Freshwater Adaptation at the Molecular Scale in the Unique Sponges of Lake Baikal

Nathan J Kenny[1]
nathanjameskenny@gmail.com

Bruna Plese[1,2]
bplese@irb.hr

Ana Riesgo[1*]
A.Riesgo@nhm.ac.uk

Valeria B. Itskovich[3*]
itskovich@mail.ru

[1] Life Sciences Department, The Natural History Museum, Cromwell Road, London SW7 5BD, UK

[2] Division of Molecular Biology, Ruđer Bošković Institute, Bijenička cesta 54, 10000, Zagreb, Croatia

[3] Limnological Institute, Siberian Branch of the Russian Academy of Science, Ulan-Batorskaya, 3, Irkutsk, 664033 Russia

Abstract:

The Lake Baikal ecosystem is unique. The largest, oldest and deepest lake in the world presents a variety of rare evolutionary opportunities and ecological niches to the species that inhabit it, and as a result the lake is a biodiversity hotspot. More than 80% of the animals found there are endemic, and they often exhibit unusual traits. The freshwater sponge *Lubomirskia baicalensis* and its relatives are good examples of these idiosyncratic organisms. Lake Baikal sponges are unusually large for freshwater sponges, and possess particular symbiotic microorganisms which aid their survival and growth. As sponges are the most abundant benthic organisms in Lake Baikal, with 13 extant species (and 2 subspecies), they play a variety of key roles in the maintenance of this unusual ecosystem.

In recent years the flora and fauna of Lake Baikal have been challenged by human activity in the region. Pollution, invasive species, temperature increases, eutrophication and falling water levels have all had deleterious effects. Sponges have suffered in particular from bleaching events, where their symbiotes are no longer retained, with knock-on effects to their survival and that of the ecosystem as a whole.

To better understand the unique biology of these sponges we have established here a variety of resources for understanding these sponges at a molecular level. We have sequenced transcriptomes from three species of Lubomirskiidae. We have also assembled a draft genome resource for *L. baicalensis*. With these resources in place, a sizable addition to our knowledge of freshwater sponge genetics, we have investigated the adaptations of these and other freshwater sponge species to their environment, investigated their symbioses, and established a range of molecular tools for understanding the molecular and population-level biology of these key species across their habitat in Lake Baikal.

Key Words: Lake Baikal, Sponges, Porifera; Transcriptome; Genome; Symbiosis

Introduction:

The Lubomirskiidae (Demospongiae: Spongillida) family of sponges is endemic to Lake Baikal, southern Siberia, and is a vital part of that unusual ecosystem. This family contains four genera and thirteen currently accepted species. As sponges are the most prominent component of the benthic assemblage of Lake Baikal, making up around 44% of its biomass (Pile et al 1997), they play a vital role within this ecological community.

The Lubomirskiidae possess a number of unique features. They are large for freshwater sponges, growing to around 1 metre in size (Kozhov 1963). Freshwater sponges depend on symbiotic associations with various chlorophyll-producing microorganisms: green algae, dinoflagellates and diatoms (Pile et al 1997). The diversification of the Lubomirskiidae is yet to be fully understood. They can be described as a “species flock” (Schröder et al 2003), and recent studies have found discrepancies between molecular and morphological data used to define species boundaries in Baikalian sponges (Itskovich et al 2015). Further study of the population structure and speciation of these species is therefore merited, especially at the genetic level.

Lubomirskia baicalensis is the most abundant and best studied species of the Lubomirskiidae, and its common name, Lake Baikal sponge, reflects this. It is found throughout the lake, but its morphology varies according to depth (Kozhov 1963). It grows as a mat in shallower water, while in deeper areas of the lake it increases in height, potentially growing to more than a metre, and begins to branch. It can form dense aggregates in favourable conditions. Together with other sponges, it forms beds providing the habitat for a range of other species. Like other members of the Lubomirskiidae, *L. baicalensis* possesses symbiotic dinoflagellates related to *Gymnodinium sanguineum* (Kaluzhnaya et al 2011, 2012, Annenkova et al 2011, Chernogor et al 2013, Kaluzhnaya and Itskovich 2015).

Lake Baikal itself is highly unusual. It is the largest freshwater lake in the world by volume, the deepest and the oldest (Timoshkin, 2001, Rusinek et al., 2012a). This has allowed a large number of species to adapt to its unique conditions. A minimum of 1,000 species of plants and 2,500 species of animals are found there, of which more than four fifths are endemic (Timoshkin, 2001; Rusinek et al., 2012b). This high level of biodiversity is due to its age and its relative isolation, as the lake has exhibited approximately the same conditions for the past 2 – 4 million years (Kozhova and Izmest'eva, 1998), separated from other major freshwater ecosystems (Timofeyev, 2010). Temperatures vary widely from summer to winter, as the lake itself freezes over in winter (Kozhov 1963). The water temperature in the lake in regions deeper than 250m is a constant 3.3–4.3°C (Shimaraev et al., 1994). In summer, temperatures in

sheltered shallow regions can reach as high as 20° C (Pomazkina et al 2012), although this is unusual and a maximum temperature of 17° C is more common (Kozhova and Izmest'eva, 1998). Coupled with high oxygen levels throughout the water column and generally oligotrophic conditions (Kozhov, 1963), these circumstances provide a challenging environment, to which the animals native to Lake Baikal have adapted.

In the last century, the Lake Baikal ecosystem has been subject to a variety of negative environmental influences. This includes pollution from both agriculture and industry, varying water levels as a result of damming and agricultural irrigation, eutrophication as a result of fertilizer run off, and subsequent algal blooms, including those by invasive species (for example, Romanova et al 2015, Ciesielski et al 2016, Kasimov et al 2017). It is not yet understood how these impact on the endemic flora and fauna of the lake. Sponges, with their filter-feeding lifestyle, and particularly the Lubomirskiidae, with their reliance on symbiotes, may be particularly vulnerable to such pollution.

Bleaching in particular has been a major issue for the Lubomirskiidae in recent decades, especially in southern and central Baikal (Kaluzhnaya and Itskovich 2015, Khanaev et al 2017). Cases of novel diseases have also been reported, and may be related to bleaching (e.g. Kaluzhnaya and Itskovich 2015, 2017, Denikina et al 2016, Kulakova et al 2017, Itskovich et al 2018). These changes are often associated with widespread changes to the presence and the ratio of sponge associated bacterial communities, in ways that are only beginning to be understood. It is vital to understand the speciation patterns and diversity of the Lubomirskiidae in order to best understand the impact of disease, manage conservation efforts and ensure the future survival of the Baikal ecosystem (Khanaev et al 2017). However, to date our tools to accomplish this are limited, and the evolutionary patterns and taxonomy of these species is subject to some dispute.

The biology of Baikal-endemic species has been studied from a variety of angles in the past decades (e.g. Rusinek et al., 2012b). However, only in the last few years has it been possible to study the molecular adaptations of the species found there. The rise of “omic” technologies has allowed the study of even recalcitrant and rare species. The affordability of these and associated computational needs has made such an approach attractive as a means to solve a diverse range of biological problems (Goodwin et al 2016). This has lead to the investigation of a number of Baikal species using next-generation sequencing (e.g. Rivarola-Duarte et al 2014, Romanova et al 2016, Naumenko et al 2017). Such technologies have been highly successful in investigating Poriferan diversity (e.g. Riesgo et al 2014, Fernandez-Valverde et al 2015, Kenny et al 2017 and others). However, to date, there have been no

studies on sponges from Lake Baikal, and no freshwater sponge genomes are presently extant in the published record.

Modern tools for molecular ecology would allow us to properly study the evolutionary patterns underlying the speciation and population dynamics of the Lubomirskiidae and aid their conservation. Previous taxonomic work, based on a limited number of mitochondrial and nuclear markers, have been unable to disentangle these patterns (Schroder et al 2003, Itskovich et al 2015). Such resources also enable the study of a range of facets of the biology of these unique species. It would also be useful to understand more about their adaptation to freshwater and their responses to climatic changes.

To increase our understanding of the biology of these unique sponges from a range of perspectives, we have assembled and analysed a number of novel sequence resources. These include transcriptomes of three Baikal-endemic species, *Baikalospongia bacillifera, Lubomirskia abietina* and *L. baicalensis*. Alongside these samples we have also assembled a draft genome for *L. baicalensis.* These resources have allowed us to identify a number of key changes in genes involved in freshwater adaptation in these species, as well as aspects of the everyday biology of these animals. In particular, we have investigated the specific molecular signatures of adaptation in freshwater sponges. This will allow us to understand the biology of the idiosyncratic sponges of Lake Baikal, and freshwater sponges more generally, and to understand the impacts of anthropic activity on the biology of these vital components of this unique ecosystem.

Results and Discussion

Sequencing

Sequencing results were of mixed success rates, and full statistics can be seen in Table 1. Our transcriptomic sequencing was of reasonably good quality, with *B. bacillifera* (A2) and *L. abietina* (A10) of good average quality (Phred median 30 <) through to the final base in both read directions. The *L. baikalensis* (A8) sample, was of less robust quality, possessing reads which declined in base call confidence score towards the end of the read. The use of trimmomatic accounted for all of these issues, with few reads removed from samples A2 and A10, while sample A8 had 40.84% of read pairs removed, either due to poor quality or short read length after truncation. Where possible, single unpaired reads were retained from these filtered read pairs. Unpaired reads were used in assembly, but not in read mapping.

Our genomic sequencing required attention before use in genome assembly, due to a large number of apparent singleton *k*-mers (Fig 2), which could be indicative of read errors. Trimmomatic was applied to these reads first, removing and truncating large quantities of reads, leaving just over half of the original read pairs and bases. The 99,040,746 (53.68%) of pairs remaining after Trimmomatic were fed into rCorrector, which corrected 67,842,587 bases in 34,392,948 read pairs. rCorrector also discarded 17,998,543 pairs of reads, which were assessed as being unfixable. This left a total of 81,042,203 read pairs for further analyses. This process is highly stringent, and in a more complete genome assembly low complexity sequence could be spanned using long reads and a variety of library sizes, but for our purposes, this ensured the quality of sequence data used in assembly, and thus the reliability of assembled data. Both clean and original reads have been uploaded to the NCBI SRA, with accession number PRJNA431612.

Assembly Metrics and Overview

Metrics related to the assemblies of the transcriptomes studied in this work can be seen in Table 2. Individual assemblies were made using Trinity for all transcriptome datasets. Given that we wish to consider both bacterial and metazoan sequences in the present study, it is worth considering the relative GC% of reads and our assemblies. The GC% of bacterial genomes is often, but not always, higher than that of metazoans, and higher GC% can empirically be an indicator of higher levels of bacterial sequence in a sample. The GC% of our paired transcriptomic reads remained stable at 49% for all samples both before and after read

cleaning. This stability indicates that all samples were of comparable composition, and none contained any obvious levels of contamination relative to the others. The GC% of our genomic reads declined slightly with read cleaning, from 43.5 to 41%. This finding is confirmed in our Blobplots (Fig 3) which show that the small amount of bacterial sequence has a very similar GC% to our assembled reads. This prohibits us from "binning" bacterial sequence by GC% to construct separate assemblies, but bacterial sequence could be recognised by BLAST similarity, as discussed elsewhere in this work.

