Genotype-phenotype modeling of light ecotypes in *Prochlorococcus* reveals genomic signatures of ecotypic divergence

Evan Brenner[1^], Charmie Vang[1^], Connah Johnson[2], Janani Ravi[1,*]

[1]Department of Biomedical Informatics, University of Colorado Anschutz, Aurora, CO, USA; [2]Physical and Computational Science Directorate, Pacific Northwest National Laboratory, Richland, WA, USA. ^Co-primary authors. *Corresponding author: janani.ravi@cuanschutz.edu

Abstract

Prochlorococcus is a cyanobacterial genus that exhibits photosynthetic capacity and remarkable genetic diversity. We analyze how *Prochlorococcus* genomics relate to high vs. low light environment adaptations, applying traditional comparative genomics and machine learning (ML) approaches to connect genotypes to phenotypes. We downloaded nearly 1,000 *Prochlorococcus* genomes from NCBI with information on their light adaptation ecotype (high-light/low-light) and depth of isolation using metadata through JGI. Average nucleotide identity analysis and traditional pangenome generation tools struggle to capture the cyanobacterial core genome, but despite its scant conservation, we clearly observe a sharp separation of the taxon by its light utilization preferences — that is, its light ecotypes. A range of classical ML models trained to predict ecotype achieve exceptional binary classification accuracy even when predicting on partial genomes (Matthews Correlation Coefficient = 0.81 – 0.98), while regression models trained to predict the depth of isolation performed poorly, with relatively high root mean square error values (40.8 – 45.3m). For ecotype prediction, top features for the best-performing models included photosynthesis-associated genes and pathways, as well as some novel markers of unknown function. Our research findings recapitulate the extreme genetic versatility in *Prochlorococcus* and find that variable genetic markers allow excellent classification accuracy, and therefore, ecotype prediction, even for incomplete metagenomic assemblies, emphasizing the specialization and separation among these cyanobacterial ecotypes.

Introduction

Prochlorococcus is the most abundant photosynthetic organism in the ocean, utilizing its energy for a major fraction of global carbon cycling[1]; these bacteria are considered top candidates for supporting the bioeconomy as chassis for bioengineering interventions[2,3], and are well-known for their ecotypic divergence into high-light (HL) and low-light (LL) adapted clades[1,4–6]. *Prochlorococcus* is also noteworthy for its remarkable genetic repertoire, characterized by a relatively small but flexible genome punctuated by dizzyingly vast numbers and types of accessory genes[7] — Biller *et al*. reports that every new *Prochlorococcus* strain sequenced adds 160 new genes to a pangenome[3,4]. Low-light (LL) ecotype isolates show more flexibility than high-light (HL) isolates[4], but all contain an abundance of rare genes that may only ever have been observed in each single isolate and whose functions are generally unknown[3]. Among these variable organisms are different clades and subtypes, and it appears that *Prochlorococcus marinus* may push the boundaries of what we consider a bacterial species[4,6,7]. In the face of this, researchers have found that the HL/LL division serves as a fairly robust dividing line[4].

The genotype-to-phenotype relationship remains one of the central challenges in microbial genomics, especially for uncultivated or environmentally derived taxa[8]. While advances in sequencing yielded an abundance of microbial genomes, linking genomic content to ecological

or physiological traits remains complex, particularly when genomes are fragmented, poorly annotated, or lack associated metadata[8,9].

Traditionally, genotype-to-phenotype studies rely on comparative genomics approaches such as core and accessory genome analysis, gene enrichment statistics, and targeted gene annotation[8,10,11]. More recently, machine learning (ML) has emerged as a powerful tool for predicting phenotypes directly from genomic data[8,12,13]. These models can incorporate vast numbers of features and capture subtle, nonlinear patterns in genomic variation, offering the potential to link incomplete or draft assemblies with ecological function.

Here, we test what such approaches can yield in a model organism for microbial niche partitioning, *Prochlorococcus marinus*[4,5]. Its extreme genetic diversity makes it an ideal system to evaluate the boundaries of predictive power for pangenome-informed ML approaches.

We find that even when we train models on more complete cyanobacterial genomes and test them against fragmented, partial genomes, our ML models excel at predicting light ecotype, but they perform poorly when predicting depth of isolation. These results support that an isolate's light adaptation ecotype is robustly connected to its genomic content, that precise environmental isolation source is less detectable by genomics, and that the wide range of features identified as predictors support the vast phylogenetic distance others have reported in *Prochlorococcus*.

