibly downwardly biased:

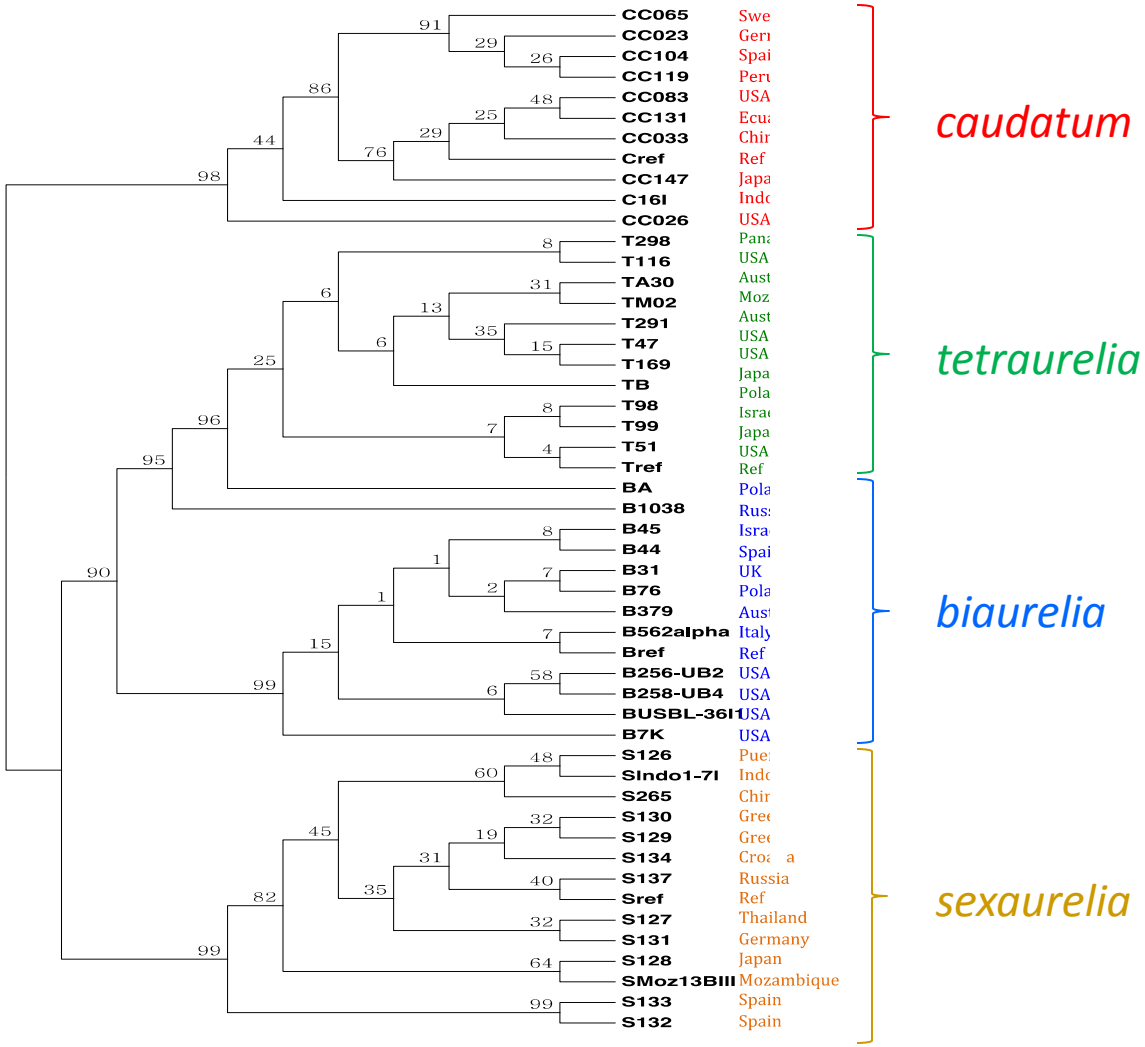
Probably some selection on silent sites.

Population-genomic Sequencing – Parul Johri, Tom Doak, Sascha Krenek

Species	Locations
P. tetraurelia (11)	Australia (Littlehampton, Melbourne), Israel (Ein Afek), Japan (Honshu Island, Moriko City), Mozambique (Mafambiasse), Panama (Empire Range), Poland (Kraków), Spain (Madrid), USA (Berkeley, Bloomington, Spencer)
P. biaurelia (12)	Australia (Tasmania Island), Israel (Dan), Italy (Milan), Poland (Kraków), Russia (Syktykar), Spain (Avila), UK (Rieff), USA (Boston, Pinehurst, Spencer)
P. sexaurelia (13)	China (Beijing), Croatia (Krka River), Germany (Stuttgart), Greece (Loannina Lake), Indonesia (Palu), Japan (Yamaguchi), Mozambique (Mafambiasse), Puerto Rico, Russia (Astrahan Nature Reserve), Spain (Seville), Thailand (Phuket)
P. caudatum (12)	China (Beijing), Ecuador (Loja), Germany (Plön), Indonesia (Palu), Japan (Kanzawa), Norway (Etnedal), Peru, Portugal (Elvas), Spain (Embalse de Contreras), Sweden (Avesta), USA (Baton Rouge, Lake Monroe)
P. multimicronucleatum (11)	Australia (Sydney), Brazil (Rio de Janeiro), Germany (River Nuthe), Greece (Dorjan Lake), Italy (Perugia, Pisa), Mozambique (Chicamba Real Dam), Portugal (Coimbrã), South Africa (Stellenbosch), USA (Bloomington), Russia (Tulun)

Thanks to: Thomas Berendonk, Eike Dusi, Sergei Fokin, Alexey Potekhin, Eva Pryzbos

Cohesive Species Phylogenies Determined from Whole-genome Sequencing



caudatum

tetraurelia

biaurelia

sexaurelia

Table 1 Genome statistics for *P. caudatum* as compared to *P. biaurelia*, *P. tetraurelia*, and *P. sexaurelia*

	P. caudatum	*P. biaurelia*	*P. tetraurelia*	*P. sexaurelia*
Genome size (Mb)	30.5	77.0	72.1	68.0
Genes	18,509	39,242	39,521	34,939
Gene length (exons + introns) (bp)	1,445.3	1,456.4	1,431.3	1,460.6
Exons/gene	3.5	3.6	3.3	3.6
Average exon length (bp)	399.0	377.9	418.8	379.3
Average intron length (bp)	24.7	31.4	24.2	30.3
Intergenic length (bp)	110.0	335.9	261.3	418.3
Genomic GC content (%)	28.2	25.8	28.0	24.1

Lengths given for genes, exons, introns, and intergenic regions are genome averages. Regions containing gaps were removed from the analysis before calculating averages.

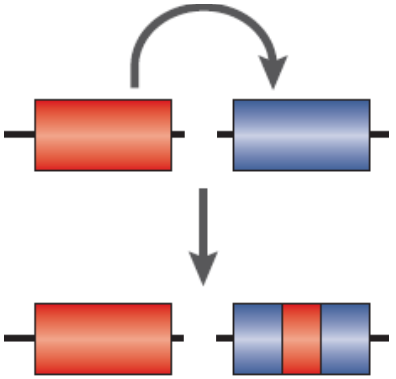
- The two most recent whole-genome (WGD) duplication events preceded the emergence of the *Aurelia* species.