The quantity of this bacterial contamination (or symbiont sequence) was, however, quite small, less than 1/50th of the quantity of sponge data for our transcriptomes, as can be seen most clearly in Fig 4, which shows the distribution of our data by Superkingdom and Phylum. The proportion of bases mapped to Bacteria in our genomic sequencing was higher, around 5%, and may well consist of symbiont sequence of utility to future studies.

Summary statistics relating to genome assemblies can be seen in Table 3. A number of assembly algorithms were assessed for their utility, with SPAdes being the best performed by several metrics. For our purposes, the higher median contig size, larger number of long (>1kb) contigs and higher number of bases in these 1kb+ contigs (155,282,664 bp) meant that an increased number of long contigs was available for our work. That SPAdes outperformed other assemblers was not in itself a surprise, as its more recent publication and paired assembly graph approach meant that it incorporates methods not always found in other assemblers. However, other assemblers were better performed by some metrics (for example, ABySS generated the largest single contig). The addition of long read data or multiple library sizes would allow us to further scaffold this data in the future. Assemblies, including preliminary and alternate forms, have been uploaded to Figshare, with DOI: and URL: 10.6084/m9.figshare.6819812, https://figshare.com/s/fe36239c32bbf7342756.

We utilized GenomeScope to gain an understanding of the coverage of our dataset, potential levels of non-sponge DNA, and to estimate the genome size of *L. baikalensis.* The results of this analysis can be seen in Fig2. The inferred genome size of *L. baicalensis,* between 558 and 565 million base pairs in size, is considerably larger than that of many previously published sponge genomes (Table 4). At the same time, the mitochondrial genome of this sponge is the largest found in demosponges (Lavrov et al 2012). The genome of *Amphimedon queenslandica* was estimated to be approximately 167 million base pairs in haploid size (Srivistava et al 2010) and *Tethya wilhelmina* possesses approximately 125 megabase pairs (Francis et al 2017). Estimates of 75 sponge species using traditional techniques (Jeffrey et al 2013) ranged from 0.04–0.63 pg in haploid size, or approximately 40-630 megabases. Smaller

genomes were much more common, with only 6 genomes larger than 400 megabases. Our genome is therefore definitely larger than the average sponge genome. The freshwater sponges in that latter dataset, three members of the Spongillidae, ranged in size from 0.31 to 0.36 pg. The size of *L. baicalensis'* genome is therefore not solely due to a freshwater environment, although this may play a part in its larger size. The unique demographic history and environment of Lake Baikal, and the unique evolutionary history of the lineage leading to *L. baicalensis,* will have played roles in the expansion of the genome of this sponge, in ways that will only be understood through examination of datasets such as the one presented here.

The number of bases recovered by our SPAdes assembly (209,989,122 at 500 bp minimum contig size) falls short of the estimated genome size of this sponge by a considerable margin, but would comfortably contain the coding sequence of the average sponge genome. It is likely that repetitive sequences are poorly represented in our dataset, due to the single small fragment size used in sequencing. To estimate how much of the coding set of genes could be present in our genomic assembly, we used BLASTN megablast (-evalue 0.000001 -num_threads 8 -max_target_seqs 1 -outfmt 6) to ascertain how many of the contigs present in our *L. baicalensis* transcriptome were present in the genomic assembly. 77,717 of the 81,951 contigs in the *L. baicalensis* transcriptome possessed a hit in the genome at this stringency. This strongly suggests that the transcribed cassette is well-represented by the genomic assembly.

Annotation of assemblies

To test the content of both our genomic and transcriptomic assemblies, we used the BUSCO ("Basic Universal Single Copy Orthologue") complements of highly conserved genes. In particular, we compared our datasets to the metazoan and eukaryote BUSCO complements, consisting of 978 and 303 genes respectively. The recovery of these gene families in our datasets was almost complete (Table 5), and suggests that our transcriptomic assemblies are excellent molecular resources for future work. Of the 303 genes in the eukaryotic BUSCO set, the maximum number of missing genes is 4 in our transcriptomic complements, and 17 in our genome assembly. Similarly, of the 978-strong metazoan BUSCO complement, the maximum number of missing genes is 46 in our transcriptomes, and 169 in our genome assembly. This compares favourably with the *A. queenslandica* genome. In the published cDNA set for that species, 1.6% of the eukaryote set (5 genes) is missing, and 4.9% (49) of the metazoan complement. Our transcriptomic assemblies therefore are more complete in terms of gene recovery than that published resource, while our genome is almost as complete, despite its low

read coverage depth and limited library complexity. Given this level of recovery of the BUSCO set, these resources are therefore likely to contain the vast majority of the genetic cassette of these species, with any genes expressed at the time of sampling likely to be recovered by our assembly.

We performed automatic annotation of the transcriptomic resources detailed here using BLAST2GO, Interproscan and ANNEX. The complete annotations are attached to this manuscript as Supplementary File 1. In summary, of 80,829 total contigs, *B. bacillifera* (A2) possessed 13,454 fully annotated contigs, while 58,890 contigs possessed no BLAST hit. For *L. abietina* (A10), 93,299 contigs included 14,623 fully annotated sequences, with 69,862 had no BLAST similarity at the given thresholds. In *L. baicalensis* (A8), these numbers were 81,877 contigs, 14,772 annotations, and 57,476 without significant similarity. The number of annotated contigs is therefore relatively stable between all of our samples, and as can be seen in Fig 4, tend to be longer than those without significant similarity, as between 75 and 80% of the assembly (by nucleotide size) is incorporated into contigs with a BLAST hit.

It is important to note that novelties, such as novel genes specific to Lake Baikal sponges, will not be well described by this annotation process, which depends on similarity in the first instance for annotation to occur.

For all these species, the most commonly top-hit species by BLAST searches was the sponge *Amphimedon queenslandica* with 11,328, 12,236 and 12,358 hits respectively. The affinity of this species, the first sequenced Poriferan, with our samples is obvious. As *A. queenslandica* is also a demosponge, and has its genome on the nr database, it is unsurprising that it is the species with the highest number of BLAST top-hits. The top five most hit species for all three samples also included *Orbicella faveolata, Stylophora pistillata, Exaiptasia pallida* and *Branchiostoma belcheri,* although the order in which these occurred slightly changes from sample to sample. The first three of these species are cnidarians, and *B. belcheri* is a cephalochordate. These species have a relatively slow rate of molecular evolution, likely resulting in their similarity to our sequences under BLAST search.

There is no clear signal of specific microbial contamination of our read data, which would be obvious at this point in the form of bacterial species with high numbers of “top hits”. However, some bacterial sequence is present, particularly that of the likely symbiont (Wilson et al 2014) *Candidatus* Entotheonella gemina*,* with 88, 76 and 85 hits respectively for our three transcriptomes. This species is therefore likely the most commonly present bacteria in our sample, or at least the one with sequences present in Genbank for assignment. Other bacterial

species are also present, although less commonly. The fuller bacterial content of our resources is discussed in the next section of this manuscript.

Dinoflagellates are represented, albeit well down the list of commonly top-hit species, with *Symbiodinium microadriaticum* represented by 53, 66 and 51 sequences respectively. It is possible that some of the sequences without blast results represent dinoflagellate sequence, as the species noted to be symbiotes of Lake Baikal sponges, such as those of the genus *Gyrodinium* (Annenkova et al 2011), are not well represented in the nr database.

Please note that we have not performed full gene prediction or annotation on the genome resource, as the relatively low contiguity and N50 of this resource (a consequence of low coverage and a lack of long read data) precludes most genes from occurring in their full length on a single contig. However, some information on the annotation of this resource can be derived from BLAST searches performed for Blobtools analysis, which incorporates an annotation step (see Fig 3,4). For full coding gene identification we recommend the use of our transcriptomic resources, while the genome will be of use for microsatellite identification, mitochondrial genomics, additional sequence determination and other research into the wider biology of these species.

Bacterial Sequence

Our analyses revealed a small amount of bacterial sequence was present in our datasets. Fig 4 shows this clearly, with a maximum of 2.4% in our transcriptomes, and 5.44% in our genomic resource. The larger proportion of bacterial content in the genomic sample (SPAdes 500 bp+ assembly) is likely the result of poly A selection procedures in our transcriptomic samples, which will preferentially target eukaryotic mRNA.

We used the results of our Blobtools annotation pathway to understand the makeup of the bacterial sequence within our SPAdes 500 bp+ genome sample. The bacterial sequences that were present and identifiable in our genomic sample had a relatively high abundance but low N50 when compared to eukaryote-annotated sequence, with more contigs (30,243 vs 27,115 annotated contigs) but a much smaller N50, 1,607 bp vs 5,877 bp. This indicates to us that bacterial sequences present represent a subsample of complete bacterial diversity, and that no one bacterial species predominates to the exclusion of others.

Our *L. baicalensis* genome resource most commonly contained Proteobacterial sequence (9,880 annotated contigs, correlating to 1.9% of the reads mapped to our assembly) but also contained more than 1000 contigs with similarity to Bacteroidetes, Verrucomicrobia, Actinobacteria and Cyanobacteria, in order of decreasing occurrence. Less commonly,

Firmicutes, *Candidatus* Tectomicrobia (with a high N50, 3,788 bp), Spirochaetes were observed in more than 100 contigs. All of these groups, with the exception of *Candidatus* Tectomicrobia, had a much lower N50 than the eukaryote (5,877 bp) and poriferan (5,088 bp) figures. This indicates that bacterial genomes were not inordinately represented in our data, and therefore were not well-assembled. Consequently, it can be inferred that no single bacterial species predominates in our dataset, with *Candidatus* Tectomicrobia the most common.

The content of our transcriptomes was largely consistent from sample to sample. Our species top hit data revealed that *Candidatus* Entotheonella gemina was the most commonly hit individual species by BLASTP identity (Phylum: *Candidatus* Tectomicrobia). Our Blobtools analysis, incorporating BLASTX, was able to categorise the number of contigs further. These results are available in full in Supplementary File 2.

In *B. bacillifera* (A2) Proteobacteria were the most commonly observed bacterial Phylum, with 392 contigs with N50 1,275. *Candidatus* Tectomicrobia was behind, with 157 contigs of 1,145 bp N50. In order from most common to least, Bacteroidetes, Firmicutes, Actinobacteria, Cyanobacteria and Verrucomicrobia were also present in at least double digit numbers of contig, the lattermost with an exceptionally high N50 (7,259 bp). Only 9 contigs of Chlamydiae were present, but their high N50 (9.999 bp) and the obligately intracellular nature of these make them worth noting.In *L. abietina* (A10) this result was mirrored, with 407 Proteobacterial contigs (N50 1,537 bp) and 203 from *Candidatus* Tectomicrobia (1,121 bp). However, in *L. abietina,* the order of commonality in other bacterial groups is (most -> least) Actinobacteria, Firmicutes, Bacteroidetes, Cyanobacteria, Verrucomicrobia (N50 7,270) and Chloroflexi. Only 8 Chlamydiae contigs were present, but the N50 of these, 10,037 was again striking. *L. baicalensis* (A8) showed the same general trend, with one exception. Actinobacteria is the most commonly observed clade of bacteria in this sample, with 1,589 contigs of N50 297. Proteobacterial contigs (N50 1,087 bp) and *Candidatus* Tectomicrobia sp (191 contigs, 1,167 bp) are again well represented. Firmicutes, Bacteroidetes, Cyanobacteria, Verrucomicrobia (with a much smaller N50 than the other two samples, 976 bp) and Chlamydiae (N50 only 1,618 bp) then round out the most common bacterial content in decreasing order of occurrence. Representatives of these phyla were identified previously in the *L. baicalensis* community (Kaluzhnaya et al 2012).

