

Universality of the DNA methylation codes in Eucaryotes

Benoît Aliaga¹, Ingo Bulla^{1,2,3}, Gabriel Mouahid¹, David Duval¹ and Christoph Grunau^{1*}

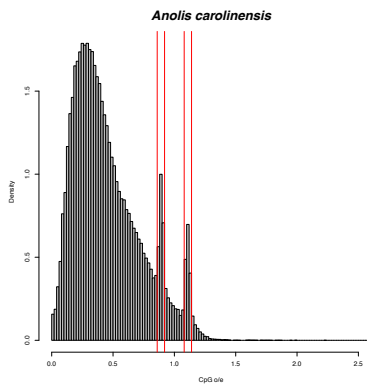
(1) Univ. Perpignan Via Domitia, IHPE UMR 5244, CNRS, IFREMER,
Univ. Montpellier, F-66860 Perpignan, France

(2) Institute for Mathematics and Informatics, University of Greifswald,
Greifswald, Germany

(3) Department of Computer Science, ETH Zürich, Zürich, Switzerland

Supplementary figures

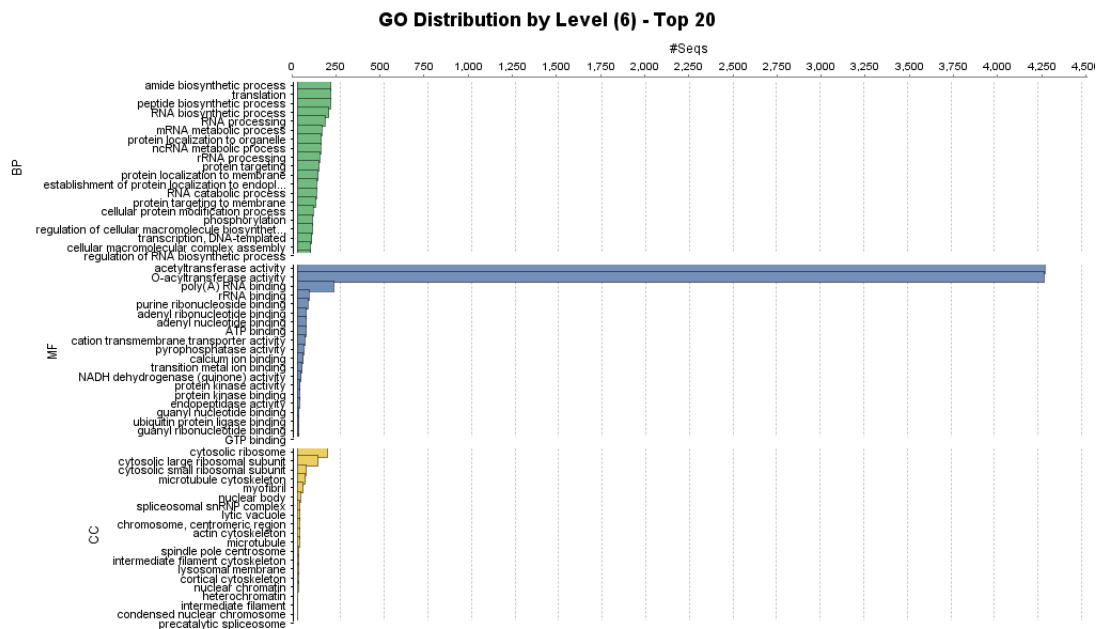
Supplementary figure 1 : DNA contamination in *Anolis carolinensis*



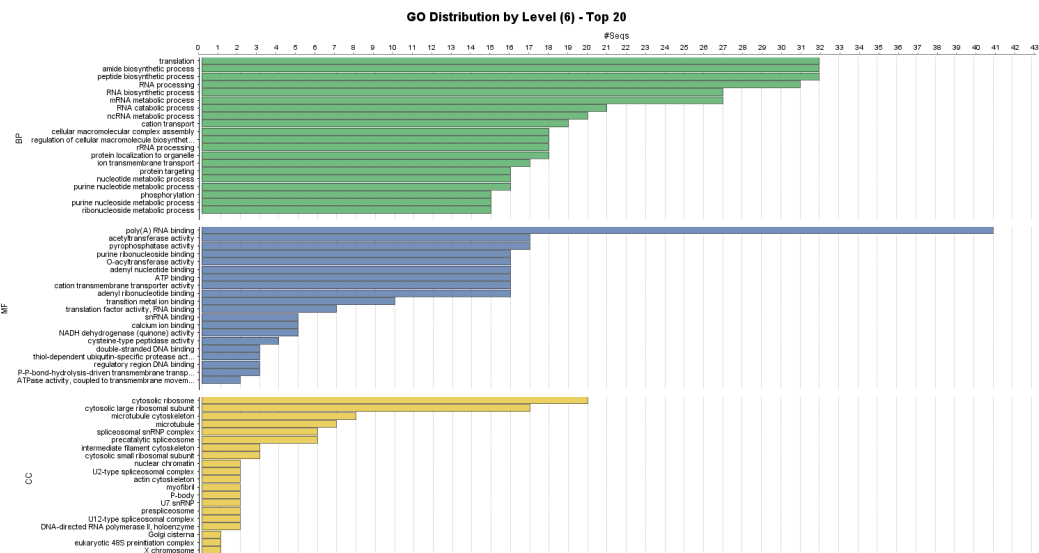
Supplementary figure 1.1: the CpGo/e profile of *Anolis carolinensis* from dbEST and its two additional peaks (peak 01-left and peak 02-right)

When we compared the dbEST profile with CleanEST and CDS profiles, two additional peaks occurred in the dbEST profile. We reasoned that the sequences present in the two peaks were not from *Anolis carolinensis* but were contaminant. In order to verify this hypothesis, we did a gene ontology research in these two additional peaks. We isolated and extracted DNA sequences from the dbEST fasta files. We used Blast2go for gene ontology search¹.

The first peak (7,030 sequences with a CpGo/e ratio between 0.92 and 1.08) contains a chloramphenicol O-acetyltransferase used in bacterial cloning vectors. The second peak (4,922 sequences with a CpGo/e ratio between 1.14 and 1.22) present homologies with sequences from apicomplexans (plasmodium), and platyhelminths suggesting presence of such parasites in the initial biological sample.

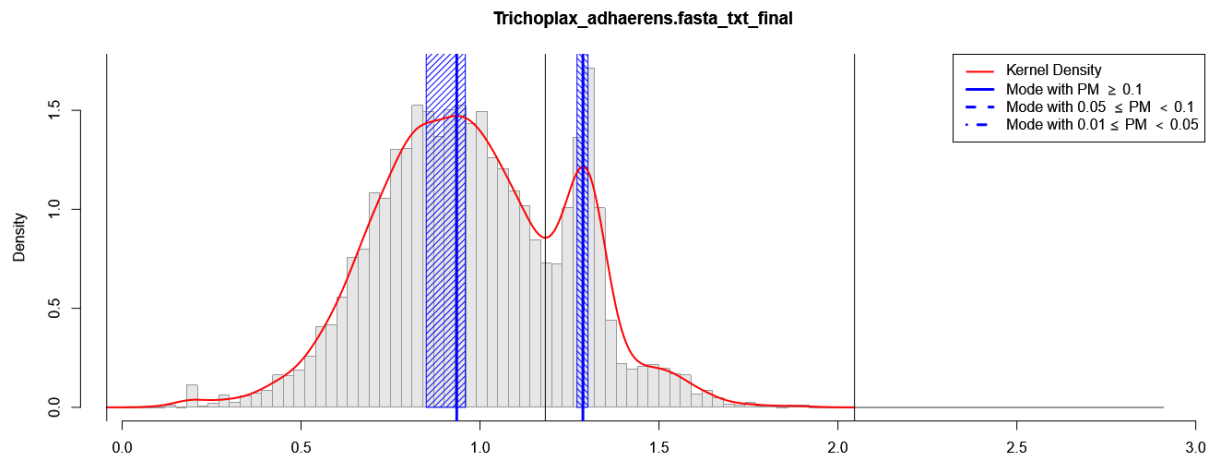


Supplementary figure 1.2: Gene ontology distribution in the Peak 01 (0.92-1.08).



