Supplementary material
Detecting molecular signatures of phenotypic convergence

Olivier Chabrol, Pierre Pontarotti, Manuela Royer-Carenzi
and Gilles Didier
Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France

May 11, 2017

1 Convergent genes for $\gamma = 10^{-4}$

Table 1 displays the genes detected convergent by setting parameter γ to 10^{-4}.

Table 2 displays the most significant GO-terms associated to genes detected for $\gamma = 10^{-4}$. Enrichment p-values are not corrected for multiple-testing with regard to the number of GO-terms.

Rank	RefSeq ID	Length	Corrected p-value	Number of convergent sites	Convergent sites
1	KMT2A	1683	9.58×10^{-21}	15	142, 245, 247, 303, 420, 902, 921, 945, 949, 1050, 1077, 1213, 1381, 1382, 1664
2	SCN3A	1313	3.71×10^{-7}	7	421, 677, 735, 742, 838, 1006, 1266
3	**SLC26A5**	330	9.98×10^{-5}	4	128, 164, 226, 248
4	BRINP3	469	3.04×10^{-4}	4	124, 128, 221, 267
5	**TMC1**	469	2.43×10^{-4}	4	46, 82, 227, 380
6	CBL	476	2.15×10^{-4}	4	419, 421, 422, 432
7	PLCL1	720	9.49×10^{-4}	4	211, 252, 267, 620
8	ELF2	248	1.95×10^{-3}	3	11, 17, 223
9	PTPRZ1	934	2.06×10^{-3}	4	444, 476, 691, 913
10	TESK2	280	2.24×10^{-3}	3	116, 226, 243
11	**DFNB59**	325	3.18×10^{-3}	3	97, 163, 175
12	RIC1	1149	3.48×10^{-3}	4	274, 318, 626, 937
13	RELN	2631	4.10×10^{-3}	5	310, 360, 579, 1320, 1439
14	WDR26	388	4.24×10^{-3}	3	367, 378, 385
15	MALT1	442	5.84×10^{-3}	3	41, 397, 415
16	DICER1	1401	5.66×10^{-3}	4	218, 998, 1034, 1161
17	EDC4	518	8.25×10^{-3}	3	197, 410, 411
18	KLHL5	534	8.53×10^{-3}	3	98, 128, 315
19	XPO6	776	2.44×10^{-2}	3	163, 518, 631
20	TAOK3	795	2.49×10^{-2}	3	51, 122, 334
21	FNDC1	133	2.62×10^{-2}	2	15, 22
22	TTC21B	864	2.89×10^{-2}	3	363, 481, 519
23	COL9A3	144	2.81×10^{-2}	2	47, 111
24	ZNF143	149	2.88×10^{-2}	2	96, 114
25	TNKS2	944	3.30×10^{-2}	3	415, 654, 866
26	RNF13	172	3.54×10^{-2}	2	35, 166
27	FGF5	177	3.61×10^{-2}	2	14, 22
28	ANXA1	182	3.68×10^{-2}	2	23, 103
29	SOS1	1023	3.60×10^{-2}	3	401, 483, 934
30	MED1	1045	3.70×10^{-2}	3	418, 489, 875
31	EXOSC7	193	3.74×10^{-2}	2	170, 179
32	LMAN2L	193	3.62×10^{-2}	2	23, 67
33	NCOA2	1070	3.61×10^{-2}	3	531, 934, 981
34	PBRM1	1086	3.66×10^{-2}	3	115, 189, 652
35	PIP4K2A	200	3.55×10^{-2}	2	94, 123
36	TATDN1	201	3.49×10^{-2}	2	4, 32
37	COL11A1	1092	3.41×10^{-2}	3	156, 287, 771
38	ABAT	204	3.40×10^{-2}	2	51, 199
39	GBAS	205	3.35×10^{-2}	2	27, 198
40	KCNAB1	206	3.30×10^{-2}	2	32, 93
41	EPYC	210	3.34×10^{-2}	2	67, 198
42	ARFGEF1	1164	3.62×10^{-2}	3	1136, 1143, 1161
43	CALB1	223	3.59×10^{-2}	2	40, 218
44	LGI2	223	3.51×10^{-2}	2	137, 204
45	AHCTF1	1190	3.61×10^{-2}	3	194, 446, 910
46	NOL4	232	3.63×10^{-2}	2	152, 224
47	CCDC88A	1203	3.56×10^{-2}	3	1026, 1036, 1111
48	LARGE	243	3.82×10^{-2}	2	102, 157
49	SLC44A1	247	3.86×10^{-2}	2	74, 237
50	CCDC129	249	3.85×10^{-2}	2	24, 125
51	KCNJ15	258	4.05×10^{-2}	2	6, 237
52	NOP58	262	4.09×10^{-2}	2	246, 255
53	DHX32	269	4.23×10^{-2}	2	61, 124
54	**OTOF**	270	4.18×10^{-2}	2	112, 198
55	BBS4	271	4.14×10^{-2}	2	73, 252
56	PIP5K1A	280	4.34×10^{-2}	2	231, 242
57	CREBBP	1368	4.27×10^{-2}	3	695, 1211, 1347
58	BICC1	291	4.52×10^{-2}	2	176, 212
59	AMPH	297	4.63×10^{-2}	2	245, 279
60	CHD6	1441	4.72×10^{-2}	3	823, 1405, 1412
61	SCG3	309	4.84×10^{-2}	2	148, 277
62	ERO1A	311	4.82×10^{-2}	2	132, 196
63	HNF4G	311	4.75×10^{-2}	2	77, 155
64	GAS2L3	313	4.73×10^{-2}	2	290, 291
65	MPP6	320	4.87×10^{-2}	2	91, 288

