0.1 Measuring DNA complexity

0.1.1 The complexity ratio and complexity value

Given a sequence, $seq = AACCTTCGTAGCATGG$:

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

Char. list				MTF
G	a	t	c	0
G	**A**	t	c	1
A	g	**T**	c	2
T	a	g	**C**	3
C	t	a	**G**	3
G	c	t	**A**	3
A	g	c	**T**	3
T	a	g	**C**	3
C	t	a	**G**	3
G	c	t	**A**	3
A	g	c	**T**	3
T	a	g	**C**	3
C	t	a	**G**	3
G	c	**T**	a	2
T	g	c	**A**	3
A	t	g	**C**	3

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

Table 0.1: CR explained by example.

These three tables summarizes the steps needed to obtain the final sequence of number from which we will finally compute Shannon's entropy. 1) The table on the left corresponds to the BWT. Original sequence is rotated sequentially (first character moved to back) resulting in different strings, as many as characters in the sequence. The resulting sequences are then sorted in lexicographic order. The "*I.*" column corresponds to the Index of this ordering (e.g.: the third sequence here in original order "#" takes the fifth position in lexicographic order). 2) The table in the center corresponds to the result of the BWT, that is the last character of previous sequences ordered as explain. 3) The table on the right corresponds to the application of the MTF algorithm. Starting from a sequence of all character named here "Char. list" (our four nucleotides in this case), the MTF will get the index of the current nucleotide from the BWT (upper case bold letter) in the "Char. list". In a second step, for the next iteration, MTF will transform the "Char. list" bringing to front the character present in previous BWT result (upper case letter).

Finally, we compute the Shannon's entropy of the values obtained by the MTF that results in our CR (the CV is obtained by multiplying CR by the length of the sequence).

i

i

S

N_i

i

S

$length(s)$

S

$$p_{(i)} = \frac{N_i}{length(s)} \tag{0.1}$$

$$E(s) = -\sum_{i=0}^{3} p_{(i)} \times log_4(p_{(i)}) \tag{0.2}$$

i

$$CR(s) = E(MTF(BWT(s))) \tag{0.3}$$

S

$$CV(s) = E(MTF(BWT(s))) \times length(s) \tag{0.4}$$

0.1.2 Complexity in strings

Genomic sequences

Annotation of repetitive elements

Human texts

Complexity in windows

0.1.3 Simulations

$1\times$

$1.0e^{-08}$

$2.5e^{-08}$

$7.1e^{-09}$

0.2 Measuring dynamics of genetic species

0.2.1 Genomes

0.2.2 Mining of Genetic Elements

Repetitive Elements

Functional Elements

0.2.3 Randomization of genetic elements

Chr. name	Original length	Corrected length	Percentage le
1	249,240,621	225,200,000	90.35%
2	243,188,741	237,670,000	97.73%
3	197,961,181	194,230,000	98.12%
4	191,044,271	187,270,000	98.02%
5	180,901,928	177,090,000	97.89%
6	171,048,878	167,050,000	97.66%
7	159,128,663	154,640,000	97.18%
8	146,302,151	142,290,000	97.26%
9	141,151,937	120,130,000	85.11%
10	135,524,747	131,040,000	96.69%
11	134,946,516	130,310,000	96.56%
12	133,841,891	129,970,000	97.11%
13	115,109,733	95,500,000	82.96%
14	107,289,415	87,910,000	81.94%
15	102,521,389	81,520,000	79.52%
16	90,290,985	78,640,000	87.10%
17	81,195,208	77,700,000	95.70%
18	78,017,245	74,580,000	95.59%
19	59,118,983	55,460,000	93.81%
20	62,962,324	59,430,000	94.39%
21	48,119,895	35,100,000	72.94%
22	51,244,541	34,790,000	67.89%
X	155,260,558	150,230,000	96.76%
Y	59,033,288	22,520,000	38.15%

Table 0.2: Transformation of Chromosome size. Example of changes in estimation of chromosome length after removing regions with no GEs for Human chromosome 1.