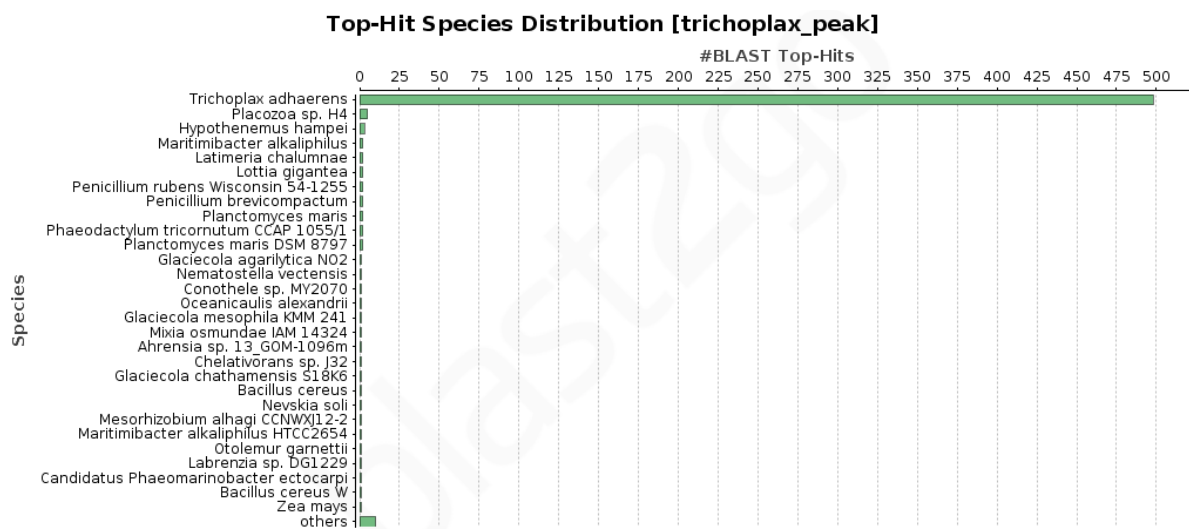
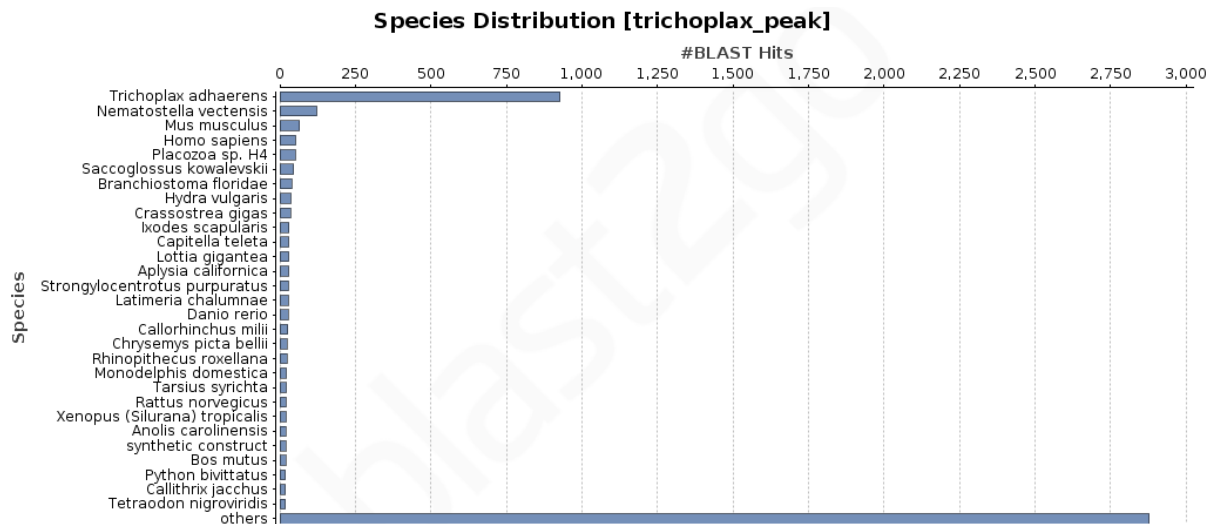
Supplementary figure 1.2: Gene ontology distribution for the peak O2 (1.14-1.22)

Supplementary figure 2: symbiont detection in *Trichoplax adhaerens*



Supplementary figure 2.1: the CpGo/e profile of *Trichoplax adhaerens* from dbEST and its additional peak (CpGo/e 1.20 - 1.40)

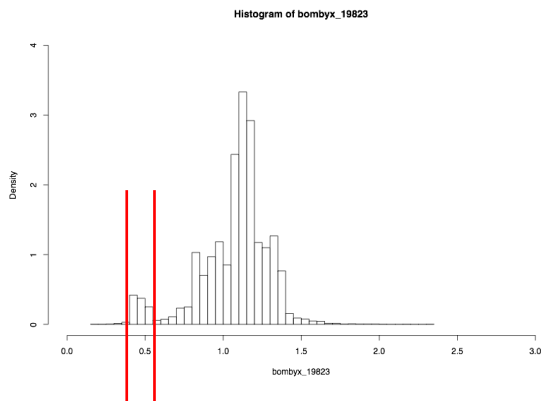
When we compared the dbEST profile with CDS profile, one additional peak appeared in the dbEST profile (1,609 sequences with a CpGo/e ratio between 1.20 and 1.40). We suggested that sequences present in this peak were not from *Trichoplax adhaerens* but were contaminants from intracellular bacteria². In order to verify this hypothesis, we did a gene ontology research in this additional peak. We isolated and extracted DNA sequences from the dbEST fasta files and confirmed our hypothesis.



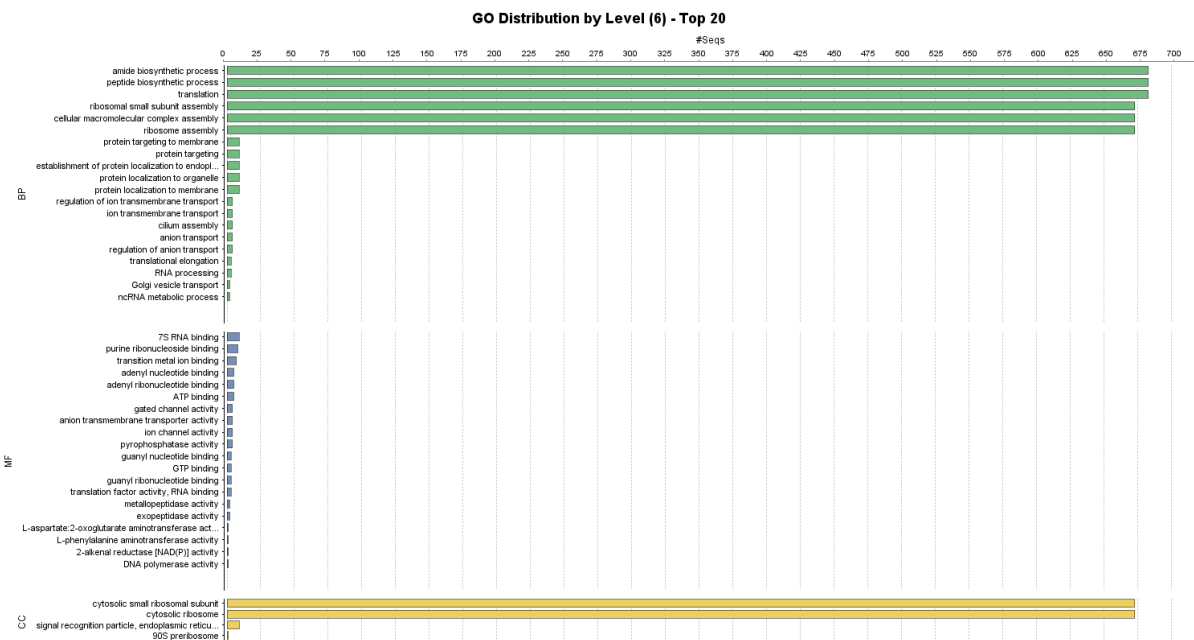
Supplementary figure 2.2: distribution of species distribution in the additional peak in the ESTs of *Trichoplax adhaerens* at a CpGo/e ratio of 1.14 - 1.22

Supplementary figure 3: transcript detection in a *Bombyx mori* ovary library

When we compare the cleanEST library #19823 (*Bombyx mori* ovary tissue) profile with the CDS profile, one additional peak was detected with a CpGo/e ratio between 0.40 and 0.60. We suspected the sequences present in this additional peak were contaminants. In order to verify this hypothesis, we isolated and extracted 769 sequences incorporated in this peak and did a gene ontology research with Blast2go¹.