Table 1: Genes detected convergent with $\gamma = 10^{-4}$ between dolphin and microbat. The genes previously reported for showing convergence and/or adaptation in echolocation are in bold.

Fisher's exact p-value	GO ID	Description	Genes
1.08×10^{-4}	GO:0030375	thyroid hormone receptor coactivator activity	MED1, NCOA2
1.79×10^{-4}	GO:0043025	neuronal cell body	SCN3A, BRINP3, PTPRZ1, DFNB59, SOS1, NCOA2, KCNAB1, CALB1
2.24×10^{-4}	GO:0021591	ventricular system development	DICER1, TTC21B, BBS4
4.36×10^{-4}	GO:0007605	sensory perception of sound	SLC26A5, TMC1, DFNB59, COL11A1, OTOF
6.37×10^{-4}	GO:0035640	exploration behavior	KMT2A, ABAT
6.37×10^{-4}	GO:0035864	response to potassium ion	KMT2A, SLC26A5
1.06×10^{-3}	GO:2000273	positive regulation of receptor activity	MED1, NCOA2
1.08×10^{-3}	GO:0030425	dendrite	BRINP3, PTPRZ1, RELN, DICER1, NCOA2, CALB1, ERO1A
1.27×10^{-3}	GO:0003007	heart morphogenesis	SOS1, PBRM1, COL11A1
1.46×10^{-3}	GO:0007611	learning or memory	PTPRZ1, KCNAB1, CALB1
1.57×10^{-3}	GO:0090303	positive regulation of wound healing	ANXA1, ARFGEF1
1.57×10^{-3}	GO:0030837	negative regulation of actin filament polymerization	ARFGEF1, BBS4
1.57×10^{-3}	GO:0016307	phosphatidylinositol phosphate kinase activity	PIP4K2A, PIP5K1A
1.57×10^{-3}	GO:0016308	1-phosphatidylinositol-4-phosphate 5-kinase activity	PIP4K2A, PIP5K1A
1.57×10^{-3}	GO:0051057	positive regulation of small GTPase mediated signal transduction	RELN, SOS1
1.89×10^{-3}	GO:0030216	keratinocyte differentiation	ANXA1, MED1, PIP5K1A
2.12×10^{-3}	GO:0005737	cytoplasm	KMT2A, SCN3A, SLC26A5, CBL, PLCL1, ELF2, PTPRZ1, TESK2, RIC1, RELN, WDR26, MALT1, DICER1, EDC4, KLHL5, XPO6, TAOK3, FNDC1, TTC21B, TNKS2, RNF13, ANXA1, EXOSC7, NCOA2, PIP4K2A, KCNAB1, ARFGEF1, CALB1, AHCTF1, CCDC88A, NOP58, OTOF, BBS4, PIP5K1A, CREBBP, BICC1, AMPH, GAS2L3
2.19×10^{-3}	GO:0061512	protein localization to cilium	TTC21B, BBS4
2.39×10^{-3}	GO:0021766	hippocampus development	PTPRZ1, RELN, BBS4
2.89×10^{-3}	GO:0032743	positive regulation of interleukin-2 production	MALT1, ANXA1
2.89×10^{-3}	GO:1900087	positive regulation of G1/S transition of mitotic cell cycle	ANXA1, CREBBP
2.89×10^{-3}	GO:0042975	peroxisome proliferator activated receptor binding	MED1, CREBBP
3.70×10^{-3}	GO:0035855	megakaryocyte development	MED1, PIP4K2A
3.70×10^{-3}	GO:0010001	glial cell differentiation	RELN, FGF5
4.14×10^{-3}	GO:0005654	nucleoplasm	KMT2A, ELF2, TESK2, WDR26, EDC4, FNDC1, ZNF143, ANXA1, MED1, NCOA2, PBRM1, TATDN1, ARFGEF1, AHCTF1, SLC44A1, NOP58, PIP5K1A, CREBBP, CHD6, HNF4G
