





## A specific class of infectious agents isolated from bovine serum and dairy products and peritumoral colon cancer tissue

Ethel-Michele de Villiers , Karin Gunst, Deblina Chakraborty , Claudia Ernst, Timo Bund  and Harald zur Hausen 

Episomal-Persistent DNA in Cancer- and Chronic Diseases, Deutsches Krebsforschungszentrum, Heidelberg, Germany

### ABSTRACT

The *in silico* analyses of 109 replication-competent genomic DNA sequences isolated from cow milk and its products (97 in the bovine meat and milk factors 2 group – BMMF2, and additional 4 in BMMF1) seems to place these in a specific class of infectious agents spanning between bacterial plasmid and circular ssDNA viruses. Satellite-type small plasmids with partial homology to larger genomes, were also isolated in both groups. A member of the BMMF1 group H1MBS.1 was recovered in a distinctly modified form from colon tissue by laser microdissection. Although the evolutionary origin is unknown, it draws the attention to the existence of a hitherto unrecognized, broad spectrum of potential pathogens. Indirect hints to the origin and structure of our isolates, as well as to their replicative behaviour, result from parallels drawn to the Hepatitis deltavirus genome structure and replication.

**ARTICLE HISTORY** Received 16 March 2019; Revised 17 July 2019; Accepted 17 July 2019

**KEYWORDS** Infectious bovine meat milk factors; new class; plasmid DNA; colon tissue; pathogenesis

### Introduction



A large set of novel viral and phage genomes has been identified during the past several years. These were identified through analyses of the human microbiome [1,2] or of samples from other sources. Metagenomic analyses are commonly applied [3–5], although the application of more conservative techniques also led to the identification of a number of new infectious agents [6–8]. Analyses mostly relied on comparisons to known agents or the identification of a replication gene in the assembled genome. Unfortunately, several studies concentrated only on the presence of known infectious entities, while paying less attention to identifying yet unknown infectious agents [9,10]. One difficulty originates from the discovery of chimeric genomes. This weakened the strict demarcation between certain types or groups of infectious agents resulting in studies investigating their phylogenetic origin [11–13].


We approached the question of as yet unknown agents from a different perspective: published data on the epidemiology of colon and breast cancer suggested milk and meat factors (BMMFs) derived from dairy cattle as presumably species-specific risk for these cancers after consumption of products from these animals [8,14,15]. Therefore, we initiated studies to isolate infectious agents from bovine sera

and commercially available cow milk and its products. Based on our previous experience with single stranded DNA virus genomes [6,16] we initiated experiments by gradient density fractionation of sera followed by rolling circle amplification of DNA obtained. Putative Rep genes were identified as part of the DNA sequences obtained by *in silico* comparisons to available sequences. Amplification using inverse PCR with back-to-back primers in the rep gene led to the isolation of full and partial circular DNA genomes from bovine sera [17].

This was extended to samples from commercially available cow milk for the presence of specific circular single-stranded DNA genomes. Four additional isolates were obtained from human brain and serum (from patients with multiple sclerosis). A total of 18 full-length circular single-stranded DNA molecules (~1100–3000 nucleotides) were cloned and sequenced [17–20]. We divided the initial 18 isolates into four different groups BMMF1 through BMMF4, according to their molecular characteristics [14]. Three of these groups revealed a remarkable degree of similarity to *Acinetobacter baumannii* and *Psychrobacter* plasmids. The fourth group comprised 3 isolates being representatives of *Genomoviridae* (*Gemycircularvirus*).

The isolation of these circular DNA molecules, in part closely related to bacterial plasmids, raised the question whether they represented bacterial

**CONTACT** Ethel-Michele de Villiers  e.devilliers@dkfz.de  Episomal-Persistent DNA in Cancer- and Chronic Diseases, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 280, Heidelberg 69120, Germany

 Supplemental data for this article can be accessed at <https://doi.org/10.1080/22221751.2019.1651620>

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group, on behalf of Shanghai Shangyixun Cultural Communication Co., Ltd  
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

contaminants or whether these were bacteria-related sequences adapted to infect and replicate in vertebrate and human cell. Presently representatives of all four groups have been analyzed for expression of genetic activity and replication in human cells [14,21].

We further extended our search for additional DNA plasmids in the BMMF1 and BMMF2 groups in the present study. We also describe the detection by mass spec of peptides specific for 3 new BMMF2 isolates, as well as the recovery of a modified BMMF1 isolate H1MSB.1, from colon tissue. The intention of this communication is to draw the attention to the existence of a hitherto unrecognized, broad spectrum of potential pathogens.

## Materials and methods

**Samples:** The isolates were obtained from milk (10) and dairy products (8) bought from local supermarkets. To avoid a biased selection, we chose milk in different types of containers (glass, carton, etc.), and from many different providers. In addition, we tested crème fraiche, cream, curd, butter and yoghurt.

**DNA extraction:** DNA was extracted with phenol/chloroform as previously described [19]. Samples were diluted 1:1 with proteinase K buffer (0.2M Tris-HCl, pH7.5, 25 mM EDTA, 0.3M NaCl, 2%SDS) and digested overnight with proteinase K (final concentration 1 mg/ml, Sigma-Aldrich) at 37°C. Ethanol precipitation followed after the phenol/chloroform extraction. DNA was extracted this way from the initial 4 milk samples, whereas all other samples were subsequently subjected to an additional purification using a Nucleospin Gel and PCR Clean up Kit (Machery and Nagel) according to the manufacturer's protocol. DNA was finally eluted in 5 mM Tris-HCl, pH 8.0.

**Rolling circle amplification** was performed as described [17]. Template DNA (50 ng) was incubated in a total volume of 10 µl (1x phi29 DNA polymerase buffer) with 25 µM Exo-resistant Random primer (Thermo Fisher Scientific) at 95°C for 3 min followed by cooling on ice. This sample was diluted to 20µl by adding BSA (0,4 mg/ml), 0,75 mM dNTPs (Takara) each and 10U phi29 DNA Polymerase (New England Biolabs) in 1x phi29 DNA polymerase buffer. Incubation followed at 30°C for 18 h and 10 min at 65°C before stopping the reaction at 15°C.

**Polymerase chain reaction:** A 3 µl template from the RCA amplified DNA was used in the polymerase chain reaction (PCR) using TAKARA Taq enzyme and the accompanying solutions according to the manufacturer's protocol (LA TAKARA). The respective back-to-back primers were added at a final concentration of 0,2 mM each. Primers were designed based on the conserved regions of HCBI1, HCBI2 and HCBI7 for the BMMF2 group and for the BMMF1 group based on the conserved region of the replication gene.

PCR was performed by using touchdown protocols specific for the respective primer pairs:

MBB2 primers for the BMMF1 group: Mbb1 (MSBI1.176bbF296 forward) –5'-TGCAGAAATTGCCCCTCGACT-3' and Mbb2 (MSBI1.176bbR295 reverse) –5'-AACAATGGGGAAGAAGTCAAAGG-3'. The first round of 5 cycles amplification was performed at 30 s melting at 94°C, 1 min annealing at 64°C and 2 min elongation at 72°C, followed by a second round of 5 cycles using 62°C annealing and a third round of 30 cycles using 60°C for annealing. The final round of 10 min at 72°C followed.

Amplification for the BMMF2 group was performed as described for the BMMF1 group using the MBB2 primers, except that the touchdown temperatures were 58°C, 56°C and 54°C for the LSconA primer and 56°C, 54°C and 52°C for the LSconB primer and the elongation time was prolonged to 3 min at 72°C per cycle. The back-to-back primer pairs used for the BMMF2 group were: LSconA5p (forward) –5'-AAGGCAGATCAACACAGG-3' and LSconA3p (reverse) – 5'-AGCAGATTGCAAAGCCTG-3', LSconB5p (forward) – 5'-CAACACAGGGATAGAA-TAAC-3' and LSconB3p (reverse) – 5'- ATCTGCCT-TAGCAGATTGC-3'. PCR products were separated by gel electrophoresis, visible fragments excised and DNA extracted using the Nucleospin Gel and PCR Clean up Kit (Machery and Nagel) according to the manufacturer's protocol. DNA was finally eluted in 5 mM Tris-HCl, pH 8,0. Eluted DNA was ligated and cloned into pCR2.1 using the TA-Cloning Kit and transformed into *E.coli* – all according to the manufacturer's protocol (Invitrogen™ TA Cloning™ Kit, with pCR™2.1 Vector and One Shot™ TOP10F' Chemically Competent *E. coli* (Invitrogen)). Sequencing of all clones was performed by GATC. This company applies cycle sequencing technology on ABI3730XL sequencing machines. Quality Phred20 is achieved for reads up to 1100 nucleotides in length.

**Laser Dissection Microscopy (LDM):** Formalin-fixed-paraffin-embedded sections with a thickness of 8µm were mounted onto LDM-compatible membrane slides (MembraneSlide 1.0 PEN NF, Carl Zeiss) and stained with anti-Rep antibodies [15]. Antibody-stained tissue regions were specifically marked and dissected using a high intensity laser beam (PALM microbeam microscope, Carl Zeiss). The selected tissue pieces together with the underlying membrane were laser-catapulted into the caps of special opaque adhesive cap microfuge tubes (500 microliter, Carl Zeiss). The tissue pieces were subsequently dissolved in ATL lysis buffer (QIAamp DMA Mini Kit, Qiagen) and homogenized by pipetting before transferring into a fresh 1.5 ml microfuge tube. An equal volume of Chelex resin (Chelex® 100 Resin, mesh 200-400, Bio-Rad, 5% suspension in water) was added in addition to 10 µg/ml Proteinase K (Sigma), followed by incubation overnight in a

thermomixer at 56°C at 750 rpm. The reaction was stopped by vortexing for 10 s and incubation for 8 min at 99°C in a thermomixer. Chelex beads were removed by centrifugation and the DNA containing supernatant collected for further processing. RCA was performed as described above, followed by PCR amplification using NnXn or NoXo back-to-back primers as previously described [19]. Products were cloned and sequenced as described above.

