Gene number expansion and contraction in vertebrate genomes with respect to invertebrate genomes

Anuphap Prachumwat and Wen-Hsiung Li

Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, IL 60637

Functional bias in vertebrate and invertebrate gene groups (gene type number *vs.* gene copy number)

In general, the patterns of functional bias for the number of gene types and for the number of gene copies are similar, though more significantly differential categories are observed for gene copy number than for gene type number because there are more gene copies than gene types (Fig. 6 and Supplemental Fig. 4S). However, for some functional categories we observe different functional bias patterns between gene type number and gene copy number. For the biological process categories, the magnitude of gene copy number overrepresentation in *Vonly* for each of the signal transduction, multicellular organismal development and "biological_process" categories is larger than the magnitude of gene type number overrepresentation. Moreover, the gene copy numbers for the signal transduction and "biological_process" categories are significantly overrepresented in *Vonly*, *VFProk*, and *V.A50* and underrepresented in *V.MCL* (Supplemental Fig. 4S), whereas the gene type numbers in these two categories are significantly overrepresented only in *Vonly* and *V.A50* (Fig 6). The gene copy number in the regulation of biological process category is overrepresented in both *Vonly* and *V.MCL* but underrepresented in *V.A50* and *VFProk* (Supplemental Fig 4S). However, the number of gene types in this category is overrepresented in *Vonly* but underrepresented in *V.MCL* (Fig. 6). The inconsistent patterns of functional bias between the number of gene types and the number of gene copies are observed only for the genes involved in transcription. In

this category, gene types are significantly overrepresented in *Vonly* but significantly underrepresented in *V.MCL*; however, gene copies are significantly underrepresented in *Vonly* but overrepresented in *V.MCL* (Fig 6 vs. Supplemental Fig. 4S).

For the cellular component categories, the functional bias patterns for gene type number and for gene copy number are very similar. However, different patterns are observed for the proteins localized to chromosome or to cytoskeleton: the number of gene copies is significantly *under*represented in *Vonly* and *over*represented in *V.MCL*, but the number of gene types is *over*represented in *Vonly* and *under*represented in *V.MCL* (Supplemental Figs 5S and 6S).

In addition, a similar comment applies to functional bias patterns for the numbers of the invertebrate gene copies and gene types (Supplemental Figs 7S vs. 8S and Figs. 9S vs. 10S). In particular, the opposite functional bias patterns between the number of gene copies and the number of gene types are observed for the invertebrate genes involved in the transcription, multicellular organismal development, and anatomical structure morphogenesis categories, where the number of gene copies is significantly overrepresented in *Ionly* (and underrepresented in *I.MCL*) but the number of gene types is underrepresented in *Ionly* (and overrepresented in *I.MCL*) (Figs 7S and 8S).

In summary, except for some functional categories, the functional bias patterns for the gene copy number and for the gene type number are similar.

Supplemental Tables:

Table 1S. The numbers of genes (and gene types [a]) in invertebrate gene groups.

Invertebrate gene group	ciona	sea urchin	yellow fever mosquito	African mosquito	fruitfly	worm
I.MCL	9151 (4497)	16958 (5592)	9192 (4334)	7997 (4290)	8125 (4362)	7428 (3575)
I.A50	137 (88)	415 (120)	335 (94)	302 (97)	194 (70)	185 (35)
Ionly	4895 (3937)	4104 (2777)	5873 (4113)	4885 (3943)	5701 (4812)	12420 (7818)
Ionly genes shared by ≥ 2 invertebrate genomes	141 (66)	201 (93)	2868 (1617)	2273 (1562)	1406 (938)	98 (53)
All genes	14183 (8522)	21477 (8489)	15400 (8541)	13184 (8330)	14020 (9244)	20033 (11428)

[a] number of gene types is the sum of the number of singletons and the number of gene families.

Supplemental Figures:

Figure 1S. Proportions of invertebrate genes grouped by homology in the 12 vertebrate genomes (4 fish plus 8 land vertebrate genomes). For each invertebrate genome on the x-axis, the invertebrate gene groups are represented, from the top to bottom, by segments of a bar for the genes that can be found in ≥1 genome of the 12 invertebrate genomes with standard homology search criteria (denoted as *I.MCL*), for those that are not in *I.MCL* but have homolog in the vertebrate genomes when the homology search criteria are relaxed (denoted as *I.A50*), and for those that are in none of the above two groups (denoted as *Ionly*).

Figure 2S. Proportions of the invertebrate *I.MCL*, *I.A50*, and *Ionly* genes that are singletons, in the two-gene families, or in the multigene families.

Figure 3S. The slopes from a simple linear model (*A*) and a robust linear model (*B*) between the average vertebrate family size (n_v) against the average invertebrate family size (n_i) for each GOSlim cellular component category in the *V.MCL* gene families. The error bar represents a 95% confidence interval of a slope from 1,000 bootstrap replicates. The slope from the original data and the mean of the bootstrapped slopes for each category are indicated by the circular and triangular points, respectively. The x-axis shows the cellular component categories with the *p*-value < 0.01 for a null hypothesis of the estimated slopes = 0 in either the simple linear model or the robust linear model. These categories are ordered by the mean of their bootstrapped slopes in panel *A*. Most of the proteins with GOSlim in the "cellular_component" category are those in extracellular matrix, synapse, viral capsid, and viral envelope. Furthermore, most of these gene families are also assigned to other GOSlim categories (mainly to cell, cytoplasmic membrane-bound vesicle, plasma membrane, cytoplasm, extracellular space or extracellular region).

Figure 4S. Functional bias in each GOSlim biological process category is shown for the *V.MCL* and *Vonly* groups (represented, respectively, by the circular and triangular points) in the top panel and for the *V.A50* and *VFProk* gene groups (represented, respectively, by the diamond and square points) in the bottom panel. The significant

functional bias categories at the 5% false discovery rate are marked by red S. The magnitude for the under- or over-represented gene copies (below or above 0, respectively) from the overall average is indicated on the y-axis. Most of the proteins with GOSlim in the "biological_process" category are those involved in cell adhesion, response to stimulus, sensory perception of smell, immune response, homophilic cell adhesion, and defense response. Furthermore, most of these gene families are also assigned to other GOSlim categories (mainly to signal transduction, regulation of biological process, response to stress or multicellular organismal development).