4.59×10^{-3}	GO:0050910	detection of mechanical stimulus involved in sensory perception of sound	TMC1, COL11A1
4.59×10^{-3}	GO:0008589	regulation of smoothened signaling pathway	TTC21B, CREBBP
4.59×10^{-3}	GO:0046488	phosphatidylinositol metabolic process	PIP4K2A, PIP5K1A
4.82×10^{-3}	GO:0004386	helicase activity	DICER1, ANXA1, DHX32, CHD6
5.57×10^{-3}	GO:0031018	endocrine pancreas development	ANXA1, HNF4G
5.57×10^{-3}	GO:0000242	pericentriolar material	TNKS2, BBS4
5.57×10^{-3}	GO:0042974	retinoic acid receptor binding	MED1, NCOA2
5.57×10^{-3}	GO:0046966	thyroid hormone receptor binding	MED1, NCOA2
5.57×10^{-3}	GO:0005086	ARF guanyl-nucleotide exchange factor activity	ARFGEF1, AMPH
6.64×10^{-3}	GO:0048041	focal adhesion assembly	TESK2, PIP5K1A
6.64×10^{-3}	GO:0004712	protein serine/threonine/tyrosine kinase activity	TESK2, RELN
6.64×10^{-3}	GO:0007616	long-term memory	RELN, CALB1
6.77×10^{-3}	GO:0005578	proteinaceous extracellular matrix	PTPRZ1, RELN, COL9A3, COL11A1, EPYC
7.79×10^{-3}	GO:0090102	cochlea development	SLC26A5, OTOF
9.03×10^{-3}	GO:0035162	embryonic hemopoiesis	KMT2A, MED1
9.03×10^{-3}	GO:0016922	ligand-dependent nuclear receptor binding	MED1, NCOA2

Table 2: Most significant GO terms for the genes of Table 1.

2 Influence of parameter γ

Figures 1 to 5 display the number of genes detected convergent for all pairs of species by setting the site significance threshold γ to 10^{-3}, 5×10^{-4}, 10^{-4}, 5×10^{-5} and 10^{-5} respectively as well as the enrichment p-values of the convergent genes with regard to the audition-related GO-terms.

Remark that the number of detected genes tends to decrease with γ, though there are exceptions. As an example, we found 137 convergent genes with $\gamma = 10^{-3}$ and 142 with $\gamma = 5 \times 10^{-4}$ for the pair alpaca-cow. By construction, the number of significant sites of a given gene always decreases with γ but a smaller number of significant site may lead to a lower p-value as γ decreases. The lowest enrichment p-value is obtained with $\gamma = 5 \times 10^{-5}$ for the echolocating pair but it is still lower than 5% for $\gamma = 10^{-5}$.

The significance of audition enrichment obtained for $\gamma = 10^{-3}$ are not conclusive (Figure 1). A possible explanation is that the null distribution used for estimating the significance of sites is too far from the "real" underlying one for this level of probability.

