Informational, Ecological and System Approaches for Complete Genome Analysis

François Serra[1]
Supervised by: Hernán Dopazo[1]

[1]Evolutionary Genomic Laboratory,
Centro de Investigación Príncipe Felipe

12th December 2012

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

VNIVERSITAT DE VALÈNCIA

Index

Index

Classification of evolutionary changes

- Evolutionary changes affect the inherited characteristics of a biological population
- **Changes are not directional**, evolution may accept or reject them through the elimination of individuals
- According to the influence of natural selection, we can classify **changes** in three categories:
  - **advantageous**: individuals carrying them may improve their fitness and thus their chances to produce offspring
  - **deleterious**: result in a loss of fitness, and may be removed from populations
  - **neutral**: they represents the narrow area that segregates advantageous and deleterious changes, they may not affect the fitness, and thus, may escape natural selection

Classification of evolutionary changes

- Evolutionary changes affect the inherited characteristics of a biological population
- **Changes are not directional**, evolution may accept or reject them through the elimination of individuals
- According to the influence of natural selection, we can classify **changes** in three categories:
  - **advantageous**: individuals carrying them may improve their fitness and thus their chances to produce offspring
  - **deleterious**: result in a loss of fitness, and may be removed from populations
  - **neutral**: they represents the narrow area that segregates advantageous and deleterious changes, they may not affect the fitness, and thus, may escape natural selection

Classification of evolutionary changes

- Evolutionary changes affect the inherited characteristics of a biological population
- **Changes are not directional**, evolution may accept or reject them through the elimination of individuals
- According to the influence of natural selection, we can classify **changes** in three categories:
  - **advantageous**: individuals carrying them may improve their fitness and thus their chances to produce offspring
  - **deleterious**: result in a loss of fitness, and may be removed from populations
  - **neutral**: they represents the narrow area that segregates advantageous and deleterious changes, they may not affect the fitness, and thus, may escape natural selection

Classification of evolutionary changes

- Evolutionary changes affect the inherited characteristics of a biological population
- **Changes are not directional**, evolution may accept or reject them through the elimination of individuals
- According to the influence of natural selection, we can classify **changes** in three categories:
  - **advantageous**: individuals carrying them may improve their fitness and thus their chances to produce offspring
  - **deleterious**: result in a loss of fitness, and may be removed from populations
  - **neutral**: they represents the narrow area that segregates advantageous and deleterious changes, they may not affect the fitness, and thus, may escape natural selection

Classification of evolutionary changes

- Evolutionary changes affect the inherited characteristics of a biological population
- **Changes are not directional**, evolution may accept or reject them through the elimination of individuals
- According to the influence of natural selection, we can classify **changes** in three categories:
  - **advantageous**: individuals carrying them may improve their fitness and thus their chances to produce offspring
  - **deleterious**: result in a loss of fitness, and may be removed from populations
  - **neutral**: they represents the narrow area that segregates advantageous and deleterious changes, they may not affect the fitness, and thus, may escape natural selection

Classification of evolutionary changes

- Evolutionary changes affect the inherited characteristics of a biological population
- **Changes are not directional**, evolution may accept or reject them through the elimination of individuals
- According to the influence of natural selection, we can classify **changes** in three categories:
  - **advantageous**: individuals carrying them may improve their fitness and thus their chances to produce offspring
  - **deleterious**: result in a loss of fitness, and may be removed from populations
  - **neutral**: they represents the narrow area that segregates advantageous and deleterious changes, they may not affect the fitness, and thus, may escape natural selection

Importance of neutral changes in molecular evolution

- Since their definition by Charles Darwin, neutral changes were though to be **minor actors** in the process of evolution.
- First molecular data and the estimation of mutation rate much higher than expected
  - Jack Lester King and Thomas H. Jukes, and independently Motoo Kimura to propose the *neutral theory of molecular evolution*.

Importance of neutral changes in molecular evolution

- Since their definition by Charles Darwin, neutral changes were though to be **minor actors** in the process of evolution.
- First molecular data and the estimation of mutation rate much higher than expected
  - Jack Lester King and Thomas H. Jukes, and independently Motoo Kimura to propose the *neutral theory of molecular evolution*.

Importance of neutral changes in molecular evolution

- Since their definition by Charles Darwin, neutral changes were though to be **minor actors** in the process of evolution.
- First molecular data and the estimation of mutation rate much higher than expected
  - Jack Lester King and Thomas H. Jukes, and independently Motoo Kimura to propose the *neutral theory of molecular evolution*.

Neutrality in Ecology

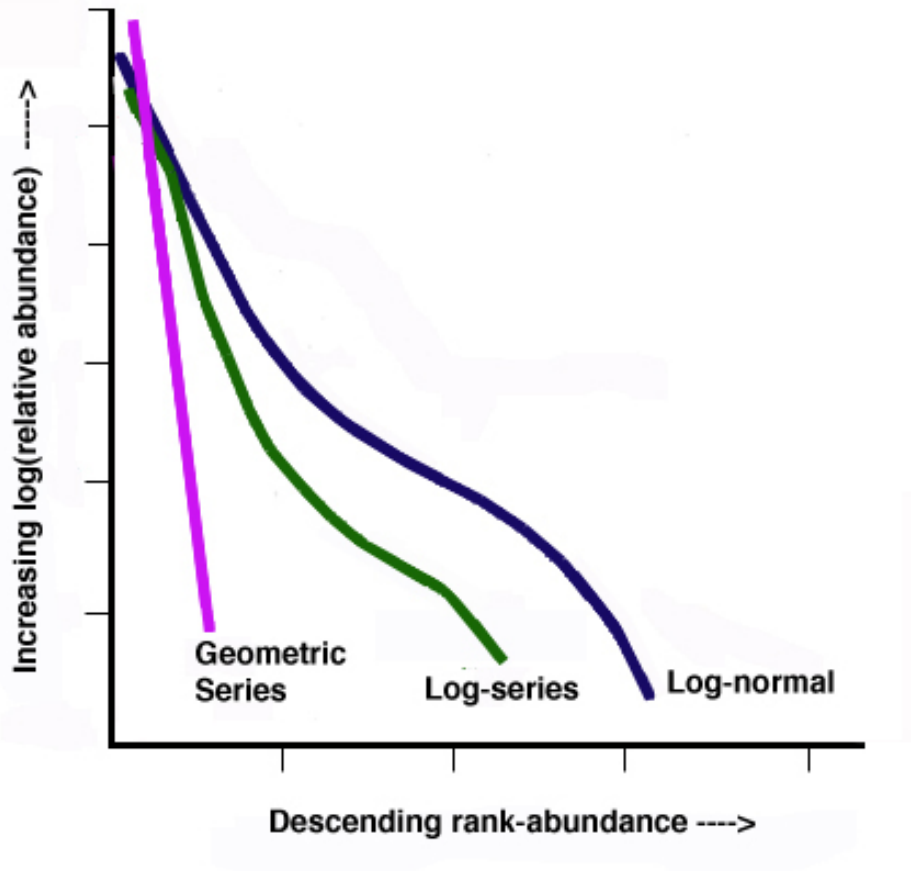
neutral distribution of species' abundances in ecosystems

- Models used were based on the concept of ecological niches. Differences observed between species were though to be mostly adaptive
- Statistical models able to reproduce the pattern of distribution of species' abundances in ecosystems (Motomura, Fisher, Preston)
- **neutral models** had to wait for the late 20^{th} century (Hubbell)

Neutrality in Ecology

neutral distribution of species' abundances in ecosystems

- Models used were based on the concept of ecological niches. Differences observed between species were though to be mostly adaptive
- Statistical models able to reproduce the pattern of distribution of species' abundances in ecosystems (Motomura, Fisher, Preston)
- **neutral models** had to wait for the late 20th century (Hubbell)

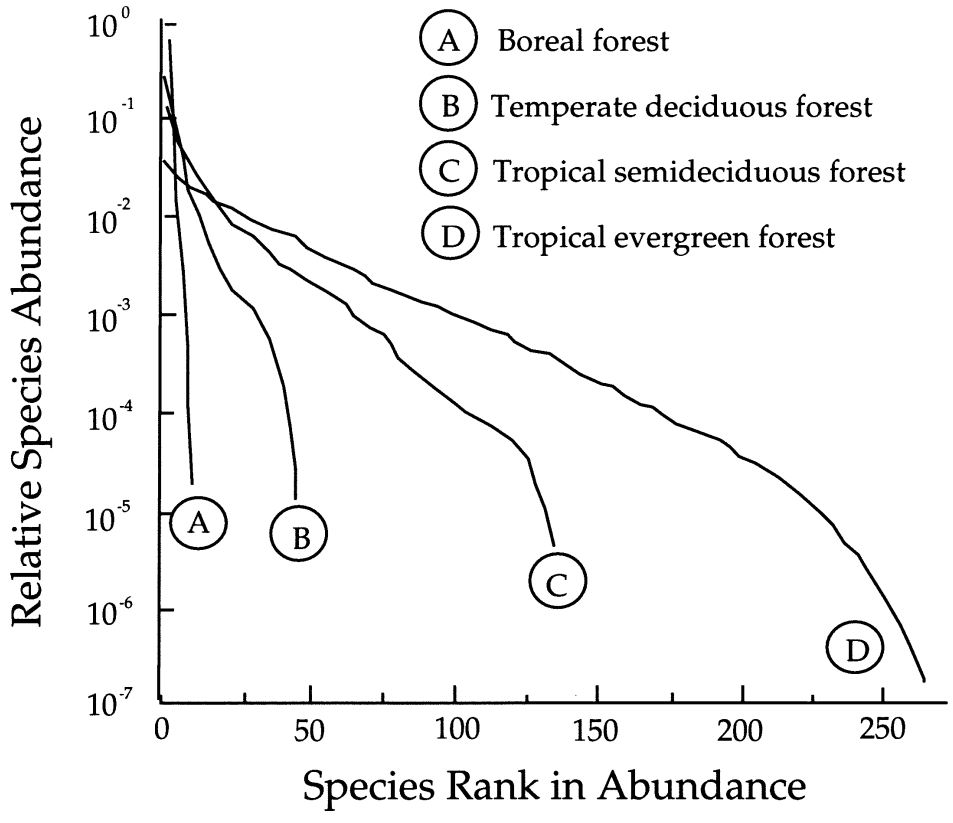


adapted from *[Magurran - (1988)]*

Neutrality in Ecology

neutral distribution of species' abundances in ecosystems

- Models used were based on the concept of ecological niches. Differences observed between species were though to be mostly adaptive
- Statistical models able to reproduce the pattern of distribution of species' abundances in ecosystems (Motomura, Fisher, Preston)
- **neutral models** had to wait for the late 20th century (Hubbell)

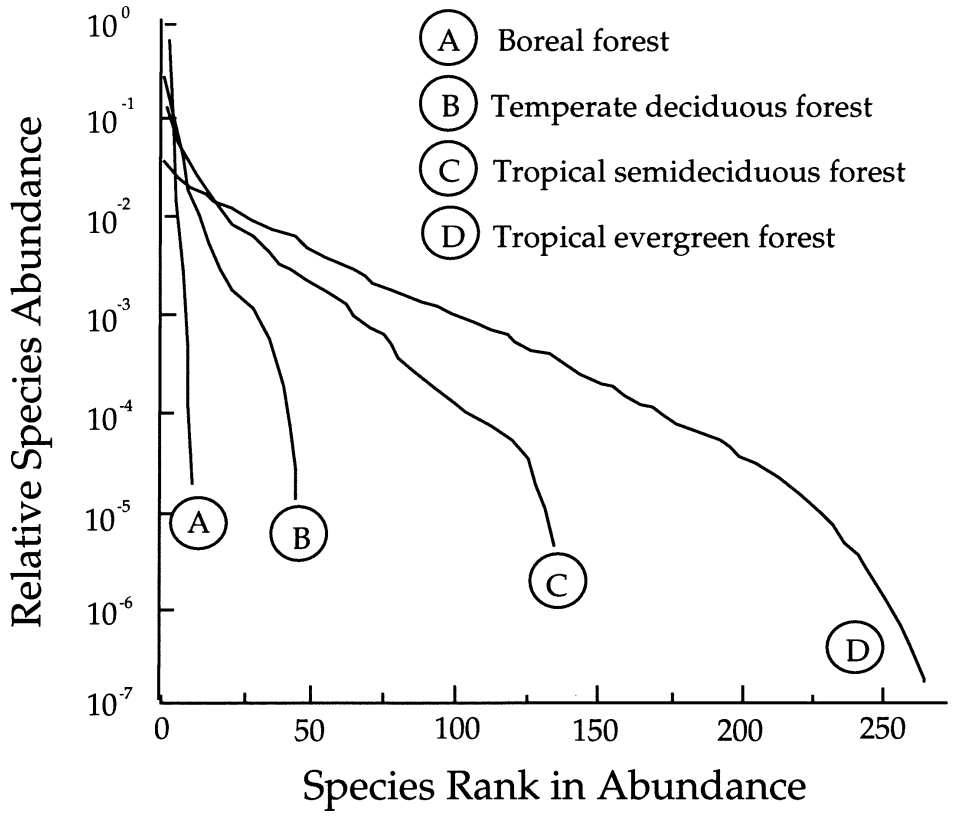


[Hubbell - (2001)]

Neutrality in Ecology

neutral distribution of species' abundances in ecosystems

- Models used were based on the concept of ecological niches. Differences observed between species were though to be mostly adaptive
- Statistical models able to reproduce the pattern of distribution of species' abundances in ecosystems (Motomura, Fisher, Preston)
- **neutral models** had to wait for the late 20th century (Hubbell)

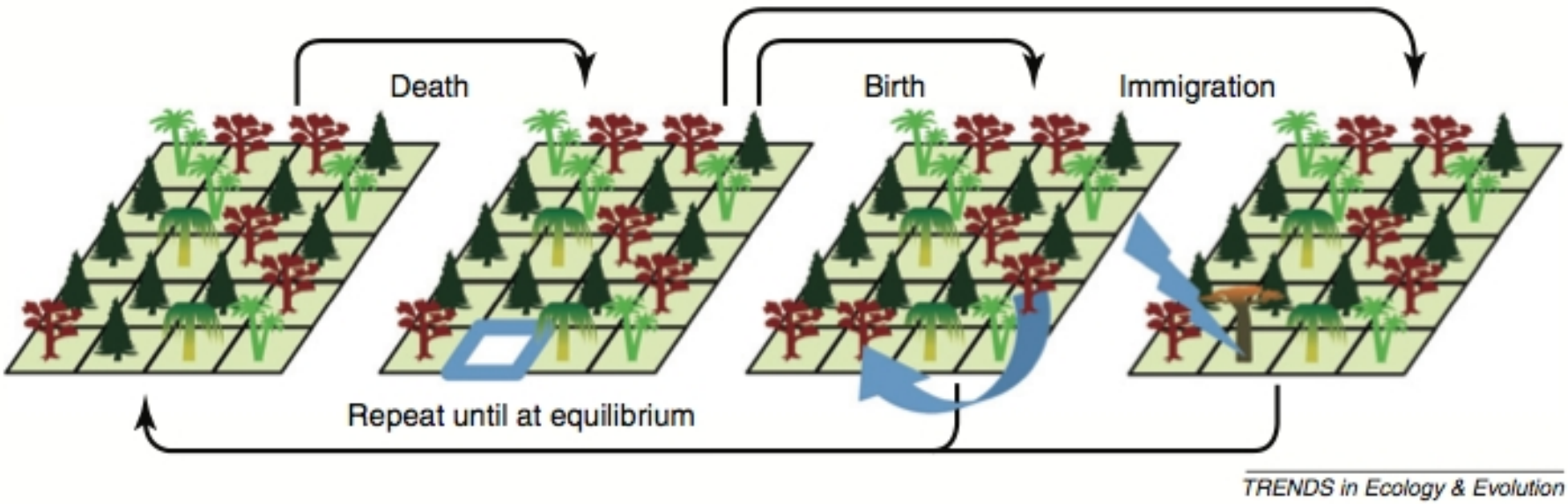


[Hubbell - (2001)]

Neutrality in Ecology

Unified Neutral Theory of Biodiversity and Biogeography (UNTB)

Assumes that diversity in a local community of individuals is maintained by migration from the metacommunity at a constant rate (m). Births and deaths in the local community occur at constant rates per generation, **regardless of the species**



[Rosindell et al. - *Trends in Ecology Evolution (2011)]*

Index

Objectives – Overview

The work presented here consists in three different approximations in the analysis of genomes, each expressing a neutral model and its corresponding deviations.

- **Informational**: where genomes are viewed as simple sequences with a given informational content. We will define a common structure of genomes and show how they are characterized by a universal quasi-random structure.
- **Ecological**: here a genetic elements populating genomes are assimilated to biological species in ecosystems. We will apply ecological models on genomes in order to expose the neutral pattern behind their composition.
- **System**: focusing on proteins working together to complete a function we present a new strategy to find natural selection fingerprint at this level of study.

Objectives – Overview

The work presented here consists in three different approximations in the analysis of genomes, each expressing a neutral model and its corresponding deviations.

- **Informational**: where genomes are viewed as simple sequences with a given informational content. We will define a common structure of genomes and show how they are characterized by a universal quasi-random structure.
- **Ecological**: here a genetic elements populating genomes are assimilated to biological species in ecosystems. We will apply ecological models on genomes in order to expose the neutral pattern behind their composition.
- **System**: focusing on proteins working together to complete a function we present a new strategy to find natural selection fingerprint at this level of study.

Objectives – Overview

The work presented here consists in three different approximations in the analysis of genomes, each expressing a neutral model and its corresponding deviations.

- **Informational**: where genomes are viewed as simple sequences with a given informational content. We will define a common structure of genomes and show how they are characterized by a universal quasi-random structure.
- **Ecological**: here a genetic elements populating genomes are assimilated to biological species in ecosystems. We will apply ecological models on genomes in order to expose the neutral pattern behind their composition.
- **System**: focusing on proteins working together to complete a function we present a new strategy to find natural selection fingerprint at this level of study.

Objectives – Overview

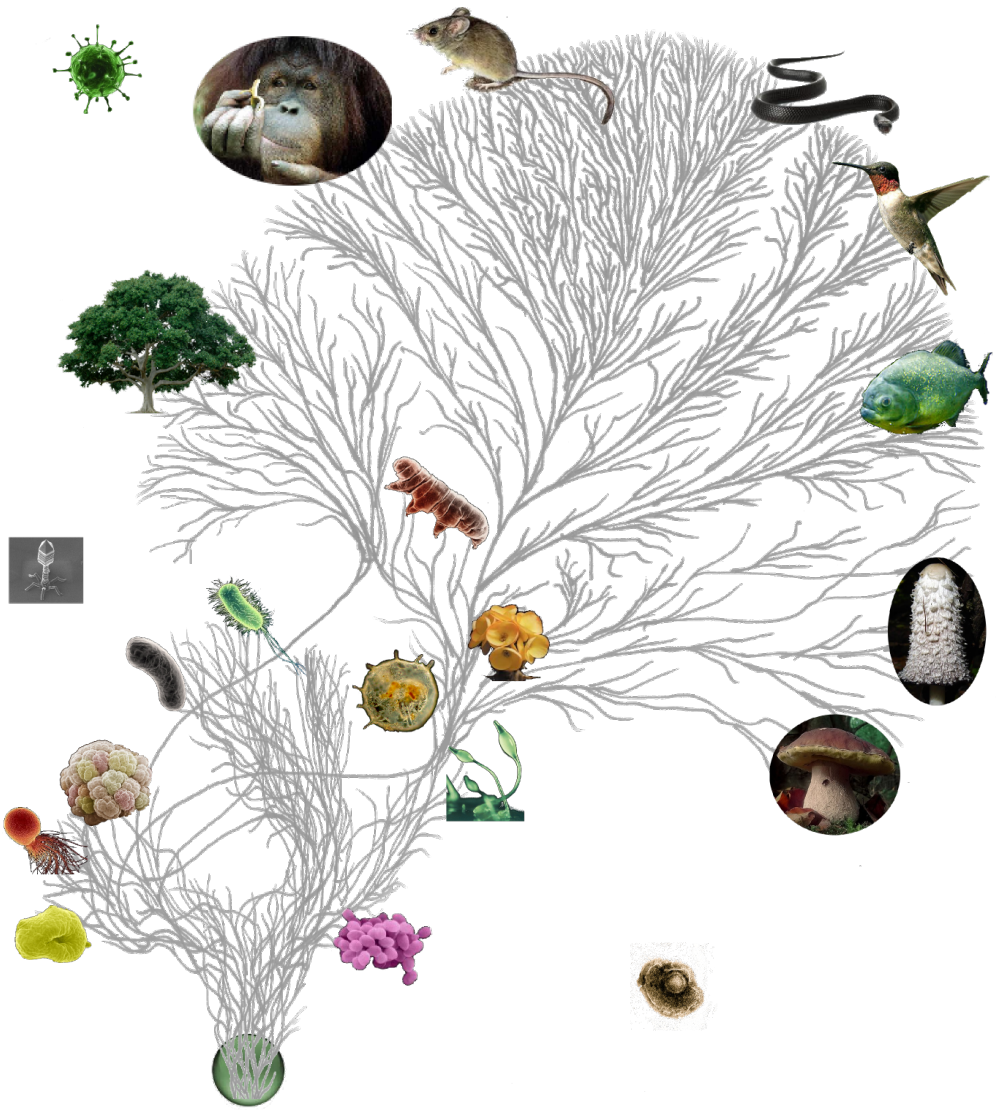
The work presented here consists in three different approximations in the analysis of genomes, each expressing a neutral model and its corresponding deviations.

- **Informational**: where genomes are viewed as simple sequences with a given informational content. We will define a common structure of genomes and show how they are characterized by a universal quasi-random structure.
- **Ecological**: here a genetic elements populating genomes are assimilated to biological species in ecosystems. We will apply ecological models on genomes in order to expose the neutral pattern behind their composition.
- **System**: focusing on proteins working together to complete a function we present a new strategy to find natural selection fingerprint at this level of study.

