Vampires in the oceans: predatory cercozoan amoebae in marine habitats.

Cédric Berney, Sarah Romac, Frédéric Mahé, Sébastien Santini, Raffaele Siano, and David Bass

Supplementary File 1

Detailed Materials and Methods for the BioMarKs sampling and 454 sequence data curation

Sampling

Through the BioMarKs consortium (http://www.BioMarKs.org/), six European coastal stations were sampled offshore: (1) Oslo, Norway (GPS position 59°16'N, 10°43'E) in September 2009 and June 2010, (2) Naples, Italy (GPS position 40°48.5'N, 14°15'E) in October 2009 and May 2010, (3) Blanes near Barcelona, Spain (GPS position 41°40'N, 2°48'E) in February 2010, (4) Roscoff, France (GPS position 48°46'N, 3°57'W) in April 2010, (5) Varna, Bulgaria (GPS position 43°10'N, 28°50'E) in May 2010, and (6) Gijon, Spain (GPS position 43°40'N, 5°35'W) in September 2010. More details of the environmental conditions at the sampling sites are available in Logares *et al*., 2012. Each station was sampled over three depth levels: sediment, deep chlorophyll maximum (DCM), and sub-surface. The same sampling protocol was used at all sites (as described by Rodriguez-Martinez *et al*., 2012). Briefly, sediment samples were taken with sediment cores. Small aliquots of the surface sediment material (~1 cm^3 on the surface of the sediment core) were frozen and stored at -80 °C for molecular analysis. For water samples, between 30 and 50 litres of seawater were collected using Niskin bottles attached to a CTD rosette at the surface and at the deep chlorophyll maximum (DCM) level. Water samples were pre-filtered with 2000 µm pore size filters, then size-fractioned using different pore size polycarbonate filters of 142 mm diameter, into three size fractions: between 20 and 2000 µm, between 3 and 20 µm, and between 0.8 and 3 µm. Filtration time was kept below 30 min. to minimise RNA degradation. All filters were flash frozen and stored at -80°C for further analysis.

DNA/RNA extraction and 454 sequencing

For water column samples, total DNA and RNA were extracted simultaneously from the same filters using the NucleoSpin RNA L kit (Macherey-Nagel, Düren, Germany). For sediment samples, DNA and RNA were isolated using PowerMax Soil DNA Isolation kit and PowerSoil total RNA Isolation kit (MoBio, USA), respectively. DNA and RNA quality were confirmed using gel electrophoresis (on 1.5% agarose gel) and quantified with a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies Inc, Wilmington, DE, USA). To avoid DNA contamination in the RNA extractions, DNAse from the TurboDNA kit (Ambion, Carlsbad, CA, USA) was used to remove any trace of DNA. Extracted RNA was immediately reverse transcribed into cDNA using the RT Superscript III kit with random primers (Invitrogen, Carsbad, CA, USA) following the protocol outlined by the manufacturer. The universal eukaryote primers TAReuk454FWD1 (5'-CCAGCASCYGCGGTAATTCC-3') and TAReukREV3 (5'-ACTTTCGTTCTTGATYRA-3') were used to amplify the V4 region (around 400bp) of the SSU rDNA (see Stoeck et al. 2010) by PCR. The primers were adapted for 454 sequencing with the configuration A–adapter–tag (of 7 or 8 bp)–forward primer and a B–adapter–reverse primer, as outlined in the manufacturers' instructions. PCRs were performed in 25 µl volumes of 1X MasterMix Phusion High-Fidelity DNA polymerase (Finnzymes, Espoo, Finland), 0.35 µM of each primer, 3% DMSO, and 5 ng of template of DNA or cDNA. PCR reactions consisted of an initial denaturation step at 98°C for 30s, followed by 10 cycles of 10s at 98°C, 30s at 53°C and 30s at 72°C, and then 15 cycles of 10s at 98°C, 30s at 48°C and 30s at 72°C. All PCR products were conducted in triplicate, checked using agarose gel electrophoresis, pooled and purified using NucleoSpin Extract II (Macherey-Nagel, Düren, Germany), eluted in 30µl of elution buffer, and quantified using NanoDrop ND-1000 spectrophotometer. The total final amount of pooled PCR amplicons for 454 sequencing was about 1 µg. Amplicon sequencing was carried out on a 454 GS FLX Titanium system (454 life sciences, Branford, CT, USA) installed at Genoscope in France (http://www.genoscope.cnrs.fr/spip/).

