Paper Title:

Are There Seven Symbols for the Nucleotide-Based Genetic Code?

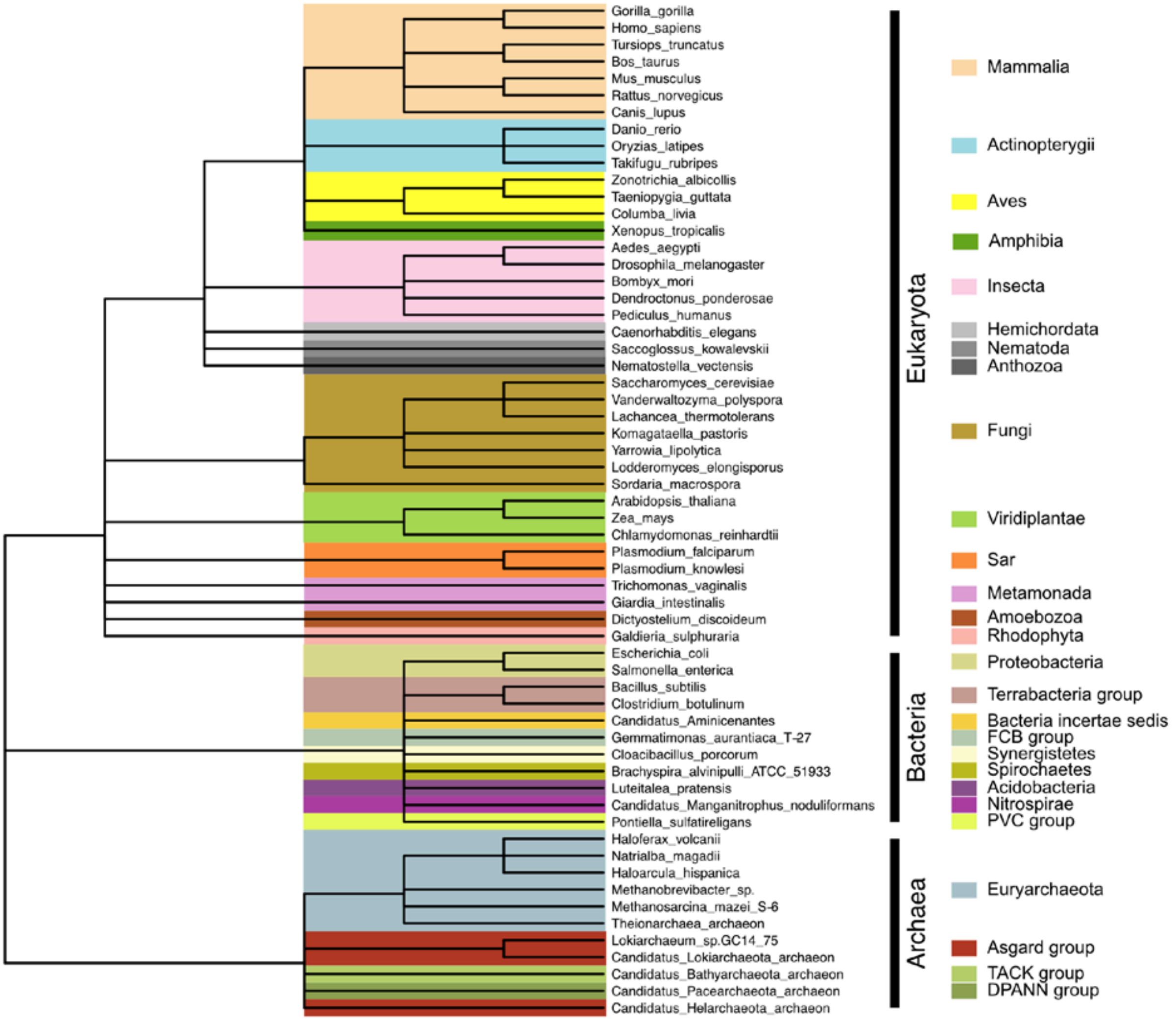
Authors:
Adam Kłóś
The Pontifical University of John Paul II, Kraków, Poland, ORCID: 0000-0001-6382-665X, adam.klos@upjp2.edu.pl.
Przemysław M. Płon
Jagiellonian University, Kraków, Poland, ORCID: 0000-0002-0261-3439, Scopus ID: 6602754825, przemyslaw.plonka@uj.edu.pl