0.2.4 Ecolopy

$K(D, A)$

0.2.5 Neutral Ecological models

Ewens sampling formula

θ

ν

$$\theta = 2J_M\nu \tag{0.5}$$

θ

$$Pr\{S, n1, n2, \ldots, n_S|\theta\} = \frac{J_M!\theta^S}{1^{\phi 1}2^{\phi 2}\ldots J_M^{\phi J_M}\phi_1!\phi_2!\cdots\phi_{J_M}!\prod_{k=1}^{J_M}(\theta+k-1)} \tag{0.6}$$

n_i

i

a

$$\mathcal{L} = \frac{\theta^S}{\prod_{k=1}^{J_M}(\theta + k - 1)} \tag{0.7}$$

θ

θ

Etienne sampling formula

$m = 1$

$m < 1$

I

$$m = \frac{I}{I + J - 1} \tag{0.8}$$

$$P[D|\theta, m, J] = \frac{J!}{\prod_{i=1}^{S} n_i \prod_{j=1}^{J} \phi_J!} \frac{\theta^S}{(I)_J} \sum_{A=S}^{J} K(D, A) \frac{I^A}{(\theta)_A} \tag{0.9}$$

$K(D, A)$

$$K(D,A) := \sum_{\{a_1,\ldots,a_s \mid \sum_{i=1}^{S} a_i = A\}} \prod_{i=1}^{S} \frac{\bar{s}(n_i,a_i)\bar{s}(a_i,1)}{\bar{s}(n_i,1)} \tag{0.10}$$

$$S_{(n,m)} = S_{(n-1,m-1)} - (n-1) \times S_{(n-1,m)} \tag{0.11}$$

$K(D, A)$

θ

m

$$P[D|\theta, m, J] = \frac{J!}{\prod_{i=1}^{S} n_i \prod_{J=1}^{J} \Phi_J!} \frac{\theta^S}{(\theta)_J} \times \sum_{A=S}^{J} \left(K(D, A) \frac{(\theta)_J}{(\theta)_A} \frac{I^A}{(I)_J} \right) \tag{0.12}$$