Methods

Curating cyanobacterial genomes and phenotypes. We curated cyanobacterial genomes from the DOE-funded Joint Genome Institute[14] (JGI), of which 501 had light ecotype metadata (categorized as high-light or low-light) and 594 had isolation depth metadata. After subsetting to samples that included both light and ecotype information, we retained 498 isolate genomes. We retrieved raw genomic assemblies via the NCBI Datasets command-line utility[15] using the provided NCBI assembly accession IDs. Many assemblies failed to pass through NCBI's annotation pipeline or contained minimal annotation, prompting us to re-annotate all genomes using Bakta (v11.11.3) with default parameters to ensure standardized gene prediction[16].

Comparative genomics. To evaluate genome quality, we ran BUSCO (v6.0.0) against the cyanobacteria lineage dataset and recorded completeness scores[17]. Genomes with >25% BUSCO completeness (n = 435) were used to construct a pangenome using PanTA, a rapid gene count matrix generation tool[18]. A full pangenome matrix was constructed across all training and testing genomes. Assemblies with higher BUSCO scores (>75%; n = 183) were labeled for ML model training, while those with intermediate BUSCO scores (25–75%; n = 252) were used for testing—genomes with <25% BUSCO completeness were not used for this analysis. Using the uwot[19] package in R[20], UMAP dimensionality reduction was performed to assess clustering of ecotypes and depth. In parallel, we conducted an all-vs-all genomic similarity analysis using FastANI to examine overall nucleotide-level divergence across ecotypes and assess whether HL and LL clades separate consistently in alignment-free space[21].

Model training. To predict light phenotypes, we trained random forest (RF) and logistic regression (LR) classifiers on the *Prochlorococcus* pangenome matrix using ranger[22] and glmnet[23–25]. Towards assessing the generalizability of our approach, we trained the models only on high-completeness (>75% BUSCO completeness) genomes; this model was then applied to the lower-quality test set (25%–75% BUSCO completeness). To help mitigate class imbalance, we down-sampled the model training set to include an equal number of HL and LL genomes.

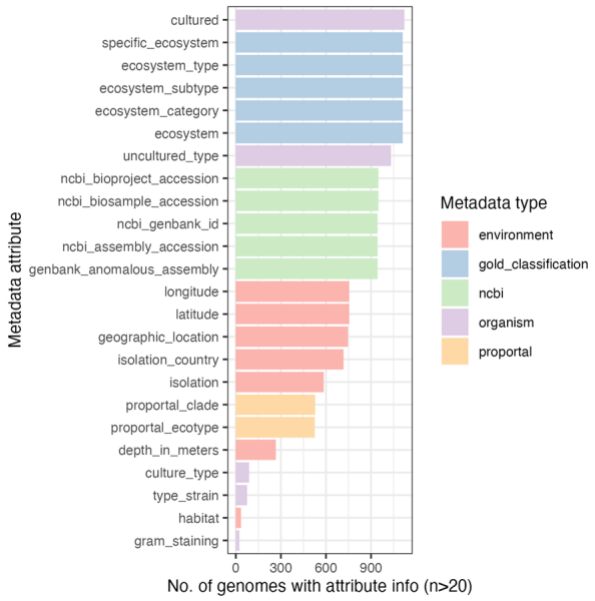

Figure 1: Distribution of metadata availability. In the set of ~1000 samples with metadata posted through JGI, the proportion of metadata completeness per isolate varied. Metadata field completeness per metadata type (color) and specific attribute (rows) are shown.

Model evaluation. Given the extreme feature space size and an overwhelming abundance of genes identified in only one or two genomes (expected in *Prochlorococcus*[4]), we constructed reduced feature matrices by creating a list of genes present in <5% of both HL and LL subsets – if the frequency was below this cutoff in both ecotypes, the gene was removed from the reduced pangenome. Finally, the tidymodels[26] recipe implementation of PCA (num_comp=50) was used on the reduced feature space to further reduce the number of features. These two reduced subsets were used for modeling separately through the same workflow as the whole feature space. For ecotype classification, performance was assessed using the Matthews correlation coefficient (MCC), a robust metric for binary classification with imbalanced classes[27]. For regression models that predict sample origin depth in meters, we report performance using root mean square error (RMSE) to quantify the error between predicted and actual values. Each model type (e.g., logistic regression, PCA feature space, ecotype classifier) was run 10 times with randomly generated seeds, and the mean and standard deviation of performance were taken across all runs to enumerate the model variance.

Interpreting key features. We assessed feature importance values for each model run using the vip R package[28]. Features that repeatedly appeared within multiple model iterations, and those features seen across multiple different model types, were considered higher confidence. We processed representative protein sequences for genes of interest into domains and motifs using MolEvolvR[29], and queried for homologs using BLAST[30]. We processed and analyzed data on the University of Colorado's Alpine HPC cluster.