Krzysztof Baczyński
Jagiellonian University, Kraków, Poland, ORCID: 0009-0005-9569-2496,
krzysztof.baczynski@gmail.com.

Corresponding author
Przemysław Mieszko Płonka,
przemyslaw.plonka@uj.edu.pl
+48 660 369 069

Supplementary Information

Figure S1. The Taxonomy of Organisms Used in the Study



Phylogenetic tree with annotations showing domains and some classes, groups of organisms.

Table S1. Genome Assembly Used in the Study

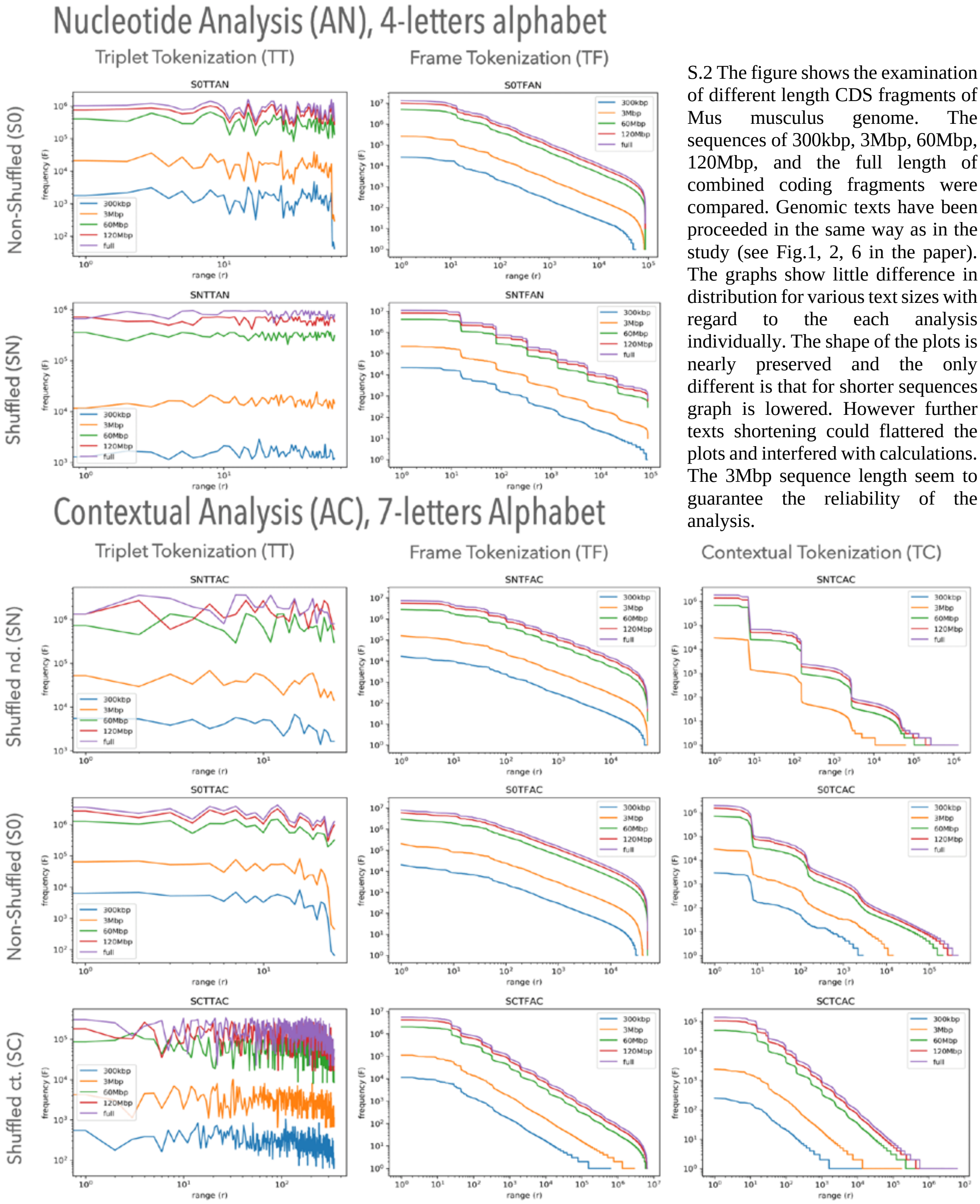
	S.C.	Organism	Taxid	Assembly
1.	AA	Aedes aegypti	7159	GCF_002204515.2_AaegL5.0_cds_from_genomic.fna.gz
2.	AT	Arabidopsis thaliana	3702	GCF_000001735.4_TAIR10.1_cds_from_genomic.fna.gz
3.	BA	Brachyspira alvinipulli ATCC 51933	1408430	GCF_000518245.1_ASM51824v1_cds_from_genomic.fna.gz
4.	BM	Bombyx mori	7091	GCF_014905235.1_Bmori_2016v1.0_cds_from_genomic.fna.gz
5.	BS	Bacillus subtilis	1423	GCF_000385985.1_Bacillus_subtilis_PS216_cds_from_genomic.fna.gz
6.	BT	Bos taurus	9913	GCF_002263795.1_ARS-UCD1.2_cds_from_genomic.fna.gz
7.	CA	Candidatus Aminicenantes	2052149	GCA_011773565.1_ASM1177356v1_cds_from_genomic.fna.gz
8.	CB	Clostridium botulinum	1491	GCF_017330945.1_ASM1733094v1_cds_from_genomic.fna.gz
9.	CBA	Candidatus Bathyarchaeota archaeon	2026714	GCA_014894645.1_ASM1489464v1_cds_from_genomic.fna.gz