Curation of the sequence data and taxonomic assignment

Raw sequence reads were trimmed and filtered according to the following procedure. Multiplexed reads were separated into sample bins based on their tag sequences. Only reads with exact forward and reverse primer sequences were retained. Furthermore, reads presenting a computed global error rate above 1% were rejected. Strictly identical reads within a given sample bin were merged to reduce the volume of data and avoid redundant computation using a pipeline of in-house perl scripts ("dereplication" step). Reads were classified into higher-level taxonomic groups by comparison with a reference database based on SILVA 108 release (http://www.arb-silva.de/) and containing 20,972 reference SSU rDNA sequences checked with KeyDNAtools (http://keydnatools.com/) and curated by Laure Guillou. Sequence taxonomic paths were standardised to eight ranks: Domain, Supergroup, Division, Class, Order, Family, Genus, and Species. Taxonomic annotations in the reference database were reviewed by group specialists, members of the BioMarKs consortium. In order to accelerate subsequent analyses, the V4 region was extracted from the reference database; any partial sequence that didn't span the whole fragment amplified with primers TAReuk454FWD1 and TAReukREV3 was excluded from the reference database. Chimera detection was run with the UCHIME module of USEARCH (Edgar, 2010; Edgar et al., 2011) and with ChimeraSlayer (Haas et al., 2011), both *de novo* and by using the sequences in the reference database. The filtered environmental reads were then mixed with the reference sequences and iteratively clustered with UCLUST (Edgar, 2010) using increasingly stringent clustering thresholds from 85% to 99%. Taxonomic assignation then proceeds backwards using the obtained clusters, as follows: all reads in any cluster containing a reference sequence at a clustering threshold of 99% receive the taxonomic assignment of that reference sequence (or are assigned to the last common ancestor of the reference sequences if there are several in that cluster). If the cluster contains no reference sequence, its parent cluster (at a clustering threshold of 98%) is tested for the presence of a reference sequence. The process is repeated until one of the parent clusters contains one or more reference sequence(s) and the reads in that cluster can be taxonomically assigned. If the initial parent cluster (at a clustering threshold of 85%) is reached without finding any reference sequence, the reads in that cluster are declared as taxonomically unknown.

Identification of vampyrellid reads

The taxonomic assignation process described above identified 1930 total sequences (1356 unique sequences after dereplication) that were classified as vampyrellids. Once crudely aligned, further visual checks of these sequences in search for contradicting taxonomic signatures (as described in Berney *et al*., 2004) identified 72 additional probable chimeric sequences that were excluded. Because sequences labelled as vampyrellids in the reference database were restricted to the terrestrial clade A, it was likely at this stage that additional vampyrellid reads from more basal lineages could be found in the dataset, that could only be attributed to higher taxonomic levels or were declared taxonomically unknown. Blastn searches were used to identify these. We didn't rely on a particular similarity threshold in these Blastn searches, because overall similarity scores depend largely on sequence divergence, and phylogenetically close but divergent sequences can have significantly lower Blastn scores than conserved sequences from totally distinct lineages. Instead we used phylogenetic analysis to test the position of all retrieved sequences. To maximise chances of identifying even the most divergent vampyrellid reads in the data, we used an iterative process: all sequences from our new isolates and from phylogenetically identified vampyrellid environmental clones in GenBank were used as Blastn seeds, and all newly found vampyrellid sequences in the BioMarKs reads were used as seeds in further Blastn searches, until we were confident that most if not all vampyrellid sequences had been found. Visual chimera checks were performed on these sequences as described above. 908 additional vampyrellid sequences were identified that way (840 after dereplication), leading to a total of 2766 non-chimeric vampyrellid reads in the BioMarKs data (2125 after dereplication). These 2766 sequences have been deposited as a sequence read archive in the European Nucleotide Archive (project accession number PRJEB1843, "BioMarKs V4 Vampyrellids": http://www.ebi.ac.uk/ena/data/view/ERP002516).

