

1 **Preempting Pandora's Box: *Blastocystis* Subtypes Revisited**

2

3 Christen Rune Stensvold^{1*}, C. Graham Clark²

4

5 1: Statens Serum Institut, Copenhagen, Denmark

6 2: London School of Hygiene and Tropical Medicine, London, UK

7

8 *Correspondence: run@ssi.dk

9

10 *Blastocystis* is a genetically diverse intestinal protist colonising both human and
11 non-human hosts. By 2013, 17 subtypes had been acknowledged. Since then,
12 nine more subtypes have been proposed. We argue that several recently
13 proposed subtypes are invalid. We also revisit recommendations regarding the
14 requirements for annotating sequences as new subtypes.

15

16 In 2007, an article was published that sought to clarify the nomenclature applied
17 to genetic variants of *Blastocystis* [1]. This stramenopile is probably the most
18 widespread non-fungal microeukaryote present in the human gastrointestinal
19 tract. Remarkable genetic diversity had been uncovered by numerous groups
20 working independently around the world, each of which had introduced its own
21 naming scheme for the genetic variants detected. A consensus was reached
22 that proposed the existence of nine genetic groups of *Blastocystis* in humans
23 and named them 'subtypes'. The identifications were based primarily on
24 differences among the small subunit ribosomal RNA (SSU) gene sequences. It
25 was recognised at that time that most of the nine subtypes were also found in
26 other mammals and birds, but that most *Blastocystis* from reptiles and amphibia
27 fell outside these groups.

28

29 This subtype system (Box 1) has proven very useful, and has been adopted
30 almost universally among those performing research into this organism. In
31 2007, the majority of samples analysed had been of human origin. Inevitably,
32 once additional hosts started to be examined in significant numbers, new
33 subtypes were quickly identified. By 2013, no fewer than eight more subtypes
34 had been proposed, and all of them had been identified in non-human hosts
35 [2]. Subsequently, an additional nine have been reported, also in non-human
36 hosts. However, we are concerned that the evidence on which some of the
37 post-2013 subtypes have been based is insufficient and potentially misleading.
38 Indeed, we believe that some of the new subtypes are the result of experimental
39 artifacts. The aim of the current review is to evaluate the validity of the

40 seventeen post-2007 subtypes and propose minimum criteria for the future
41 naming of new subtypes.

42

43 **Subtypes described between 2007 and 2013**

44 Subtype 10 was described in 2009 based on sequences from two non-
45 overlapping regions of the SSU gene [3]. A complete gene sequence for ST10
46 (KC148207) was obtained only four years later [2], and subsequently ST10 has
47 gone on to be recognised as a very common subtype in cattle, sheep and other
48 artiodactyls worldwide [3]

49

50 Subtypes 11 and 12 were detected initially in zoo animals and were based on
51 the sequence of about 60% of the SSU gene [4]. A near-complete sequence of
52 ST12 was actually deposited in GenBank a year later (EU427515), but this was
53 not recognised until recently because of the way that BLASTⁱ ranks sequence
54 matches. No complete sequence of ST11 is yet available, to our knowledge.

55

56 The absence of a full-length sequence for ST11 is potentially problematic. The
57 'missing' region of the gene is one that is commonly used for subtype
58 identification, the so-called "barcode region" [5]. A novel barcode sequence
59 might be proposed as representing a new subtype when in fact it is actually the
60 missing region of ST11. This situation is not farfetched, as a similar
61 misidentification happened with ST13. A barcode sequence previously reported
62 as a variant of ST5 [6] actually proved to be the barcode region of ST13 when
63 a full length sequence for the latter was described in 2013 [2].

64

65 ST13 to ST17 are based on almost full-length SSU gene sequences obtained
66 from a variety of non-human hosts [2, 7]. So, with the exception of ST11, all the
67 new subtypes reported between 2007 and 2013 are represented by full- or
68 almost full-length SSU gene sequences. Some were derived from sequencing
69 of cloned PCR products, others from direct sequencing of PCR products, but
70 all have now been isolated multiple times, usually in multiple different hosts and
71 by several independent researchers, and they form discrete clades in
72 phylogenetic trees. We have no doubt that ST11 to ST17 are all 'real'.

