Preempting Pandora's Box: *Blastocystis* Subtypes Revisited

Christen Rune Stensvold[1]*, C. Graham Clark[2]

1: Statens Serum Institut, Copenhagen, Denmark

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

*Correspondence: run@ssi.dk

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

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

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

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

Subtypes described between 2007 and 2013

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

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

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

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

Subtypes described after 2013

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

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

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

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

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

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

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

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

Clearly, sequence length and reliability are critical to the process of allocating sequences to subtypes of *Blastocystis*. In the case of sequences that may represent novel subtypes, near-complete SSU gene sequences should be generated before assigning a number and phylogenetic analyses involving a standard set of reference sequences [iii] should be used in the investigation. Invalid subtypes must be kept to a minimum in order not to undermine the subtype terminology. To this end we recommend that STs 18–20 and ST22 be rejected, while STs 21 and 23–26 need to be investigated further to generate full-length gene sequences – we acknowledge that the latter five subtypes are likely to be confirmed as new but, at present, it is not clear how these five are related to previously described subtypes.

Conclusion

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

Box 1: Subtyping *Blastocystis* – proposed guidelines

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

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

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

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

- New STs should undergo phylogenetic analysis to ensure they do not nest within previously known STs.
- The most recent *Blastocystis* reference set of ST sequences [iii] should be used for phylogenetic analyses.

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

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

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

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

200
201

Resources

[i] https://blast.ncbi.nlm.nih.gov/Blast.cgi

[ii] https://www.pubmlst.org/blastocystis

[iii] http://entamoeba.lshtm.ac.uk/blastorefseqs.htm

References

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