Figure 5S. Functional bias in each GOSlim cellular compartment category is shown for the *V.MCL* and *Vonly* groups (represented, respectively, by the circular and triangular points) in the top panel and for the *V.A50* and *VFProk* gene groups (represented, respectively, by the diamond and square points) in the bottom panel. The significant functional bias categories at the 5% false discovery rate are marked by red S. The magnitude for the under- or over-represented families (below or above 0, respectively) from the overall average is indicated on the y-axis. Most of the proteins with GOSlim in the "cellular_component" category are those in extracellular matrix, synapse, viral capsid, and viral envelope. Furthermore, most of these gene families are also assigned to other GOSlim categories (mainly to cell, cytoplasmic membrane-bound vesicle, plasma membrane, cytoplasm, extracellular space or extracellular region).

Figure 6S. Functional bias in each GOSlim cellular component category is shown for the *V.MCL* and *Vonly* groups (represented, respectively, by the circular and triangular points) in the top panel and for the *V.A50* and *VFProk* gene groups (represented, respectively, by the diamond and square points) in the bottom panel. The significant functional bias categories at the 5% false discovery rate are marked by red S. The magnitude for the under- or over-represented gene copies (below or above 0, respectively) from the overall average is indicated on the y-axis. Most of the proteins with GOSlim in the "cellular_component" category are those in extracellular matrix, synapse, viral capsid, and viral envelope. Furthermore, most of these gene families are also assigned to other GOSlim categories (mainly to cell, cytoplasmic membrane-bound vesicle, plasma membrane, cytoplasm, extracellular space or extracellular region).

Figure 7S. Functional bias in each GOSlim biological process category is shown for the *I.MCL*, *I.A50* and *Ionly* groups (represented, respectively, by the circular, diamond and triangular points). The significant functional bias categories at the 5% false discovery rate are marked by red S. The magnitude for the under- or over-represented families (below or above 0, respectively) from the overall average is indicated on the y-axis. Most of the proteins with GOSlim in the "biological_process" category are those involved in cell adhesion, homoiothermy, sensory perception of smell, homophilic cell adhesion, immune response, response to stimulus, ciliary or flagellar motility, defense response, and visual perception. Furthermore, most of these gene families are also assigned to other GOSlim categories (mainly to response to stress, response to abiotic stimulus, signal transduction, regulation of biological process, multicellular organismal development, cell differentiation, response to external stimulus, or anatomical structure morphogenesis).

Figure 8S. Functional bias in each GOSlim biological process category is shown for the *I.MCL*, *I.A50* and *Ionly* groups (represented, respectively, by the circular, diamond and triangular points). The significant functional bias categories at the 5% false discovery rate are marked by red S. The magnitude for the under- or over-represented gene copies (below or above 0, respectively) from the overall average is indicated on the y-axis. Most of the proteins with GOSlim in the "biological_process" category are those involved in cell adhesion, homoiothermy, sensory perception of smell, homophilic cell adhesion, immune response, response to stimulus, ciliary or flagellar motility, defense response, and visual perception. Furthermore, most of these gene families are also assigned to other GOSlim categories (mainly to response to stress, response to abiotic stimulus, signal transduction, regulation of biological process, multicellular organismal development, cell differentiation, response to external stimulus, or anatomical structure morphogenesis).

Figure 9S. Functional bias in each GOSlim cellular compartment category is shown for the *I.MCL*, *I.A50* and *Ionly*groups (represented, respectively, by the circular, diamond and triangular points). The significant functional bias categories at the 5% false discovery rate are marked by red S. The magnitude for the under- or over-represented families (below or above 0, respectively) from the overall average is

indicated on the y-axis. Most of the proteins with GOSlim in the "cellular_component" category are those in extracellular matrix, viral nucleocapsid, viral capsid, synapse and viral envelope. Less than half of these gene families are also assigned to other GOSlim categories (e.g., cytoplasm, cell, extracellular region, plasma membrane, lysosome, intracellular, or proteinaceous extracellular matrix).

Figure 10S. Functional bias in each GOSlim cellular compartment category is shown for the *I.MCL*, *I.A50* and *Ionly*groups (represented, respectively, by the circular, diamond and triangular points). The significant functional bias categories at the 5% false discovery rate are marked by red S. The magnitude for the under- or over-represented gene copies (below or above 0, respectively) from the overall average is indicated on the y-axis. Most of the proteins with GOSlim in the "cellular_component" category are those in extracellular matrix, viral nucleocapsid, viral capsid, synapse and viral envelope. Less than half of these gene families are also assigned to other GOSlim categories (e.g., cytoplasm, cell, extracellular region, plasma membrane, lysosome, intracellular, or proteinaceous extracellular matrix).

Figure 1S.

