



# Article Ubiquitous Micro-Modular Homologies among Genomes from Viruses to Bacteria to Human Mitochondrial DNA: Platforms for Recombination during Evolution?

Stefanie Weber <sup>1,†</sup>, Christina M. Ramirez <sup>2,†</sup> and Walter Doerfler <sup>1,3,\*</sup>

- <sup>1</sup> Institute for Clinical and Molecular Virology, University Hospital, Friedrich-Alexander University Erlangen-Nürnberg, Schlossgarten 4, 91054 Erlangen, Germany; siweber@gmx.de
- <sup>2</sup> Department of Biostatistics, UCLA School of Public Health, Los Angeles, CA 90095-1772, USA; cr@ucla.edu
- <sup>3</sup> Institute of Genetics, University of Cologne, 50674 Cologne, Germany
- \* Correspondence: walter.doerfler@t-online.de
- + These authors contributed equally to this work.

**Abstract:** The emerging *Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2)* and its variants have raised tantalizing questions about evolutionary mechanisms that continue to shape biology today. We have compared the nucleotide sequence of *SARS-CoV-2* RNA to that of genomes of many different viruses, of endosymbiotic proteobacterial and bacterial DNAs, and of human mitochondrial DNA. The entire 4,641,652 nt DNA sequence of *Escherichia coli K12* has been computer-matched to *SARS-CoV-2* RNA. Numerous, very similar micro-modular clusters of 3 to 13 nucleotides lengths were detected with sequence identities of 40 to >50% in specific genome segments between *SARS-CoV-2* and the investigated genomes. These clusters were part of patch-type homologies. Control sequence comparisons between 1000 randomly computer-composed sequences of 29.9 kb and with the A, C, G, T base composition of *SARS-CoV-2* genome versus the reference Wuhan *SARS-CoV-2* sequence showed similar patterns of sequence homologies. The universal A, C, G, T genetic coding mode might have succeeded in evolution due in part to its built-in capacity to select for a substantial reservoir of micro-modular domains and employ them as platforms for integrative recombination. Their role in *SARS-CoV-2* interspecies transition and the generation of variants appears likely, but their actual involvement will require detailed investigations.

**Keywords:** micro-modular sequence identities; *SARS-CoV-2* RNA compared to genomes of many viruses, protobacteria, *Escherichia coli*; patch type sequence similarities; the 4,641,652 nt DNA sequence of *Escherichia coli K12*; micro-modular identities in 1000 randomized *SARS-CoV-2* genomes; four letter genetic encoding system; signals for evolution, annealing and integrative recombination

# 1. Introduction

Like many researchers [1–3], we have followed the active replication and worldwide spread of *SARS-CoV-2*, the time course of rapidly emerging new mutations and its more complex variants [4,5]. These analyses led to the finding that more than 50% of the new mutations were C to T transitions, but there was a compensatory "repair" of T and A base pairs to C and G's (Table 13 in [5]) to counteract the otherwise deleterious depletion of cytosines in the sequence. The possibility exists that the cellular APOBEC cytosine deaminase might be involved in the high incidence of C to U (T) transitions [6].

In search of an independent approach to the study of basic nucleotide sequence design, vestiges of evolution as it were, different nucleotide sequences were juxtaposed to investigate fundamental nucleotide sequence commonalities across species barriers. Here we report on marked nucleotide sequence homologies between *SARS-CoV-2* RNA and the genomes of many classes of viruses, the genomes of proteobacteria, of bacteria, and of human mitochondrial DNA. The entire 4,641,652 nt DNA sequence of *E. coli K12* 



Citation: Weber, S.; Ramirez, C.M.; Doerfler, W. Ubiquitous Micro-Modular Homologies among Genomes from Viruses to Bacteria to Human Mitochondrial DNA: Platforms for Recombination during Evolution? *Viruses* 2022, *14*, 885. https://doi.org/10.3390/v14050885

Academic Editor: Eugene V. Ryabov

Received: 25 February 2022 Accepted: 20 April 2022 Published: 24 April 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https:// creativecommons.org/licenses/by/ 4.0/). has been computer-matched to SARS-CoV-2 RNA, and similar patchy sequence identity patterns have been observed. These similarities manifest themselves as micro-modules of two to maximally 13 nucleotide clusters. Could these sequence coincidences be explained statistically, when taking into account the fundamental biochemical design of genetic information that is limited to only four nucleosides? Initially five, subsequently expanded to 1000, randomized sequences were created in silico with the only premise of comparable sequence lengths (29.9 kb) and base composition to that of SARS-CoV-2 RNA. And indeed, very similar patches of sequence identities among up to 1000 randomized DNA sequences and of each of these to the SARS-CoV-2 sequence were noted at comparable frequencies, possibly with shorter homologies. So, has it all been statistically predictable? Certainly in the affirmative. However, in the information storage system of DNA or RNA that has successfully sustained evolution to this day, nucleotide sequence patches have facilitated base-pairing among genomes of very different derivations. Hence, the actual existence of micro-modular connectedness among vastly different genomes—though statistically predictable-still facilitates segmental sequence exchanges between genomes across the genetic universe. This genetic repertoire might have played a decisive role during evolution and still function as the molecular engine for exchanges among nucleotide sequence patches by whatever recombination-based mechanisms. In biology, the statistically predicted still holds highly significant functional alternatives.

The mechanisms operative during the emergence of SARS-CoV-2 from a poorly identified source to the worldwide pandemic pathogen with the ability to generate an abundance of highly replication-efficient mutants and variants have left us with a major enigma. Could principles of structure design in genomes in general offer a possible clue for further investigations? We have set out to compare the nucleotide sequences of the major groups of viruses, of proteobacteria, of *Escherichia coli K12*, and of the presumably bacterial plasmid-derived mitochondrial DNA of *Homo sapiens* to the nucleotide sequence of SARS-*CoV*-2. We found complex patterns of nucleotide sequence homologies at very similar levels among all genomes investigated. We have quantitated these extensive homologies and recorded the number of 3- to 9-mers, in some cases of up to 13-mers, of continuous sequence identities between the SARS-CoV-2 genome and the genomes selected for study. Aside from numerical documentation, complex patchy patterns of sequence identities have been found that might carry structural signals, e.g., for integrative recombination and nucleic acid-protein interactions. In control experiments, 1000 randomized sequences of the size and base composition of the SARS-CoV-2 genome were similarly quantitated and revealed very similar micro-modular patch homologies. We pursue the notion that the four-letter system of encoding genetic information has proved successful during evolution partly due to its inherent capability of providing identity platforms for integrative recombination and specific DNA-protein interactions.

#### 2. Materials and Methods

All DNA sequences on which this study is based were taken from the NCBI database, and table accession numbers were summarized in Table S1. For sequence comparisons, VectorNTI Advance 11 was employed. The default settings of the align tool of the software were used (pairwise alignment: K-tuple size: 2; number of best diagonals: 4; window size: 4; gap penalty: 5; gap opening penalty: 15; gap extension penalty: 6.66; multiple alignment: gap opening penalty: 15; gap extension penalty: 6.66; gap separation penalty range: 8; % identity for alignment delay: 40; transition weighting: 0). We have performed sequence comparisons with these settings for years.

Many sequence alignments have been included under Supplemental Materials, e.g., as Figure S1C from which Figure 1C was taken as a screen shot. Similar selections of complete sequence alignments are Figures S1–S4. These alignments will be available from the authors upon request.



**Figure 1.** (**A**) *SARS-CoV-2* original Wuhan and delta spike sequences versus fiber protein of *human adenovirus type 2* (*Ad2*). Segmental homology 45.1%. (**B**) Viral models of *human adenovirus type 2* and *SARS-CoV-2*. (**C**) Sequence alignment between Wuhan *SARS-CoV-2* and *adenovirus type 12* (*Ad12*). The yellow highlight designates sequence identities between the genomes of *SARS-CoV-2* and *Ad2* (**A**) or *Ad12* (**C**). The blue highlight stands for sequence identities between the nucleotide sequences of the SARS-CoV-2 spike in the WUHAN isolate and its delta variant.

The random and shuffle sequences were created using the Sequence Manipulation Suite (https://www.bioinformatics.org/sms2/about.html, accessed on 9 September 2021). Further technical details have been published previously [4,5]. We then used the PRSS

software to shuffle the random sequences 1000 times and then estimated the statistical significance from the shuffled alignment scores using E values. The E values ranged from 0.2–1 with a median value of 0.53. The sequences were aligned as described previously. The number of matches in the alignment were counted using the stringer package in R Studio (1.2.5033).

# 3. Results

#### 3.1. Problems Studied

We have set out to investigate overall similarities in basic genome architecture among genomes from very different organisms. How closely related are their sequences? Our study has focused on comparisons between different viral, proteobacterial, and bacterial genomes to *SARS-CoV-2* RNA. Since genetic information storage is based on DNA or RNA molecules, in which the sequence of only four nucleotides carries the entire genetic message, sequence similarities can be expected solely on statistical grounds. It will be much harder, however, to search beyond statistical considerations and explain the frequently observed micro-modular clusters that resemble one another among seemingly unrelated genomes. Of course, all genomes are interconnected through their common evolutionary history.

Table 1 lists the viral, proteobacterial, bacterial, and human mitochondrial genomes studied in this analysis. Levels of mutual overall sequence homologies can vary, but sequence homologies in selected segments of different genomes range between 44 and 47%, and in some cases reach > 50% (see Figures 1–4 in [6]). In comparisons between sequences in selected segments of a given genome and that of *SARS-CoV-2* RNA, clusters between 3 and up to 13 nucleotides of identical sequences were registered (Figures 1–4 in [6]). Similarly, extensive sequence homologies, at on average 44.5%, between *SARS-CoV-2* and randomized DNA sequences have also been documented (Table 2).

**Table 1.** Nucleotide sequences that have been compared to the *SARS-CoV-2* sequence: percentages of sequence homologies to *SARS-CoV-2*. Nucleotide sequences from different viruses, to proteobacteria, bacteria, and to human mitochondrial DNA have been analyzed for sequence homologies, overall or in selected segments, to the *SARS-CoV-2* genome.

Genomes	<b>Overall Homology</b>	In Selected Segments		
Archaea Viruses				
Acidianus rod-shaped virus 1	41.4%	46.2%		
Nitrosopumilus spindle-shaped virus	45.8%	47%		
Sulfolobus turreted icosahedral virus	30.3%	45.5%		
Bacteriophages				
Escherichia virus lambda ( $\lambda$ )	30.6%	45.3%		
Escherichia virus T4	9%	46.4%		
Escherichia virus mu	39.7%	44.2%		
Plant Viruses				
Alternaria brassicicola endornavirus	13.1%	44.5%		
Carrot mottle mimic umbravirus (ds-RNA)	5.9%	45%		
Fig badnavirus 1	11.8%	45.5%		
Insect Virus				
Autographa californica Nuclear Polyhedrosis Virus	11.2%	45.6%		
Human Viruses				
Adenovirus type 2	38.4%	43.2%		
Adenovirus type 5	40%	43.4%		
Adenovirus type 12	43.2%	44.9%		

A

В

Genomes	<b>Overall Homology</b>	In Selected Segments
Hepatitis B Virus	5.4%	45.7%
Human Immunodeficiency Virus 1	16.6%	45.8%
Bacterial Genomes (Segments)		
Escherichia coli K12 (129,062–161,212 nt)		45.2%
Candidatus Carsonella ruddii strain BT	8.5%	46%
Mitochondrial DNA—homo sapiens	27.3%	45.3%



**Figure 2.** (**A**) Sequence alignment between Wuhan *SARS-CoV-2* and *HIV-1*. (**B**) Sequence alignment between Wuhan *SARS-CoV-2* and *Sulfolobus turreted icosahedral virus 1* (*STIV1*). The yellow highlight designates sequence identities between the genomes of *SARS-CoV-2* and *HIV* (**A**) and *Sulfolobus turreted icosahedral virus* (*STIV1*) (**B**), respectively.