This indicates to us that the general content of our resources is consistent. The most likely individual bacterial symbiont species within our samples is clearly related to *Candidatus* Entotheonella gemina (Phylum: *Candidatus* Tectomicrobia). It is striking that this bacterial species made the transition to freshwater to join its host, although the contemporaneity of this has not been established here. The functions provided by this species therefore remain vital,

even in the markedly differing freshwater environment (Wilson et al 2014, Liu et al 2016), and may be necessary for the survival of the host. This is the first clear example of a sponge symbiont that is present in both of these two environments, a fact that clearly merits further investigation.

Symbiont and Other Sequence Content

Symbiodinium microadriaticum was represented by53, 66 and 51 BLAST top-hits respectively in *B. bacillifera* (A2) *L. abietina* (A10) and *L. baicalensis* (A8). Of these hits, most were represented in two or more of the three species, with 7 represented in all three species. Only 3 (OLQ15661.1 OLP92806.1 OLP93934.1), 1 (OLP92903.1) and 2 (OLQ11623.1 OLP83851.1) sequences from the nr database present in only one of the three species. The degree of overlap between the hits in our samples leads us to conclude that all species of sponge make use of the same diatom symbionts (likely the genus *Gyrodinium* (Annenkova et al 2011)) and that the same genes are expressed at high levels by these. However, they may not be common within the samples at this point of the year, or more genes would be identified in our transcriptome samples.

We were also able to use our Blobplot results to examine the contents of our genomic resource. A total of 104 contigs, with an N50 length of 2,651 bp, were identified as belonging to the Bacillariophyta by blast similarity. This is not a high number, and indicates that the symbionts are not present in high numbers within our samples, at least at the time of sampling.

Many of the best assembled sequences present in our genomic data derive not from diatoms, but instead from unicellular algaes. These represent another potential symbiotic species that should be investigated further. In our genomic sample, the Chlorophyta are the second most commonly-hit Phylum after Porifera, with 1,359 very well assembled contigs of N50 21,412 bp - an N50 fourfold higher than that for our sponge data. In our transcriptomes, 189 *L. baicalensis*(A8) but only 14 and 21 contigs in *L. abietina* (A10) and *B. bacillifera* (A2) respectively. As our genomic and transcriptomic sample for *L. baicalensis*are derived from the same tissue, we suspect that a member of the Chlorophyta was particularly abundant in that specimen, and could have been performing a symbiotic role, as has been reported previously in freshwater sponges (Sand-Jensen & Pedersen 1994).

We also note, as mentioned in the bacterial section above, species with similarity to *Entotheonella* (Wilson et al 2014) are also present. These findings, alongside those published previously, suggest the microbial symbiota of the sponges of Lake Baikal are diverse and

species dependent, with individual sponges taking advantage of the benefits of symbiotes in a situational manner which is ripe for future investigation.

Mitochondrial Genomes and Phylogeny of Sponges Studied

The mitochondrial genomes of several Lake Baikal sponges have been published previously (Lavrov 2010, Lavrov et al 2012, Maikova et al 2016). They display a number of special characteristics, notably the proliferation of small “hairpin” inverted repeats, which act as a source of variation. To confirm the identity of the sponges sequenced here, and to identify regions of difference at the DNA and coding level, we studied the sequence of the mitochondrial genomes of the three species investigated, and compared them to previously published work. Of the three species included in this manuscript, only *L. abietina* has not been sequenced previously. The *L. abietina* sequences have been uploaded to Genbank with accession numbers MH697685-MH697697.

Of the sequences presented here, that of *L. baicalensis* (both transcriptomic and genomic) seems closely related to that of the previously sequenced example of this species, while *B. bacillifera* differs somewhat from the example in Genbank, which appears more similar to *B. intermedia.* The small number of changes between these sequences, however, may render the differences inconsequential compared to morphological evidence (as suggested by low posterior probability of a real clade occurring between these sequences, which could be collapsed to a polytomy), and indels also add data, as discussed below. The phylogeny shown in Fig 5 seems to indicate paraphyly in both the *Lubomirskia* and *Baikalospongia*, albeit in some cases with limited posterior probability support. This mirrors the phylogeny previously observed in Lavrov et al 2012 (Fig 6 of that paper), adding extra taxa to the outline presented there. *Swartschewskia papyracea* and *Rezinkovia echinata* seem to render the *Lubomirskia* paraphyletic, while the *Baikalospongia* seem to be paraphyletic as a mixed paraphyletic sister group of this clade. From the evidence presented in this figure, taxonomic revision of these clades appears warranted.

The phylogeny shown in Fig 5, and particularly the branch lengths, does not fully show the level of divergence of the Lake Baikal clade, as gaps and indel regions have been excluded for the purpose of phylogenetic analysis. Several of these sites show clear diagnostic differences between the sponges of Lake Baikal and other freshwater species. One of these regions, that of *nad1* can be seen in Fig 5B, by way of example. This alignment clearly shows the disparity between freshwater sponges of Lake Baikal and those from further afield, but also (in position 826) shows how a diagnostic difference separates *L. baicalensis* and *R. echinata*

from the other species examined here. A fuller examination of the poriferan diversity of Lake Baikal, considering both morphological and genetic evidence, is therefore likely warranted to resolve these species into a taxonomically and systematically cohesive framework.

Nuclear gene-derived phylogeny

As part of the process of studying the signatures of selection within freshwater sponges, the Phylotreepruner output alignments of all 3,222 genes to be tested were concatenated, and used to construct a robust phylogeny for our species, spanning 3,242,264 amino acid residues although all gaps and indels were removed for phylogenetic inference. This phylogeny (and our tests of signatures of selection) included the three species of Lake Baikal examined here, alongside three other members of the Spongillida (*Spongilla lacustris, Eunapius fragilis* and *Ephydatia muelleri*) and five marine outgroup species.

Phylogenetic inference was performed by MrBayes using a by-gene partitioned GTR analysis, and the resulting phylogeny can be seen in Fig 6A. In this phylogeny, the freshwater sponges form a distinct clade corresponding to the Order Spongillida (Manconi and Pronzato, 2002), with maximal posterior probability support. *Mycale phyllophila*is noted as the outgroup to this clade, suggesting Poecilosclerids are more closely related to Spongillida than the other species represented here, albeit with weak (0.72) posterior probability. The three species of Haplosclerid examined form a monophyletic group with posterior probability of 1, and the sole member of the Dictyoceratida (and Keratosa) present in our sample, *Ircinia fasciculata*, is used to root the tree shown here. This is in agreement with the view of inter-relationships put forward in Morrow and Cárdenas 2015 (see Fig 2 of that work).

Within the Spongillida, *Lubomirskia* forms a monophyletic grouping, with *Baikalospongia*as the sister clade. This is in contrast to the mitochondrial tree, and likely represents the increased depth of data available to distinguish these clades. *E. muelleri* is strongly supported as the outgroup to the Lubomirskidae, with *E. fragilis* and *S.lacustris* placed in a monophyletic group as sister to the (Lubomirskidae + *E. muelleri*) clade. This mirrors the results of prior, individual and several gene/ITS based phylogenetic studies in freshwater sponges, e.g. (Meixner et al 2007, Erpenbeck et al 2011). As such, the monophyly of freshwater sponges, and the hypothesis of the independent evolution of endemic sponge species locally, from cosmopolitan founder species, seems secure (Meixner et al 2007).

Signatures of Molecular Adaptation to Freshwater Conditions

In order to discern how freshwater sponge species have evolved to cope with their environments, we have utilized the Hyphy and PAML software suites to identify genes where there are marked signals of selection in freshwater sponges when compared to the ancestral condition. The phylogeny of sponges used to conduct these tests, the concatenated nuclear gene phylogeny, can be seen in Fig 6A, and is discussed further above. In total, 3,222 single copy orthologous alignments remaining after pruning paralogous sequences with Phylotreepruner, henceforth referred to as “orthogroups” were tested from the 11 transcriptomic resources we used in this experiment.

We used the concilience of several tests to validate our findings, to avoid artifactual or spurious results. In particular, we used CodeML, Busted and aBSREL to test for signatures of selection using a combination of approaches (branch-site model, alignment-wide episodic diversifying selection, and adaptive branch-site random effects likelihood, respectively). These methods use slightly varying means to detect selection, so a combinatorial approach is most suitable for ensuring that identified selective events are well supported by evidence, rather than the product of branch-specific signals or errors that could yield false positives (Venkat et al 2018).

A total of 277, 324 and 252 orthogroups were identified as bearing signatures of selection by CodeML, Busted and aBSREL respectively. These represent 566 unique orthogroups, with 287 of them overlapping in some way, as can be seen in Fig 6B. The fact that 566 orthogroups, □ of those tested, show some sign of selection is suggestive of broad scale changes in sequence as a consequence of adaptation to a freshwater lifestyle. Of these, 61 orthogroups were identified by all 3 of our tests, and it was these that we concentrated on in further analyses.

Of these 61 sequences, only one was un-annotatable by BLAST identity to previously described genes. Annotation was aided by the inclusion of the most recent *A. queenslandica* resource in our test dataset, and the sole unannotatable orthogroup represents a newly identified sequence not originally annotated in the *A. queenslandica* genome, which nonetheless is present in all the sponge species studied here. These orthogroups, the details of which are given in Supplementary File 3, represent a variety of gene families.

We investigated the nature of the changes, and the exact sites under selection pressure, in these sequences by performing tests in MEME (Mixed Effects Model of Evolution, Murrell et al 2012, 2015), as well as by investigating the BEB (Bayes Empirical Bayes, Yang et al 2005) results from CodeML analysis. Of the 61 orthogroups identified as significant by all 3 tests described above, all but one possessed at least one amino acid site noted as under significant

evolutionary pressure by either our BEB or MEME test. The sole exception, OG0002186, is annotated to be *U3 small nucleolar ribonucleoprotein MPP10-like*. These results are also listed in Supplementary File 3, although please note that site number refers to that position in the alignment, and not an individual amino acid (as this will vary from species to species). This extra, almost perfect, support from site-based tests gives us even more confidence in the veracity of these results, and the strength of the selection pressures brought to bear on these genes.