*In silico analyses:* All sequences obtained were compared to nucleotide sequences available in Genbank by using BLASTN and the putative proteins to BLASTP (swissprot and swissprot\_splicevar). In addition, Domainsweep [22] was used to identify the domain architecture within a protein sequence by using different database searches including blocks-searcher, cddsearch, cathscan, pfamascan, printscan, prodomblast, prosite, smartscan, superfamily and tigrfamscan.

Predicted genes were obtained with an in-house ORFmap program, as well as with the gene identification program GENSCAN which was developed by Chris Burge from Stanford University [23]. It analyzes genomic DNA sequences from human, other vertebrates, invertebrates and plants. Other more sensitive prediction tools (Augustus [24] or Genemark [25]), were considered not suitable for our purposes, as they rely on calibration with species-specific training sets. Genscan was chosen because information is provided on the presence of polyA signals and our isolates [21] had been shown to replicate in human cells.

*Phylogenetic trees* were developed using maximum likelihood based on aligned (MUSCLE algorithm) sequences – both full-length nucleotide genomes, as well as the core region of Rep. Phylogenetic and molecular evolutionary analyses were conducted using MEGA version X [26].

Data availability: Genome sequencing files are available at the European nucleotide archive (<https://www.ebi.ac.uk/ena>) under Project PRJEB30101. Accession numbers LR215494-LR215600 were allocated for the respective sequences (re Table 1).

## Results

### ***In silico* characterization of additional BMMF1 (bovine meat and milk factor) isolates**

The BMMF1 group of isolates consists of 13 circular DNA genomes [14,19,27] (Suppl. Table 1). Mutual characteristics included the presence of an initiator replication protein RepB, an iteron-like tandem repeat region (3 × 22 nt plus a partial repeat of 17 /18nt, in one isolate 4 complete tandem repeats) and a conserved palindromic structure located upstream of the repeats which may serve as a putative origin of

replication. The RepB protein and the iterons were similar to those found in bacterial plasmids, whereas the putative origin of replication structure was similar to the nonanucleotide stem-loop origin of replication in single-stranded plant and animal viruses [3,14]. The genomes of all the isolates shared nucleotide identity to various plasmids of *Acinetobacter baumannii* and one isolate HCBI5.173 nucleotide identity to plasmids of *Psychrobacter spp.*

Here we report the characterization of additional 5 isolates in our extended analyses of milk samples (Suppl. Table 1). The BMMF1 isolates described in this study are C1MI.15M.1 (2040nt) (with subtype C1MI.15M.2, 2041nt, 99.9% nucleotide identity, one additional 57aa putative ORF), C1MI.9M.1 (1935nt) (with subtype C1MI.9M.2, 1934nt, 99.7% nucleotide identity, missing 51aa putative ORF), as well as C1MI.3M.1 (1767nt), were all obtained after DNA amplification using back-to-back primers located upstream of the Rep gene. C1MI.3M.1 shares 90% nucleotide identity to the uncultured bacterium plasmid HD4bpcirc (additional accession numbers in Suppl. Table 2) and 83% to our H1MSB.2 isolate by BLASTN analyses. C1MI.15M shares 99% nucleotide identity to uncultured prokaryote plasmid pRGRH0677 from rat gut metagenome, 80% to H1MSB.2 and 78% to pHD4bpcirc, whereas C1MI.9M shares 93% nucleotide identity to the same plasmid pRGRH0677 and 79% to both H1MSB.2 and pHD4bpcirc. These isolates share lower nucleotide identity to the majority of the previous group. Similar to other members of the BMMF1 group, they contain conserved direct repeats (iterons) (4 × 22 nucleotides, ATAAGACGACACTTACCTACCA) located at nt164-251 in C1MI.15M and at nt162-249 in C1MI.9M. Additional conserved repeats (4 × 9nt GGTTTTTAA) are located at nt87-121 in C1MI.9M. Similarly, iterons are located at nt164-250 (4 × 22 ATATCACACCGTTTACCCATCA) in C1MI.3M.

The putative Rep proteins of C1MI.15M and pRGRH0677 are identical (100%), whereas C1MI.9M shares 94% similarity to the rat gut plasmid (Figure 1). The putative Rep protein of C1MI.3M.1 is 95% similar to the Rep protein of pHD4bpcirc, 85% to H1MSB.2Rep and 81% to pRGRH0677Rep. *In silico* analyses of this putative Rep protein point to RepA, as well as to RepC function, differing here from other BMMF1 Rep proteins. In contrast, the putative Rep proteins of the previous group of BMMF1 isolates are only between 55 and 56% similar to these above-mentioned isolates.

The putative 225aa protein from the second large ORF in the C1HB.4 genome (BMMF1) shared 66% identity in 227aa to a mobilization protein (DOUH-C7\_ACIBA 381aa) identified in *Acinetobacter*. This protein could be involved in plasmid recombination [30]. It similarly shared 91% identity in 225aa to an

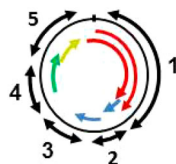
**Table 1.** Characteristics of BMMF2 isolates described in this study.

BMMF2 isolates	Genome size (nucleotides)	Putative proteins (positive strand) (size amino acid)					Acc. number	Similarities and other
		Region 1 (rep)	Region 2	Region 3	Region 4	Region 5		
<i>C2MI.15B.11</i>	2362	341, 307		55	138	61,97	LR215563	
<b><i>C2MI.15B.12</i></b>	2362	341, 307		55	138	61, 97	LR215564	
<i>C2MI.15B.3</i>	2362	341, 307		55	138	97	LR215555	
<i>C2MI.15B.4</i>	2362	341, 307		55	138	61, 97	LR215556	
<b><i>C2MI.10A.2</i></b>	2504	307, 341, 107	60		138	97	LR215593	
<i>C2MI.10A.1</i>	2505	244, 278, 74	60		138	61, 97	LR215592	3 peptides identical in mas spec
<i>C2MI.5A.2</i>	2363	341, 307		77	128	64, 97	LR215501	
<b><i>C2MI.5A.4</i></b>	2362	368, 307		77	128	64,97	LR215503	
<b><i>C2MI.9A.1</i></b>	2405	309, 353	62	56, 52	103	97	LR215538	
<i>C2MI.1A.1</i>	2293	370, 309		63, 57	102	97	LR215583	
<b><i>C2MI.1A.3</i></b>	2294	147, 208,187		63, 57	102	97	LR215585	
<i>C2MI.7A.4</i>	2313	184, 218, 158		50	51, 101	119	LR215516	
<b><i>C2MI.7A.5</i></b>	2315	261, 90, 58		50	51, 102	119	LR215517	
<b><i>C2MI.7A.6</i></b>	2315	344, 310		50	51, 102	119	LR215518	
<i>C2MI.15B.5</i>	2279	344, 310	62, 52	77	121	115	LR215557	
<i>C2MI.15B.6</i>	2279	344, 310	52, 55	67	121	115	LR215558	
<b><i>C2MI.15B.7</i></b>	2279	344, 310	59	52	119, 59	97	LR215559	
<b><i>C2MI.15B.8</i></b>	2277	344,310	52, 55	67	121	122, 51	LR215560	
<i>C2MI.15B.13</i>	2278	245,279,127	52, 55	77	121	115	LR215565	
<b><i>C2MI.15B.1</i></b>	2275	344, 310	52	66	121	137	LR215553	
<i>C2MI.15B.15</i>	2275	344, 310	52	66	121	137	LR215567	
<b><i>C2MI.5A.1</i></b>	2257	344,310	52	77	121	114	LR215500	Replaces HCBI2.170
<i>C2MI.5A.3</i>	2257	344, 310	52	77	121	114	LR215502	
<i>C2MI.5B.5</i>	2257	344, 310	52	77	121	114	LR215504	
<i>C2MI.5B.8</i>	2257	344, 310	52	77	121	114	LR215507	
<b><i>C2MI.15B.2</i></b>	2312	344, 310	61	52	119, 59	97	LR215554	
<b><i>C2MI.7B.16</i></b>	2486	338, 304		55, 65 63, 67	119, 59	99	LR215528	
<b><i>C2MI.9B.5</i></b>	2486	338, 304		55, 63, 65, 67	119, 59	99	LR215542	
<i>C2MI.16B.3</i>	2478	338, 304		70	59, 119	99	LR215573	
<b><i>C2MI.16B.4</i></b>	2478	338, 304		69, 59	59, 119	99	LR215574	
<i>C2MI.16B.5</i>	2478	338, 304		70	59, 119	99	LR215575	
<i>C2MI.16B.6</i>	2478	338, 304		70	59, 119	99	LR215576	
<i>C2MI.16B.12</i>	2479	338, 304		70	60, 108	99	LR215582	
<b><i>C2MI.5B.13</i></b>	2392	344, 310		81	118	96	LR215512	
<i>C2MI.7A.1</i>	2103	302, 268	59	53, 52	118	135	LR215513	
<i>C2MI.7A.2</i>	2102	344, 310	59	53, 52	118	135	LR215514	
<i>C2MI.7A.3</i>	2102	344, 310		52, 53	118	135	LR215515	
<i>C2MI.7A.7</i>	2102	344, 310		52, 53	118	135	LR215519	
<b><i>C2MI.7B.12</i></b>	2102	344, 310		52, 53	118	135	LR215524	
<i>C2MI.7B.14</i>	2102	344, 310		52, 53	118	135	LR215526	
<i>C2MI.16B.1</i>	2234	344, 310	58		118	96	LR215571	
<i>C2MI.16B.2</i>	2233	245, 279, 84	58		118	96	LR215572	
<i>C2MI.16B.8</i>	2234	344, 310	58		118	96	LR215578	
<b><i>C2MI.16B.10</i></b>	2234	344, 310	58		118	96	LR215580	
<b><i>C2MI.9B.14</i></b>	2593	386, 419	79	54	123	70, 96	LR215551	
<b><i>C2MI.15B.9</i></b>	2590	419, 386	77		115	97	LR215561	
<i>C2MI.15B.10</i>	2590	419, 386	79		115	97	LR215562	
<b><i>C2MI.9B.15</i></b>	2554	155, 122, 254, 88	77	52	109	115	LR215552	
<i>C2MI.7B.11</i>	2564	387, 354		75	115, 56	115	LR215523	
<i>C2MI.7B.13</i>	2565	414, 381		75	115, 56	115	LR215525	
<i>C2MI.7B.15</i>	2563	221, 188, 228		74	56, 115	115	LR215527	
<b><i>C2MI.7B.17</i></b>	2564	414, 381		75	56, 115	115	LR215529	
<i>C2MI.7B.18</i>	2565	121, 88, 334		75	56, 115	115	LR215530	
<b><i>C2MI.15B.14</i></b>	2776	419, 386	76		103	56, 116	LR215566	
<i>C2MI.15B.18</i>	2778	419, 386	76		103	56, 110	LR215570	RNA helicase
<b><i>C2MI.16B.9</i></b>	2777	369, 67	76		103	56, 116	LR215579	
<i>C2MI.16B.11</i>	2782	342, 106, 73	76		103	63	LR215581	
<i>C2MI.5B.9</i>	2736	419, 386	77		103	56, 116	LR215508	
<b><i>C2MI.5B.10</i></b>	2736	419,386	77		103	56, 116	LR215509	