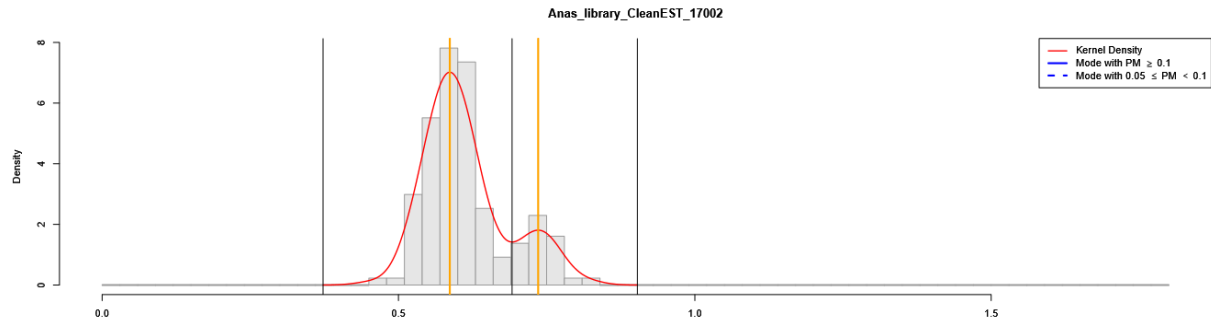
Supplementary figure 3.1: the CpGo/e profile of *Bombyx mori* from cleanEST library n°19823 and its additional peak (included between 0.40 and 0.60)



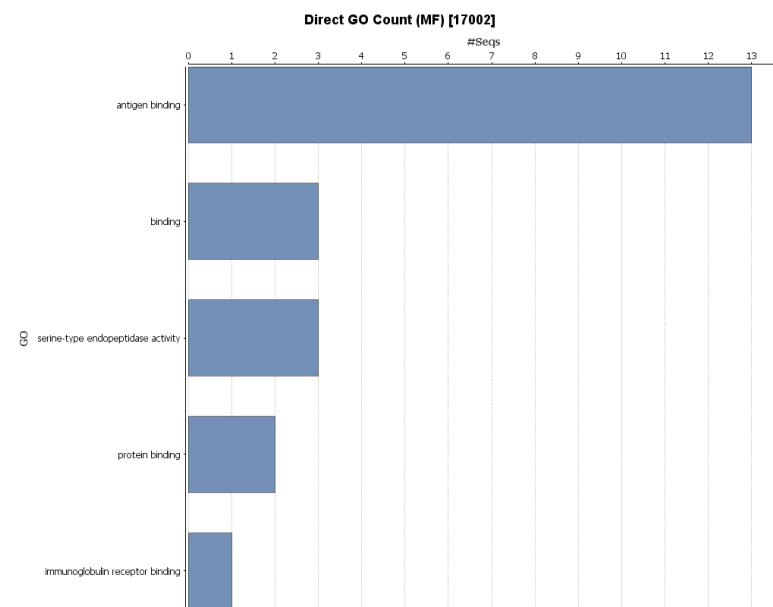
Supplementary figure 3.2: Gene ontology distribution for the additional peak (incorporated between a CpGo/e ratio between 0.40 and 0.60) in *Bombyx mori* ovary library (cleanEST n°19823)

Supplementary figure 4: transcript from a tissue in *Anas platyrhynchos* (spleen)

This dataset is from CleanEST library #17002 and represent 146 sequences. We applied notes to this dataset and obtained a CpGo/e profile. We performed a gene ontology analysis with blast2go¹. Principal GO term was “antigen binding” suggesting a bias toward immunoglobulins in the EST data set.



Supplementary figure 4.1: the CpGo/e profile of cleanEST library #17002 (146 sequences) from a tissue in *Anas platyrhynchos* (spleen)



Supplementary figure 4.2: gene ontology of cleanEST library n°17002 (146 sequences) from a tissue in *Anas platyrhynchos* (spleen)

References

1. Conesa, A. *et al.* Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
2. Driscoll, T., Gillespie, J. J., Nordberg, E. K., Azad, A. F. & Sobral, B. W. Bacterial DNA sifted from the *Trichoplax adhaerens* (Animalia: Placozoa) genome project reveals a putative rickettsial endosymbiont. *Genome Biol. Evol.* **5**, 621–645 (2013).

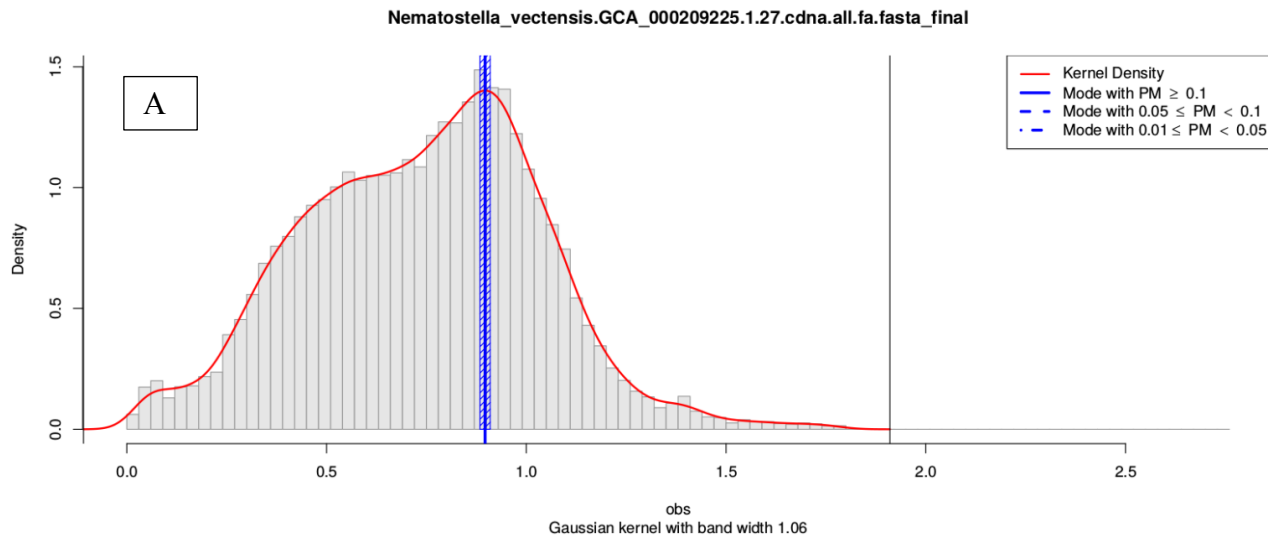
Supplementary figure 05: link between CpGo/e and gene expression level: a statistical analysis

When we compared profiles with two peaks (bimodality), we observed differences for three invertebrate and one plant species (*Crassostrea gigas*, *Nasonia vitripennis*, *Nematostella vectensis* and *Oryza sativa*) depending on the origin of the data: cDNA and dbEST/cleanEST. Since in the cDNA data set each gene is represented only once by its genomic coding sequence while in dbEST/cleanest potentially multiple sequence can exist for the same gene, we hypothesized that this could introduce a bias in the dataset by RNA abundance.