10	CE	Caenorhabditis elegans	6239	GCF_000002985.6_WBcel235_cds_from_genomic.fna.gz
11	CHA	Candidatus Helarchaeota archaeon	2719382	GCA_013375455.1_ASM1337545v1_cds_from_genomic.fna.gz
12	CHE	Chlamydomonas reinhardtii	3055	GCF_000002595.1_v3.0_cds_from_genomic.fna.gz
13	CL	Cnis lupus	9612	GCF_012295265.1_UNSW_AlpineDingo_1.0_genomic.gff.gz
14	CLA	Candidatus Lokiarchaeota archaeon	2053489	GCA_014729315.1_ASM1472931v1_cds_from_genomic.fna.gz
15	CLL	Columba livia	8932	GCF_000337935.1_Cliv_1.0_cds_from_genomic.fna.gz
16	CMN	Candidatus Manganitrophus noduliformans	2606439	GCF_012184425.1_ASM1218442v1_cds_from_genomic.fna.gz
17	CP	Cloacibacillus porcorum	1197717	GCF_001701045.1_ASM170104v1_cds_from_genomic.fna.gz
18	CPA	Candidatus Pacearchaeota archaeon	2026773	GCA_002690445.1_ASM269044v1_cds_from_genomic.fna.gz
19	DD	Dictyostelium discoideum	44689	GCF_000004695.1_dicty_2.7_genomic.gff.gz
20	DM	Drosophila melanogaster	7227	GCF_000001215.4_Release_6_plus_ISO1_MT_cds_from_genomic.fna.gz
21	DP	Dendroctonus ponderosae	77166	GCF_000355655.1_DendPond_male_1.0_cds_from_genomic.fna.gz
22	DR	Danio rerio	7955	GCF_000002035.6_GRCz11_cds_from_genomic.fna.gz
23	EC	Escherichia coli	562	GCF_000008865.2_ASM886v2_cds_from_genomic.fna.gz
24	GA	Gemmatimonas aurantiaca T-27	379066	GCF_000010305.1_ASM1030v1_cds_from_genomic.fna.gz
25	GG	Gorilla gorilla	9593	GCF_008122165.1_Kamilah_GGO_v0_cds_from_genomic.fna.gz
26	GL	Giardia lamblia	5741	GCF_000002435.2_UU_WB_2.1_cds_from_genomic.fna.gz
27	GS	Galdieria sulphuraria	130081	GCF_000341285.1_ASM34128v1_cds_from_genomic.fna.gz
28	HH	Haloarcula hispanica	51589	GCF_000223905.1_ASM22390v1_cds_from_genomic.fna.gz
29	HS	Homo sapiens	9606	GCF_000001405.39_GRCh38.p13_cds_from_genomic.fna.gz
30	HV	Haloferax volcanii	2246	GCF_000025685.1_ASM2568v1_cds_from_genomic.fna.gz
31	KP	Komagataella pastoris	4922	GCA_001708105.1_ASM170810v1_genomic.gff.gz