(Continued)

Table 1. Continued.

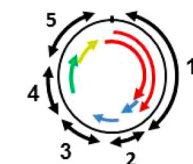
Putative proteins (positive strand) (size amino acid)



BMMF2 isolates	Genome size (nucleotides)	Region 1 (rep)	Region 2	Region 3	Region 4	Region 5	Acc. number	Similarities and other
<i>C2MI.5B.11</i>	2735	419, 386	77		103	56, 116	LR215510	
<b><i>C2MI.8B.7</i></b>	2736	419, 386	77		103	56, 116	LR215537	
<b><i>C2MI.9B.10</i></b>	2832	229, 262, 142	71		143	75, 116	LR215547	5 specific peptides in mass spec
<b><i>C2MI.16B.7</i></b>	2566	330, 363	66	74	146	97	LR215577	RNA-pol
<i>C2MI.5B.12</i>	2567	218, 251, 181	74		116	97	LR215511	
<i>C2MI.15B.17</i>	2952	420, 386	70	114	148	56, 116	LR215569	Deaminase(116)
<i>C2MI.15B.16</i>	2824	420, 386	71, 51		102	116, 56	LR215568	5 spec. peptides in mass spec Deaminase(116) Rep -
<b><i>C2MI.8B.4</i></b>	2850	424, 390	71	117	51, 102	116, 57	LR215534	RNA-pol
<b><i>C2MI.9B.11</i></b>	2537	420, 386	71		103	98	LR215548	S4 binding
<i>C2MI.8A.1</i>	3090	445, 401	105		103	56, 97	LR215531	RNA-pol
<i>C2MI.8A.2</i>	3090	445, 401, 94	105	74, 60	103	56, 97	LR215532	
<i>C2MI.8A.3</i>	3090	445, 401, 94	105	74, 60	103	56, 97	LR215533	
<i>C2MI.8B.5</i>	3090	44, 401, 94	105	74, 60	103	56, 97	LR215535	
<b><i>C2MI.8B.6</i></b>	3090	445, 401, 94	105	74aa, 60	103	56, 97	LR215536	
<i>C2MI.4A.1</i>	2661	397,428,92	87	119	128	65, 104	LR215589	
<i>C2MI.4A.2</i>	2661	397, 428, 82	87	119	128	65, 104	LR215590	
<b><i>C2MI.4B.3</i></b>	2661	397, 431, 82	87	119	128	65, 104	LR215591	
<b><i>C2MI.7A.8</i></b>	2460	401, 435, 94	74	51	114	96	LR215520	
<i>C2MI.7A.9</i>	2460	401, 435, 94	74	51	114	96	LR215521	
<i>C2MI.7A.10</i>	2460	401, 435, 94	74	51	114	96	LR215522	
<i>C2MI.5B.6</i>	2376	402, 441, 61	74	55	136	97	LR215505	
<i>C2MI.5B.7</i>	2375	140, 179, 246	74	55	136	97	LR215506	
<i>C2MI.13B.1</i>	2301	441, 401, 61	74		61, 112	112	LR215594	
<b><i>C2MI.13B.2</i></b>	2300	436, 397, 61	63	55	61, 112	112	LR215595	
<i>C2MI.13B.3</i>	2301	441, 402, 61	91	55	61, 112	112	LR215596	
<b><i>C2MI.9B.6</i></b>	2301	441, 402, 61	74	55	61, 112	112	LR215543	
<b><i>C2MI.9B.4</i></b>	2278	188, 227, 243	86		118	96	LR215541	
<i>C2MI.9B.7</i>	2279	438, 399	86		118	96	LR215544	
<i>C2MI.9B.8</i>	2279	438, 399	86		118a	96	LR215545	
<i>C2MI.9B.9</i>	2279	438, 399	98		118	96	LR215546	
<b><i>C2MI.9A.3</i></b>	2365	434, 400	60		152	99	LR215540	RNA-helicase
<i>C2MI.9B.12</i>	2367	431,400,89	60		144	99	LR215549	
<i>C2MI.9B.13</i>	2366	431, 400, 89	60		144	99	LR215550	
<b><i>C2MI.9A.2</i></b>	2366	431, 400, 89	60		144	99	LR215539	
<i>C2MI.3A.1</i>	2356	436, 402			183	97	LR215587	
<b><i>C2MI.3A.2</i></b>	2356	436, 402			183	97	LR215588	
<b><i>C2MI.1A.2</i></b>	2296	401, 435, 89	107		61, 112	112	LR215584	
<i>C2MI.1A.4</i>	2296	401, 435	107		61, 112	112	LR215586	

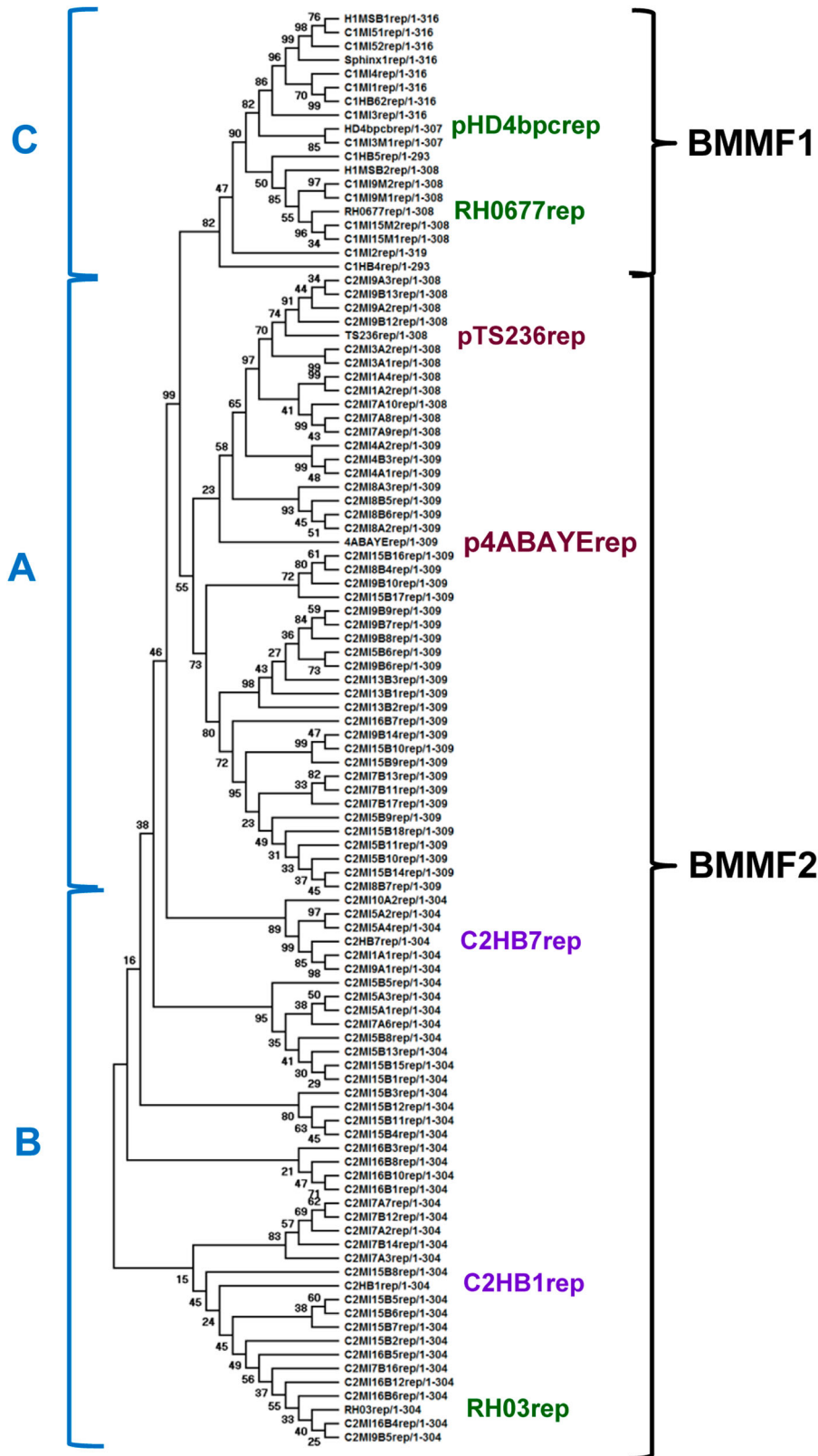
Notes: Isolates spaced together – >90% identity; isolates of neighbouring groups in italics – >80% identity; Bold – represented on the condensed phylogenetic tree (Figure 2). RNA-pol – Rep = RNA-polymerase II Rpb4 core; S4 binding – Rep = RNA binding S4 domain; RNA-helicase – Rep = ATP-dependent RNA-helicase; C2MI.5A.1 – cattle group2 milk isolate 5A.1. Samples 5 and 13 – dairy products, whereas all other samples are from dairy milk.