0.2.6 Model optimization

θ

θ

m

θ

m

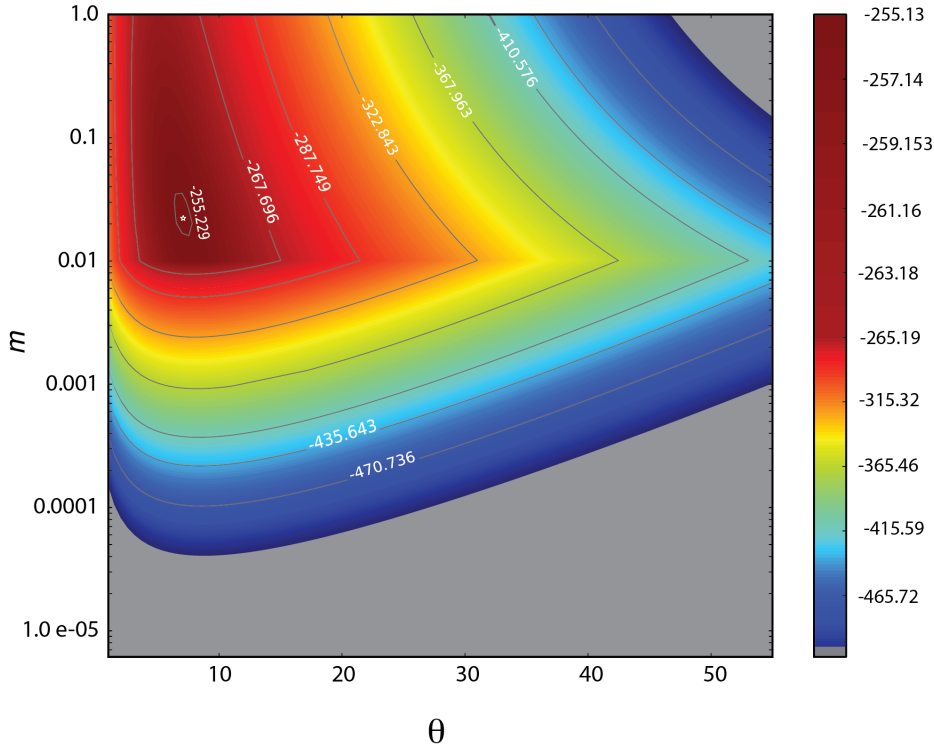

Figure 0.1: Maximum likelihood inference of neutral parameters. Log likelihood surface as a function of migration rate (m), and the fundamental biodiversity number (θ) for *D. rerio* chromosome 19. Dark red color shows regions of the surface where parameters maximize the probability to explain abundances and diversity of genetic elements in the chromosome. Likelihood ratio tests favored Etienne in contrast to Ewens sampling formula to explain the observed data in the chromosome.

0.2.7 Model testing

m

0.2.8 Test for neutrality

S

θ

m

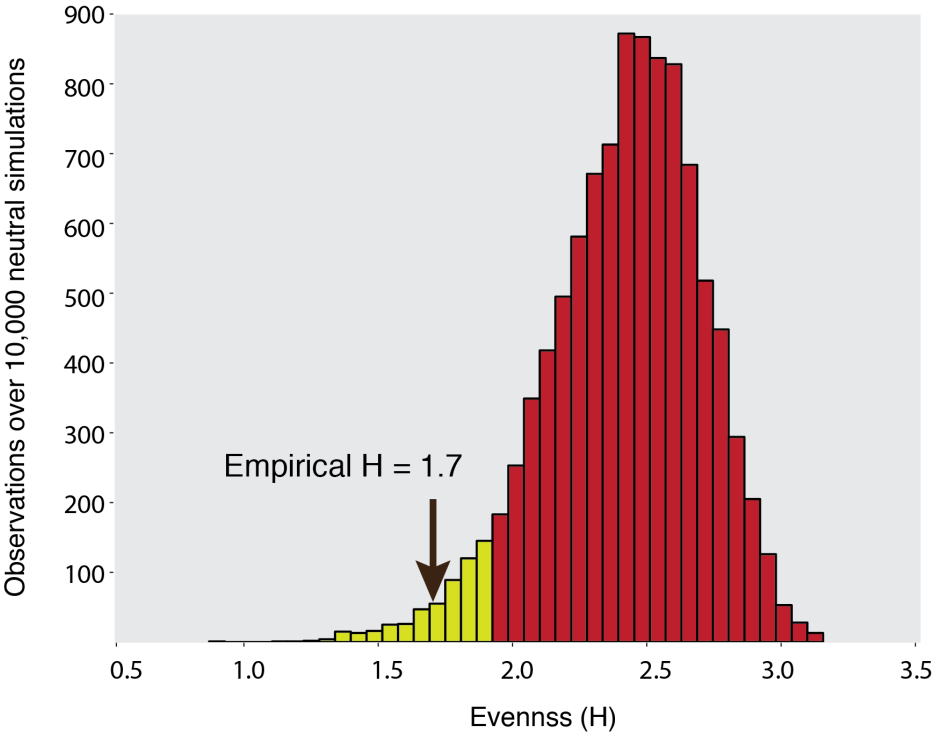

Figure 0.2: Comparing simulated and empirical evenness.
Neutrality test statistically compares simulated null distribution of H with the empirical H value of the corresponding chromosome. In this case, the null distribution of H values was derived from 10,000 neutral simulations of A. gambiae chromosome 2L, with neutral parameters (θ and m) optimized by ML using Etienne sampling formula. Yellow and red bars display 5% and 95% of the simulated neutral data, respectively. Although in this case neutrality was rejected (p= 0.01), posterior correction by multiple testing favored the null neutral hypothesis (q= 0.21).