A

В

								Sec	tion 106
	8401	,8410	8420	8430	8440	8450	8460	8470	8480
SARS-CoV-2 AbEV1	A <mark>CICAA</mark> G <mark>C</mark> GA <mark>AA</mark>	AACACTAGTT AAATCGAGCT	GC <mark>A</mark> ACT <mark>G</mark> CAG TG <mark>A</mark> CGG <mark>G</mark>	AAGC <mark>IGAAC</mark> GCII <mark>IGAAC</mark>	ITGCAAAGAA ITGCA <mark>C-</mark> GAA	TGTGTCCTT TGGGCTGCT	AG <mark>ACA-ATGT</mark> C AA <mark>A</mark> AA <mark>CA</mark> GC <mark>T</mark> A	TTATCTAC <mark>I TT</mark> TGAC <mark>A</mark> AT	TTTATT ATTGAA
	8/81	8/190	8500	8510	8520	8530	8540	8550	107 Non 107
SARS-CoV-2 AbEV1	TCAGC GGCTA	LGCTCG <mark>G</mark> CA LGAAAA <mark>G</mark> GGT <mark>A</mark>	AG <mark>GGTT</mark> TGTT AC <mark>GGTT</mark> ACCA	G <mark>A</mark> TTCA <mark>GA</mark> T T <mark>A</mark> AGTG <mark>GA</mark> A	GTAGAAA <mark>C</mark> TA CCGACGC <mark>C</mark> CT	A <mark>AGATGTTG</mark> G <mark>ATA</mark> TTG	TIGAAIGICII TIGAAC <mark>G</mark> GICA	AAATTGT AGGGCTTAG	CACATO CGGAAC
								Sec	tion 108
CARC C-V.2	8561	,8570	8580	,8590	8600	,8610	8620	8630	8640
AbEV1	AAICIG AAAGGG	ACATAGAAGT AGCGC <mark>GA</mark> CCA	A <mark>TAC</mark> ACCT <mark>GA</mark>	TAGTIGIAA TACG <mark>T</mark> C <mark>T</mark> GC	G <mark>AA</mark> GGGG <mark>G</mark>	TCA <mark>AC</mark> GGTG	AACAAAGIIGA AAC <mark>GAAGI</mark> CCI	TGT <mark>C</mark> GA <mark>G</mark> GA Sec	AATAG-
	8641	,8650	8660	,8670	,8680	,8690	,8700	8710	8720
SARS-CoV-2 AbEV1	GTGAC GTGAC	C <mark>TTG</mark> GTGCTTG STTA <mark>G</mark> AACTTC	TAT <mark>TGA</mark> CTG <mark>T</mark> TGA <mark>T</mark> T <mark>A</mark> TGT <mark>T</mark>	<mark>AG</mark> T <mark>G</mark> CGC <mark>AG</mark> A <mark>G</mark> ATACA	<mark>gtcatatta</mark> a <mark>gtca</mark> aa <mark>a</mark> aa	<mark>TGC</mark> GCAGGI <mark>TGC</mark> TTTACC	TAGCAAA <mark>A</mark> AG <mark>TC</mark> CCAAGTT <mark>A</mark> CT <mark>TC</mark>	AC <mark>AA</mark> CATTG AG <mark>AA</mark> G <mark>A</mark> CAT	CTTT <mark>GA</mark> GCCC <mark>GA</mark>
	0704	0700	0740	0750	0700	0770	0700	Sec	tion 110
CARE CaV 2	8/21 TATCC2	BISU		JOI 00		0//U	B/80	77777777	7 N N C T T
AKS-COV-2 AbEV1	<mark>G</mark> C	ACGGTARAGA	-CG <mark>C</mark> <mark>GTC</mark> G	TTCGGA <mark>G</mark> GA	GC <mark>A</mark> G <mark>T</mark> CGAT <mark>A</mark>	T <mark>a</mark> ctggga <mark>a</mark>	AA <mark>TA</mark> C <mark>IG</mark> G <mark>I</mark> C-		TTAGGA
	8801	,8810	8820	,8830	,8840	,8850	8860	8870	8880
ARS-CoV-2 AbEV1	A <mark>C</mark> CTTI G <mark>C</mark> CA	TTA <mark>AGT</mark> T <mark>G</mark> AC <mark>A</mark> AGG <mark>AGT</mark> G <mark>G</mark> CA <mark>A</mark>	IGIGCAA <mark>C</mark> TA ICCCGIG <mark>C</mark> AG	C <mark>TA</mark> GACAAG T <mark>TA</mark> TCTG <mark>A</mark> A	TTG <mark>T</mark> TA <mark>A</mark> TGT ACA <mark>T</mark> GG <mark>A</mark> ACG	TGTA <mark>A</mark> C <mark>AAC</mark> AAAT <mark>A</mark> A <mark>AAC</mark>	<mark>AAA</mark> GATA <mark>GC</mark> AC AC <mark>A</mark> ACCT <mark>GC</mark> CG	TTAAGGGTG CTCAT <mark>GGT</mark> A	GTAA <mark>AA</mark> GGTC <mark>AA</mark> tion 112
	8881	8890	8900	8910	8920	8930	8940	8950	8960
ARS-CoV-2 AbEV1	TT <mark>G</mark> TTA GA <mark>G</mark> CAT	AAT <mark>AATTGGT</mark> - ICA <mark>AA</mark> A <mark>T</mark> A <mark>GT</mark> G	TG <mark>AAGCA</mark> GTT TA <mark>AA</mark> ACAAGG	A <mark>ATTAAA</mark> GT G <mark>A</mark> GA <mark>AAA</mark>	TACACTIGIG TACAAGGCAI	TT <mark>C</mark> CTTTTT AC <mark>C</mark> ACACCA	IGTT <mark>GC</mark> TG <mark>C</mark> TAT ATAC <mark>GC</mark> GA <mark>C</mark> AC <mark>T</mark>	TTTCTATTT AATAACCTG	AATAAC AT <mark>A</mark> G
	90475	90495	80.406	9050	7 80	510	80520	Section 1	163
AcNPV	TTGAA	AATAAAAAA	ACAAC <mark>GA</mark> ATC	GATTTCATC	TIGT <mark>A</mark> ACAAT	TGCGCTTC	AAACACAC	GTTTTCAA-	AGC
SARS-CoV-2	-TCTAT	IGC <mark>I</mark> TT <mark>AA</mark> GG <mark>C</mark>	<mark>a</mark> tttt <mark>ga</mark> tga	AGG <mark>T</mark> AAT <mark>I</mark> G	IGAC <mark>A</mark> CATTA	AAA <mark>G</mark> AAAT <mark>A</mark>	CTTGT <mark>CA</mark> CATA	CAATTGTTG — Section 1	T <mark>G</mark> A 1164
	89552	89562	89573	,8958	34 ,89	595	89606	89617 89	9628
AcNPV SARS-CoV-2	G <mark>G</mark> TC <mark>G</mark> I T <mark>G</mark> AT <mark>G</mark> A	TAA <mark>AT</mark> GT <mark>TAAA</mark> ATT <mark>AT</mark> T- <mark>T</mark> C <mark>AA</mark>	CCTT <mark>AA</mark> ACTG TAAA <mark>AA</mark> GGAC	TATTGTAAT TGG	CTGTAAGCGC TATGATTT	ACAT <mark>G</mark> GTGC TGTA <mark>G</mark> AAAA	CATT <mark>C</mark> G <mark>ATATA</mark> A AC <mark>C</mark> C <mark>A</mark> G <mark>ATA</mark> I	ACC <mark>TTATAAT</mark>	ATG
	89629	89639	89650	8966	51	672	89683	89694 89	9705
AcNPV	AACGAT	TCCAATTCTC	TGTTGATTAC	GCGTTTGGC.	AGCGCAAATA	CTGTCCAGA	AACATGCAAAC	GGTGGATGT	GAT
								- Section 1	1166
	89706	89716	89727	8973	88 ,89	749	89760	89771 89	9782
AcNPV SARS-CoV-2	TGTTGA GCCATG	A <mark>CGA</mark> C <mark>A</mark> AAAC <mark>G</mark> G <mark>CGA</mark> A <mark>A</mark> TGCT <mark>G</mark>	C <mark>T</mark> CAGT <mark>TTGG</mark> G <mark>T</mark> ATTG <mark>TTGG</mark>	AAG <mark>A</mark> AAA TGT <mark>A</mark> CTGAC	AA <mark>TAGACA</mark> CG AT <mark>TAGA</mark> T <mark>A</mark> A-	T <mark>TGA</mark> CCAGO - <mark>T</mark> CAAGATO	ATGG <mark>TG</mark> T <mark>T</mark> GGC TCAA <mark>TG</mark> G <mark>T</mark> AAC	TGTAAATAG TGGTATGAT	CC- T
	00702	90702	00004	0001	15 900	906	00027	Section 1	116/
AcNPV	-CGCC	CAATCGCCGC	CGCGGGTAAC	ATCCAGC	CCTGGCCGCA	TCGATCATT	AAAATAACAG	CAAAATGGT	GGG
SARS-CoV-2	T <mark>CG</mark> GT <mark>G</mark>	ATT <mark>TC</mark> ATACA	AA <mark>C</mark> CACGCCA	.GGTAGTG <mark>GA</mark>	GT <mark>T</mark> CCTGTTG	TAGATTCTI	ATTATTCATTC	TTAATGCCT	A 168
	89860	,89870	89881	,8989	2 ,89	903	89914	89925 89	9936
AcNPV SARS-CoV-2	C <mark>A</mark> ACG <mark>A</mark> T <mark>A</mark> TTA <mark>A</mark>	TTTTGAAATG CC <mark>T</mark> T	CGAT <mark>A</mark> CAA <mark>C</mark> G GACC <mark>A</mark> GGG <mark>C</mark> T	TGTTGC <mark>G</mark> TA <mark>T</mark> TAACT <mark>G</mark> C <mark>A</mark>	T <mark>G</mark> GCCGT <mark>C</mark> G <mark>T</mark> - <mark>G</mark> AGTCA <mark>C</mark> A <mark>T</mark>	T <mark>TT</mark> TGTTAA G <mark>TT</mark> GACACI	A <mark>GCATT</mark> ATC <mark>C</mark> CA GAC <mark>TT</mark> -AA <mark>C</mark> AA	AGTATTACA AGCCTTACA	ACG TTA
	80037	89947	80059	8006	0 90	980	89991	Section 1	1013
AcNPV	AGACGA	LCCGCCGGTTT	AGT <mark>TGC</mark> CGAA	ATAGAAAGT	ATC <mark>T</mark> GTTG <mark>C</mark>	AA <mark>TATCA</mark> AA	ATTATGTAAAC	CAAGGCAAT	TAT
ADC C.U.A	ACTOC	ATTTCTTAAA	TATCACTTC	2000222020	ACCTTANAA	TCTTCACC	CTTATTT	AAATATTCC	CAT

**Figure 3.** (**A**) Sequence alignment between the genomes of Wuhan *SARS-CoV-2* and of *Alternaria brassicicola endornavirus* (*AbEV1*). (**B**) Sequence alignment between the genomes of Wuhan *SARS-CoV-2* and of *Autographa californica nuclear Polyhedrosis Virus* (*AcNPV*). The yellow highlight designates sequence identities between the genome of *Alternaria brassicola endornavirus* (*AbEV1*) (**A**), or *Autographa californica Nuclear Polyhedrosis Virus* (*AcNPV*) (**B**) with the genome of *SARS-CoV-2*.