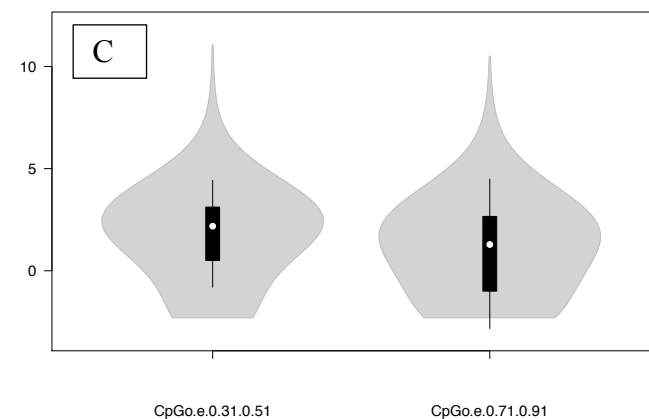
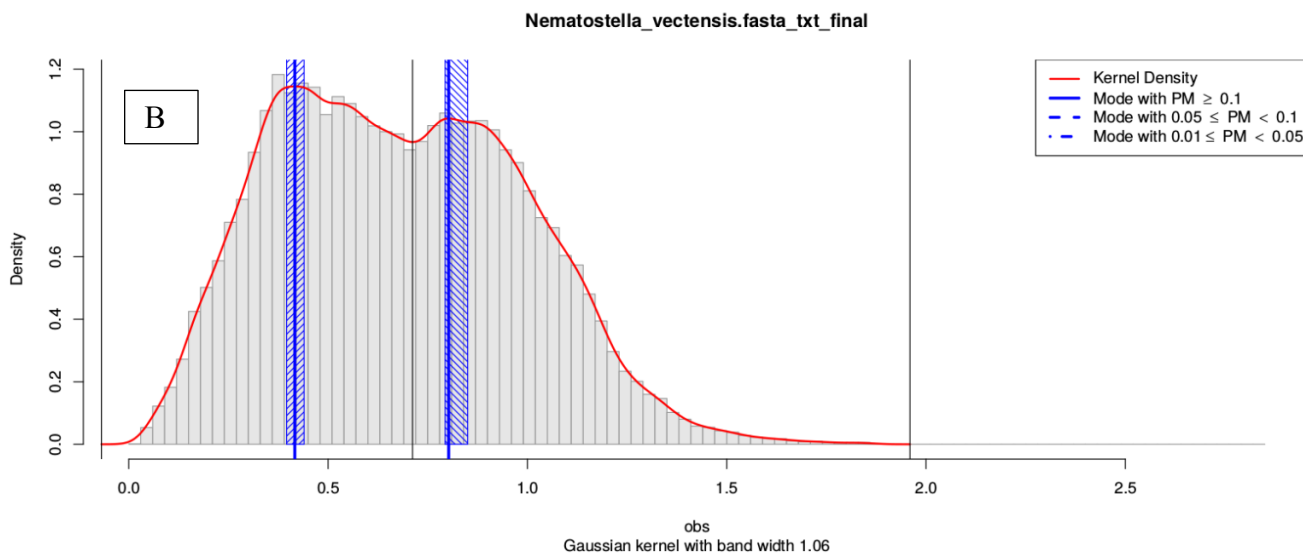
To test this hypothesis, we downloaded RNA-seq raw data from European Nucleotide Archive (<http://www.ebi.ac.uk/ena/> and/or NCBI, details in supplementary file 3). We filtered the reads with a Phred quality score ≥ 26 . The filtered reads were mapped on their references genomes with RNA STAR⁴. We estimated the FPKM (Fragments Per Kilobase Million) with Cufflinks⁵. We then compared RNA-Seq FPKM for the genes under the two modes. Our results (presented below) show that there are significant differences in FPKM for the two mode positions for *Nasonia vitripennis*, *Crassostrea gigas* and *Oryza sativa*. Genes predicted methylated with *Notos*¹ are more transcribed. This goes in line with earlier observations in other species that gene body methylation is associated with higher transcription^{2,3}. We concluded that gene expression difference is probably the origin of the bias in the EST datasets.

Supplementary figure 05. 1: Statistical analysis between CpGo/e in CDS extracted from genome and RNA seq expression level in *Nematostella vectensis*

cDNA



dbEST



CpGo/e

dbEST peak 1: 0.416

dbEST peak 2: 0.802

cDNA peak: 0.897

CpG o/e $0.31 \leq x \leq 0.51$: 3,838 genes

CpG o/e $0.71 \leq x \leq 0.91$: 5,789 genes

RNA-Seq FPKM

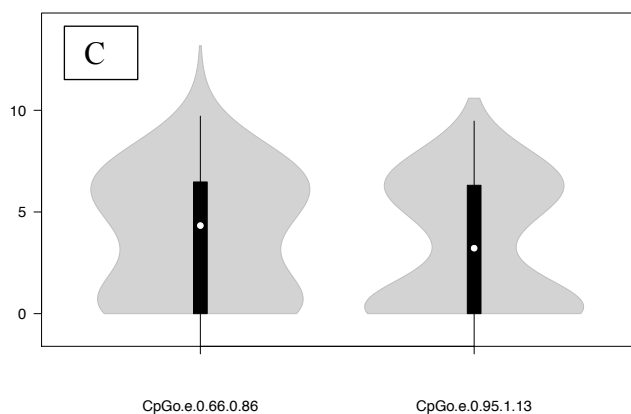
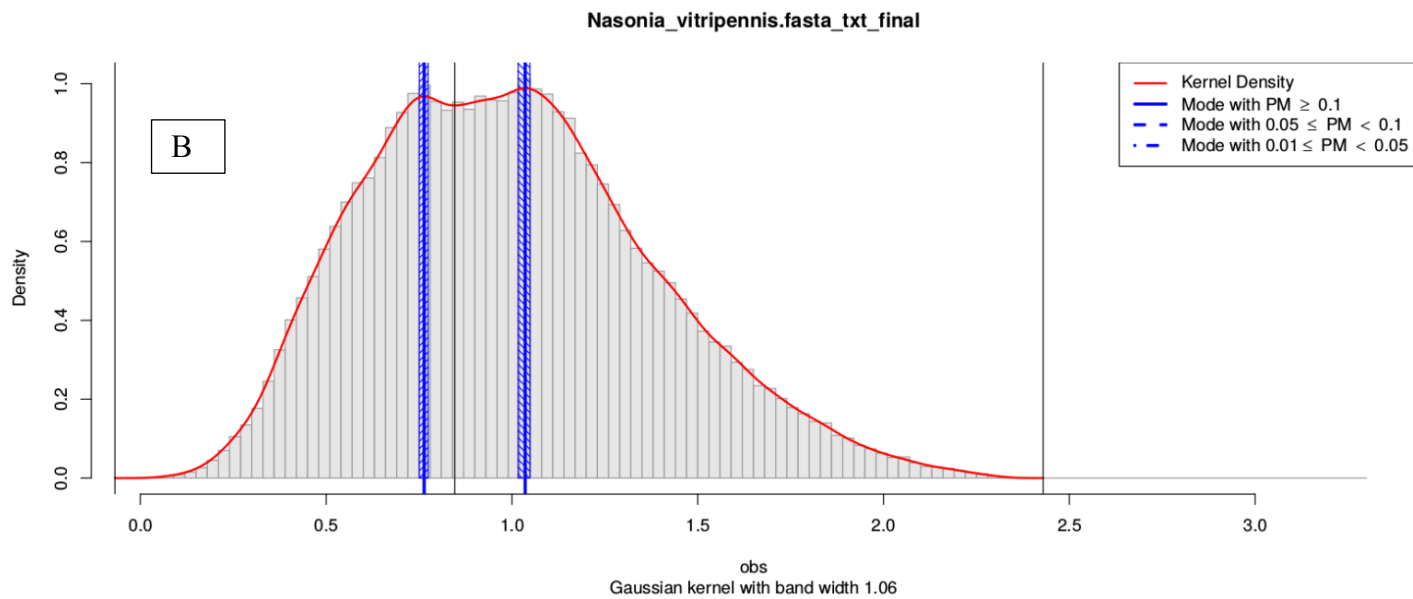
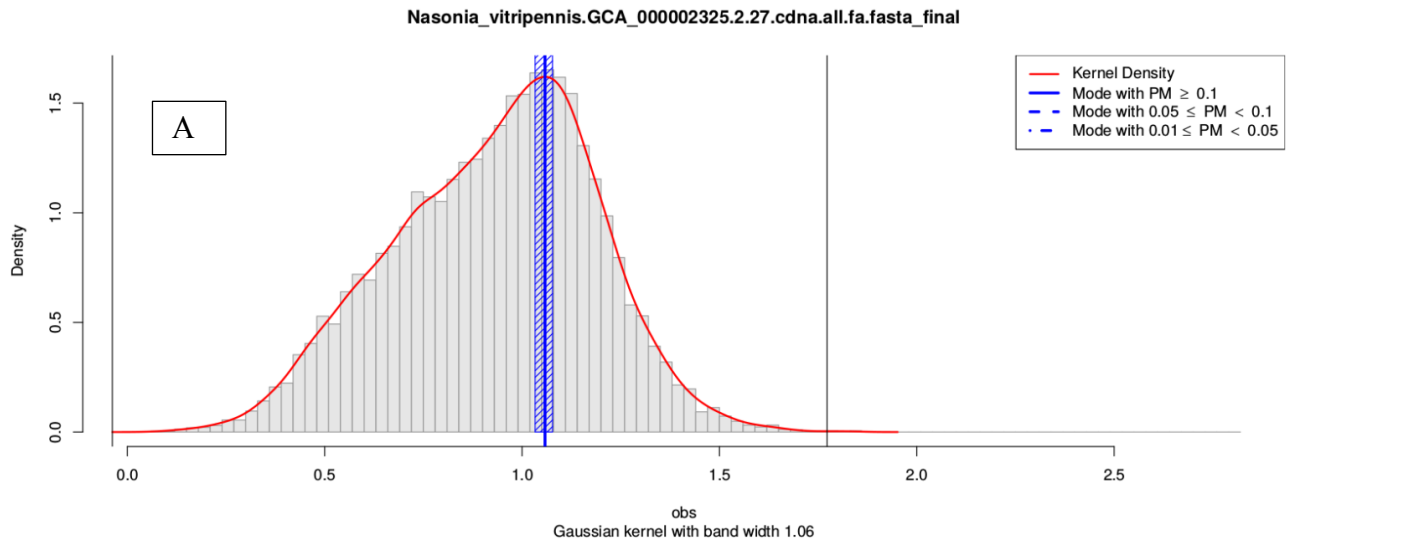
median CpG o/e $0.31 \leq x \leq 0.51$: 8.78

