ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Characterization of DNA methylation as a function of biological complexity via dinucleotide inter-distances

This is the submitted version (pre peer-review, preprint) of the following publication:

*Published Version:*
Paci, G., Cristadoro, G., Monti, B., Lenci, M., Degli Esposti, M., Castellani, G.c., et al. (2016). Characterization of DNA methylation as a function of biological complexity via dinucleotide inter-distances. PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY OF LONDON SERIES A: MATHEMATICAL PHYSICAL AND ENGINEERING SCIENCES, 374(2063), 1-11 [10.1098/rsta.2015.0227].

*Availability:*
This version is available at: https://hdl.handle.net/11585/548697 since: 2017-05-12

*Published:*
DOI: http://doi.org/10.1098/rsta.2015.0227

*Terms of use:*

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

12 October 2024

# Characterization of DNA methylation as a function of biological complexity via dinucleotide inter-distances SUPPLEMENTARY MATERIAL

Giulia Paci[1], Giampaolo Cristadoro[3], Barbara Monti[4], Marco Lenci[2,3], Mirko Degli Esposti[3], Gastone C. Castellani[1,2] and Daniel Remondini*[1,2]

[1]Department of Physics and Astronomy, University of Bologna, Viale B. Pichat 6/2, 40127 Bologna, Italy
[2]INFN, Bologna Unit, Viale B. Pichat 6/2, 40127 Bologna, Italy
[3]Department of Mathematics, University of Bologna, Piazza di Porta S. Donato 5, 40126 Bologna, Italy
[4]Department of Pharmacy and Biotechnology, University of Bologna, Via S. Donato 15, 40127 Bologna, Italy

*Corresponding author

Table 1: List of the organisms and their DNA sequence repository website.

| **Organism** | **Repository** |
|---|---|
| Human Adenovirus 54 | www.ncbi.nlm.nih.gov/nuccore/253761974 |
| Apis mellifera (honey bee, release 4.5) | hymenopteragenome.org/beebase/q=download_sequences |
| Bos taurus (cow) | www.ensembl.org/Bos_taurus/Info/Index |
| Caenorhabditis elegans (round worm) | www.ensembl.org/Caenorhabditis_elegans/Info/Index |
| Canis familiaris (dog) | www.ensembl.org/Canis_familiaris/Info/Index |
| Ciona intestinalis (sea vase) | www.ensembl.org/Ciona_intestinalis/Info/Index |
| Danio rerio (Zebrafish) | www.ensembl.org/Danio_rerio/Info/Index |
| Drosophila Melanogaster (Fruit Fly) | www.ensembl.org/Drosophila_melanogaster/Info/Index |
| Equus caballus (horse) | www.ensembl.org/Equus_caballus/Info/Index |
| Escherichia Coli | www.genome.wisc.edu |
| Homo Sapiens (man, release hg19) | hgdownload.cse.ucsc.edu/downloads.html#human |
| Macaca mulatta (rhesus monkey) | www.ensembl.org/Macaca_mulatta/Info/Index |
| Monodelphis domesticus (opossum) | www.ensembl.org/Monodelphis_domestica/Info/Index |
| Mus musculus (mouse) | www.ensembl.org/Mus_musculus/Info/Index |
| Oikopleura diotica (tunicate) | www.genoscope.cns.fr/externe/GenomeBrowser/Oikopleura |
| Ornithorhyncus anatinus (platypus) | www.ensembl.org/Ornithorhynchus_anatinus/Info/Index |
| Pan troglodytes (chimpanzee) | www.ensembl.org/Pan_troglodytes/Info/Index |
| Rattus norvegicus (rat) | www.ensembl.org/Rattus_norvegicus/Info/Index |
| Saccharomyces cerevisiae R64-1-1 | www.ensembl.org/Saccharomyces_cerevisiae |
| Tetraodon nigroviridis (puffer fish) | www.ensembl.org/Tetraodon_nigroviridis/Info/Index |
| Tribolium castaneum (beetle) | metazoa.ensembl.org/Tribolium_castaneum/Info/Index |

Table 2: Power-law fit of all human dinucleotide distributions. For each dinucleotide, the fit parameters $b$, the goodness of fit $r^2$, the P-value of the normalized Chi-square test $P(\chi^2)$ are shown. All errors are expressed as 95% confidence intervals, and rounded to the first significant digit. Only the dinucleotide CG distribution is significantly non compatible with a power-law distribution.