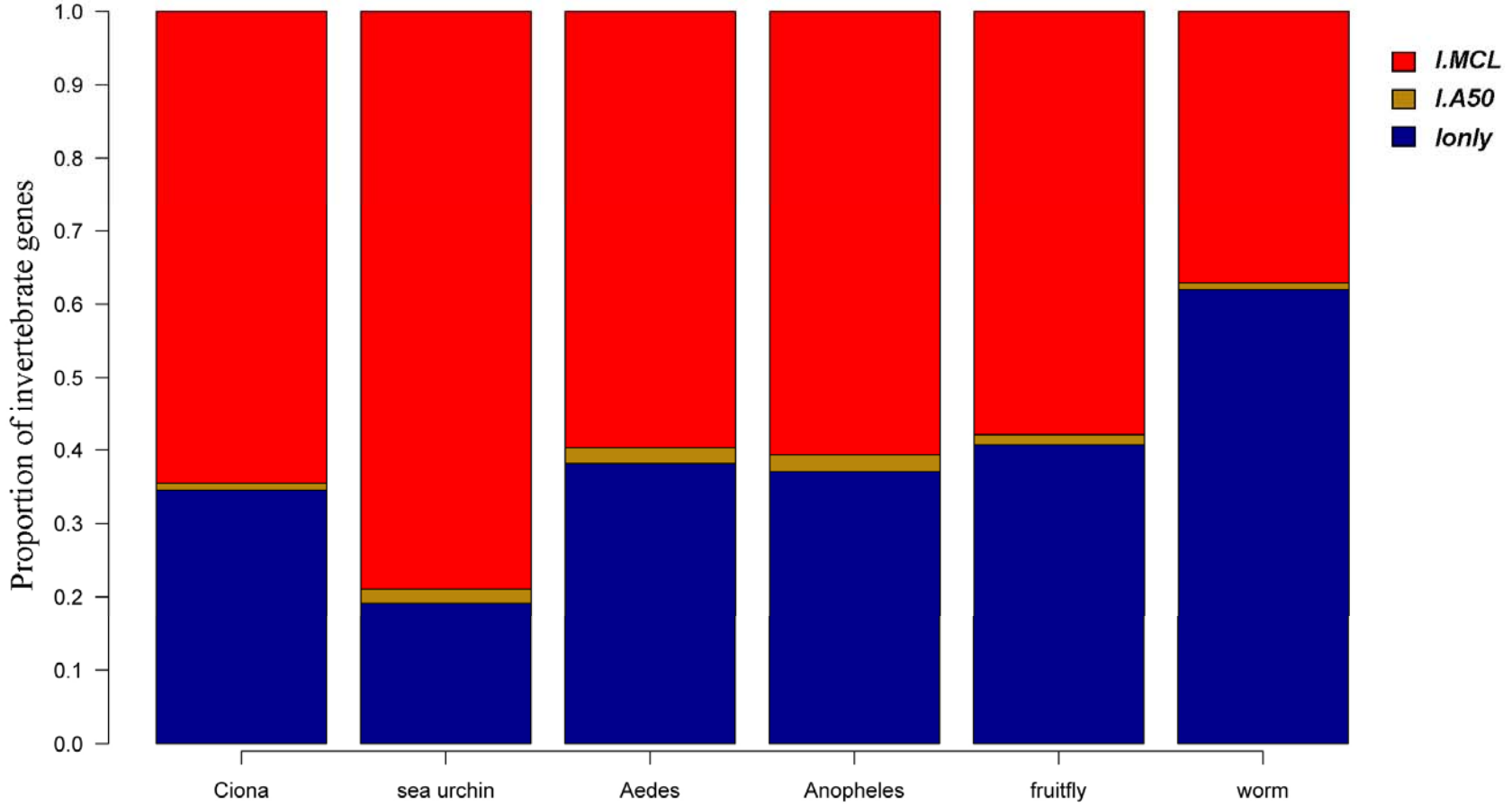

Proportion of invertebrate genes
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Ciona
sea urchin
Aedes
Anopheles
fruitfly
worm
I.MCL
I.A50
Ionly

Figure 2S.

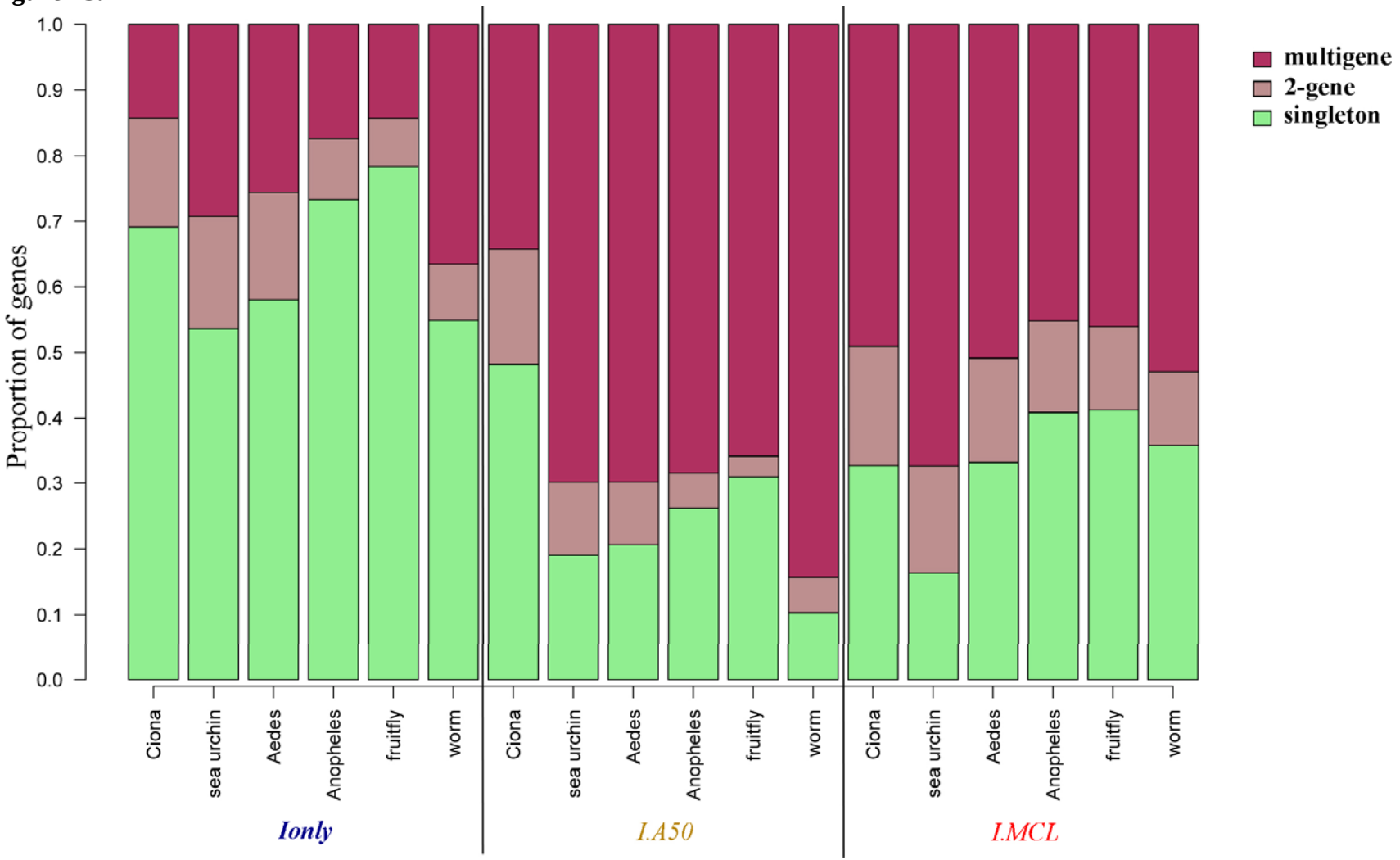

Proportion of genes
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
multigene
2-gene
singleton
Ciona
sea urchin
Aedes
Anopheles
fruitfly
worm
Ionly
I.A50
I.MCL

Figure 3S.