0.3 Detection of selective pressure at molecular level

0.3.1 Orthology prediction

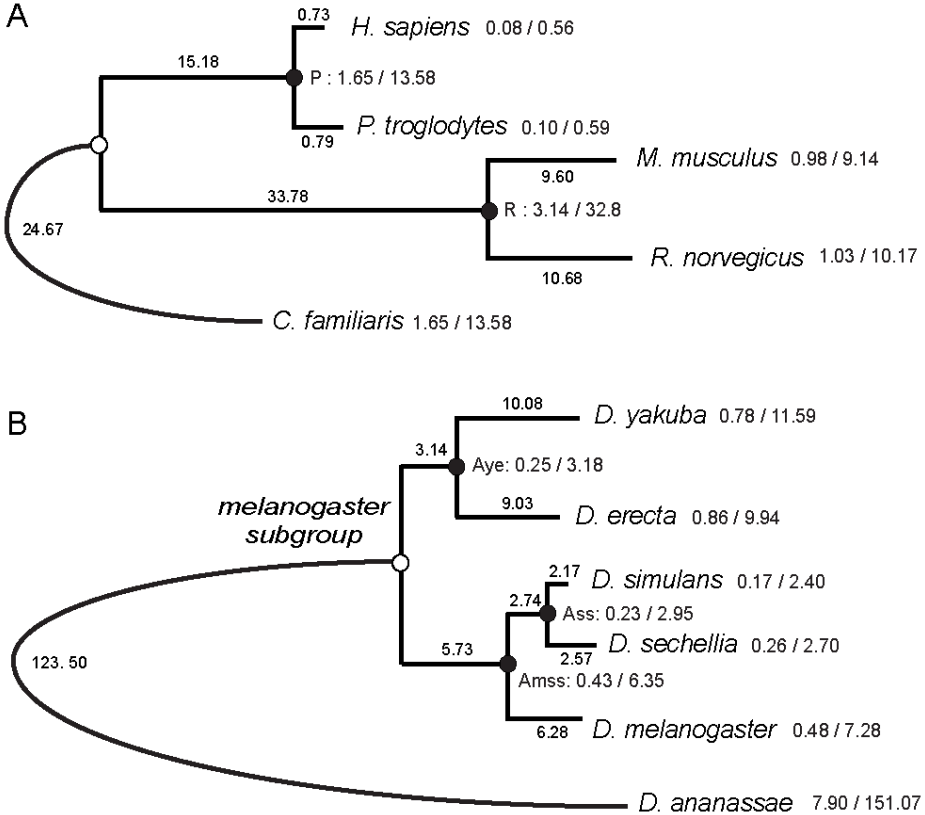

Figure 0.3: Mammals and *melanogaster* group phylogeny.
Numbers on internal and external nodes represent the median number of non-synonymous and synonymous substitutions per codon (dN/dS) estimated from all the coding sequences compared in mammal (A) and Drosophila (B) genomes. Branch lengths and rates were multiplied by 100. Ancestral estimation of parameters was done in primates (P), rodents (R), D. yakuba and D. erecta (Aye), D. simulans and D. sechellia (Ass), and D. melanogaster, D. simulans and D. sechellia (Amss). C. familiaris and D. ananassae were chosen as outgroup species in the corresponding tree.