Index

Adaptive raise in organism complexity

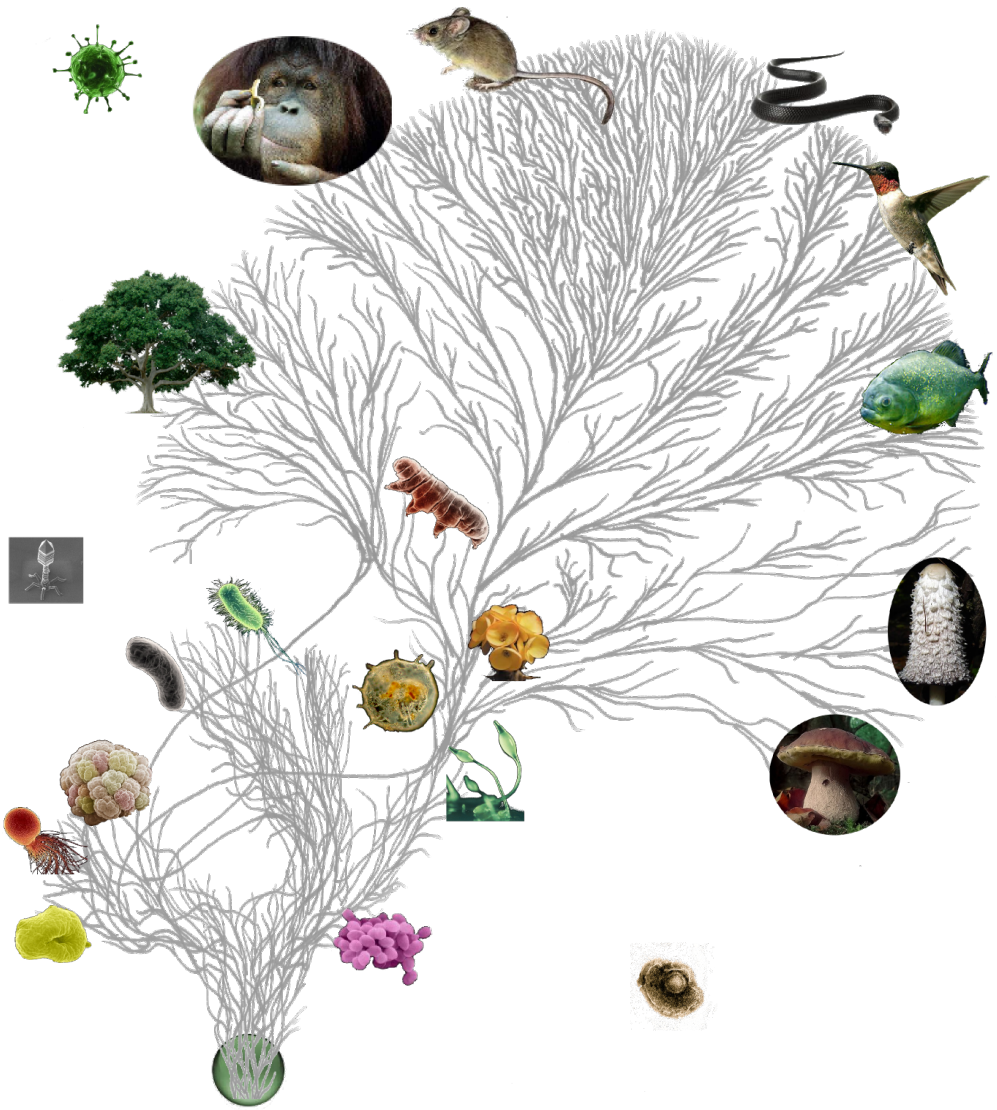
- Darwinian evolution does not encompass directional change nor global adaptive advance
- Not easy to define biological complexity
- At genomic level we would expect that the quantity of hereditary information is proportional to the level of complexity



adapted from [Maddison & Schulz - (2007)]

Adaptive raise in organism complexity

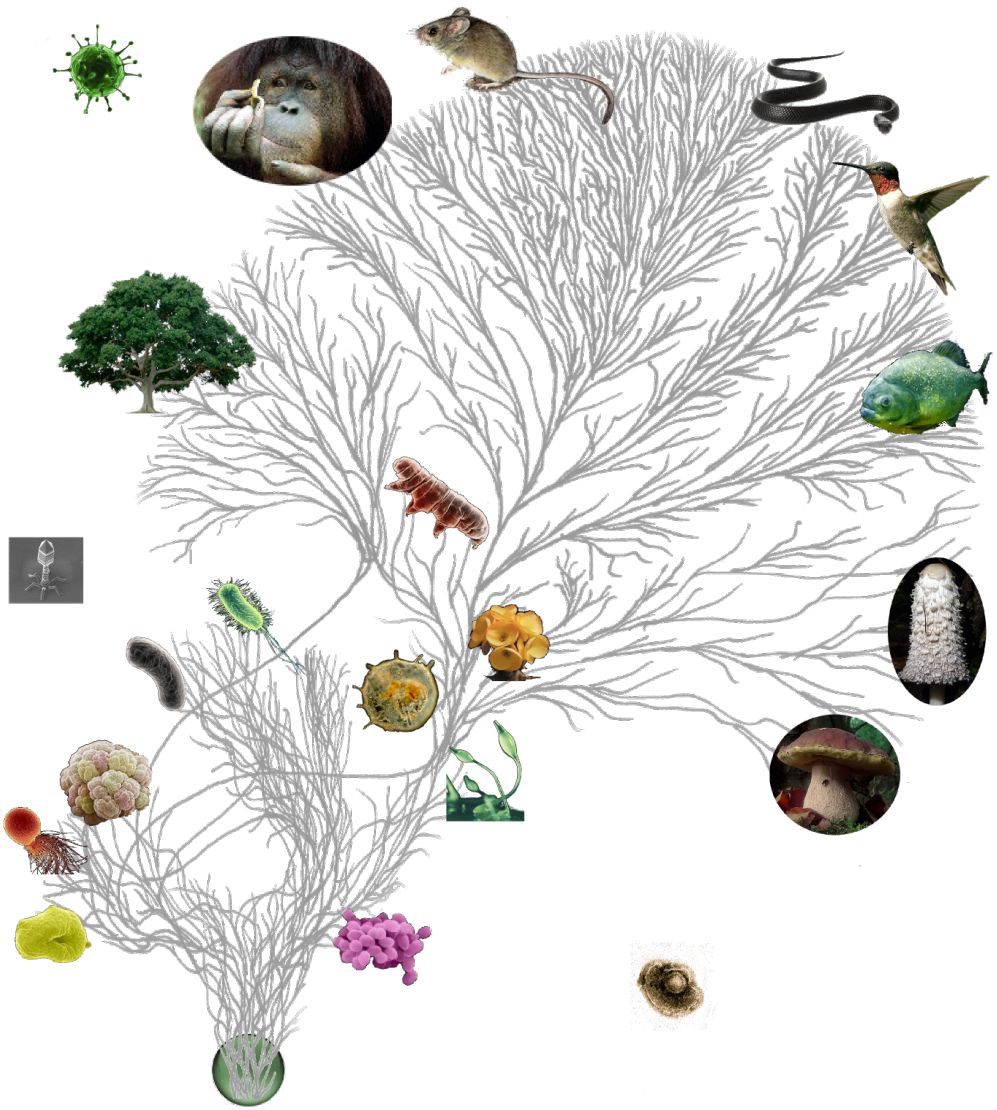
- Darwinian evolution does not encompass directional change nor global adaptive advance
- Not easy to define biological complexity
- At genomic level we would expect that the quantity of hereditary information is proportional to the level of complexity



adapted from [Maddison & Schulz - (2007)]

Adaptive raise in organism complexity

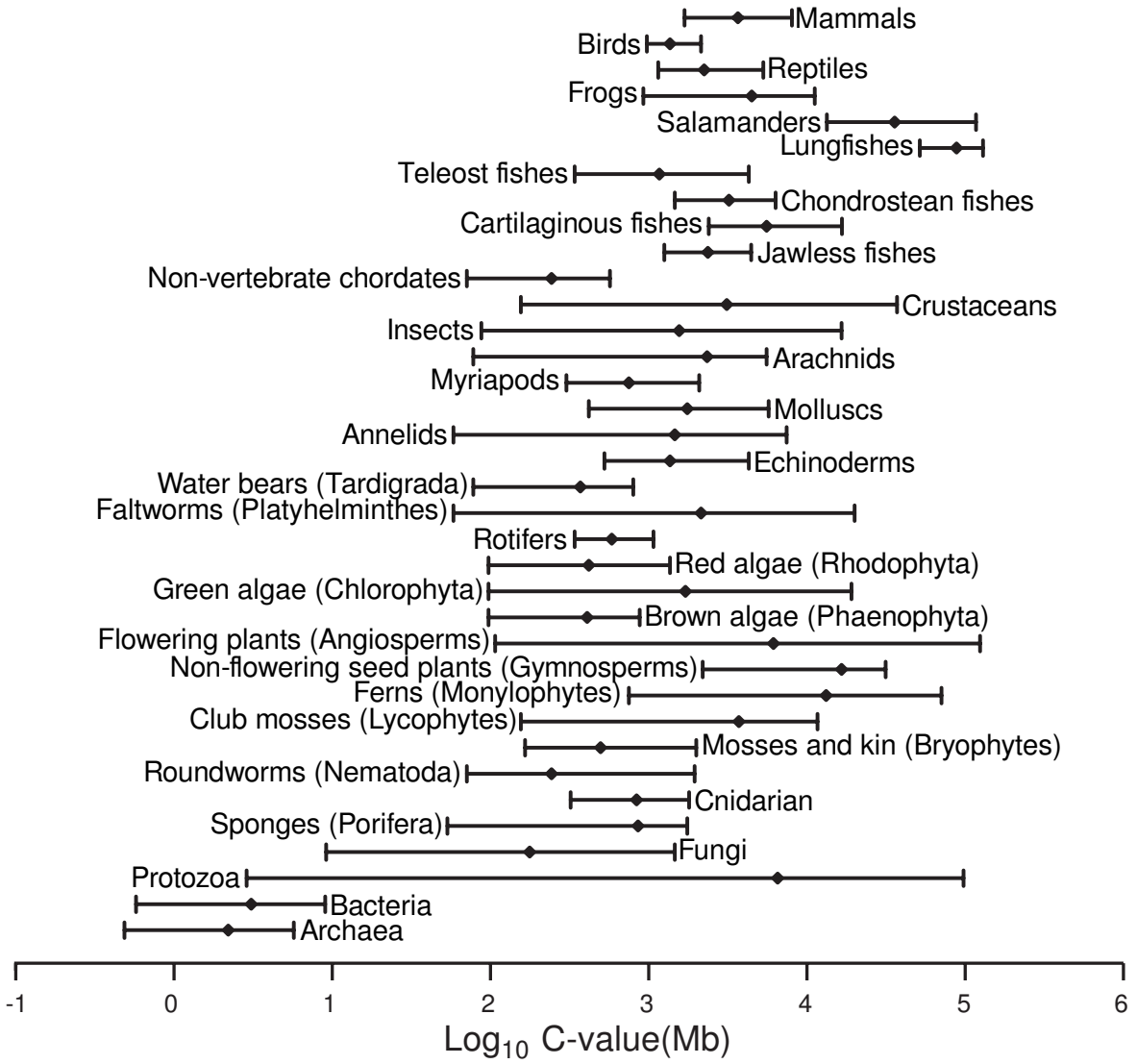
- Darwinian evolution does not encompass directional change nor global adaptive advance
- Not easy to define biological complexity
- At genomic level we would expect that the quantity of hereditary information is proportional to the level of complexity



adapted from [Maddison & Schulz - (2007)]

Challenging the C-value paradox

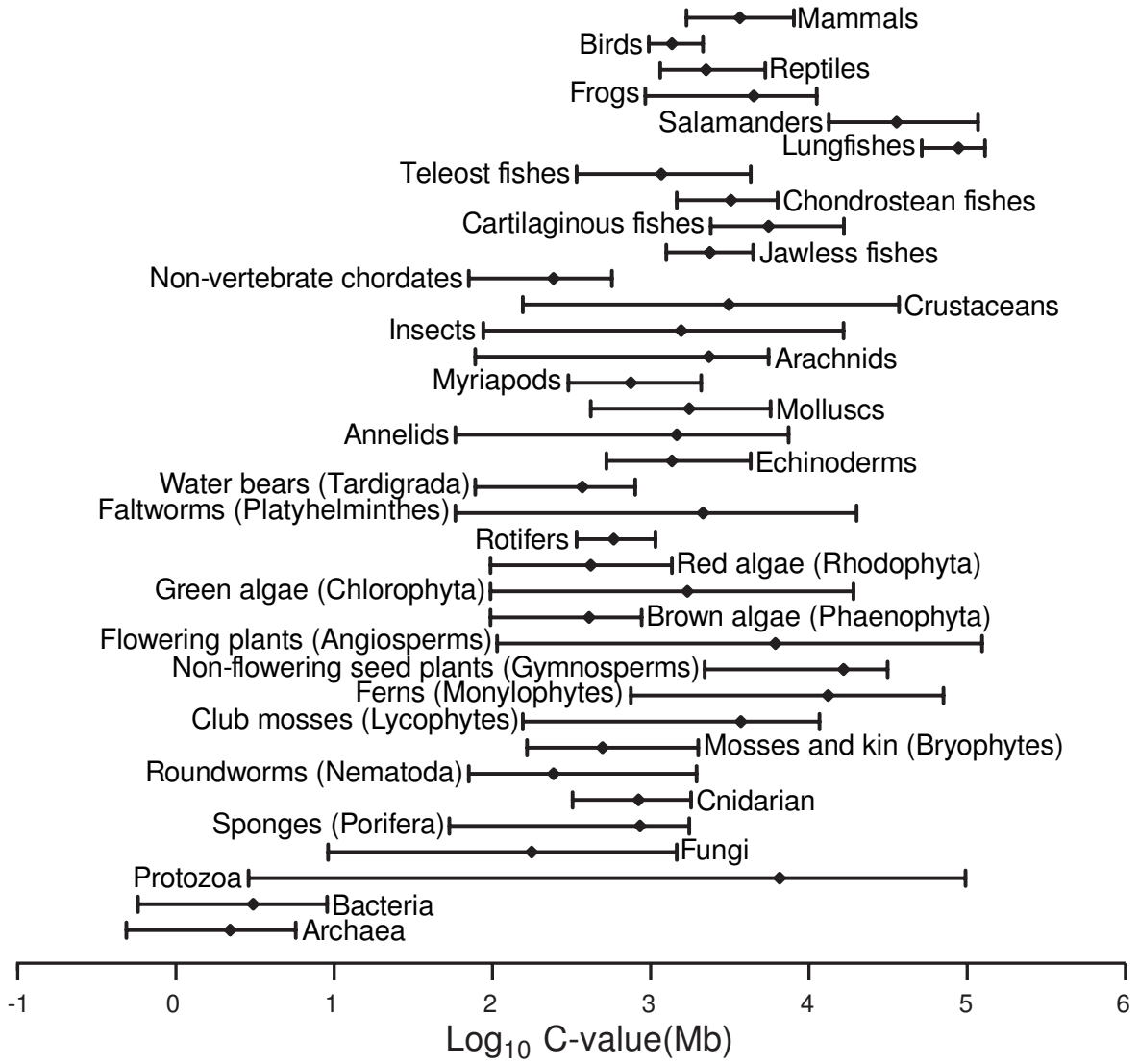
- DNA content (C-value) do not correlate with biological complexity
- What can be summarized as:
  - the search of patterns in the informational content of genomes
  - the dynamics underlying the distribution and appearance of non-coding DNA



[Gregory - Nature Reviews Genetics (2005)]

Challenging the C-value paradox

- DNA content (C-value) do not correlate with biological complexity
- What can be summarized as:
  - the search of patterns in the informational content of genomes
  - the dynamics underlying the distribution and appearance of non-coding DNA



[Gregory - Nature Reviews Genetics (2005)]

Informational content

- In order to normalize genomes DNA content, instead of working with direct measures of genome size, we work with informational content.
- Informational content corresponds to DNA complexity and may be related to the compressed genome size
- Indeed the complexity value we use is based on classical algorithms of data compression

Index

The Genome/Chromosome Complexity Value

Given a sequence, $seq = AACCTTCGTAGCATGG$:

#	Rotating sequence	*I.*	BWT	Char. list				MTF
0	AACCTTCGTAGCATG**G**|	0	**G**	**G**	a	t	c	0
1	ACCTTCGTAGCATGG|**A**	1	**A**	G	**A**	t	c	1
2	CCTTCGTAGCATGG|A**A**	5	**T**	A	g	**T**	c	2
3	CTTCGTAGCATGG|AA**C**	7	**C**	T	a	g	**C**	3
4	TTCGTAGCATGG|AAC**C**	15	G	C	t	a	**G**	3
5	TCGTAGCATGG|AACC**T**	13	A	G	c	t	**A**	3
6	CGTAGCATGG|AACCT**T**	6	T	A	g	c	**T**	3
7	GTAGCATGG|AACCTT**C**	11	C	T	a	g	**C**	3
8	TAGCATGG|AACCTTC**G**	12	G	C	t	a	**G**	3
9	AGCATGG|AACCTTCG**T**	2	A	G	c	t	**A**	3
10	GCATGG|AACCTTCGT**A**	9	T	A	g	c	**T**	3
11	CATGG|AACCTTCGTA**G**	4	C	T	a	g	**C**	3
12	ATGG|AACCTTCGTAG**C**	3	G	C	t	a	**G**	3
13	TGG|AACCTTCGTAGC**A**	14	T	G	c	**T**	a	2
14	GG|AACCTTCGTAGCA**T**	10	A	T	g	c	**A**	3
15	G|AACCTTCGTAGCAT**G**	8	C	A	t	g	**C**	3

$\Rightarrow CR(seq) = E(MTF(BWT(seq))) = E(0, 1, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 2, 3, 3) = 0.593$

- 54 species (virus to mammals) with genomes from 1.6Kb to 3.4Gb

The Genome/Chromosome Complexity Value

Given a sequence, $seq = AACCTTCGTAGCATGG$:

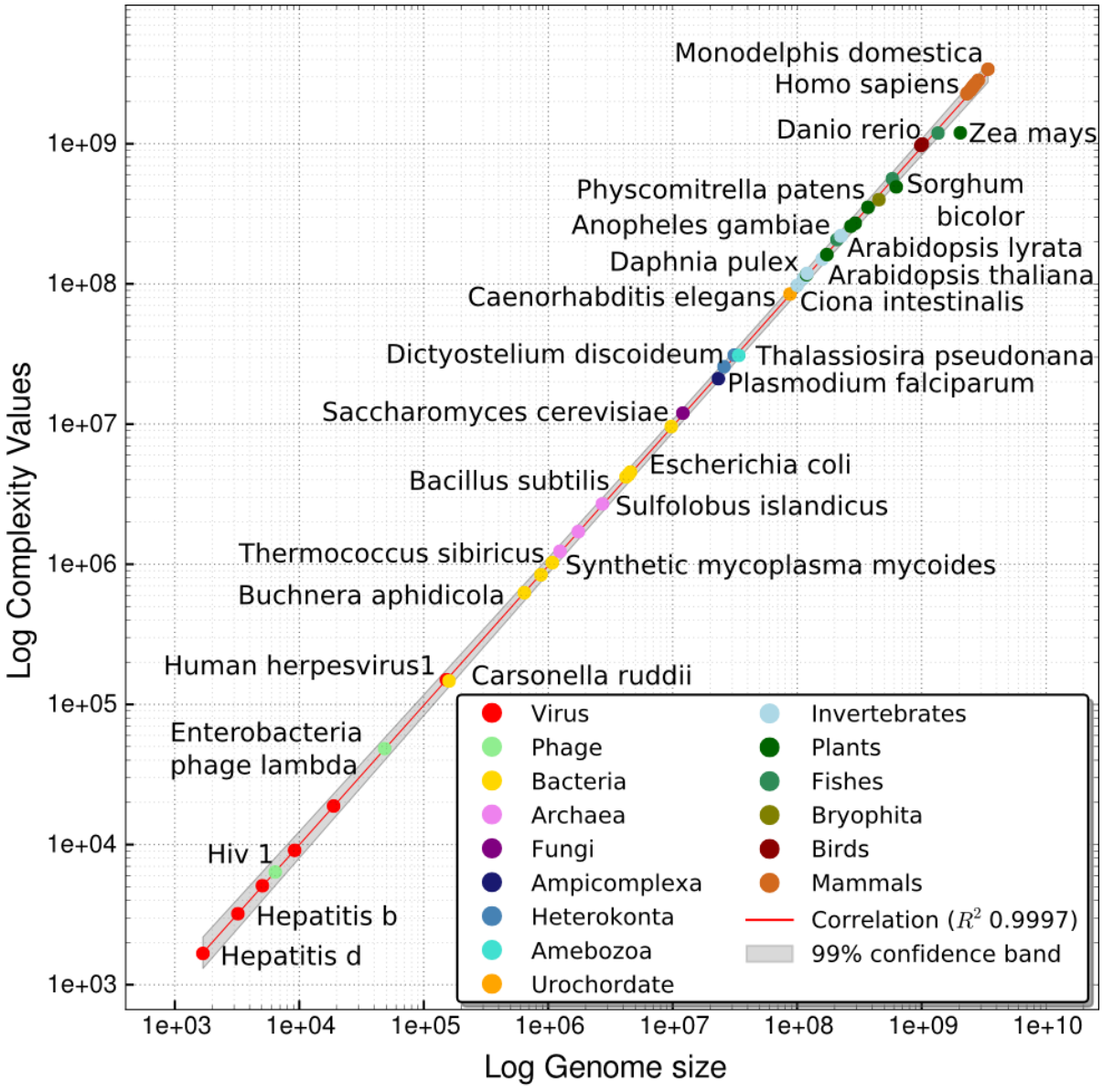
#	Rotating sequence	*I.*	BWT	Char. list				MTF
0	AACCTTCGTAGCATG**G**|	0	G	G	a	t	c	0
1	ACCTTCGTAGCATGG|**A**	1	A	G	A	t	c	1
2	CCTTCGTAGCATGG|A**A**	5	T	A	g	T	c	2
3	CTTCGTAGCATGG|AA**C**	7	C	T	a	g	C	3
4	TTCGTAGCATGG|AAC**C**	15	G	C	t	a	**G**	3
5	TCGTAGCATGG|AACC**T**	13	A	G	c	t	**A**	3
6	CGTAGCATGG|AACCT**T**	6	T	A	g	c	**T**	3
7	GTAGCATGG|AACCTT**C**	11	C	T	a	g	**C**	3
8	TAGCATGG|AACCTTC**G**	12	G	C	t	a	**G**	3
9	AGCATGG|AACCTTCG**T**	2	A	G	c	t	**A**	3
10	GCATGG|AACCTTCGT**A**	9	T	A	g	c	**T**	3
11	CATGG|AACCTTCGTA**G**	4	C	T	a	g	**C**	3
12	ATGG|AACCTTCGTAG**C**	3	G	C	t	a	**G**	3
13	TGG|AACCTTCGTAGC**A**	14	T	G	c	**T**	a	2
14	GG|AACCTTCGTAGCA**T**	10	A	T	g	c	**A**	3
15	G|AACCTTCGTAGCAT**G**	8	C	A	t	g	**C**	3

$\Rightarrow CR(seq) = E(MTF(BWT(seq))) = E(0, 1, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 2, 3, 3) = 0.593$

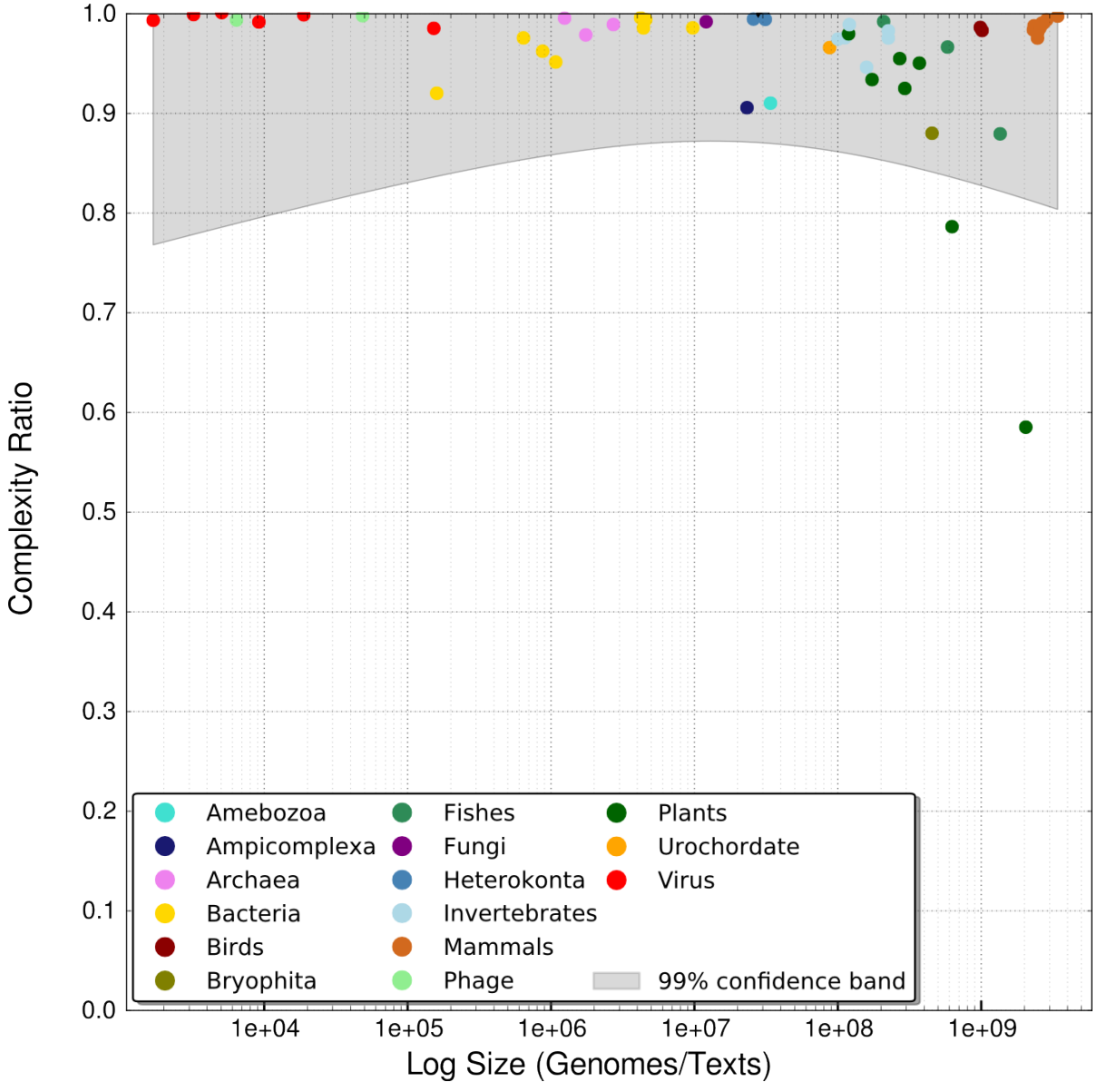
- 54 species (virus to mammals) with genomes from 1.6Kb to 3.4Gb