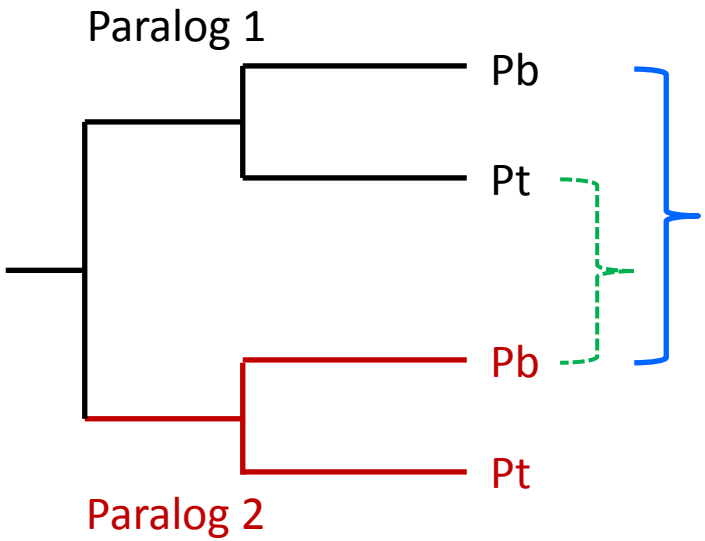
- These WGD events are unique to the *Aurelia* lineage and not shared by *P. caudatum* or *P. multimicronucleatum*.

- Very little post-WGD duplication: only ~300 tandem duplicates in each species, and only 8% of these are lineage specific.

- *P. caudatum* has ~1050 genes not found in any *Aurelia* species, and the latter have ~350 genes not shared by *P. caudatum*.

Evidence of On-going Gene Conversion Between Paralogs: Implications for the Age of WGD Events





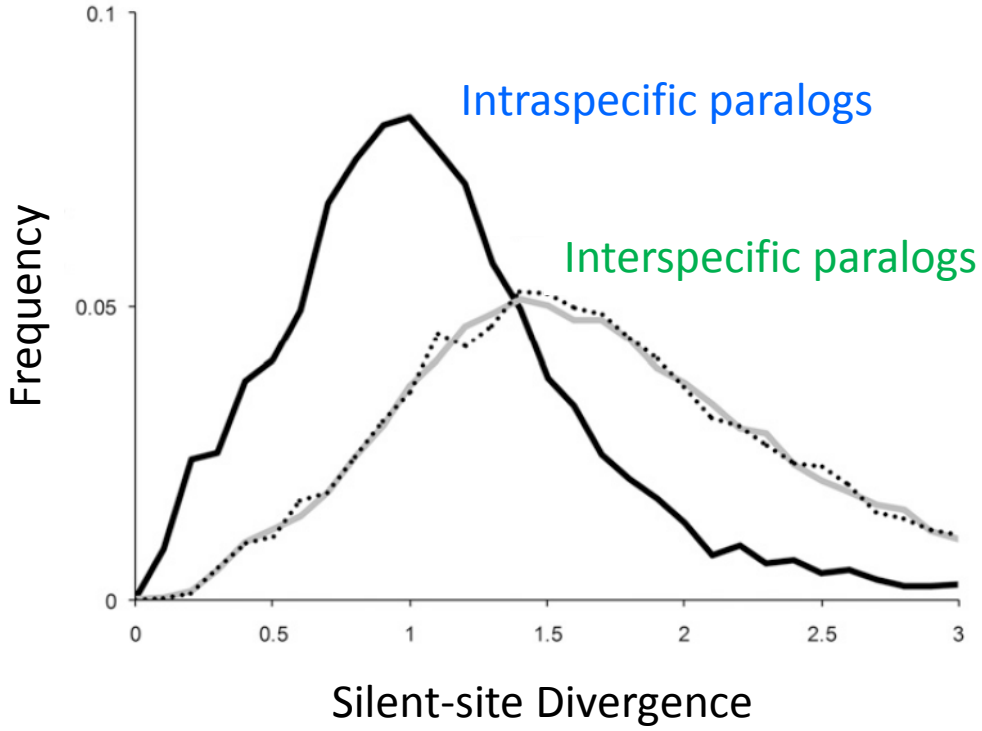
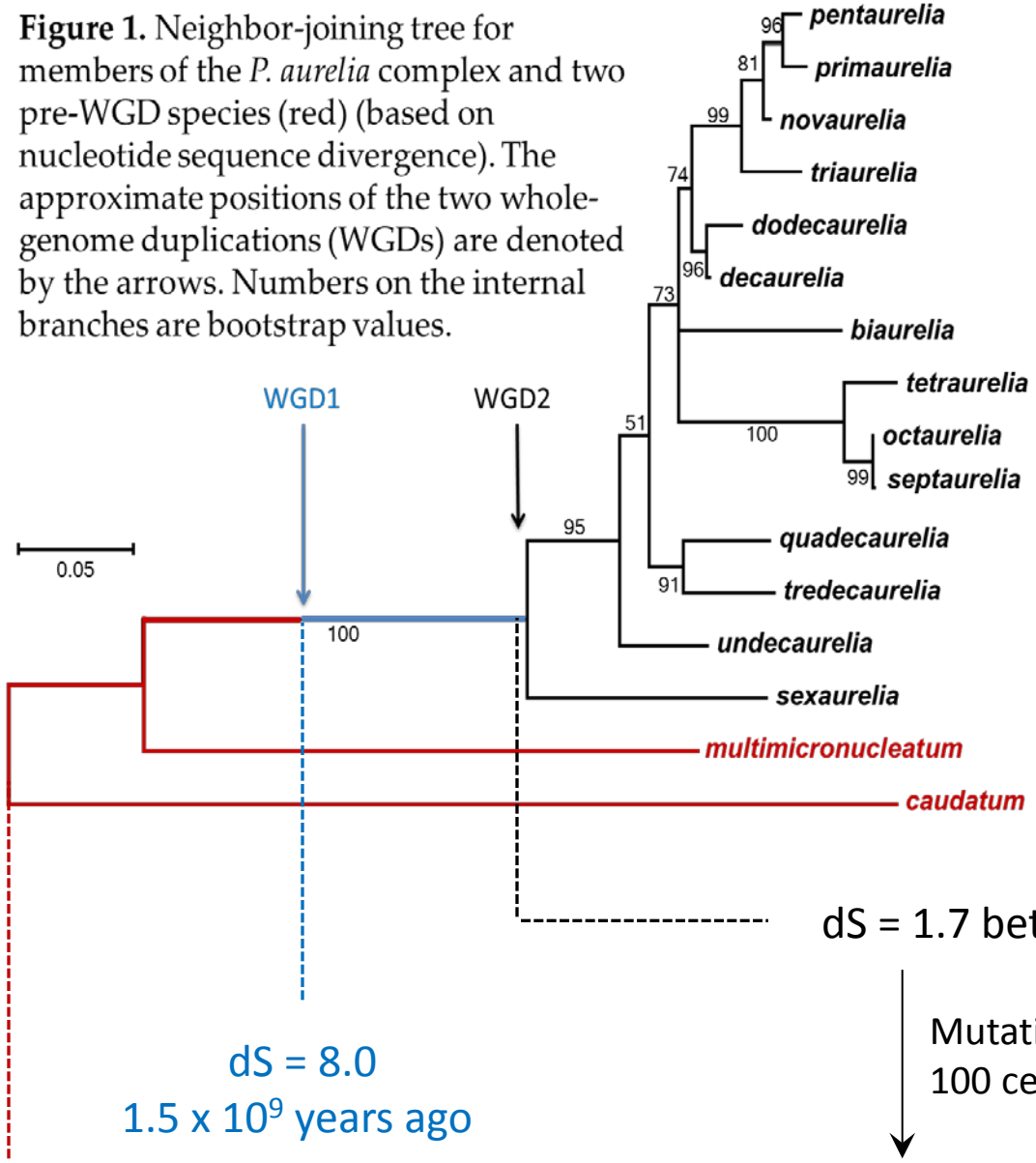


Figure 1. Neighbor-joining tree for members of the *P. aurelia* complex and two pre-WGD species (red) (based on nucleotide sequence divergence). The approximate positions of the two whole-genome duplications (WGDs) are denoted by the arrows. Numbers on the internal branches are bootstrap values.