The 61 orthogroups highlighted by this analysis correspond to transcription factors, structural proteins, membrane transport molecules and a variety of other gene families. This diversity reflects the wholesale nature of changes required by adaptation to a novel freshwater environment. To highlight some genes in particular, transporter genes such as *zinc transporter 2* (OG0000120) and *importin subunit alpha-6* (OG0001194), transcription factors such as *aristaless-related homeobox* (OG0006570), structural peptides such as *actin 3* (OG0000142) are all implicated.

However, despite this diversity, some genes have clear roles in homeostasis and membrane transport that would be under particular pressure as a result of the adaptation to a freshwater environment. Several of the genes identified by our analysis are transmembrane or membrane-associated proteins, such as the aforementioned transporters, *integrin alpha-9* (OG0001031), *neurobeachin* (OG0000141), *UNC93-like protein* (OG0005914) and t*weety homolog 2 (*OG0005907). Several are also involved in the Rho GTPase pathway, including *vav-1* (OG0000757), *rho GTPase-activating protein 39* (OG0001117) and *rho GTPase-activating protein 2* (OG0002195). Given the widespread roles of Rho GTPases in maintaining homeostasis and cellular functionality (Bustelo et al 2007) their modification to a freshwater environment under selective pressure would be necessary.

We examined our BEB results to see whether some amino acids were more likely to be the site of selection than others. Of the 424 sites identified by BEB across the 61 orthogroups, the most likely amino acid to exhibit positive selection is lysine (K), which occurs 38 times. Serine (S) occurred 33, leucine (L) 29, and aspartic acid (D) and alanine (A) occurred 28 times. These amino acids contain a variety of charged, polar and hydrophobic side chains, and thus these changes are likely to reflect a variety of changes within these molecules. In contrast, tryptophan (W) was the least likely to be positively selected, occurring only 4 times, likely a result of the large side chain which would prove difficult to incorporate into protein structures. Cysteine (C), 7, and phenylalanine (F) and tyrosine (Y), with only 10 occurrences, were similarly

less likely to be subject to positive selection, likely as a result of their tendency to form cross links and large side chains respectively.

These molecules and their roles will be studied in more detail in future research papers, and this work is underway. However, the large number of changes and the diversity of the genes affected reflect the profound changes which must have occurred across the genome of freshwater sponges in order to adapt to the markedly different osmotic, climatic and environmental differences posed by freshwater environments. As all freshwater sponge species examined here are descended from the same common ancestor, we are unable to use convergent signal to verify that these signals of selection are the result of freshwater specific cues. They could therefore derive from other selection pressures within this clade in the time since their stem lineage diverged from their sister taxa. However, as these freshwater sponges share these signatures of positive selection, they are likely to be necessary for their survival to the present day.

Previous “omic” studies looking into adaptation to freshwater environments have tended to focus on metazoans with complex body plans, for example fish (Mäkinen et al 2008) and prawns (Rahi et al 2017). The molecular adaptations of these species to less saline environments have often been found to be focused on genes expressed in gills and kidneys, as these are the organs most involved in maintaining homeostasis under the contrasting osmotic pressure found there. Sponges must utilise such adaptations across the entirety of their bodies, without the benefit of impermeable membranes, shells or skins. It is unsurprising that sponges exhibit profound changes across the majority of their molecular repertoires as a result of the pressures of freshwater environments. As sponges only populated freshwater on a single occasion, despite being widespread in marine environments, this transition cannot have been straightforward. Resources such as those described here will allow us to understand the demands of this process, and contrast them with those encountered by independent transition events to the same environment.

Conclusion:

The animals of Lake Baikal in general, and the sponges in particular, possess a number of unique biological features of interest to science. The advent of relatively low cost NGS techniques has at last allowed these to be understood from a molecular standpoint, providing a range of information for use in studies into the adaptation and population-level diversity of these species. Foundational research, such as that presented here, will allow this research to proceed with a firm empirical basis.

Here we have presented both data, in the form of three deep transcriptomic resources and a draft genome, to further investigate the biology of these fascinating creatures. Lake Baikal sponges have much to teach us about the myriad demands of freshwater adaptation and their unique habitat, and this research will be fruitful for many years to come.

Materials and Methods:

Sponge Collection, Nucleic Acid Extraction, Library Construction and Next Generation Sequencing

Sponge samples from healthy samples of the three species studied here were collected by SCUBA diving in September 2015 in the Southern basin of Lake Baikal near the village of Bolshie Koty (west coast, 51°53'54.4"N 105°04'15.3"E) at 10m depth. Samples of tissue for RNA extraction were placed in IntactRNA (Evrogen) immediately on return to the surface, incubated at 4° C overnight, before transfer to a -80° C freezer for long term storage. Samples for DNA extraction were placed in 100% ethanol and stored at -20° C. The *L. baicalensis* sample for genomic study was taken from the same individual as the mRNA sample.

RNA extraction was performed using a ExtractRNA kit (Evrogen).AnAgilent Technologies 2100 Bioanalyzer or 2200 TapeStation were used to establish that the RNA Integrity Number (RIN) value was greater than or equal to 8 for samples for sequencing.RNA libraries were constructed by Macrogen using the TruSeq RNA Sample Prep Kit v2. The HiSeq2000 platform (Illumina, USA) was used for paired-end sequencing.

DNA extraction was performed using a standard CTAB method. DNA quality was assessed by running a subsample on a 1% agarose gel and the quantity of DNA was measured using a NanoVue (GE Healthcare). DNA libraries were constructed by Macrogen using a TruSeq DNA PCR-free library kit (350bp insert size). An Illumina Hiseq X10 platform was used for sequencing, performed by Macrogen. Initial assessment of read quality and de-multiplexing of reads was performed by the provider according to their proprietary procedures. Paired-end reads were made available for download from their server, with no unpaired orphan reads retained by the process.

Initial Quality Assessment

The FastQC program (Andrews 2010) was used to perform initial quality assessment of all reads. The RNA reads required some cleaning, which was performed using Trimmomatic 0.33 (Bolger et al 2014) with the following settings: ILLUMINACLIP:../Adaptors.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:30. The Adaptors.fa file was adjusted to include the adaptor sequences specific to each read pair. Several problems were observed with DNA reads in particular, as detailed in the results. To remedy these, Trimmomatic was first run as described above, before rCorrector (Song and Florea 2015) was used to correct reads. Read pairs for which one of the reads was unfixable were removed using the

FilterUncorrectablePEfasta.py script (https://github.com/harvardinformatics/). FastQC was then re-run to confirm the quality of reads before assembly. Further, seqtkfqchk (https://github.com/lh3/seqtk) was used to ascertain additional basic metrics.

Assembly and Assessment

Assemblies of RNAseq data were performed using Trinity version 2013_08_14 (Grabherr et al 2011). Samples from *B. bacillifera* (A2) *L. abietina* (A10) and *L. baicalensis* (A8) were assembled separately. For our genomic data, a variety of assembly methods were trialed at a range of *k* mer sizes. Velvet 1.2.10 (Zerbino and Birney 2008), ABySS 2.0.2 (Simpson et al 2009), SOAPdenovo2 2.04 (Luo et al 2012) and SPAdes 3.9.1 (Bankevich et al 2012) were assayed as noted in the results. Default settings were used with the exception of a minimum contig size of 200 and a minimum coverage of 3 when possible. As these resources were to be used for the identification of symbiote and bacterial sequence, no cleaning was performed after assembly.

Numerical metrics relating to assembly were recovered using basic perl and python scripts, available from the authors on request. BUSCO v1.1b1 (Simao 2015) was used to assess the content of both the genome and transcriptomic assemblies, with the completeness of these measured relative to the eukaryotic and metazoan Basic Universal Single Copy Orthologue (BUSCO) cassettes. *K* mers were counted using Jellyfish (Marcais and Kingsford 2011) and GenomeScope (Verture et al 2017) was used to assay the genome size, heterozygosity and coverage of the *L. baicalensis* genome sample.

To assess the content of these resources further, Blobtools (Laetsch and Blaxter 2017) was run, using Diamond (Buchfink et al 2015) to run BLASTx searches for assignment of taxon identity, against the NCBI nr database. Bowtie2 (Langmead and Salzberg 2012) was used for read mapping. Blobtools was then run using the *de novo* genome pathway, and with mapping against a local copy of TaxID files.

Gene Identification and Annotation

For individual genes, tBLASTn (Altschul et al 1990) was used on a locally constructed database to putatively identify genes using sequences of known homology downloaded from Genbank. To further confirm identity, the sequences thus identified were reciprocally BLASTed (BLASTx) against the NCBI nr database online using BLASTx.

For total annotation, the longest ORF for each contig was translated using the getORF.py python script, retaining only the longest ORF, followed by BLASTP annotation

(Altschul et al 1990) against the nr protein database. These results were imported into Blast2GO Pro (Gotz et al 2008) followed by InterPro scanning, mapping, annotation (including ANNEX augmentation) and enzyme code mapping. The complete annotations for these resources are attached as Supplementary File 1.

Mitochondrial Sequence Identification, Annotation and Phylogenetic Analysis

TBLASTN (Altschul et al 1990) was used to identify mitochondrial sequences (-db_gencode 4) using sequences of known homology from other sponges. This recovered complete circular mitochondrial sequence after manual alignment and removal of extraneous sequences. Annotation was performed using the MITOS2 webserver (Bernt et al 2013), with the "04 - Mold/Protozoan/Coelenterate" setting used as the translational code for sponge data. Manual curation was performed to confirm start/stop codon identity, using homology to known genes to collaborate start/stop location. Mitochondrial phylogeny was inferred based on nucleotide sequence of all of these genes (both protein coding and rRNA), alongside those of known homology downloaded from Genbank. These were aligned on a gene-by-gene basis using MAFFT (Katoh et al 2002), and the G-INS-i approach. After concatenation with FASConCat-G (Kuck and Longo 2014) gaps and regions of poor alignment were excluded and phylogenetic inference was performed using partitioned Bayesian analysis in MrBayes (Ronquist and Huelsenbeck 2003), with models chosen using mixed models in the first instance. The first 25% of samples were discarded as 'burn-in', before the remaining samples were used to generate the figures shown here. Alignments are displayed in Geneious (Kearse et al 2012) for figure generation.

Selection Test

Transdecoder (Haas et al 2013), Orthofinder (Emms and Kelly 2015, MCL and MAFFT options) and Phylotreepruner (Kocot et al 2013) were run sequentially to identify a dataset of orthologous gene alignments. Final alignments, after pruning, are represented by only one sequence per species (i.e. no paralogous sequence). Concatenated amino acid sequence from these alignments were used as the basis for partitioned Bayesian analysis in MrBayes (Ronquist and Huelsenbeck 2003) to determine a species tree for selection tests, with a GTR+4G model applied to each gene in an independent partition. PAL2NAL (Suyama et al 2006) was then run orthogroup-by-orthogroup to find the CDS region corresponding to the aligned protein sequence, and generate a nucleotide alignment for performing selection tests.