Results

Given the dominant roles cyanobacteria play in aquatic ecosystems and carbon cycling, questions about genomic signatures of their niche adaptation have high importance towards understanding both global ecology and their repurposing for bioengineering. However, metadata availability (source: JGI ProPortal) paired with sequenced *Prochlorococcus* genomes (NCBI) constrains most questions in our present dataset (**Figure 1**). We started with the classic and essential adaptation to light (high and low light ecotypes; HL, LL), followed by most likely co-adaptations to depth.

We observed that the primary challenge with *Prochlorococcus* is the lack of a 'core genome' by classical pangenome generation tools — even when limiting to highly BUSCO complete genomes, PanTA and other tools all struggled to capture a core genome due to the extreme sequence and gene content diversity across genomes. Even pangenomes subsetted to HL and LL ecotypes failed to return core genes, further highlighting the phenotypic spread in the dataset. Despite this, the full pangenome provided a feature set of 77,427 total gene features, with 75,414 found in under 15% of genomic isolates.

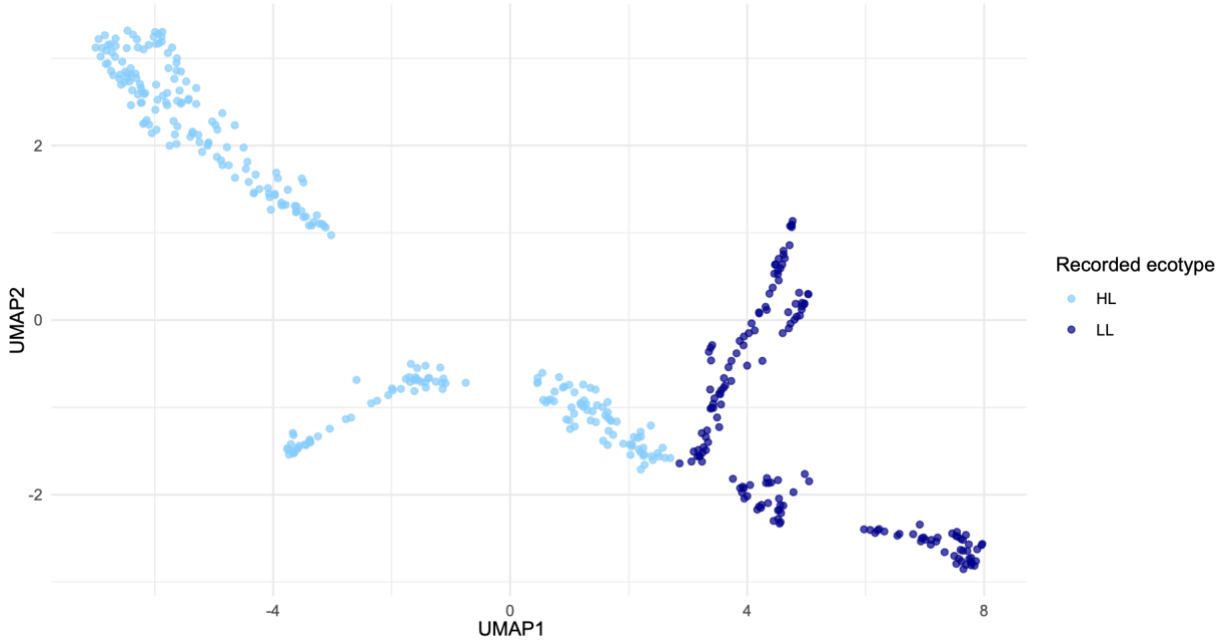

Figure 2: UMAP of genome content for *Prochlorococcus*. We performed unsupervised clustering of the PanTA pangenome content of all analyzed isolates and then colored points by their recorded ecotype (high-light, HL; low-light, LL). A strong separation between ecotypes is clear, with subclusters possibly reflecting the *Prochlorococcus* clades that have been previously described in the literature.

To capture the overall global structure of genome variations, we employed unsupervised clustering by UMAP. When clustering encoded pangenome content per isolate, we found a striking separation of HL and LL ecotypes, even when fragmented genomes were included, suggesting that the ecotype signal is well-embedded in pangenomic variation (**Figure 2**). Given these strongly separated clusters, we would expect ML models to perform well, which we observed across ecotype prediction iterations. Models consistently successfully predicted LL genomes, but most misclassified a subset of HL genomes as LL (**Figure 3A**; light blue triangles). These misclassifications were less common as the feature space was reduced, with the fewest errors occurring in the PCA models (**Figure 3B,C**). The repeated misclassifications restricted to a small, consistent subset of HL genomes may represent transitional or misclassified ecotypes or may potentially stem from rarer incidents of horizontal gene transfer across ecotypes.