Index

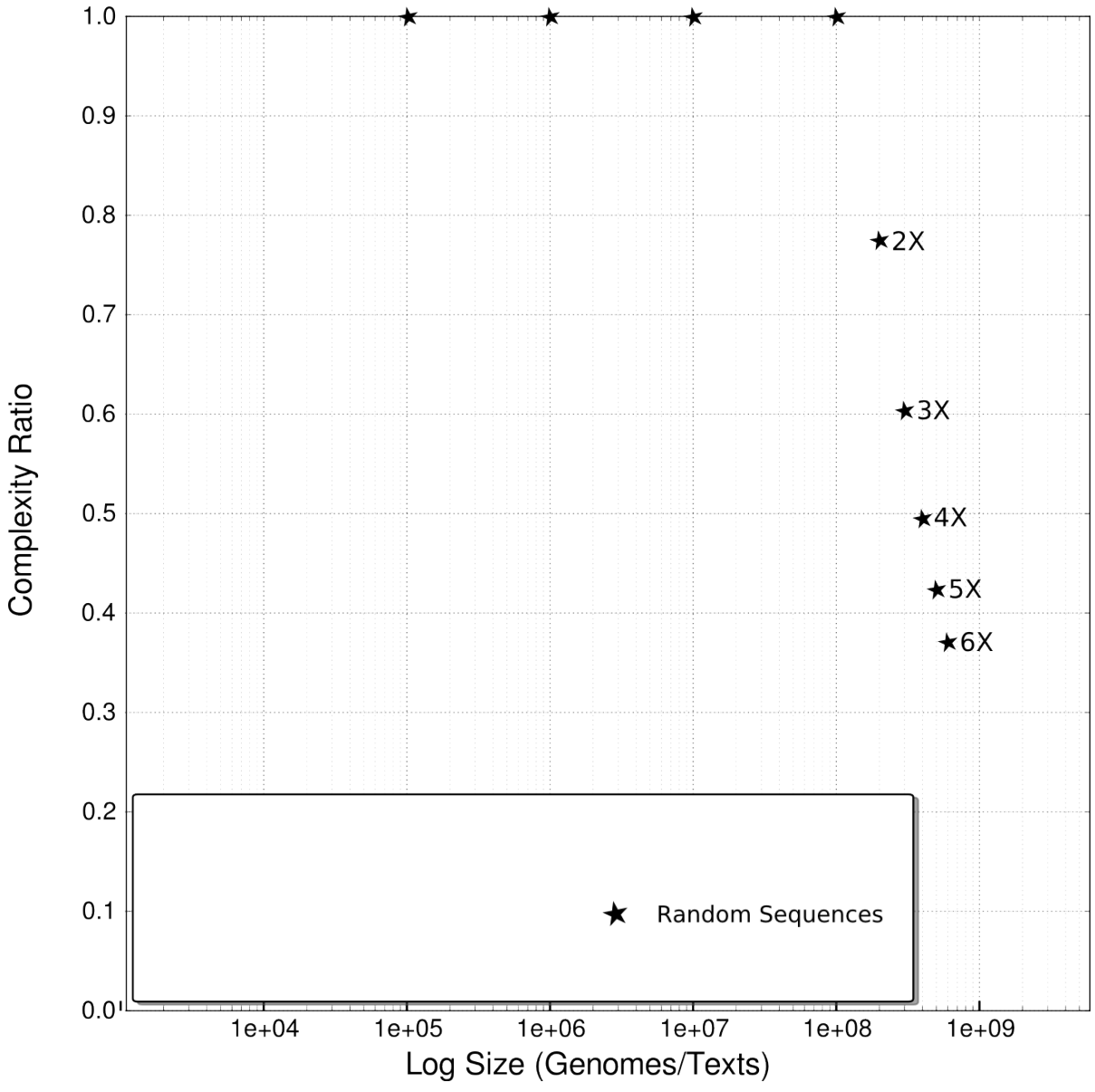
Complexity Value of genomes



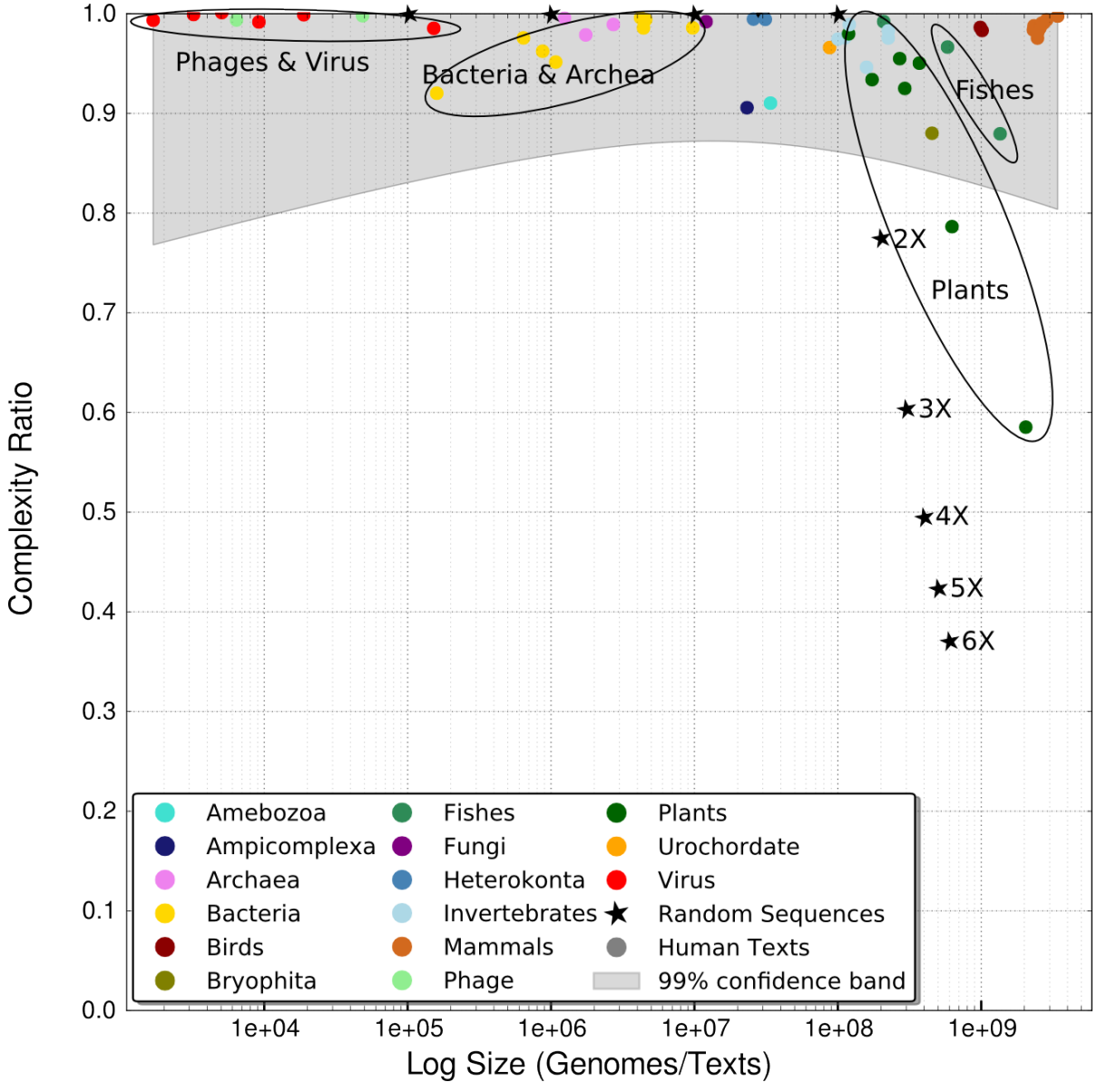
Complexity Ratio of genomes



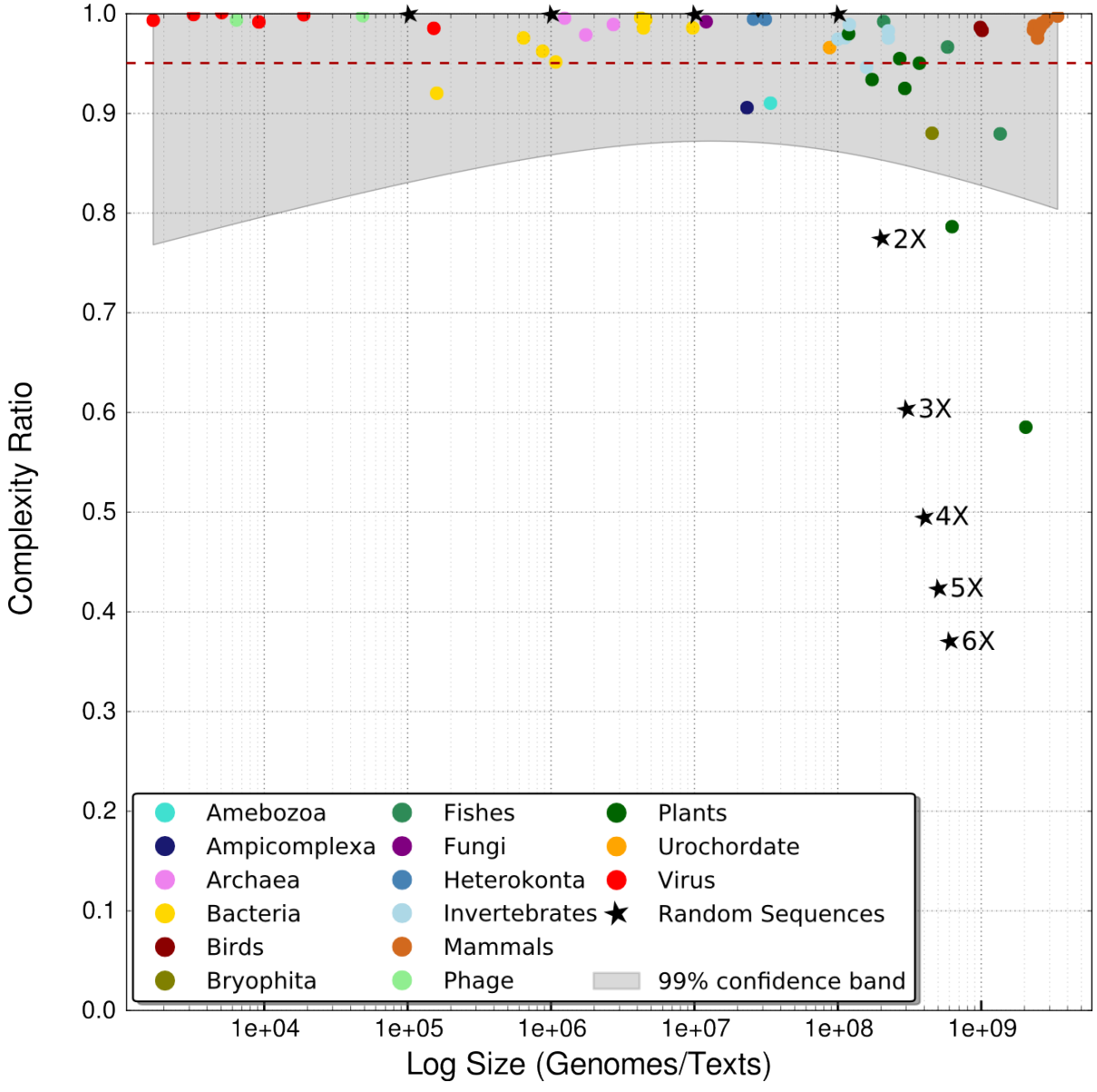
Complexity Ratio of genomes



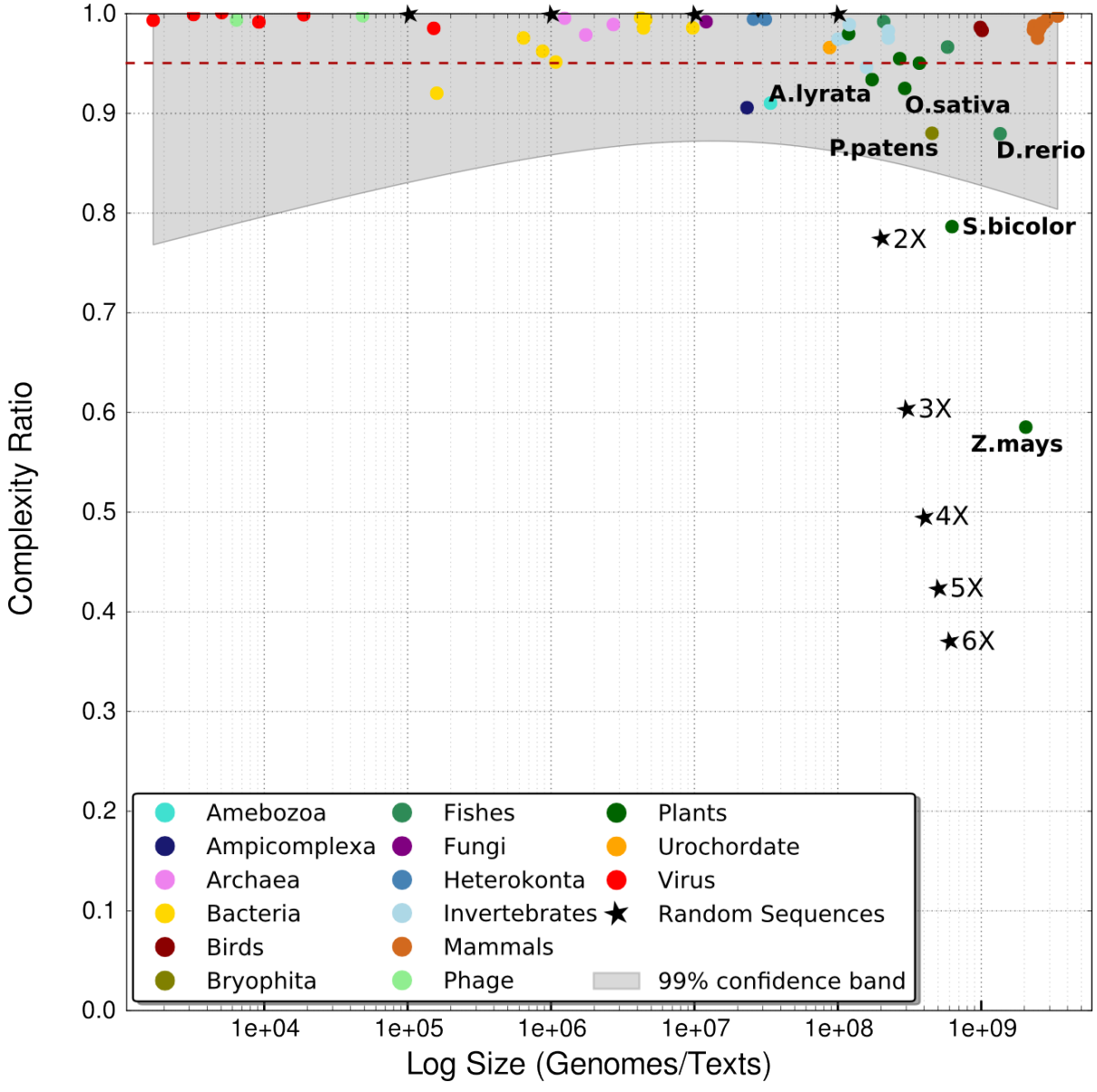
Complexity Ratio of genomes



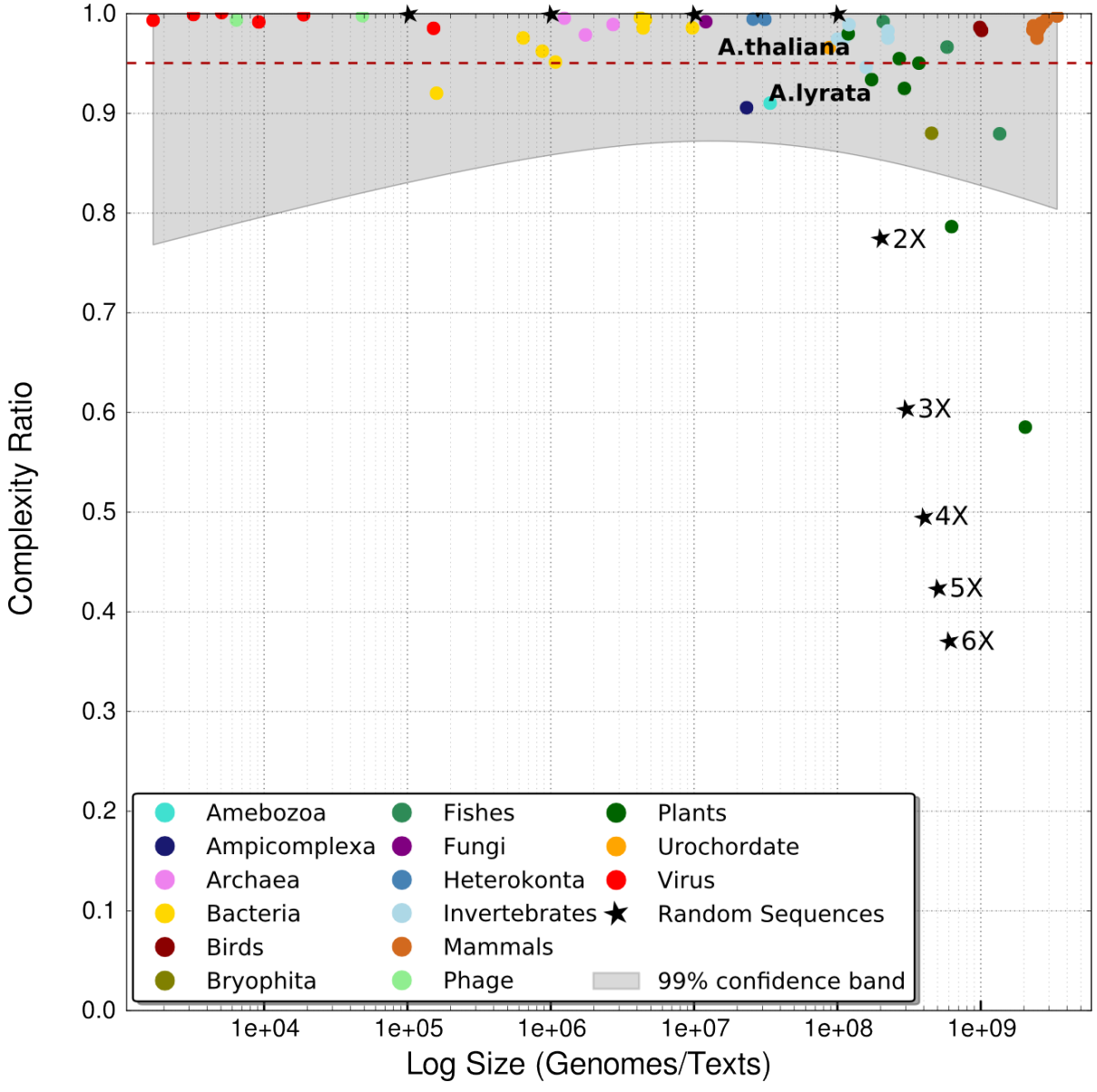
Complexity Ratio of genomes



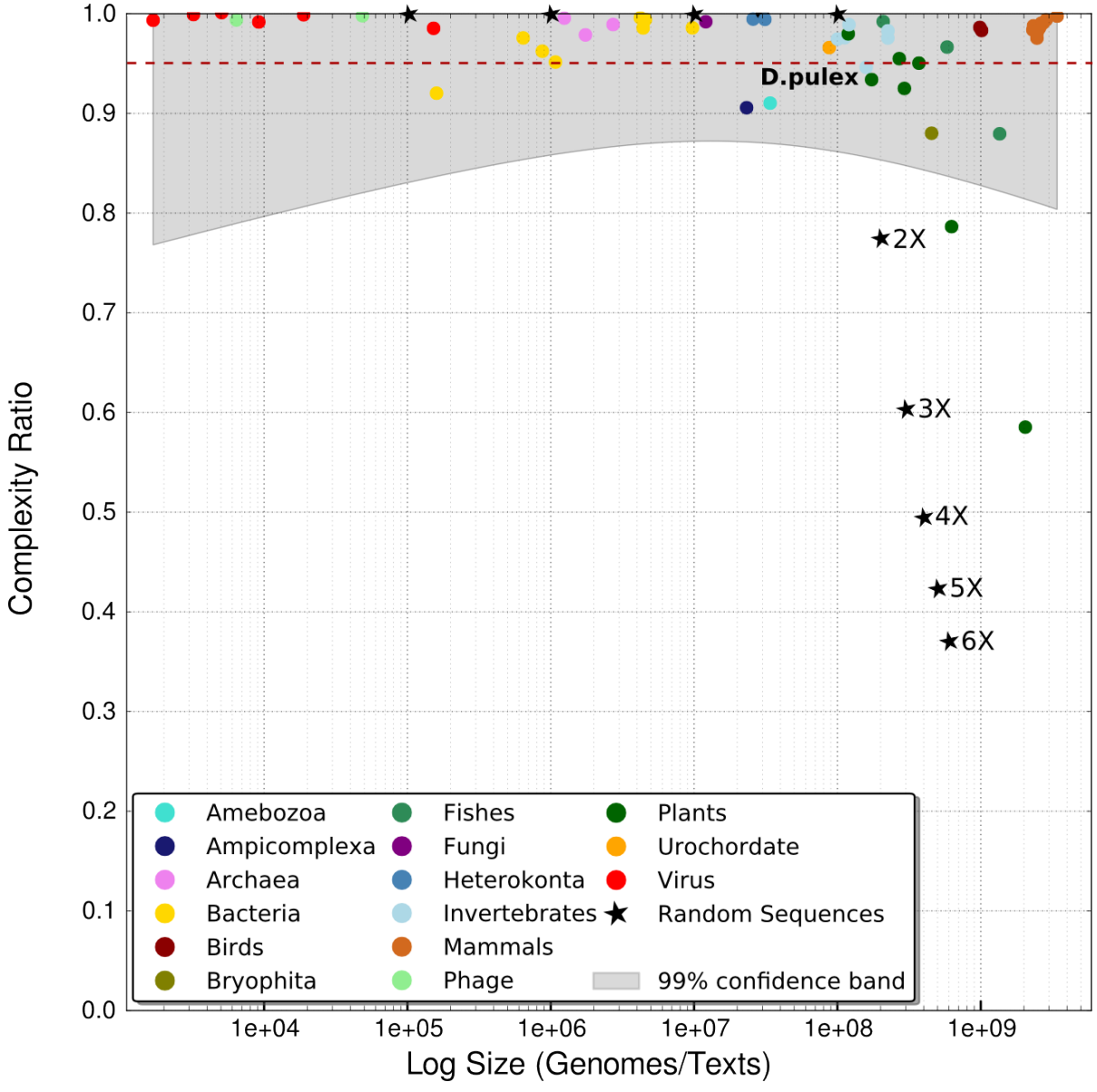
Complexity Ratio of genomes



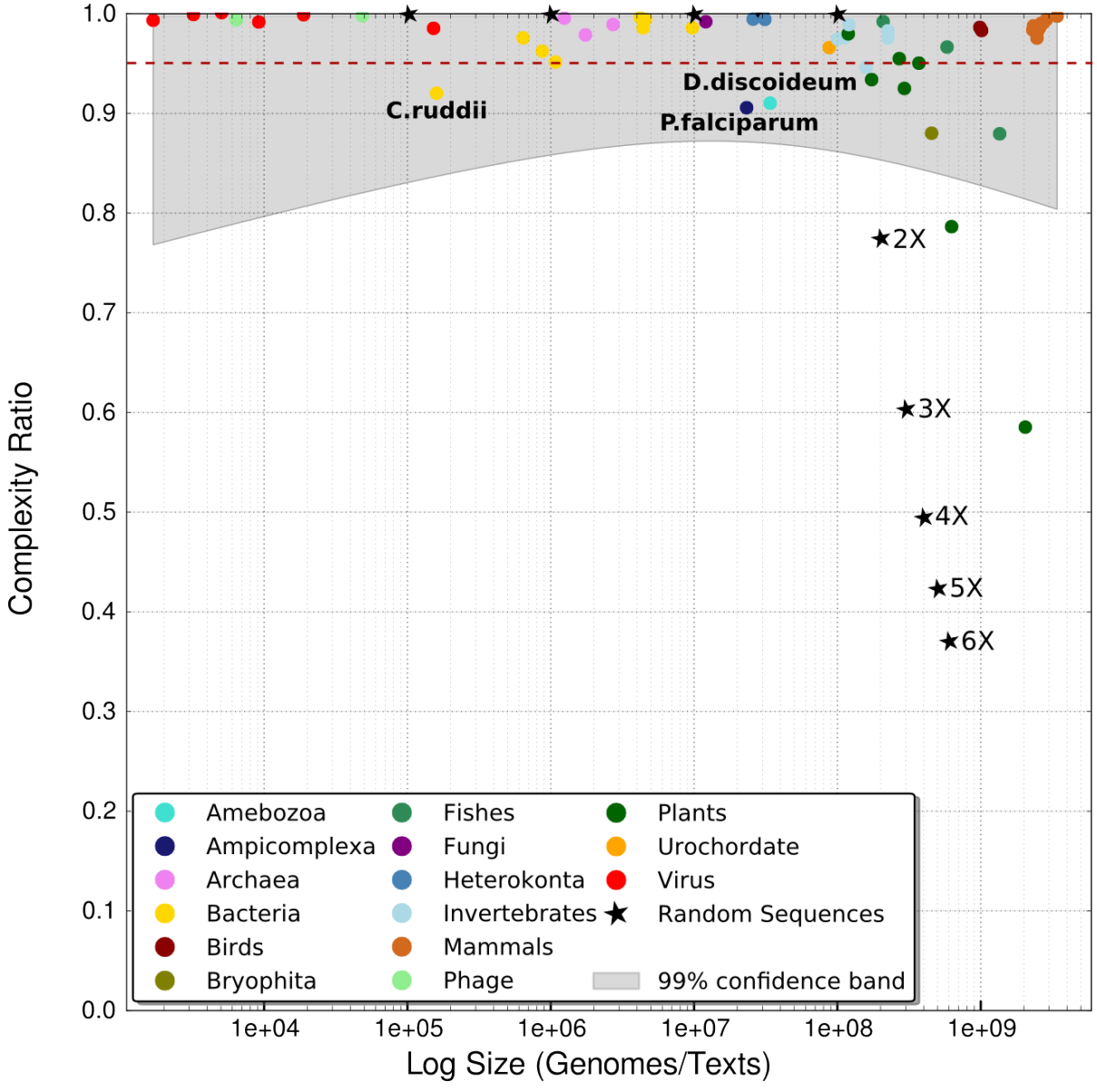
Complexity Ratio of genomes



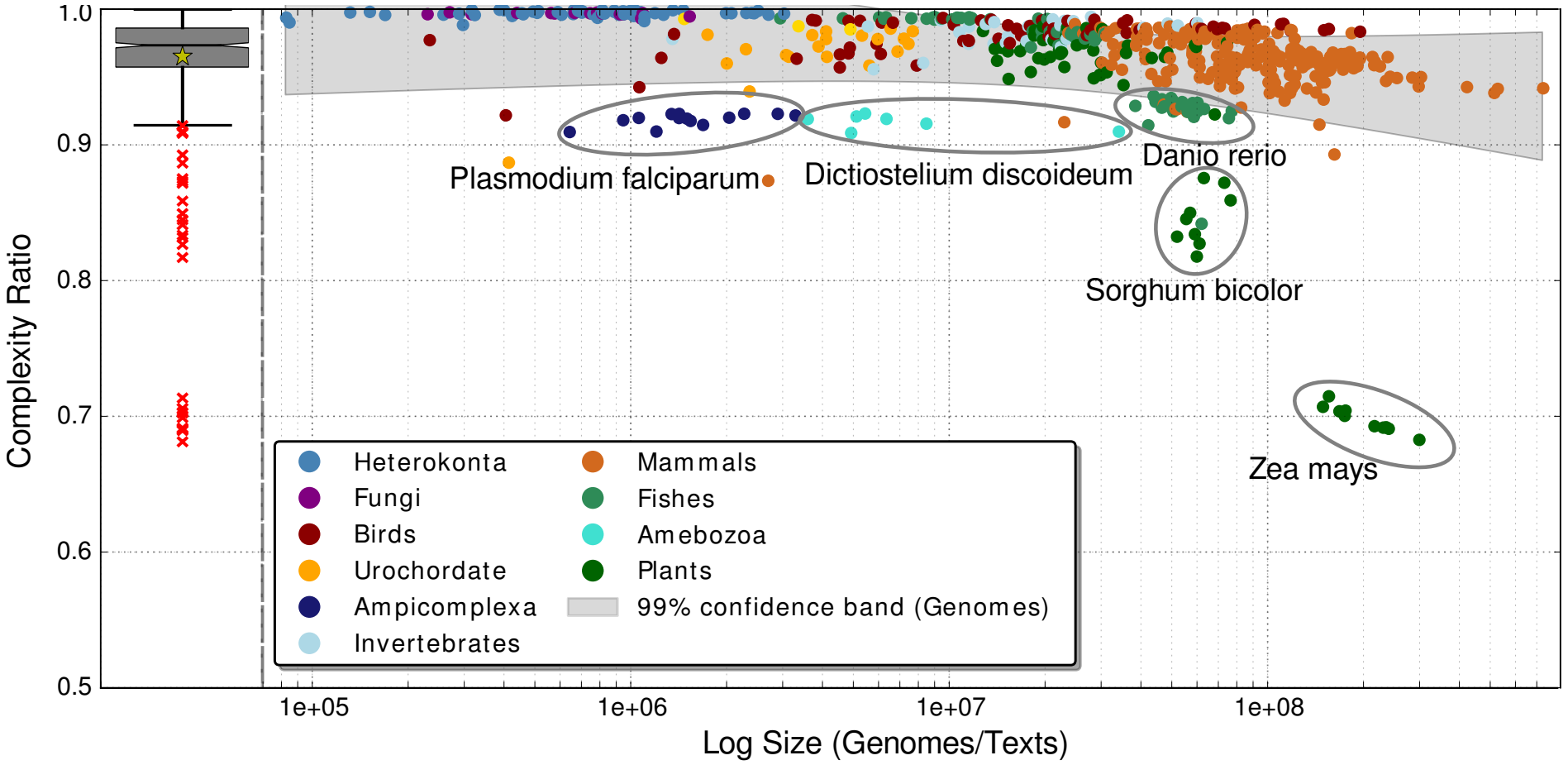
Complexity Ratio of genomes



Complexity Ratio of genomes



Complexity Ratio of chromosomes



Hypotheses

- These observations led us to hypothesize that:
  - A quasi-random combinatorial structure of DNA is a universal feature of non-polyploid genomes
  - The fate of polyploid genomes is to reach almost maximal CR. Increases in CR occur as a function of time
  - Genome complexity only increases through rounds of duplications followed by the divergence of duplicates during evolution
- These hypotheses can be nullified in some specific cases:
  - Genomes of recent polyploid species evidencing high CR, e.g.: species that have undergone significant genome reduction
  - Genomes evidencing a non-random DNA structure (low CR), due to a very strong GC content bias

Hypotheses

- These observations led us to hypothesize that:
  - A quasi-random combinatorial structure of DNA is a universal feature of non-polyploid genomes
  - The fate of polyploid genomes is to reach almost maximal CR. Increases in CR occur as a function of time
  - Genome complexity only increases through rounds of duplications followed by the divergence of duplicates during evolution
- These hypotheses can be nullified in some specific cases:
  - Genomes of recent polyploid species evidencing high CR, e.g.: species that have undergone significant genome reduction
  - Genomes evidencing a non-random DNA structure (low CR), due to a very strong GC content bias

Hypotheses

- These observations led us to hypothesize that:
  - A quasi-random combinatorial structure of DNA is a universal feature of non-polyploid genomes
  - The fate of polyploid genomes is to reach almost maximal CR. Increases in CR occur as a function of time
  - Genome complexity only increases through rounds of duplications followed by the divergence of duplicates during evolution
- These hypotheses can be nullified in some specific cases:
  - Genomes of recent polyploid species evidencing high CR, e.g.: species that have undergone significant genome reduction
  - Genomes evidencing a non-random DNA structure (low CR), due to a very strong GC content bias

Hypotheses

- These observations led us to hypothesize that:
  - A quasi-random combinatorial structure of DNA is a universal feature of non-polyploid genomes
  - The fate of polyploid genomes is to reach almost maximal CR. Increases in CR occur as a function of time
  - Genome complexity only increases through rounds of duplications followed by the divergence of duplicates during evolution
- These hypotheses can be nullified in some specific cases:
  - Genomes of recent polyploid species evidencing high CR, e.g.: species that have undergone significant genome reduction
  - Genomes evidencing a non-random DNA structure (low CR), due to a very strong GC content bias

Hypotheses

- These observations led us to hypothesize that:
  - A quasi-random combinatorial structure of DNA is a universal feature of non-polyploid genomes