0.3.2 Alignments refinement and filters

0.3.3 Evolutionary analysis

0.3.4 GSSA, evolutionary and statistical simulations

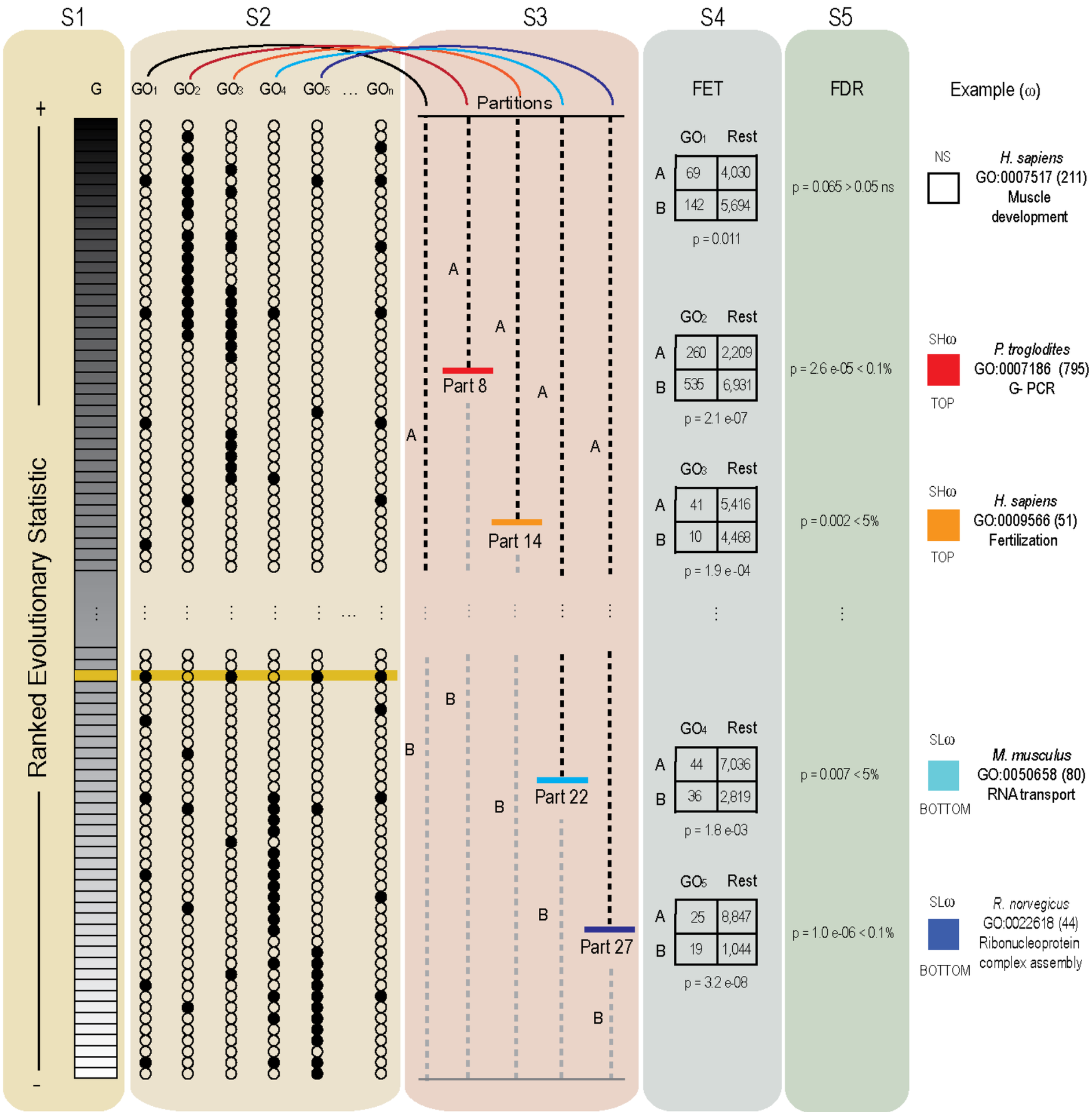
S1
S2
S3
S4
S5
G
GO1 GO2 GO3 GO4 GO5 … GOn
Partitions
Ranked Evolutionary Statistic
+
-
A
B
Part 8
Part 14
Part 22
Part 27
FET
GO1 Rest
A 69 4,030
B 142 5,694
p = 0.011
GO2 Rest
A 260 2,209
B 535 6,931
p = 2.1 e-07
GO3 Rest
A 41 5,416
B 10 4,468
p = 1.9 e-04
GO4 Rest
A 44 7,036
B 36 2,819
p = 1.8 e-03
GO5 Rest
A 25 8,847
B 19 1,044
p = 3.2 e-08
FDR
p = 0.065 > 0.05 ns
p = 2.6 e-05 < 0.1%
p = 0.002 < 5%
p = 0.007 < 5%
p = 1.0 e-06 < 0.1%
Example (ω)
NS
H. sapiens
GO:0007517 (211)
Muscle development
SHω
TOP
P. troglodites
GO:0007186 (795)
G- PCR
SHω
TOP
H. sapiens
GO:0009566 (51)
Fertilization
SLω
BOTTOM
M. musculus
GO:0050658 (80)
RNA transport
SLω
BOTTOM
R. norvegicus
GO:0022618 (44)
Ribonucleoprotein complex assembly