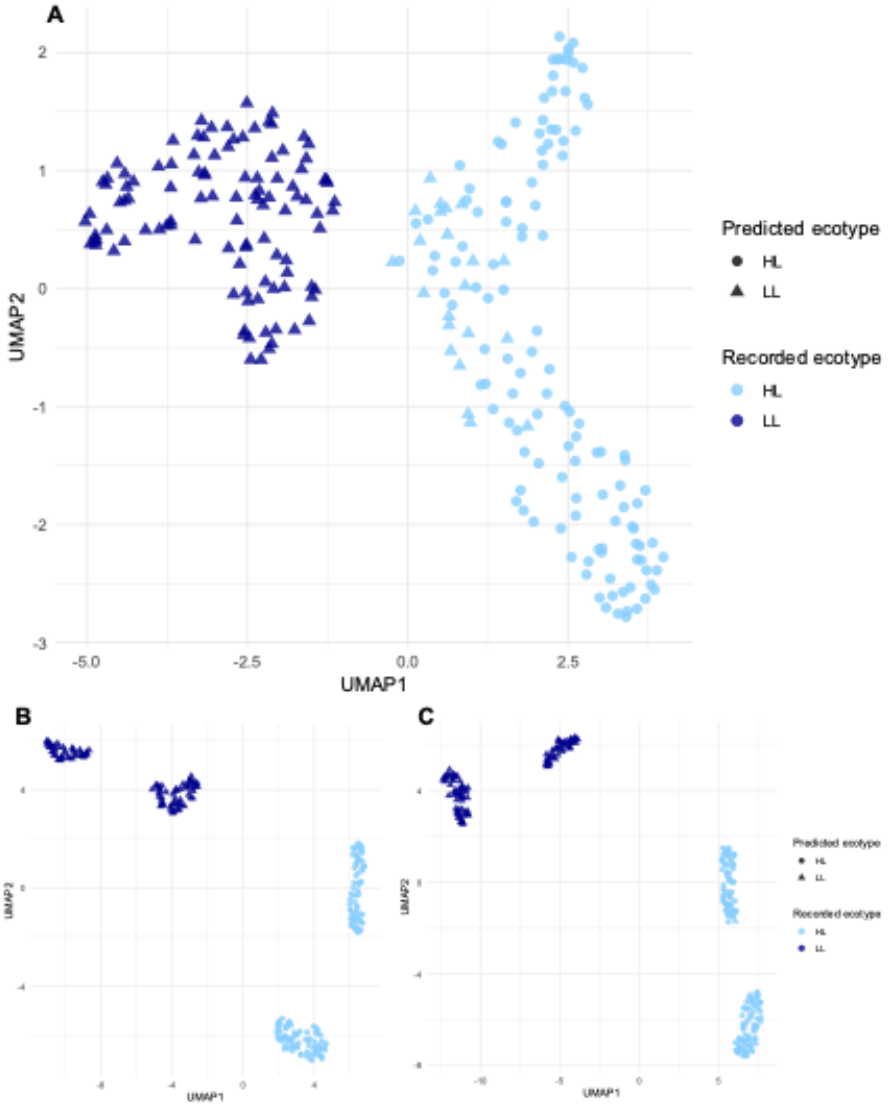

Figure 3: UMAP of test set *Prochlorococcus* genome content with ML classifications. After machine learning modeling, we compared model predictions (shapes) with the recorded ecotypes from JGI (colors). **A)** A representative plot from a single logistic regression model run using the full feature space. Note that misclassifications only occur ir one direction – some HL isolates are incorrectly predicted to be LL isolates (light blue triangles, right cluster). Across models, this pattern of unidirectional misclassification was always observed. **B)** PCA models with random forest, and **C)** PCA models with logistic regression, which together showed the best cluster separation and the fewest misclassifications.