73

74 **Subtypes described after 2013**

75 Recently, subtypes numbered 18 through 26 have been proposed [8=10].
76 However, we do not believe that all of these are real and will discuss below the
77 different factors we have considered in reaching our conclusions. In particular,
78 we believe that some of them are actually molecular chimaeras and will briefly
79 describe how these are generated and how to recognise them.

80

81 Chimaeras arise during PCR amplification, usually when there are two distinct
82 subtypes in the DNA sample and when there is incomplete replication of a DNA
83 strand during a cycle. After denaturation in the next cycle, one single-stranded
84 partial product may anneal to a single stranded product derived from a different
85 subtype; this is possible due to the extensive sequence similarity in some
86 regions of the gene. Extension then results in a PCR product combining
87 sequences from the two sources (subtypes, or even different organisms). The
88 conservation of SSU genes means there can be sufficient similarity to allow
89 binding even between products derived from distantly related organisms.

90

91 Chimaeras are generally only detected when the PCR products are cloned
92 before sequencing, although they are also common in sequence data obtained
93 by Next Generation Sequencing. Where a PCR product is sequenced directly
94 using a dideoxynucleotide-based chain termination method, the chimaera
95 sequences present will be 'diluted out' because the sequence obtained is the
96 average of all the products in that reaction, and so the sequence read will be
97 that of the major product of the reaction. Only when single products from that
98 mixture are studied in isolation will chimaeras be detected.

99

100 In the original *Blastocystis* 'barcoding' publication of Scicluna et al. [5], a
101 sequence was identified in GenBank (AF538348) where the 5' and 3' ends
102 clearly derived from different subtypes. Several of the newly described
103 subtypes also appear to be chimaeras. ST19 [10] is similar to the example
104 above. The 5' half is 99% identical to ST3 sequences while the 3' half is 99%
105 identical to ST1 sequences. In contrast, in the sequence designated ST18 [10],
106 the 3' end shows no similarity to other *Blastocystis* at all, while the 5' end shows
107 over 90% identity to several *Blastocystis* subtypes. Similarly, for ST20 [10] the
108 very 5' end (130 bp) does not match any organisms, while the remainder is 96%
109 identical to ST5. For ST22 [10], the 5' end matches ST14 with 95% identity,
110 while the 3' end shows 99% identity to ST10. Each of these 'subtypes' was
111 reported only on one occasion and is represented by only a single sequence –
112 this is as would be expected from an artifact.

113

114 There are other *Blastocystis* sequences in GenBank that have not been
115 allocated to subtypes but are also chimaeras. For example, MH496651 is
116 partially *Blastocystis*, partially plant. Other examples include MH489079, which
117 appears to be mostly from a banana, and MH496654, which has a 5' end with
118 a 100% match to ST13 but a 3' end that has no similarity to *Blastocystis*, and
119 so on.

120

121 In contrast, subtypes 21 and 23-26 have all been isolated multiple times and in
122 most cases by research groups working in different countries (Table 1); this
123 strongly suggests that the sequences are not artifacts. However, all consist of
124 incomplete SSU gene sequences.

125

126 This raises the question of defining boundaries between subtypes. How
127 different does a sequence need to be before it can be considered a new
128 subtype? With incomplete sequences it is not possible to be prescriptive,
129 because regions of the SSU gene exhibit differing degrees of conservation and
130 therefore differ in the percentage divergence between subtypes. For this
131 reason, we previously recommended designating sequences as new subtypes
132 only if >80% of the SSU gene has been sequenced and if that sequence
133 diverges by more than 4% from previously sequenced complete *Blastocystis*
134 SSU genes [11]. Intra-subtype variation differs between subtypes but can be
135 up to 3% in, for example, ST1 and ST2, which is why the 4% cut-off was
136 selected. A particular issue is being seen in the cluster of subtypes that includes
137 ST5 and STs 12-14. Several of the proposed new STs are related to sequences

138 in this region of the tree and when their partial sequences are incorporated in
139 the phylogenetic analysis, the established clade structure breaks down.

140

141 Clearly, sequence length and reliability are critical to the process of allocating
142 sequences to subtypes of *Blastocystis*. In the case of sequences that may
143 represent novel subtypes, near-complete SSU gene sequences should be
144 generated before assigning a number and phylogenetic analyses involving a
145 standard set of reference sequences ⁱⁱⁱ should be used in the investigation.