CODEML was run in PAML (Muse and Gaut 1994, Yang 2007) to test null vs alternative hypotheses as to gene-level selection, and differences in LnL used as the basis for χ^2 tests of level of significance. These were corrected for multiple comparison FDR using Benjamini and Hochberg correction (Benjamini and Hochberg 1995, Benjamini and Yekutieli 2001). Bayes Empirical Bayes (BEB) values in the ALT output were also extracted and used to identify sites under selection (Yang et al 2005).

HyPHy was run, with BUSTED, aBSREL and MEME tests of branch-level and site-level selection respectively (Pond et al 2005, Murrell et al 2012, Murrell et al 2015, Smith et al 2015). Sequences were annotated by BLASTP identity to the nr database.

Acknowledgements:

The authors thank the members of their laboratories for their many helpful discussions and support. We thank editors and reviewers for their time and thoughtful contributions towards this manuscript. This work was supported bythe ADAPTOMICS MSCA [IF750937] with a grant to NJK under the Horizon 2020 program. This study was done in the framework of the Basic project VI.50.1.4 (0345-2016-0002) by Federal Agency for Scientific Organizations and partially supported by RFBR grants No 17-04-01598, 17-44-388103.

Data Note:

Raw reads are available from the NCBI SRA under accession PRJNA431612. Assemblies, including preliminary and alternate forms, have been uploaded to Figshare, with DOI: and URL: 10.6084/m9.figshare.6819812, https://figshare.com/s/fe36239c32bbf7342756.

Contribution of Authors:

VBI conceived of the study, gathered specimens and performed RNA and DNA extractions. NJK, BP, AR and VBI designed and performed experiments. BP performed mitochondrial data analysis. NJK performed assembly and bioinformatic analysis. All authors wrote and agreed on the final form of the manuscript.

References:

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., 1990. Basic local alignment search tool. *Journal of molecular biology*, *215*(3), pp.403-410.

Andrews S. FastQC: A quality control tool for high throughput sequence data [Internet]. 2010.http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Annenkova NV, Lavrov DV, Belikov SI. Dinoflagellates associated with freshwater sponges from the ancient Lake Baikal. *Protist*. 2011 Apr 1;162(2):222-36.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D. and Pyshkin, A.V., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, *19*(5), pp.455-477.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological).* 1995 Jan 1:289-300.

Benjamini, Y. and Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pp.1165-1188.

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., ... & Stadler, P. F. (2013). MITOS: Improved de novo metazoan mitochondrial genome annotation. *Molecular phylogenetics and evolution*, *69*(2), 313-319.

Bielawski JP, Baker JL, Mingrone J. 2016. Inference of Episodic Changes in Natural Selection Acting on Protein Coding Sequences via CODEML. *Curr Protoc Bioinformatics* 54:6.15.1–6.15.32.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*.; 2014;30: 2114–2120.

Buchfink, B., Xie, C. and Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND. *Nature methods*, *12*(1), p.59.

Bustelo, X.R., Sauzeau, V. and Berenjeno, I.M., 2007. GTP□binding proteins of the Rho/Rac family: regulation, effectors and functions in vivo. *Bioessays*, *29*(4), pp.356-370.

Chernogor, L., Denikina, N., Kondratov, I., Solovarov, I., Khanaev, I., Belikov, S. and Ehrlich, H., 2013. Isolation and identification of the microalgal symbiont from primmorphs of the endemic freshwater sponge *Lubomirskia baicalensis* (Lubomirskiidae, Porifera*). European journal of phycology*, 48(4), pp.497-508.

Ciesielski, T.M., Pastukhov, M.V., Leeves, S.A., Farkas, J., Lierhagen, S., Poletaeva, V.I. and Jenssen, B.M., 2016. Differential bioaccumulation of potentially toxic elements in benthic and pelagic food chains in Lake Baikal. *Environmental Science and Pollution Research*, *23*(15), pp.15593-15604.

Darriba, D., Taboada, G.L., Doallo, R. and Posada, D., 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, *27*(8), pp.1164-1165.

Denikina, N.N., Dzyuba, E.V., Bel'kova, N.L., Khanaev, I.V., Feranchuk, S.I., Makarov, M.M., Granin, N.G. and Belikov, S.I., 2016. The first case of disease of the sponge *Lubomirskiabaicalensis:* Investigation of its microbiome. *Biology Bulletin*, *43*(3), pp.263-270.

Emms, D.M. and Kelly, S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16:157

Erpenbeck D, Weier T, De Voogd NJ, Wörheide G, Sutcliffe P, Todd JA, Michel E. Insights into the evolution of freshwater sponges (Porifera: Demospongiae: Spongillina): barcoding and phylogenetic data from Lake Tanganyika endemics indicate multiple invasions and unsettle existing taxonomy. *Molecular Phylogenetics and Evolution*. 2011 Oct 1;61(1):231-6.

Excoffier, L. and Lischer, H.E., 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources*, *10*(3), pp.564-567.

Fernandez-Valverde, S.L., Calcino, A.D. and Degnan, B.M., 2015. Deep developmental transcriptome sequencing uncovers numerous new genes and enhances gene annotation in the sponge *Amphimedon queenslandica*. *BMC genomics*, *16*(1), p.387.

Francis, W.R., Eitel, M., Vargas, S., Adamski, M., Haddock, S.H., Krebs, S., Blum, H., Erpenbeck, D. and Wörheide, G., 2017. The Genome Of The Contractile Demosponge *Tethya wilhelma* And The Evolution Of Metazoan Neural Signalling Pathways. *bioRxiv*, p.120998.

Goodwin, S., McPherson, J.D. and McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), pp.333-351.

Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11:725-736.

Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic acids research*. 2008 Apr 29;36(10):3420-35.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* United States; 2011;29: 644–652.

Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic acids research*. 2010 Jul 1;38(12):e131.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols*. 2013 Aug;8(8):1494.

Itskovich, V.B., Kaluzhnaya, O.V., Veynberg, E. and Erpenbeck, D., 2015. Endemic Lake Baikal sponges from deep water. 1: Potential cryptic speciation and discovery of living species known only from fossils. *Zootaxa*, *3990*(1), pp.123-137.

Itskovich, V.B., Shigarova, A.M., Glyzina, O.Y., Kaluzhnaya, O.V., Borovskii, G.B., 2018. Heatshock protein 70 (Hsp70) response to elevated temperatures in the endemic Baikal sponge *Lubomirskia baicalensis*. *Ecological Indicators*, 88, pp. 1-7.

Jeffery, N.W., C.B. Jardine, and T.R. Gregory (2013). A first exploration of genome size diversity in sponges. *Genome* **56**: 451-456

Kaluzhnaya, O.V., Itskovich, V.B. and McCormack, G.P., 2011. Phylogenetic diversity of bacteria associated with the endemic freshwater sponge *Lubomirskia baicalensis*. *World Journal of Microbiology and Biotechnology*, 27(8), pp.1955-1959.

Kaluzhnaya, O.V., Krivich, A.A., and Itskovich, V.B., 2012. Diversity of 16S rRNA genes in metagenomic community of the freshwater sponge *Lubomirskia baicalensis*. *Russian journal of genetics*, *48*(8), pp.855-858.

Kaluzhnaya, O.V. and Itskovich, V.B., 2015. Bleaching of Baikalian sponge affects the taxonomic composition of symbiotic microorganisms. *Russian journal of genetics*, *51*(11), pp.1153-1157.

Kaluzhnaya, O.V. and Itskovich, V.B., 2017. Molecular identification of filamentous cyanobacteria overgrowing the endemic sponge Lubomirskia baicalensis. *Inland Waters*, *7*(3), pp.267-271.

Kasimov, N., Karthe, D. and Chalov, S., 2017. Environmental change in the Selenga River—Lake Baikal Basin. *Regional Environmental Change* 17(7):1945–1949 .

Katoh, K., Misawa, K., Kuma, K.I. and Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, *30*(14), pp.3059-3066.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Mentjies, P., & Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data.*Bioinformatics*, *28*(12), 1647-1649.

Kenny, N.J., de Goeij, J.M., de Bakker, D.M., Whalen, C.G., Berezikov, E. and Riesgo, A., 2018. Towards the identification of ancestrally shared regenerative mechanisms across the Metazoa: A Transcriptomic case study in the Demosponge *Halisarca caerulea*. *Marine Genomics*. S1874-7787(17) 30232-5

Khanaev, I.V., Kravtsova, L.S., Maikova, O.O., Bukshuk, N.A., Sakirko, M.V., Kulakova, N.V., Butina, T.V., Nebesnykh, I.A. and Belikov, S.I., 2017. Current state of the sponge fauna (Porifera: Lubomirskiidae) of Lake Baikal: Sponge disease and the problem of conservation of diversity. *Journal of Great Lakes Research*.

Kocot KM, Citarella MR, Moroz LL, Halanych KM. PhyloTreePruner: a phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. *Evolutionary Bioinformatics.* 2013 Jan;9:EBO-S12813.

Kozhov, M., 1963. Lake Baikal and its life, vol. XI of. *Monographia Biologicae*. Dr. W. Junk Publishers, The Hague

KozhovaO. M. &Izmest'evaL. R. (Eds) Lake Baikal – Evolution and Biodiversity (1998). Backhuys Publishers, Leiden.

Kulakova, N.V., Sakirko, M.V., Adelshin, R.V., Khanaev, I.V., Nebesnykh, I.A. and Pérez, T., 2017. Brown Rot Syndrome and Changes in the Bacterial Community of the Baikal Sponge *Lubomirskia baicalensis*. *Microbial ecology*, pp.1-11.

Kück P, Longo GC. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Frontiers in zoology.* 2014 Dec;11(1):81.

Laetsch, D.R. and Blaxter, M.L., 2017. BlobTools: Interrogation of genome assemblies. *F1000Research*, *6*.

Langmead, B. and Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), pp.357-359.

Lavrov, D.V., Maikova, O.O., Pett, W. and Belikov, S.I., 2012. Small inverted repeats drive mitochondrial genome evolution in Lake Baikal sponges. *Gene*, *505*(1), pp.91-99.

Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*.; 2011;12: 323.

Liu F, Li J, Feng G, Li Z. New genomic insights into “Entotheonella” symbionts in *Theonellas winhoei*: Mixotrophy, anaerobic adaptation, resilience, and interaction. *Frontiers in microbiology.* 2016 Aug 25;7:1333.

Lohse, M., Drechsel, O., & Bock, R. (2007). OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Current genetics*, *52*(5-6), 267-274.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y. and Tang, J., 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, *1*(1), p.18.

Mäkinen, H.S., Cano, J.M. and Merilä, J., 2008. Identifying footprints of directional and balancing selection in marine and freshwater three□spined stickleback (*Gasterosteus aculeatus*) populations. *Molecular Ecology*, *17*(15), pp.3565-3582.

Marçais, G. and Kingsford, C., 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), pp.764-770.

Manconi, R.; Pronzato, R. (2002). Suborder Spongillina subord. nov.: Freshwater sponges. Pp. 921-1020. In Hooper, J. N. A. & Van Soest, R. W. M. (ed.) Systema Porifera. A guide to the classification of sponges. 1 (Kluwer Academic/ Plenum Publishers: New York, Boston, Dordrecht, London, Moscow).