Sequence heterogeneity across sites in the targeted V4 region

Automated clustering methods with a given similarity threshold are extremely useful in cases where very large amounts of sequences are analysed and/or little information is available on the levels of sequence variation between real species for the molecular marker analysed. Their main limitation, however, resides in the difficulty of distinguishing sequence differences likely to correspond to PCR and sequencing errors from genuine sequence variation between distinct but closely related organisms. In this study, we were only analysing 2125 unique sequences and could benefit from the exhaustive information available on sequence heterogeneity across sites in the SSU rDNA (see for instance Wuyts *et al*., 2001). This allowed us to use a non-automated clustering approach (see below). The V4 fragment amplified with primers TAReuk454FWD1 and TAReukREV3 starts in the pseudoknot formed by helices 20 and 21, between the 5' stem of helix 21 and the 3' stem of helix 20, and ends at the base of the 5' stem of hairpin 27 (as numbered in Wuyts *et al*., 2000). After exclusion of the primer sites, the first 49 and last 99 nucleotide positions of the amplified fragment (from the 3' stem of helix 21 to the 5' stem of helix 23, and from the 5' stem of helix 24 to the 3' stem of hairpin 24, respectively) are highly conserved, often across all eukaryotes, and certainly across all known vampyrellids. These portions of the amplicons happen to correspond to regions where sequencing errors were most likely to have occurred (in particular the ones due to homopolymers with pyrosequencing), making them easy to identify as such. On the other hand, our exhaustive experience of eukaryotic SSU rDNA sequences indicates that substitutions between closely related species are only ever observed in a few "variability hotspots" of the V4 fragment, corresponding to helices and hairpins E23-1 to E23-7, the terminal loop at the end of hairpin E23-12, and the pseudoknot formed by helices E23-13 and E23-14. Remaining portions of the V4 region exhibit intermediate levels of sequence variability but importantly they are usually of conserved length and sequence within lower-level lineages.

Clustering of the vampyrellid reads

Our non-automated clustering approach was therefore based on the following criteria: (1) indels and single-nucleotide substitutions in the first 49 and last 99 nucleotide positions of the sequences were ignored as highly likely PCR and sequencing errors (indels outside of the first 49 and last 99 nucleotide positions of the sequences were all found to be associated with homopolymers and also conservatively ignored); (2) for V4 reads to be considered as distinct SSU-types, at least three substitutions had to be observed in total across at least two of the "variability hotspots" mentioned above, or two substitutions in a single "variability hotspot" if they were present in several sequences each from at least two independent samples; (3) a single sequence present in a single sample was considered a distinct phylotype only if it differed by at least five substitutions from all other sequences, in at least three "variability hotspots"; and (4) single-nucleotide substitutions in the remaining V4 portions of intermediate sequence variability were addressed individually and discarded as likely PCR and sequencing errors whenever they did not co-occur with a higher proportion of substitutions in the "variability hotspots" (as a result most were conservatively discarded). Our clustering approach resulted in 461 clusters (designated as "SSU-types" in the main text); they are provided in fasta format in Supplementary File 2. Supplementary Table S4 provides the correspondence between the 461 SSU-types and the 2125 dereplicated sequences they comprise, as well as the samples they came from (which are described in Supplementary Table S5). In the few cases where there were only two or three differences between SSU-types in a group of closely related clusters, the SSU-types were given the same number followed by small-case letters a, b, c, etc. to highlight the possibility that they might not really be distinct OTUs (see Figure 5, Supplementary Figure S1, and Supplementary Table S4). The average length of the targeted V4 SSU rDNA fragment is 388 bp. in vampyrellids, with 95% of the clusters between 385 and 391 bp. Our clustering approach therefore implies an *a posteriori* similarity threshold above 99%. However it is based on highly stringent and informed criteria, so that we believe it is biologically sensible.