32	LE	Lodderomyces elongisporus	36914	GCF_000149685.1_ASM14968v1_cds_from_genomic.fna.gz
33	LO	Lokiarchaeum sp. GC14_75	1538547	GCA_000986845.1_ASM98684v1_cds_from_genomic.fna.gz
34	LP	Luteitalea pratensis	1855912	GCF_001618865.1_ASM161886v1_cds_from_genomic.fna.gz
35	LT	Lachancea thermotolerans	381046	GCF_000142805.1_ASM14280v1_cds_from_genomic.fna.gz
36	MB	Methanobrevibacter sp.	66852	GCA_017652345.1_ASM1765234v1_cds_from_genomic.fna.gz
37	MM	Mus musculus	10090	GCF_000001635.27_GRCm39_cds_from_genomic.fna.gz
38	MMA	Methanosarcina mazei S-6	213585	GCF_010706455.1_ASM1070645v1_cds_from_genomic.fna.gz
39	NM	Natrialba magadii	13769	GCF_000337875.1_ASM33787v1_cds_from_genomic.fna.gz
40	NV	Nematostella vectensis	45351	GCF_000209225.1_ASM20922v1_genomic.gff.gz
41	OL	Oryzias latipes	8090	GCF_002234675.1_ASM223467v1_cds_from_genomic.fna.gz
42	PF	Plasmodium falciparum	5833	GCF_000002765.5_GCA_000002765_cds_from_genomic.fna.gz
43	PH	Pediculus humanus	121225	GCF_000006295.1_JCVI_LOUSE_1.0_genomic.gff.gz
44	PK	Plasmodium knowlesi	5850	GCF_000006355.2_GCA_000006355.2_cds_from_genomic.fna.gz
45	PS	Pontiella sulfatireligans	2750658	GCF_900890705.1_Pontiella_sulfatireligans_F21_T_draft_genome_cds_from_genomic.fna.gz
46	RN	Rattus norvegicus	10116	GCF_015227675.2_mRatBN7.2_cds_from_genomic.fna.gz
47	SC	Saccharomyces cerevisiae	4932	GCF_000146045.2_R64_cds_from_genomic.fna.gz
48	SE	Salmonella enterica	28901	GCF_000006945.2_ASM694v2_cds_from_genomic.fna.gz
49	SK	Saccoglossus kowalevskii	10224	GCF_000003605.2_Skow_1.1_genomic.gff.gz
50	SM	Sordaria macrospora	5147	GCF_000182805.2_ASM18280v2_cds_from_genomic.fna.gz
51	TG	Taeniopygia guttata	59729	GCF_003957565.2_bTaeGut1.4.pri_cds_from_genomic.fna.gz
52	TH	Theionarchaea archaeon	2747605	GCA_019008485.1_ASM1900848v1_cds_from_genomic.fna.gz
53	TR	Takifugu rubripes	31033	GCF_901000725.2_fTakRub1.2_cds_from_genomic.fna.gz