median CpG o/e $0.71 \leq x \leq 0.91$: 3.52

Mood's median test: p-value = 0.4677

y axis: log(FPKM)

Supplementary figure 06: RNA-seq analysis *Nasonia vitripennis*



CpGo/e

dbEST peak 1: 0.764

dbEST peak 2: 1.035

cDNA peak: 1.058

CpG o/e $0.66 \leq x \leq 0.86$: 3,571 genes

CpG o/e $0.95 \leq x \leq 1.13$: 4,799 genes

RNA-Seq FPKM

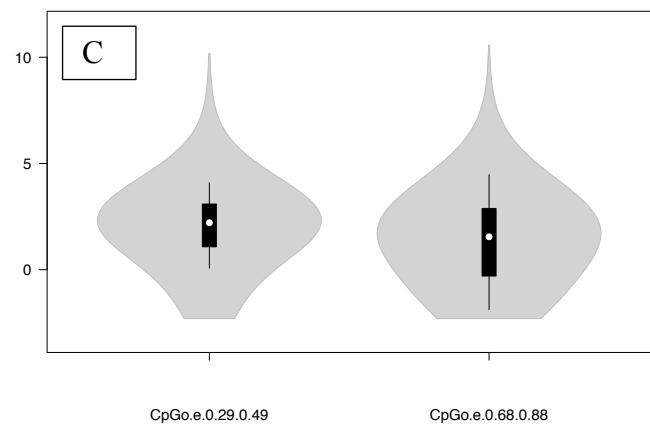
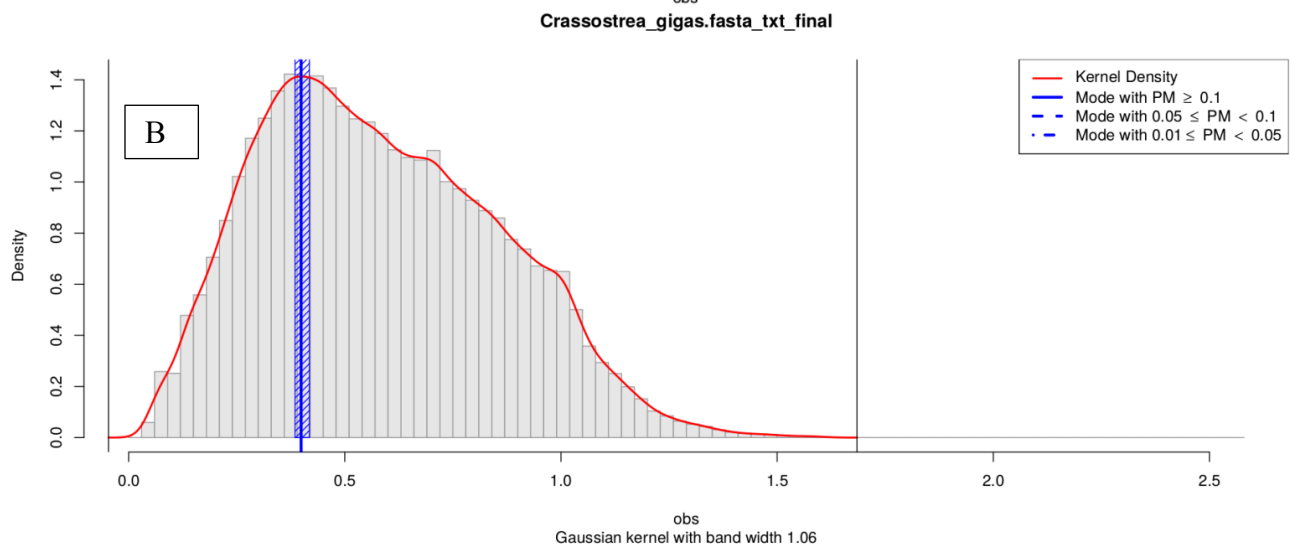
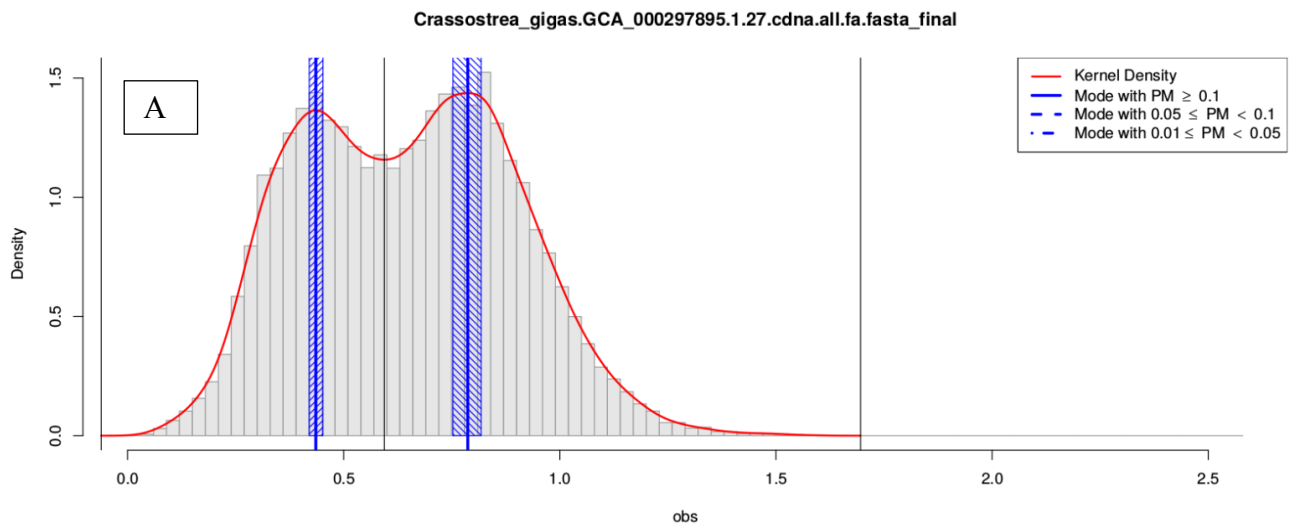
Median CpG o/e $0.66 \leq x \leq 0.86$: 75