		150952	150962	150974	150995	150906	159907	S	ection 2077		
A Ca.C.	Ca. C. rudii strain BT rudii strain 135 557 - 167 442 pt	TAATGCTA TAATGCTA	ATACCITGGA ATACCITGGA	TAGAT <mark>AA</mark> AAA TAGAT <mark>AA</mark> AAA	AGTAAATTT - T AGTAAATTT - T	AGTCAAACAA AGTCAAACAA	159907 AAGAAGAATGG AAGAAGAATGG	AAATCATCTT AAATCATCTT	CTGAAGCT CTGAAGCT		
en. e.	SARS-CoV-2	GCT <mark>T</mark> AT <mark>T</mark>	AT <mark>GTGGGTT</mark> A	.TCTTC <mark>AA</mark> CCT	<mark>ag</mark> g <mark>a</mark> ct <mark>ttt</mark> c1	AT <mark>TAAAAT</mark> ATA	AATG <mark>A</mark> AATGG	<mark>aa</mark> c- <mark>cat-</mark> ia S	ection 2078		
	Ca. C. rudii strain BT	159930 <mark>TCTAAAA</mark> 1	,159940 I <mark>atiaataa</mark>	159951, AAAAAT <mark>TAT</mark> A	159962, Attgattcaaa	159973, TT <mark>GTGTAAGAA</mark>	159984, ATAGCAT <mark>CTTT</mark>	,1599 AAGAIGICAI	95 160006 TCAC <mark>A</mark> GTA		
Ca. C.	<i>rudii strain</i> 135,557 - 167,442 nt SARS-CoV-2	G- <mark>TAGACI</mark> G- <mark>TA</mark> GACI	I <mark>atiaataa</mark> Ig <mark>t</mark> gc <mark>a</mark> c <mark>t</mark> tg	ACCCTC <mark>T</mark> CTC	<mark>AIIGAII</mark> CAAA <mark>A</mark> <mark>ga</mark> aa <mark>caaa</mark>	<mark>GIGIAAGA</mark>	ATAGCAT <mark>CT</mark> II C <mark>C</mark> G <mark>II</mark>	AAGAIGICAI GAAAICCI S	TCACAGTA TCACTGTA ection 2079		
	Ca C rudii strain BT	160007		160028					72 160083		
Ca. C.	rudii strain 135,557 - 167,442 nt SARS-CoV-2	<mark>TTTTTTA</mark> GA <mark>A</mark>	A <mark>TAAAATTTA</mark> AA <mark>A</mark> GG <mark>A</mark> ATCT	AT <mark>A</mark> AAA <mark>ATAT</mark> ATCAAACTTC	TAA <mark>TA</mark> TT <mark>A</mark> ATA TAAC-TTTAGA	ATATATATGA GTCCAACCAA	A <mark>ATTATACATC</mark> C <mark>a</mark> ga <mark>at</mark> CT <mark>at</mark> t	ATAAATTTGI G <mark>T</mark> T <mark>A</mark> GA <mark>TT</mark> TC	TTATATAG CTA-ATAT		
		160084	160094	,160105	,160116	,160127	,160138	S ,1601	ection 2080 49 160160		
Ca. C.	Ca. C. rudii strain BT rudii strain 135,557 - 167,442 nt SARS-CoV-2	TTAAAAA TTAAAAA TACAAACT	I <mark>AATGAAAAA IAATGAAAAA ITG<mark>TG</mark>CCCTT</mark>	AATTCTAATT AATTCTAATT TTGGTG <mark>AA</mark> GT	TAATITTAACA TAATITTAACA TTT <mark>T</mark> AACGC <mark>CA</mark>	ICCATATAATG ICCATATAATG ICCAGATT-TG	P <mark>ATCTG</mark> GAAAA P <mark>ATCTG</mark> GAAAA C <mark>ATCTG</mark>	TTAAAAATTI TTAAAAATTI TTTATGCTIG	TTGT <mark>AGGT</mark> TTGT <mark>AGGT</mark> GAAC <mark>AGG</mark> A		
		160161	160171	160182	,160193	160204	,160215	5 <u>1602</u>	ection 2081 26 160237		
Ca. C.	Ca. C. rudii strain BT rudii strain 135,557 - 167,442 nt SARS-CoV-2	AGA AT AGA AT AGA GA <mark>AT</mark> O	I <mark>AAAAAAAGT</mark> I <mark>AAAAAAAGT</mark> C <mark>A</mark> GC <mark>AA</mark> CT <mark>GT</mark>	<mark>TTAATTAATA TTAATTAATA G<mark>T</mark>TGC<mark>T</mark>G<mark>AT</mark>T</mark>	ATAAAATTTTT ATAAAATTTTT <mark>AT</mark> TCTG <mark>T</mark> CC <mark>T</mark> -	ACAATGITTAC ACAATGITTAC ATATAA <mark>TT</mark> CC(	SCATAGGAGAT SCAT <mark>AGGAGAT</mark> SCATCATTTTC	CAATTATT-G CAATTATT-G CACTTTTAAG	TGGGGTGC TG <mark>GGGTGC</mark> TGTTA <mark>TG</mark> G		
	Co. C. mulii atuain PT	160238	160248	160259	160270	160281	,160292	S 21603	ection 2082 03 160314		
Ca. C.	rudii strain 135,557 - 167,442 nt SARS-CoV-2	AG <mark>CIGAAC</mark> AG <mark>CIGAAC</mark> AGTGICIC	CCTA <mark>T</mark> TAAA <mark>A</mark> CCTA <mark>T</mark> TAAA <mark>A</mark> CCTACTAAA-	. <mark>GATITITA</mark> GA <mark>TI</mark> AAAT <mark>GA</mark>	AATATTAATAA TC <mark>T</mark> CTGCTTTA	GTAATATATAT CTAATGTCT-A	ATG <mark>AAA</mark> ATT <mark>A</mark> A ATGCAGATTCA	ATTTIGIGGI ATTTIGT <mark>GGI</mark> <mark>TTIGT</mark> AA-	GTTAAAAA - TTAGAGG		
В		153833	,15384	46 ,1538	357 ,153	868 ,153	879 ,153	3890 ,1	Section 53901	1877 153914	
	<i>E. coli K12</i> 1 - 200,000 nt	CTTTTGGI AG <mark>T</mark> G <mark>T</mark> ATC	TGAAGIIIII C <mark>G</mark> G <mark>AGI</mark> CGC <mark>I</mark>	AACGCCACCA AAAACTATCA	GATT <mark>IGCATCI</mark> TTAC <mark>IG</mark> A <mark>A</mark> AG <mark>I</mark>	GTT <mark>T</mark> ATGCTTG CAG <mark>T</mark> CCA <mark>C</mark> CGC	GGAACAGGAAG CT <mark>A</mark> TGCAAT <mark>A</mark> C	AGAATCAGCA AAAAC <mark>CA</mark> CCG	TCGGTGTGTT STCGGTGCT Section	GCTGA GA-GA 1878	
	SARS-CoV-2	153915	153923	153934	,153945	,153956	,153967	153978,		153996 AATGA	
	<i>E. coli K12</i> 1 - 200,000 nt	GA <mark>A</mark> ACT <mark>TO</mark>	ACGGC <mark>IGI</mark> I	atcg <mark>c</mark> tat <mark>t</mark> t	gcag <mark>a</mark> aa <mark>tt</mark> ga	ACCTGCCAGC	TTCCCC <mark>A</mark> TG <mark>G</mark>	TG <mark>ACTC</mark> A <mark>TA</mark> G	CTGGCA <mark>TA</mark> — Section	AC 1879	
	SARS-CoV-2	<u>153997</u> T <mark>CTC</mark> TG <mark>C</mark> I	154, T <mark>tact</mark> a <mark>a</mark> tg	011 ,154 <mark>TCTATGC</mark> A <mark>G</mark> A	4022 ,15 - <mark>TTCATT</mark> TG <mark>T</mark> A	4033 ,15 .A <mark>T-TAG</mark> AGGTO	4044 ,1 ATGA <mark>A</mark> GT <mark>C</mark> AG	54055 Ac <mark>aa</mark> atcg <mark>c</mark> t	154066 CC <mark>AG</mark> G <mark>G</mark> CA	154078 AACTG	
	<i>E. coli K12</i> 1 - 200,000 nt		ATAACICAAA	TCTTTECTE	CTTTATTCATC	GTATAGCCAGT	ATTCACGCT-	ATAACTGACG	GAGCGTT — Section	1880	
	SARS-CoV-2 E. coli K12 1 - 200.000 nt	GAAAGATI TATAGCCA	154088 IG <mark>CIGAI</mark> TA <mark>I</mark> A-CIGCIGCI	AATTATAAAT AACGTTGAGT	,154110 TACCAGATGAT TGGTTATTACC	TTACAGGCTG CTTAAAGTCAC	154132 GCGTTATAGCT CTGCT-TAT	,15414, TG <mark>GA</mark> ATICI <mark>A</mark> TIGAGTAICA	ACAATCTT ATACTCTG	GATTC GAAAC	
		154161	.15	4176 .1	54187 .1	54198 .1	54209	154220		1881	
	<i>SARS-CoV-2</i> <i>E. coli K12</i> 1 - 200,000 nt	- <mark>TA</mark> AG <mark>GT</mark> I C <mark>T</mark> GAAGTA	IG <mark>GIGGIA</mark> AC <mark>GI</mark> I <mark>GI</mark> IC <mark>A</mark>	A <mark>T</mark> TAT <mark>AATTA</mark> G <mark>T</mark> GCC <mark>AA</mark> G <mark>TA</mark>	CC <mark>TGT</mark> AT <mark>A</mark> GAT AT <mark>TTT</mark> TC <mark>A</mark> ATT	T <mark>G</mark> TT <mark>TAG</mark> GA <mark>A</mark> G G <mark>G</mark> AA <mark>T</mark> G <mark>G</mark> TG <mark>A</mark> A	ST <mark>CTAA</mark> TC <mark>TCA</mark> AA <mark>CTAA</mark> GA <mark>T</mark> AA	<mark>A</mark> AC <mark>CTTT</mark> TGA ACG <mark>CT</mark> A <mark>T</mark> CGT	.G <mark>AGA</mark> GATA C <mark>AG</mark> T <mark>G</mark> TCG	TTT <mark>C</mark> A C <mark>C</mark> G	
	CARC C-U C	154243	154253	154264	154275	154286	154297	1543	Section	1882 154324	
	<i>E. coli K12</i> 1 - 200,000 nt	ACTGAAAI T <mark>CT</mark> TC <mark>A</mark> TI	ICTATCAG-G IC <mark>CAT</mark> G <mark>A</mark> AC <mark>G</mark>	CCGGTAGCAC CT <mark>GG</mark> GCA <mark>C</mark> TG	ACCTIGTAATG AC <mark>ACIGTA</mark> <mark>G</mark>	G <mark>IG</mark> IIGAAGGI C <mark>IG</mark> CCCC <mark>AGG</mark> #	TTTTAATTGIT. A <mark>T</mark> G <mark>TA</mark> C <mark>T-GIT.</mark>	ACTTTCCT ACT <mark>GT</mark> AG <mark>CC</mark> A	TTACAATC ATAGAGTA Section	AT ATTGC	
	SARS-CoV-2	154325154		54341	154352	154363	154374		TTTTGAAC	154406	
	<i>E. coli K12</i> 1 - 200,000 nt	TACG <mark>AT</mark> TI	T <mark>TGTCC</mark> GGA <mark>A</mark>	GCCCAGTAAT	C <mark>G</mark> GACCAAC <mark>I</mark> T	CCGGAAAGATA	AAATGAACCG	TAATCTTTT	TCTCAAAT	TTCAA	
C		14876	.148	90 .1490	00 1491	0 14920	14930	14940	S	ection 17	76 960
C	SARS-CoV-2 11,168 - 29,055 nt H. sapiens mitochondrion	ATAAACTI CCT <mark>A</mark> TCT	I <mark>TGTA</mark> AGAAT A <mark>T-TA</mark> CTCTC	AATAATGAGG Atcgc <mark>t</mark> acct	CTT <mark>TG</mark> GCTTTG CCC <mark>TG</mark> ACAA-G	CTGGAAATGCC C-GCCT <mark>AT</mark> AGC	GT <mark>TC</mark> C <mark>AAAAA</mark> AC <mark>TC</mark> G <mark>AA</mark> TAA	CC <mark>CAT</mark> TAC TT <mark>C</mark> TT <mark>CTC</mark> AC	TT <mark>TA</mark> TG <mark>A</mark> T CC <mark>TA</mark> AC <mark>A</mark> G	GC <mark>CAAC</mark> GT <mark>CAAC</mark>	T A C T
		14961	14970	,14980	,14990	15000	,15010	15020	15030 Se	ection 17 150	77 045
	SARS-CoV-2 11,168 - 29,055 nt H. sapiens mitochondrion	TIT <mark>ICI</mark> II CGC <mark>I-I</mark> CO	IG <mark>CIGGCATA</mark> CC <mark>C</mark> ACC <mark>C</mark> T <mark>TA</mark>	<mark>CTAA</mark> TTG <mark>T</mark> TA CTAACAT <mark>T</mark> AA	<mark>CGA</mark> CT <mark>AT</mark> IGIA <mark>CGA</mark> AA <mark>AT</mark> A	T <mark>acc</mark> ttacaai - <mark>acc</mark> cc <mark>ac</mark> cc <mark>i</mark>	AGTGI <mark>AAC</mark> II ACTAA <mark>A</mark> CCC	C <mark>IICAA</mark> I A <mark>II</mark> A <mark>AA</mark> CGCC	IGI <mark>CA</mark> IIA IG <mark>GCA</mark> GCC	CTTCAG GGAAGC	GI CI
	SARS-CoV-2 11 168 - 29 055 pt	15046			70 ,1508	0 ,15090	15100	15110		20 151	78 130
	H. sapiens mitochondrion	AT <mark>T</mark> C <mark>GCA</mark> G	GG <mark>A</mark> TTTC <mark>TC</mark> A	TAC <mark>T</mark> AAC <mark>AA</mark>	CATTTCCCCCG	CATCCCCC-T	CC <mark>A</mark> AAC <mark>AA</mark> CA	ATCCCCCTCT	ACC-TAAA	ACTCAC	: A 0
	SARS-CoV-2 11,168 - 29,055 nt	<u>15131</u> TTG <mark>T</mark> ATTA	,15140 A <mark>CACAGTTAC</mark>	,15150 TICA <mark>CT</mark> TCA <mark>G</mark>	,15160 <mark>act</mark> at <mark>ta</mark> ccag	15170 CIG <mark>IA</mark> CI <mark>C</mark> AAC	15180 TCAAT <mark>T</mark> GAG <mark>T</mark>	15190 <mark>a</mark> cag <mark>a</mark> c <mark>ac</mark> tg	15200 GTG <mark>T</mark> TG <mark>AA</mark>	152 C <mark>A</mark> TGT <mark>T</mark>	215 A (
	H. sapiens mitochondrion	ccc <mark>1</mark>	- <mark>C</mark> G <mark>C</mark> T <mark>GT</mark> C <mark>AC</mark>	TTTC <mark>CT</mark> AG- <mark>G</mark>	<mark>act</mark> tc <mark>ta</mark> acag	CCC <mark>TA</mark> GA <mark>C</mark> C	<mark>TCAA</mark> C <mark>T</mark> ACC <mark>T</mark> .	AACC <mark>A</mark> - <mark>AC</mark> AA	.act <mark>t</mark> aa <mark>aa</mark> S	T <mark>A</mark> AAA <mark>T</mark> ection 18	с <mark>с</mark> 80
	SARS-CoV-2 11,168 - 29,055 nt	15216	152. ATCT <mark>ACA</mark> ATA	30 ,1524 A <mark>AATTGT</mark> TGA	40 1525	0 15260	15270, AAT <mark>TCACACA</mark> A	15280	1529	90 153 <mark>gg</mark> ag <mark>tt</mark>	300 G
	n. supiens mitochondrion	15204	15210	15200	15220	15340	15350	15260		ection 18	81
	SARS-CoV-2 11,168 - 29,055 nt H. saviens mitochondrion	TAATCO	AGTAATGGA	ACCAATTTAT CCCTACTCCT	GATGAACCGAC CCTAGACCTAA	GAC <mark>GACTA</mark> CT CCTGACTAGA	GCGTGCCTTT AAGCTATTAC	G <mark>TAA</mark> GC <mark>ACAA</mark> CTAAA-ACAA	GC <mark>TGA</mark> TGA TTTCAC-A	GTACGA GCACCA	AC A2
		15386	.154	00 .154	10 .1542	0 .15430	.15440	.15450	S	ection 18	82 47(
	SARS-CoV-2 11,168 - 29,055 nt H. saviens mitochondrion	TTATGTAC TCTCCAC	TCATTCGTT	TCGGAAGAGA ACCTCA-ACC	CAGGTACGTTA	ATAGTTAATAG	CGTACTTCTT CTTTACTTCCT	TTTCTTGCTT CT-CTTTCTT	TCGTGGTA -CTTCCCA	TTCTTG CTCATC	