Regions of the genome:



BMMF2 "small" isolates	Genome size (nucleotides)	Acc. number	Other
<i>C2MI.9As.2</i>	697	LR215600	
<i>C2MI.10As.1</i>	697	LR215597	
<i>C2MI.5As.1</i>	697	LR215598	Identical in overlap with C2MI5B.10
<i>C2MI.7As.1</i>	697	LR215599	almost identical in overlap with C2MI.7B.17

Notes: C2MIs.5A.1 – cattle group2 milk small isolate 5A.1. Isolates grouped – >90% identity.



**Figure 1.** Molecular phylogenetic analysis of the overlapping region of all Rep proteins of both BMMF1 and BMMF2. The evolutionary history was inferred by using the Maximum Likelihood method and JTT matrix-based model [28]. The bootstrap consensus tree inferred from 500 replicates [29] is taken to represent the evolutionary history of the taxa analysed [29]. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [29]. Initial tree(s) for the heuristic search were obtained by applying the Neighbour-Joining method to a matrix of pairwise distances estimated using a JTT model. This analysis involved 104 amino acid sequences. There were a total of 319 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [26].

uncharacterized protein (N9GYE3\_AIHA uncharacterized, 377aa).

The single-stranded *Geminiviridae* (*Begomovirus* and *Mastrevirus* genera) and *Nanoviridae* of plants are associated with small circular ssDNA components which have a nonanucleotide stem-loop origin of replication and encode a Rep protein. They range from 970nt (nanoviruses) to ca 1400 nucleotides (geminiviruses) in size and are dependent on a helper virus for their spread. These alphasatellites have recently been classified into a new family of viruses *Alphasatellitidae* [31]. The functions of these alphasatellites are not clear, although it has been suggested that they contribute to symptoms by influencing transcription rates. The genomes of TT viruses (family *Anelloviridae*) undergo rearrangement during propagation to lead to the formation of small replication-competent circular DNA genomes (micro-TTV, ca 700-900nt) [6]. These micro-TTVs encode in part new putative proteins with similarity to cellular proteins suggesting a role in molecular mimicry [6]. The identification of the novel BMMF1 isolates reported in this study also led to the identification of a small circular DNA molecule C1MI.3Ms of 461nt in size. These small circular genomes resulted after PCR amplification using back-to-back primers as described for the full-length genomes. This genome shares 89% nucleotide identity with the plasmid pRGRH0677 isolated from rat gut and 78% and 77% nucleotide identity to H1MSB2 and pHD4bpcirc, respectively. The iteron repeats ( $4 \times 22$ nt, ATAAGAC-GAGACTTACCTACCA) present are identical to that in the genomes of C1MI.15M and C1MI.9M. *In silico* analyses revealed a putative Rep protein of 75 amino acids sharing >95% amino acid similarity to putative Rep proteins of C2MI.15M and C1MI.9M. BLASTP analyses indicate a 97% similarity to the first 50 amino acids of a protein in the Tunicate *Oikopleura dioica* [32] and lower similarity to uncharacterized RepB proteins.

### **In silico characterization of BMMF2 genomes and putative proteins**

The circular DNA genomes (HCBI1, HCBI2 and HCBI7, Suppl. Table 1) were isolated from serum of healthy cattle [17]. These isolates were grouped as BMMF2 based on their nucleotide identity to the previously described Sphinx2.36 [14,33]. In this study, we used two sets of back-to-back primers to generate the full-length circular genomes of 97 isolates (Table 1 including accession numbers) ranging from 2102 to 3090 nucleotides. BLASTN comparisons to available DNA databanks, revealed similarity mainly to bacterial plasmids (*A. baumannii* str). AYE plasmid p4ABAYE [34], *A. baumannii* strain A85 plasmid pA85-1 [35], and *A. baumannii* strain DS002 plasmid pTS236

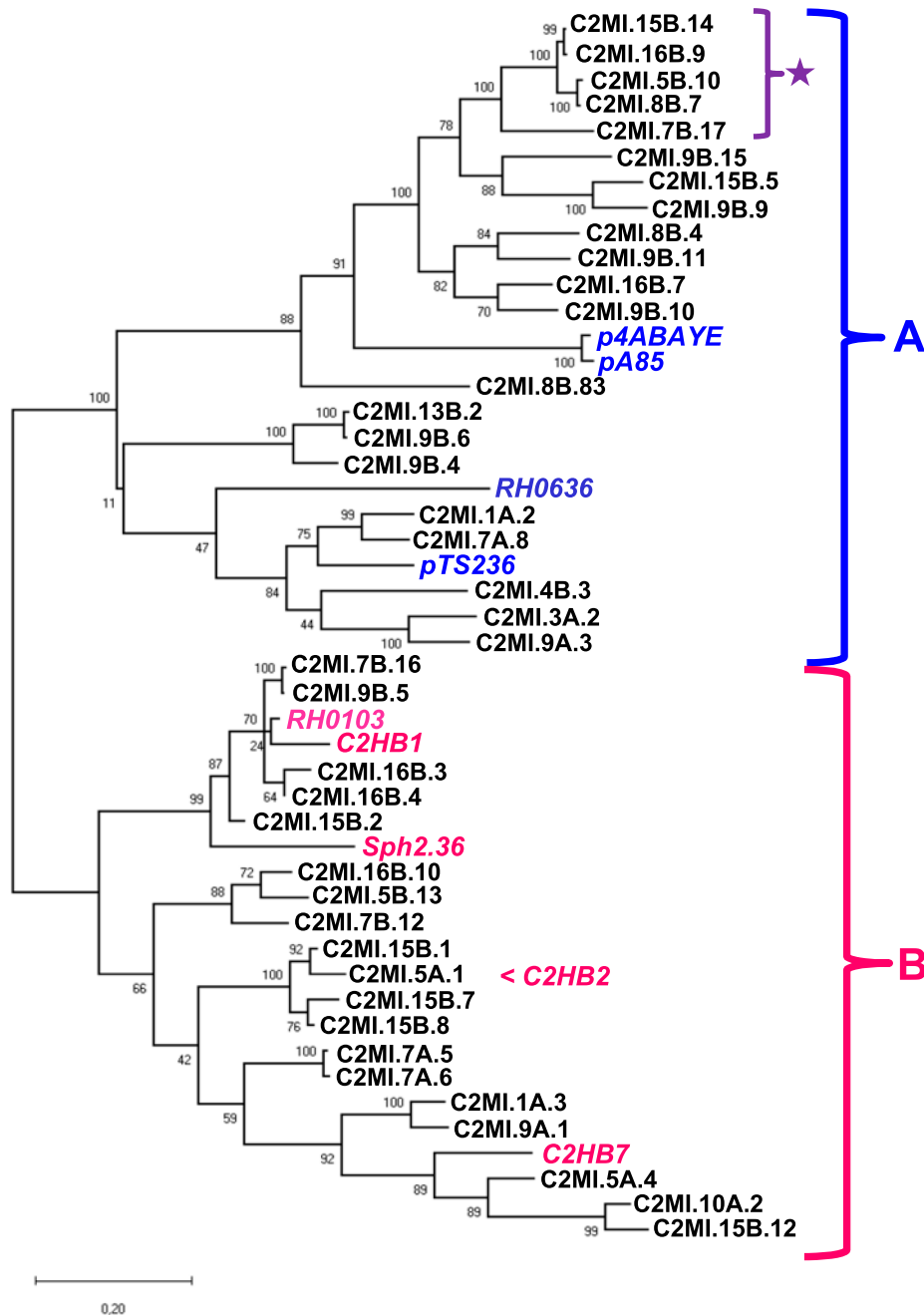
[36], but also to ssDNA isolates from Rat gut (pRGRH0103 and pRGRH0636) [37] (Suppl. Table 2). We performed phylogenetic analyses in order to determine relationships between isolates and to facilitate our selection of isolates with which to continue in-depth biological studies.