Meixner, M.J., Lüter, C., Eckert, C., Itskovich, V., Janussen, D., von Rintelen, T., Bohne, A.V., Meixner, J.M. and Hess, W.R., 2007. Phylogenetic analysis of freshwater sponges provide evidence for endemism and radiation in ancient lakes. *Molecular phylogenetics and evolution*, *45*(3), pp.875-886.

Morrow C, Cárdenas P. Proposal for a revised classification of the Demospongiae (Porifera). *Frontiers in Zoology.* 2015 Dec;12(1):7.

Müller, W.E., Belikov, S.I., Kaluzhnaya, O.V., Perović□Ottstadt, S., Fattorusso, E., Ushijima, H., Krasko, A. and Schröder, H.C., 2007. Cold stress defense in the freshwater sponge *Lubomirskia baicalensis*. *The FEBS journal*, *274*(1), pp.23-36.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL (2012) Detecting Individual Sites Subject to Episodic Diversifying Selection. *PLoS Genet* 8(7): e1002764. https://doi.org/10.1371/journal.pgen.1002764

Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, Eren K, Pollner T, Martin DP, Smith DM, Scheffler K, Kosakovsky Pond SL. 2015. Gene-Wide Identification of Episodic Selection. *Molecular Biology and Evolution* 32:1365–1371.

Muse, S.V. and Gaut, B.S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* 11:715-724.

Narum, S.R., 2006. Beyond Bonferroni: less conservative analyses for conservation genetics. *Conservation Genetics*, *7*(5), pp.783-787.

Naumenko, S.A., Logacheva, M.D., Popova, N.V., Klepikova, A.V., Penin, A.A., Bazykin, G.A., Etingova, A.E., Mugue, N.S., Kondrashov, A.S. and Yampolsky, L.Y., 2017. Transcriptome-based phylogeny of endemic Lake Baikal amphipod species flock: fast speciation accompanied by frequent episodes of positive selection. *Molecular ecology*, *26*(2), pp.536-553.

Pile, A.J., Patterson, M.R., Savarese, M., et al., Trophic effects of sponge feeding within Lake Baikal's littoral zone: 2. Sponge abundance, diet, feeding efficiency, and carbon flux, *Limnol. Oceonogr.*, 1997, vol. 42, no. 1, pp. 178–184.

Pomazkina, G., Kravtsova, L. and Sorokovikova, E., 2012. Structure of epiphyton communities on Lake Baikal submerged macrophytes. *Limnological Review*, *12*(1), pp.19-27.

Pond SLK, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21:676–679.

Rahi ML, Amin S, Mather PB, Hurwood DA. (2017) Candidate genes that have facilitated freshwater adaptation by palaemonid prawns in the genus *Macrobrachium*: identification and expression validation in a model species (*M. koombooloomba*) *PeerJ* 5:e2977

Rice, P., Longden, I. and Bleasby, A., 2000. EMBOSS: the European molecular biology open software suite.

Riesgo, A., Farrar, N., Windsor, P.J., Giribet, G. and Leys, S.P., 2014. The analysis of eight transcriptomes from all poriferan classes reveals surprising genetic complexity in sponges. *Molecular biology and evolution*, *31*(5), pp.1102-1120.

Rivarola□Duarte, L., Otto, C., Jühling, F., Schreiber, S., Bedulina, D., Jakob, L., Gurkov, A., Axenov□Gribanov, D., Sahyoun, A.H., Lucassen, M. and Hackermüller, J., 2014. A first glimpse at the genome of the Baikalian amphipod Eulimnogammarus verrucosus. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, *322*(3), pp.177-189.

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010 Jan 1;26(1):139-40.

Romanova, E.V., Kravtsova, L.S., Izhboldina, L.A., Khanaev, I.V. and Sherbakov, D.Y., 2015. Identification of filamentous green algae from an area of local biogenic pollution of Lake Baikal (Listvennichnyi Bay) using SSU 18S rDNA. *Russian Journal of Genetics: Applied Research*, *5*(2), pp.118-125.

Romanova, E.V., Aleoshin, V.V., Kamaltynov, R.M., Mikhailov, K.V., Logacheva, M.D., Sirotinina, E.A., Gornov, A.Y., Anikin, A.S. and Sherbakov, D.Y., 2016. Evolution of mitochondrial genomes in Baikalian amphipods. *BMC genomics*, *17*(14), p.1016.

Ronquist, F. and Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, *19*(12), pp.1572-1574.

Rusinek OT, Takhteev VV, Gladkochub DP, Khodzher TV, Budnev NM, editors. 2012a. Baicalogy. Book 1. Novosibirsk: Nauka

Rusinek OT, Takhteev VV, Gladkochub DP, Khodzher TV, Budnev NM, editor. 2012b. Baicalogy. Book 2. Novosibirsk: Nauka.

Sand-Jensen K, Pedersen MF. Photosynthesis by symbiotic algae in the freshwater sponge*, Spongilla lacustris. Limnology and Oceanography*. 1994 May 1;39(3):551-61.

Schröder, H.C., Efremova, S.M., Itskovich, V.B., Belikov, S., Masuda, Y., Krasko, A., Müller, I.M. and Müller, W.E.G., 2003. Molecular phylogeny of the freshwater sponges in Lake Baikal. *Journal of Zoological Systematics and Evolutionary Research*, *41*(2), pp.80-86.

Shimaraev, M.N., Verbolov, V.I., Granin, N.G., Sherstyankin, P.P., 1994. Physical Limnology of Lake Baikal: A Review. BICER Publishers, Irkutsk and Okayama.

Simao FA, Waterhouse RM, Ioannidis P, Kriventseva E V, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*.; 2015;31: 3210–3212.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J. and Birol, I., 2009. ABySS: a parallel assembler for short read sequence data. *Genome research*, *19*(6), pp.1117-1123.

Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. 2015. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Molecular Biology and Evolution* 32:1342–1353.

Song, L. and Florea, L., 2015. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience*, *4*(1), p.48.

Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M.E., Mitros, T., Richards, G.S., Conaco, C., Dacre, M., Hellsten, U. and Larroux, C., 2010. The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature*, *466*(7307), pp.720-726.

Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research.* 2006 Jul 1;34(suppl_2):W609-12.

Thomas GWC, Hahn MW, Hahn Y. 2017. The effects of increasing the number of taxa on inferences of molecular convergence. *Genome Biology and Evolution* 9(1):213–221.

Timoshkin OA. 2001. Lake Baikal: fauna diversity, problems of its immiscibility and origin, ecology and exotic communities. In: Timoshkin OA, editor. Index of animal species inhabiting Lake Baikal and its catchment area. Novosibirsk: Nauka Publishers. p 16–73.

Timofeyev MA. 2010. Ecological and physiological aspects of adaptation to abiotic environmental factors in endemic Baikal and Palearctic amphipods. Tomsk: Tomsk State University.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M. and Rozen, S.G., 2012. Primer3—new capabilities and interfaces. *Nucleic acids research*, *40*(15), pp.e115-e115.

Venkat A, Hahn MW, Thornton JW. Multinucleotide mutations cause false inferences of lineage-specific positive selection. *Nature ecology & evolution.* 2018 Jul 2:1.

Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J. and Schatz, M.C., 2017. GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics*, p.btx153.

Wilson, M.C., Mori, T., Rückert, C., Uria, A.R., Helf, M.J., Takada, K., Gernert, C., Steffens, U.A., Heycke, N., Schmitt, S. and Rinke, C., 2014. An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature*, *506*(7486), p.58.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* 24:1586–1591.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution* 15:496–503.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution* 22:1107–1118.

Zerbino, D.R. and Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, *18*(5), pp.821-829.

Zhang J. 2005. Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. *Molecular Biology and Evolution* 22:2472–2479.

Figures and Legends

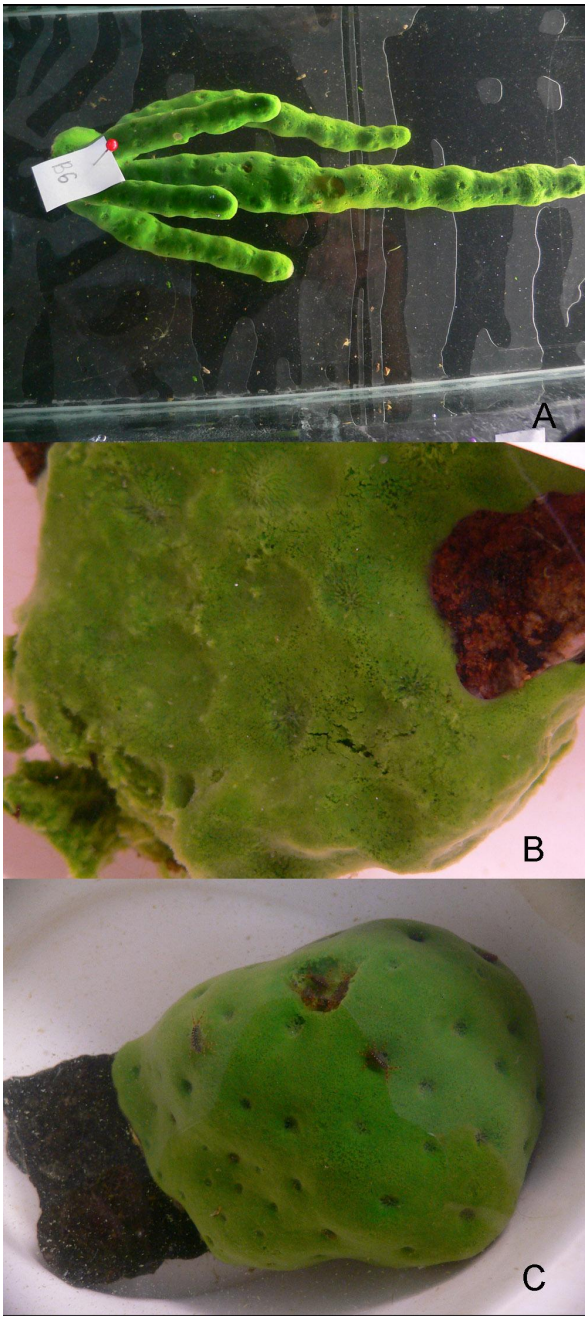

Figure 1: Specimens used for RNA and gDNA extraction. A) *Lubomirskia baicalensis*. B) *Lubomirskia abietina*. C) *Baikalospongia bacillifera*.