Figure 0.5:

Summary of the steps developed by the GHRA:**Summary of the steps developed by the GHRA**GHRA can be roughly described in a series of five steps S0 to S4.: S0, create proof of a genuine recording in an evolutionary context.- S1, assign functional classes to all the found genes. S2, apply a fixed number of partitions on the ranked list. S3, proceed with a Fisher's exact test (FEXT) in each partition. S4, adjust p-values by FDR. See for a full description. Colored boxes (red, orange, cyan and blue) represent functional modules with genes significantly accumulated (FAT-FEXT and FT-FDR) at the corresponding subranges of a list (top and bottom), and described with significantly high (HH) and low (LL) values of the evolutionary variable (variable) respectively. Within segments in non-significant association (NS). Rectangles show the subranges (S2) categories with significant and non-significant distributions of the variable. As parenthesis, the total number of genes corresponding to the title term is shown. For GMS, the functions seems to be accumulated with the arrangement of the genes. In the example (GO:0009570) proteins 60 in human (not shown in the picture) reported the lowest p-value ($p = 0.011$) although it was not significant after FDR correction ($FDR = 0.085$). Figure 2: and lower left sides of the ranked list. Red represents both sides of the specified partition number. Remember GO categories (GO:0110165) show the association of dark dots with values located at the top (significant high - value - HH), and at the bottom (significant low - value - LL) of the list (for GO:0009570 and GO:0110165, respectively). In example (FDR) found the most significant p-value for proteins 6, 14, 22 and 23 for GO:0009570, GO:0009506, GO:0009536 and GO:0009514 (Ribosome, ferrous, nucleoid and cell pariete, respectively).