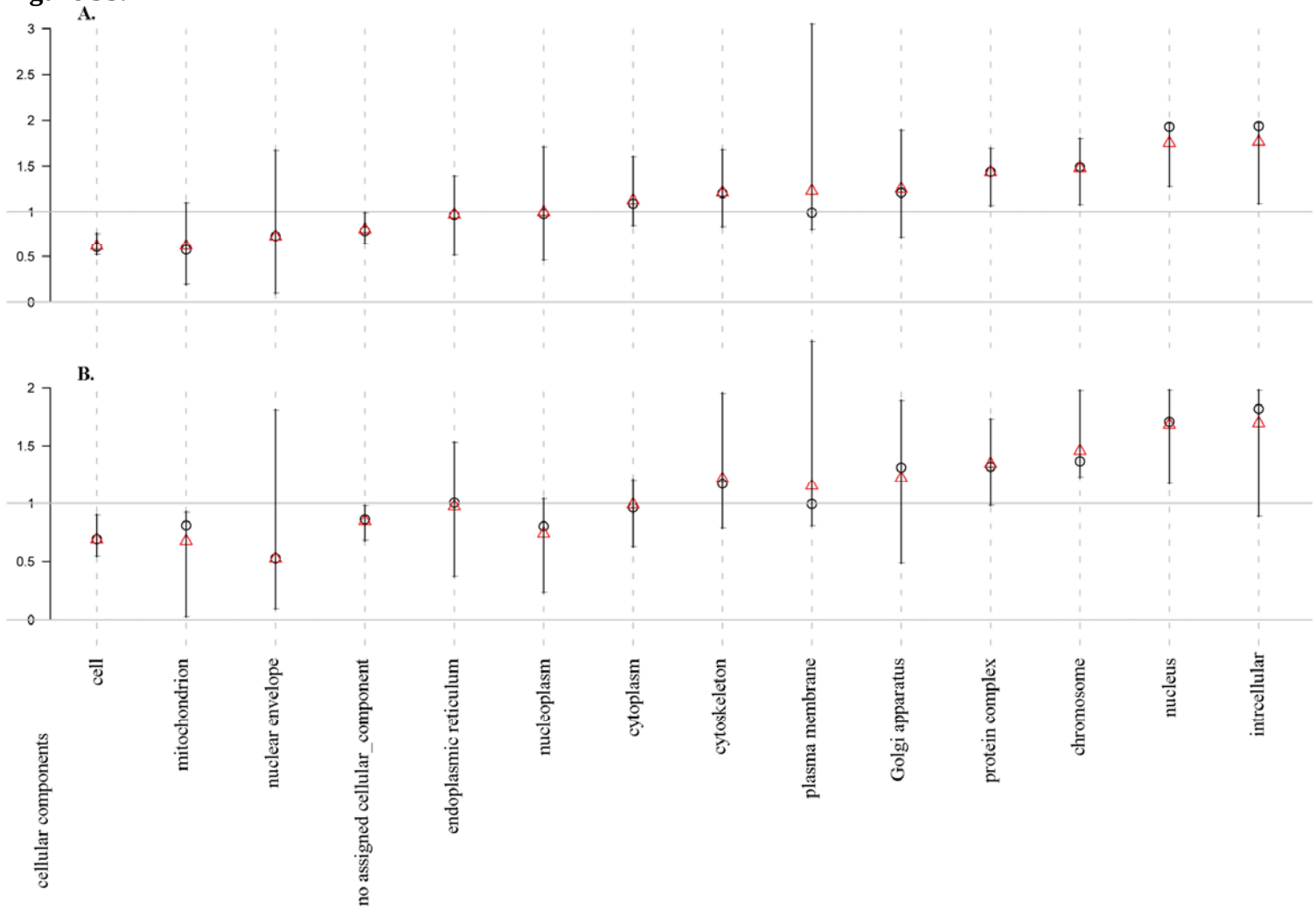

A.
3
2.5
2
1.5
1
0.5
0
B.
2
1.5
1
0.5
0
cellular components
cell
mitochondrion
nuclear envelope
no assigned cellular_component
endoplasmic reticulum
nucleoplasm
cytoplasm
cytoskeleton
plasma membrane
Golgi apparatus
protein complex
chromosome
nucleus
intreellular

Figure 4S.