**Figure 4.** (**A**) Sequence alignment between Wuhan *SARS-CoV-2* and *Candidatus Carsonella ruddii*. (**B**) Sequence alignment between Wuhan *SARS-CoV-2* and *E. coli K12*. (**C**) Sequence alignment between Wuhan *SARS-CoV-2* and human mitochondrial DNA. The yellow highlight designates sequence identities between the genomes of *SARS-CoV-2* and two segments of the proteobacterium *Candidatus Carsonella ruddii* (**A**), a segment of the *Escherichia coli K12* genome (**B**) or mitochondrial DNA of *Homo sapiens* (**C**). The blue highlight stands for sequence identities between the nucleotide sequences of two different segments of the *Candidatus Carsonella ruddii genome*.

Random DNA	Homology	
Random DNA 1 versus SARS-CoV-2	44.7%	
Random DNA 2 versus SARS-CoV-2	44.8%	_
Random DNA 3 versus SARS-CoV-2	44.4%	Average: 44.5%
Random DNA 4 versus SARS-CoV-2	44.5%	_
Random DNA 5 versus SARS-CoV-2	44.3%	_

**Table 2.** Randomized computer-generated nucleotide sequences of 30 kbp and their sites of sequence coincidences with the sequence of the *SARS-CoV-2* genome.

# 3.2. Comparisons of Selected Segments of SARS-CoV-2 RNA with Segments of High Sequence Homologies in the Genomes of Diverse Viral and Non-Viral Families

In Figures 1–4, selected frame shots from total sequence alignments between genomes of many different viruses, proteobacteria, bacteria, plus human mitochondrial DNA and *SARS-CoV-2* RNA were reproduced. In Figures 1–4, sequence identities in specific genome parts were highlighted in yellow. Data compilations in Table 1 present overall sequence similarities across the entire genomes of the listed species and the *SARS-CoV-2* Wuhan reference RNA sequence. Table 1 also contains sequence homologies in selected frame shots with high values of sequence identities. Finally, data from a control series that aligned the *SARS-CoV-2* RNA sequence to computer-randomized nucleotide sequences of 29.9 kbp length, have been presented as comparisons to the five computer-generated sequence artifacts (Table 2).

- 1. Comparisons of the Wuhan *SARS-CoV-2* nucleotide sequences to sequences from diverse virus families
- (i) Genomes of SARS-CoV-2 and human adenovirus types 2 and 12

In Figure 1A, the sequence of *SARS-CoV-2* RNA from the coding part of the viral spike protein between nucleotides 1863 and 2365, both from the Wuhan reference sequence and that of the delta variant, were aligned with *human adenovirus type 2* (*Ad2*) sequence between nucleotides 31,081 and 31,668, a segment coding for part of the *Ad2* fiber protein. In the analyzed genome segment, the Wuhan and delta *SARS-CoV-2* sequences differed in only one nucleotide in position 31,149. Segmental homology/sequence identity in the depicted section between the *SARS-CoV-2* and *Ad2* sequences was 43.2%. The lengths of identical micro-modular elements between the two biologically unrelated sequences ranged from n = 2 [frequent] to n = 7 [to n = 10 with one interruption, see bottom array].

We also compared the amino acid sequences of the *Ad2* fiber and the *SARS-CoV-2* spike proteins (https://pubmed.ncbi.nlm.nih.gov/; accessed on 9 September 2021). There were no common sequences, not even at the level of oligopeptides. The four-letter system of DNA/RNA encoding represents an amazingly simple framework at the statistical level but facilitates a high level of micro-modular homologies and, above all, has generated a recombinatorial platform for the enormous complexity of life.

This overall theme—occurrence of micro-modular sequence identities among unrelated viral nucleotide sequences and those from bacteria or human mitochondrial DNA will be documented by a series of relevant examples.

The alignment of the adenovirus type 12 sequence between nucleotides 5989 and 6640 to the 6112-6790 region in *SARS-CoV-2* RNA revealed further micro-modular sequence homologies between n = 2 and n = 6 nucleotides with several closely-spaced alternating spots of homology and non-homologies that are thought to favor recombinatorial integration events (Figure 1C).

(ii) Genome comparison: SARS-CoV-2 and human immunodeficiency virus-1 (HIV-1)

Figure 2A depicts the sequence alignment between the designated regions of the *SARS-CoV-2* and *HIV-1* genomes with an overall homology of 45.8%. Apart from the extent of this homology, this arbitrarily selected frame shot demonstrates the staccato-like

alternations between short modules with and those without sequence homologies. In a different field of our earlier research, i.e., the characterization of sequence elements at the sites of integrative recombination between *Ad2* or *Ad12* DNA and (hamster) cellular DNA in *Ad2*-transformed hamster cells or in *Ad12*-induced hamster tumor cells, that carry multiple copies of integrated *Ad2* or *Ad12* DNA, very similar patchy sequence homologies have been described (see Section 5).