146 Invalid subtypes must be kept to a minimum in order not to undermine the
147 subtype terminology. To this end we recommend that STs 18–20 and ST22 be
148 rejected, while STs 21 and 23–26 need to be investigated further to generate
149 full-length gene sequences – we acknowledge that the latter five subtypes are
150 likely to be confirmed as new but, at present, it is not clear how these five are
151 related to previously described subtypes.

152

153 **Conclusion**

154 While we recommend rejecting STs 18-20 and ST22, we do not believe it is a
155 good idea to reuse these ST numbers in the future, as this will only generate
156 confusion in the literature. We recommend keeping ST21 and STs 23-26 until
157 further data lead to them being confirmed or rejected. The next new subtype
158 should therefore be named ST27 and we recommend to anyone aiming to
159 report a sequence as representing a new subtype that they follow the
160 guidelines in Box1.

161 **Box 1: Subtyping *Blastocystis* – proposed guidelines**

162 Application of the subtype system for *Blastocystis* relies on our ability to obtain
163 accurate identification while allowing for a certain amount of variation. The
164 terminology should be sufficiently detailed to permit identification of major
165 groups that may differ in epidemiology, host specificity, and potentially variation
166 in virulence.

167 The 10 subtypes known to colonise humans (subtypes 1–9 and 12) are easily
168 differentiated using e.g. barcode sequences [5, 12] and querying these against
169 the Blastocystis Subtype (18S) and Sequence Typing (MLST) Databasesⁱⁱ.
170 More than 90% of human *Blastocystis* belongs to subtypes 1, 2, 3, and 4 [13].
171 Many hosts still await sampling, so new subtypes and hosts of *Blastocystis*
172 likely await discovery. Subtype calling of non-human *Blastocystis* should be
173 carried out with caution and not be based solely on top BLAST hits in the NCBI
174 Database.

175 When potentially new subtypes are discovered, we recommend the following:

- 176 • New STs should be based on $\geq 80\%$ of the ca. 1800 bp SSU gene.
- 177 • New STs should normally differ by $\geq 4\%$ from previously known
178 STs.
- 179 • New ST sequences should be checked for chimaerism using
180 appropriate software; separate BLAST analysis of each end, at a
181 minimum.
- 182 • Standard primer sequences for amplifying and sequencing PCR
183 products should be used, such as those mentioned in studies by e.g.,
184 Stensvold et al. [14], and Santin et al [15].

185 • New STs should undergo phylogenetic analysis to ensure they do
186 not nest within previously known STs.

187 • The most recent *Blastocystis* reference set of ST sequences ⁱⁱⁱ
188 should be used for phylogenetic analyses.

189 We encourage researchers to contact the authors
190 (crs@blastocystis.net) before proposing a new subtype. The authors will
191 gladly provide an opinion as to whether they believe it qualifies as a new
192 subtype, indicate the subtype number to be used and add the sequence
193 to the reference set ⁱⁱⁱ. If all proposals adhere to this procedure, there will
194 be very little risk that two different variants will have the same subtype
195 number. It is planned that at the 3rd International *Blastocystis*
196 Conference (Crete, 2021) a community subtype working group will be
197 established to take on this responsibility going forward.

Table 1. Novel subtypes of *Blastocystis* published after 2013 that are probably valid based on analysis of currently available data.

Proposed subtype	Host and accession number in GenBank	# SSU bases	Comment
Subtype 21	<i>Kobus ellipsiprymnus</i> KY823403	896	A region of 335 bp is shared between sequences from the two sources, with 99% identity. Samples from China and N. America
	<i>Bos taurus</i> MH634461 MH634462	480	
Subtype 23	<i>Bos taurus</i> MH634463 MH634464, MH634465 MH634466 MK244936	477–479	All sequences to date are from N. America
Subtype 24	<i>Bos taurus</i> MH634467 MH634468 MH634469 MK244942 MK244937 MK244938 MK244939 MK244940 MK244941	478–480	>99% identity, samples from N. America and Belgium and multiple hosts, but also 94–97% identity to ST14
	HF569224	439	
	<i>Ovis aries</i> HF569209 HF569219	439	
	<i>Lama glama</i> HF569216	439	
Subtype 25	<i>Bos taurus</i> MH634470 MK244943 MK244944	475–480	>99% identity, samples from N. America and Belgium and multiple hosts, but also 97.5% identity to ST14
	<i>Ovis aries</i> HF569213	440	