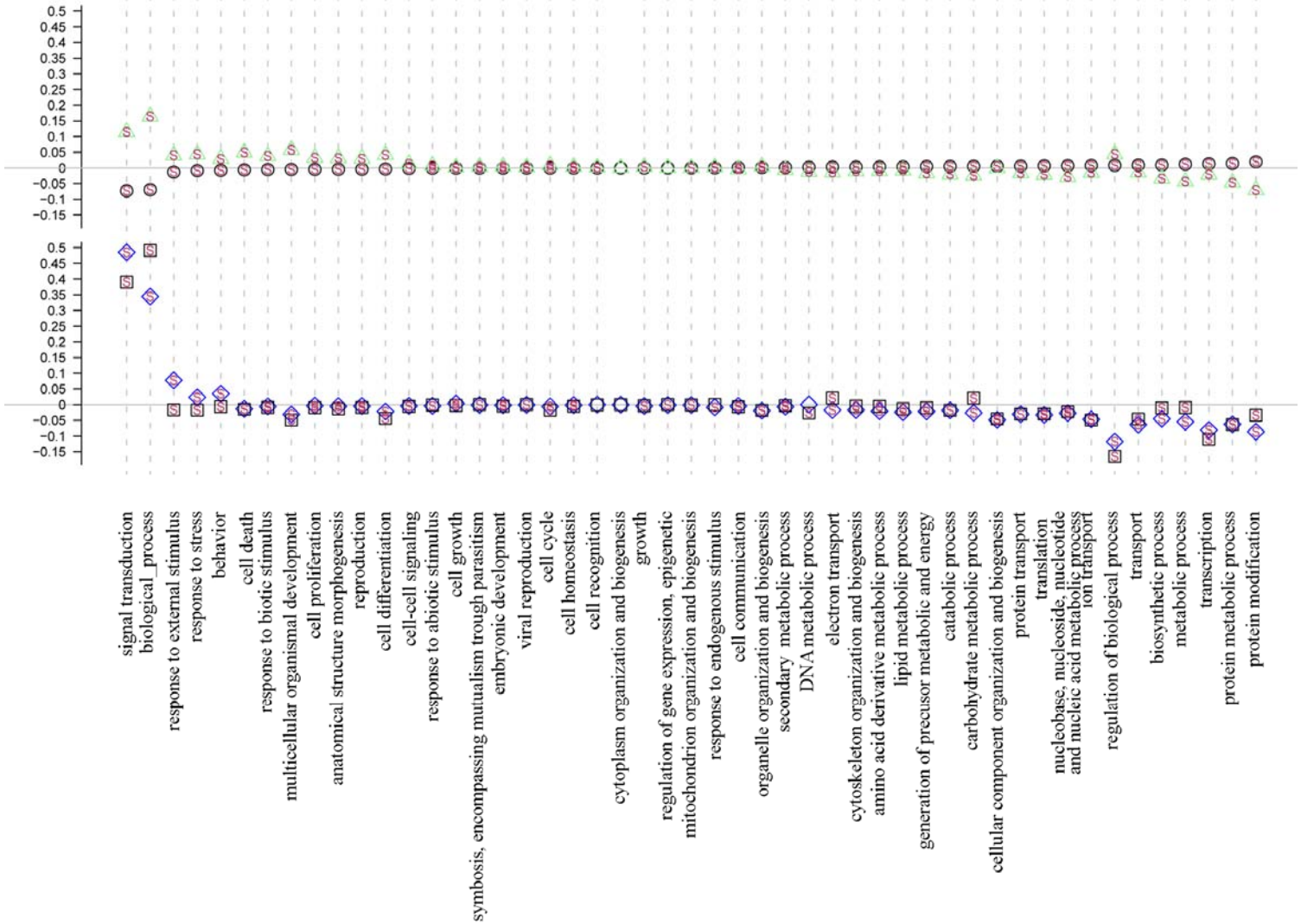

0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−0.05
−0.1
−0.15
signal transduction
biological_process
response to external stimulus
response to stress
behavior
cell death
response to biotic stimulus
multicellular organismal development
cell proliferation
anatomical structure morphogenesis
reproduction
cell differentiation
cell-cell signaling
response to abiotic stimulus
cell growth
symbosis, encompassing mutualism trough parasitism
embryonic development
viral reproduction
cell cycle
cell homeostasis
cell recognition
cytoplasm organization and biogenesis
growth
regulation of gene expression, epigenetic
mitochondrion organization and biogenesis
response to endogenous stimulus
cell communication
organelle organization and biogenesis
secondary metabolic process
DNA metabolic process
electron transport
cytoskeleton organization and biogenesis
amino acid derivative metabolic process
lipid metabolic process
generation of precusor metabolic and energy
catabolic process
carbohydrate metabolic process
cellular component organization and biogenesis
protein transport
translation
nucleobase, nucleoside, nucleotide
and nucleic acid metabolic process
ion transport
regulation of biological process
transport
biosynthetic process
metabolic process
transcription
protein metabolic process
protein modification

Figure 5S.

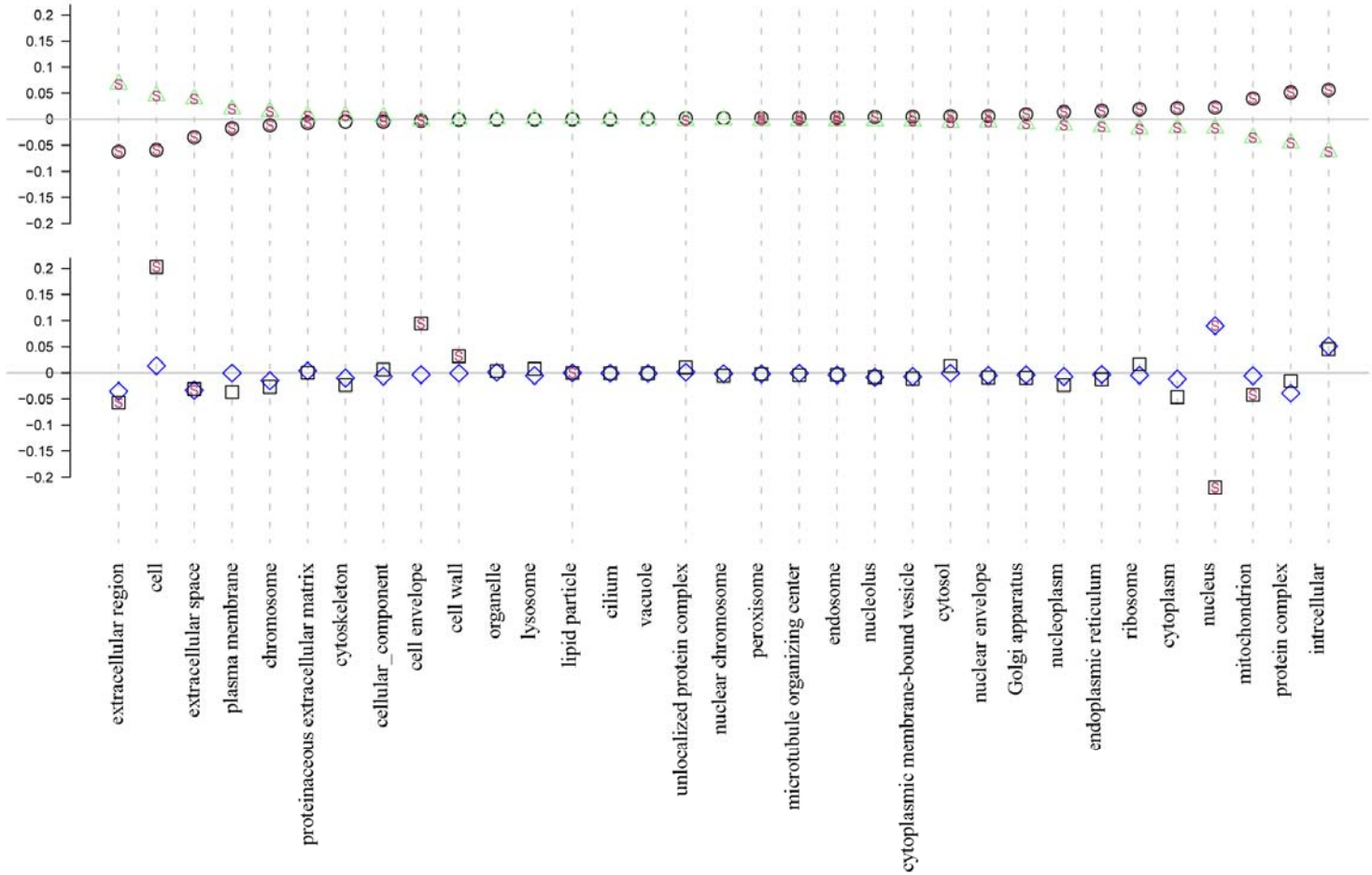

0.2
0.15
0.1
0.05
0
−0.05
−0.1
−0.15
−0.2
extracellular region
cell
extracellular space
plasma membrane
chromosome
proteinaceous extracellular matrix
cytoskeleton
cellular_component
cell envelope
cell wall
organelle
lysosome
lipid particle
cilium
vacuole
unlocalized protein complex
nuclear chromosome
peroxisome
microtubule organizing center
endosome
nucleolus
cytoplasmic membrane-bound vesicle
cytosol
nuclear envelope
Golgi apparatus
nucleoplasm
endoplasmic reticulum
ribosome
cytoplasm
nucleus
mitochondrion
protein complex
intrcellular

Figure 6S.