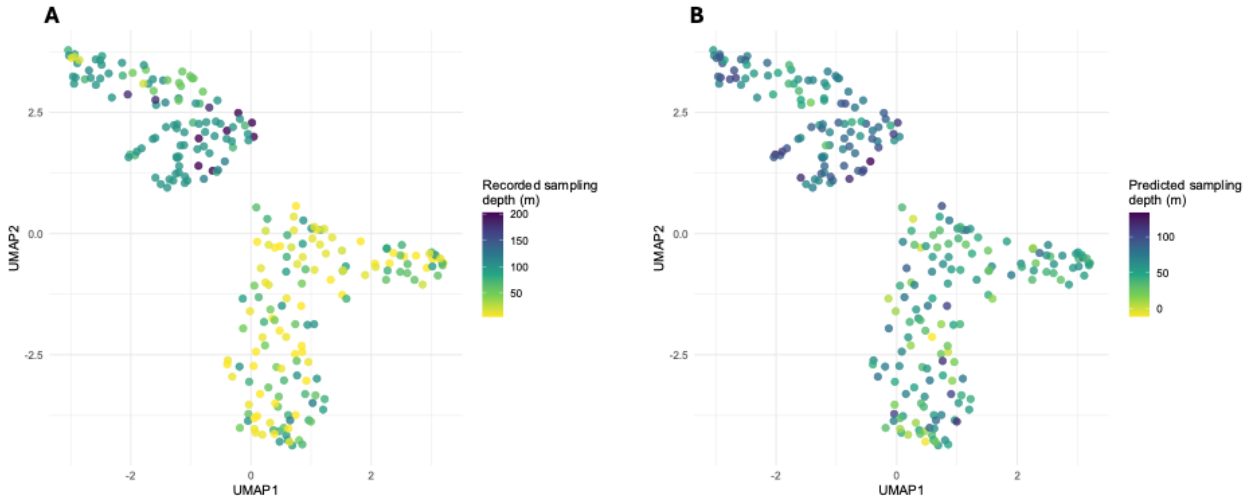

Figure 4: UMAP of test set *Prochlorococcus* genome content colored by depth of isolation. A) When clusters were colored by their reported depth of isolation, one cluster tended to be found at greater depths (corresponding to the LL ecotype cluster), while the other tended to by isolated closer to the surface (corresponding to the HL ecotype cluster). However, we observed a significant amount of variation within and across clusters. **B)** Regression models performed poorly at predicting the depth of isolation, tending to assign fairly uniform values to each cluster. While these did reflect the reality of sampling depth, the root mean square error for these models was >20% of the total variation possible in the dataset.

In contrast to the high predictability of light-adapted subpopulations, regression models predicting depth showed poor performance (RF: 40.7m–45.3m; LR: 41.0m–43.6m). Reducing the feature space (from ~77K to ~7K) marginally improved LR accuracy but yielded worse RF performance. Given that 476 of 498 samples fell between 0 and 150m, this level of error (~30% of the typical range) suggests major flaws with any identified predictors. A UMAP colored by depth shows that while HL and LL isolates tend to cluster at higher and lower depths, there is considerable within-cluster variation with some LL isolates collected near the surface and vice versa (**Figure 4A**). Model predictions of depth are relatively uniform within each cluster by contrast (**Figure 4B**).

Overall, light-based ecotype prediction models performed exceptionally well. Those using the full and reduced feature spaces performed equivalently when predicting ecotype, but PCA models performed best (**Figure 5**). Logistic regression-based models tended to slightly outperform the random forest approach, but both were robust.

Conversely, we noted very little difference when comparing model types for errors in depth prediction (**Figure 6**). These models showed poor performance regardless of model type (LR, RF) or feature elimination strategy.

The top predictors driving performance of ecotype classification showed low per-feature importance scores and variation between the models, suggesting that performance was not well-explained by any individual, specific genes. We therefore selected only the best-performing model approaches (LR PCA and RF PCA) and extracted predictors from the top PC in both models.

Across 10 randomly seeded model runs, we chose the genes that appeared most frequently in this subset as the highest confidence feature set.

These features included a high-light inducible family protein, a conserved hypothetical protein, the UV-B stress response GTPase *obgE*, an M3 peptidase encoded adjacent to *ndhG*[31], and another conserved hypothetical protein Ycf35 (possibly involved in photosynthesis).

Finally, we visualized divergence using an all-vs-all FastANI analysis. HL and LL ecotypes formed completely distinct genomic similarity clusters (**Figure 7**). These findings corroborate the presence of deep ecological and evolutionary separation between HL and LL clades, consistent with prior phylogenetic studies[4–6]. Due to the fragmented nature of many of these genomes, some also show overall low similarity to others due to a few regions to effectively compare. FastANI intentionally does not report ANI values below ~75% – blank pairs represent a value at or below this point, meaning the true distances between genomes may still be greater than what the software is designed to calculate.

In summary, while *Prochlorococcus* diversity is unusually extreme among bacteria, genomics-based classification of ecotype by ML yields excellent accuracy, even when models are tested by predicting on incomplete genomes. Predictions for other metadata variables, like depth of isolation, yielded poor performance, suggesting that if genetic predictors of depth as a niche exist, they are not captured by our current dataset or analysis.

