nature portfolio

https://doi.org/10.1038/s41586-026-10828-6

Supplementary information

Brown bullhead catfish melanoma represents a novel transmissible cancer

In the format provided by the authors and unedited

Table of Contents

Section 1. RNA preparation, sequencing, and Sequence Availability

We included an additional 18 tissue samples (Supplementary Table 3) collected in July of 2015 from Lake Memphremagog in Hospital Cove (HC, n=5 N and n=9 T; Blazer et al. 2020) and Ticklenaked Pond (TP, n=4 N). These samples were interrogated for transcriptomics. Tissues collected for RNA work were placed in RNAlater® (ThermoFisher Scientific, Waltham, MA) and stored frozen. RNA extraction from skin tissue was done at EESC as described in Iwanowicz et al. (Iwanowicz et al., 2022).

Total RNA samples, from the 2015 fish, were quantified using an Agilent Bioanalyzer 2100 (Agilent, Santa Clara, CA) followed by poly-A enrichment and sequence library construction using the Illumina TruSeq stranded mRNA kit (Illumina, San Diego, CA). Libraries were paired end sequenced (2 x 150 bp) on a Illumina HiSeq 4000 instrument at the University of Maryland School of Medicine, Institute for Genome Sciences.

Raw short read RNA sequence data for the 2015 Ticklenaked Pond and Lake Memphremagog samples are available on NCBI (BioProjects PRJNA1284832 and PRJNA1284812; SRA Accessions: SRR34320463 - SRR34320479, SRR34328784 - SRR36648244, and PV874414.1 and GCA_051624135.1).

Section 2. DNA Sequencing Results and Sequence Availability

2.1 DNA Short Read Sequencing Summary

We generated over 5.1 billion paired end reads for the 8 reference samples collected in 2022 from Lake Champlain, the Connecticut River, and Lake Bomoseen, the 2024 Hermon Pond, Maine samples, and the 2025 Village Pond, New Hampshire samples (Supplementary Tables 1-2 and 4). The mean number of paired-end reads was 465.2 M and the median was 553.2 M. The minimum number of paired-end reads for a sample was 75 M and the maximum number of paired-end reads per sample was 732.6 M.

We generated over 11.9 billion paired end reads for 112 tissue samples collected from Lake Memphremagog in 2019 and 2023 (Supplementary Tables 1-2 and 4). The mean number of paired-end reads was 106.8 M and the median was 99.75 M. The minimum number of paired-end reads for a sample was 45.6 M and the maximum number of paired-end reads per sample was 221.6 M.

Raw short read DNA sequence data for the reference and Lake Memphremagog samples are available on NCBI (BioProjects PRJNA1284832 and PRJNA1284812; SRA Accessions: SRR34320463 - SRR34320479, SRR34328784 - SRR36648244, and PV874414.1 and GCA_051624135.1).

2.2 DNA Long Read Sequencing Summary
We generated over 326.83 Gbp for the eight reference samples (Supplementary Tables 1-2 and 5). The mean number was 40.85 Gbp and the median was 28.3 Gbp. The minimum number of bp for a sample was 18.59 Gbp and the maximum number of bp for a sample was 128.25 Gbp.

We generated over 1,309.65 Gbp for 26 tissue samples collected from Lake Memphremagog in 2023 (Supplementary Tables 1-2 and 5). The mean number was 50.37 Gbp and the median was 45.0 Gbp. The minimum number of bp for a sample was 26.88 Gbp and the maximum number of bp for a sample was 138.85 Gbp.

Raw long read DNA sequence data for the reference and Lake Memphremagog samples are available on NCBI (BioProjects PRJNA1284832 and PRJNA1284812; SRA Accessions: SRR34320463 - SRR34320479, SRR34328784 - SRR36648244, and PV874414.1 and GCA_051624135.1).

Section 3. Mitochondrial Genome Assembly and Annotation

The mitochondrial genome of HL4 is 16,513 bp, which is consistent with the builds of existing mitochondrial genomes deposited in NCBI (e.g. NC_042499.1, and MF621733.1). The mitochondrial genome build had a base coverage of 644.4x and kmer coverage 200.6x. This mitochondrial build can be found with the following NCBI identifier (PV874414.1).