0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−0.05
−0.1
−0.15
−0.2
cell
extracellular region
extracellular space
plasma membrane
organelle
cell envelope
lysosome
proteinaceous extracellular matrix
lipid particle
cell wall
cilium
vacuole
nuclear chromosome
microtubule organizing center
cellular_component
nucleolus
endosome
peroxisome
unlocalized protein complex
cytosol
nuclear envelope
cytoplasmic membrane-bound vesicle
Golgi apparatus
nucleoplasm
chromosome
endoplasmic reticulum
cytoskeleton
cytoplasm
ribosome
mitochondrion
protein complex
intrcellular
nucleus

Figure 7S.

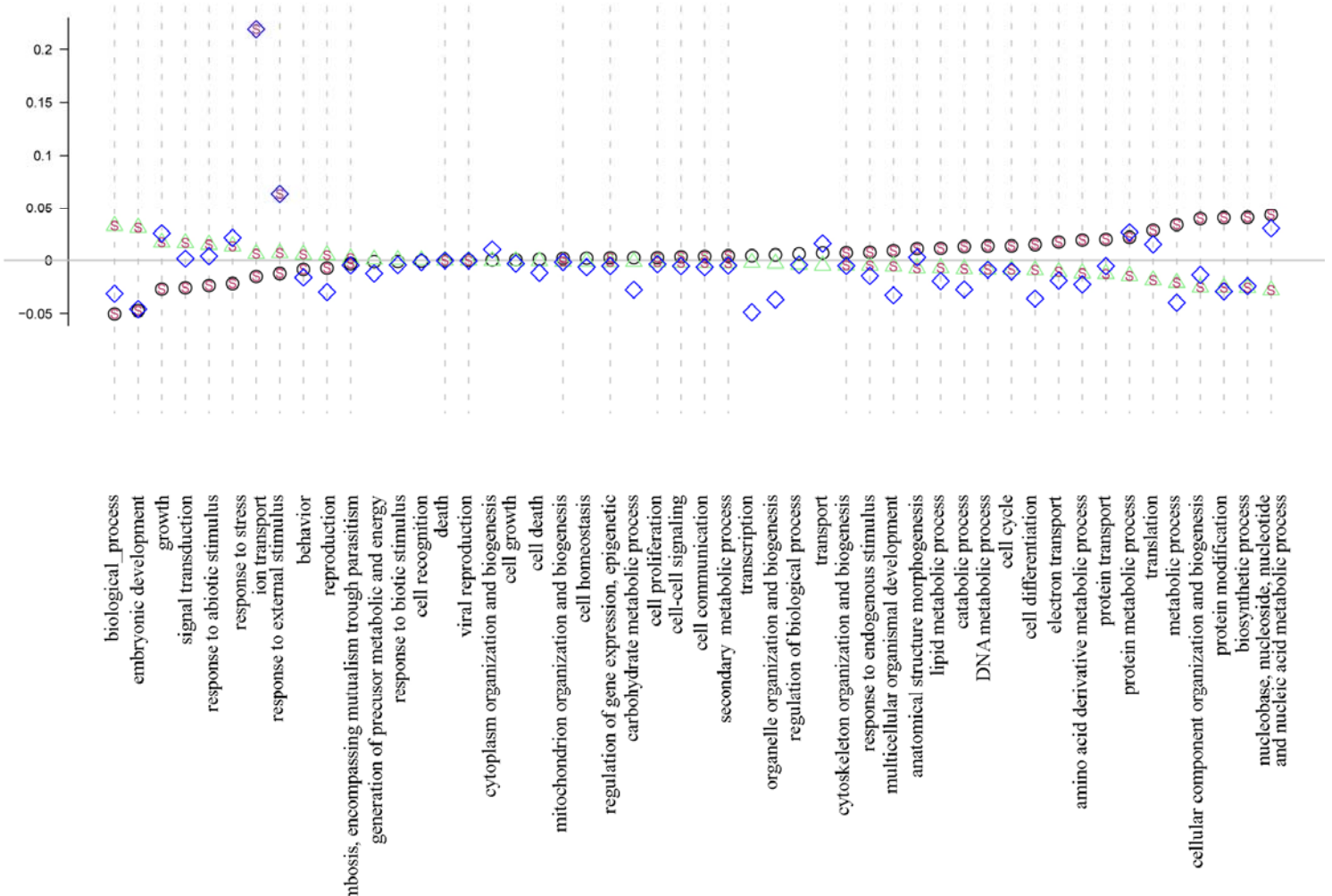

0.2
0.15
0.1
0.05
0
−0.05
biological_process
embryonic development
growth
signal transduction
response to abiotic stimulus
response to stress
ion transport
response to external stimulus
behavior
reproduction
symbosis, encompassing mutualism trough parasitism
generation of precusor metabolic and energy
response to biotic stimulus
cell recognition
death
viral reproduction
cytoplasm organization and biogenesis
cell growth
cell death
mitochondrion organization and biogenesis
cell homeostasis
regulation of gene expression, epigenetic
carbohydrate metabolic process
cell proliferation
cell-cell signaling
cell communication
secondary metabolic process
transcription
organelle organization and biogenesis
regulation of biological process
transport
cytoskeleton organization and biogenesis
response to endogenous stimulus
multicellular organismal development
anatomical structure morphogenesis
lipid metabolic process
catabolic process
DNA metabolic process
cell cycle
cell differentiation
electron transport
amino acid derivative metabolic process
protein transport
protein metabolic process
translation
metabolic process
cellular component organization and biogenesis
protein modification
biosynthetic process
nucleobase, nucleoside, nucleotide
and nucleic acid metabolic process

Figure 8S.