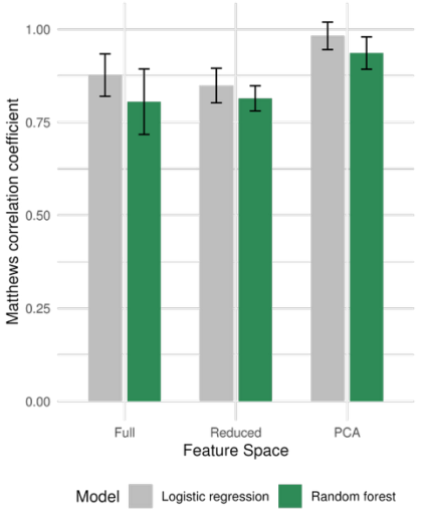

Figure 4: Mean model performance for ecotype prediction. For each model type, 10 randomly seeded models were trained, binary classification performance was recorded by the Matthews correlation coefficient for each iteration, and mean and standard deviation per model type is shown. PCA models outperform others, but all return high predictive performance.

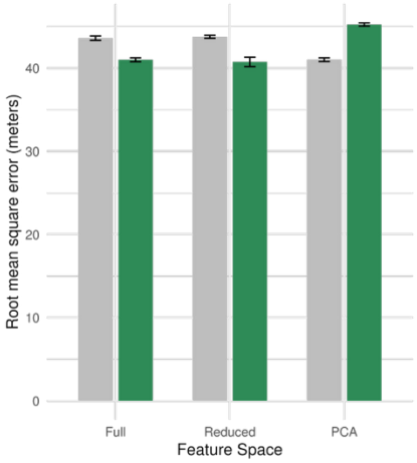

Figure 5: Mean regression model root mean square error for depth prediction. For each model type, 10 randomly seeded models were trained, with performance recorded by root mean square error (RMSE). All models showed high error relative (~40m) relative to the overall variation observed in the dataset. No particular model type appeared to do better in this task, suggesting any signal predicting depth of isolation may not be captured by our modeling approach or dataset.

Discussion

Our results demonstrate that simple, classical ML models trained on gene presence/absence data can accurately infer ecological phenotypes from draft-quality or metagenomic assemblies. The strong separation in both gene content (pangenome) and nucleotide identity (FastANI) supports that ecotype is a deeply encoded and readily recoverable signal in *Prochlorococcus* genomes[4–6].

Considering pangenomes including multiple ecotypes reliably have very few core genes, even when subsetted to genomes with high BUSCO completeness, we find that current pangenome pipeline clustering methods are inadequate to capture the diversity in these cyanobacterial