We included only one representative of a group of isolates sharing 90–99% nucleotide identity in the phylogenetic tree shown in Figure 2, as well as isolates originating from different dairy samples. Two main clades were seen – one clade sharing 50% to 57% nucleotide identity to the *A. baumannii* plasmid p4ABAYE (Figure 2 clade A) and between 40% and 50% to the second clade (Figure 2 clade B). The nucleotide identity between pRGRH0636 and pRGRH0103, two circular ssDNA genomes identified by metagenomic analyses from Rat gut, is 39%. The latter shares between 80 and 92% nucleotide identity to C2HB1, C2MI.15B.2, C2MI.16B.3, C2MI.7.16 and C2MI.9B.5.

HCBI2.170 was isolated from bovine serum by density gradient fractionation followed by RCA amplification of restricted DNA fragments. Despite using a series of back-to-back primers on the known sequence [17], we failed to rescue a larger fragment. An isolate obtained in the present study resulted in the full-length genome, i.e. HCBI2.170 and C2MI.5A.1 (2257nt) shared identical nucleotides in the overlapping part of the genomes. C2MI.5A.1 now replaces HCBI2.170.

*In silico* analyses of putative open reading frames (ORF, >50aa) were performed on each genome by using the following alternative start codons: ATG, CTG, ATA, ATT, ACG, GTG, TTG and ATC [39]. This resulted in the identification of mainly 2 additional major ORFs on the positive DNA strand (Table 1), as well as putative proteins ranging from 101 to 138 amino acids and 96–148 amino acids in size, respectively. Additional smaller ORFs were present. Multiple ORFs (average number 2–10, up to 125 amino acids) were identified on the reverse strand. BLASTP analyses of the latter putative proteins revealed no or very weak similarity to known proteins. Further experimental investigation is needed to determine whether splicing events may result in combined, functional proteins. Transcription of the negative DNA strand was previously noted for the BMMF1 group of isolates [21].

An interesting result was the isolation of smaller circular DNA molecules 697nt in size and sharing 93-99% nucleotide identity. These were isolated from 3 different milk samples and one yoghurt sample. *In silico* analyses of C2MIs.5A.1, C2MIs.7A.1, C2MIs.9A.2, and C2MIs.10A.1 revealed 2 overlapping putative ORFs on the positive strand which could encode for 224 and 248 amino acid replication proteins, whereas the negative strand could encode for putative proteins of 53, 58 and 81 amino acids in C2MIs.5A.1 and



**Figure 2.** Phylogenetic tree of the nucleotide sequences of representatives of each BMMF2 isolate in comparison to the plasmids of *A. baumannii* p4ABAYE, pA85-1 and pTS236, genomic isolates from rat gut pRGRH0103 (RH103) and pRGRH0636 (RH0636) in clade A, as well as Sphinx2.36 and our previous BMMF1 isolates C2MI1 and C2HB7 in clade B. The recently isolated C2MI.5A.1 represents the full-length genome of C2HB2. The star indicate BMMF2 isolates of which the nucleotide sequence is almost identical in the overlapping nucleotide sequence of the respective “small” BMMF2 isolate. The evolutionary history was inferred by using the Maximum Likelihood method and Tamura-Nei model [38]. The bootstrap consensus tree inferred from 500 replicates [29] is taken to represent the evolutionary history of the taxa analysed [29]. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [29]. Initial tree(s) for the heuristic search were obtained by applying the Neighbour-Joining method to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach. This analysis involved 48 nucleotide sequences. There were a total of 3170 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [26].

C2MIs.7A.1, and 64 amino acids in C2MIs.9A.2 and C2MIs.10A.1. The nucleotide sequences of C2MIs.5A.1 and C2MIs.7A.1 were almost identical to the overlapping regions of the full-length genomes C2MI5B.10 and C2MI.7B.17, respectively. The rep proteins differed from those of the full-length genomes

counterparts, mainly in N-terminal 33aa of 248aa rep proteins and 7aa of both 224aa and 248aa proteins in the C-terminal. In addition, single amino acid differences were noted between the 2 groups of small circular genomes.

As the aim of our study is to identify, isolate and associate unknown infectious agents to the aetiology



of colon cancer, we performed mass spec analyses of protein samples from colon cancer and normal tissue, as well as laser microdissection of these samples. Mass spec analyses of protein samples from colon cancer and normal colon tissue (Bund et al., in preparation), resulted in identification of multiple specific peptide motifs originating from 3 BMMF2 isolates. Specific peptide motifs could be allocated, 5 each to C2MI.15B.16 and C2MI.9B.10 (Table 2). Interestingly, several of these amino acid motifs indicating a possible functional involvement, were located on the negative DNA strand. Isolates C2MI.10A.1 and C2MI.10A.2 share 99.7% nucleotide identity, but 3 specific peptide motifs were identified in mass spec to C2MI.10A.1, whereas C2MI.10A.2 harboured only 2 of these 3 motifs. BMMF isolates sharing >90% identity in their nucleotide sequences, were initially regarded as variants of a single genome. Based on the possible biological significance of such small differences, we decided to include 97 BMMF2 isolates for more detailed analyses. In addition, we recovered a member of the BMMF1 group from colon tissue by laser microdissection. Cloning and sequencing confirmed a H1MSB.1 genome sequence modified from the original isolate MSB1.176 (now H1MSB.1) from an autopsy brain sample of a multiple sclerosis patient (Figure 4(a)). The nucleotide modifications included mainly C/T or T/C or less often, A/G or G/A transitions. These led to alterations in the putative gene organization with an additional 62aa putative protein in LD10.154 (Figure 4). Domainsweep analyses indicated a weak similarity to a domain (diarrhea inducing) of a non-structural glycoprotein of Rotaviruses.

### Characteristics specific for putative BMMF replication proteins

Bacterial plasmids are classified based on nucleotide homology of their replicase genes [40]. We identified 2 overlapping putative Rep genes in the BMMF2 isolates differing between 30aa to 50aa in length. Putative Rep proteins of all BMMF2, as well as BMMF1 isolates were compared. A core conserved region (329aa) was used in the phylogenetic analyses. BMMF2 rep proteins were grouped into 2 clades (A and B), whereas BMMF1 rep proteins formed a third clade (C) (Figure 1). Clades A and B shared 60–65% similarity. Rep proteins in clade A shared 69–90% similarity within the clade and 70–82% to the p4ABAYE rep. Similarly, clade B

rep proteins shared 80–90% within the clade and 59–62% to p4ABAYE rep. The rep proteins of BMMF1 (clade C) are more diverse and shared almost no similarity (10–12%) to clades A and B. The rep proteins of H1MSB.1, C1MI5.1, and C1MI.5.2 are identical and share 97% identity to that of Sphinx1.76 [33]. These all share 75–88% similarity to C1MI.1, C1MI.2, C1MI.3, C1MI.4 and C1HB.6.2 reps. This group of rep proteins share only 45% identity to C1HB.4rep, 37–39% to C1HB.5rep and 58% to H1MSB.2rep. The Rep proteins of the newly identified BMMF1 isolates described in this study, share 50–62% to the other BMMF1 isolates.

*In silico* structure prediction [41] of putative BMMF Reps revealed a partial structural homology of Rep with known replication initiator proteins of the Rep super-families 1, 2 and 3. Catalytically active motifs previously identified in the Reps of Geminivirus and other known ssDNA viruses were identified in the BMMF2 Reps [3,42,43]. Motif I (consensus aa FLTYP) is generally less conserved and also partially represented in BMMF2 Reps (T(F/L)(T/S)V(K/R)N). Motif II (consensus HxH) is important for metal ion binding and is – with one exception C2MI.5B.10rep – present as H(V/L/I)H in all BMMF2 reps. Motif III containing the active site tyrosine is present as consensus YALK in all BMMF2 reps. Motif positions in the C2HBI7Rep, taken as representative, are as follows: motif I at aa 110–115, motif II at 172–174 and motif III at aa 227–230. The lengths of amino acid spacers between the motifs in C2HBI7Rep are roughly 60aa between motifs I and II and 50aa between motif II and III. Putative Walker A, B and C (SF3 helicase) motifs are not conserved amongst the BMMF2 Reps. Structure predictions based on amino acid sequences of BMMF1 Reps show similarity to RepB within the Rep superfamily 3 which also harbours RepE, RepC and RepA. Crystallography based on the N-terminal domain of the H1MSB.1 Rep (aa1-136) resulted in a structure with a resolution of 1.53 [44]. The overall structure showed a remarkable similarity with the structure of the RepA protein of *Pseudomonas syringae* despite sharing only 28% amino acid sequence identity.