Subtype 26	<i>Bos taurus</i>		>98% identity, samples from N. America, Thailand and Belgium and multiple hosts
	MH634471	447–480	
	MH634472		
	MH634473		
	MH634474		
	MH634475		
	MH634476		
	MH634477		
	MH634478		
	MK244945		
	MK244946		
	MK244947		
	MK244949		
	MK244948		
	MK244950		
	MK244951		
	MK244952		
MK244953			
HF569225	438		
<i>Ovis aries</i>			
HF569204	439		
Hosts			
unstated			
MH104960	1077		
MH104964	1077		
MH104966	1086		

200
201

202 **Resources**

- 203 ⁱ <https://blast.ncbi.nlm.nih.gov/Blast.cgi>
204 ⁱⁱ <https://www.pubmlst.org/blastocystis>
205 ⁱⁱⁱ <http://entamoeba.lshtm.ac.uk/blastorefseqs.htm>

206

207 **References**

- 208 1. Stensvold, C.R. et al. (2007) Terminology for Blastocystis subtypes--a
209 consensus. *Trends Parasitol* 23 (3), 93-6.
210 2. Alfellani, M.A. et al. (2013) Genetic diversity of Blastocystis in livestock and
211 zoo animals. *Protist* 164 (4), 497-509.
212 3. Stensvold, C.R. et al. (2009) Subtype distribution of Blastocystis isolates from
213 synanthropic and zoo animals and identification of a new subtype. *Int J Parasitol*
214 39 (4), 473-9.
215 4. Parkar, U. et al. (2010) Molecular characterization of Blastocystis isolates from
216 zoo animals and their animal-keepers. *Vet Parasitol* 169 (1-2), 8-17.
217 5. Scicluna, S.M. et al. (2006) DNA barcoding of Blastocystis. *Protist* 157 (1), 77-
218 85.
219 6. Petrášová, J. et al. (2011) Diversity and host specificity of Blastocystis in
220 syntopic primates on Rubondo Island, Tanzania. *Int J Parasitol* 41 (11), 1113-20.
221 7. Fayer, R. et al. (2012) Detection of concurrent infection of dairy cattle with
222 Blastocystis, Cryptosporidium, Giardia, and Enterocytozoon by molecular and
223 microscopic methods. *Parasitol Res* 111 (3), 1349-55.
224 8. Maloney, J.G. et al. (2019) Zoonotic and genetically diverse subtypes of
225 Blastocystis in US pre-weaned dairy heifer calves. *Parasitol Res* 118 (2), 575-
226 582.
227 9. Maloney, J.G. et al. (2019) Next generation amplicon sequencing improves
228 detection of Blastocystis mixed subtype infections. *Infect Genet Evol* 73, 119-
229 125.
230 10. Zhao, G.H. et al. (2017) Molecular characterization of Blastocystis sp. in
231 captive wild animals in Qinling Mountains. *Parasitol Res* 116 (8), 2327-2333.
232 11. Clark, C.G. et al. (2013) Recent developments in Blastocystis research. *Adv*
233 *Parasitol* 82, 1-32.
234 12. Stensvold, C.R. and Clark, C.G. (2016) Molecular identification and subtype
235 analysis of Blastocystis. *Curr Protoc Microbiol* 43, 20A.2.1-20A.2.10.
236 13. Alfellani, M.A. et al. (2013) Variable geographic distribution of Blastocystis
237 subtypes and its potential implications. *Acta Trop* 126 (1), 11-18.
238 14. Stensvold, C.R. et al. (2011) Increased sampling reveals novel lineages of
239 Entamoeba: consequences of genetic diversity and host specificity for taxonomy
240 and molecular detection. *Protist* 162 (3), 525-41.
241 15. Santín, M. et al. (2011) Development of a new PCR protocol to detect and
242 subtype Blastocystis spp. from humans and animals. *Parasitol Res* 109 (1), 205-
243 12.
244

245