	Cow	Alpaca	Dolphin	Megabat	Microbat	Elephant	Human	Marmoset
Alpaca	137							
Dolphin	2	90						
Megabat	316	202	291					
Microbat	470	223	318	3				
Elephant	505	258	397	277	360			
Human	404	229	398	239	240	29		
Marmoset	364	264	214	192	219	29	0	
Mouse	512	338	375	274	557	282	27	44

	Cow	Alpaca	Dolphin	Megabat	Microbat	Elephant	Human	Marmoset
Alpaca	1.02×10^{-1}							
Dolphin	1.00	8.16×10^{-1}						
Megabat	2.00×10^{-1}	7.02×10^{-1}	3.57×10^{-2}					
Microbat	7.41×10^{-1}	6.41×10^{-1}	1.93×10^{-1}	1.00				
Elephant	4.86×10^{-1}	8.00×10^{-1}	3.88×10^{-1}	9.19×10^{-1}	8.73×10^{-1}			
Human	5.19×10^{-1}	3.93×10^{-1}	4.08×10^{-1}	1.85×10^{-2}	4.29×10^{-1}	1.00		
Marmoset	3.74×10^{-1}	2.06×10^{-1}	8.65×10^{-1}	3.30×10^{-1}	2.11×10^{-1}	6.46×10^{-1}	1.00	
Mouse	4.44×10^{-2}	6.21×10^{-1}	3.15×10^{-1}	7.33×10^{-1}	6.23×10^{-1}	8.55×10^{-1}	6.20×10^{-1}	4.61×10^{-1}

Figure 1: Number of genes detected convergent with a site significance threshold $\gamma = 10^{-3}$ for all pairs of species of the dataset (top) and their enrichment p-values with regard to the audition-related GO-terms (bottom).