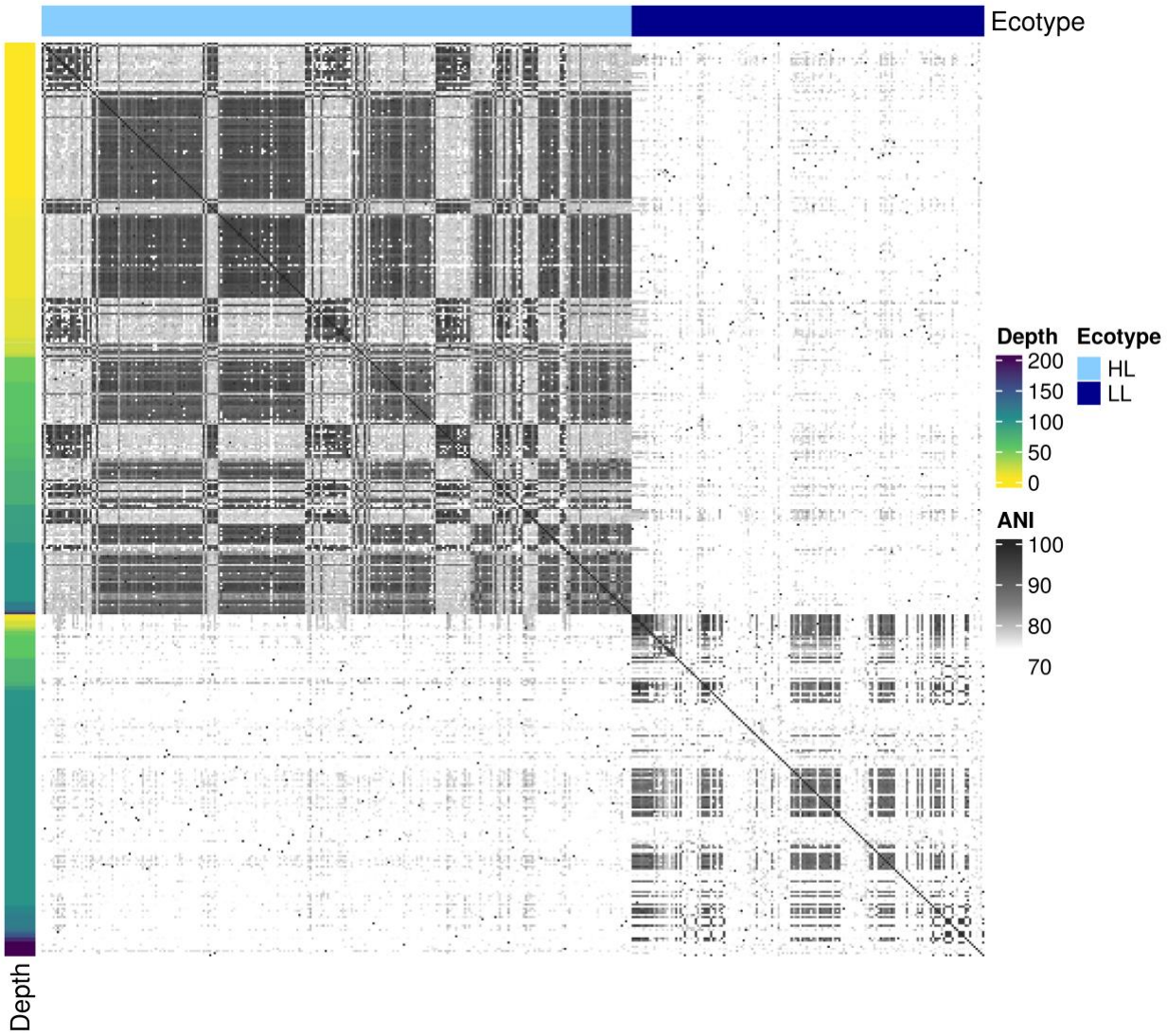

Figure 6: Pairwise average nucleotide identity heatmap of *Prochlorococcus* genomes. Using FastANI, we created an all-vs.-all comparison of average nucleotide identity and plotted the results as a heatmap. Self-comparisons are along the diagonal (100% ANI). The order of comparisons was selected by ecotype, then sorted by depth. Two distinct blocks composed of a high-light set of genomes and a low-light set of genomes are apparent, though considerable diversity exists within these blocks as well. HL genomes tend to share greater ANI, while LL genomes share lower ANI, reflecting the diversity seen in these ecotypes as described in the literature.

isolates, but HMM-based orthology detection (i.e., as used by BUSCO) may be better suited for this. Nevertheless, despite the dilution of signal in our gene matrices by the fragmentation of gene homologs into many separate clusters, predictive performance remains excellent. When subsetting to only gene predictors that continually arise from top models using multiple ML approaches, a small number of genes remain. Of these genes, most have apparent roles in differentiating high and low-light lifestyles – a putative high-light inducible protein with a light-harvesting complex domain (PFAM00504); a hypothetical protein Ycf35 suspected to play a role in photosynthesis (containing DUF1257); the photosynthesis-essential *obgE*[32], induced by UV-B stress and protection, and even an M3 family peptidase encoded directly adjacent to the photosystem-supporting Ndh-1 complex member *ndhG*[31]. However, a 93aa conserved hypothetical protein has no suspected functions annotated, contains no domains, encodes no

predicted secondary structures, and shows scattered conservation across only isolates in *Prochlorococcus*. Genes like these are a major fraction of cyanobacterial diversity, and while these results here suggest it may have a role in differentiating ecotypes and light adaptation, it is our understanding that no other characterization of this gene has been published. Our analysis may be combined with other phylogeny-based methodologies to suggest an evolutionary trajectory for the development of circadian regulation in cyanobacteria and pathway rewiring for enhanced bioproduction[33,34].

Regardless of the specific features and their precise mechanistic contributions, further interrogation of these and other ecotype-specific gene features behind good predictive performance would greatly benefit from refined homology assignment, but these genes already support the previously observed evolutionary divergence and specialization of *Prochlorococcus* ecotypes[4,6]. We do not expect that the minimal set we discuss here is an exhaustive list, but instead offer that the vast library of cyanobacterial genes without known function could be associated with specific phenotypes and traits using ML approaches. Towards this goal, supporting and expanding ongoing sampling efforts for cyanobacteria, developing new methodologies to better cluster and differentiate the complex pangenome space of *Prochlorococcus*, and investigating novel approaches to expand on the evolutionary trajectory of bacterial species of interest are all worthwhile and necessary directions that research into cyanobacteria and beyond will greatly benefit from.

Acknowledgments

This work was enabled through collaboration and continuous discussion with the NW-BRaVE teams at Pacific Northwest National Labs (PNNL); we would like to particularly thank Drs. Ruonan Wu, Margaret Cheung, and David Pollock.