# (iii) Genome comparison: SARS-CoV-2 and of an archaea phage

*Archaea* are extremophiles that constitute the third branch of living organisms [7]. They have been categorized together with bacteria as prokaryotes and are distinguished from eukaryotes by the absence of a distinct nucleus. There are a considerable number of archaea phages whose nucleotide sequences have also been compared to that of *SARS-CoV*-2 RNA (Table 1). Overall sequence homologies range from 30 to 45, and in selected genome segments can reach 47%.

As an example, Figure 2B presents a screen shot of the genome segment of Wuhan reference *SARS-CoV-2* RNA between nucleotides 17,921 and 18,480 that encodes part of ORF1b. It likely codes for the *SARS-CoV-2* exoribonuclease (nsp14) and/or other nsp (non-structural) proteins involved in *SARS-CoV-2* replication [8]. This *SARS-CoV-2* genome segment has been aligned to the genome segment between nucleotides 13,472 and 14,001 of the archaea phage *Sulfolobus turreted icosahedral virus*. In this particular genome segment, there is 45.5% sequence homology between the two viral genomes with micro-modular sequence patches of between n = 2 and n = 5 (6) nucleotides, the latter encompassing one mismatch.

#### (iv) Genome comparison: SARS-CoV-2 and plant Alternaria brassicola endornavirus

In selected segments, there is also up to 44.5% nucleotide sequence homology between the genomes of *SARS-CoV-2* and the plant virus *Alternaria brassicola endornavirus* (*AbEV1*) (Table 1). In the screen shot segment shown in Figure 3A, sequence homology between the two genomes amounts to 44.5%. There are distinct patterns of micro-modular arrays with sequence tracks of 2 to 10 nucleotides identity. It might be the staccato-like alternations of short sequence identities and non-homologies that can serve as target sites for recombinatorial mechanisms.

 (v) Genome comparison: SARS-CoV-2 and insect Autographa californica nucleo- polyhedrovirus (AcNPV)

As shown in Figure 3B and described in its legend, patchy genome homologies between 2 and 6 nucleotides also exist between these taxonomically unrelated viruses. Again, several sites of alternating homologous and non-homologous regions generate recognizable patterns with potentially functional significance.

- 2. Homologies between SARS-CoV-2 RNA and proteobacterial and bacterial genomes
- (i) Genome comparison: SARS-CoV-2 and Candidatus Carsonella ruddii strain BT

Psyllids are primitive plant-specific and plant-feeding insects that harbor obligate endosymbionts like *Candidatus Carsonella ruddii*, a proteobacterium with one of the smallest known genomes, a circular DNA molecule of 174,004 nt [9]. We have aligned this genome with that of *SARS-CoV-2* and registered 8.5% overall and 46% homologies in specific segments of the viral and proteobacterial genomes (Table 1).

In the *SARS-CoV-2* genome segment between nucleotides 22,352 and 22,773 located in its *spike* protein coding region that was compared to sequence positions 159,853 to 160,314 in the *Candidatus Carsonella ruddii* strain BT genome, sequence homology reached 44.8% (Figure 4A).

#### (ii) Genome comparison: SARS-CoV-2 and Escherichia coli strain K12

Significant sequence homologies between *SARS-CoV-2* RNA and segments of the *Escherichia coli K12* genome were also documented (Table 1). The graph in Figure 4B visualizes one prominent example with the nucleotide 22,572 to 23,116 *spike*-coding region

in the *SARS-CoV-2* genome exhibiting 47.3% homology to the *E. coli K12* genome region between nucleotides 153,328 and 153,888. Comparable findings were documented in the adjacent genome segments (data not shown). (i) 48% homology to nucleotide segment 4317 to 4845 in *SARS-CoV-2 PL proteinase* (ORF1a) to an *E. coli K12* genome region of nucleotides 255,598 to 256,156. (ii) 46.9% homology to nucleotide segment 22,300 to 22,813 in *SARS-CoV-2 spike* protein to an *E. coli K12* genome region of nucleotides 270,286 to 270,851. In both comparisons the lengths of micro-modular homologies vary between n = 2 and n = 7. More extensive comparisons between the entire *E. coli* nucleotide sequence and the *SARS-CoV-2* genome will be described in Section 4.

3. Genome comparison: SARS-CoV-2 RNA to human mitochondrial DNA.

Current concepts pursue the notion that during evolution, human mitochondrial DNA might have been derived from a prokaryotic endosymbiont, possibly related to bacterial plasmid DNA. The data in Figure 4C reveal patchy nucleotide sequence homologies of 45.3% in the SARS-CoV-2 region 11,168–29,055 nt compared to the entire mitochondrial DNA sequence. The latter falls in the mitochondrial DNA region coding for the ND5 [*NADH dehydrogenase*] gene. The lengths of the patchy sequence homologies in this segment range between two and seven nucleotides. There are multiple segments that exhibit closely spaced short sequence identities interrupted by one or two non-congruent nucleotides. As suggested in the description of micro-modular sequence arrays with several viral genomes and *SARS-CoV-2* RNA, similar patchy sequence homology arrangements have been observed at junction sites between adenoviral or *SV40* DNA [10] and cellular DNA at sites of adenoviral DNA integration due to recombinatorial integration events [11–16].

4. Genome comparison: *SARS-CoV-2* RNA to the entire *Escherichia coli K12* chromosome of 4,641,652 nucleotide pairs

The comparison between the 29,903 nucleotide sequence of *SARS-CoV-2* and the entire *Escherichia coli K12* nucleotide sequence of 4,641,652 nucleotide pairs produced a statistically very rigorous insight into nucleotide sequence relationships. The data in Figure 5 and Figure S5A,B (Supplemental Materials) confirm the occurrence of n = 3-mers to n = 10-mers, as described above, for many different sets of sequences but fail to reflect the actual complexities of the micro-modular arrays (as apparent from the sequence alignments in Figures 1–4).



**Figure 5.** Histograms of nucleotide sequence identities between the *SARS-CoV-2* genome and different segments of the entire *Escherichia* (*E*) *coli K*12 DNA as indicated in individual panels. For details of the programmed analyses of the entire *E. coli K*12 genome see Materials and Methods. Very similar data obtained for the *E. coli* DNA segments 1 to 1.8 million and 3,400,001 to 4,641,652 have been deposited under Supplemental Materials (Figure S5A,B).

5. Statistical analysis (bar graphs) of inter-genomic micro-modular homology clusters of 3 to 9 and up to rarely occurring 13 nucleotides.

The actual patterns of homologies between the different analyzed genomes are characterized by high complexity. There are constantly shifting runs of sequence homologies and non-homologies that change from segment to segment. A realistic approach to their evaluation can be reached by scrolling through the entire sequence alignments of individual genomes with the nucleotide sequence of SARS-CoV-2 (Supplemental Materials Figures S1–S4). As a manageable method of analyses, the number of 3- to 9-mers, plus the very rare 10- to 13-mers of homologies were counted and subsequently in part analyzed by a computer program. The results of this quantitation for the following eleven organisms (Table 1 and Figures 1-4), selected as examples, were converted to the bar graphs in (Figure 6): Sulfolobus turreted icosahedral virus, an Archaea virus (Figure 6A), Alternaria brassicicola betaendorna virus, a plant virus (Figure 6B), Fig badna virus 1, another plant virus (Figure 6C), Autographa californica Nuclear Polyhedrosis Virus, AcNPV, a baculovirus of hymenoptera (Figure 6D), Bacteriophage lambda of Escherichia coli K12 (Figure 6E), Human Adenovirus type 12 (Figure 6F), Human Hepatitis B virus, HBV (Figure 6G), Human Immunodeficiency Virus type 1, HIV (Figure 6H), the Proteobacterium Candidatus Carsonella (Figure 6I), Escherichia coli K12 (human endo-bacterium), sub-segment comprising nucleotides 255,328–290,948 (Figure 6J), and Homo sapiens, mitochondrial DNA (Figure 6K). The numerical data for the composition of these graphs has been submitted as Table S2 under Supplemental Materials. The graphs drawn from a control computer-aided evaluation of n-mers of homology between the entire 4,641,652 nucleotide sequence of Escherichia coli DNA and SARS-CoV-2 RNA, were presented in Figure 5 and Figure S5A, B. The graphs are very similar to the one presented in Figure 6.



Figure 6. Cont.



Figure 6. Table S2. Bar graphs. As described under Methods, the individual nucleotide sequences as indicated in the Table/Figures were aligned to the DNA-converted RNA sequence of SARS-CoV-2. We analyzed these complex patterns of sequence identities by counting the total number of 3 to up to maximally 13-mers in each of the sequence alignments chosen for representative organisms. As apparent from the nucleotide sequence alignments [see Supplementary Material Figures S1–S4], sequence identities were highlighted in yellow. Individual identities of n-mers were quantitated and expressed as percent of total 3- to n-mers, as indicated. (A-K) summarize the results in the form of bar graphs, in Table S1 the actual numbers of 3- to 13-mers have been compiled. The bar graphs did not, however, reflect the complexity of actual patterns of sequence identities in individual classes that wee, however, apparent in the sequence alignments in Figures 1-4. The nucleotide sequences of the following organisms were aligned with the genome sequence of SARS-CoV-2: (A) Circular doublestranded DNA with 17,663 nucleotides of the Archaea phage Sulfolobus turreted icosahedral virus. (B) RNA with 10,290 nucleotides of the plant virus Alternaria brassicicola betaendornavirus. (C) DNA with 7140 nucleotides of the plant virus Fig badna virus 1. (D) Double-stranded DNA with 133,894 base pairs of the hymenopteran insect virus Autographa californica Nuclear Polyhedrosis Virus. (E) Doublestranded DNA with 48,502 nucleotide pairs of Escherichia coli bacteriophage lambda. (F) Double-stranded DNA with 34,125 nucleotide pairs of human adenovirus type 12. (G) Partly double-stranded DNA with 3182 nucleotide pairs of human Hepatitis B Virus (HBV). (H) Single-stranded RNA with 9719 bp of Human Immunodeficiency Virus-1 (HIV-1). (I) Circular double-stranded DNA of the proteobacterium Candidatus Carsonella with 174,004 nucleotide pairs. (J) Sequence sub-segment comprising nucleotides 255,328 to 290,948 (arbitrary selection) from the 4,608,319 bp double-stranded circular genome of Escherichia coli K12. (K) Circular double-stranded DNA of Homo sapiens mitochondrial DNA with 16,269 base pairs.

Among the micro-modular clusters of sequence identities, 3-mers (54 to 78), 4-mers (18 to 27), and 5-mers (6 to 14) were found in all inspected sequence alignments. Very long stretches of identities (10-, 11-, 13-mers) were solitary and found only in some of the alignments (Figure 6A–C,I–K). Aside from statistical considerations, the occurrence of all homologies, particularly the longer ones, facilitates sequence recognition in DNA-protein interactions and provides platforms for integrative recombination. The frequencies of distinct micro-modular clusters do not, however, sufficiently reflect the actual mosaic of modular patterns and their unknown recognition functions (See alignments in Figures S1–S4). In these patterns, homologies alternate with non-homologies leading to patchy sequence arrangements that might have signal potential.