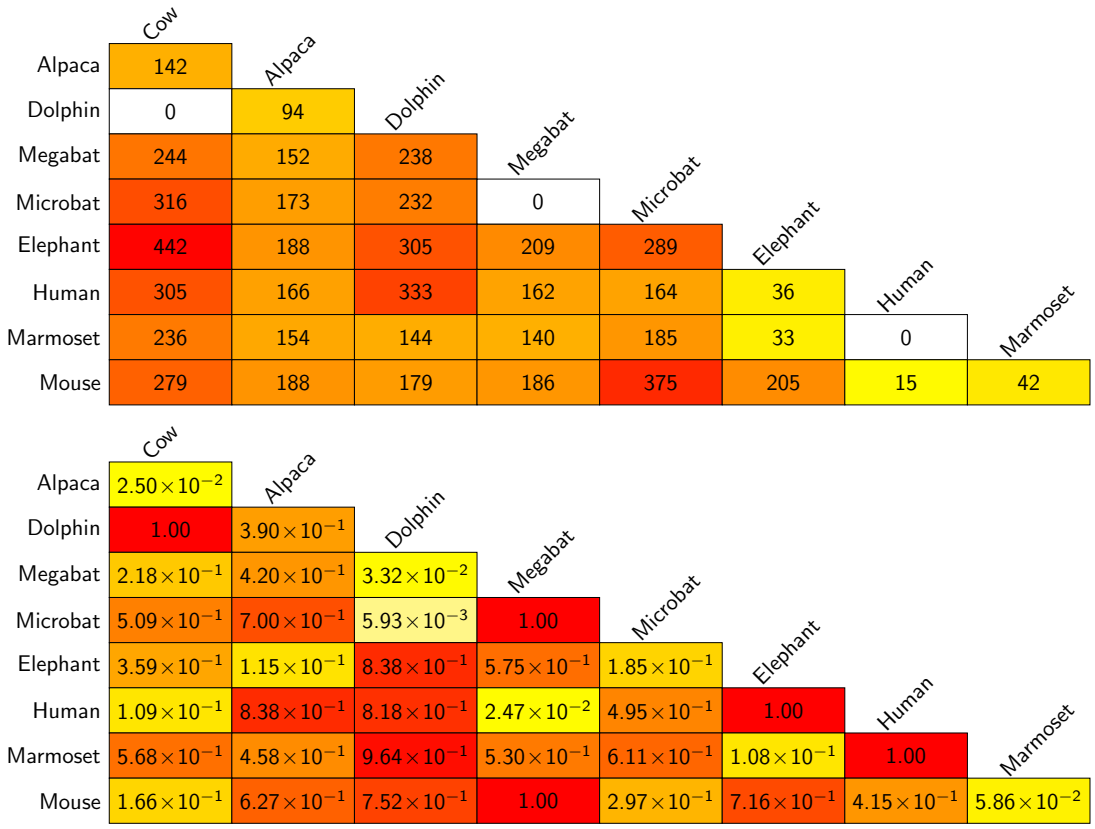

Figure 2: Number of genes detected convergent with a site significance threshold $\gamma = 5 \times 10^{-4}$ for all pairs of species of the dataset (top) and their enrichment p-values with regard to the audition-related GO-terms (bottom).

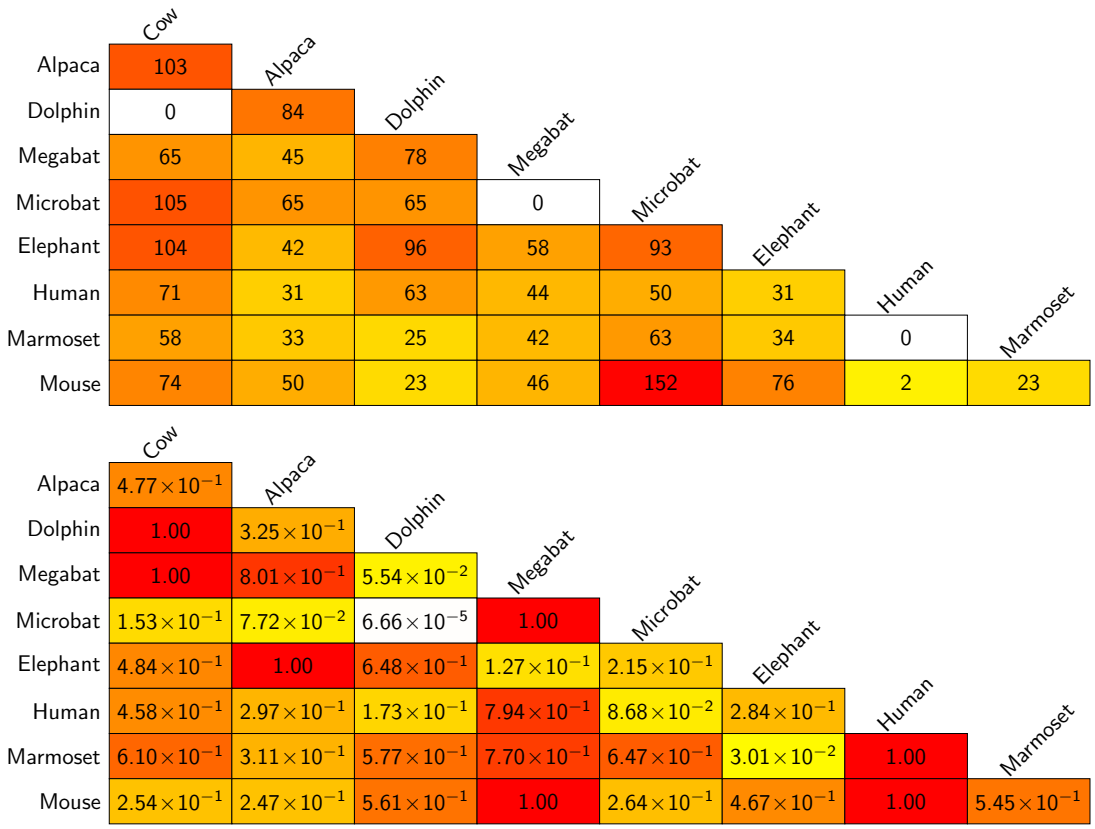

Figure 3: Number of genes detected convergent with a site significance threshold $\gamma = 10^{-4}$ for all pairs of species of the dataset (top) and their enrichment p-values with regard to the audition-related GO-terms (bottom).