Section 4. Nuclear Genome Assembly and Annotation

4.1 Nuclear Genome Assembly Results
The draft nuclear genome of HL4 is 852,833,594 bp in length with a GC content of 39.91 and an N50 of 8,448,107 bp. The long-read coverage of the genomes is 72x and the short-read coverage is 162x. The genome consists of 2469 scaffolds, where 90% of the genome is represented by 141 scaffolds with the largest being 25,653 Kbp. A draft assembly and annotation of HL4 can be found in NCBI (PRJNA1284889, GCA_051624135.1).

4.2 Nuclear Genome Annotation Results
Based on the BUSCO version 5.4.7 actinopterygii database (Manni et al., 2021) the HL4 genome contains 97.3% complete BUSCOs (95.7% of the expected complete and single copy BUSCOs, 1.5% of the BUSCOs are complete but duplicated), 0.7% of the BUSCOs are fragmented, and 2.1% of the BUSCOs are missing. The draft genome,

annotated by EGAPx (https://github.com/ncbi/egapx, v0.4.1-alpha) contains 26,650 genes, and 43,693 mRNA transcripts.

Section 5. Mitochondrial Phasing of RNA sequence samples

We phased mitochondria from RNA sequences. These samples did not have paired healthy and lesion tissue samples. In some cases the allele frequency for a variant site was within 45-65% of the total reads for that site making it difficult to assign SNVs to host or tumor mitochondria. We left these samples out of the maximum likelihood tree of phased mitochondrial but included them in the neighbor joining tree of mitochondrial genotypes.

Section 6. Nuclear Analyses of differentially shared SNVs based on two methods: GATK's HaplotypeCaller and Mutect2

We used a somatic variant caller, Mutect2, to validate our tumor-specific results on a smaller dataset of 6 paired tumor and normal tissues with the greatest sequencing depth. We found a larger number of tumor-specific SNVs using GATK's HaplotypeCaller (n= 907,715 tumor-specific SNV sites) than (Mutect2 n=609,745 tumor-specific SNV sites, however they had a comparable distribution of variant sharing across fish tumor samples. The analysis of 6 fish produced more shared SNVs (using either method) than the analysis of the 16 paired tumor and normal tissues. This was largely due to increased read depth which minimized the filtering of SNVs; more sites had at least 10x coverage in 90% of the samples.

Nuclear genomic DNA was analyzed for single nucleotide variants and the variant allele frequency in tumors summarized (Supplementary Table 6). In addition, Pearson's correlations statistics for normalized copy number variation between all combinations of paired tumor tissue (Supplementary Tables 7 and 8) and between all combinations of paired normal tissue (Supplementary Table 9) are summarized and reported.

Section 7. Screening for Microorganisms in Tissue Samples

7.1 Methods

We looked for evidence of microorganisms (bacteria, fungi, viruses) in fish tissues. This was done by taking the tissue sample DNA and RNA sequences that did not map to the HL4 reference genomes and assembling partial and complete genomes from those reads. We used the nextflow (Di Tommaso et al., 2017) workflow nf-core/mag (Ewels et al., 2020; Krakau et al., 2022) to generate assemblies (SPAdes and MEGAHIT) (Li et al., 2016; Nurk et al., 2017) and run classification of mobile genetic elements (virus and plasmid) using geNomad on assembled contigs (Camargo et al., 2023) and

classification of prokaryotic and eukaryotic organisms on sequencing reads with Kraken2 (Wood et al., 2019).

Kraken2 data were imported into R using phyloseq (McMurdie & Holmes, 2013). We explored differential abundance of taxa between tumor and normal samples for each dataset (RNA or DNA) using DESeq2 (Love et al., 2014). To remove false positives, we filtered out organisms with less than 50 counts across the data set, and that were present in fewer than a quarter of the sample subset prior to running the analysis.

7.2 Viral Results

We identified viruses from 10 Classes of viruses (Supplementary Table 10). The vast majority were viruses of prokaryotes and were widely distributed across samples (∫.). We did not find any viruses specific to tumor samples. We did recover a Calicivirus previously identified in the RNA of vt02_05 (normal skin samples) and also found it in vt02_23 (a tumor skin sample) (Iwanowicz et al., 2022).

Of the viruses found in our fish samples, the majority were viruses of prokaryotes. We identified three viral lineages that are known to infect animals, and two that generally infect Eukaryotes. Of the Viruses of Animals, a member of Amarillovirales was found in a single tumor tissue, a member of Blubervirales was found in a single normal tissue, and a member of Orthopolintovirales was found in a single normal tissue. Of the viruses of Eukaryotes, a member of the Picornavirales (see the comment above) was found in a single tumor and a single normal tissue. A member of the Ortervirales was found in four normal tissues.