Ages of the Two Most Recent WGD Events

With 1000 cell divisions per year, these estimates should be divided by 10.

dS = 1.7 between interspecific paralogs

Mutation rate = 2.6×10^{-11} / site / division
100 cell divisions / year

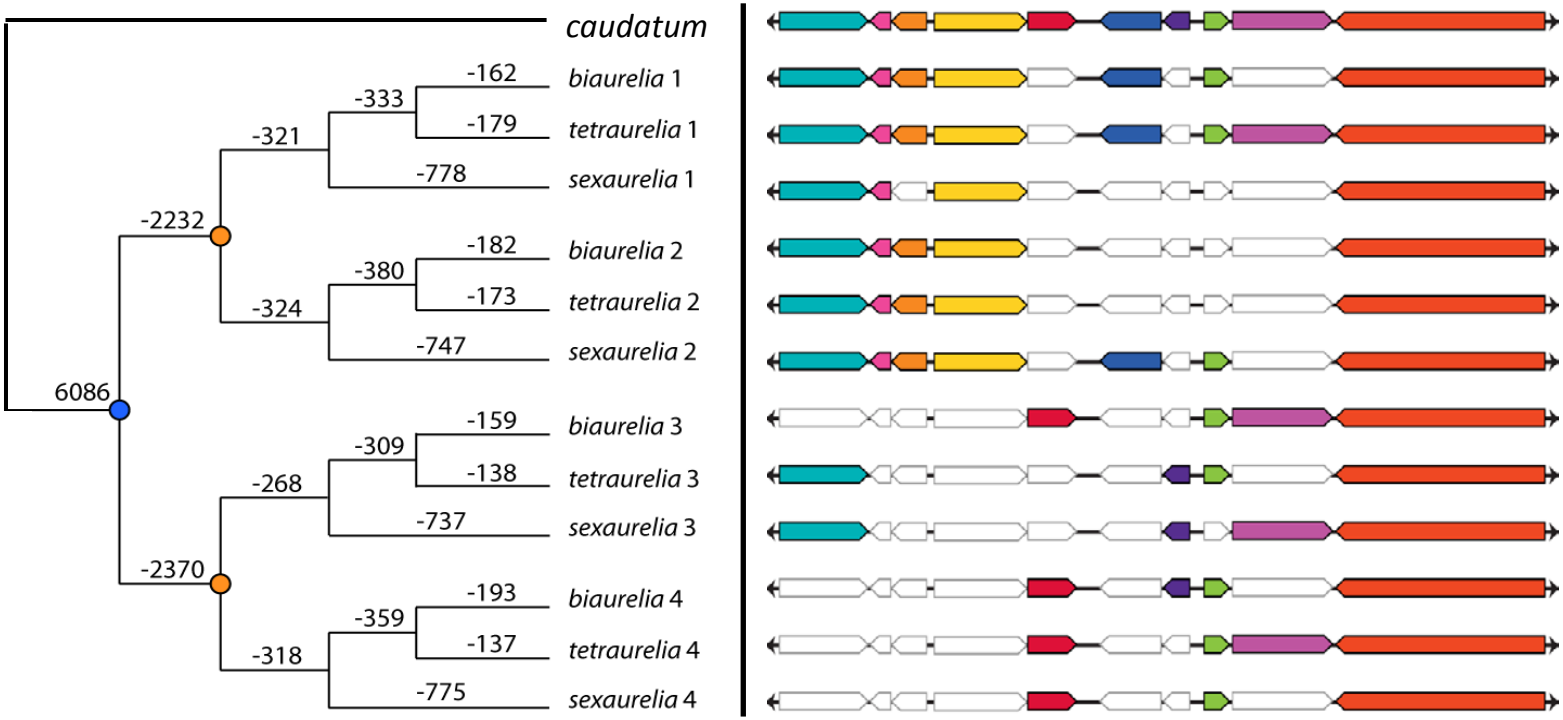
~32 billion cell divisions ≈ ~320×10^6 years ago

dS = 8.0
1.5×10^9 years ago

2.5×10^9 years?

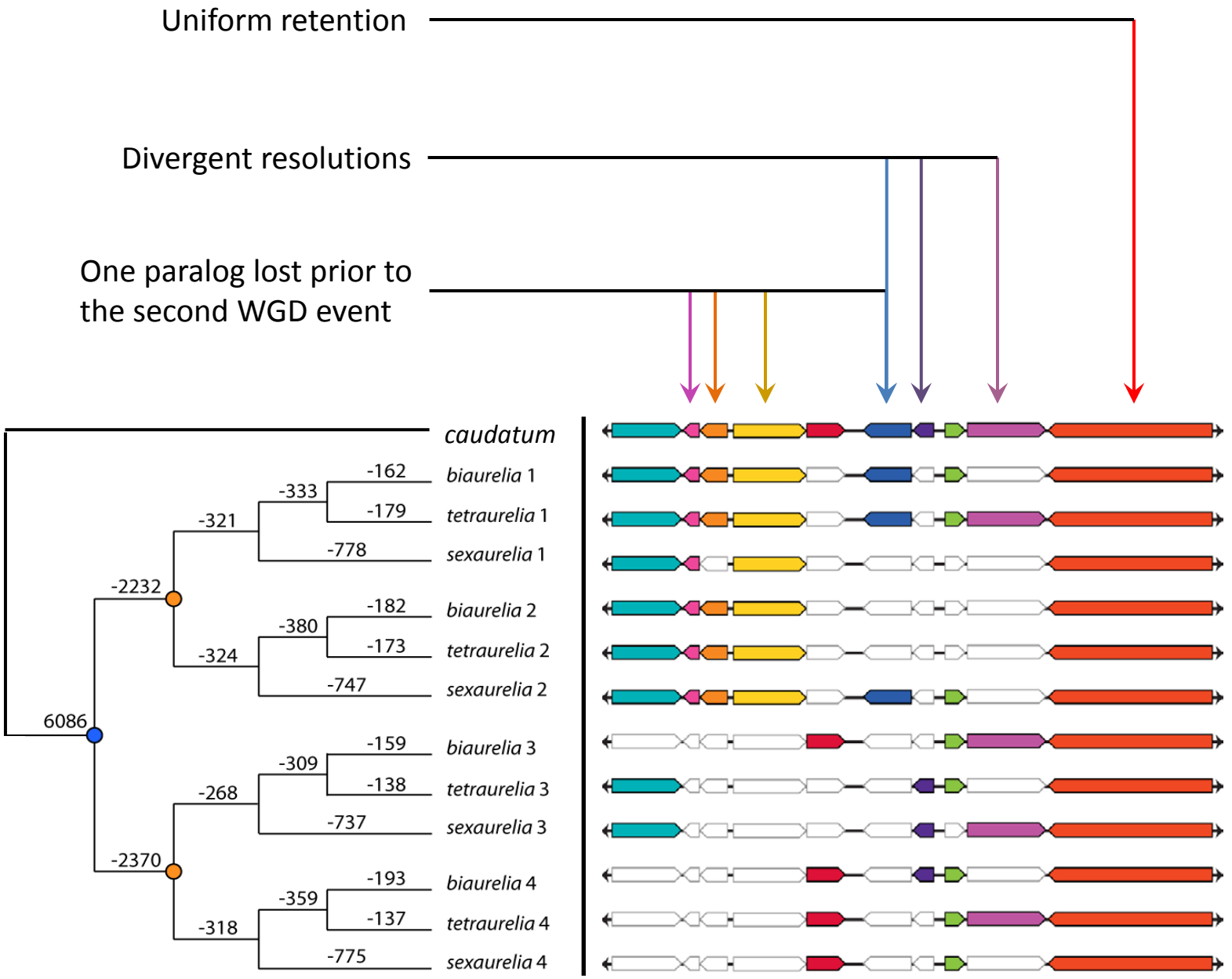
A Powerful Resource for Studying the Evolutionary Demography of Gene Duplicates

- Almost no post-WGD chromosomal rearrangements.
- Almost no post-WGD single-gene duplications.
- Almost no transposable elements.