  - The fate of polyploid genomes is to reach almost maximal CR. Increases in CR occur as a function of time
  - Genome complexity only increases through rounds of duplications followed by the divergence of duplicates during evolution
- These hypotheses can be nullified in some specific cases:
  - Genomes of recent polyploid species evidencing high CR, e.g.: species that have undergone significant genome reduction
  - Genomes evidencing a non-random DNA structure (low CR), due to a very strong GC content bias

Hypotheses

- These observations led us to hypothesize that:
  - A quasi-random combinatorial structure of DNA is a universal feature of non-polyploid genomes
  - The fate of polyploid genomes is to reach almost maximal CR. Increases in CR occur as a function of time
  - Genome complexity only increases through rounds of duplications followed by the divergence of duplicates during evolution
- These hypotheses can be nullified in some specific cases:
  - Genomes of recent polyploid species evidencing high CR, e.g.: species that have undergone significant genome reduction
  - Genomes evidencing a non-random DNA structure (low CR), due to a very strong GC content bias

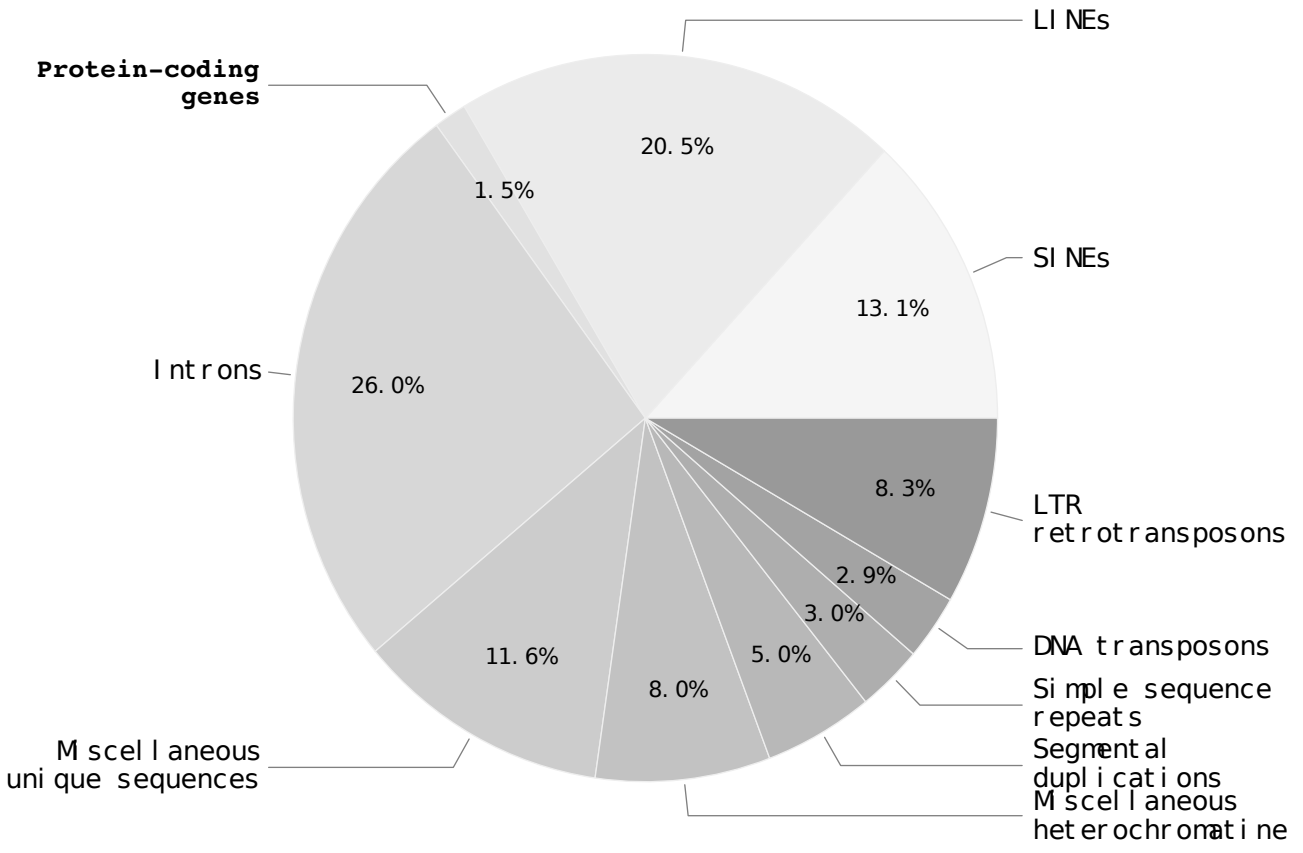
Hypotheses

- These observations led us to hypothesize that:
  - A quasi-random combinatorial structure of DNA is a universal feature of non-polyploid genomes
  - The fate of polyploid genomes is to reach almost maximal CR. Increases in CR occur as a function of time
  - Genome complexity only increases through rounds of duplications followed by the divergence of duplicates during evolution
- These hypotheses can be nullified in some specific cases:
  - Genomes of recent polyploid species evidencing high CR, e.g.: species that have undergone significant genome reduction
  - Genomes evidencing a non-random DNA structure (low CR), due to a very strong GC content bias

Index

Variation of families of genetic elements in eukaryote

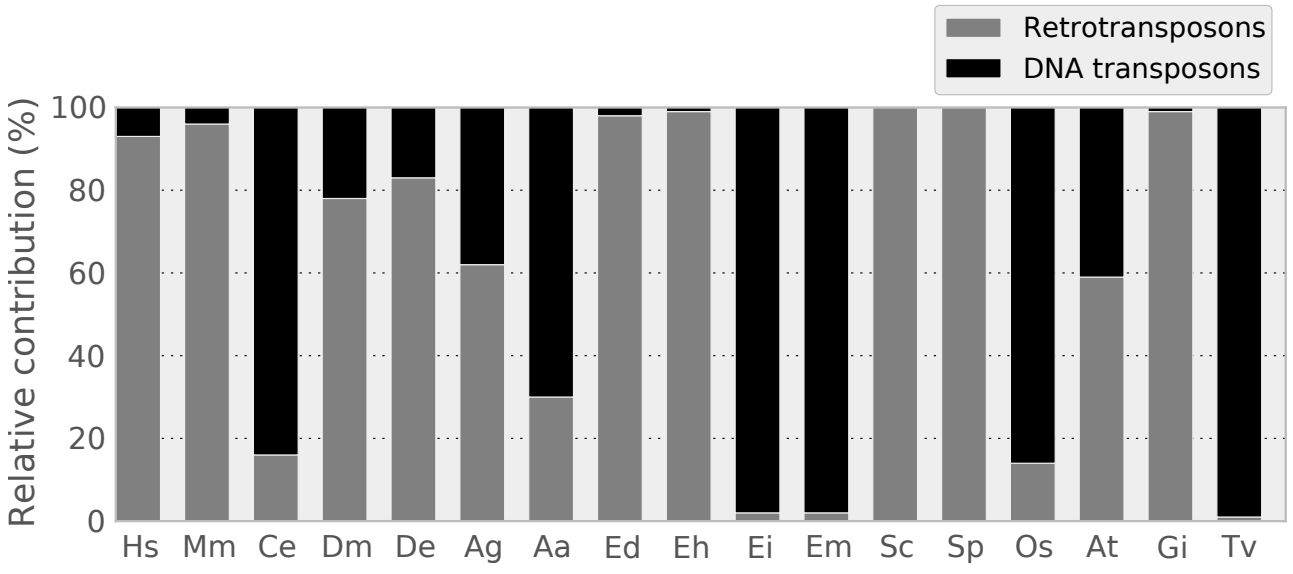
- Genomes are populated with genetic elements, coding or non-coding
- Proportions of these families of genetic elements vary greatly between species



adapted from *[Lander* et al. - *Nature (2001)]*

Variation of families of genetic elements in eukaryote

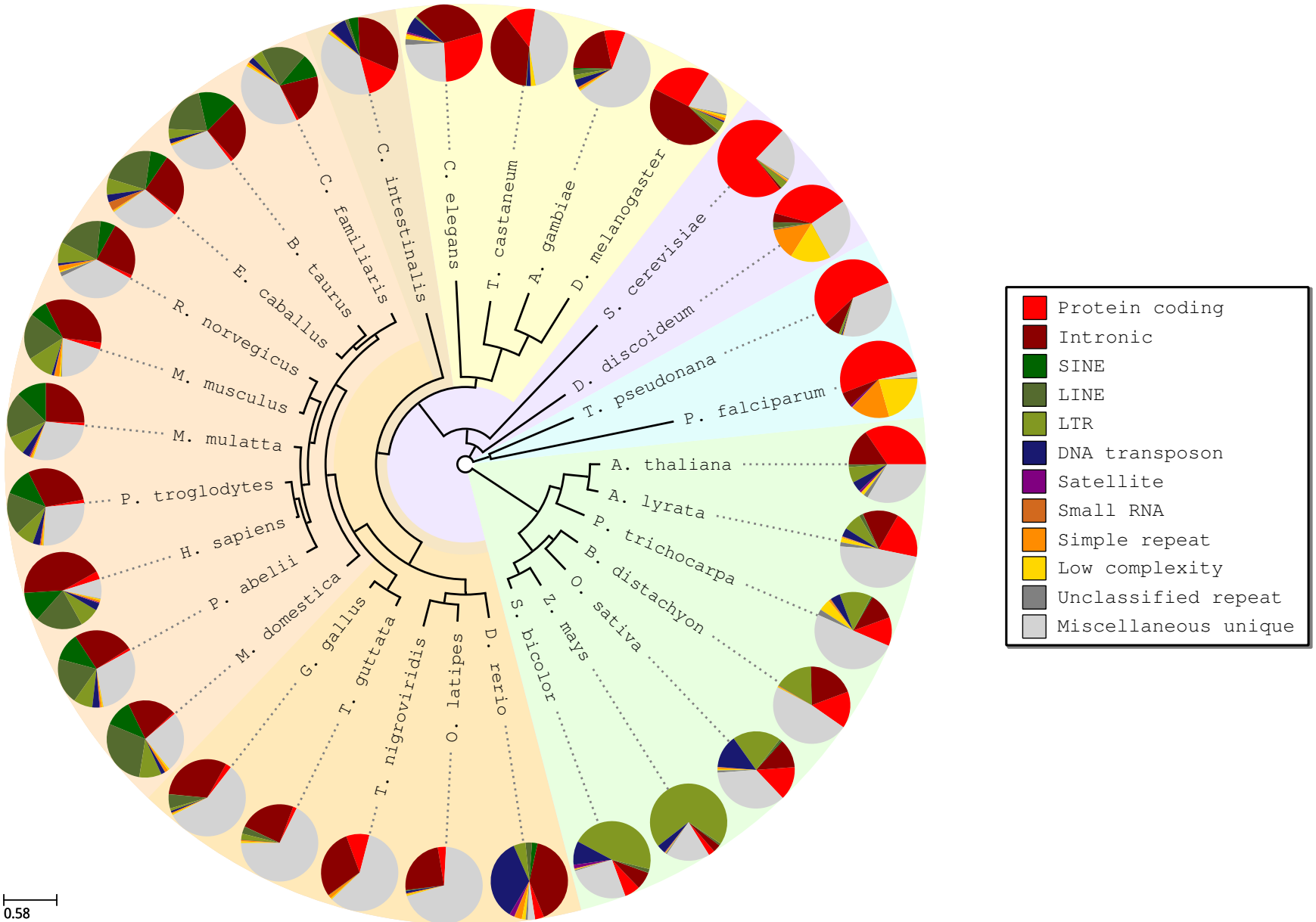
- Genomes are populated with genetic elements, coding or non-coding
- Proportions of these families of genetic elements vary greatly between species



adapted from *[Pritham - The Journal of Heredity (2009)]*

Index

Distribution of genetic species within biological species



Non-random distribution of genetic species

- Test for the simplest hypotheses: **each of the genetic species (GSs) are found in equal proportions in each chromosome of a given genome**
- We generated 1,000 random genomes for each of the 31 biological species
- Test the differences between the GSs in the 1,000 shuffling and the observed value in 548 chromosomes
- In less than **4% of the cases GS were found in the expected proportion**

Non-random distribution of genetic species

- Test for the simplest hypotheses: **each of the genetic species (GSs) are found in equal proportions in each chromosome of a given genome**
- We generated 1,000 random genomes for each of the 31 biological species
- Test the differences between the GSs in the 1,000 shuffling and the observed value in 548 chromosomes
- In less than **4% of the cases GS were found in the expected proportion**

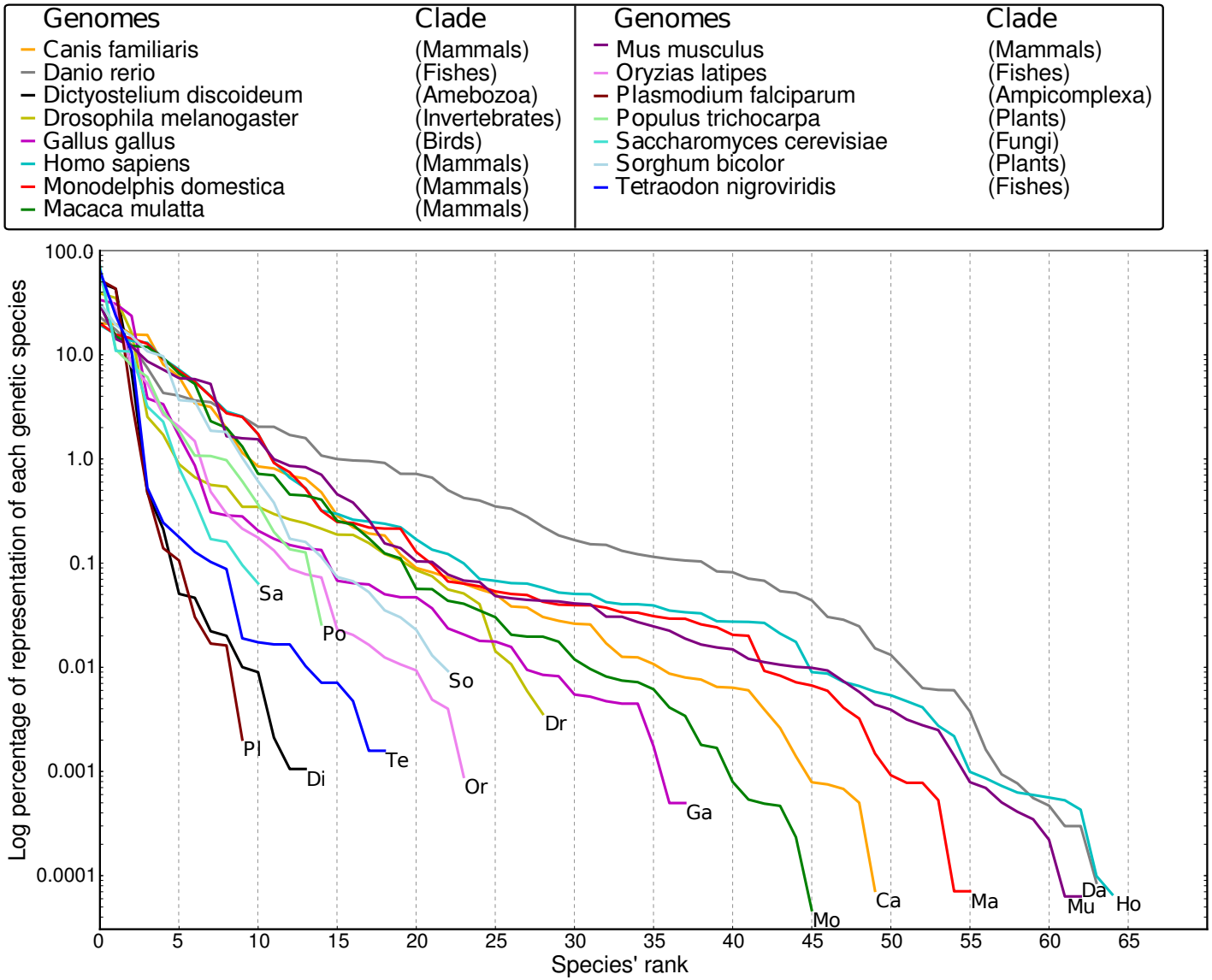
Non-random distribution of genetic species