Median CpG o/e $0.95 \leq x \leq 1.13$: 24

Mood's median test: p-value = 1.789e-05

y axis: log(FPKM)

Supplementary figure 07: RNA-seq analysis *Crassostrea gigas*



y axis: log(FPKM)

CpGo/e
dbEST peak 1: 0.399

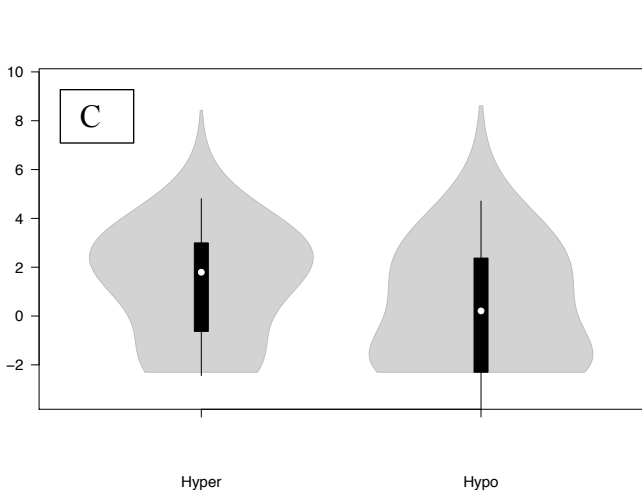
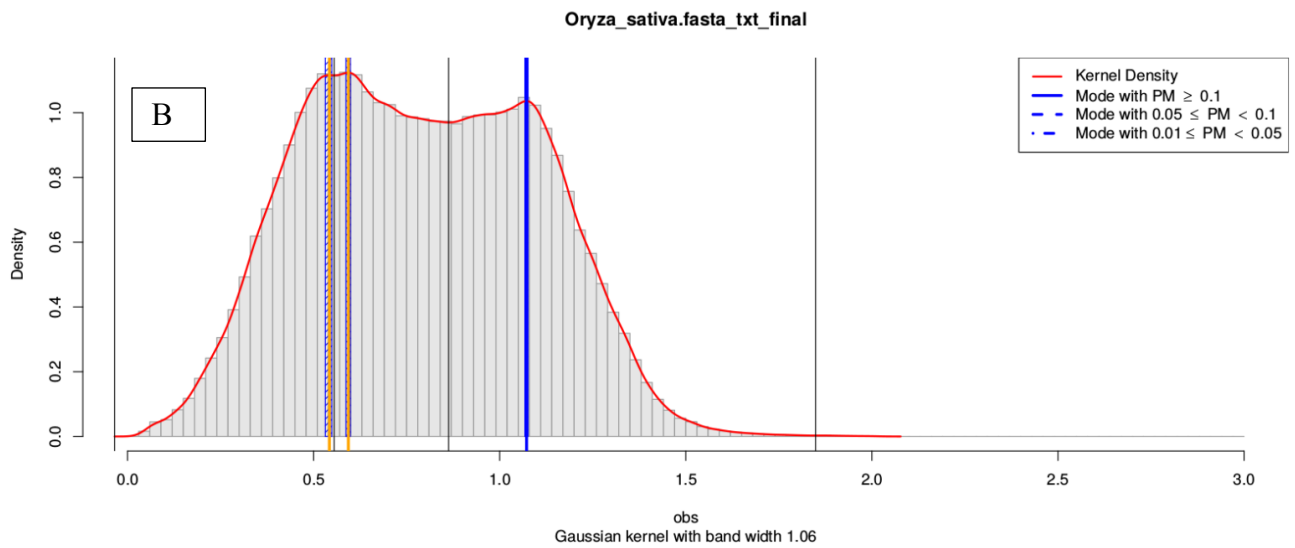
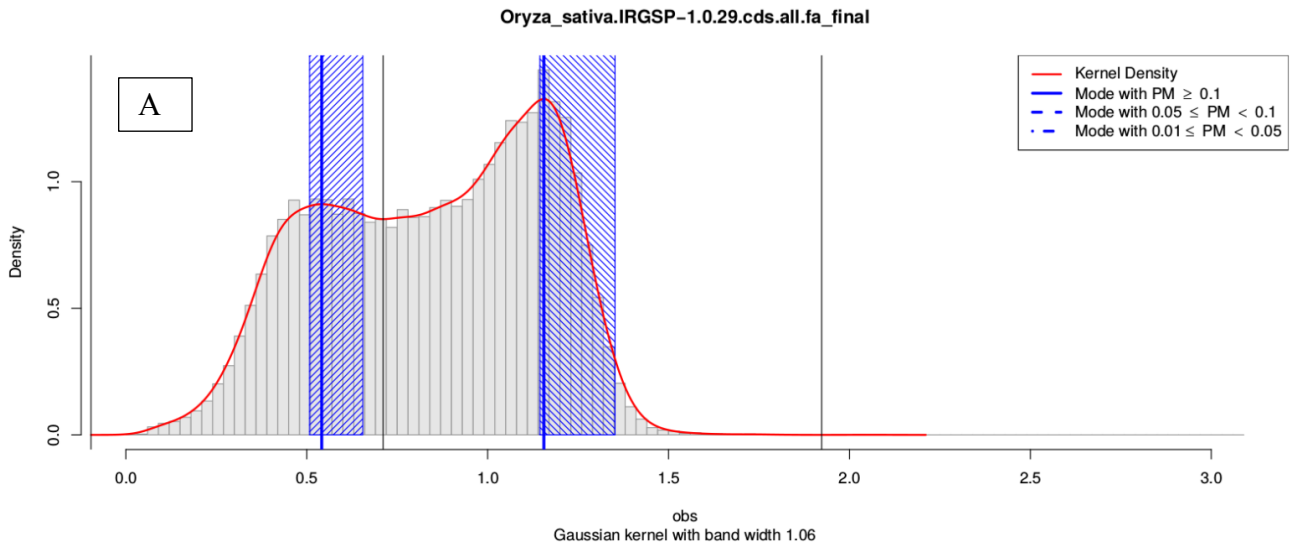
cDNA peak 1: 0.435
cDNA peak 2: 0.787

CpG o/e $0.299 \leq x \leq 0.499$: 6,481 genes
CpG o/e $0.687 \leq x \leq 0.887$: 7,085 genes

RNA-Seq FPKM
median CpG o/e $0.299 \leq x \leq 0.499$: 9.022
median CpG o/e $0.687 \leq x \leq 0.887$: 4.6

Mood's median test: p-value < 2.2e-16

Supplementary figure 08: RNA-seq analysis *Oryza sativa*



CpGo/e

dbEST peak 1: 0.54
dbEST peak 2: 1.07

cDNA peak 1: 0.541
cDNA peak 2: 1.155

CpG o/e $0.44 \leq x \leq 0.64$: 6,481 gènes
CpG o/e $0.97 \leq x \leq 1.17$: 7,085 gènes