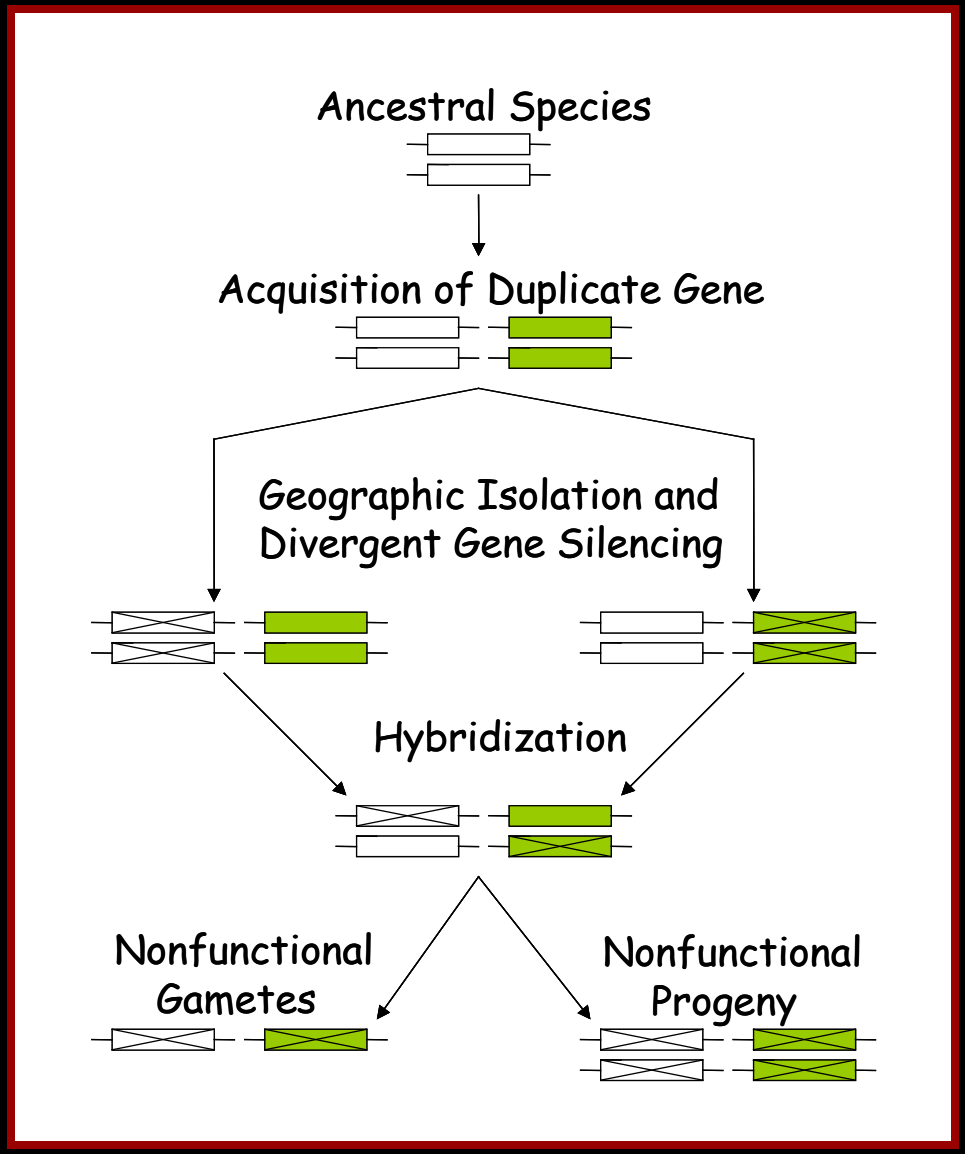


Patterns of gene loss for 6086 unambiguous co-orthologous genes present prior to the *aurelia*-specific whole-genome duplication events.

Inferring the Fates of Gene Duplicates

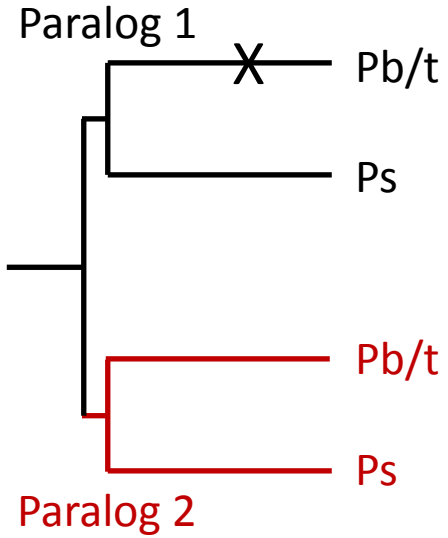


Stochastic, Divergent Losses of Duplicate Genes Leads to the Passive Origin of Reproductive Isolating Barriers by Inducing Microchromosomal Rearrangements

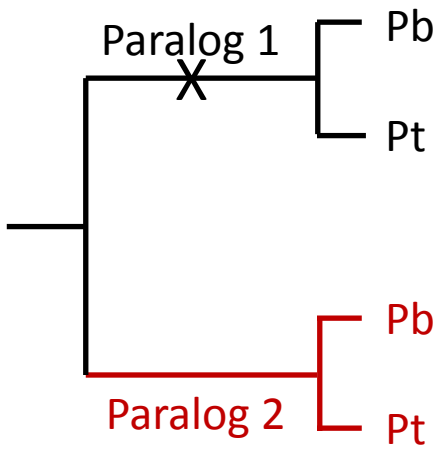


Dramatic Repatterning of Chromosomes by Divergent Resolution Drives Ongoing Emergence of Post-mating Isolating Barriers Among the *Aurelia* Lineages