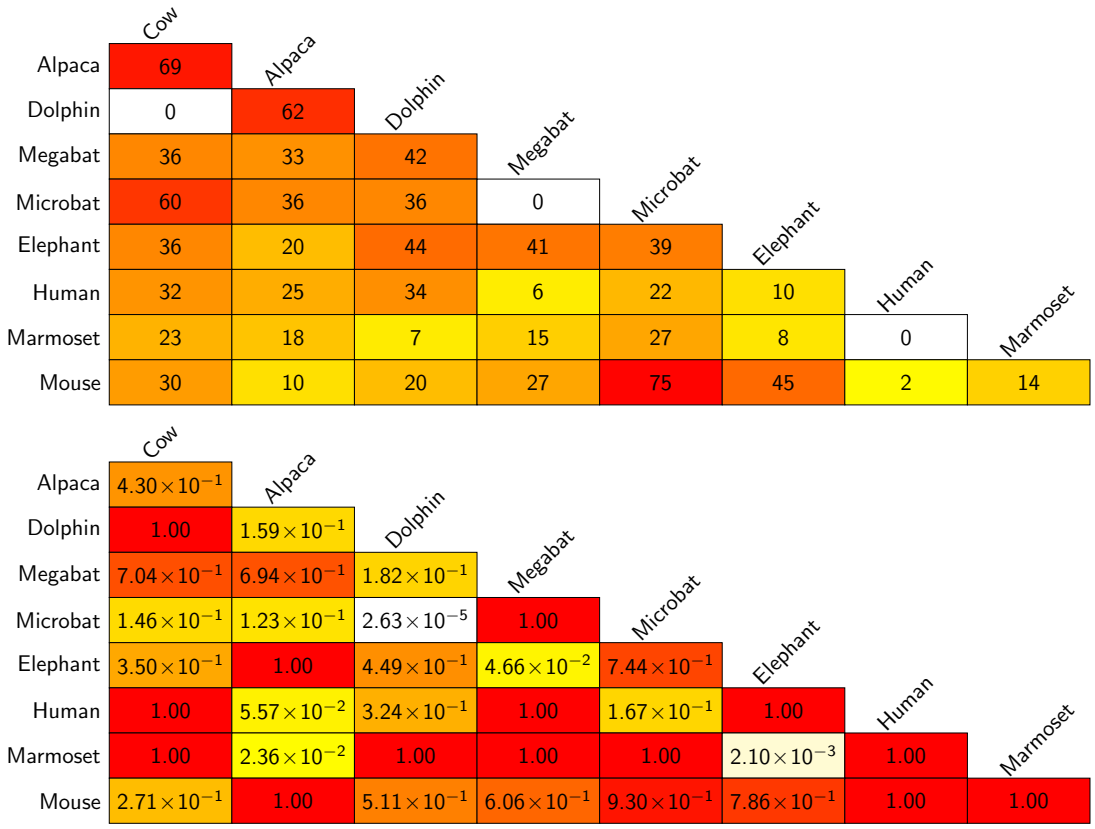

Figure 4: Number of genes detected convergent with a site significance threshold $\gamma = 5 \times 10^{-5}$ for all pairs of species of the dataset (top) and their enrichment p-values with regard to the audition-related GO-terms (bottom).