RNA-Seq FPKM

CpG o/e $0.44 \leq x \leq 0.64$: 5.916
CpG o/e $0.97 \leq x \leq 1.17$: 1.133

Mood's median test: p-value < 2.92e-12

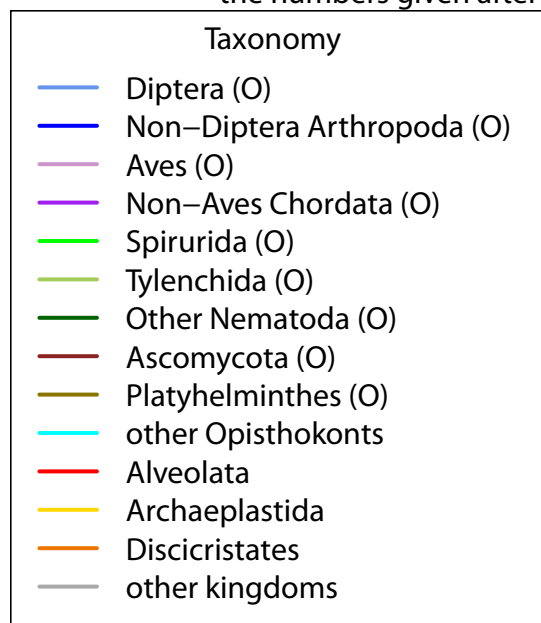
y axis : log(FPKM)

References:

1. Bulla, I. *et al.* Notos - a galaxy tool to analyze CpN observed expected ratios for inferring DNA methylation types. *BMC Bioinformatics* **19**, 105 (2018).
2. Bewick, A. J. & Schmitz, R. J. Gene body DNA methylation in plants. *Curr. Opin. Plant Biol.* **36**, 103–110 (2017).
3. He, X.-J., Chen, T. & Zhu, J.-K. Regulation and function of DNA methylation in plants and animals. *Cell Res.* **21**, 442–465 (2011).
4. Dobin, A. & Gingeras, T. R. Mapping RNA-seq Reads with STAR. in *Current Protocols in Bioinformatics* 11.14.1-11.14.19 (John Wiley & Sons, Inc., 2015). doi:10.1002/0471250953.bi1114s1
5. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).

Supplementary figure 9 : clustering of species according to CpG o/e ratio distributions and methylation types

the numbers given after the species names are number of modes, mode position, absolute Q50 mode skewness, and standard deviation