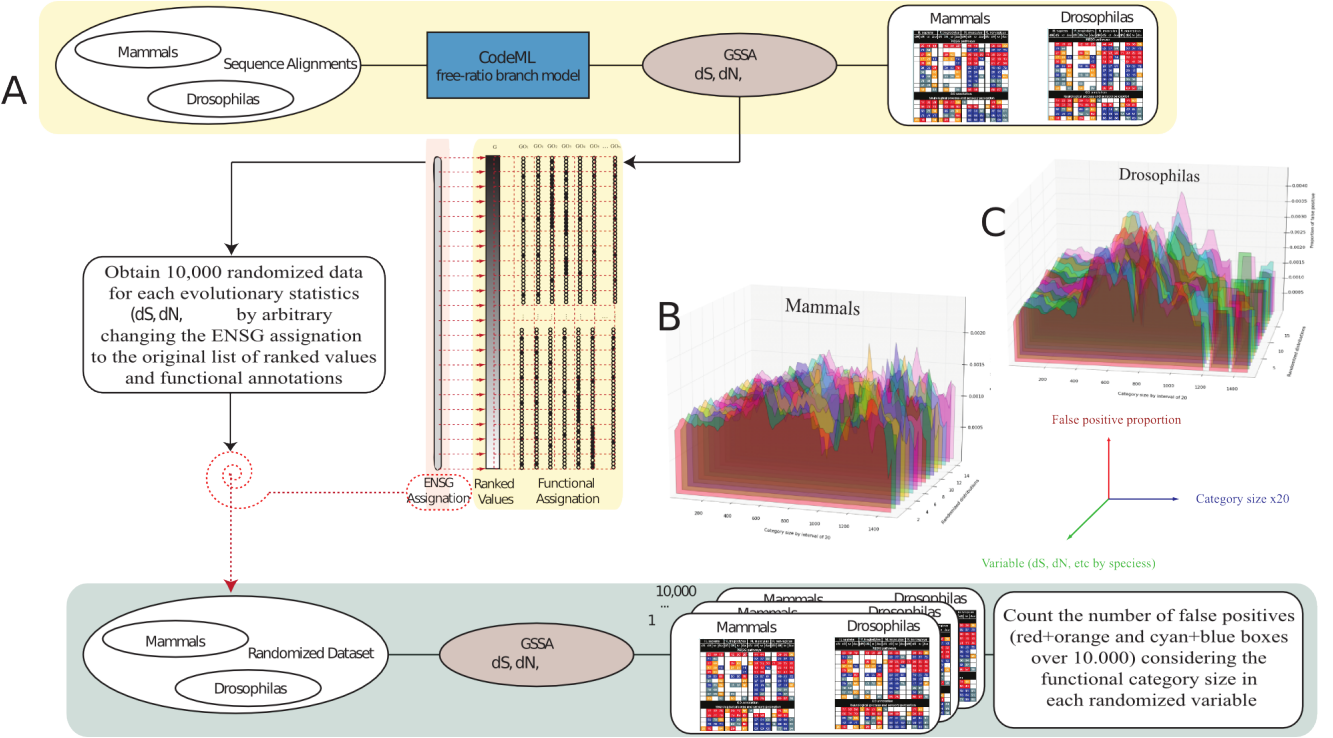
A
Mammals
Drosophilas
Sequence Alignments
CodeML
free-ratio branch model
GSSA
dS, dN,
Mammals
Drosophilas
Obtain 10,000 randomized data for each evolutionary statistics (dS, dN, by arbitrary changing the ENSG assignation to the original list of ranked values and functional annotations
ENSG Assignation
Ranked Values
Functional Assignation
B
Mammals
C
Drosophilas
False positive proportion
Category size x20
Variable (dS, dN, etc by speciess)
Mammals
Drosophilas
Randomized Dataset
GSSA
dS, dN,
10,000
1
Mammals
Drosophilas
Count the number of false positives (red+orange and cyan+blue boxes over 10.000) considering the functional category size in each randomized variable

Figure 0.7:

[Randomisation experiment.]**Randomisation experiment.**(A) The pipeline shows the steps followed to tests possible biases attributed to the size of the functional category, the magnitude of change in evolutionary rate and the proportion of genes experiencing a rate change in the GSSA. The proportion of false positive results never reached 5% (FDR) in mammals (B) and Drosophila (C).