- Test for the simplest hypotheses: **each of the genetic species (GSs) are found in equal proportions in each chromosome of a given genome**
- We generated 1,000 random genomes for each of the 31 biological species
- Test the differences between the GSs in the 1,000 shuffling and the observed value in 548 chromosomes
- In less than **4% of the cases GS were found in the expected proportion**

Non-random distribution of genetic species

- Test for the simplest hypotheses: **each of the genetic species (GSs) are found in equal proportions in each chromosome of a given genome**
- We generated 1,000 random genomes for each of the 31 biological species
- Test the differences between the GSs in the 1,000 shuffling and the observed value in 548 chromosomes
- In less than **4% of the cases GS were found in the expected proportion**

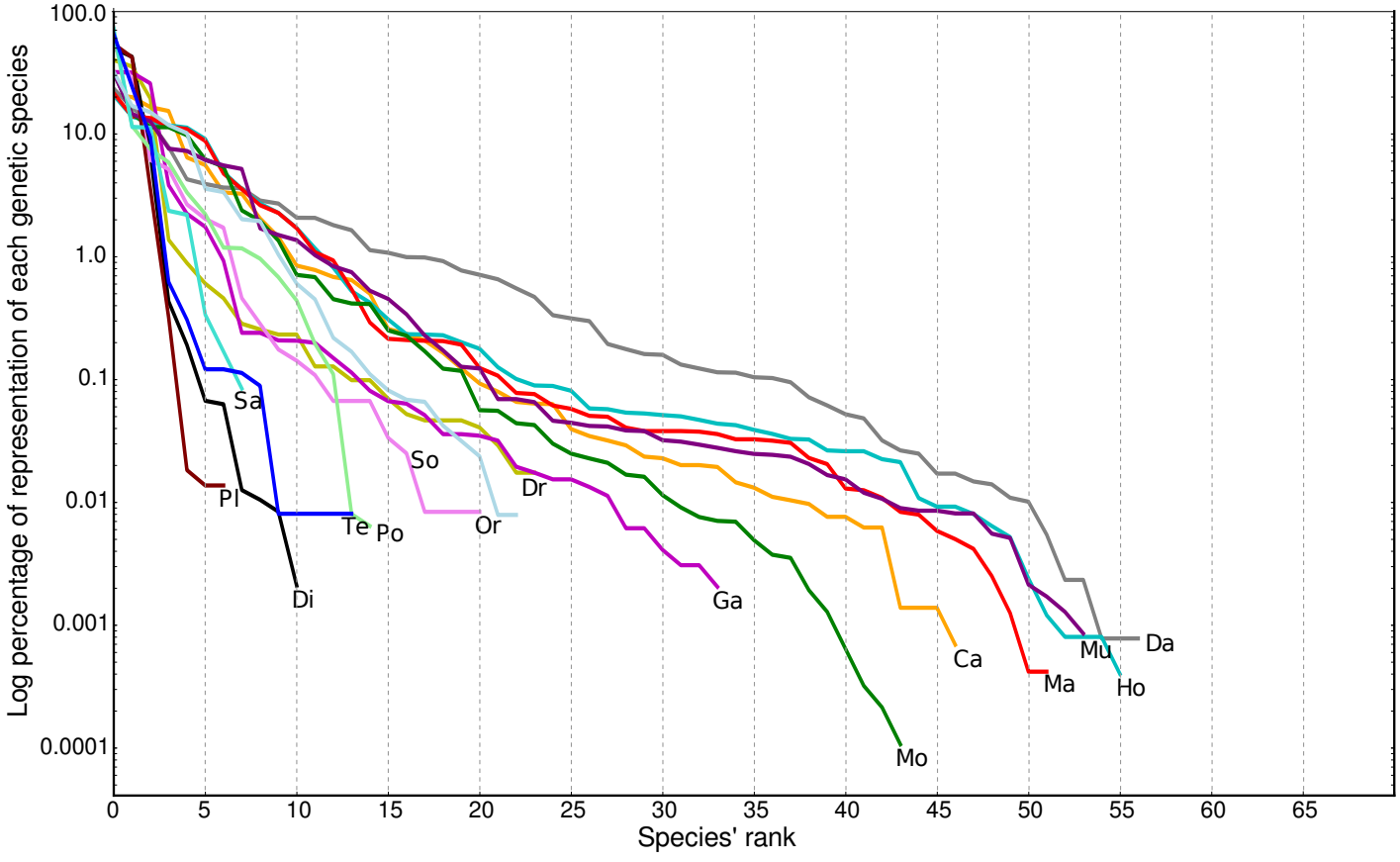
Counterbalanced GSs abundances in genomes



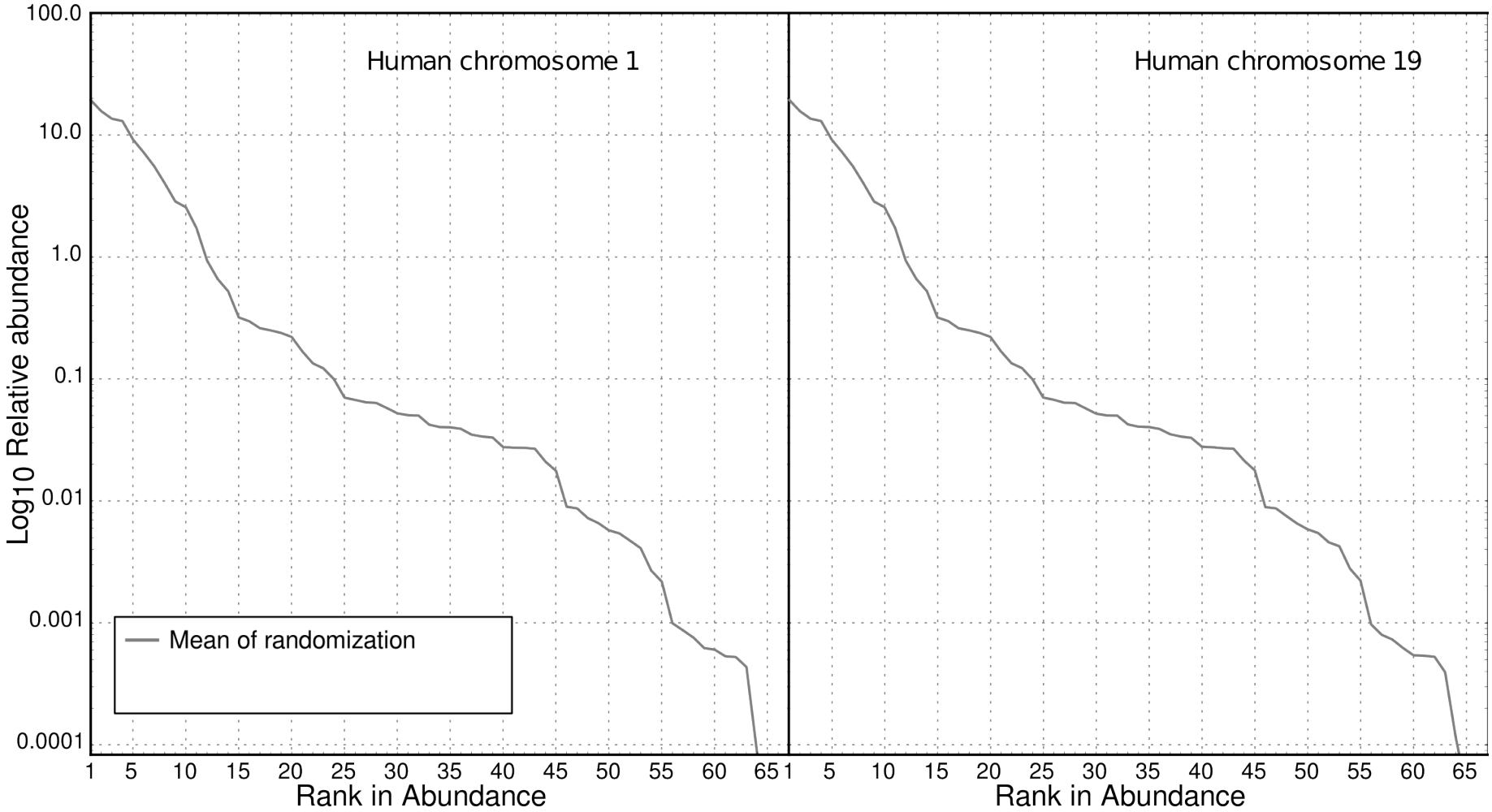
Counterbalanced GSs abundances in Chromosomes

Genomes	Chrm.	Clade
Canis familiaris	1	(Mammals)
Danio rerio	7	(Fishes)
Dictyostelium discoideum	2	(Amebozoa)
Drosophila melanogaster	3R	(Invertebrates)
Gallus gallus	1	(Birds)
Homo sapiens	1	(Mammals)
Monodelphis domestica	1	(Mammals)
Macaca mulatta	1	(Mammals)

Genomes	Chrm.	Clade
Mus musculus	1	(Mammals)
Oryzias latipes	3	(Fishes)
Plasmodium falciparum	14	(Ampicomplexa)
Populus trichocarpa	1	(Plants)
Saccharomyces cerevisiae	IV	(Fungi)
Sorghum bicolor	2	(Plants)
Tetraodon nigroviridis	1	(Fishes)

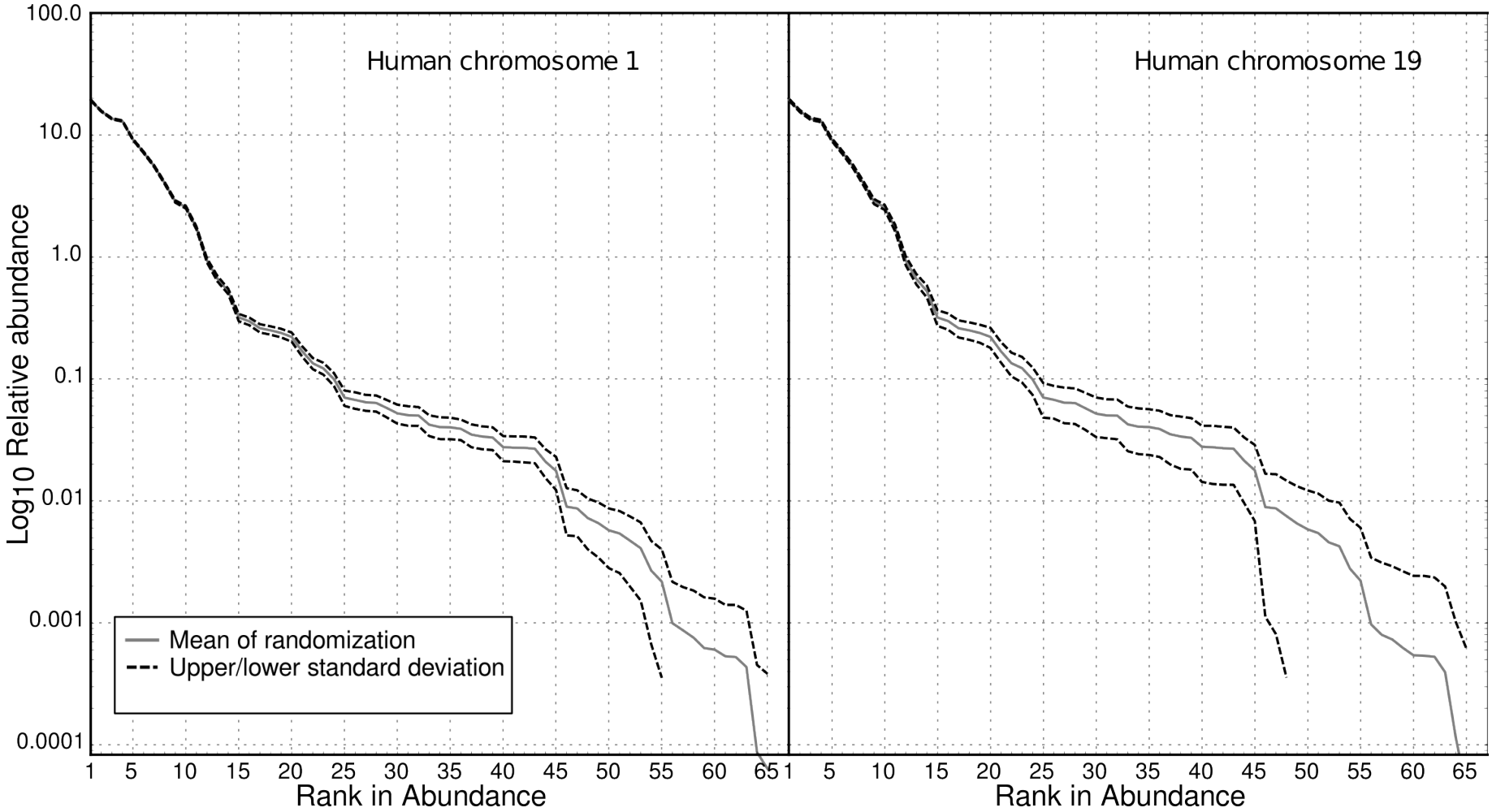


Relative species abundance: observed vs expected



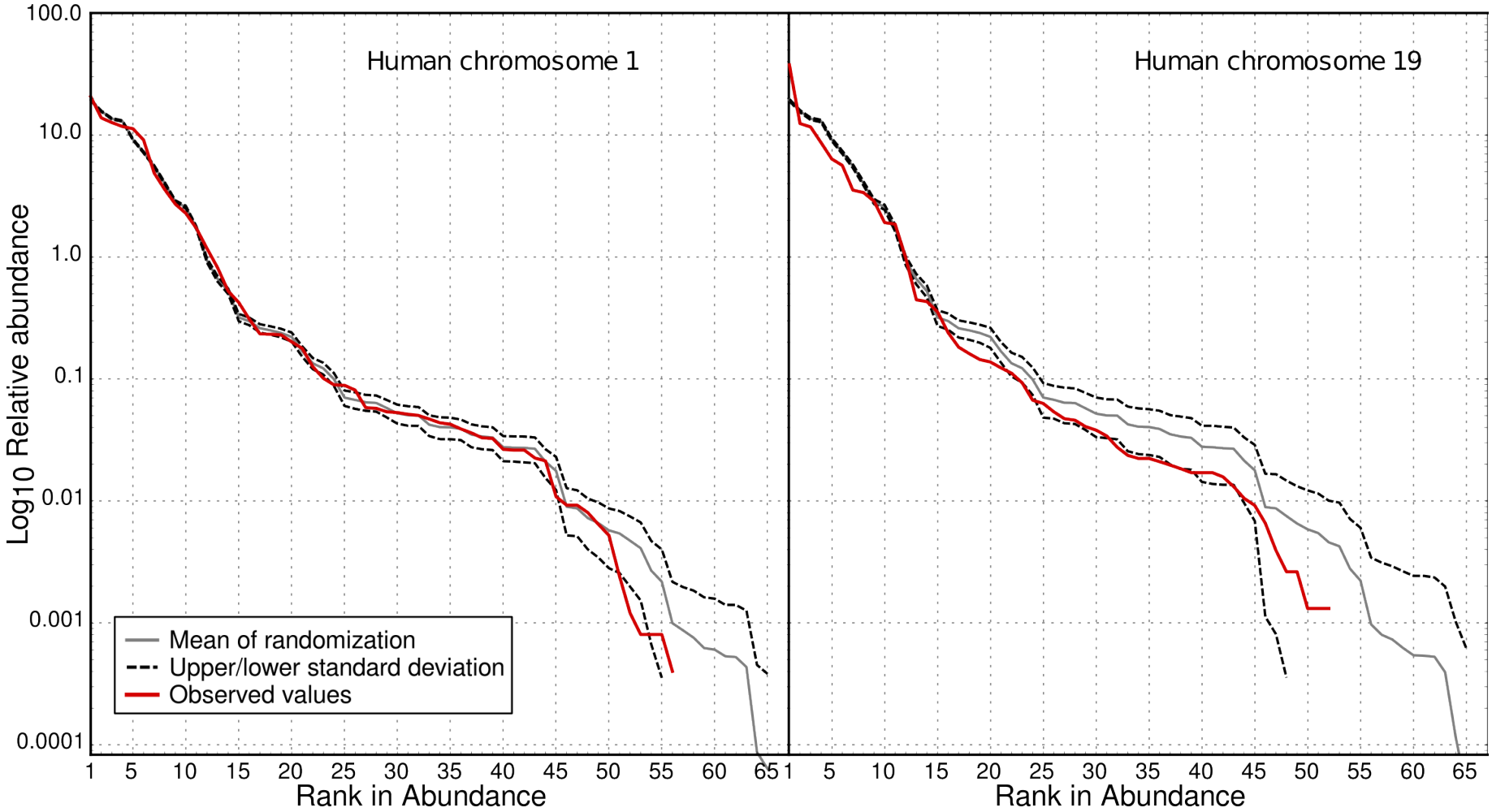
- Kolmogorov test red vs grey lines. 76 out of 548 chromosomes showing different distributions → **86% of chromosomes fitting**

Relative species abundance: observed vs expected



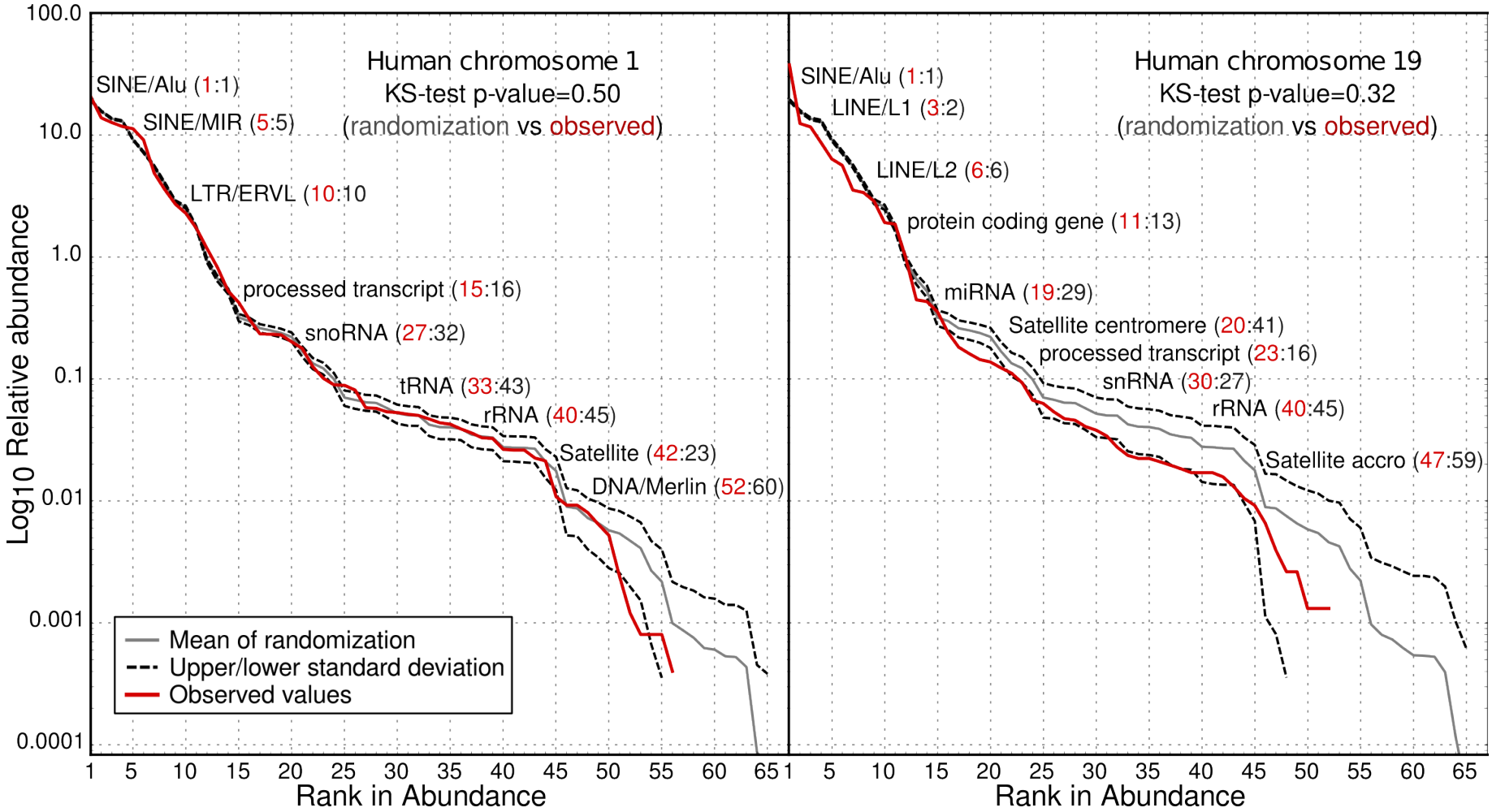
- Kolmogorov test red vs grey lines. 76 out of 548 chromosomes showing different distributions → **86% of chromosomes fitting**

Relative species abundance: observed vs expected



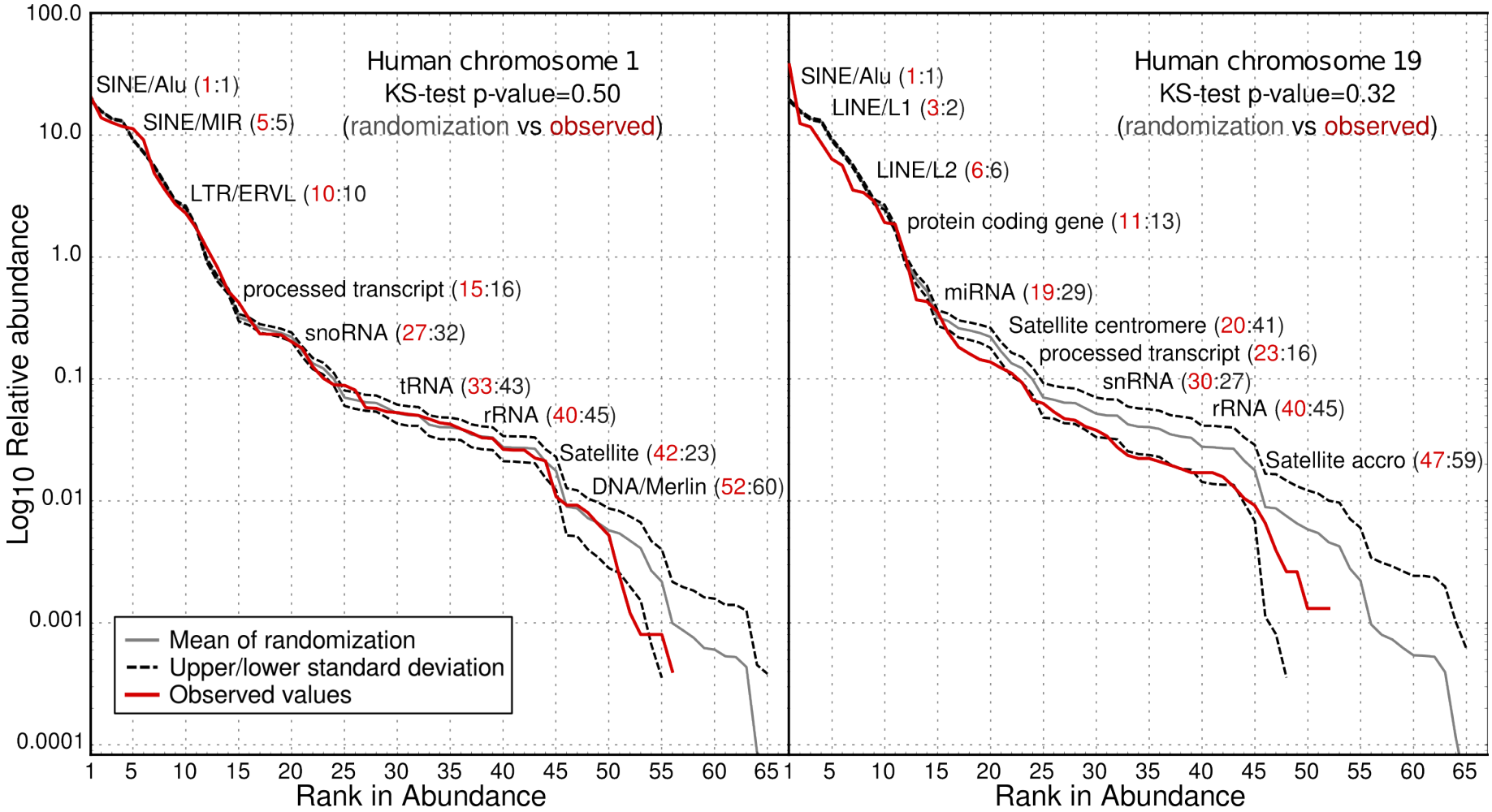
- Kolmogorov test red vs grey lines. 76 out of 548 chromosomes showing different distributions → **86% of chromosomes fitting**