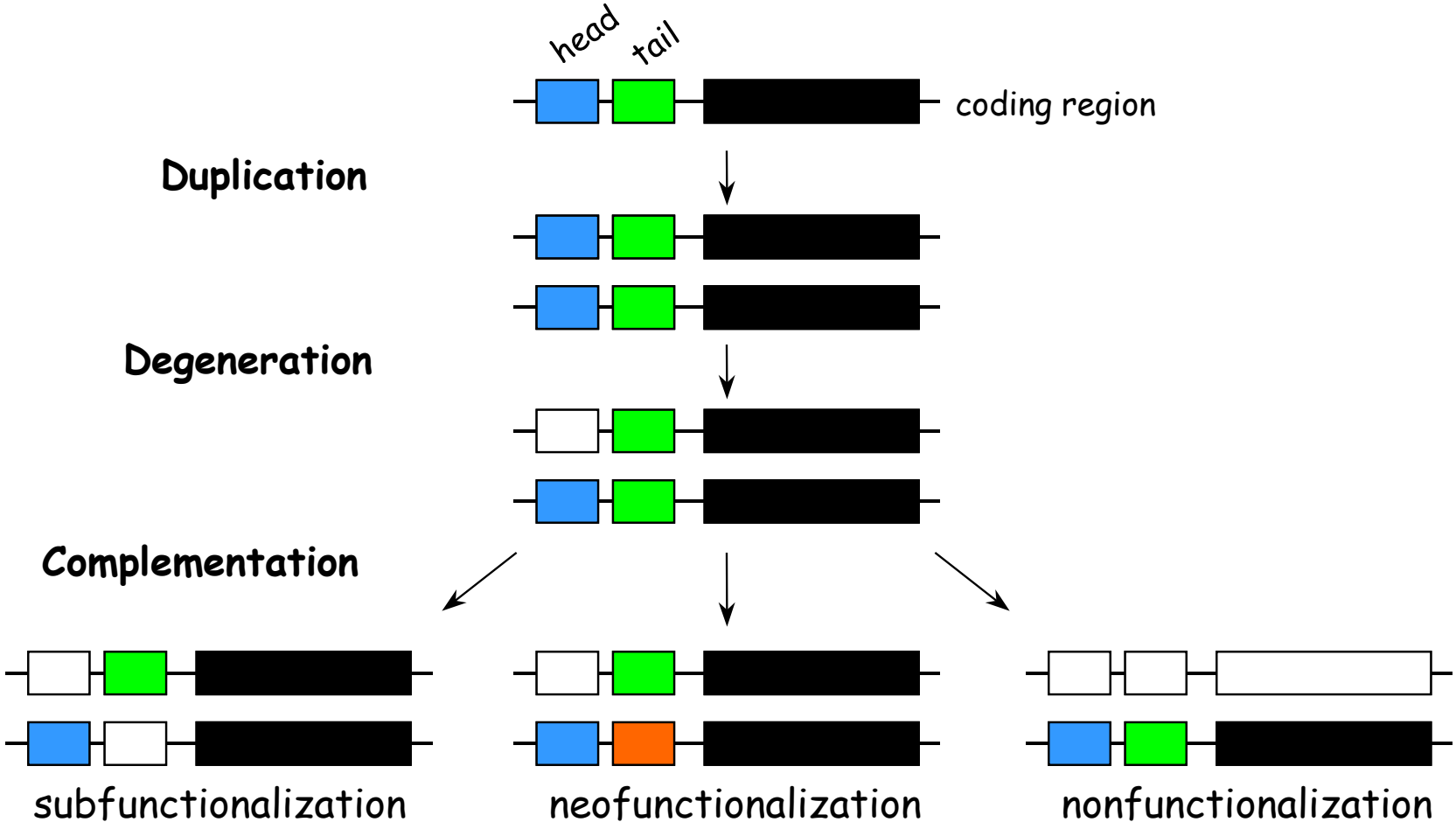
- 2312 observed cases of divergent resolution between *P. sexaurelia* and *P. biaurelia / tetraurelia*.
- 2741 cases of ancestral / parallel loss.
- This balance is expected if the split between *sexaurelia* and *biaurelia / tetraurelia* occurred shortly after the last whole-genome duplication event.



- 113 divergent resolutions between *P. biaurelia* and *P. tetraurelia*.
- 5421 cases of ancestral / parallel loss.

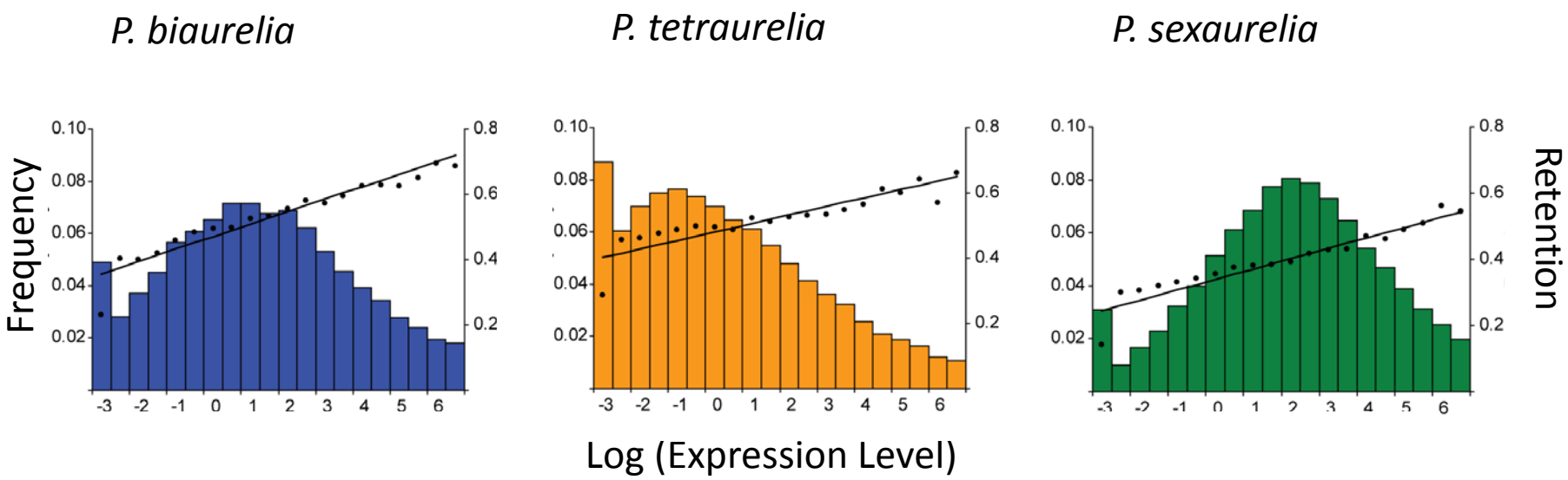


The DDC Model



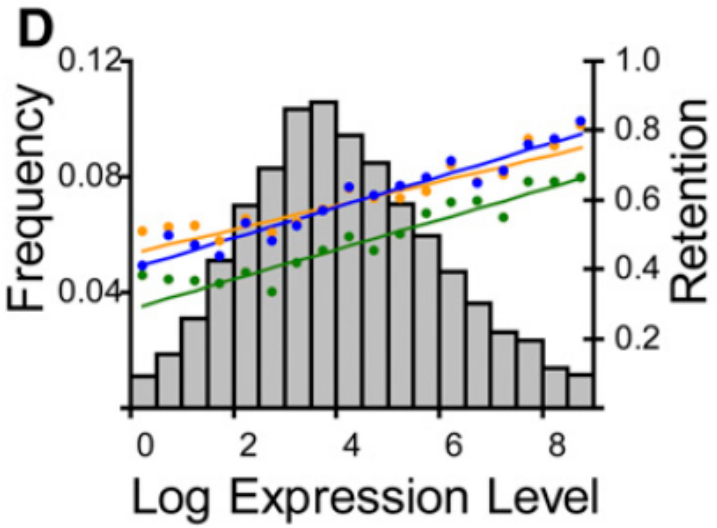
Force et al. (Genetics, 1999)

Expression Level is the Strongest Predictor of Duplicate-gene Retention Probability