#### 6. Control analyses with randomly computer-synthesized 29.9 kb DNA

The data presented above suggest significant patch type micro-modular homologies between the 29.9 kb Wuhan sequence of *SARS-CoV-2* that was chosen as a highly relevant reference genome to conduct sequence comparisons in a current biomedical context and a selection of very different genomes. What controls could be performed to reflect on the biological significance of the 44.2 to 47% sequence homologies between *SARS-CoV-2* and a spectrum of viral, proteobacterial, bacterial, and human mitochondrial DNA sequences? Obviously, there is no planet B harboring life with a completely different evolutionary history. Instead, five computer-generated, random nucleotide sequences of 30 kb lengths (Figure 7C) were chosen for comparisons among these artefactual sequences and of some of the randomized sequences to the Wuhan *SARS-CoV-2* sequence (Figure 7A) or to the sequence of *bacteriophage T4* (Figure 7B).

A selected frame in Figure 7A shows the sequence alignment between *SARS-CoV-2* sequence and one of the randomized nucleotide sequences and demonstrates the occurrence of very similar patchy sequence homologies at about comparable frequencies as discussed in Figures 1–4. In another example, a 30 kb randomized sequence was compared to the DNA sequence of *bacteriophage T4* (Figure 7B). The nucleotide sequence of *bacteriophage T4* DNA (nucleotides 5858 to 6442) was aligned to the sequence of a 30 kb computer-generated randomized sequence (nucleotides 5803 to 6390). Sequence homologies of micro-modular patches of 2 to 10 nucleotides lengths amounted to 49%. Numbers represent lengths of homologous matches between the two sequences, dots stand for non-homologies. In Figure 7C five randomized 30 kb nucleotide sequences have been aligned to each other and again exhibit micro-domains of homologies among each other.

This series of random sequence controls was expanded in two ways. First, the randomization was now based on the 29.903 nucleotide length and additionally on the specific nucleoside composition of the *SARS-CoV-2* RNA molecule. Second, a total of 1000 such randomized sequences were analyzed by the program described in Section 2. Figure 8 documents that the frequency distributions of 3- to 9- (10-)-mers of sequence identities among 1000 randomized A, C, G, T sequences followed the same patterns very closely as shown in Figure 5A–K and Figure 6 for comparisons between the *SARS-CoV-2* and the analyzed nucleotide sequences in this report. Hence, the data further support the notion that platforms of annealable nucleotide sequences can be attributed to the inherent characteristics of the four letter code that generated all of biology.

		21650 21670	21690	21600 21	700 21710	21720	21720	- Section 239
A	SARS-CoV-2 Random DNA 09.09.2021	GCATAAIGAIGAAIG GIGCICCITIAAI	ICGCAAAATATACI ICCCTAAATGACC	CCAACTGTGTCAATA ATGGACG <mark>CA</mark> GGO	ATTAAACACATTAA SC <mark>TT</mark> GG <mark>AC</mark> GCCCTAA	CATTAGCTGTA CTCTA-CTATA	CCCTATAAT SACGC <mark>T</mark> GG <mark>T</mark> GGC	-ATGAGA-GTT CAGTAGACGTT - Section 240
	SARS-CoV-2 Random DNA 09.09.2021	21750 21760 ATACATITIGGIGCI ATICCIGGGGGACCC	21770 GG <mark>T</mark> ICIG <mark>A</mark> T7 AA <mark>I</mark> CGACGACA <mark>A</mark> G3	21780 217 LAAGGAGTTGCACC LAAAGCTTCGCACC	90 <u>21800</u> GGTAC <mark>AG</mark> CTGTTTT GG <mark>AG</mark> GC <mark>G</mark> AACO	,21810 T <mark>aagacagtggt</mark> Gaac <mark>agactag</mark> a	21820 Igcct <mark>acgg</mark> g Iaaag <mark>a</mark> tgacga	21830 21840 TACGCTGCTTG TGGGGGCTGG
	<i>SARS-CoV-2</i> Random DNA 09.09.2021	21841 21850 TC <mark>GATT</mark> CAGATCTTA <mark>GATT</mark> TGTCCG <mark>T</mark> GT	,21860 ATG <mark>ACT</mark> TTG <mark>TCTC</mark> ACC <mark>ACT</mark> ACA <mark>TCTC</mark>	21870 2188 Gatgcagattcaac G-tccagctgta	0 <u>21890</u> TIT <mark>GATIG</mark> GIG <mark>A</mark> II TGG <mark>GGIIG</mark> C <mark>A</mark> AG	21900 GIGCA <mark>ACIG</mark> IA GGAGI <mark>A</mark> G <mark>IG</mark> AG	21910 2 AT <mark>ACAGCTAAT</mark> TA <mark>AC</mark> C <mark>GC</mark> A <mark>AAT</mark>	- Section 241 1920 21931 AAA <mark>T-G</mark> GGA <mark>TC</mark> CTG <mark>TCG</mark> ATG <mark>TC</mark>
	SARS-CoV-2 Random DNA 09 09 2021	21932 21940	21950 2 TATGTACGACCC	21960 21970	21980	21990	22000 22	- Section 242 D10 22022 TTACATTTGTG GATAGTGTTTG
	SARS-CoV-2	22023 22030 GGTTTA <mark>TA</mark> CAACAAA	22040 22	2050 22060	22070	22080 22	2090 ,2210 30 <mark>16a1</mark> 0 <mark>1</mark> 17a1	- Section 243 00 22113 AAGCTCATGGG
	Random DNA 09.09.2021	22114 22120	22130 221	CATGGCATGAATA	22160 2	-ACIGIGCCGCI:	111160161ACAG	ATCGATATGCC - Section 244 ) 22204
	SARS-C0V-2 Random DNA 09.09.2021	22205 22210	22220 2223	0 22240	22250 22	2260 2227	GATGTAAATAT GAATGTAAATAT 70 22280	CCACTATTCCG - Section 245 22295
	<i>SARS-CoV-</i> 2 Random DNA 09.09.2021	AC <mark>C</mark> AC <mark>G</mark> CGAACAA) AG <mark>C</mark> TG <mark>G</mark> GTT <mark>ACA</mark> GTC	A <mark>TAG</mark> ATG <mark>GTTAIG</mark> I C <mark>T</mark> G <mark>G</mark> GGC <mark>GTT</mark> T <mark>TG</mark> A	CCA <mark>TGC</mark> AT <mark>GC</mark> AAA <mark>T</mark> AGT <mark>TGC</mark> CG <mark>GC</mark> <mark>T</mark> O	ACATATITIGGAGO ACGG <mark>AT</mark> ATAACGII	G <mark>AATA</mark> CA <mark>AAT</mark> CC) ( <mark>ACTA</mark> TG <mark>ACT</mark> TG)	A <mark>ATT</mark> CAGT <mark>TG</mark> TC 5 <mark>ACT</mark> ATAG <mark>TG</mark> CA	TIC <mark>CTATIC</mark> TT <mark>Caat</mark> gCaa
В	Enterobacteria phage T4 Random DNA	6091 61	05 6116	6127	,6138 TTCAATAGTAACTI	6149	6160	ection 71 6177
	Enterobacteria phage T4	6178 £	6193 (620 TC <mark>TC</mark> TTCCT <mark>GC</mark> TCT	4 <u>6215</u>	,6226 CTGATAT <mark>GGA</mark> TT <mark>T</mark> I	,6237	6248	Ction 72 6264 CTTCGT
	Random DNA Enterohacteria nhage T4	6265 6270	6281 62	TGAGGCTCAACCGC 92 <u>,6303</u> CGTCAAGGATATCA	6314	6325	TACGCTCCGAA 	ACTTAAG ection 73 6351 IGICATA
	Random DNA	GCTTTGCGAGTTTGA	6369 6	T <mark>GTCAAGG</mark> T <mark>TTCC</mark> G 380 ,6391	GT <mark>GTCTGC</mark> CTCGTC β402	CAGCATCTCGG	STATG <mark>C</mark> GG <mark>IG</mark> TG S S	GTA <mark>CA</mark> AC ection 74 6438
	Enterobacteria phage T4 Random DNA	ATTTTAGTAGGACGCC AGCTCAGC-GGGGGGCC	CAAGATAATCAGO GIG <mark>G</mark> I <mark>C</mark> GATI <mark>G</mark> G	ATCTAAAATTTTAA A-CICTCACCGAGG	AAGGCICCAIACCI AGGGAICICGCI	AAACGICGIGCA AGAICCIAIA 6501	TAGATTCCGTT GAATTCACGT 6512	ATCAGTA AGAGGGA ection 75 6525
	Enterobacteria phage T4 Random DNA	TGGTCTTTAATAAAA ACGACGTAATTCCCA	A <mark>TT</mark> AGCGAG <mark>GC</mark> AGG	TTTATTTTT TGTTCCCACTCTTG	AATIGATACO G <mark>AT</mark> G <mark>GA</mark> CAAATGCO	ATT <mark>TA</mark> TTAACTT AAA <mark>TA</mark> GCT <mark>A</mark> GGO	ICAAAT <mark>ITA</mark> AT <mark>A</mark> SGGGCC <mark>ITA</mark> GG <mark>A</mark> SGGGCC	CCATTA SATACAA ection 76
	Enterobacteria phage T4 Random DNA	6526 ,6534 ATAAATTICCTICCA ACAAIGGIICCIGCA	,6545 STAAGTIG <mark>IGCCG</mark> I A <mark>IA</mark> ICGA- <mark>IG</mark> G <mark>CG</mark> A	6556 65 CTTCAGTAATTTCA GTTGTAATT-CA	67 6578 cgaaa <mark>taca</mark> ccata ttgcc <mark>taca</mark> ggtaa	,6589 AATT <mark>GGCT</mark> GTT TTGA <mark>GG</mark> T <mark>T</mark> CG <mark>T</mark> 3	6600 TAT <mark>CACCG</mark> ACTT ACA <mark>CTCGG</mark> GG <mark>T</mark> A	6612 ICTCATA AAAG <mark>AT</mark> G ection 77
	Enterobacteria phage T4 Random DNA	6613 6622 CACATAAACAGAAGT GGGGGGAGATCATT	, <mark>6633</mark> Daa <mark>giga-</mark> giaaac Cg- <mark>giga</mark> c <mark>giaaa</mark> i	, <mark>6644 6</mark> T <mark>igctagtatgti</mark> c g <mark>t</mark> tac <mark>ag</mark> a <mark>a</mark> gtc <mark>t</mark> i	655 6666 C <mark>TITTTGAACTACT</mark> A T <mark>T</mark> CG <mark>T</mark> AT <mark>ATAACT</mark> G	,6677 ICCA <mark>AAT</mark> IT <mark>GGAI</mark> ATT <mark>AAT</mark> GA <mark>G</mark> C <mark>AI</mark>	6688 6688 CAA <mark>ATAA-TAC</mark> CGC <mark>AAAAGTAC</mark>	6699 ATCTT <mark>CA</mark> FAGCA <mark>C</mark> T
С	25057	25070 250	080 ,25090	25100	25110 2512	0 ,25130	25140	Section 233 25150 25164
	Random DNA 1 CACAAGAGC Random DNA 2 GDCC GGCC Random DNA 3 TCGC AGGC Random DNA 4 TCGC CAGT Random DNA 5 CCGC ACGC	IGITIATITACICCI GTITCACCACIC IIGITAIGAATO GIGGGCATAIIG CCAGG <mark>AII</mark> T <mark>CI</mark> C	CCGTATTCGGGGT ACAGTTA-GATGG ACCGGTGACGT GGTGGGTTG GGGGGGGCTG	GCACCCATTGCT CCACGATGTGTAC -CACTCACGGAGT GCAATTACGCTCT CCAAGATGACGAC	TETTTGTCGTCTG ATATCTATATTGCC AGGCCAGTGTAGG AGACACGTGAAGG AG-CATCTGTCCC	TCGTGTTGAAT CTATG- <mark>ACAA</mark> TG ACCGT <mark>ACAATT</mark> SCT <mark>CACCT</mark> G <mark>AGAA-T</mark>	GTGTATCATCG AGTTTATTAAG GCGAAAAAT CGTCCAGGCAGT AGATA <mark>AA</mark> CCT <mark>AT</mark>	CGC GGA CATTACTGGGG CATGTCTC - TATACCG- AATGTGGAATTATGAAAC GATGTGACGTATAGGT- GTC CG-GGTCTCGA- Catton 224
	25165 25165 25170   Random DNA 1 AIAGTITAA   Random DNA 2 TCCSGGGGG   Random DNA 3 CCSGGGGG   Random DNA 4 GCAGAAGTC   Random DNA 5 CCSACAGG	25180 TTO CAATAGGCCTAT TTAGCT TTAAATTATG TCAGT GTAAAGCCCT TAAATCAACTTCTT GGCTT GCTACGGCTT	25190 2520 TGGGGTACTOTC CATGTTGATCTG - ACCAGCTAGTG TGGAACCACT- GGGAAGAAACGG	0 25210 ACCAGGAAAGAA ATCACCCC ACTCCGAGCTATCC TAGGGATAGT TAACTCAGGATT	25220 25 CGTGATCGATAATO CACTGACGTGA CACTAGCGTGA CACCACCGGTGTAACO GACCACCGATA	2230 2524 CAACGCGGATTT 	0 25250 AGACACAAAG-G GATATGTCGTG AATCATIGT- ACCCCAACGGTG AGAGT <mark>ATGG</mark> -G	25260 25272 AAG CITEICCCANTA AAG GOA CICGIOCO AAAAAA COCCICCICCA AAAAAAAAA CIGIICCCCAAAAAAAAAAAAAAAAAAAAA
	25273 252 Random DNA 1 TGT TTGC Random DNA 2 TG - CGCA Random DNA 3 CTAGAGGCT Random DNA 4 CGCTCCCC Random DNA 5 G CGCCT	280 25290 TAATT-COASTCTSCC AATAAGTSCSAATCCC GATCCASTGSAA GCGCA-CSASACTCAT ITGCG-TSASACTCAT TTGCG-TSASAAGCA-	25300 25 IGGTCCGAAGTIGA IAG-CCGITGTC-C CCGCGCGC-1 GACTCCGCGAAGGI CCGCGAGGTCA1	310 25320 CGACGGATCCTTC TATCAGTCGTGT CACGCCAGACCT CGTCCCCGATCGC TCTCACACACGCT	25330 CGGAATCGGAGACGT CGGTCCGGAACGT GGGTCCGGAAAGGT ACT-CTCCGTAGGA ACAGCTCCCGAGCG	25340 25 TTGTAGTGTGC TATACTGCGACA CATGCCTAAACC AGGCCGAGGATGC ATGCGTAGAAC	5350 2530 ACACCTACACCG AGAGAAA TCATGA TAAATATT TGGGTATG	CO 25370 25380 CC AGC CC CT CA CAT CC AGC CT CT CA GAC ACCACCT CT CC AGCC ATECTA CT CA CCT ACCACCT CT CA CCT ACCACCT CT CA CCT CT TTA CT CC CC ACCT Section 236
	25381 Random DNA 1 AACCOTTC Random DNA 2 ACTAGAACG Random DNA 3 CGATGGTTG Random DNA 4 CGATGGTTG Random DNA 5 AATAGATT	25390 25400 CAACC COTAGINIT IACTC DEGGGGG SGCTC AAAIGA ATAT DOGGCGTI GTATCGACCCCCTAT	25410 IGC <mark>GAA</mark> GCGCTIGG SATGGIC-CGCCCC ACTGAA <mark>GICI</mark> I ACTGAAGICGIGIGI ITTGAAGICGIGIGAGGI	25420 ,2543 TACTCTTCCCAAA TGCCAAGCTACAG TCTATGCCAAGA TCTATGACCAAAT TCCA-GTCCAACG	0 25440 LCGT GAACTA-ACA TAT GAGCA CGT GCGCA TAT GGGCA TAT GGGCA CAGACTGCTACA	25450 GCGTIGCAIG GCACTIAAACO CCCAC CCCA-ACO CCCA-ACO CCCACAGI	25460 2 GGGGAACCACCA GGTCCACCGCAA T-CGGAGTCCA T-CGATGTTGT TACGATTTCCA	5470 25488 TATAT GAAG TAGCCCTATCACAA ACCCGACTGAGAA AACTTGAACAATACA GCATTACTGGTCTTGGT