Relative species abundance: observed vs expected



- Kolmogorov test red vs grey lines. 76 out of 548 chromosomes showing different distributions → **86% of chromosomes fitting**

Relative species abundance: observed vs expected



- Kolmogorov test red vs grey lines. 76 out of 548 chromosomes showing different distributions $\rightarrow$ **86% of chromosomes fitting**

Neutrality of GSs distribution in chromosomes

Unified Neutral Theory of Biodiversity and Biogeography for genetic species

- We test the fit of GSs' distributions in the UNTB

Species	Chr	*J*	*S*	θ	*m*	P-val
T. castaneum	7	7,865	18	2.12	—	**0.01**
A. gambiae	X	21,215	42	6.97	0.037	**0.03**
G. gallus	9	6,621	32	4.28	—	**0.05**
D. melanogaster	X	20,787	26	2.86	—	**0.09**
T. nigroviridis	3	8,505	17	1.96	—	**0.11**
M. musculus	14	143,018	59	6.58	0.149	**0.15**
P. trichocarpa	2	32,946	15	2.23	0.009	**0.16**
O. latipes	19	7,223	21	2.57	—	**0.17**
H. sapiens	17	93,105	52	6.51	0.065	**0.18**
M. mulatta	16	84,626	50	5.95	0.119	**0.23**
S. cerevisiae	II	640	9	1.37	—	**0.26**
D. discoideum	1	26,650	14	1.36	—	**0.27**
D. rerio	1	105,305	56	8.17	0.016	**0.29**
C. familiaris	1	144,103	47	5.28	0.093	**0.32**
P. falciparum	13	18,738	10	0.95	—	**0.38**
M. domestica	2	675,788	44	4.46	0.031	**0.43**
S. bicolor	1	37,626	23	2.86	0.067	**0.68**

- Neutrality can not be rejected for any of the chromosome studied

Neutrality of GSs distribution in chromosomes

Unified Neutral Theory of Biodiversity and Biogeography for genetic species

- We test the fit of GSs' distributions in the UNTB

Species	Chr	J	S	θ	m	P-val
T. castaneum	7	7,865	18	2.12	—	**0.01**
A. gambiae	X	21,215	42	6.97	0.037	**0.03**
G. gallus	9	6,621	32	4.28	—	**0.05**
D. melanogaster	X	20,787	26	2.86	—	**0.09**
T. nigroviridis	3	8,505	17	1.96	—	**0.11**
M. musculus	14	143,018	59	6.58	0.149	**0.15**
P. trichocarpa	2	32,946	15	2.23	0.009	**0.16**
O. latipes	19	7,223	21	2.57	—	**0.17**
H. sapiens	17	93,105	52	6.51	0.065	**0.18**
M. mulatta	16	84,626	50	5.95	0.119	**0.23**
S. cerevisiae	II	640	9	1.37	—	**0.26**
D. discoideum	1	26,650	14	1.36	—	**0.27**
D. rerio	1	105,305	56	8.17	0.016	**0.29**
C. familiaris	1	144,103	47	5.28	0.093	**0.32**
P. falciparum	13	18,738	10	0.95	—	**0.38**
M. domestica	2	675,788	44	4.46	0.031	**0.43**
S. bicolor	1	37,626	23	2.86	0.067	**0.68**

- Neutrality can not be rejected for any of the chromosome studied

Neutrality of GSs distribution in chromosomes

Unified Neutral Theory of Biodiversity and Biogeography for genetic species

- We test the fit of GSs' distributions in the UNTB

Species	Chr	*J*	*S*	θ	*m*	P-val
T. castaneum	7	7,865	18	2.12	—	**0.01**
A. gambiae	X	21,215	42	6.97	0.037	**0.03**
G. gallus	9	6,621	32	4.28	—	**0.05**
D. melanogaster	X	20,787	26	2.86	—	**0.09**
T. nigroviridis	3	8,505	17	1.96	—	**0.11**
M. musculus	14	143,018	59	6.58	0.149	**0.15**
P. trichocarpa	2	32,946	15	2.23	0.009	**0.16**
O. latipes	19	7,223	21	2.57	—	**0.17**
H. sapiens	17	93,105	52	6.51	0.065	**0.18**
M. mulatta	16	84,626	50	5.95	0.119	**0.23**
S. cerevisiae	II	640	9	1.37	—	**0.26**
D. discoideum	1	26,650	14	1.36	—	**0.27**
D. rerio	1	105,305	56	8.17	0.016	**0.29**
C. familiaris	1	144,103	47	5.28	0.093	**0.32**
P. falciparum	13	18,738	10	0.95	—	**0.38**
M. domestica	2	675,788	44	4.46	0.031	**0.43**
S. bicolor	1	37,626	23	2.86	0.067	**0.68**

- Neutrality can not be rejected for any of the chromosome studied

Neutrality of GSs distribution in chromosomes

Unified Neutral Theory of Biodiversity and Biogeography for genetic species

- We test the fit of GSs' distributions in the UNTB

Species	Chr	*J*	*S*	θ	*m*	P-val
T. castaneum	7	7,865	18	2.12	—	0.01
A. gambiae	X	21,215	42	6.97	0.037	0.03
G. gallus	9	6,621	32	4.28	—	0.05
D. melanogaster	X	20,787	26	2.86	—	0.09
T. nigroviridis	3	8,505	17	1.96	—	0.11
M. musculus	14	143,018	59	6.58	0.149	0.15
P. trichocarpa	2	32,946	15	2.23	0.009	0.16
O. latipes	19	7,223	21	2.57	—	0.17
H. sapiens	17	93,105	52	6.51	0.065	0.18
M. mulatta	16	84,626	50	5.95	0.119	0.23
S. cerevisiae	II	640	9	1.37	—	0.26
D. discoideum	1	26,650	14	1.36	—	0.27
D. rerio	1	105,305	56	8.17	0.016	0.29
C. familiaris	1	144,103	47	5.28	0.093	0.32
P. falciparum	13	18,738	10	0.95	—	0.38
M. domestica	2	675,788	44	4.46	0.031	0.43
S. bicolor	1	37,626	23	2.86	0.067	0.68

- Neutrality can not be rejected for any of the chromosome studied

Neutrality of GSs distribution in chromosomes

Unified Neutral Theory of Biodiversity and Biogeography for genetic species

- We test the fit of GSs' distributions in the UNTB

Species	Chr	*J*	*S*	θ	*m*	P-val
T. castaneum	7	7,865	18	2.12	—	**0.01**
A. gambiae	X	21,215	42	6.97	0.037	**0.03**
G. gallus	9	6,621	32	4.28	—	**0.05**
D. melanogaster	X	20,787	26	2.86	—	**0.09**
T. nigroviridis	3	8,505	17	1.96	—	**0.11**
M. musculus	14	143,018	59	6.58	0.149	**0.15**
P. trichocarpa	2	32,946	15	2.23	0.009	**0.16**
O. latipes	19	7,223	21	2.57	—	**0.17**
H. sapiens	17	93,105	52	6.51	0.065	**0.18**
M. mulatta	16	84,626	50	5.95	0.119	**0.23**
S. cerevisiae	II	640	9	1.37	—	**0.26**
D. discoideum	1	26,650	14	1.36	—	**0.27**
D. rerio	1	105,305	56	8.17	0.016	**0.29**
C. familiaris	1	144,103	47	5.28	0.093	**0.32**
P. falciparum	13	18,738	10	0.95	—	**0.38**
M. domestica	2	675,788	44	4.46	0.031	**0.43**
S. bicolor	1	37,626	23	2.86	0.067	**0.68**

- Neutrality can not be rejected for any of the chromosome studied

Neutrality of GSs distribution in chromosomes

Unified Neutral Theory of Biodiversity and Biogeography for genetic species

- We test the fit of GSs' distributions in the UNTB

Species	Chr	*J*	*S*	θ	*m*	P-val
T. castaneum	7	7,865	18	2.12	—	**0.01**
A. gambiae	X	21,215	42	6.97	0.037	**0.03**
G. gallus	9	6,621	32	4.28	—	**0.05**
D. melanogaster	X	20,787	26	2.86	—	**0.09**
T. nigroviridis	3	8,505	17	1.96	—	**0.11**
M. musculus	14	143,018	59	6.58	0.149	**0.15**
P. trichocarpa	2	32,946	15	2.23	0.009	**0.16**
O. latipes	19	7,223	21	2.57	—	**0.17**
H. sapiens	17	93,105	52	6.51	0.065	**0.18**
M. mulatta	16	84,626	50	5.95	0.119	**0.23**
S. cerevisiae	II	640	9	1.37	—	**0.26**
D. discoideum	1	26,650	14	1.36	—	**0.27**
D. rerio	1	105,305	56	8.17	0.016	**0.29**
C. familiaris	1	144,103	47	5.28	0.093	**0.32**
P. falciparum	13	18,738	10	0.95	—	**0.38**
M. domestica	2	675,788	44	4.46	0.031	**0.43**
S. bicolor	1	37,626	23	2.86	0.067	**0.68**

- Neutrality can not be rejected for any of the chromosome studied

Index

The single gene approach

- Functional categories enriched in **PSGs** ($\omega = \frac{dN}{dS} > 1$)
- No set of PSGs were find to be significantly enriched

The single gene approach

- Functional categories enriched in **PSGs** ($\omega = \frac{dN}{dS} > 1$)
- No set of PSGs were find to be significantly enriched

The single gene approach

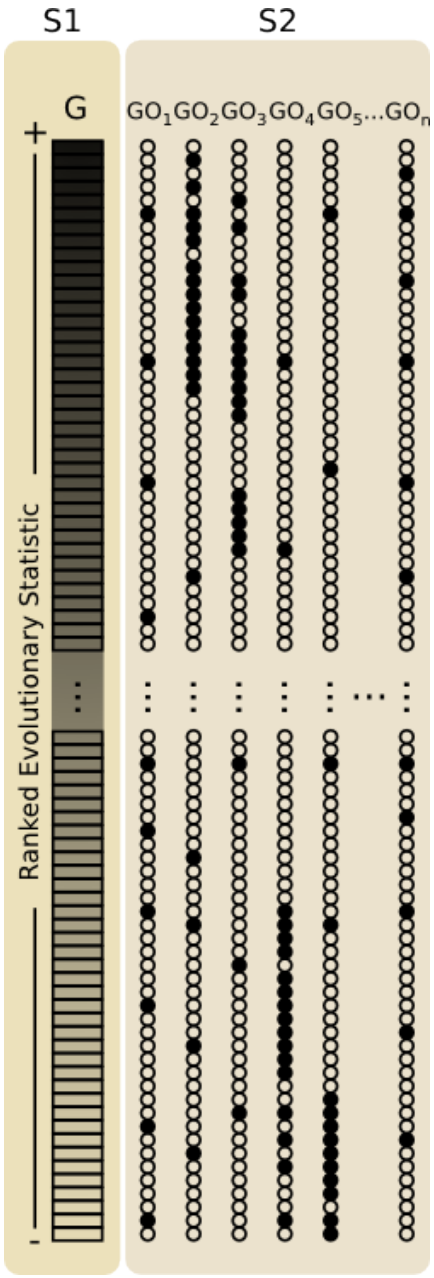
- Functional categories enriched in **PSGs** ($\omega = \frac{dN}{dS} > 1$)
- No set of PSGs were find to be significantly enriched

[Arbiza et al. - *PLoS Computational Biology (2006), Clark* et al. - *Science (2003), Nielsen* et al. - *PLoS Biology (2005)]*

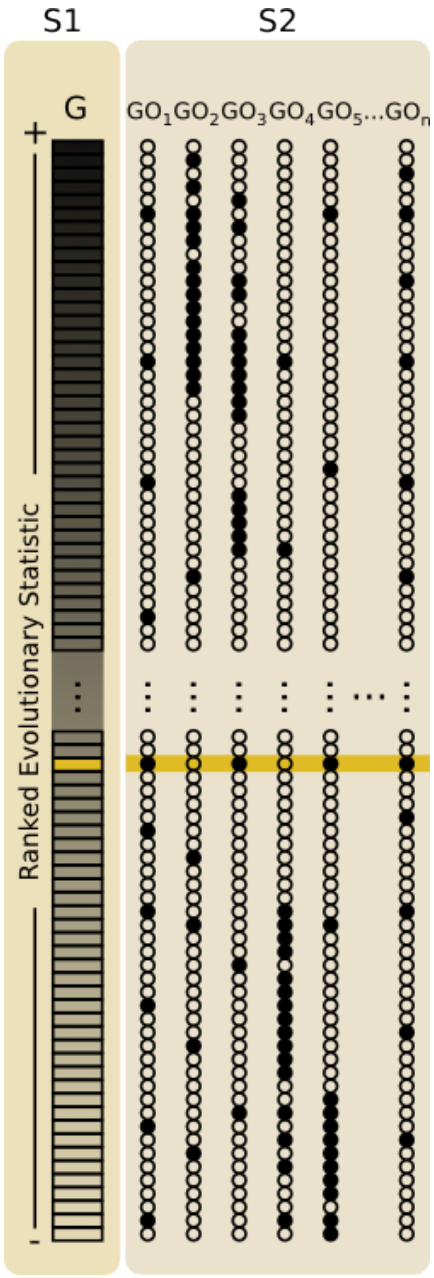
Index

Gene-Set Selection Analysis (GSSA)

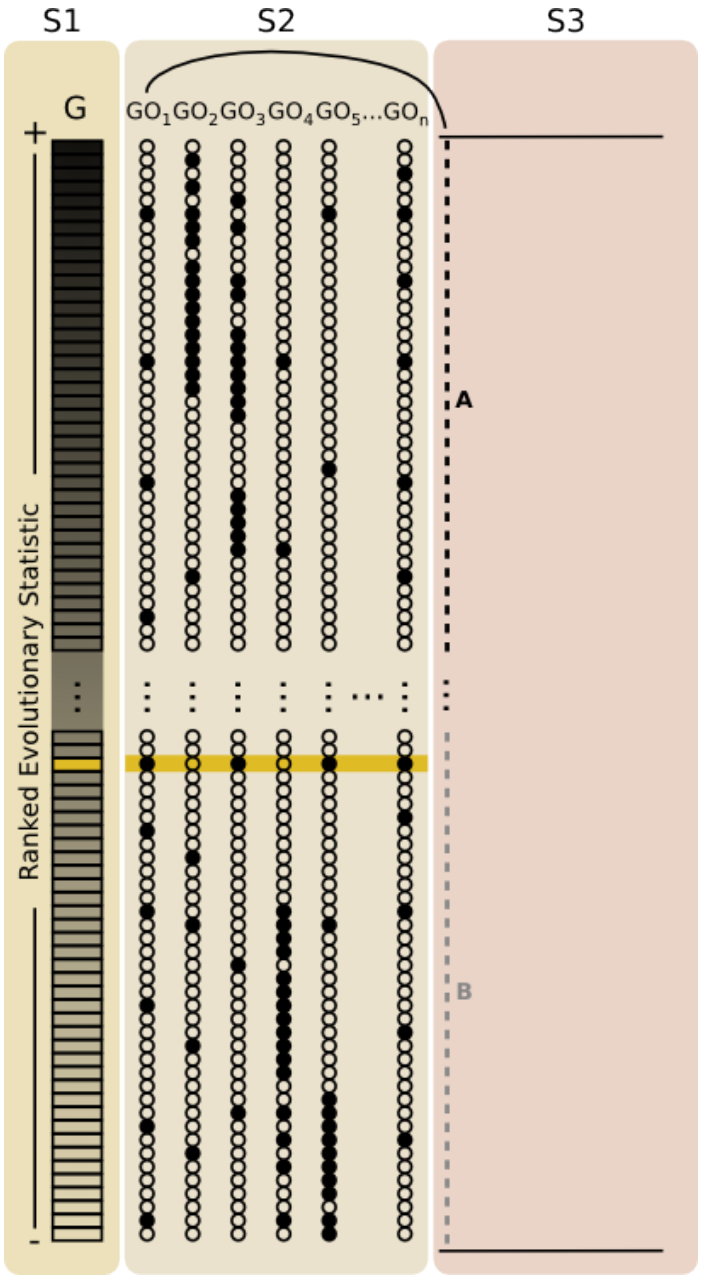
Gene-Set Selection Analysis (GSSA)



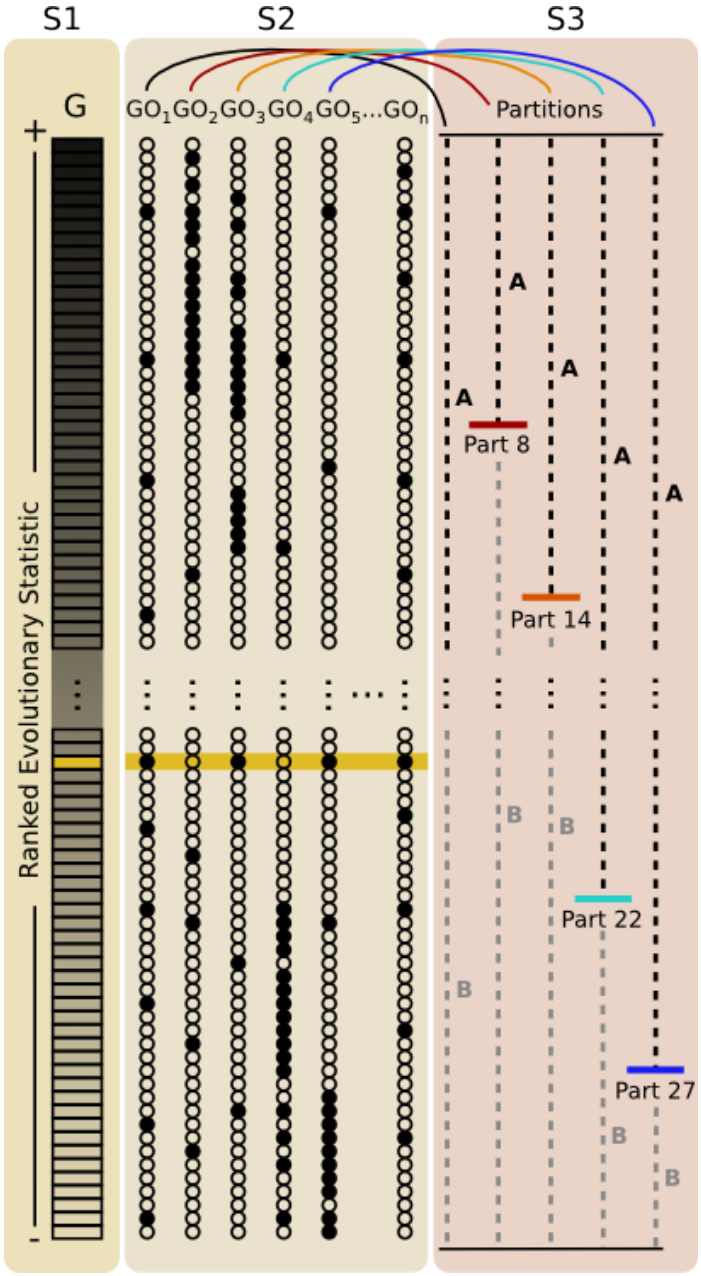
Gene-Set Selection Analysis (GSSA)



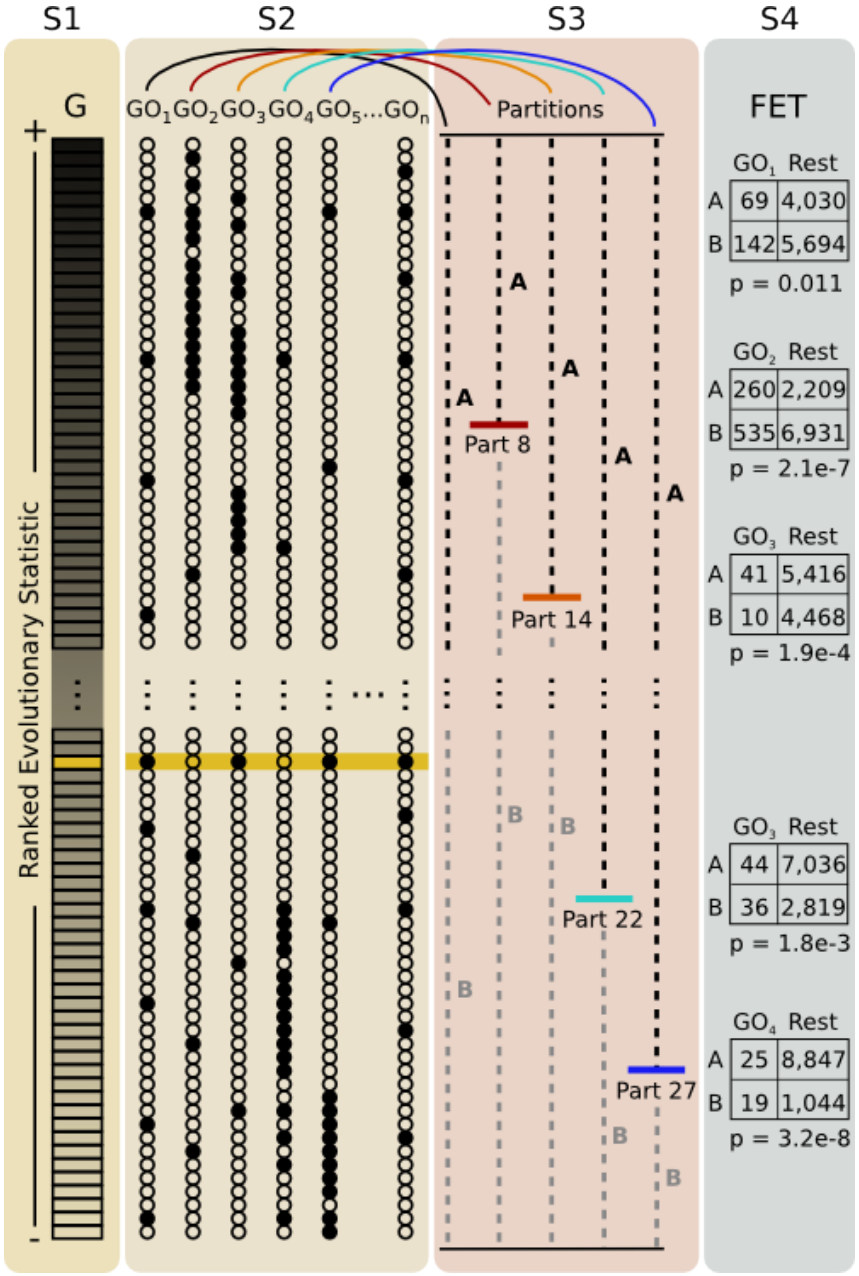
Gene-Set Selection Analysis (GSSA)