The putative rep proteins of 14 BMMF2 isolates did not result only from one continuous ORF, but were compiled of more than one truncated ORF. BLASTP analyses of the smaller putative proteins indicated similarity to the Rep gene. A similar phenomenon was observed in a previous study on TT viruses

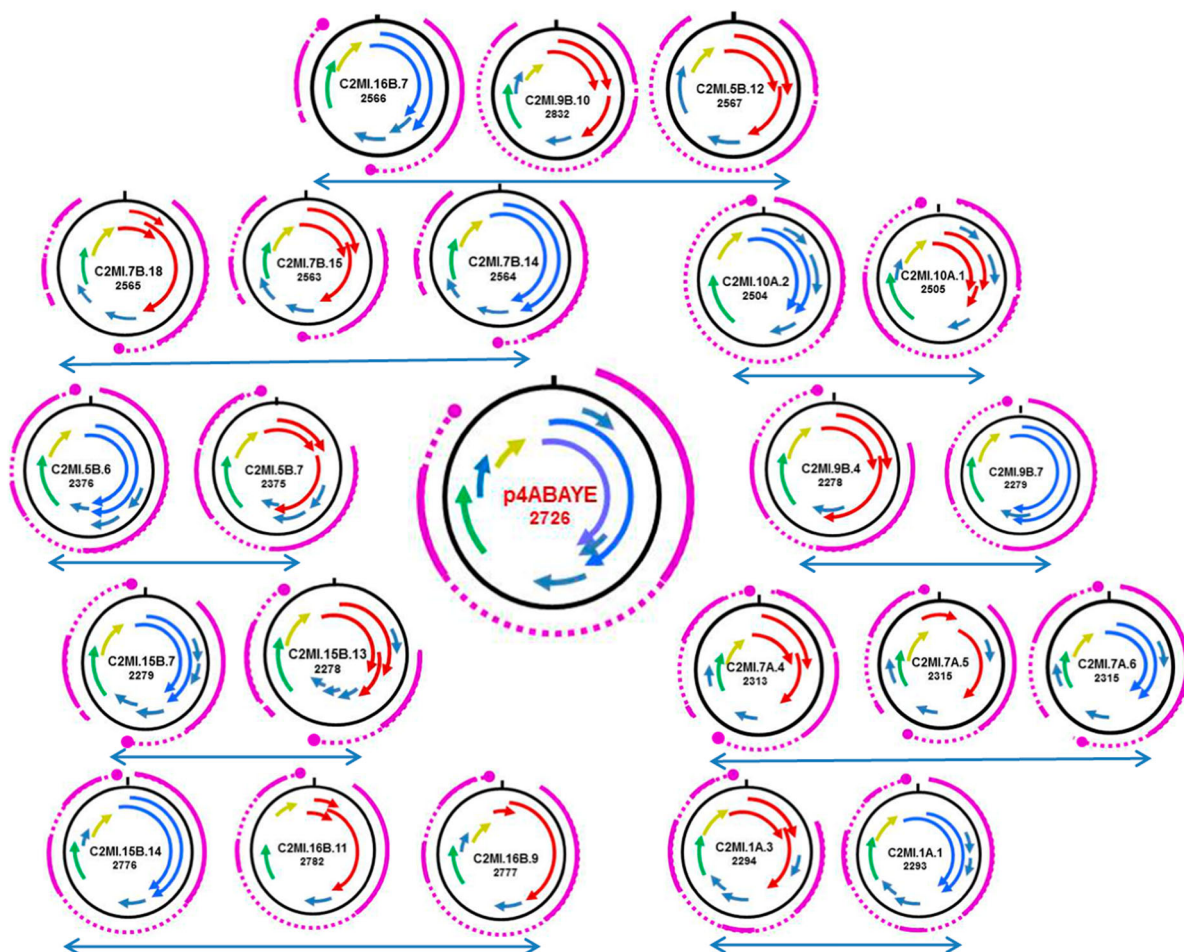
**Table 2.** Peptides in 3 BMMF2 isolates identified by mass spec in colon tissue.

Peptide motifs	C2MI.10A.1 (positive strand)	C2MI.9B.10 (* strand)	C2MI.15B.16 (* strand)
Motif 1	IHLSLTFFLK	TVLTVNGLR (positive)	LITNSLR (reverse)
Motif 2	ILELSNDR	ISTVIFRFR (reverse)	LTPFGVMCK (reverse)
Motif 3	LEGLLLNTFTQQGGQNK	VNAVTRAFETK (reverse)	MPFLCLLVIR (reverse)
Motif 4		YIFFRMHASAFK (reverse)	PEARAPHGMLTR (reverse)
Motif 5		NLLISTVIFRFRVF (reverse)	TATYEVSSLFFLLLDITNSGDK (positive)

(*Anelloviridae*). Here the premature stop codons had been verified and resulted from confirmed single nucleotide differences between independent isolates of the same TTV type [16]. In order to ascertain whether potential introns were present in BMMF2 genomes (with and without continuous Rep genes), we performed GENSCAN analyses [23]. This prediction tool was developed to identify genes in prokaryotes, eukaryotes and plants and was most suitable for our present purpose. We aligned the resulting genes to the respective gene maps obtained on the full-length nucleotide genomes. The GENSCAN result of *A. baumannii* plasmid p4ABAYE is included (Figure 3, Suppl. Table 3). The closest related isolate with an intact Rep ORF is presented with each isolate with truncated Rep ORFs. The presence in almost every isolate of a polyadenylation (poly(A)) signal (restricted to 6nt) is very notable in the GENSCAN products, including in the plasmid p4ABAYE. The polyA signal (AATAAA) of the latter was identical to the majority of putative poly(A) signals in the BMMF2 group. Variations (ATAAAA, AAATAA) were detected in 9 other

isolates. Notable differences were AATCAA in C2MI.5A.2 and CTAAA in C2MI.16B.2.

No significant hits were obtained in the majority of BLASTP analyses performed on most of the additional putative proteins identified in various BMMF2 isolates. DomainSweep analyses provide a compilation of protein characteristics resulting from identification of protein structures by employing different protein database search methods and scanning a number of protein domain/family databases [22,45]. The minority of additional putative proteins identified in individual isolates shared very weak similarities (i.e. not significant similarities, but nevertheless included in the first 10 listed for the respective search method used in Domainsweep) to proteins with known functions. In view of the aim of the present study, we regard these data as indicative of possible functions, but needs to be verified by further investigation. Interestingly a few putative Rep proteins were assigned very weak similarities to RNA helicase, to the RNA-binding S4 domain and to RNA polymerase II Rpb4 core domain, as well as a putative protein from region 5

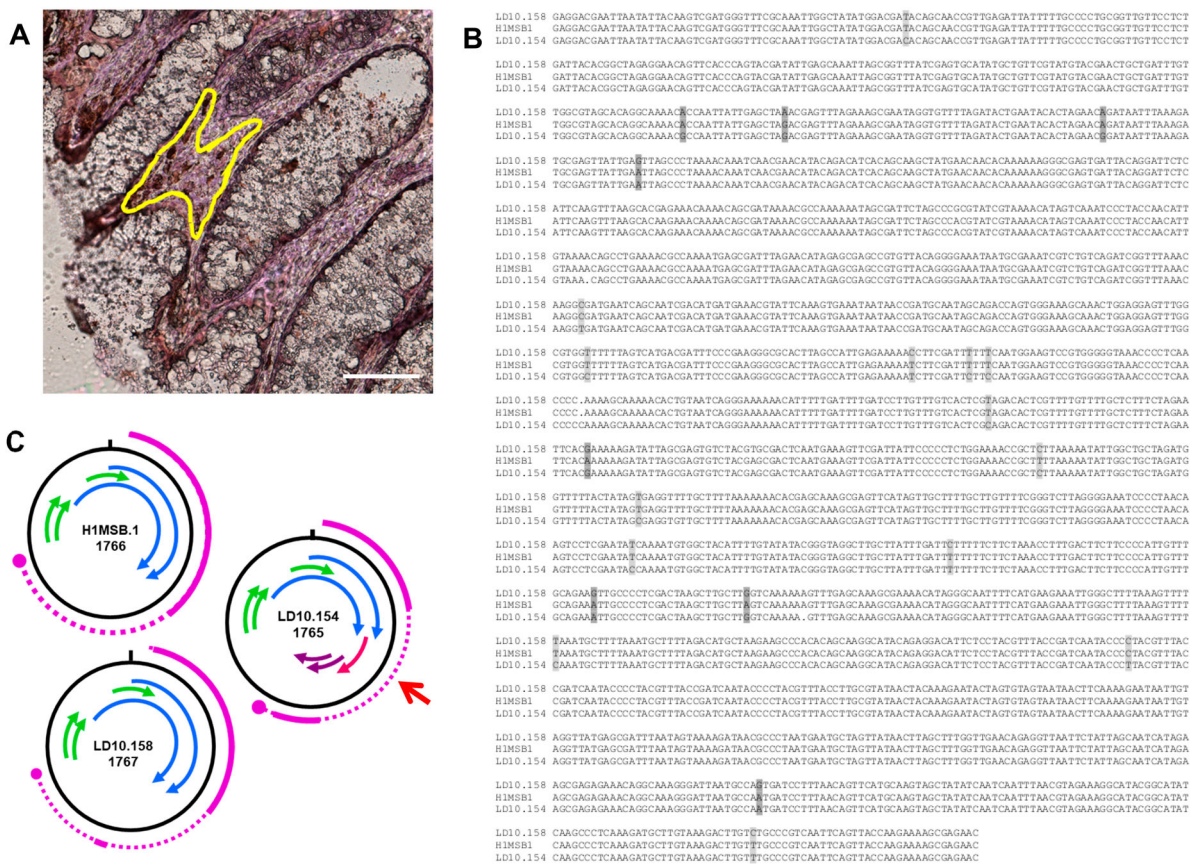


**Figure 3.** Schematic presentation of the ORF maps of BMMF2 isolates displaying truncated ORFs which encode for putative Rep proteins (arrows in red) in comparison to closely related isolates with intact Rep ORFs (blue arrows). The ORF map of p4ABAYE is shown in the middle. Arrows in same colour (green or yellow) represent ORFs which encode putative proteins showing high similarity between isolates. The GENSCAN analyses indicating the resulting splicing transcriptional events are indicated in purple. The large purple dot represents the poly A signal. In some cases no poly A signal is present for the respective transcript.