| **Dinucleotide** | **b** | $r^2$ | $P(\chi^2)$ |
|---|---|---|---|
| AA | -3.1 ± 0.2 | 0.98 | 1 |
| AC | -3.7 ± 0.2 | 0.94 | 1 |
| AG | -2.9 ± 0.2 | 0.94 | 1 |
| AT | -3.5 ± 0.1 | 0.99 | 1 |
| CA | -3.1 ± 0.2 | 0.93 | 1 |
| CC | -3.6 ± 0.2 | 0.98 | 0.99 |
| CG | -2.7 ± 0.4 | 0.83 | 0.00082 |
| CT | -3.0 ± 0.2 | 0.96 | 1 |
| GA | -3.2 ± 0.2 | 0.96 | 1 |
| GC | -4.1 ± 0.2 | 0.98 | 0.99 |
| GG | -3.6 ± 0.2 | 0.97 | 0.99 |
| GT | -3.8 ± 0.3 | 0.95 | 1 |
| TA | -3.6 ± 0.2 | 0.98 | 0.99 |
| TC | -3.2 ± 0.2 | 0.96 | 1 |
| TG | -3.3 ± 0.2 | 0.96 | 1 |
| TT | -2.9 ± 0.1 | 0.98 | 1 |

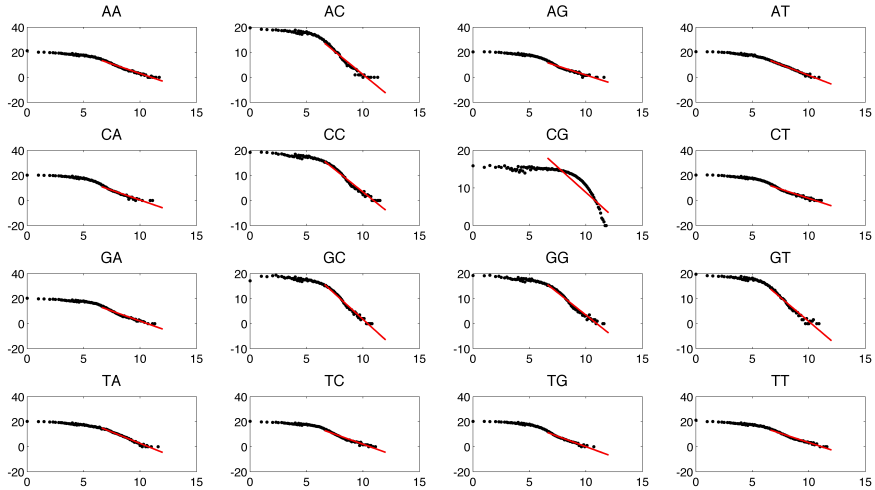


Figure 1: Double logarithmic plot of the dinucleotide distance distributions for human, together with the power-law fit (red line). The curves were fitted in the tails ($d > 90$, corresponding to $x = 6.5$ in logarithmic scale).

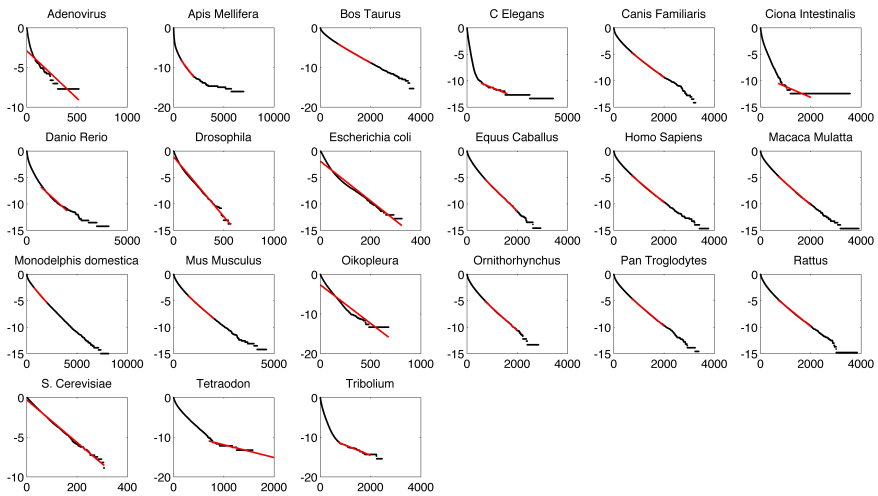


Figure 2: Plot of the cumulative distributions for all the studied organisms in semi-logarithmic scale, together with the exponential fit (red line). The curves were fitted in the interval $700 < d < 2000$.