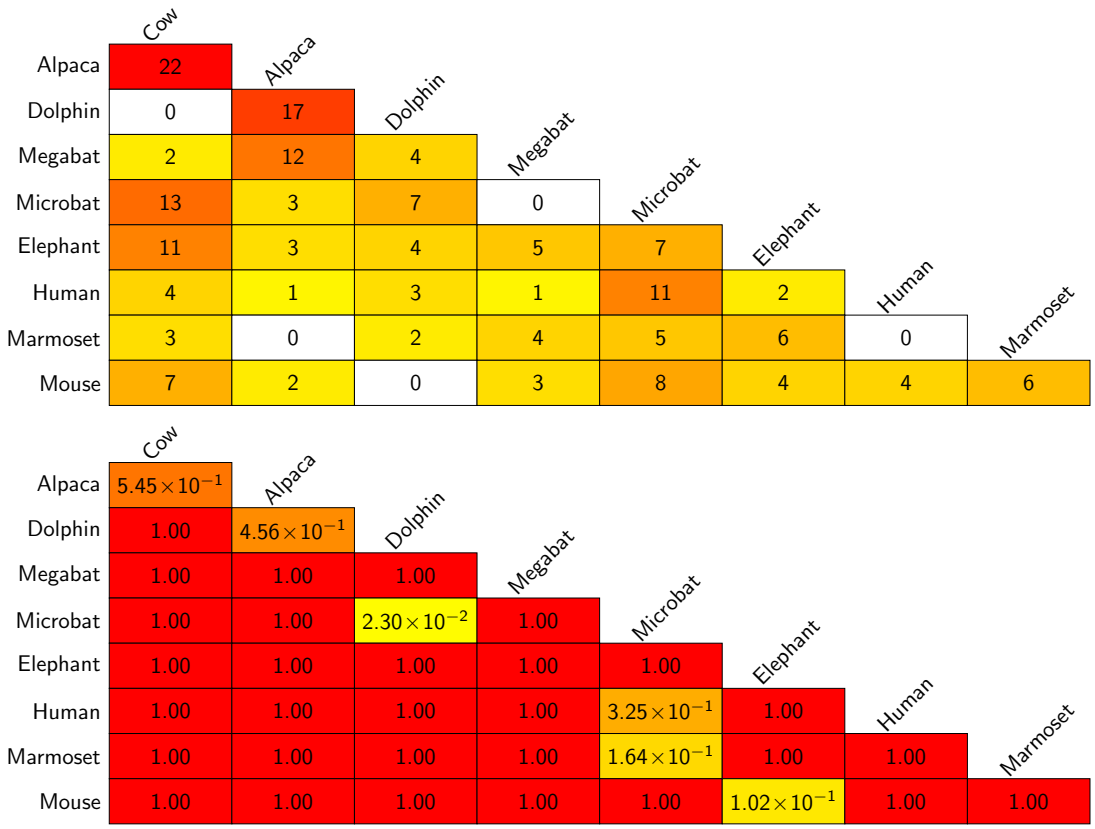

Figure 5: Number of genes detected convergent with a site significance threshold $\gamma = 10^{-5}$ for all pairs of species of the dataset (top) and their enrichment p-values with regard to the audition-related GO-terms (bottom).

3 Software

We provide a package containing two software:

- `msd` detects molecular signatures of convergence events by using the convergence index.
- `enr` computes GO terms enrichments of a list of genes.

3.1 Detection pipeline – `msd`

Software `msd` implements the whole detection pipeline. On our configuration (two Intel® Xeon(R) CPU E5-2680 v2 2.80GHz × 40), it took about 10 hours for treating the 6,332 genes of the dataset. Most of the computational time is spent for simulating empirical distributions of the convergence indexes for all alignments/genes.

```
NAME
msd - detection of molecular signatures of convergence events

SYNOPSIS
msd [OPTIONS] [FILE TREE] [FILE CONV.] [ALIGN. FILE 1] [ALIGN. FILE 2] ...

DESCRIPTION
return a table where each line displays the result of the detection of molecular signature
of each alignment files [ALIGN. FILE i] in Fasta format with the regard to the tree in Newick format
[FILE TREE] and the character stored as a table in [FILE CONV.].

Options are

-n [FILE]
set the optimisation options to those contained in [FILE].
-m [FILE]
set the evolution model to that contained in [FILE].
-o [NAME]
set the name of the output file to [NAME]. By default, it is 'output.txt'.
-e [VALUE]
set the threshold under which a site is considered significant.
-e [VALUE]
save an extra file named 'significant.txt', which contains all the names of alignments
with a corrected p-value under [VALUE].
-t [NUMBER]
set the maximal number of simultaneous threads to [NUMBER].
-h
display help
```

3.2 GO-terms enrichment – `enr`

Software `enr` is just another implemention of the hypergeometric "Fisher's exact" test of GO-terms enrichment.

```
NAME
enr - compute GO terms enrichment of a list of genes.

SYNOPSIS
enr [OPTIONS] [FILE GO] [FILE LIST]

DESCRIPTION
save a file containg the GO terms associated to at least a gene of [FILE LIST]
according to [FILE GO].

-o [FILE]
load the GO descriptors from [FILE].
-b [FILE LIST]
set the background set of genes to those of [FILE LIST].
-h
display help
```