Correspondence between SSU-types and protist "species"

The relatively few detailed studies of "species"-level differences between microbial eukaryotic lineages suggest that the phylogenetic resolution offered by SSU rDNA even in the V4 "variability hotspots" may be too low for making such distinctions. ITS rDNA or a protein-coding gene such as cytochrome oxidase I that evolve faster than SSU rDNA would be better "species" markers in protists (see reviews in Bass & Boenigk, 2010 and Boenigk et al., 2012). In some lineages, available data suggest that phenotypically and/or ecologically distinct organisms with identical SSU rDNA sequences can indeed be discriminated genetically using such faster evolving markers (see, e.g., Bass et al., 2007). Of course at present we do not have the necessary data to test whether this is true for vampyrellids as well. The only available pieces of information that relate to this issue are (1) the fact that we never observed intra-genomic SSU rDNA heterogeneity between copies of the gene within our isolates when sequencing multiple clones of the same PCR amplicon, even in the V4 "variability hotspots" - and these isolates cover almost the full phylogenetic range of vampyrellid sequences revealed in our study; and (2) the level of sequence divergence between the closest pairs of SSU-types from our clustering approach corresponds to that observed between phenotypically distinct soil isolates (see Figure 5). Together with the stringency of our clustering approach, these observations suggest that the sequence diversity found in the *BioMarKs* data does actually represent biologically distinct lineages, rather than intra-genomic variation or methodological artefact, and that we are not over-estimating the organismal diversity behind the SSU-types defined. Therefore we believe that unless/until proven otherwise an approach comparable to that used in other protist lineages for vampyrellids is the most transparent way to present the results of this study.

Cited references

Bass D, Boenigk J. (2010) Everything is Everywhere: a 21st Century De-/Reconstruction. Invited chapter for Systematics Association Special Volume: Biogeography of Microscopic Organisms: Is Everything Everywhere? (ed. D Fontaneto), Ch. 6. The Systematics Association, Cambridge University Press.

Bass D, Richards TA, Matthai L, Marsh V, Cavalier-Smith T. (2007). DNA evidence for global dispersal and probable endemicity of protozoa. *BMC Evol Biol* **7**: 162.

Berney C, Fahrni J, Pawlowski J. (2004). How many novel eukaryotic 'kingdoms'? Pitfalls and limitations of environmental DNA surveys. *BMC Biol* **2**: 13.

Boenigk J, Ereshefsky M, Kerstin Hoef-Emden K, Mallet J, Bass D. (2012) Concepts in protistology: species definitions and boundaries. *Eur J Protistol* **48**: 96-102.

Edgar RC. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**: 2194–2200.

Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, *et al*. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res* **21:** 494-504.

Logares R, Audic S, Santini S, Pernice MC, de Vargas C, Massana R. (2012). Diversity patterns and activity of uncultured marine heterotrophic flagellates unveiled with pyrosequencing. *ISME J* **6**: 1823-1833.

Rodriguez-Martinez R, Rocap G, Logares R, Romac S, Massana R. (2012). Low evolutionary diversification in a widespread and abundant uncultured protist (MAST-4). *Mol Biol Evol* **29**: 1393-1406.

Stoeck T, Bass D, Nebel M, Christen R, Jones MD, Breiner HW, *et al*. (2010). Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol Ecol* **19 (Suppl 1)**: 21-31.

Wuyts J, De Rijk P, Van de Peer Y, Pison G, Rousseeuw P, De Wachter R. (2000). Comparative analysis of more than 3000 sequences reveals the existence of two pseudoknots in area V4 of eukaryotic small subunit ribosomal RNA. *Nucleic Acids Res* **28**: 4698-4708.

Wuyts J, Van de Peer Y, De Wachter R. (2001). Distribution of substitution rates and location of insertion sites in the tertiary structure of ribosomal RNA. *Nucleic Acids Res* **29**: 5017-5028.