(116aa) to a specific subgroup of CDD/CDA-like deaminases [46] (Table 1). The H1MSB.1 isolate recovered from colon tissue displayed mainly C/T modifications. These transitions led to alterations in the genome structure (Figure 4) resulting in putative genes which could be involved in pathogenesis. Small ssDNA viruses experience higher mutation rates when in a changing environment such as chronic infections or host immune defence [47]. These modifications could also lead to more virulent viral strains inducing severe disease [48]. RNA editing of hepatitis D virus occurs via C/T or on the negative strand as a functional A-to-G substitution [49]. The unexpected similarities to RNA helicase and deaminase, the perfect matching peptides (identified in mass spec) which are allocated to the negative strand of the respective genome, as well as the two overlapping putative Rep genes, posed the question to the origin of these BMMF2 isolates which were initially all obtained by DNA amplification. These characteristics are very conspicuous and stimulated speculation as to whether parallels can be drawn to the structure and replication of Hepatitis delta virus [50], but this requires further investigation.

In addition, the following two results deserve discussion: C2HB7 (BMMF2) codes for a putative 228aa protein sharing a weak similarity to the bacterial ribon-helix-helix CopG protein which acts as a transcriptional repressor on the Rep gene [51]. The putative 291aa protein of C1MI.1 (BMMF1) also shares weak similarity to the CopG protein. Surprisingly this 228aa putative protein of C2HB7 displays similarity to viral proteins as well, i.e. it shares weak homology to the capsid protein of narcissus mosaic RNA virus reflecting recombination between DNA and RNA viruses during evolution [52]. Bacilladnaviruses (with a partial double-stranded DNA genome) probably acquired their capsid gene from ssRNA nodaviruses by horizontal transfer [11].

Possible recombination events within the BMMF2 group were also noted. The 196aa putative protein of C2HB7 shares 90% identity with the 183aa putative protein of C2MI.3A.2, whereas their rep proteins share only 45–55% identity (conserved core 64%). The nucleotide identity between these full-length genomes is 55%. Possible recombination events were noted in aligning the full-length nucleotide genomes of all BMMF2 isolates. This observation needs to be



**Figure 4.** Modified DNA genomes of BMMF1 H1MSB.1 were recovered by laser microdissection from the lamina propria cells of the colon from colon cancer patients. A – area dissected from the peritumoral colon tissue. Rep-stained areas are marked (yellow). Bar 100  $\mu$ m. B – Clustal analysis of the modified genomes LD10.154 and LD10.158 with the original H1MSB.1 genome indicated a number of C/T transitions (light grey shaded) and fewer A/G modifications (dark grey shaded). C – Schematic presentation of the putative gene organization of these DNA genomes. The modifications in the LD10.154 genome led to a modified gene structure (red arrow) resulting in a 62 amino acid putative protein which could be involved in pathogenesis.

confirmed by additional phylogenetic analyses which were not covered in this study.

## Discussion

Infectious agents are involved in the pathogenesis of a number of malignant diseases [8]. Based on available epidemiological data, we intensified our search for yet unknown infectious agents in additional cancer types [8,14]. We described the isolation of BMMF factors from serum obtained from healthy cattle, as well as brain and serum from patients suffering multiple sclerosis [14]. Closer characterization of these isolates not only revealed similarities to plasmids from *A. baumannii*, but included features of circular ssDNA viruses in plants and animals. The recovery of a BMMF1 isolate H1MSB1 from the peritumoral lamina propria cells of colon cancer patients with CD68 macrophage invasion [15] also demonstrates the association of these agents to infections of the colon. These results potentially place such infectious agents into a new class of pathogens [15].

*Acinetobacter baumannii* is broadly distributed in the environment and commonly found in water, soil and animals [53]. Classification of the *Acinetobacter baumannii* plasmids is based on typing of the replicons and replicate genes [40]. These plasmids typically have 4 conserved direct repeats (iterons) and a Rep gene belonging to the Rep\_3 superfamily or no iterons when comprising a Rep gene belonging to the Rep1/2 superfamily. Isolates of the BMMF1 group were related to a broader range of bacterial plasmids. They harbour iterons and have genes encoding initiator of replication proteins (RepB or Rep superfamily 3). RepB does not have an intrinsic helicase activity, but retains this activity through formation of a hexameric ring structure [54]. The latter is also a necessity for the helicase activity in eukaryotic viruses. One putative protein of isolate C1HB4 has weak similarity to a mobilization protein. In contrast, the closest related sequence to the BMMF2 isolates is plasmid p4ABAYE, the smallest of 4 plasmids of *A. baumannii* strain ABAYE. This plasmid has been described as cryptic and is classified together with plasmids pTS236 and pA85-1 in the GR14 clade [35,40,55]. Plasmid p4ABAYE is unique and seems to have followed a different evolutionary pathway than other bacterial plasmids. Distinct differences are that p4ABAYE has no iterons, has lost its mobility, with only the replication (Rep\_1 gene) and transfer functions remaining, and did not share any link with the 6 chromosomes of *A. baumannii*, i.e. does not participate in exchange of genes [53]. The BMMF2 isolates share characteristics with plasmid p4ABAYE, i.e. size of genome (2726nt), nucleotide similarity, do not have iterons and have a gene encoding a putative Rep protein belonging to the Rep superfamily 1/2.

BMMF1 and BMMF2 also display viral features. The Rep genes of recently identified CRESS (circular Rep-encoding ssDNA)-DNA viruses are highly conserved and form a hexameric ring structure [3]. These ssDNA viruses (which include the plantviruses *Geminiviridae* and *Nanoviridae*) are all characterized as having a stem-loop structure as the origin of replication. Such a putative stem-loop structure is present in BMMF1 isolates (in addition to the bacterial iterons). The second large ORF of several CRESS-DNA viruses shows some similarity to capsid proteins [3] – a similar observation was made for single BMMF2 isolates. The *in silico* observation of “truncated” ORFs which putatively code for a Rep protein may represent a similar mechanism of transcription as described in plant RNA viruses. Here the non-canonical translation includes leaky scanning, stop-codon read-through and ribosomal frameshifting [56]. The presence of polyA signal in almost all putative transcripts including that of p4ABAYE, also indicates transcription in eukaryotes.

The most striking feature of both BMMF1 and BMMF2 groups is the existence of larger full-length genomes in conjunction with smaller circular satellite-like genomes encoding for only Rep-associated genes. The bacterial iterons are also present in smaller genomes in BMMF1. Members of the recently established family *Alphasatellitidae* all have a stem-loop structure as origin of replication. Although their specific function remains unclear, their co-infection is associated with exacerbation or reduction of disease symptoms in plants [31]. Similarly, the non-coding deltasatellites (682nt) associated with genus Begomovirus of the *Geminiviridae* seem to influence the severity of disease or virus accumulation in plants indirectly [57]. Co-infection of a “large” (C1MI.1, 2523nt) and “small” (H1MB1, 1776nt) BMMF1 isolates resulted in increased transcription of the smaller genome [21]. The identification of even smaller circular genomes in both BMMF1 and BMMF2 – C1MI.3Ms (461nt with a 75aa putative Rep protein), C2MIs.5A.1, C2MIs.7A.1, C2MIs.9A.1, and C2MIs.10A.1 (697nt encoding for a putative full-length Rep protein), is intriguing. Future *in vivo* investigations are urgently needed exploring their functions, as this may indicate whether these molecules act like viral satellites influencing or even causing disease symptoms.

This group of infectious agents has been implicated in the pathogenesis of colon and breast cancer [15]. Although the evolutionary origin is unknown, it seems to fall between a unique bacterial plasmid and circular ssDNA viruses. We suspect that it originated from a bacterial plasmid which has evolved to adapt to infect and replicate in mammalian cells. The acquisition of smaller satellite-like genomes may have evolved subsequently in order to strengthen and/or

extend survival or to convey pathogenic symptoms. Indirect hints as to the origin and structure of our isolates, as well as its replicative behaviour, result from parallels drawn from Hepatitis deltavirus genome structure and replication.

Thus far, the present available analyses of human whole genome sequencing precluded the identification of originally bacterially derived plasmids as human pathogens by dismissing and discarding sequences suspected to originate from bacterial contaminants. A different perspective for investigating an unexplored spectrum of potential human pathogens seems to be highly desirable and recommended.

## Acknowledgements

We thank Imke Grewe for excellent technical assistance.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding


This study was funded by an unrestricted grant from Oryx Alpha, München.