54	TT	Tursiops truncatus	9739	GCF_011762595.1_mTurTru1.mat.Y_cds_from_genomic.fna.gz
55	TV	Trichomonas vaginalis	5722	GCF_000002825.2_ASM282v1_cds_from_genomic.fna.gz
56	VP	Vanderwaltozyma polyspora	36033	GCF_000150035.1_ASM15003v1_genomic.gff.gz
57	XT	Xenopus tropicalis	8364	GCF_000004195.4_UCB_Xtro_10.0_cds_from_genomic.fna.gz
58	YL	Yarrowia lipolytica	4952	GCF_000002525.2_ASM252v1_cds_from_genomic.fna.gz
59	ZA	Zonotrichia albicollis	44394	GCF_000385455.1_Zonotrichia_albicollis-1.0.1_cds_from_genomic.fna.gz
60	ZM	Zea mays	4577	GCF_902167145.1_Zm-B73-REFERENCE-NAM-5.0_cds_from_genomic.fna.gz

Sequence data that support the findings of this study have been deposited in the The National Center for Biotechnology Information - Genome assembly Database with the accession codes: GCF_002204515.2, GCF_000001735.4, GCF_000518245.1, GCF_014905235.1, GCF_000385985.1, GCF_002263795.1, GCA_011773565.1, GCF_017330945.1, GCA_014894645.1, GCF_000002985.6, GCA_013375455.1, GCF_000002595.1, GCF_012295265.1, GCA_014729315.1, GCF_000337935.1, GCF_012184425.1, GCF_001701045.1, GCA_002690445.1, GCF_000004695.1, GCF_000001215.4, GCF_000355655.1, GCF_000002035.6, GCF_000008865.2, GCF_000010305.1, GCF_008122165.1, GCF_000002435.2, GCF_000341285.1, GCF_000223905.1, GCF_000001405.39, GCF_000025685.1, GCA_001708105.1, GCF_000149685.1, GCA_000986845.1, GCF_001618865.1, GCF_000142805.1, GCA_017652345.1, GCF_000001635.27, GCF_010706455.1, GCF_000337875.1, GCF_000209225.1, GCF_002234675.1, GCF_000002765.5, GCF_000006295.1, GCF_000006355.2, GCF_900890705.1, GCF_015227675.2, GCF_000146045.2, GCF_000006945.2, GCF_000003605.2, GCF_000182805.2, GCF_003957565.2, GCA_019008485.1, GCF_901000725.2, GCF_011762595.1, GCF_000002825.2, GCF_000150035.1, GCF_000004195.4, GCF_000002525.2, GCF_000385455.1, GCF_902167145.1

Figure S2. Comparison of Different Text Lengths



S.2 The figure shows the examination of different length CDS fragments of Mus musculus genome. The sequences of 300kbp, 3Mbp, 60Mbp, 120Mbp, and the full length of combined coding fragments were compared. Genomic texts have been proceeded in the same way as in the study (see Fig.1, 2, 6 in the paper). The graphs show little difference in distribution for various text sizes with regard to the each analysis individually. The shape of the plots is nearly preserved and the only different is that for shorter sequences graph is lowered. However further texts shortening could flattered the plots and interfered with calculations. The 3Mbp sequence length seem to guarantee the reliability of the analysis.