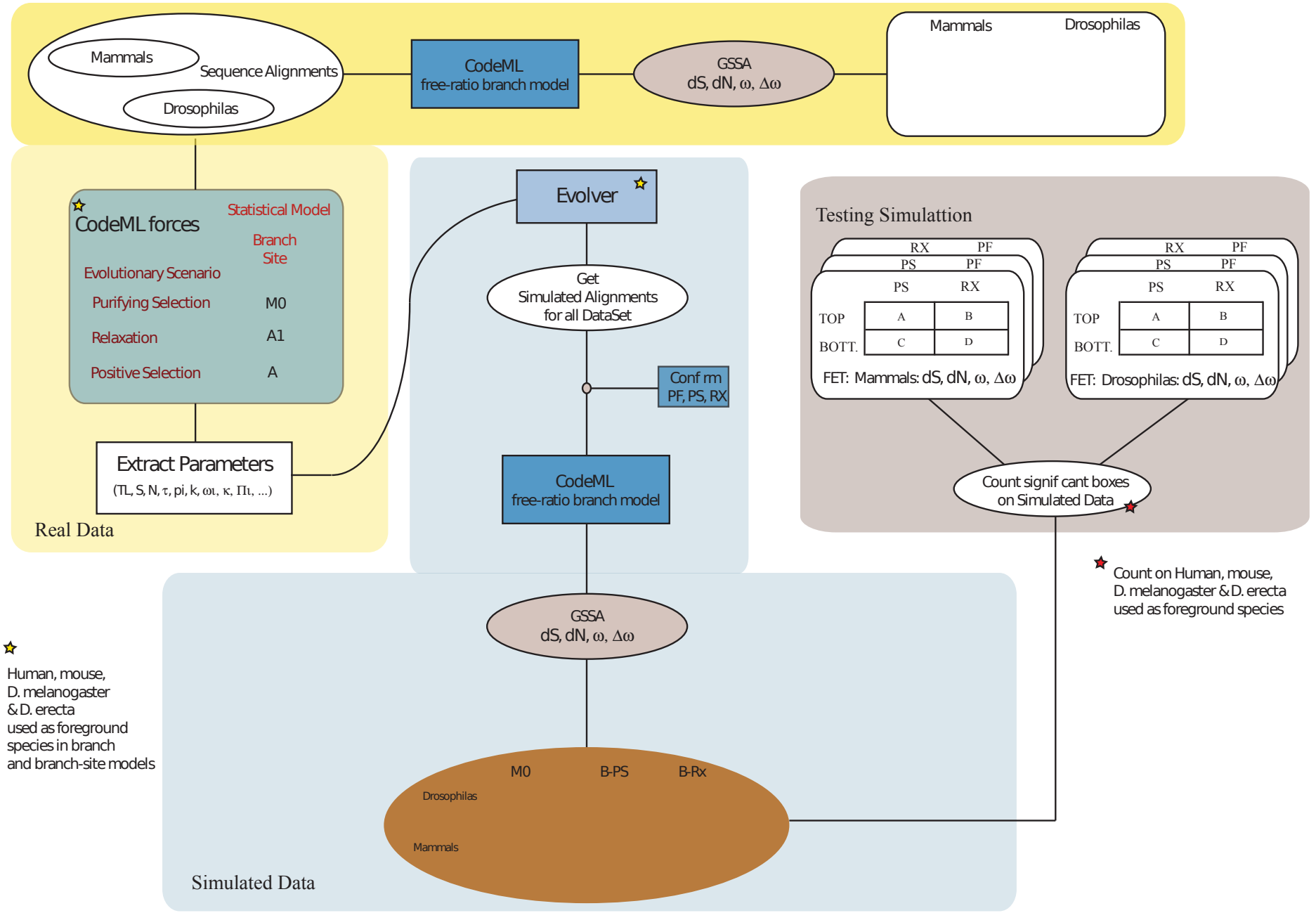

Figure 0.8: Evolutionary and statistical simulation of GSSA.
The pipeline shows the steps taken along three different spaces of analysis, the real data, the simulated data and the testing block. See Supplementary Results for a complete explanation of methods and results.

	PS		RX		PF	
	# PSG	# RXG	# PSG	# RXG	# PSG	# RXG
Homo sapiens	658	1640	11	1939	0	1
Mus musculus	1500	954	14	1565	1	0
D. melanogaster	736	630	25	1104	0	0
D. erecta	778	1292	26	1713	2	1

Table 0.3: Number of PSG and relaxed genes (RXG) in each of the simulated evolutionary scenarios.

	PS	RX	PF
PS	—	92.50%	98.50%
RX	91.10%	—	99.00%
PF	88.90%	90.60%	—

Table 0.4: Proportion of significant functional categories that are still significant (identical signs of odd-ratios) under a different evolutionary scenario.