Cluster 1

Cluster 2

Cluster 3

Cluster 4

| | | |
|-----|-------------------------------|--------------------------|
| 82 | Metarhizium anisopliae | 1 / 0.92 / 0.019 / 0.12 |
| 15 | Aspergillus oryzae | 1 / 0.88 / 0.015 / 0.12 |
| 133 | Trichoderma atroviride | 1 / 0.89 / 0.012 / 0.12 |
| 38 | Clonorchis sinensis | 1 / 0.94 / 0.005 / 0.12 |
| 78 | Magnaporthe oryzae | 1 / 0.94 / 0.004 / 0.12 |
| 39 | Coprinopsis cinerea | 1 / 0.95 / 0.001 / 0.11 |
| 73 | Leishmania major | 1 / 1.08 / 0.001 / 0.09 |
| 37 | Claviceps purpurea | 1 / 0.97 / 0 / 0.11 |
| 33 | Chlamydomonas reinhardtii | 1 / 0.94 / -0.004 / 0.09 |
| 89 | Neurospora crassa | 1 / 0.92 / -0.011 / 0.12 |
| 10 | Anopheles gambiae | 1 / 1.15 / -0.013 / 0.12 |
| 55 | Emiliania huxleyi | 1 / 1.19 / -0.016 / 0.1 |
| 130 | Toxoplasma gondii | 1 / 1.05 / -0.019 / 0.11 |
| 9 | Anopheles darlingi | 1 / 1.16 / -0.02 / 0.12 |
| 1 | Acyrtosiphon pisum | 1 / 0.9 / 0.095 / 0.34 |
| 102 | Phycomitrella patens | 1 / 0.72 / 0.047 / 0.18 |
| 87 | Necator americanus | 1 / 0.96 / 0.046 / 0.19 |
| 7 | Ancylostoma ceylanicum | 1 / 0.94 / 0.036 / 0.18 |
| 29 | Caenorhabditis remanei | 1 / 0.9 / 0.035 / 0.21 |
| 40 | Cordyceps militaris | 1 / 1.02 / 0.031 / 0.13 |
| 64 | Giardia intestinalis | 1 / 0.72 / 0.031 / 0.13 |
| 3 | Albugo candida | 1 / 1.07 / 0.025 / 0.16 |
| 25 | Caenorhabditis brenneri | 1 / 0.85 / 0.024 / 0.2 |
| 81 | Meloidogyne incognita | 1 / 0.72 / 0.023 / 0.21 |
| 48 | Drosophila ananassae | 1 / 0.85 / 0.018 / 0.18 |
| 27 | Caenorhabditis elegans | 1 / 0.91 / 0.017 / 0.2 |
| 134 | Trichoplax adhaerens | 1 / 0.89 / 0.017 / 0.19 |
| 16 | Aspergillus terreus | 1 / 0.98 / 0.015 / 0.12 |
| 24 | Brugia malayi | 1 / 0.95 / 0.014 / 0.21 |
| 90 | Nippostrongylus brasiliensis | 1 / 1.08 / 0.014 / 0.18 |
| 6 | Ancylostoma caninum | 1 / 0.93 / 0.013 / 0.17 |
| 132 | Trichinella spiralis | 1 / 1 / 0.011 / 0.22 |
| 67 | Haemonchus contortus | 1 / 0.97 / 0.01 / 0.17 |
| 110 | Schistosoma japonicum | 1 / 0.8 / 0.009 / 0.21 |
| 139 | Tuber melanosporum | 1 / 0.86 / 0.009 / 0.12 |
| 34 | Chondrus crispus | 1 / 1.03 / 0.008 / 0.16 |
| 143 | Wuchereria bancrofti | 1 / 0.97 / 0.008 / 0.22 |
| 13 | Aspergillus flavus | 1 / 0.89 / 0.007 / 0.12 |
| 91 | Onchocerca flexuosa | 1 / 1.01 / 0.007 / 0.23 |
| 104 | Postia placenta | 1 / 1.04 / 0.007 / 0.14 |
| 145 | Yarrowia lipolytica | 1 / 0.8 / 0.007 / 0.12 |
| 49 | Drosophila melanogaster | 1 / 0.89 / 0.006 / 0.15 |
| 60 | Fusarium oxysporum | 1 / 0.86 / 0.006 / 0.13 |
| 26 | Caenorhabditis briggsae | 1 / 0.92 / 0.005 / 0.22 |
| 112 | Schizosaccharomyces pombe | 1 / 0.84 / 0.004 / 0.15 |
| 126 | Tetranichus urticae | 1 / 0.69 / 0.004 / 0.17 |
| 131 | Tribolium castaneum | 1 / 1.03 / 0.002 / 0.21 |
| 70 | Heterorhabditis bacteriophora | 1 / 0.84 / 0.001 / 0.2 |
| 111 | Schistosoma mansoni | 1 / 0.8 / -0.001 / 0.19 |
| 74 | Loa loa | 1 / 0.99 / -0.002 / 0.21 |
| 14 | Aspergillus niger | 1 / 0.89 / -0.003 / 0.12 |
| 137 | Trypanosoma brucei | 1 / 0.96 / -0.003 / 0.12 |
| 109 | Saccharomyces cerevisiae | 1 / 0.76 / -0.004 / 0.15 |
| 123 | Taenia solium | 1 / 0.87 / -0.004 / 0.13 |
| 113 | Schmidtea mediterranea | 1 / 0.88 / -0.007 / 0.23 |
| 36 | Ciona savignyi | 1 / 0.84 / -0.008 / 0.23 |
| 47 | Dendroctonus ponderosae | 1 / 0.86 / -0.008 / 0.15 |
| 51 | Echinococcus multilocularis | 1 / 0.87 / -0.008 / 0.13 |
| 138 | Trypanosoma cruzi | 1 / 0.98 / -0.01 / 0.12 |
| 141 | Uncinocarpus reesii | 1 / 0.97 / -0.011 / 0.12 |
| 2 | Aedes aegypti | 1 / 1.07 / -0.013 / 0.13 |
| 50 | Echinococcus granulosus | 1 / 0.88 / -0.013 / 0.14 |
| 28 | Caenorhabditis japonica | 1 / 1.12 / -0.014 / 0.21 |
| 58 | Fasciola hepatica | 1 / 1.02 / -0.02 / 0.16 |
| 18 | Atta cephalotes | 1 / 1.1 / -0.021 / 0.27 |
| 63 | Gasterosteus aculeatus | 1 / 0.7 / -0.021 / 0.16 |
| 114 | Selaginella moellendorffii | 1 / 0.91 / -0.021 / 0.17 |
| 122 | Suillus luteus | 1 / 0.92 / -0.021 / 0.18 |
| 31 | Capitella teleta | 1 / 0.81 / -0.022 / 0.18 |
| 120 | Strigamia maritima | 1 / 1.07 / -0.024 / 0.31 |
| 118 | Solenopsis invicta | 1 / 1.13 / -0.027 / 0.26 |
| 43 | Culex quinquefasciatus | 1 / 1.13 / -0.03 / 0.12 |
| 83 | Mnemiopsis leidyi | 1 / 0.75 / -0.03 / 0.18 |
| 105 | Pristionchus pacificus | 1 / 1.07 / -0.031 / 0.19 |
| 135 | Trichuris muris | 1 / 1.11 / -0.033 / 0.16 |
| 101 | Petromyzon marinus | 1 / 0.9 / -0.035 / 0.17 |
| 19 | Babesia equi | 1 / 0.71 / -0.038 / 0.2 |
| 121 | Strongylocentrotus purpuratus | 1 / 0.48 / 0.12 / 0.22 |
| 5 | Anas platyrhynchos | 1 / 0.35 / 0.113 / 0.18 |
| 8 | Anolis carolinensis | 1 / 0.35 / 0.107 / 0.19 |
| 62 | Gallus gallus | 1 / 0.37 / 0.102 / 0.18 |
| 65 | Glycine max | 1 / 0.34 / 0.101 / 0.24 |
| 42 | Cryptosporidium parvum | 1 / 0.36 / 0.1 / 0.19 |
| 124 | Taeniopygia guttata | 1 / 0.32 / 0.095 / 0.19 |
| 59 | Ficedula albicollis | 1 / 0.34 / 0.094 / 0.19 |
| 12 | Arabidopsis thaliana | 1 / 0.61 / 0.089 / 0.21 |
| 22 | Brassica oleracea | 1 / 0.66 / 0.08 / 0.23 |
| 30 | Canis lupus familiaris | 1 / 0.38 / 0.074 / 0.2 |
| 23 | Brassica rapa | 1 / 0.66 / 0.073 / 0.22 |
| 79 | Meleagris gallopavo | 1 / 0.33 / 0.072 / 0.15 |
| 142 | Vitis vinifera | 1 / 0.38 / 0.069 / 0.21 |
| 144 | Xenopus tropicalis | 1 / 0.34 / 0.065 / 0.15 |
| 116 | Solanum lycopersicum | 1 / 0.44 / 0.064 / 0.22 |
| 117 | Solanum tuberosum | 1 / 0.43 / 0.064 / 0.22 |
| 140 | Tursiops truncatus | 1 / 0.45 / 0.059 / 0.19 |
| 106 | Prunus persica | 1 / 0.43 / 0.057 / 0.2 |
| 57 | Equus caballus | 1 / 0.39 / 0.05 / 0.18 |
| 76 | Macaca mulatta | 1 / 0.37 / 0.05 / 0.17 |
| 71 | Homo sapiens | 1 / 0.36 / 0.046 / 0.17 |
| 97 | Ovis aries | 1 / 0.42 / 0.044 / 0.18 |
| 17 | Astyanax mexicanus | 1 / 0.47 / 0.042 / 0.17 |
| 32 | Cavia porcellus | 1 / 0.39 / 0.042 / 0.18 |
| 85 | Naegleria gruberi | 1 / 0.35 / 0.033 / 0.15 |
| 92 | Oreochromis niloticus | 1 / 0.44 / 0.032 / 0.15 |
| 96 | Oryzias latipes | 1 / 0.52 / 0.032 / 0.16 |
| 21 | Bos taurus | 1 / 0.44 / 0.03 / 0.19 |
| 56 | Entamoeba histolytica | 1 / 0.29 / 0.028 / 0.16 |
| 4 | Amphimedon queenslandica | 1 / 0.34 / 0.024 / 0.23 |
| 66 | Gorilla gorilla | 1 / 0.35 / 0.024 / 0.16 |
| 99 | Paramecium tetraurelia | 1 / 0.4 / 0.021 / 0.16 |
| 84 | Mus musculus | 1 / 0.36 / 0.019 / 0.17 |
| 125 | Tetrahymena thermophila | 1 / 0.37 / 0.019 / 0.15 |
| 108 | Rhodnius prolixus | 1 / 0.68 / 0.016 / 0.19 |
| 45 | Danio rerio | 1 / 0.52 / 0.013 / 0.17 |
| 129 | Theileria parva | 1 / 0.67 / 0.01 / 0.18 |
| 80 | Meloidogyne hapla | 1 / 0.67 / 0.007 / 0.23 |
| 61 | Gadus morhua | 1 / 0.68 / 0.002 / 0.18 |
| 103 | Plasmodium falciparum | 1 / 0.67 / 0 / 0.21 |
| 94 | Oryctolagus cuniculus | 1 / 0.41 / -0.018 / 0.2 |
| 127 | Theileria annulata | 1 / 0.68 / -0.03 / 0.18 |
| 93 | Ornithorhynchus anatinus | 1 / 0.63 / -0.033 / 0.2 |
| 75 | Lottia gigantea | 2 / 0.37 / 0.219 / 0.29 |
| 136 | Triticum aestivum | 2 / 0.55 / 0.214 / 0.26 |
| 147 | Zootermopsis nevadensis | 2 / 0.4 / 0.196 / 0.26 |
| 52 | Eimeria acervulina | 2 / 0.43 / 0.163 / 0.22 |
| 35 | Ciona intestinalis | 2 / 0.66 / 0.141 / 0.28 |
| 100 | Parastromyloides trichosuri | 2 / 0.46 / 0.062 / 0.26 |
| 72 | Ixodes scapularis | 1 / 0.96 / -0.044 / 0.19 |
| 107 | Puccinia graminis | 1 / 0.92 / -0.045 / 0.21 |
| 69 | Helobdella robusta | 1 / 0.84 / -0.047 / 0.23 |
| 128 | Theileria orientalis | 1 / 0.94 / -0.048 / 0.16 |
| 44 | Danaus plexippus | 1 / 0.96 / -0.051 / 0.21 |
| 119 | Steinernema carpocapsae | 1 / 1.15 / -0.052 / 0.2 |
| 20 | Bombyx mori | 1 / 1.06 / -0.064 / 0.24 |
| 68 | Heliconius melpomene | 1 / 0.95 / -0.067 / 0.25 |
| 46 | Daphnia pulex | 1 / 1.04 / -0.096 / 0.23 |
| 86 | Nasonia vitripennis | 1 / 1.06 / -0.128 / 0.26 |
| 41 | Crassostrea gigas | 2 / 0.44 / -0.15 / 0.25 |
| 88 | Nematostella vectensis | 1 / 0.9 / -0.161 / 0.29 |
| 11 | Apis mellifera | 2 / 0.69 / -0.189 / 0.39 |
| 53 | Eimeria maxima | 2 / 0.3 / -0.2 / 0.25 |
| 146 | Zea mays | 2 / 0.76 / -0.24 / 0.26 |
| 115 | Setaria italica | 2 / 0.57 / -0.291 / 0.27 |
| 95 | Oryza sativa | 2 / 0.54 / -0.304 / 0.3 |