Table S2. Additional Information Regarding Genomic Texts That Met the KS Statistic Criteria D<0.05

		S0TCAC	S0TFAC	S0TFAN	SCTCAC	SCTFAC	SNTCAC	SNTFAC	SNTFAN
xmin	count	19	60	60	60	60	4	58	33
	mean	5,26	604,18	197,42	5,83	11,17	4,00	447,79	106,39
	std	5,55	345,34	209,26	1,95	7,29	0,00	346,79	89,67
	min	3,00	225,00	34,00	3,00	4,00	4,00	186,00	31,00
	25%	4,00	386,00	78,25	5,00	6,00	4,00	313,00	34,00
	50%	4,00	519,00	134,00	5,00	11,00	4,00	353,00	85,00
	75%	4,50	659,75	199,25	6,25	12,00	4,00	419,00	141,00
	max	28,00	2191,00	1194,0	15,00	37,00	4,00	2649,0	412,00
alpha	count	19	60	60	60	60	4	58	33
	mean	1,819	2,026	1,995	1,966	2,024	1,911	2,032	2,002
	std	0,088	0,020	0,015	0,032	0,019	0,028	0,020	0,009
	min	1,663	1,961	1,940	1,894	1,986	1,884	1,950	1,980
	25%	1,768	2,015	1,988	1,954	2,011	1,893	2,030	1,998
	50%	1,836	2,028	1,997	1,969	2,019	1,906	2,035	2,003
	75%	1,870	2,037	2,006	1,978	2,037	1,924	2,040	2,007
	max	1,994	2,077	2,021	2,044	2,068	1,948	2,096	2,022
D	count	19	60	60	60	60	4	58	33
	mean	0,038	0,013	0,008	0,012	0,008	0,029	0,022	0,019
	std	0,010	0,004	0,003	0,003	0,004	0,011	0,010	0,014
	min	0,014	0,008	0,003	0,005	0,001	0,014	0,009	0,003
	25%	0,035	0,011	0,006	0,010	0,004	0,027	0,014	0,007
	50%	0,041	0,012	0,007	0,012	0,007	0,033	0,019	0,011
	75%	0,045	0,015	0,010	0,014	0,011	0,035	0,029	0,030
	max	0,049	0,026	0,017	0,018	0,014	0,039	0,048	0,048

Parameters explanation:

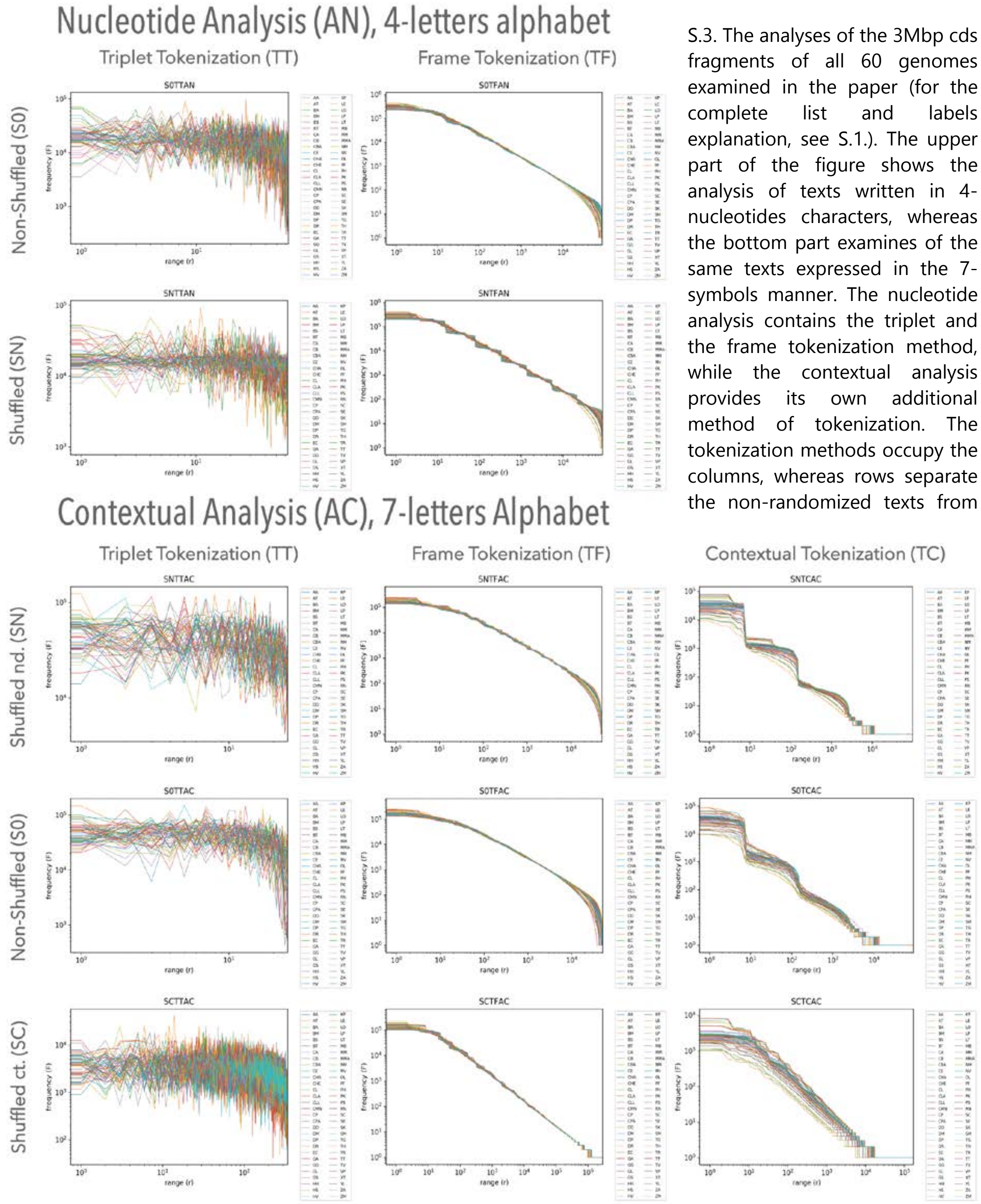
D - a minimal distance between the empirical distribution function of the sample and the cumulative distribution function of the power law distribution

xmax - the maximum value of the fitted distributions

alpha - the power law exponent or scaling parameter

xmin - The data value beyond which distributions should be fitted

Figure S3. Frequency-Rank Plots for All Studied Organisms



S.3. The analyses of the 3Mbp cds fragments of all 60 genomes examined in the paper (for the complete list and labels explanation, see S.1.). The upper part of the figure shows the analysis of texts written in 4-nucleotides characters, whereas the bottom part examines of the same texts expressed in the 7-symbols manner. The nucleotide analysis contains the triplet and the frame tokenization method, while the contextual analysis provides its own additional method of tokenization. The tokenization methods occupy the columns, whereas rows separate the non-randomized texts from

Table S3. Additional Information Regarding Random Texts That Met the KS Statistic Criteria *D*<0.05

		SCTCAC	SCTFAC
D	count	60	60
	mean	0,020	0,017
	std	0,002	0,000
	min	0,016	0,016
	25%	0,019	0,017
	50%	0,020	0,017
	75%	0,020	0,017
	max	0,025	0,018
alpha	count	60	60
	mean	1,96	2,07
	std	0,01	0,00
	min	1,94	2,06
	25%	1,95	2,06
	50%	1,96	2,07
	75%	1,96	2,07
	max	1,97	2,07
xmin	count	60	60
	mean	4	4
	std	0	0
	min	3	4
	25%	4	4
	50%	4	4
	75%	4	4
	max	5	4

Parameters explanation:

D - a minimal distance between the empirical distribution function of the sample and the cumulative distribution function of the power law distribution

xmax - the maximum value of the fitted distributions

alpha - the power law exponent or scaling parameter

xmin - The data value beyond which distributions should be fitted