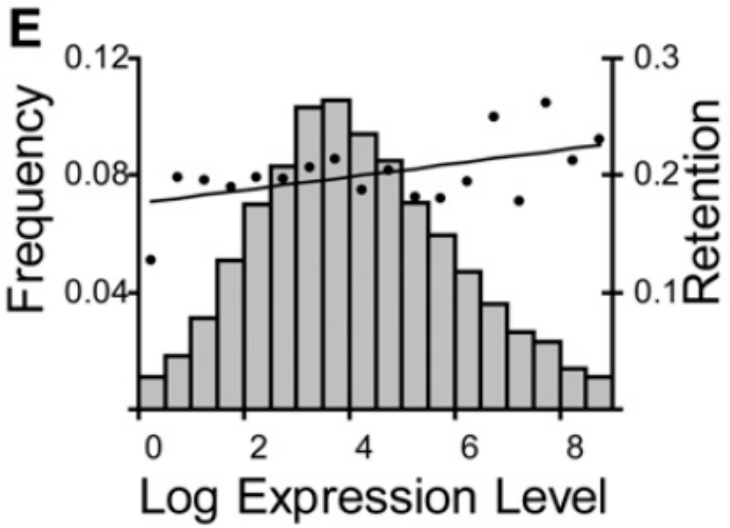


The Expression Level of *P. caudatum* Genes Predicts the Retention Probability of Duplicate Genes in the *Aurelia* Species