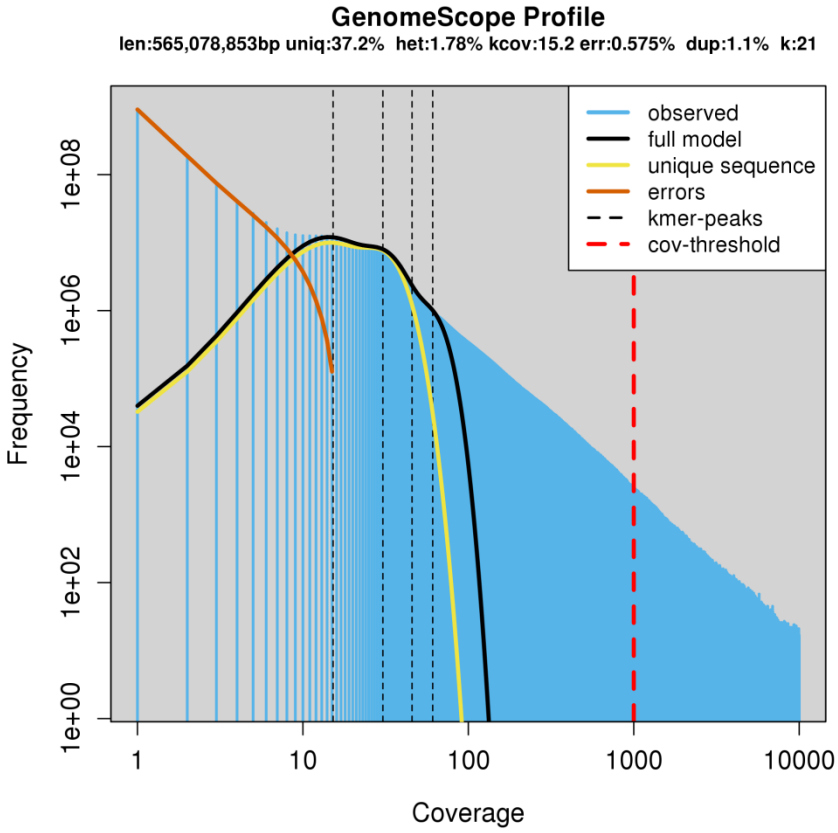

Figure 2: GenomeScope profile of *k*-mer distribution, based on Jellyfish *k*-mer counts of untrimmed and uncorrected reads. Note peak of *k*-mer coverage (15.2) and inferred genome size of 565,078,853 bp.

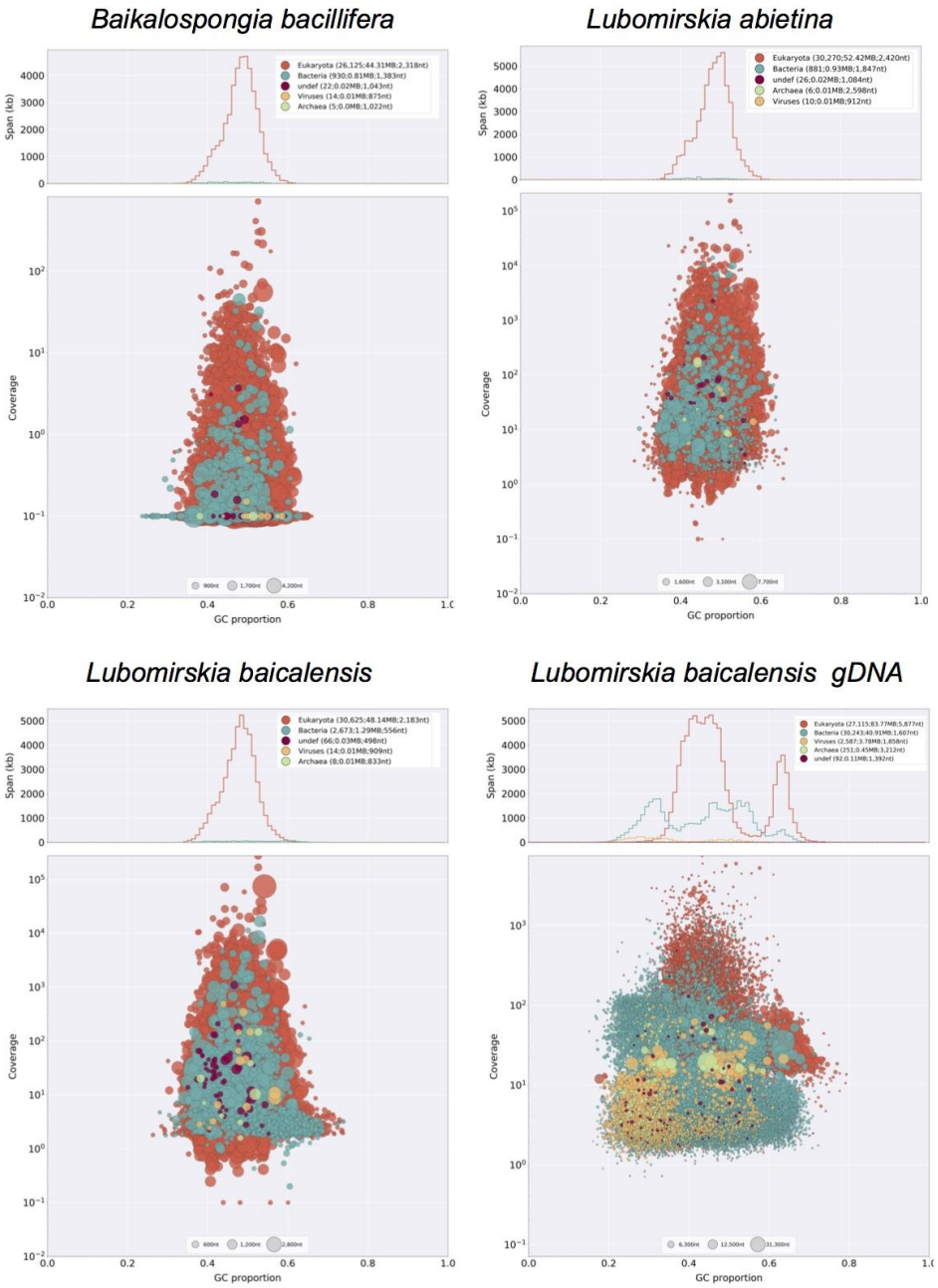

Figure 3: Blobplot results, showing distribution of annotated contigs according to GC content (x axis) and coverage (y axis) for the resources presented here. Summary statistics are also provided at the top of each panel. Note the bimodal distribution of GC content in the genome, which may represent differences between the coding and non-coding elements of the genome as observed in other species. .

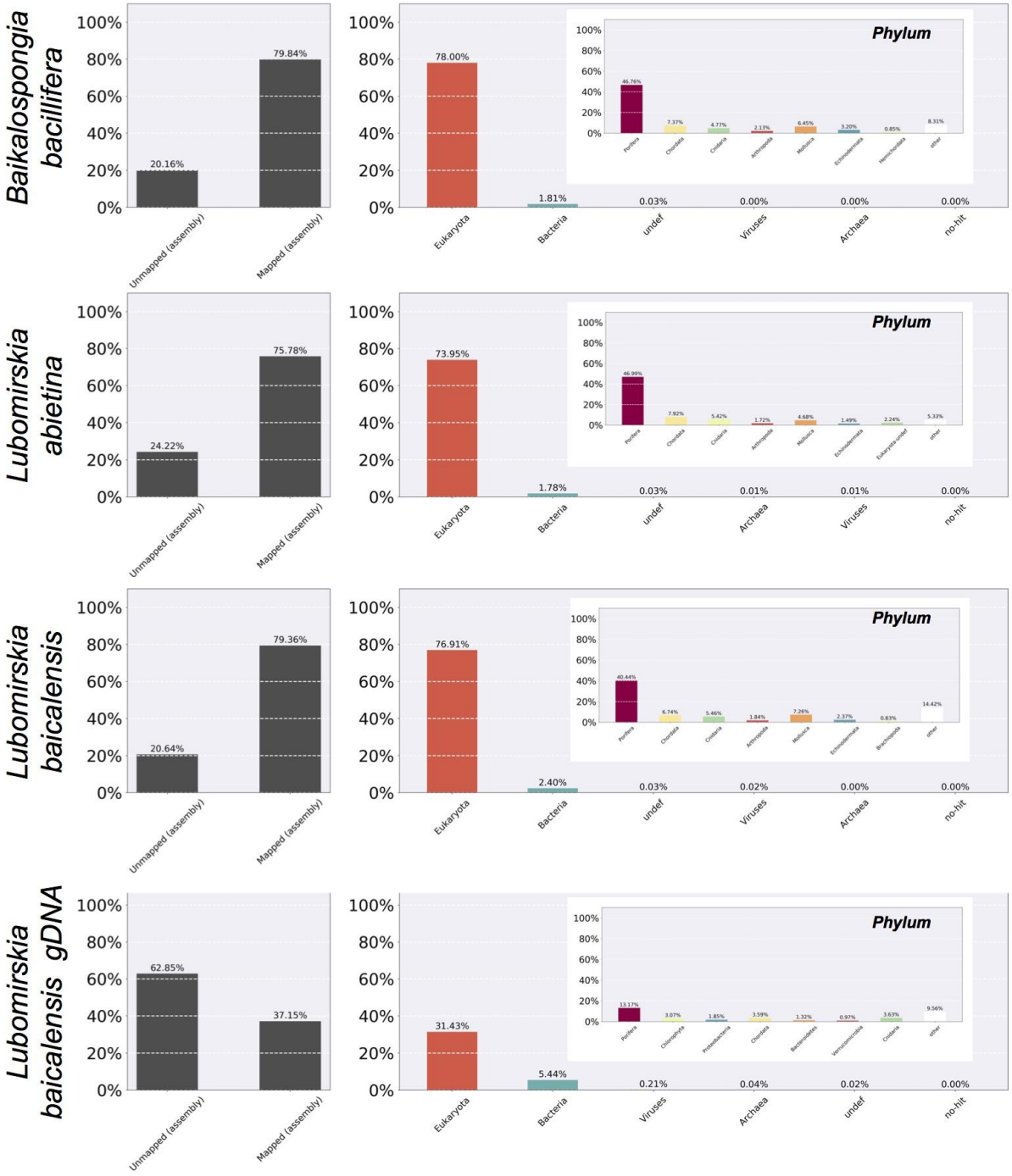

Figure 4: Mapping results for transcriptomic and genome assemblies vs. the nr database. %ages represent proportion of base pairs. At left: overall %age of contigs mapped to a known sequence. At right, Superkingdom and (inset) Phylum of these hits. Top 6 (Superkingdom)/Top 7 (Phylum) results shown.