**Figure 7.** (**A**) Sequence alignment between Wuhan *SARS-CoV-2* and a randomly computer-generated DNA. (**B**) Sequence alignment between *bacteriophage T4* and a randomly computer-generated DNA. (**C**) Sequence alignments between five randomly computer-generated DNAs. The different highlights have been explained in the legends of Figures 1–4. The blue highlights in part (**C**) indicate sequence identities among different randomized sequences.



**Figure 8.** By using programs described under Methods, 1000 randomly generated nucleotide sequences were computer-screened for sequence identities among these randomized sequences of 3- to 9-mers (up to 10-mers in rare instances).

#### 4. Discussion and Outlook

#### (i) Synopsis and interpretation of data

A comparison of the nucleotide sequences of the *SARS-CoV-2* genome to a number of viral, proteobacterial, and bacterial genomes as well as to human mitochondrial DNA has revealed common patch-like homologies with micro-domains of up to 40 and 50%, with the highest values in selected regions of these genomes. These homologies comprise between 2 to maximally 13 nucleotides, in rare instances, and are further characterized by longer stretches of sequences with alternating patch homologies interrupted, often by just one nucleotide pair of non-homology. Similar patch-like homologies, sometimes permitting the formation of stem-loop structures (Figure 9), have been previously reported from this laboratory at sites of integrative recombination between *adenovirus* DNA and cellular DNA in *Adenovirus type 2-, Ad12*-transformed rodent cells, and *Ad12*-induced hamster tumor cells [11–16]. Similar patch homologies at junction sites of integrated viral DNA and the recipient host DNA have also been described for *SV40* [10], *Hepatitis B Virus* [17,18]. Patch-type micro-domains have served as signals for heterologous integrative recombination.

Control studies on homologies between the SARS-CoV-2 genome and five artificially computer-generated randomized nucleotide sequences of 29.9 kb have revealed similar patch homologies, possibly of slightly shorter lengths. Control studies on computergenerated randomized nucleotide sequences have been extended to 1000 sequences. Their comparisons to the sequence of SARS-CoV-2 RNA yielded patterns very similar to the ones described for all comparisons described (Figure 8). In light of the results of these control studies, one might question the biological significance of the sequence comparisons reported here. In fact, however, these ubiquitously occurring homologies do exist and are inherent in the reality of a four-letter coding system on which all life is based. We submit that it would be a serious neglect to discount the possible biological significance of such wide-spread sequence homologies and their possible role in integrative recombinatorial events between very different genomes or their fragments. The recombinatorial repertoire available through these ubiquitous nucleotide sequence homologies might have played an important role during evolution. In the current SARS-CoV-2 pandemic, it is worth considering similar mechanisms that have been at work during the continuing evolution of SARS-CoV-2 from undefined animal to human hosts and the generation of ever more efficient variants during the current unabated COVID-19 pandemic. Of interest to note, *SARS-CoV-2*-superinfected *HIV*-positive, often immunocompromised individuals (compare our Figure 2A) have been purported as possible driving forces in the emergence of the *omicron variant* of *SARS-CoV-2* [19].



**Figure 9.** Stem-loop structures at the sites of linkage between *Ad12* DNA and cellular DNA from several *Ad12*-transformed hamster cell lines and *Ad12*-induced hamster or mouse tumor cell lines. A stem-loop structure at the preinsertion site in *BHK21* cells was also shown. Details are described in the text. Double-headed arrows, *Ad12* DNA-cell DNA junction sites; asterisks, *topoisomerase I* recognition sites (from [15]).

 Patch homologies and stem-loop structures between DNA molecules of very diverse origins

In this article, nucleotide sequence alignments and comparisons in a study on genomes of very different organisms—the entire spectrum of viruses, proteobacteria, bacteria (*Escherichia coli K12*), and human mitochondria, presumably of endo-symbiotic origin—have unequivocally documented that a plethora of patch-type or micro-modular homologies exist between these heterogeneous populations of genomes and the *SARS-CoV-2* genome. In addition, randomized, computer-synthesized 29.9 kbp genomes exhibit similar micro-modular sequence homologies among each other and to the *SARS-CoV-2* genome.

In a research program that was pursued over two decades, i.e., an extensive analysis of sites of integrative recombination between DNA molecules that were unlikely to meet in nature, like the DNA of *Ad12* and mammalian (hamster) cellular DNA, almost identical patch structures have been observed (Figure 9, reproduced from [15,16]). Special sequence patterns, including patchy homologies, were also reported for a number of non-homologous recombination events, e.g., the insertion site of unique and repetitive DNA fragments into the *aprt* locus in the hamster genome [20] or a hotspot for novel amplification joints in *Alu*-like repeats [21] as well as for virus-cell DNA integration sites [22,23]. These structural similarities of DNA patterns in several entirely different biological systems cannot necessarily prove that the reaction products would be the same. However, the observation calls for more detailed investigations of the biological meaning of ubiquitous sequence homologies.

(iii) The four-letter coding system and its selection during evolution—a hypothesis

The history of genetic research in the 1930s [24], reminds us that vivid discussions were then dominated by speculations about the chemical nature of the genetic principle or as Oswald Avery and colleagues, who succeeded in identifying it as DNA called it—the *transforming principle* [25]. Today, we have long since internalized that a genetic alphabet of four chemical compounds, often referred to by their initials A, C, G and T, suffices to store all information for the structure and activities for all of biology. Another important aspect in evolution has been the exchange and scavenging of genetic information among different organisms with serendipitous encounters of their genomes. Under favorable conditions, mobile genetic elements, or viral genomes, probably any form of foreign DNA, can find

entry via integrative recombination into established genomes that is often mediated by short sequence homologies. In analyses of junction sites between cellular DNA and different integrating viral genomes, it has been documented that patchy sequence homologies between the recombination partners—viral and receiving mammalian genomes—do occur and can direct the localization in integrative recombination with genomes of foreign derivation. In detail, alternations of base pairs with and without homologies in these patches provide distinct recognition signals that aid targeting possible sites for this type of heterologous recombination. A well-documented example of base pairing of short stretches of nucleotides providing enough stability to ascertain essential biological function is the 12 nucleotide homology in the cohesive termini in DNA of Escherichia virus lambda  $(\lambda)$  [26] that facilitate genome circularization as the first essential step in *lambda* DNA replication or in site-specific integration into the bacterial genome [5' GGGCGGCGACCT 3' and 5' AGGTCGCCGCCC 3']. Hydrogen bond stability, of course, is strengthened in this short sequence by the predominance of C-G base pairs [27]. We intend to pursue the role that micro-module-based sequence platforms might have played and continue to play in the emergence of SARS-CoV-2 and the selection of variants of concern. While this research was initiated with a focus on SARS-CoV-2 genetics, the data presented now bear on a much wider scope and intend to shed light on important mechanisms in past and on-going evolution.

