Background

Biofuel produced from high-lipid algae offers a renewable and carbon-neutral alternative to fossil fuels, the limited natural resource our current energy infrastructure relies on. A significant obstacle to large-scale biofuel production is the potential for unknown pathogens to infect algal ponds, causing them to become inviable in a matter of days. Here we propose a metagenomic pipeline used to classify microbial species present in multiple agal pond samples for identification of candidate pathogens. In addition to Illumina short-read sequencing, we utilize Hi-C proximity ligation sequencing for improved deconvolution of species. Our pipeline produces graphs of both the taxonomic community profile and percentage abundances of species present.

We showcase the efficacy of our method by identifying a bacterial species in the phylum Bdellovibrionota as a candidate pathogen to the algal species *Nannochloropsis oceanica.* This novel species was found to dominate the microbial communities in sick ponds and was absent in healthy ponds.

Methods

A total of four samples were collected from four unique ponds at Arizona State University; two samples were collected from healthy ponds and two samples were collected from sick ponds. Each pond sample was prepared using two library preparation methods: Illumina shotgun and Hi-C proximity ligation, for a total of 8 library preparations. All samples were sequenced using the Illumina NextSeq 550 platform to produce 150 bp paired end reads.

The workflow manager Snakemake[1] was used to manage environments and combine metagenome software into a cohesive pipeline. This pipeline ran independently for each of the 4 separate samples described above, with each analysis receiving its associated shotgun and Hi-C read datasets as input.

The software fastp[2] was used to preprocess FASTQ data, excluding reads with ≤ 40% of bases with Phred quality score ≤ 15 or fewer than 150 bases in length. Next, BWA[3] was used in conjunction with SAMtools[4] to filter out reads mapping to the *N. oceanica* (SAMN10284882) reference assembly. MetaSPAdes[5] was then used to produce a metagenome-assembled genome (MAG) for each sample using the cleaned and filtered shotgun sequence data. The software metaQUAST[6] produced summary statistics for these assemblies.

Binning of metagenome assemblies was performed independently by four separate software: CONCOCT[7], MaxBin2[8], MetaBAT2[9], and bin3C[10]. While all software utilized shotgun sequence data, only bin3C supplemented its binning with Hi-C sequence data. Using the bin outputs from all four binning software, DAS Tool[11] was implemented to synthesize a set of non-redundant, high-quality bins for each sample. With the metagenome grouped into bins, the *Genome Taxonomy Database Toolkit* (gtdbtk[12]) software was used to identify the taxonomy of each bin. Bins were also subjected to checkm[13], which quantified bin abundance. Together, these analyses produced a quantitative classification table describing species abundance for each sample.

Results

Sequencing of the shotgun libraries yielded 13.2-16.6 GB of read data each while sequencing of the Hi-C prepared libraries yielded 3-4.7 GB of read data each. Fastp removed 9%-14.3% of reads from each sequence library, regardless of type (shotgun or Hi-C). In sequence libraries extracted from healthy ponds, BWA identified 15.9%-27.5% of reads primarily mapping to the *N. oceanica* assembly while sick ponds had 2.5%-15.3% of their reads primarily aligning. MetaSPAdes produced assemblies ~656 million bp long with 328K contigs, the largest of which was 2.1 million bp.

Binning results segregated by sample type (healthy or sick). DAS Tool consistently produced the largest quantity of high-quality bins, identifying (25, 28) in healthy ponds and (51, 63) in sick ponds. Inclusion of Hi-C data, facilitated by bin3C, increased the number of bins identified by ~10% across all samples. Classification and quantification performed by gtdbtk and checkm respectively produced taxonomic profiles that segregated by sample type (healthy or sick). Notably, both sick samples contained a high abundance (~40% of normalized population) of a single unclassified species located in the Phylum Bdellovibrionota, as seen in Figure 1.