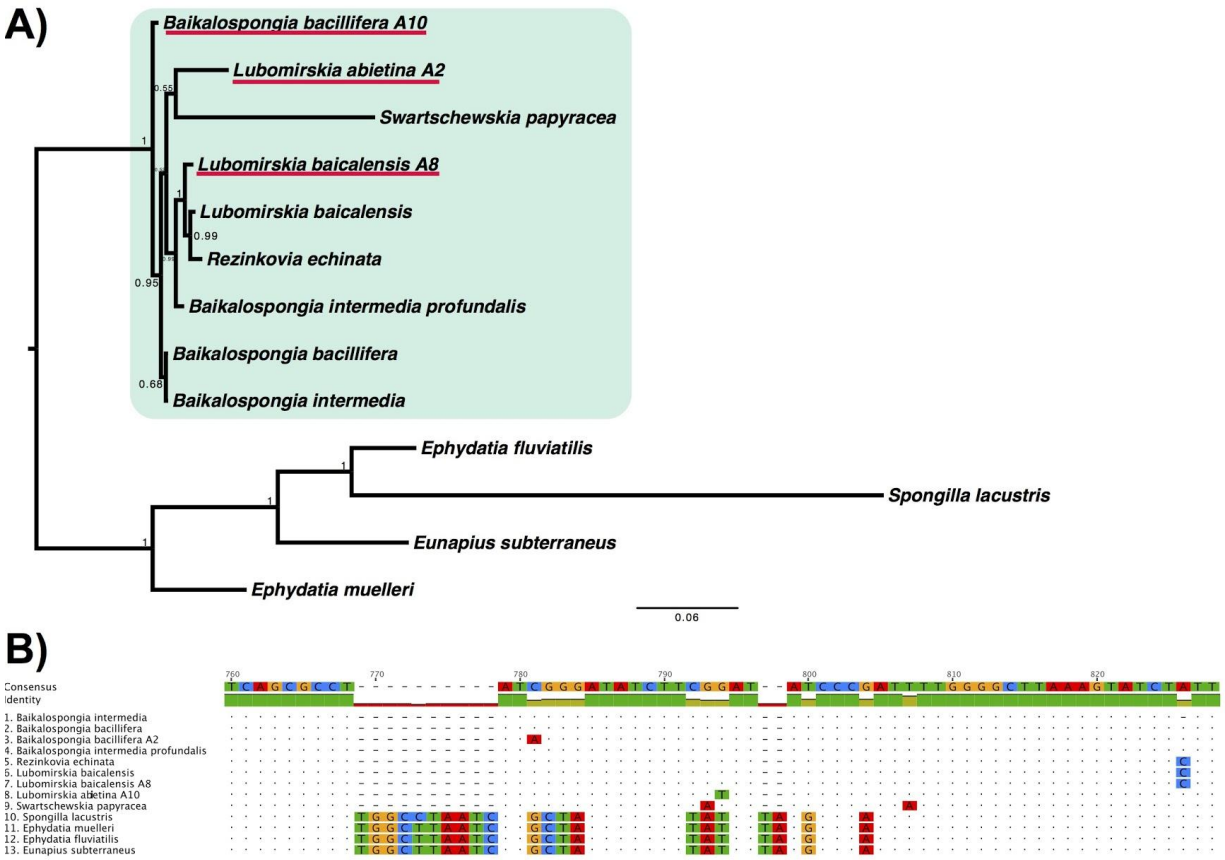

Figure 5: A) Phylogenetic relationships of a variety of freshwater sponge species, inferred using Bayesian methods, based on alignment of nucleotide sequence from mitochondrial protein coding and rRNA genes. Numbers at bases of nodes indicate posterior probability, number at base of tree (under scale bar) indicates number of changes per site at unit length. B) example of indels not used in tree, but nonetheless present in alignment. These indels (example from *Nad1* provided) provide further data that can be used to support clades, but are not represented in the phylogeny shown (as any site without sequence data is excluded from that analysis). Sponges endemic to Lake Baikal are, on the basis of this evidence, firmly supported as a monophyletic clade, to the exclusion of other freshwater sponges.

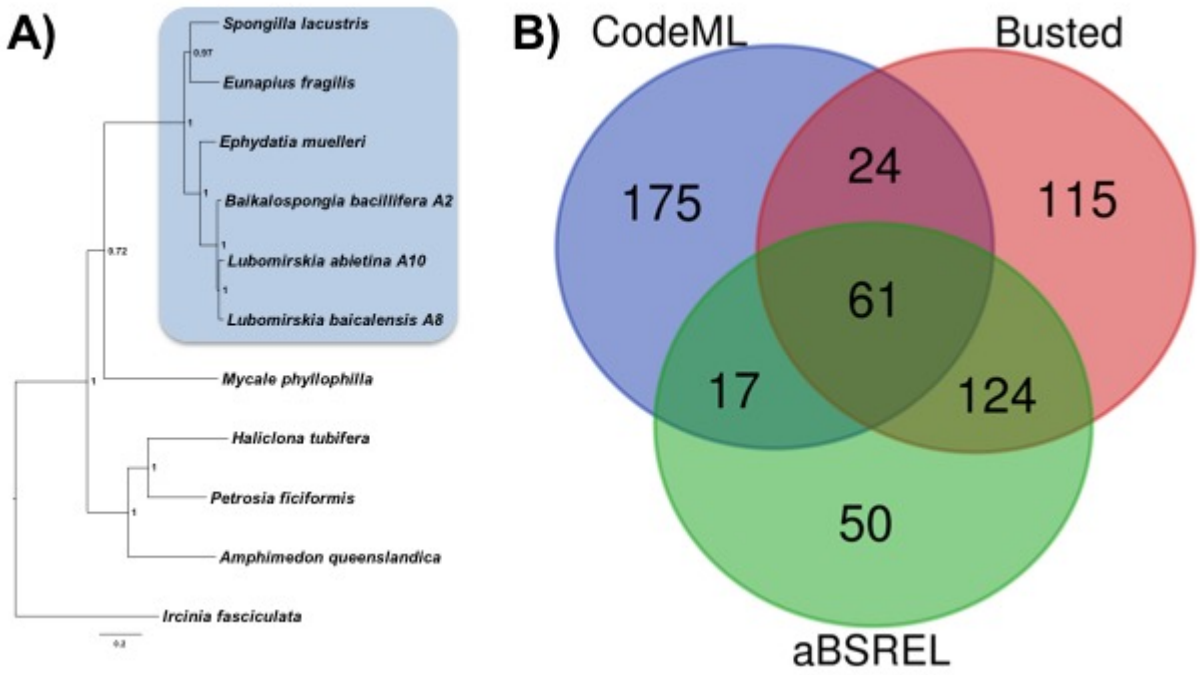

Figure 6: A) Phylogenetic relationship of freshwater sponge species and outgroup taxa, inferred from a concatenated multiple gene alignment using partitioned GTR analysis in MrBayes (3,222 genes, 3,242,264 site alignment, before trimming). This tree was used as the basis for tests of selection on the freshwater lineage, the results of which are summarised in: B) Orthogroups displaying significant results for tests of selection under several tests, shown in a Venn diagram indicating consilience of results. Note that further site-level (MEME, BEB) results are available in Supplementary File 3.

Tables

Prior to Cleaning	***Baikalospongia bacillifera* (A2)**	***Lubomirskia abietina* (A10)**	***Lubomirskia baicalensis* (A8)**	***Lubomirskia baicalensis* gDNA Reads**
Number of Read Pairs	50,694,150	52,345,337	54,439,423	184,491,682
Read Length	101	101	101	151
GC%	49	49	49	43.5
Average Quality	34.4	34.39	33.74	37.5
Total bases	10,240,218,300	10,573,758,074	10,996,763,446	55,716,487,964
After Cleaning	**A2**	**A10**	**A8**	**gDNA Reads**
Number of Read Pairs	48,445,101	50,007,487	32,204,065	81,042,203
GC% (Paired)	49	49	49	41
Average Quality (Paired)	36.25	36.25	36.15	39.75
Total Bases (Paired)	9,518,960,084	9,822,304,994	6,321,981,078	19,426,404,098
Unpaired Reads	2129241	2214127	1381971	n/a
GC% (Unpaired)	50	50	50	n/a
Average Quality (Unpaired)	33	33	32.9	n/a
Total bases (Unpaired)	183279397	190482063	117844996	n/a

Table 1: Metrics relating to reads, before and after read cleaning.

	***Baikalospongia bacillifera* (A2)**	***Lubomirskia abietina* (A10)**	***Lubomirskia baicalensis* (A8)**
Number of Trinity Transcripts	80,925	93,404	81,951
Number of Trinity 'Genes'	54,606	62,809	54,913
Min contig length:	201	201	201
Max contig length:	16,639	30,430	11,157
Mean contig length:	850.38	849.74	854.37
N50 contig length:	1,595	1,628	1,572
Number of contigs >=1kb:	20,558	22,946	21,595
Number of contigs in N50:	12,300	13,341	12,943
Number of bases in all contigs:	68,817,041	79,368,987	70,016,550
Number of bases in contigs >=1kb:	44,913,440	52,015,444	45,916,970
GC Content of contigs: (%)	46.75	46.62	46.78

Table 2: Statistics relating to transcriptome assemblies

	SPAdes, 500 min	SPAdes, 200 min	ABySS, 61mer	SOAPdenovo, 61mer	Velvet, 61mer
Min contig length:	500	200	200	200	200
Max contig length:	124,926	124,926	216,201	99,521	21,493
Mean contig length:	1553.28	681.89	729.9	544.51	464.06
Median contig length:	845	351	294	324	292
N50 contig length:	2,213	1,019	1,931	661	544
Number of contigs:	135,191	451,479	235,631	579,486	357,804
Number of contigs >=1kb:	54,728	54,728	22,065	49,540	28,260
Number of contigs in N50:	19,573	53,387	14,759	95,954	72,346
Number of bases in all contigs:	209,989,122	307,857,163	171,988,112	315,536,346	166,041,199
Number of bases in contigs >=1kb:	155,282,664	155,282,664	96,787,949	120,598,815	51,363,918
GC Content of contigs: (%)	43.68	42.64	43.85	40.94	44.39

Table 3: Genome assembly results for draft *L. baicalensis* resource using a variety of programmes. SPAdes (with a minimum size cutoff of 500 bp) used for further analyses, but other assemblies also available for download.

property	min	max
Heterozygosity	1.75375%	1.80378%
Genome Haploid Length	558,344,824 bp	565,078,853 bp
Genome Repeat Length	350,461,528 bp	354,688,339 bp
Genome Unique Length	207,883,296 bp	210,390,514 bp
Model Fit	85.7458%	98.2203%
Read Error Rate	0.574939%	0.574939%

Table 4: Genome metrics for *L. baicalensis,* computed using Genoscope, using Jellyfish-derived *k* mer counts (size = 21 bp).

Eukaryote Cassette:	***Baikalospongia bacillifera* (A2)**	***Lubomirskia abietina* (A10)**	***Lubomirskia baicalensis* (A8)**	**SPAdes Assembly**
Complete BUSCOs	299	297	291	235
-Complete (Single Copy)	192	196	178	204
-Complete (Duplicated)	107	101	113	31
Fragmentary BUSCOs	2	2	9	51
Missing BUSCOs	2	4	3	17
Total BUSCO genes	303	303	303	303
***Metazoan Cassette*:**	**A2**	**A10**	**A8**	**SPAdes Assembly**
Complete BUSCOs	918	914	904	562
-Complete (Single Copy)	555	556	515	493
-Complete (Duplicated)	363	358	389	69
Fragmentary BUSCOs	19	18	31	247
Missing BUSCOs	41	46	43	169
Total BUSCO genes	978	978	978	978

Table 5: BUSCO results for genome and transcriptome assemblies.

Supplementary Files:

Supplementary File 1: Annotations for transcriptomes (.annot, zipped).

*Supplementary File 2:*Details of Blobplot taxonomic binning (.xlsx files, zipped).

Supplementary File 3: Details of selection tests (.xlsx files, zipped).