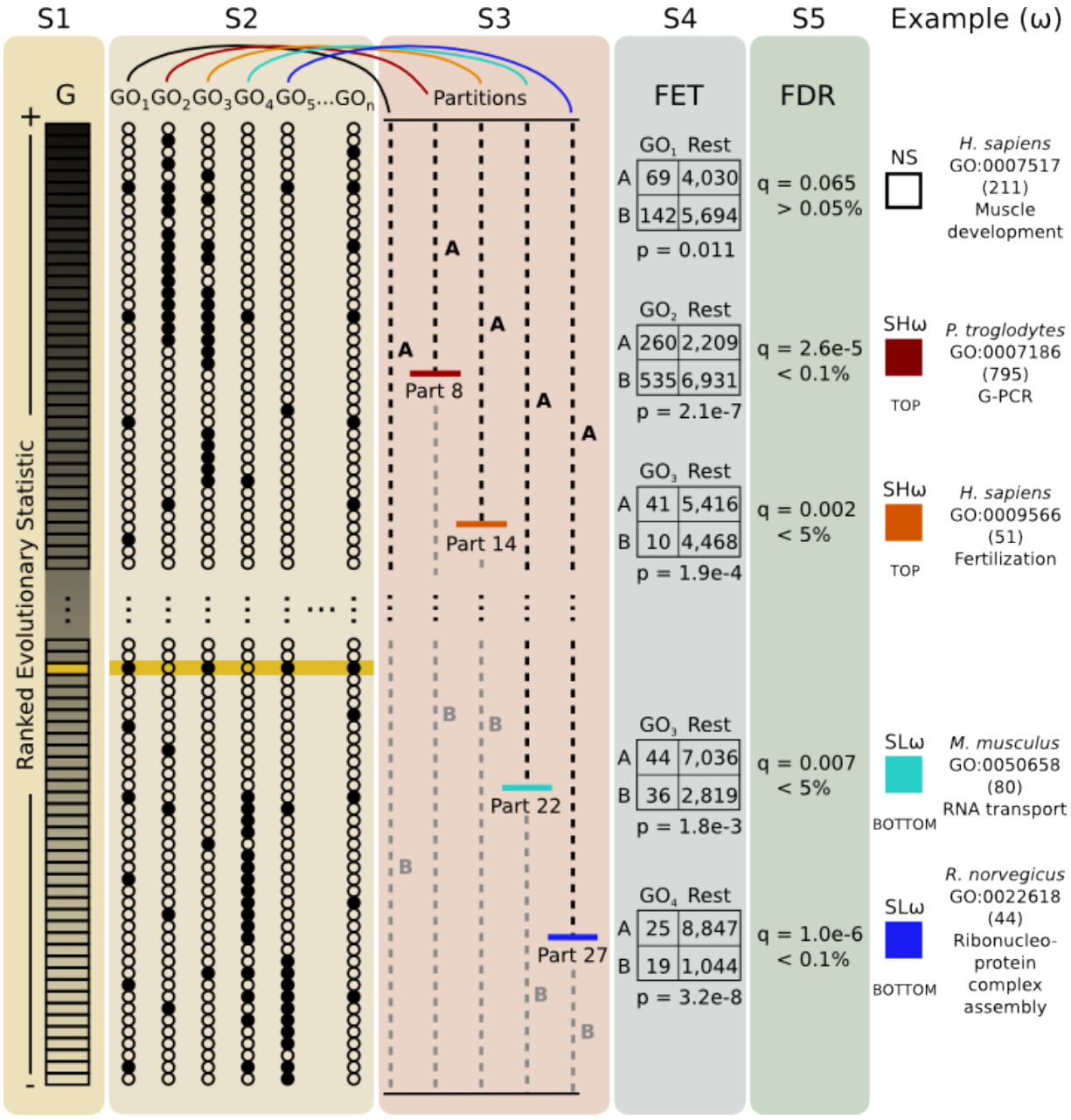
Gene-Set Selection Analysis (GSSA)



Gene-Set Selection Analysis (GSSA)

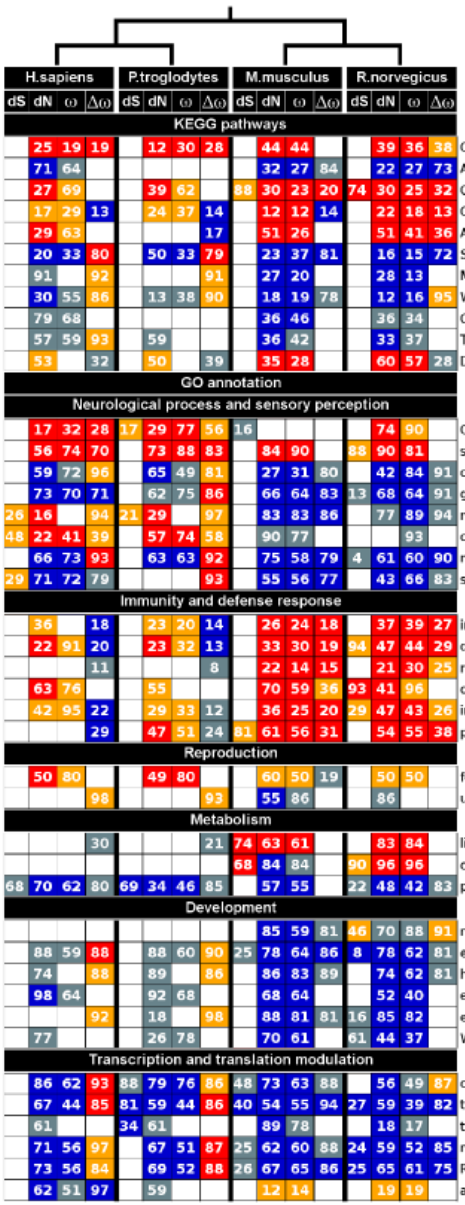


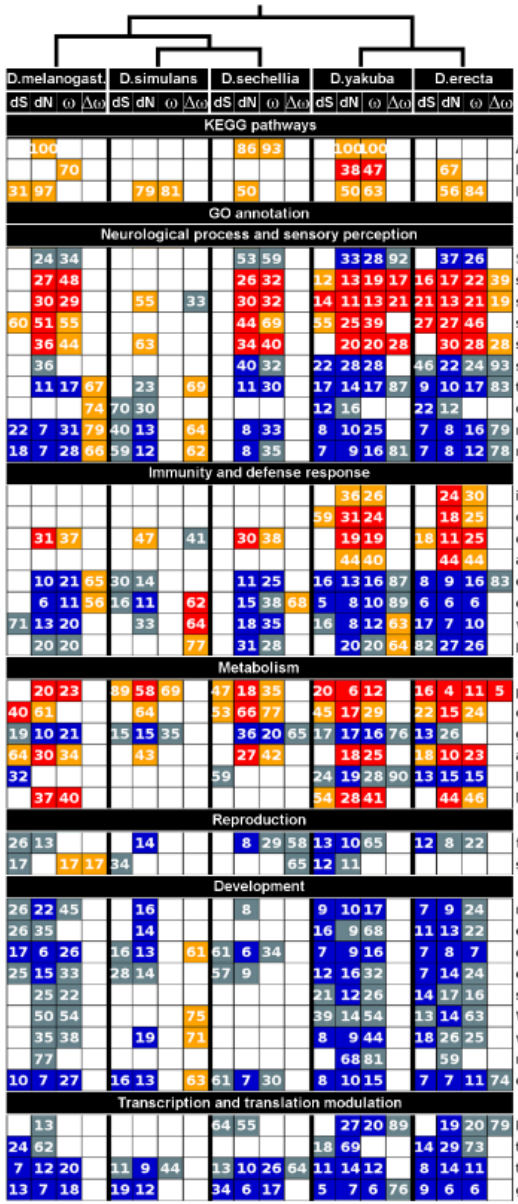
Gene-Set Selection Analysis (GSSA)