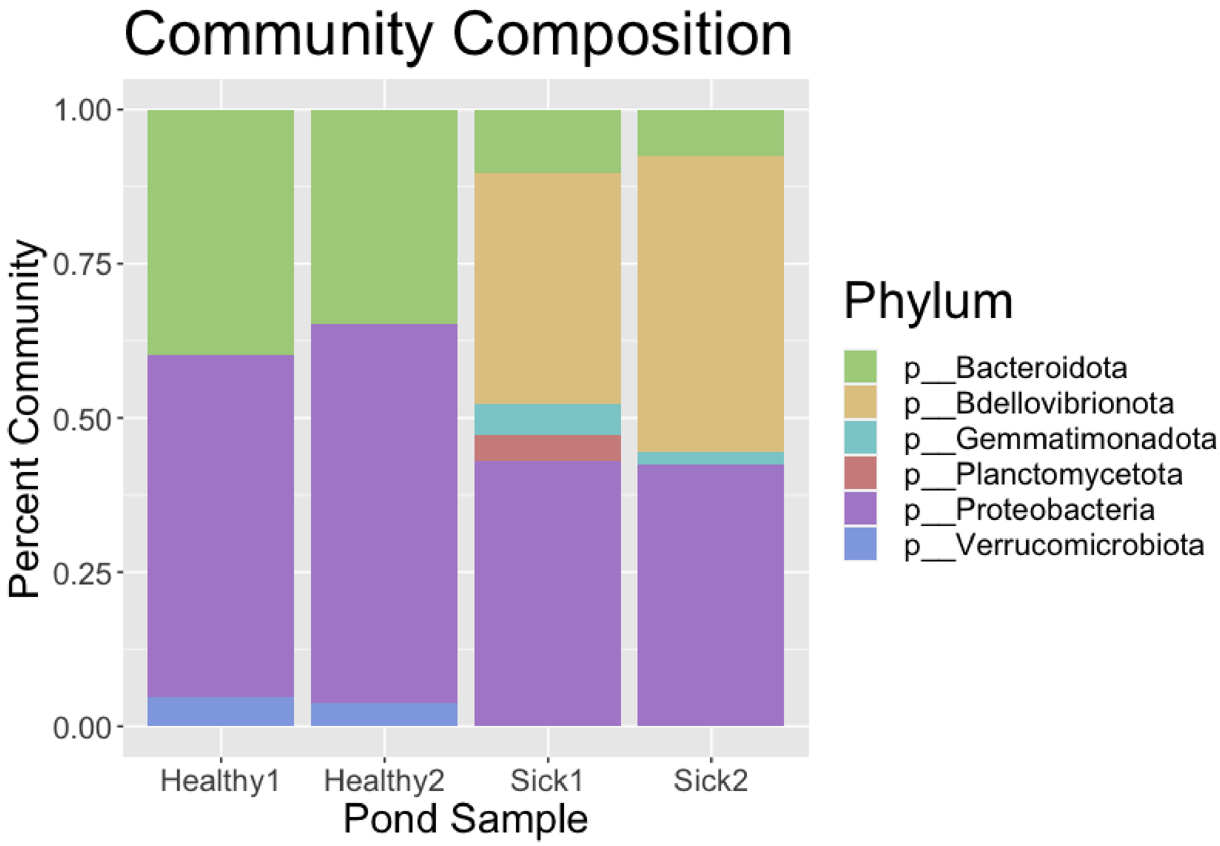

Figure 1. Algal Pond Community Composition. The percent community composition of each pond shows unique profiles for healthy and sick ponds after filtering of algal data. Sick ponds display a substantial abundance (~33%) of the phylum Bdellovibrionota, which is not observed in healthy ponds. This finding suggests that Bdellovibrionota is detrimental to algal pond health.

Discussion

This analysis demonstrates the viability of our metagenomic classification pipeline for identifying candidate pathogen species, using comparisons of healthy and sick pond samples. Additionally, we provide evidence that the addition of Hi-C sequence data to traditional shotgun sequencing has the potential to increase the number of species binned and classified. This finding has implications that may improve the difficulties associated with metagenomic analysis.

Works Cited

(1) Koster, J.; Rahmann, S. Snakemake--a Scalable Bioinformatics Workflow Engine. *Bioinformatics* **2012**, 28 (19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480.

(2) Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor. *Bioinformatics* **2018**, 34 (17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

(3) Li, H.; Durbin, R. Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* **2009**, 25 (14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

(4) Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* **2009**, 25 (16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

(5) Nurk, S.; Meleshko, D.; Korobeynikov, A.; Pevzner, P. A. MetaSPAdes: A New Versatile Metagenomic Assembler. *Genome Res* **2017**, 27 (5), 824–834. https://doi.org/10.1101/gr.213959.116.

(6) Mikheenko, A.; Saveliev, V.; Gurevich, V. MetaQUAST: Evaluation of Metagenome Assemblies. *Bioinformatics* **2016**, 32 (7), 1088-1090. https://doi.org/10.1093/bioinformatics/btv697

(7) Alneberg, J.; Bjarnason, B. S.; de Bruijn, I.; Schirmer, M.; Quick, J.; Ijaz, U. Z.; Lahti, L.; Loman, N. J.; Andersson, A. F.; Quince, C. Binning Metagenomic Contigs by Coverage and Composition. *Nature Methods* **2014**, 11 (11), 1144–1146. https://doi.org/10.1038/nmeth.3103.

(8) Wu, Y.-W.; Simmons, B. A.; Singer, S. W. MaxBin 2.0: An Automated Binning Algorithm to Recover Genomes from Multiple Metagenomic Datasets. *Bioinformatics* **2016**, 32 (4), 605–607. https://doi.org/10.1093/bioinformatics/btv638.

(9) Kang, D. D.; Li, F.; Kirton, E.; Thomas, A.; Egan, R.; An, H.; Wang, Z. MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction from Metagenome Assemblies. *PeerJ* **2019**, 7. https://doi.org/10.7717/peerj.7359.

(10) DeMaere, M. Z.; Darling, A. E. Bin3C: Exploiting Hi-C Sequencing Data to Accurately Resolve Metagenome-Assembled Genomes. *Genome Biology* **2019**, 20 (1), 46. https://doi.org/10.1186/s13059-019-1643-1.

(11) Sieber, C. M. K.; Probst, A. J.; Sharrar, A.; Thomas, B. C.; Hess, M.; Tringe, S. G.; Banfield, J. F. Recovery of Genomes from Metagenomes via a Dereplication, Aggregation and Scoring Strategy. *Nature Microbiology* **2018**, 3 (7), 836–843. https://doi.org/10.1038/s41564-018-0171-1.

(12) Chaumeil, P.-A.; Mussig, A. J.; Hugenholtz, P.; Parks, D. H. GTDB-Tk: A Toolkit to Classify Genomes with the Genome Taxonomy Database. *Bioinformatics* **2019**, btz848. https://doi.org/10.1093/bioinformatics/btz848.

(13) Parks, D. H.; Imelfort, M.; Skennerton, C. T.; Hugenholtz, P.; Tyson, G. W. CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes. *Genome Res* **2015**, 25 (7), 1043–1055. https://doi.org/10.1101/gr.186072.114.