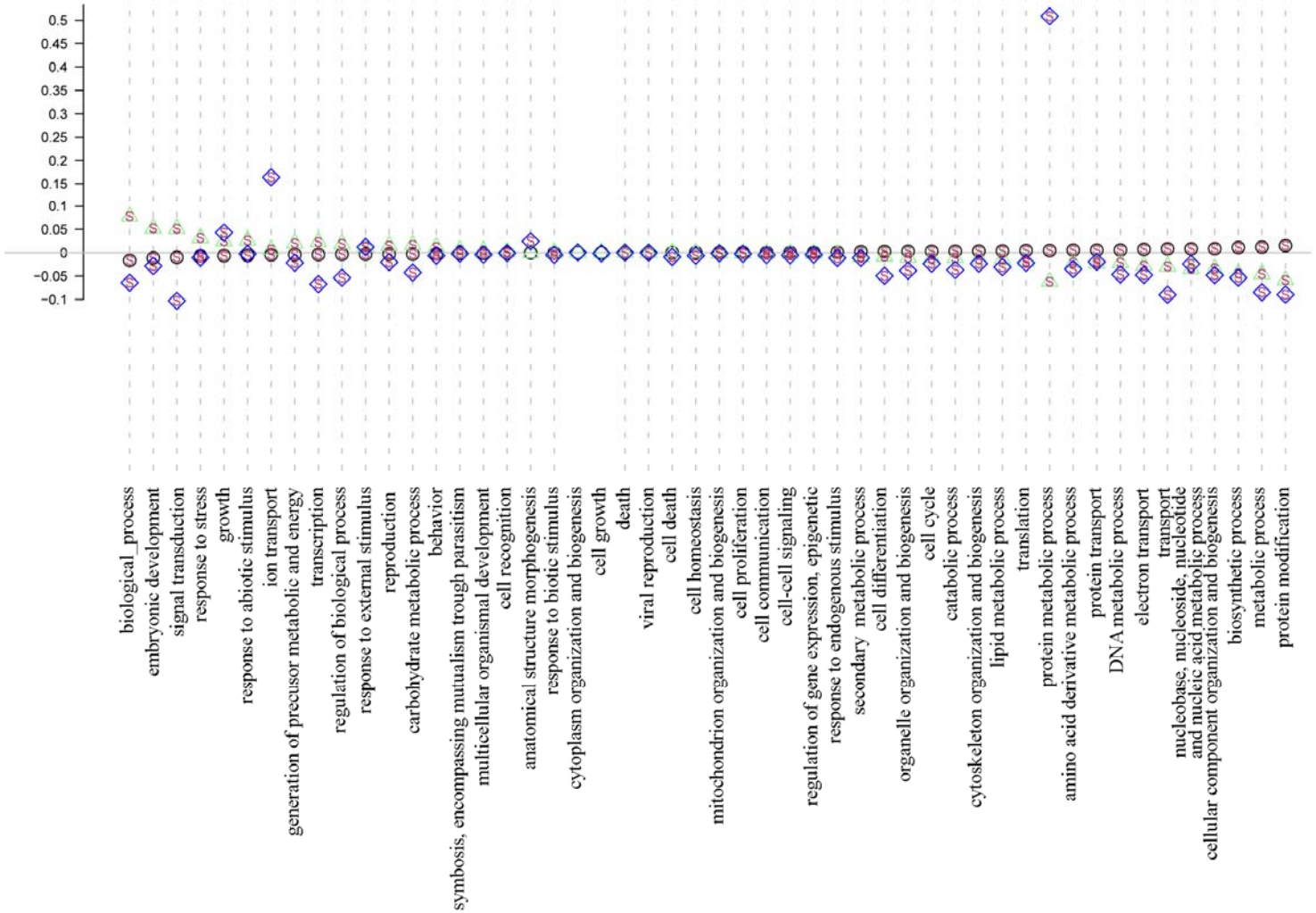

0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−0.05
−0.1
biological_process
embryonic development
signal transduction
response to stress
growth
response to abiotic stimulus
ion transport
generation of precusor metabolic and energy
transcription
regulation of biological process
response to external stimulus
reproduction
carbohydrate metabolic process
behavior
symbosis, encompassing mutualism trough parasitism
multicellular organismal development
cell recognition
anatomical structure morphogenesis
response to biotic stimulus
cytoplasm organization and biogenesis
cell growth
death
viral reproduction
cell death
cell homeostasis
mitochondrion organization and biogenesis
cell proliferation
cell communication
cell-cell signaling
regulation of gene expression, epigenetic
response to endogenous stimulus
secondary metabolic process
cell differentiation
organelle organization and biogenesis
cell cycle
catabolic process
cytoskeleton organization and biogenesis
lipid metabolic process
translation
protein metabolic process
amino acid derivative metabolic process
protein transport
DNA metabolic process
electron transport
transport
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
cellular component organization and biogenesis
biosynthetic process
metabolic process
protein modification

Figure 9S.