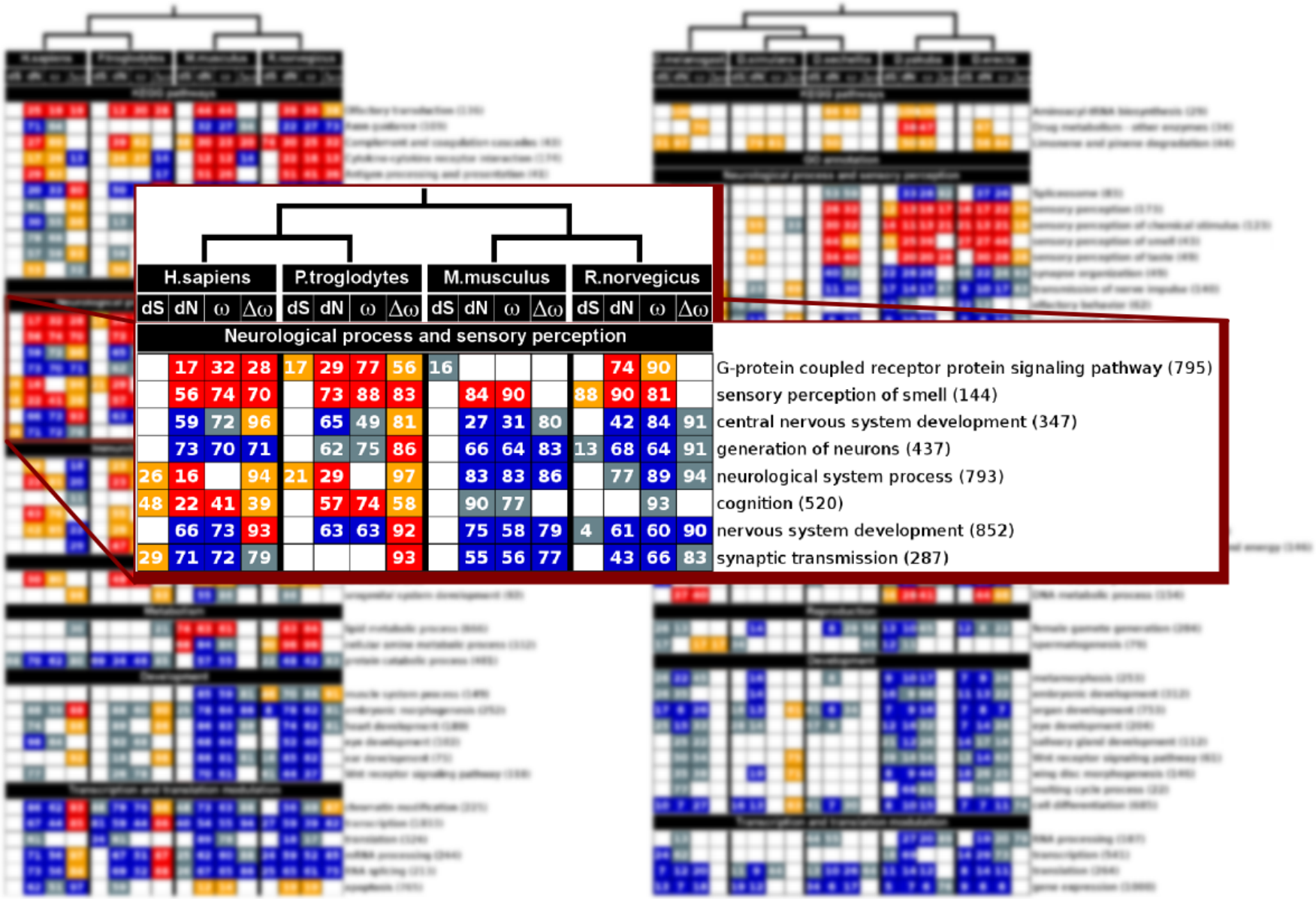
Index

GSSA – Summary results

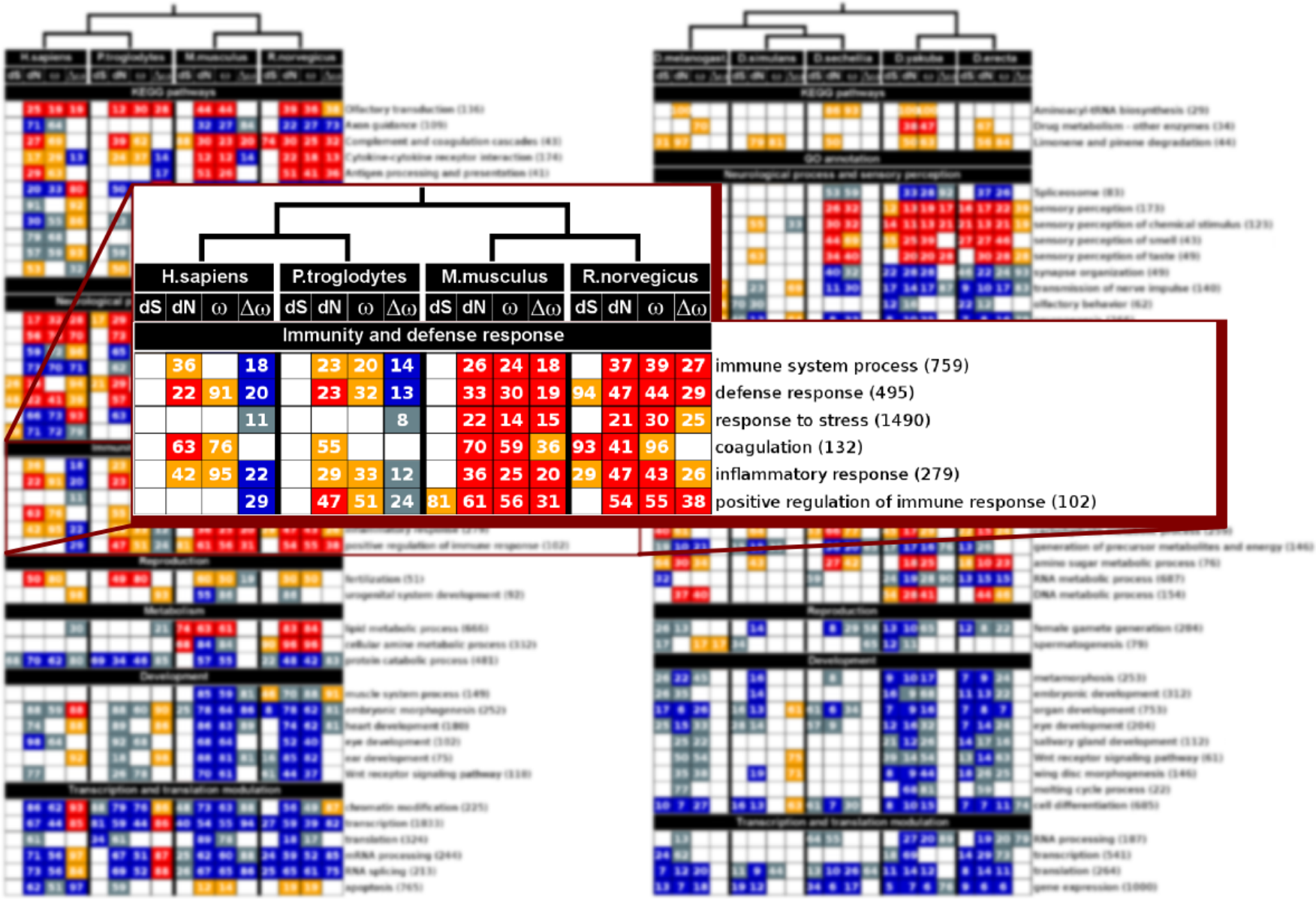




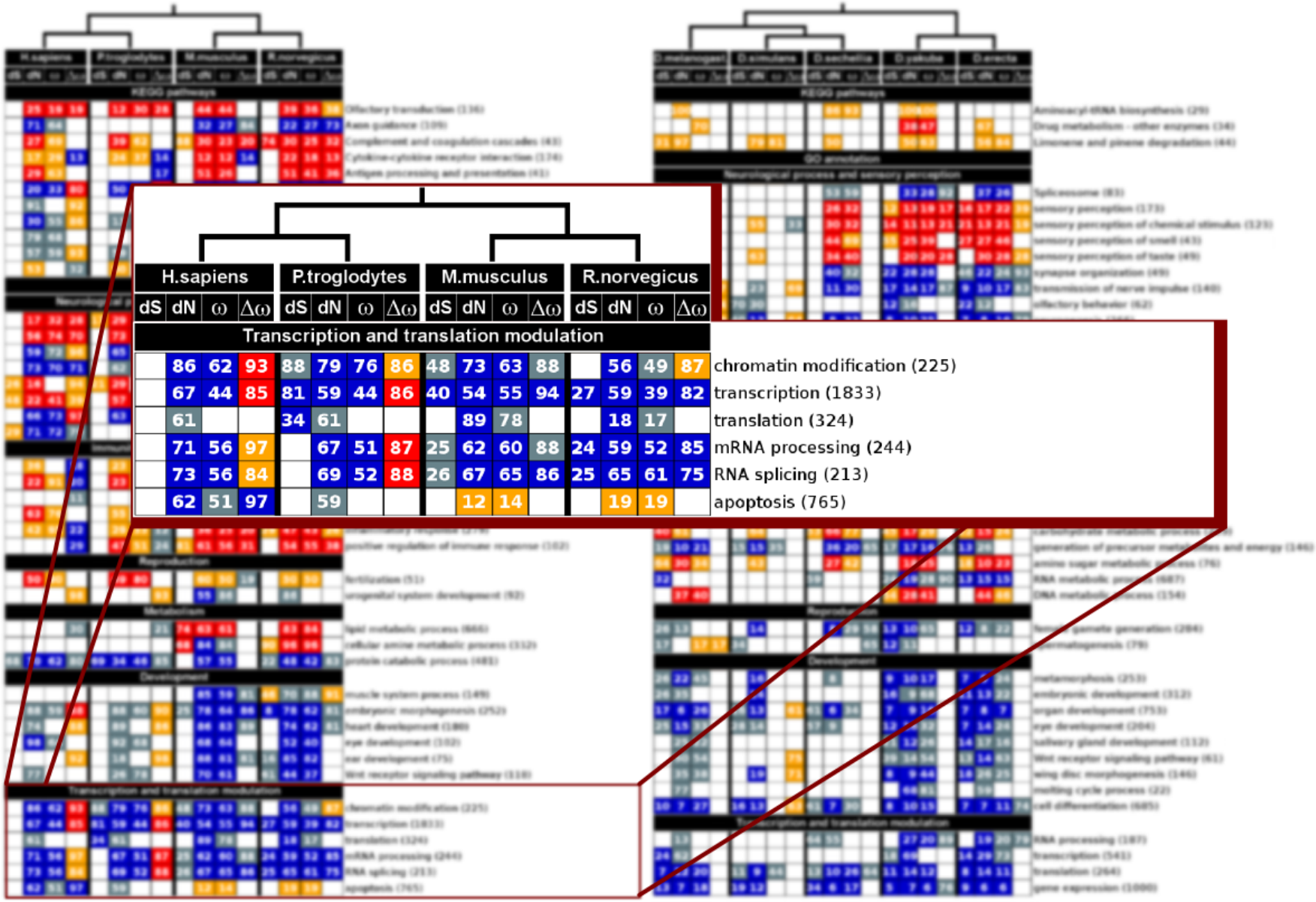
GSSA – Summary results



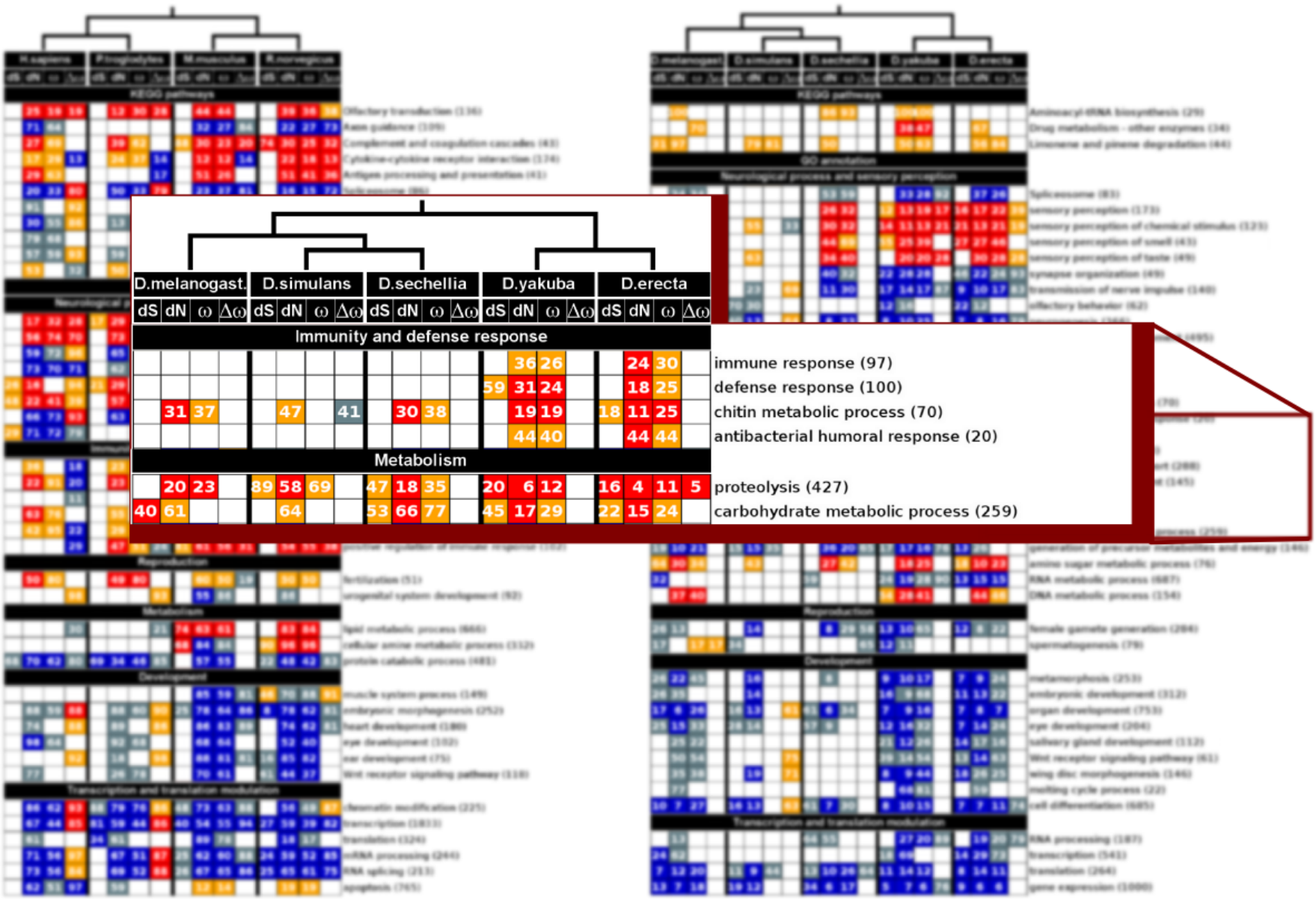
GSSA – Summary results



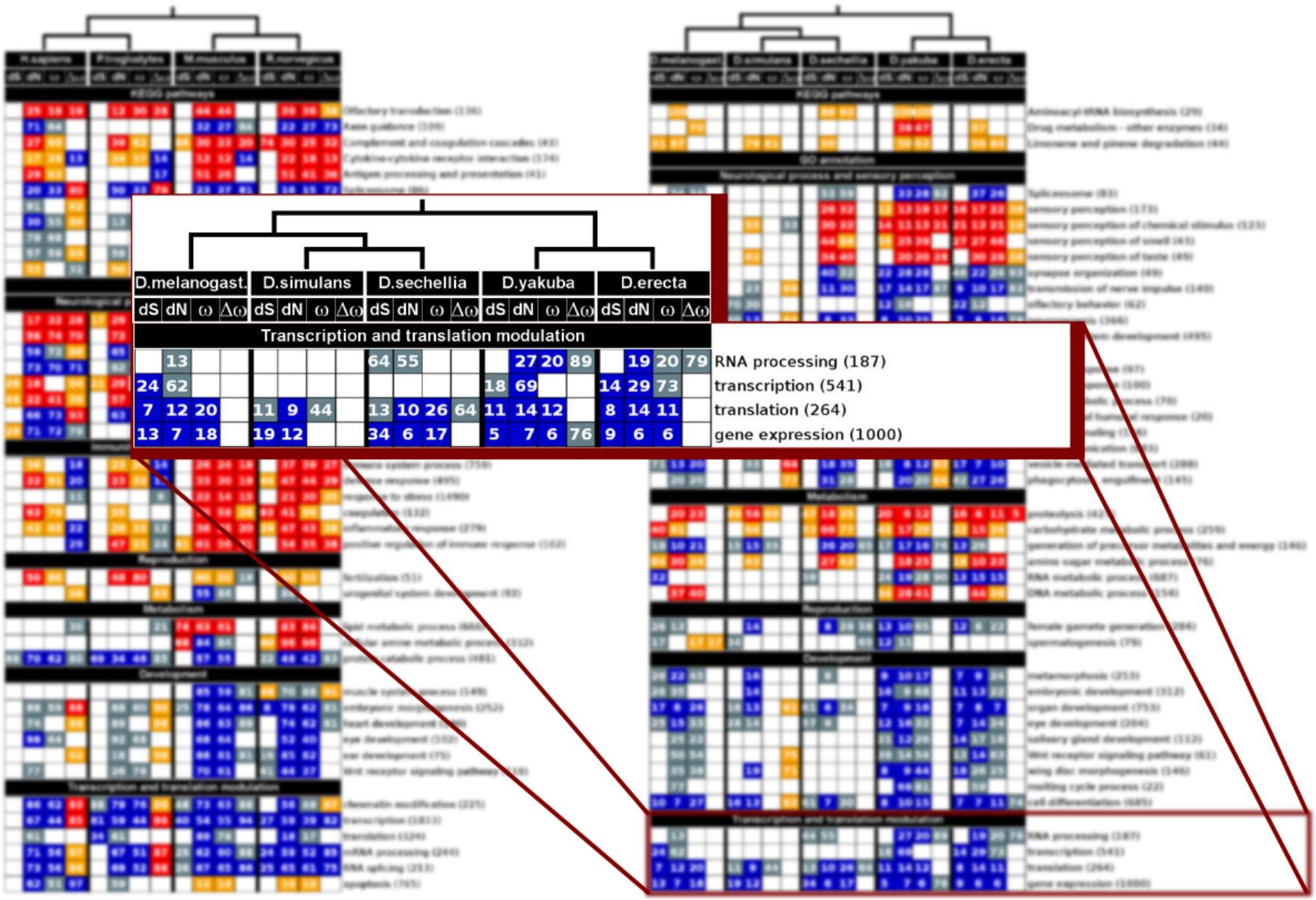
GSSA – Summary results



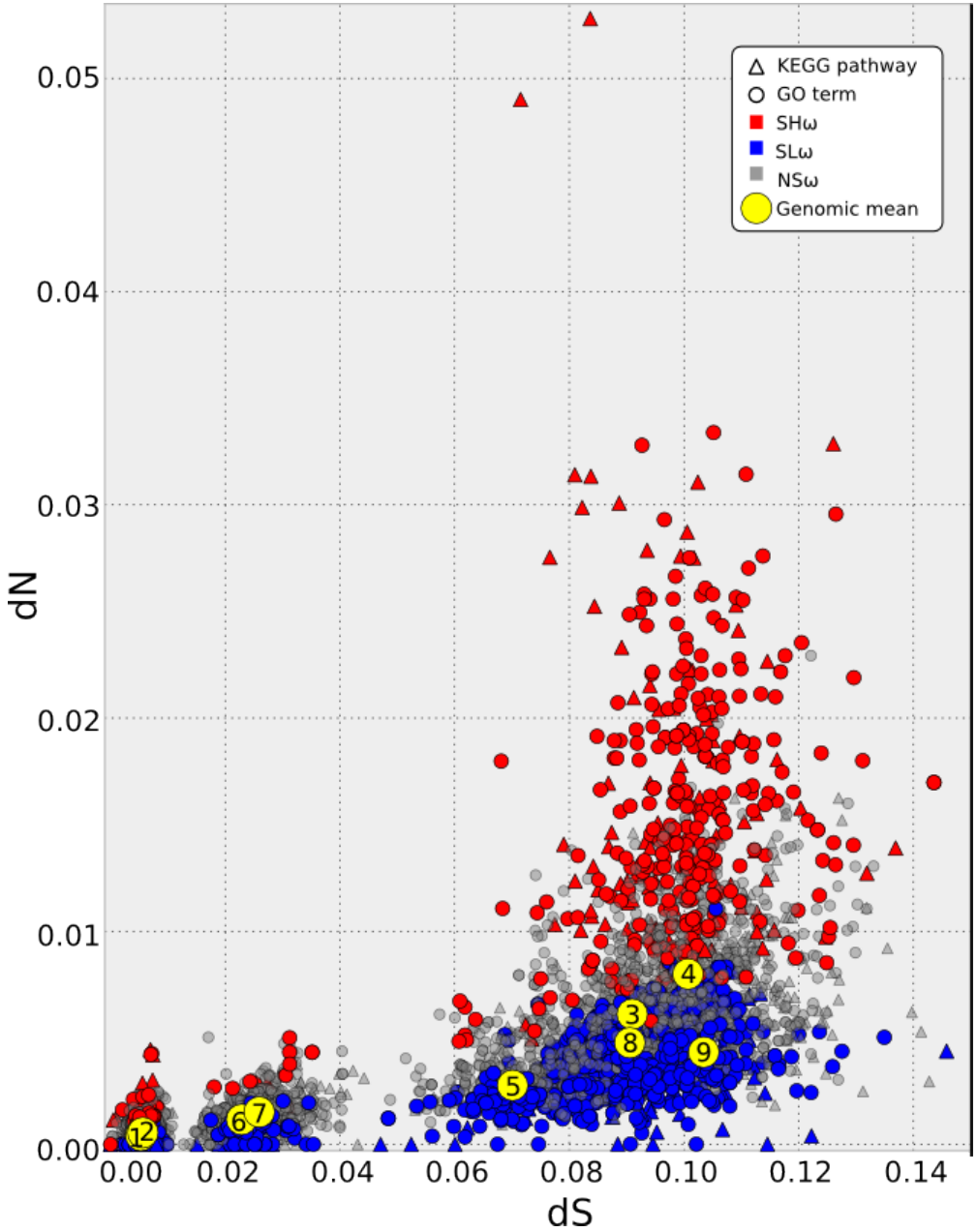
GSSA – Summary results



GSSA – Summary results

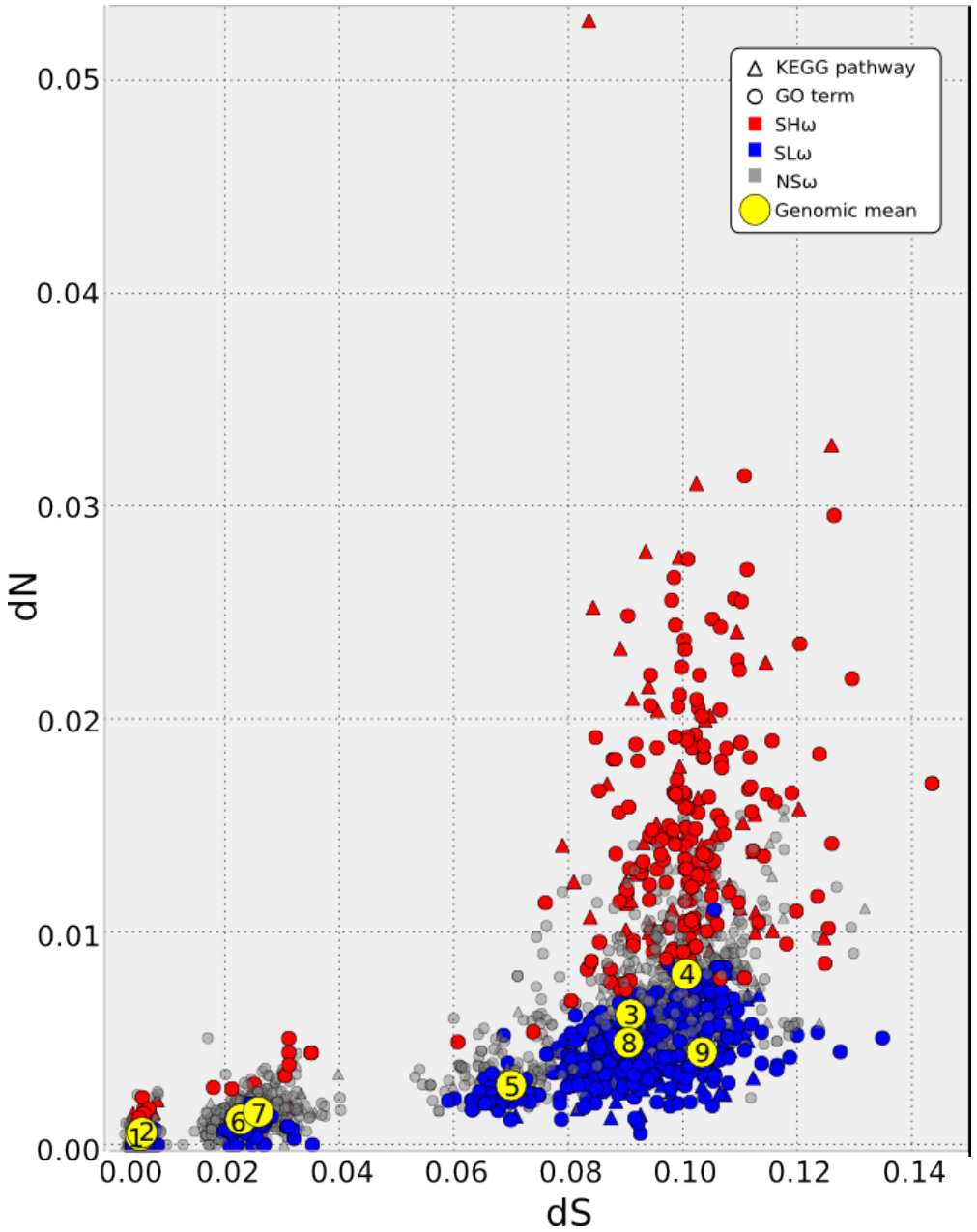


PSGS within GSSA results



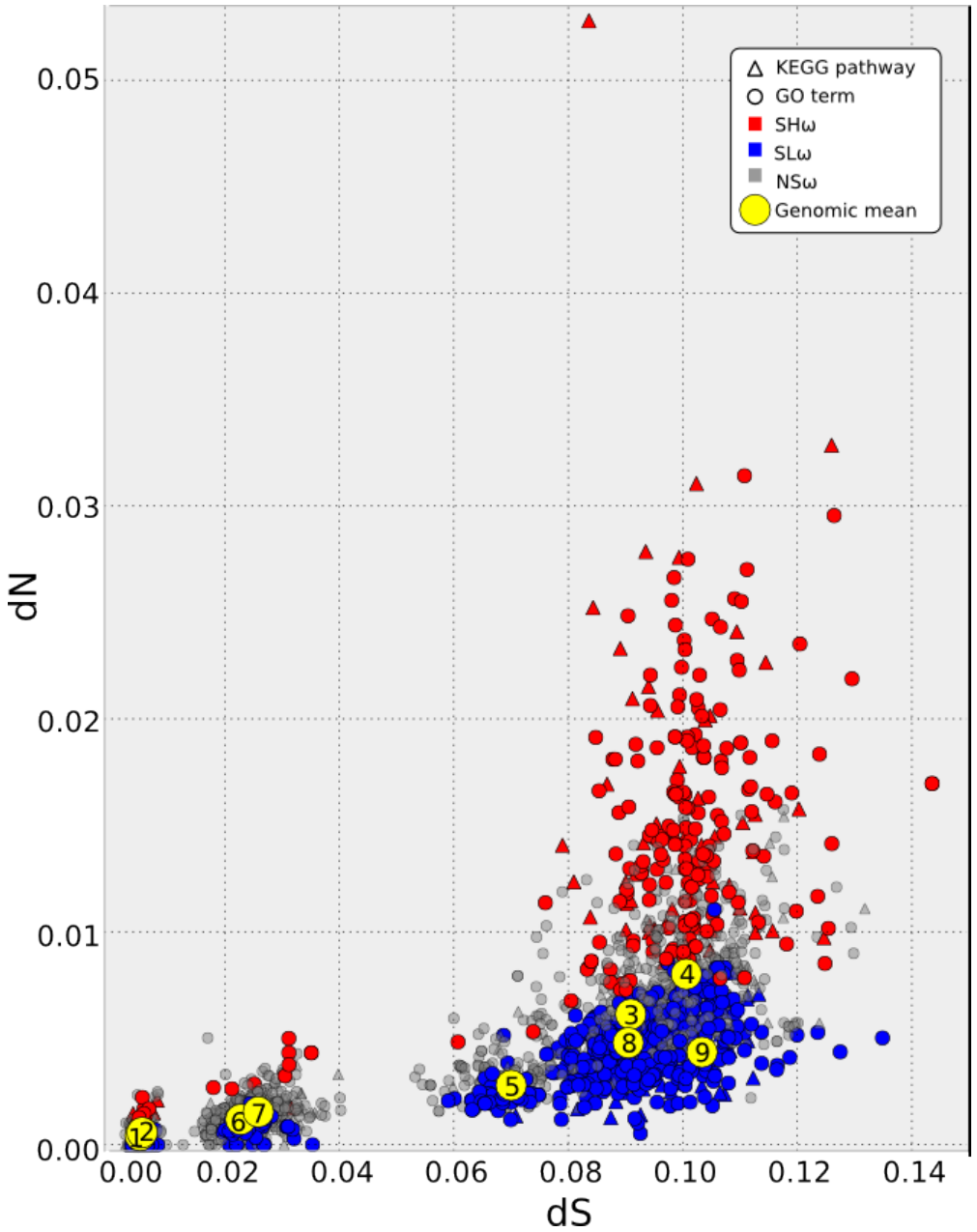
- PSGs are largely present in categories with SHω, SLω and NSω.
- PSGs where enriched in SHω categories
- However not significant for primates

PSGS within GSSA results



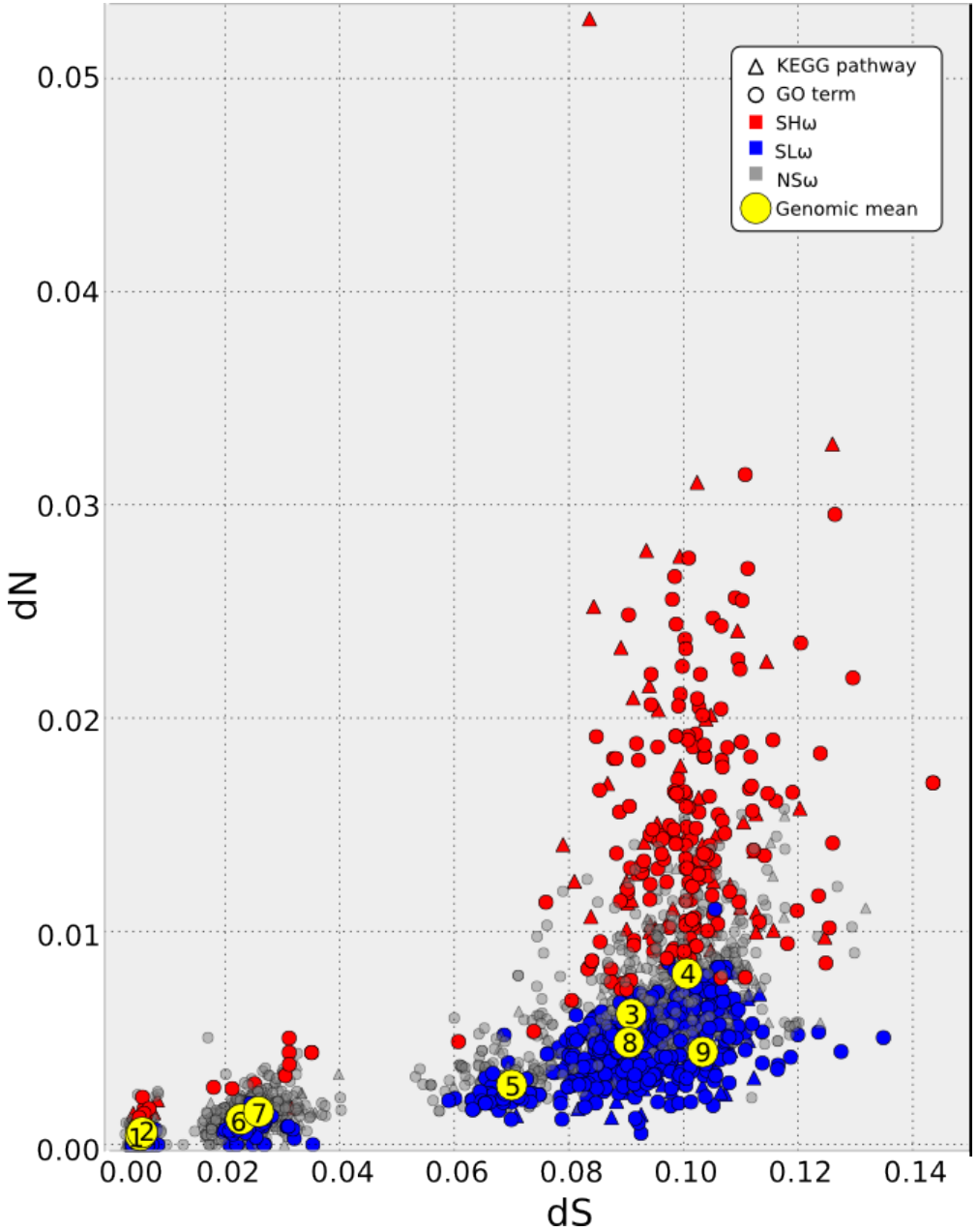
- PSGs are largely present in categories with SHω, SLω and NSω.
- PSGs where enriched in SHω categories
- However not significant for primates

PSGS within GSSA results



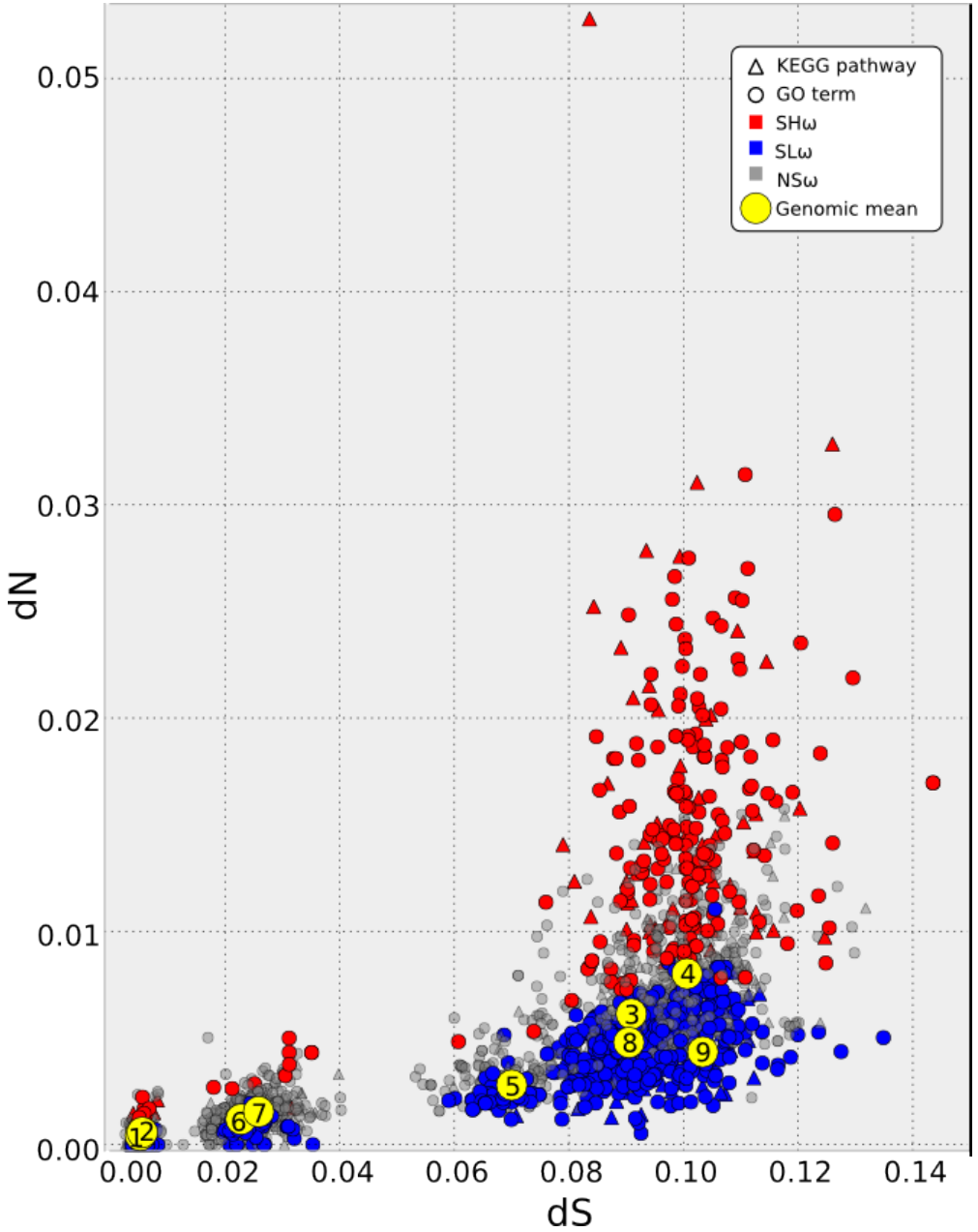
- PSGs are largely present in categories with SHω, SLω and NSω.
- PSGs where enriched in SHω categories
- However not significant for primates

PSGS within GSSA results



- PSGs are largely present in categories with SHω, SLω and NSω.
- PSGs where enriched in SHω categories
- However not significant for primates

PSGS within GSSA results



- PSGs are largely present in categories with SHω, SLω and NSω.
- PSGs where enriched in SHω categories
- However not significant for primates

Conclusions

1. In the whole diversity of life, from viruses to mammals, informational content of the genomes exhibits quasi-maximum values. Only dramatic changes such as polyploidization events, or strong biases in nucleotide contents, are able to lower genome entropy.

2. According to the observed universal adjustment of genomes to maximum complexity, we hypothesize that increases in biological complexity are the consequence of genome expansions events through duplications or polyploidization.

3. Similarly to the biological species in ecosystems, eukaryotic genomes present an heterogeneous distribution of families or “species” of genetic elements: a few are very abundant, others quite frequent, and the majority rare.

4. Likewise for ecological species-area correlation graph, we observe that, along a great diversity of eukaryote genomes, the total number of genetic species in chromosomes is proportional to chromosome length.

5. The distributions and abundances of families of genetic elements in eukaryotic genomes, either functional or repetitive, follow the expectations of the unified neutral theory of biodiversity (UNTB).

Conclusions

6. GSSA allows testing for functional biases within fast or slowly-evolving genes. This methodology successfully identified all previously reported candidate functional categories as important targets of natural selection. Moreover, given that the GSSA is not limited by the compulsory presence of positively-selected genes, it extended the list of phenotypical targets to previously undetectable ones.

7. Genes under positive selection were found to be present in functional categories evolving rapidly, slowly or without a significant trend (ω). However a significant bias towards fast evolving categories was found in rodents and *Drosophila*. Regarding to the even distribution found in primates, we hypothesize that it may be the result of small population sizes limiting the influence of natural selection; just as suggested by the theory of slightly deleterious mutations.

8. We believe that the role of genes under positive selection not only consists of the adaptive evolutionary changes of phenotypes but may also be related to other processes such as the adjustment to the deleterious mutations of other genes in a given network.

9. Throughout this thesis three primary bioinformatics tools were implemented with the idea of facilitating and extending future researches of the scientific community. These software are in line with the fields of ecology (of genomes), phylogeny, phylogenomics and the testing of evolutionary hypotheses.

Thank you :)

Bibliography I

thebibliography


LEONARDO ARBIZA, JOAQUÍN DOPAZO, AND HERNÁN DOPAZO, Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome.
PLoS Computational Biology **2**(4) (2006), p. e38.
77


ANDREW G CLARK, STEPHEN GLANOWSKI, RASMUS NIELSEN, PAUL D THOMAS, ANISH KEJARIWAL, MELISSA A TODD, DAVID M TANENBAUM, DANIEL CIVELLO, FU LU, BRIAN MURPHY, STEVE FERRIERA, GARY WANG, XIANQGUN ZHENG, THOMAS J WHITE, JOHN J SNINSKY, ET AL., Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios.
Science **302**(5652) (2003), 1960–3.
77

Bibliography II


T Ryan Gregory, Synergy between sequence and size in large-scale genomics.
Nature Reviews. Genetics **6**(9) (2005), 699–708.
27, 28


Stephen P Hubbell, *The Unified Neutral Theory of Biodiversity and Biogeography*.
Princeton University Press, 2001.
15, 16


Eric S Lander, L M Linton, Bruce W Birren, C Nusbaum, Michael C Zody, Jennifer Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, et al., Initial sequencing and analysis of the human genome.
Nature **409**(6822) (2001), 860–921.
52

Bibliography III

D. R. Maddison and K.-S. Schulz, The Tree of Life Web Project, http://tolweb.org, 2007.
24, 25, 26

Anne E Magurran, *Ecological diversity and its measurement*.
Cambridge University Press, 1988.
14

Rasmus Nielsen, Carlos Bustamante, Andrew G Clark, Stephen Glanowski, Timothy B Sackton, Melissa J Hubisz, Adi Fledel-Alon, David M Tanenbaum, Daniel Civello, Thomas J White, John J Sninsky, Mark D Adams, and Michele Cargill, A scan for positively selected genes in the genomes of humans and chimpanzees.
PLoS Biology **3**(6) (2005), p. e170.
77

Bibliography IV


ELLEN J PRITHAM, Transposable elements and factors influencing their success in eukaryotes.
The Journal of Heredity **100**(5) (2009), 648–55.
53


JAMES ROSINDELL, STEPHEN P HUBBELL, AND RAMPAL S ETIENNE, The unified neutral theory of biodiversity and biogeography at age ten.
Trends in Ecology & Evolution **26**(7) (2011), 340–8.
17