Recent WGD

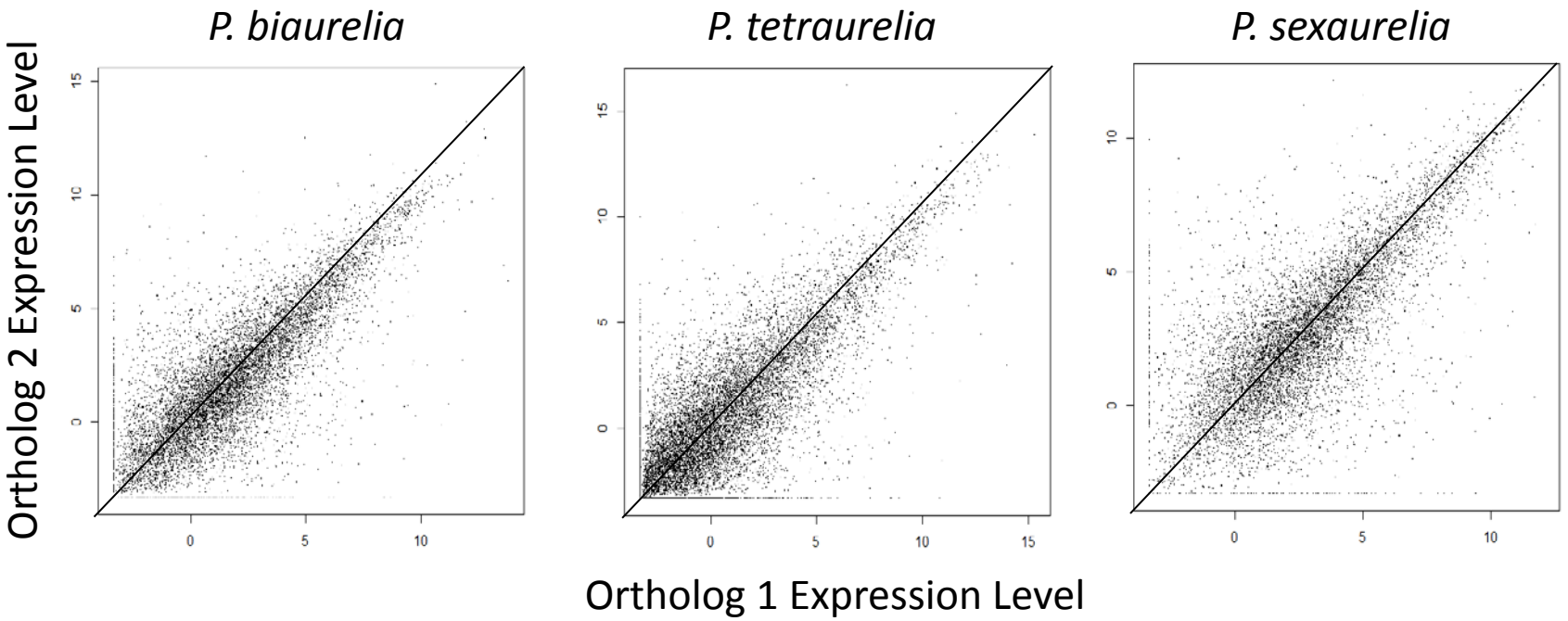


Intermediate WGD

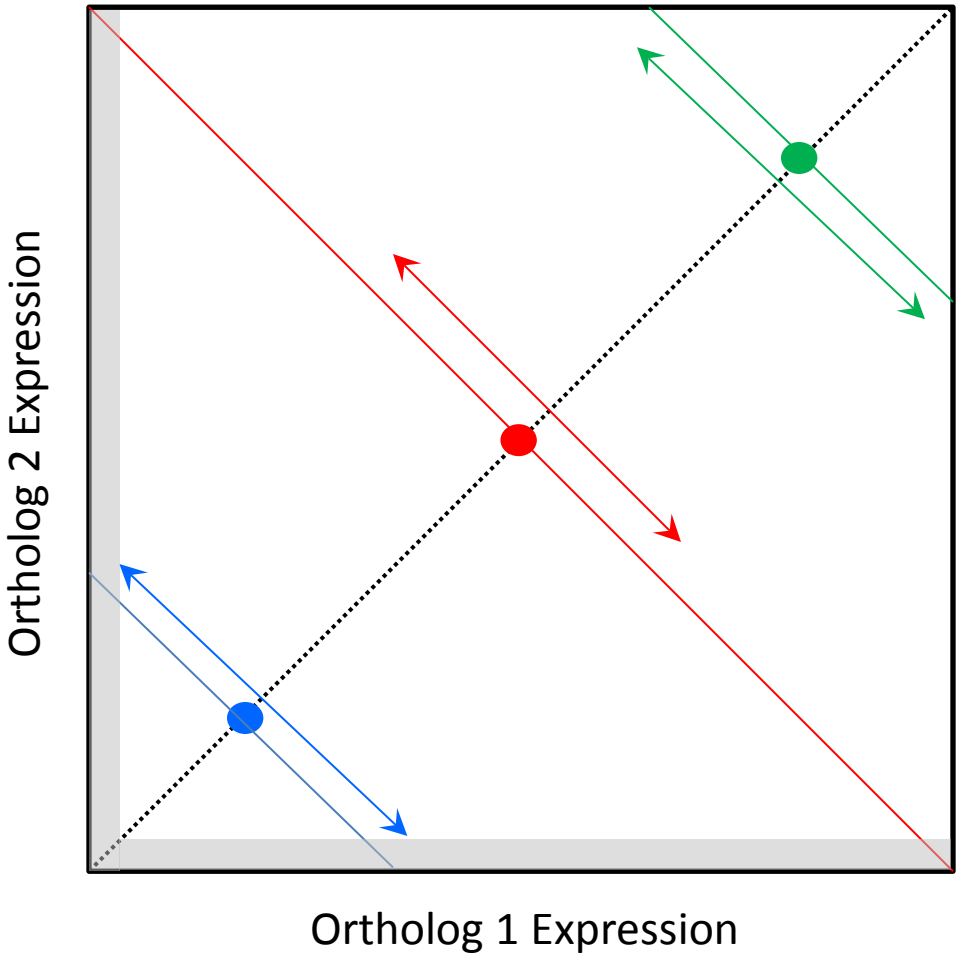


P. biaurelia ***P. sexaurelia*** ***P. tetraurelia***

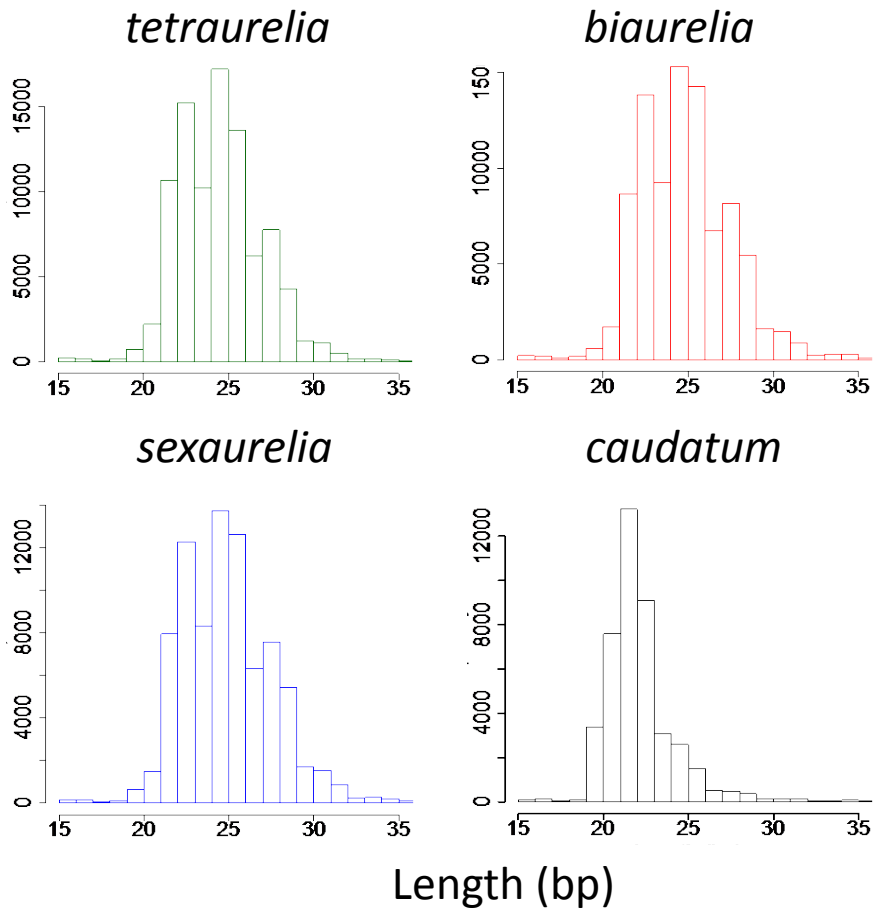
Despite saturation at silent sites, expression levels of orthologous proteins remain correlated.

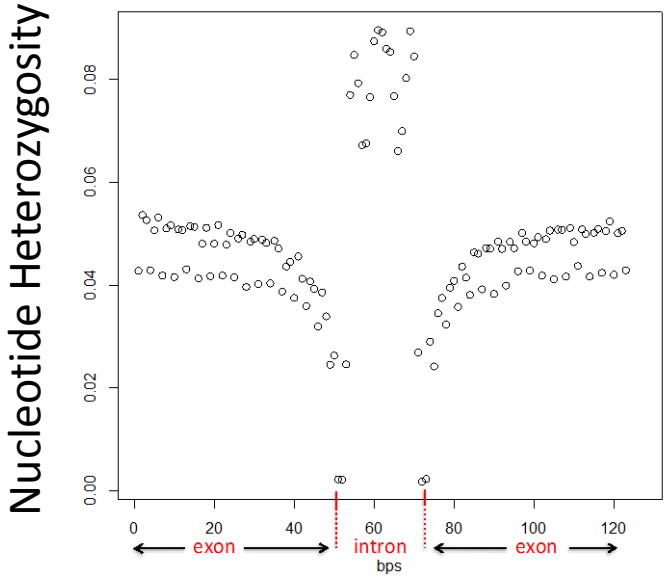


Quantitative Subfunctionalization: Joint Meandering of Paralog Expression Along the Drift Barrier



Extraordinary Constraints on Introns





Silent-site heterozygosity in protein-coding genes = 0.13

Table 1 Genome statistics for *P. caudatum* as compared to *P. biaurelia*, *P. tetraurelia*, and *P. sexaurelia*

	P. caudatum	*P. biaurelia*	*P. tetraurelia*	*P. sexaurelia*
Genome size (Mb)	30.5	77.0	72.1	68.0
Genes	18,509	39,242	39,521	34,939
Gene length (exons + introns) (bp)	1,445.3	1,456.4	1,431.3	1,460.6
Exons/gene	3.5	3.6	3.3	3.6
Average exon length (bp)	399.0	377.9	418.8	379.3
Average intron length (bp)	24.7	31.4	24.2	30.3
Intergenic length (bp)	110.0	335.9	261.3	418.3
Genomic GC content (%)	28.2	25.8	28.0	24.1

Lengths given for genes, exons, introns, and intergenic regions are genome averages. Regions containing gaps were removed from the analysis before calculating averages.

- ***Paramecium* sps. have the smallest known intron sizes in eukaryotes.**

- ***P. caudatum* has one of the shortest average intergenic distances known in eukaryotes.**

Strong Purifying Selection in Intergenic Regions

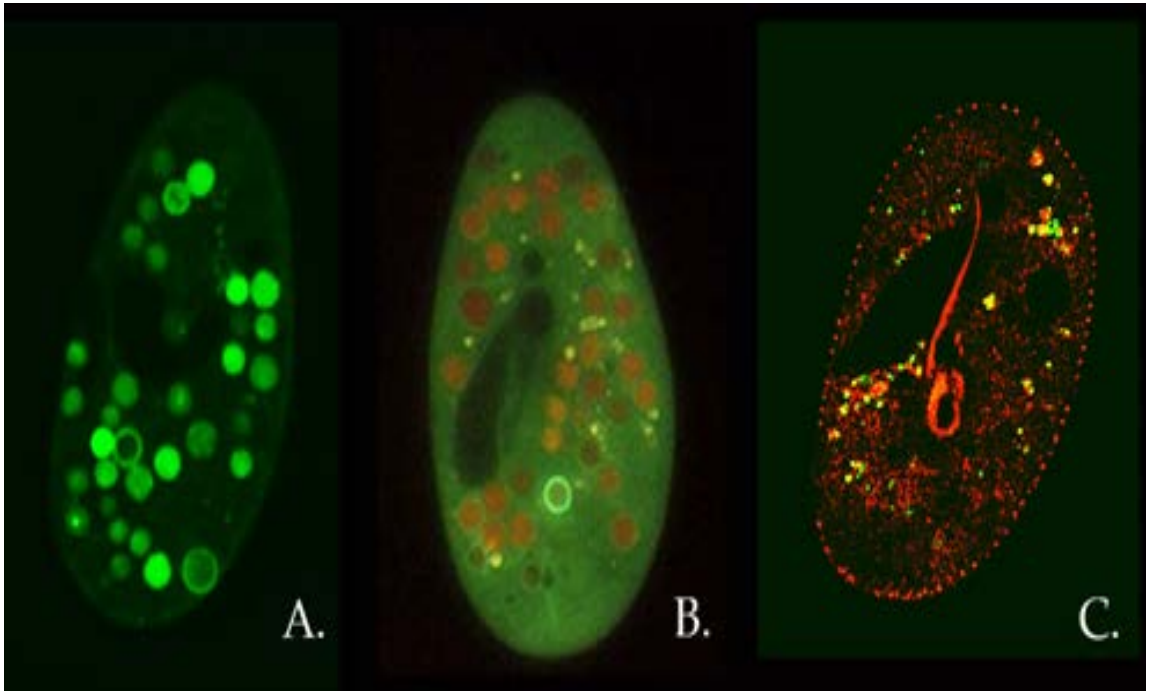
Nucleotide heterozygosities:	Silent sites	Intergenic regions
P. biaurelia	0.0079	0.0047
P. sexaurelia	0.0245	0.0155
P. tetraurelia	0.0051	0.0031
P. caudatum	0.1329	0.0311
P. multimicronucleatum	0.1589	0.0337