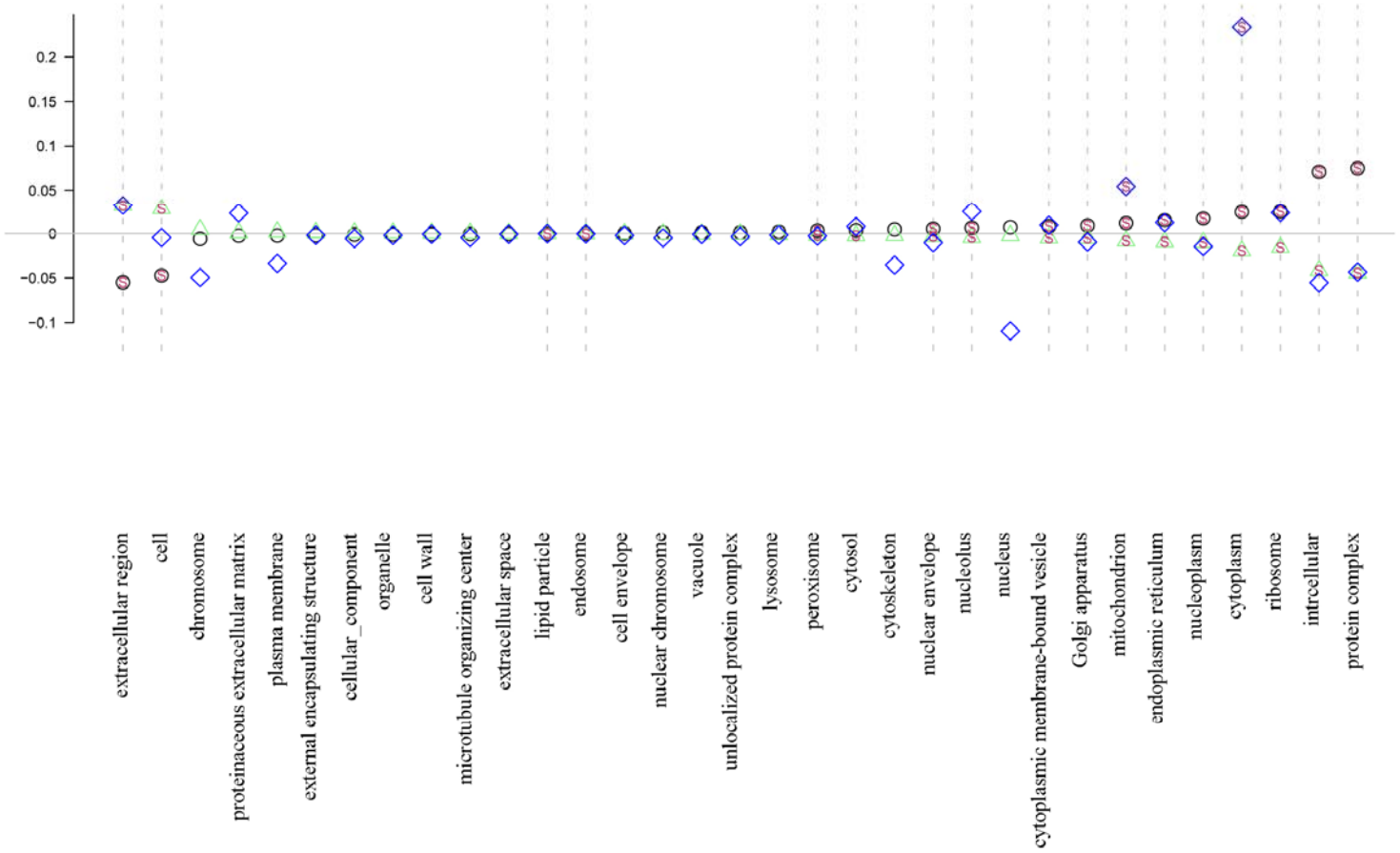
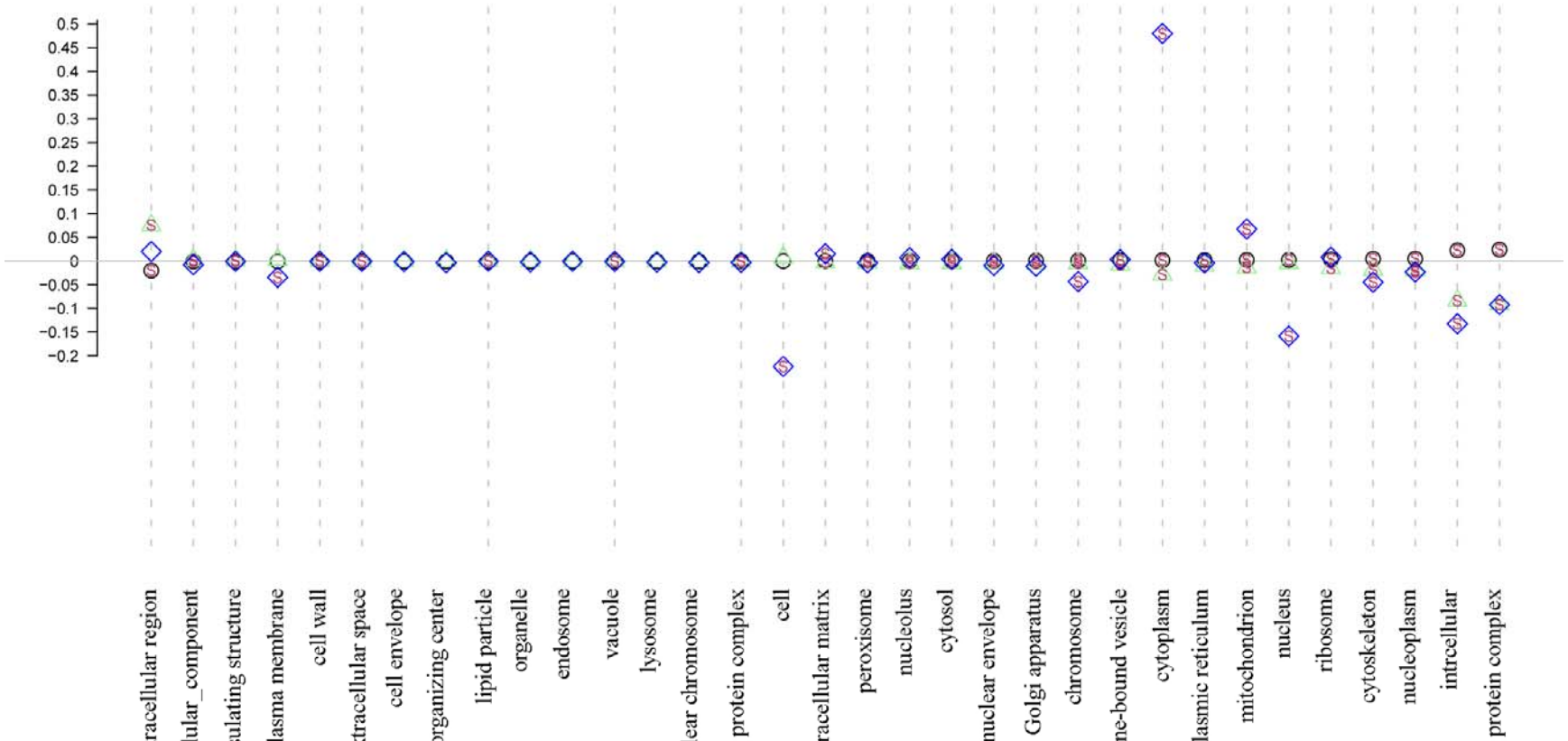

0.2
0.15
0.1
0.05
0
−0.05
−0.1
extracellular region
cell
chromosome
proteinaceous extracellular matrix
plasma membrane
external encapsulating structure
cellular_component
organelle
cell wall
microtubule organizing center
extracellular space
lipid particle
endosome
cell envelope
nuclear chromosome
vacuole
unlocalized protein complex
lysosome
peroxisome
cytosol
cytoskeleton
nuclear envelope
nucleolus
nucleus
cytoplasmic membrane-bound vesicle
Golgi apparatus
mitochondrion
endoplasmic reticulum
nucleoplasm
cytoplasm
ribosome
intrcellular
protein complex

Figure 10S.


0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−0.05
−0.1
−0.15
−0.2
extracellular region
cellular_component
external encapsulating structure
plasma membrane
cell wall
extracellular space
cell envelope
microtubule organizing center
lipid particle
organelle
endosome
vacuole
lysosome
nuclear chromosome
unlocalized protein complex
cell
proteinaceous extracellular matrix
peroxisome
nucleolus
cytosol
nuclear envelope
Golgi apparatus
chromosome
cytoplasmic membrane-bound vesicle
cytoplasm
endoplasmic reticulum
mitochondrion
nucleus
ribosome
cytoskeleton
nucleoplasm
intrcellular
protein complex