## ORCID

Ethel-Michele de Villiers  <http://orcid.org/0000-0003-4357-0567>

Deblina Chakraborty  <http://orcid.org/0000-0002-3755-6809>

Timo Bund  <http://orcid.org/0000-0002-3304-4494>

Harald zur Hausen  <http://orcid.org/0000-0003-1103-0347>

## References

- [1] Kowarsky M, Camunas-Soler J, Kertesz M, et al. Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA. *Proc Natl Acad Sci U S A*. 2017;114:9623–9628.
- [2] Fernandes MA, Verstraete SG, Phan TG, et al. Enteric virome and bacterial microbiota in children with ulcerative colitis and Crohn's disease. *J Pediatr Gastroenterol Nutr*. 2019;68:30–36.
- [3] Zhao L, Rosario K, Breitbart M, et al. Eukaryotic circular rep-encoding single-stranded DNA (CRESS DNA) viruses: ubiquitous viruses with small genomes and a diverse host range. *Adv Virus Res*. 2019;103:71–133.
- [4] Gronenborn B, Randles JW, Knierim D, et al. Analysis of DNAs associated with coconut foliar decay disease implicates a unique single-stranded DNA virus representing a new taxon. *Sci Rep*. 2018;8:5698.
- [5] Wang H, Li S, Mahmood A, et al. Plasma virome of cattle from forest region revealed diverse small circular ssDNA viral genomes. *Virology*. 2018;15:11.
- [6] de Villiers EM, Borkosky SS, Kimmel R, et al. The diversity of torque teno viruses: in vitro replication leads to the formation of additional replication-competent subviral molecules. *J Virol*. 2011;85:7284–7295.
- [7] Peretti A, FitzGerald PC, Bliskovsky V, et al. Hamburger polyomaviruses. *J Gen Virol*. 2015;96:833–839.
- [8] Zur Hausen H, de Villiers EM. Dairy cattle serum and milk factors contributing to the risk of colon and breast cancers. *Int J Cancer*. 2015;137:959–967.
- [9] Moustafa A, Xie C, Kirkness E, et al. The blood DNA virome in 8,000 humans. *PLoS Pathog*. 2017;13:e1006292.
- [10] Sadeghi M, Kaupusinsky B, Yugo DM, et al. Virome of US bovine calf serum. *Biologicals*. 2017;46:64–67.
- [11] Kazlauskas D, Dayaram A, Kraberger S, et al. Evolutionary history of ssDNA bacilladnaviruses features horizontal acquisition of the capsid gene from ssRNA nodaviruses. *Virology*. 2017;504:114–121.
- [12] Krupovic M, Zhi N, Li J, et al. Multiple layers of chimerism in a single-stranded DNA discovered by deep sequencing. *Genome Biol Evol*. 2015;7:993–1001.
- [13] Roux S, Enault F, Bronner G, et al. Chimeric viruses blur the borders between the major groups of eukaryotic single-stranded DNA viruses. *Nat Commun*. 2013;4:2700.
- [14] Zur Hausen H, Bund T, de Villiers EM. Infectious agents in bovine red meat and milk and their potential role in cancer and other chronic diseases. *Curr Top Microbiol Immunol*. 2017;407:83–116.
- [15] Zur Hausen H, Bund T, de Villiers EM. Specific nutritional infections early in life as risk factors for human colon and breast cancers several decades later. *Int J Cancer*. 2019;144:1574–1583.
- [16] Jelcic I, Hotz-Wagenblatt A, Hunziker A, et al. Isolation of multiple TT virus genotypes from spleen biopsy tissue from a Hodgkin's disease patient: genome reorganization and diversity in the hypervariable region. *J Virol*. 2004;78:7498–7507.
- [17] Funk M, Gunst K, Lucansky V, et al. Isolation of protein-associated circular DNA from healthy cattle serum. *Genome Announc*. 2014;2:e00846–14.
- [18] Gunst K, zur Hausen H, de Villiers EM. Isolation of bacterial plasmid-related replication-associated circular DNA from a serum sample of a multiple sclerosis patient. *Genome Announc*. 2014;2(4):e00847–14.
- [19] Whitley C, Gunst K, Müller H, et al. Novel replication-competent circular DNA molecules from healthy cattle serum and milk and multiple sclerosis-affected human brain tissue. *Genome Announc*. 2014;2:e00849–14.
- [20] Lamberto I, Gunst K, Müller H, et al. Mycovirus-like DNA virus sequences from cattle serum and human brain and serum samples from multiple sclerosis patients. *Genome Announc*. 2014;2:e00848–14.
- [21] Eilebrecht S, Hotz-Wagenblatt A, Sarachaga V, et al. Expression and replication of virus-like circular DNA in human cells. *Sci Rep*. 2018;8:2851.
- [22] del Val C, Ernst P, Falkenhahn M, et al. Protsweep, 2DSweep and DomainSweep: protein analysis suite at DKFZ. *Nucl Acids Res*. 2007; W444–W450.
- [23] Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*. 1997;268:78–94.
- [24] Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucl Acid Res*. 2005;33:W465–W467.
- [25] Besemer J, Borodovsky M. Genemark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucl Acid Res*. 2005;33:W451–W454.
- [26] Kumar S, Stecher G, Li M, et al. MEGA X: molecular evolutionary Genetics analysis across computing platforms. *Mol Biol Evol*. 2018;35:1547–1549.

- [27] Falida K, Eilebrecht S, Gunst K, et al. Isolation of two virus-like circular DNAs from commercially available milk samples. *Genome Announc.* 2017;5:e00266–17.
- [28] Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. *Comp Appl Biosci.* 1992;8:275–282.
- [29] Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution.* 1985;39:783–791.
- [30] Priebe SD, Lacks SA. Region of the Streptococcal plasmid pMV158 required for conjugative mobilization. *J Bacteriol.* 1989;171:4778–4784.
- [31] Briddon RW, Martin DP, Roumagnac P, et al. Alphsatellitidae: a new family with two subfamilies for the classification of geminivirus- and nanovirus-associated alphsatellites. *Arch Virol.* 2018. doi:10.1007/s00705-018-3854-2.
- [32] Denoeud F, Henriot S, Mungpakdee S, et al. Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science.* 2010;330:1381–1385.
- [33] Manuelidis L. Nuclease resistant circular DNAs copurify with infectivity in scrapie and CJD. *J Neurovirol.* 2011;17:131–145.
- [34] Vallenet D, Nordmann P, Barbe V, et al. Comparative analysis of Acinetobacters: three genomes for three lifestyles. *PLoS ONE.* 2008;3:E1805.
- [35] Hamidian M, Kenyon JJ, Holt KE, et al. A conjugative plasmid carrying the carbapenem resistance gene blaOXA-23 in AbaR4 in an extensively resistant GC1 Acinetobacter baumannii isolate. *J Antimicrob Chemother.* 2014;69:2625–2628.
- [36] Longkumer T, Kamireddy S, Muthyala VR, et al. Acinetobacter phage genome is similar to Sphinx 2.36, the circular DNA copurified with TSE infected particles. *Sci Rep.* 2013;3:2240.
- [37] Jorgensen TS, Hansen MA, Xu Z, et al. Plasmids, viruses, and other circular elements in rat gut. 2017. Available from: <https://www.biorxiv.org/content/10.1101/143420v2.abstract>
- [38] Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol.* 1993;10:512–526.
- [39] Kearse MG, Wilusz JE. Non-AUG translation: a new start for protein synthesis in eukaryotes. *Genes Develop.* 2017;31:1717–1731.
- [40] Bertini A, Poirel L, Mugnier PD, et al. Characterization and PCR-based replicon typing of resistance plasmids in Acinetobacter baumannii. *Antimicrob Agents Chemother.* 2010;54:4168–4177.
- [41] Kelley LA, Mezulis S, Yates CM, et al. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc.* 2015;10:845–858.
- [42] Campos-Olivas R, Louis JM, Clerot D, et al. The structure of a replication initiator unites diverse aspects of nucleic acid metabolism. *Proc Natl Acad Sci USA.* 2002;99:10310–10315.
- [43] Varsani A, Krupovic M. Sequence-based taxonomic framework for the classification of uncultured single-stranded DNA viruses of the family *Genomoviridae*. *Virus Evol.* 2017;3:vew037.
- [44] Kilic T, Popov AN, Burk-Körner A, et al. Structural analysis of a replication protein encoded by a plasmid isolated from a multiple sclerosis patient. *Acta Crystallogr D Struct Biol.* 2019;75:498–504. doi:10.1107/S2059798319003991.
- [45] Senger M, Flores T, Glatting K-H, et al. W2h: WWW interface to the GCG sequence analysis package. *Bioinformatics.* 1998;14:452–457. Available from: <http://www.w2h.dkfz-heidelberg.de>
- [46] Iyer LM, Zhang D, Rogozin IB, et al. Evolution of the deaminase fold and multiple origins of eukaryotic editing and mutagenic nucleic acid deaminases from bacterial toxin systems. *Nucl Acids Res.* 2011;39:9473–9497.
- [47] Duffy S, Shackleton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet.* 2008;9:267–276.
- [48] Firth C, Charleston MA, Duffy S, et al. Insights into the evolutionary history of an emerging livestock pathogen: porcine circovirus 2. *J Virol.* 2009;83:12813–12821.
- [49] Casey JL, Gerin JL. Hepatitis D virus RNA editing: specific modification of adenosine in the antigenomic RNA. *J Virol.* 1995;69:7593–7600.
- [50] Sureau C, Negro F. The hepatitis delta virus: replication and pathogenesis. *J Hepatol.* 2016;64:S102–S116.
- [51] Gomis-Rüth FX, Solá M, Acebo P, et al. The structure of plasmid-encoded transcriptional repressor CopG unliganded and bound to its operator. *EMBO J.* 1998;17:7404–7415.
- [52] Stedman KM. Deep recombination: RNA and ssDNA virus genes in DNA virus and host genomes. *Annu Rev Virol.* 2015;2:203–217.
- [53] Fondi M, Bacci G, Brillì M, et al. Exploring the evolutionary dynamics of plasmids: the Acinetobacter pan-plasmidome. *BMC Evol Biol.* 2010;10:59.
- [54] Boer DR, Ruiz-Masó JA, Rueda M, et al. Conformational plasticity of RepB, the replication initiator protein of promiscuous streptococcal plasmid pMV158. *Sci Rep.* 2016;6:20915.
- [55] Lean SS, Yeo CC. Small, enigmatic plasmids of the nosocomial pathogen, Acinetobacter baumannii: good, bad, who knows? *Front Microbiol.* 2017;8:1547.
- [56] Miras M, Miller WA, Truniger V, et al. Non-canonical translation in plant RNA viruses. *Front Plant Sci.* 2017;8:494.
- [57] Hassan I, Orillio AF, Fiallo-Olivé E, et al. Infectivity, effects on helper viruses and whitefly transmission of the deltasatellites associated with sweepoviruses (genus Begomovirus, family Geminiviridae). *Sci Rep.* 2016;6:30204.