# 5. Conclusions

- Genome structures of several organisms across the evolutionary spectrum were aligned with the SARS-CoV-2 genome sequence. Micro-modular homologies of n = 3 to n = 9, and up to n = 13 (rare case) were recorded to gain insights into generalities in the design of existing and evolving (SARS-CoV-2) genomes.
- We found abundant micro-modular identities across the spectrum of genomes and investigated and documented these data in the bar graphs depicted in Figure 6A–K.
- In a computer-aided sequence comparison, the sequence of the *SARS-CoV-2* genome was matched to the entire 4,641,652 nt DNA sequence of *Escherichia coli K12*. This analysis yielded very similar results (Figure 6 and Figure S6A,B) with respect both to the numerical n = 3 to n = 9, and up to n > 10-mer arrays and to the complex patch pattern sequence homologies, reminiscent of those described in integrative recombination reactions [11–16].
- In an extensive control project, we subjected 1000 randomly generated nucleotide sequences of the length and base composition of the SARS-CoV-2 genome to the same type of analyses for micro-modular sequence identities and found very similar patterns of micro-modular homologies, now presented in Figure 8.
- The four letter A, C, G, T genetic encoding strategy has proven exceptionally successful during 4 billion years of evolution. We submit that the system-inherent capability to generate a vast abundance of sequence and pattern identities alternating with non-identities to the tune of "patchy homologies" has provided opportunities for biochemically meaningful platforms for DNA-protein interactions and for (integrative) recombination reactions. In the course of evolutionary time periods, this system has facilitated the annealing of nucleotide sequences required for repeated extensions of initially smaller nucleotide sequences to the real genomes of enormous lengths that exist today. Moreover, the actually existing patterns (see alignments in Figures S1–S4) offer unidentified recognition signals for recombination events between otherwise unrelated genomes.
- It will take systematic investigations to explore the possibility of whether these mechanisms might have served also in the generation and selection of variants of newly emerging pathogenic viruses like *SARS-CoV-2* and others.

**Supplementary Materials:** The following supporting information can be downloaded at: https: //www.mdpi.com/article/10.3390/v14050885/s1, Figures S1–S4—total sequence alignments; Figure S5A,B—parts of the analysis of the total *E. coli* genome versus *SARS-CoV-2*; Table S1—accession numbers for analyzed sequences; Table S2—actual numbers the bar graphs in Figure 6 were derived from.

**Author Contributions:** S.W. and C.M.R. have equally contributed to this investigation. S.W. participated in the conceptualization of the project, carried out all sequence alignments and part of sequence analyses, interpreted data, and wrote part of the manuscript. C.M.R. did all statistical evaluations, interpreted data, and wrote part of the manuscript. W.D. initiated the project, participated in its conceptualization, the analysis and interpretation of data, and wrote the manuscript. All authors have read and agreed to the published version of the manuscript and the corrected proofs.

**Funding:** In its initial stage, this research was supported by the Robert Pfleger Foundation, Bamberg, Germany.

**Institutional Review Board Statement:** Not applicable. This manuscript deals exclusively with analyses of sequences from generally accessible data bases.

**Informed Consent Statement:** This article does not report data derived from studies on patients but has analyzed published nucleotide sequences.

**Data Availability Statement:** All data presented in this article will be freely available from the authors on request.

**Acknowledgments:** We acknowledge that free access to databases of nucleotide sequences has helped this analysis. W.D. and S.W. are indebted to the Institute for Clinical and Molecular Virology, Universitätsklinikum Friedrich Alexander University (FAU) Erlangen-Nürnberg for their continued support of our Epigenetics Group.

Conflicts of Interest: The authors declare that they have no conflict of interest.

# References

- 1. Boehm, E.; Kronig, I.; Neher, R.A.; Eckerle, I.; Vetter, P.; Kaiser, L. Geneva Centre for Emerging Viral Diseases. Novel SARS-CoV-2 variants: The pandemics within the pandemic. *Clin. Microbiol. Infect.* **2021**, *27*, 1109–1117. [CrossRef] [PubMed]
- Harvey, W.T.; Carabelli, A.M.; Jackson, B.; Gupta, R.K.; Thomson, E.C.; Harrison, E.M.; Ludden, C.; Reeve, R.; Rambaut, A.; COVID-19 Genomics UK (COG-UK) Consortium; et al. SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.* 2021, 19, 409–424. [CrossRef] [PubMed]
- Yusof, W.; Irekeola, A.A.; Wada, Y.; Engku Abd Rahman, E.N.S.; Ahmed, N.; Musa, N.; Khalid, M.F.; Rahman, Z.A.; Hassan, R.; Yusof, N.Y.; et al. A Global Mutational Profile of SARS-CoV-2: A Systematic Review and Meta-Analysis of 368,316 COVID-19 Patients. *Life* 2021, 11, 1224. [CrossRef] [PubMed]
- 4. Weber, S.; Ramirez, C.; Doerfler, W. Signal hotspot mutations in SARS-CoV-2 genomes evolve as the virus spreads and actively replicates in different parts of the world. *Virus Res.* **2020**, *289*, 198170. [CrossRef]
- Weber, S.; Ramirez, C.M.; Weiser, B.; Burger, H.; Doerfler, W. SARS-CoV-2 worldwide replication drives rapid rise and selection of mutations across the viral genome: A time-course study—potential challenge for vaccines and therapies. *EMBO Mol. Med.* 2021, 13, e14062. [CrossRef]
- Anant, S. Davidson NO Molecular mechanisms of apolipoprotein B mRNA editing. *Curr. Opin. Lipidol.* 2001, 12, 159–165. [CrossRef] [PubMed]
- Woese, C.R.; Kandler, O.; Wheelis, M.L. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria and Eucarya. Proc. Natl. Acad. Sci. USA 1990, 87, 4576–4579. [CrossRef]
- Jiang, Y.; Yin, W.; Xu, H.E. RNA-dependent RNA polymerase: Structure, mechanism, and drug discovery for COVID-19. *Biochem. Biophys. Res. Commun.* 2021, 538, 47–53. [CrossRef] [PubMed]
- 9. Nakabachi, A.; Yamashita, A.; Toh, H.; Ishikawa, H.; Dunbar, H.E.; Moran, N.A.; Hattori, M. The 160-kilobase genome of the bacterial endosymbiont. *Carsonella Sci.* 2006, 314, 267. [CrossRef]
- 10. Botchan, M.; Stringer, J.; Mitchison, T.; Sambrook, J. Integration and excision of SV40 DNA from the chromosome of a transformed cell. *Cell* **1980**, *20*, 143–152. [CrossRef]
- 11. Stabel, S.; Doerfler, W.; Friis, R.R. Integration sites of adenovirus type 12 DNA in transformed hamster cells and hamster tumor cells. *J. Virol.* **1980**, *36*, 22–40. [CrossRef] [PubMed]
- 12. Deuring, R.; Winterhoff, U.; Tamanoi, F.; Stabel, S.; Doerfler, W. Site of linkage between adenovirus type 12 and cell DNAs in hamster tumour line CLAC3. *Nature* **1981**, *293*, 81–84. [CrossRef] [PubMed]
- 13. Gahlmann, R.; Leisten, R.; Vardimon, L.; Doerfler, W. Patch homologies and the integration of adenovirus DNA in mammalian cells. *EMBO J.* **1982**, *1*, 1101–1104. [CrossRef]

- Doerfler, W.; Gahlmann, R.; Stabel, S.; Deuring, R.; Lichtenberg, U.; Schulz, M.; Eick, D.; Leisten, R. On the mechanism of recombination between adenoviral and cellular DNAs: The structure of junction sites. *Curr. Top. Microbiol. Immunol.* 1984, 109, 193–228. [CrossRef]
- 15. Knoblauch, M.; Schröer, J.; Schmitz, B.; Doerfler, W. The structure of adenovirus type 12 DNA integration sites in the hamster cell genome. *J. Virol.* **1996**, *70*, 3788–3796. [CrossRef]
- 16. Doerfler, W. 2000 Weinheim, New York, Chichester, Brisbane, Singapore, Toronto; Wiley-VCH: Weinheim, Germany, 2000.
- Shafritz, D.A.; Shouval, D.; Sherman, H.I.; Hadziyannis, S.J.; Kew, M.C. Integration of hepatitis B virus DNA into the genome of liver cells in chronic liver disease and hepatocellular carcinoma. Studies in percutaneous liver biopsies and post-mortem tissue specimens. N. Engl. J. Med. 1981, 305, 1067–1073. [CrossRef] [PubMed]
- Ananth, S.; Morath, K.; Trautz, B.; Tibroni, N.; Shytaj, I.L.; Obermaier, B.; Stolp, B.; Lusic, M.; Fackler, O.T. Multifunctional Roles of the N-Terminal Region of HIV-1<sub>SF2</sub>Nef Are Mediated by Three Independent Protein Interaction Sites. *J. Virol.* 2019, 94, e01398-19. [CrossRef]
- 19. Mandala, W.L.; Liu, M.K.P. SARS-CoV-2 and HIV-1: Should HIV-1-Infected Individuals in Sub-Saharan Africa Be Considered a Priority Group for the COVID-19 Vaccines? *Front. Immunol.* **2021**, *12*, 797117. [CrossRef]
- Nablbantoglu, J.; Miles, C.; Meuth, M. Insertion of unique and repetitive DNA fragments into the aprt locus of hamster cells. J. Mol. Biol. 1988, 200, 449–459. [CrossRef]
- Hyrien, O.; Debatisse, M.; Buttin, G.; Vincent, B.R.d. A hotspot for novel amplification joints in a mosaic of Alu-like repeats and palindromic A1T rich DNA. *EMBO J.* 1987, 6, 2401–2408. [CrossRef] [PubMed]
- 22. Bullock, P.; Forrester, W.; Botchan, M. DNA sequence studies of simian virus 40 chromosomal excision and integration in rat cells. J. Mol. Biol. **1984**, 174, 55–84. [CrossRef]
- 23. Romani, M.; de Ambrosis, A.; Alhadeff, B.; Purrello, M.; Gluzman, Y.; Siniscalco, M. Preferential integration of the Ad5/SV40 hybrid virus at the highly recombingenic human chromosomal site 1p36. *Gene* **1990**, *95*, 231–241. [CrossRef]
- 24. Delbrück, M. A physicist's renewed look at biology: Twenty years later. Science 1970, 168, 1312–1315. [CrossRef] [PubMed]
- Avery, O.T.; Macleod, C.M.; McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.* 1944, 79, 137–158. [CrossRef] [PubMed]
- 26. Wu, R.; Kaiser, A.D. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *J. Mol. Biol.* **1968**, *35*, 523–537. [CrossRef]
- Jaeger, J.A.; Turner, D.H.; Zuker, M. Improved predictions of secondary structures for RNA. Proc. Natl. Acad. Sci. USA 1989, 86, 7706–7710. [CrossRef]