7.3 Cellular Organism Results

Kraken2 identified 1,933 unique genera across samples. Most had few reads across samples (77% had < 10 total reads). Only 84 (4%) genera were identified at least half of the samples. We subset the data four ways and ran differential abundance analysis between groups (Supplementary Table 11); DNA from paired tumor and normal tissue, all DNA normal vs tumor samples, DNA tumor vs RNA tumor, and RNA normal vs RNA tumor tissue. We found the only genera that increased in abundance in DNA tumors relative to DNA paired normal samples and all DNA normal samples was *Homo* (as in human or *Homo sapiens*). There were no differences in genera abundances between the healthy and tumor RNA skin samples. Between the RNA and DNA tumor data sets, we found that the genera Homo and Bradyrhizobium were more abundant in the DNA dataset and the genera Cupriavidus and Streptomyces were more abundant in the RNA dataset.

Supplementary Table	Title and Legend
Supplementary Table 1	Summary of reference fish samples
Supplementary Table 2	Summary of the Lake Memphremagog DNA samples sequenced for this study. This is a subset of the fish collected during the sampling events, where not all tumor fish were sampled, and only single tumors were sampled per fish.
Supplementary Table 3	Summary of 2015 RNAseq fish samples
Supplementary Table 4	Summary of short read sequences
Supplementary Table 5	Summary of raw long read sequences
Supplementary Table 6	Summary of variant allele frequency (VAF) for tumor SNV (n=144,177) shared by at least 6 tumor tissue samples. SNV cover variant sites with at least 10X coverage for 90% of the 17 tumor normal pairs with a median coverage > 8x.
Supplementary Table 7	Pearson's correlations statistics for normalized copy number variation between all combinations of paired tumor and normal tissue.
Supplementary Table 8	Pearson's correlations statistics for normalized copy number variation between all combinations of paired tumor tissue.
Supplementary Table 9	Pearson's correlations statistics for normalized copy number variation between all combinations of paired normal tissue.
Supplementary Table 10	Evidence of viral lineages in 2023 short read and 2015 RNA sequence samples. Absent samples had no identified viruses.
Supplementary Table 11	Differential genera abundance as determined by DESeq2 - 2015 and 2023 Lake Memphremagog samples

References

Camargo, A. P., Roux, S., Schulz, F., Babinski, M., Xu, Y., Hu, B., Chain, P. S. G., Nayfach, S., & Kyrpides, N. C. (2023). *You can move, but you can't hide: Identification of mobile genetic elements with geNomad* (p. 2023.03.05.531206). bioRxiv. https://doi.org/10.1101/2023.03.05.531206

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319. https://doi.org/10.1038/nbt.3820

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P., & Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, *38*(3), 276–278. https://doi.org/10.1038/s41587-020-0439-x

Iwanowicz, L. R., Blazer, V. S., Jones, T., Bodnar, M., Eckstrom, K., Dragon, J. A., & Emerson, P. (2022). Draft Genome Sequence of a Novel Calicivirus from a Brown Bullhead (Ameiurus nebulosus) from Lake Memphremagog, Vermont/Quebec. *Microbiology Resource Announcements*, *11*(3), e01188-21. https://doi.org/10.1128/mra.01188-21

Krakau, S., Straub, D., Gourlé, H., Gabernet, G., & Nahnsen, S. (2022). nf-core/mag: A best-practice pipeline for metagenome hybrid assembly and binning. *NAR Genomics and Bioinformatics*, *4*(1), lqac007. https://doi.org/10.1093/nargab/lqac007

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., Yamashita, H., & Lam, T.-W. (2016). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*, *102*, 3–11. https://doi.org/10.1016/j.ymeth.2016.02.020

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO: Assessing Genomic Data Quality and Beyond. *Current Protocols*, *1*(12), e323. https://doi.org/10.1002/cpz1.323

McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE*, *8*(4), e61217. https://doi.org/10.1371/journal.pone.0061217

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: A new versatile metagenomic assembler. *Genome Research*, *27*(5), 824–834. https://doi.org/10.1101/gr.213959.116

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 257. https://doi.org/10.1186/s13059-019-1891-0