Regulatory Motifs Can be Detected Bioinformatically

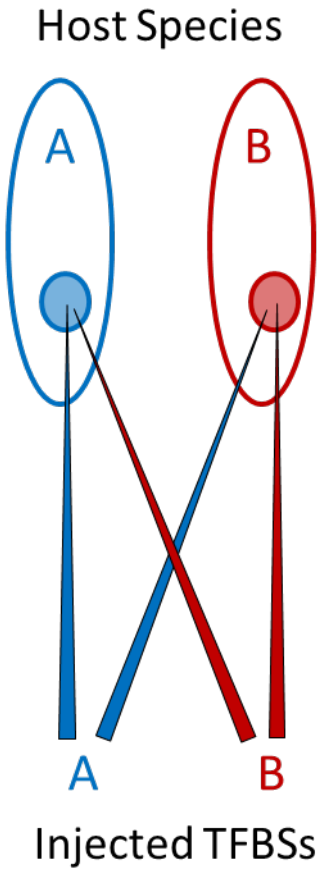
- Putative activators have positive effects that decrease with distance from the gene.
- Putative repressors have negative effects that increase with distance from the gene.

Motif	Associated Gene Number		Average Distance (bp)	Reg. Coef.	Significance
GTCGTTT	110	H	92	-0.34	2.49E-4
ATCGTTTTT	143	H	65	-0.28	8.58E-4
TGATTAAT	1601	H	92	-0.06	1.26E-2
TTCTAGAA	362	H	106	-0.24	3.81E-6
AAAGATTT	955	H	98	-0.32	2.20E-16
TGAGTAT	941	H	101	-0.07	2.16E-2
GAACTT	1498	L	115	0.07	4.07E-3
TAATTCTATTA	110	L	88	0.27	5.09E-3
TAACAG	1418	L	92	0.05	4.04E-2
C[TA]GTAGA	274	L	120	0.18	3.05E-3
ATTCTAC	624	L	123	0.06	1.49E-1
GATACA	1059	L	117	0.08	1.34E-2

Using Injected Mini-chromosomes to Study Gene Expression

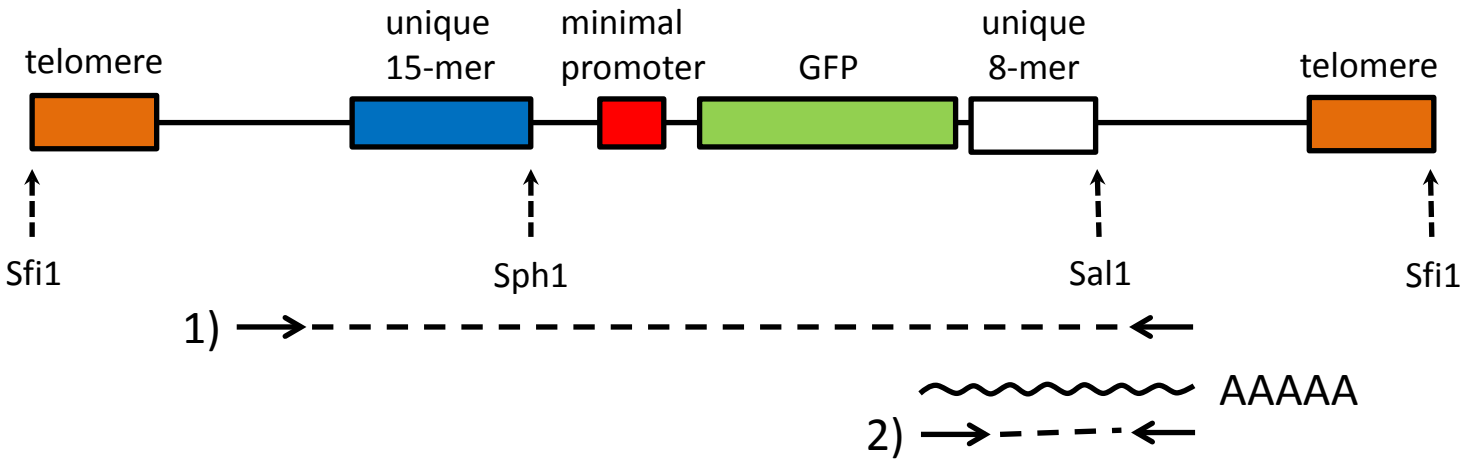


Subcellular localization of fluorescent protein fusions indicates that functional change has occurred between paralogous Rab genes in *P. tetraurelia* (Lydia Bright).



A Strategy for Identifying Transcription-factor Binding Sites

- All 4096 possible 6-mers are contained in 184 unique 15-mers.
- The full set of constructs can be co-injected into the macronucleus, and interrogated for activity with high-throughput sequencing of the bar-coded regions.



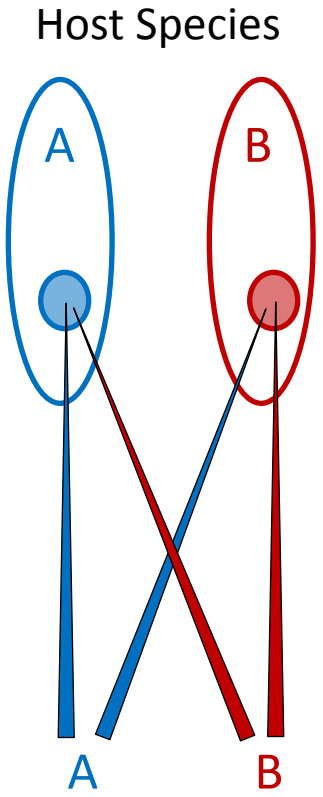
Host Cell	TFBS Source	Context	Linear model:
A	A	native use	$z_{AA} = \mu + h + f + x$
B	B	native use	$z_{BB} = \mu - h - f + x$
A	B	cross-species	$z_{AB} = \mu + h - f - x$
B	A	cross-species	$z_{BA} = \mu - h + f - x$

Full-parameter interpretation:

Grand mean:	$\mu = (z_{AA} + z_{BB} + z_{AB} + z_{BA})/4$
Host-cell (*trans*) effect:	$h = (z_{AA} + z_{AB} - z_{BB} - z_{BA})/4$
Binding-site (*cis*) effect:	$f = (z_{AA} + z_{BA} - z_{BB} - z_{AB})/4$
cis x *trans* interaction:	$x = (z_{AA} + z_{BB} - z_{AB} - z_{BA})/4$

Partial-parameter interpretation:

Net host-cell effect:	$(h + x) = (z_{AA} - z_{BA})/2$
Net binding-site effect:	$(f + x) = (z_{AA} - z_{AB})/2$



Casey McGrath

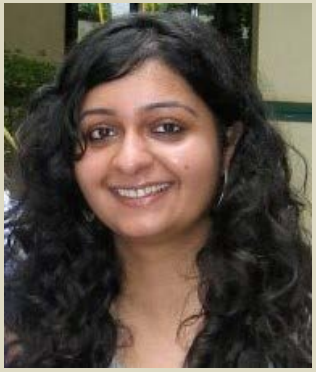
Tom Doak

Jean-Francois Gout

Lydia Bright

CC Chen

Hongan Long



Parul Johri

Georgi Marinov

Francesco Catania

John Preer

NATIONAL INSTITUTES OF HEALTH

NSF