Funding

This project was partially supported by the Northwest Biopreparedness Research Virtual Environment project (NW-BRaVE), funded by the Department of Energy, Office of Science, Biological and Environmental Research program FWP 81832. This research was performed on a project award (Enhancing biopreparedness through a model system to understand the molecular mechanisms that lead to pathogenesis and disease transmission) from the Environmental Molecular Sciences Laboratory, a DOE Office of Science User Facility sponsored by the Biological and Environmental Research program under Contract No. DE-AC05-76RL01830. Pacific Northwest National Laboratory is a multi-program national laboratory operated by Battelle for the DOE under Contract DE-AC05-76RL01830. This work utilized the Alpine high-performance computing resource at the University of Colorado Boulder. Alpine is jointly funded by the University of Colorado Boulder, the University of Colorado Anschutz, Colorado State University, and the National Science Foundation (award 2201538). Also, start-up funds from the University of Colorado Anschutz awarded to JR supported JR, EPB, and CKV.

Bibliography

1. Cai, L., Li, H., Deng, J., Zhou, R. & Zeng, Q. Biological interactions with Prochlorococcus: implications for the marine carbon cycle. *Trends Microbiol.* **32,** 280–291 (2024).
2. Motwalli, O., Essack, M., Jankovic, B. R., Ji, B., Liu, X., Ansari, H. R., Hoehndorf, R., Gao, X., Arold, S. T., Mineta, K., Archer, J. A. C., Gojobori, T., Mijakovic, I. & Bajic, V. B. In silico screening for candidate chassis strains of free fatty acid-producing cyanobacteria. *BMC Genomics* **18,** 33 (2017).
3. Laurenceau, R., Bliem, C., Osburne, M. S., Becker, J. W., Biller, S. J., Cubillos-Ruiz, A. & Chisholm, S. W. Toward a genetic system in the marine cyanobacterium Prochlorococcus. *Access Microbiol.* **2,** (2020).
4. Biller, S. J., Berube, P. M., Lindell, D. & Chisholm, S. W. Prochlorococcus: the structure and function of collective diversity. *Nat. Rev. Microbiol.* **13,** 13–27 (2015).
5. Johnson, Z. I., Zinser, E. R., Coe, A., McNulty, N. P., Woodward, E. M. S. & Chisholm, S. W. Niche Partitioning Among *Prochlorococcus* Ecotypes Along Ocean-Scale Environmental Gradients. *Science* **311,** 1737–1740 (2006).
6. Becker, J. W., Pollak, S., Berta-Thompson, J. W., Becker, K. W., Braakman, R., Dooley, K. D., Hackl, T., Coe, A., Arellano, A., LeGault, K. N., Berube, P. M., Biller, S. J., Cubillos-Ruiz, A., Van Mooy, B. A. S. & Chisholm, S. W. Novel isolates expand the physiological diversity of Prochlorococcus and illuminate its macroevolution. *mBio* **15,** e0349723 (2024).
7. Berube, P. M., Biller, S. J., Hackl, T., Hogle, S. L., Satinsky, B. M., Becker, J. W., Braakman, R., Collins, S. B., Kelly, L., Berta-Thompson, J., Coe, A., Bergauer, K., Bouman, H. A., Browning, T. J., De Corte, D., Hassler, C., Hulata, Y., Jacquot, J. E., Maas, E. W., Reinthaler, T., Sintes, E., Yokokawa, T., Lindell, D., Stepanauskas, R. & Chisholm, S. W. Single cell genomes of Prochlorococcus, Synechococcus, and sympatric microbes from diverse marine environments. *Sci. Data* **5,** 180154 (2018).
8. Lees, J. A., Mai, T. T., Galardini, M., Wheeler, N. E., Horsfield, S. T., Parkhill, J. & Corander, J. Improved Prediction of Bacterial Genotype-Phenotype Associations Using Interpretable Pangenome-Spanning Regressions. *mBio* **11,** e01344-20 (2020).
9. Kemble, H., Nghe, P. & Tenaillon, O. Recent insights into the genotype–phenotype relationship from massively parallel genetic assays. *Evol. Appl.* **12,** 1721–1742 (2019).
10. Brynildsrud, O., Bohlin, J., Scheffer, L. & Eldholm, V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* **17,** 238 (2016).
11. Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J. A., Gladstone, R. A., Lo, S., Beaudoin, C., Floto, R. A., Frost, S. D. W., Corander, J., Bentley, S. D. & Parkhill, J. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **21,** 180 (2020).
12. Ghosh, A., Vang, C. K., Brenner, E. P. & Ravi, J. Unlocking antimicrobial resistance with multiomics and machine learning. *Trends Microbiol.* S0966842X25001465 (2025). doi:10.1016/j.tim.2025.04.017
13. Jovel, J. & Greiner, R. An Introduction to Machine Learning Approaches for Biomedical Research. *Front. Med.* **8,** 771607 (2021).
14. Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., Smirnova, T., Grigoriev, I. V. & Dubchak, I. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res.* **42,** D26–D31 (2014).
15. O'Leary, N. A., Cox, E., Holmes, J. B., Anderson, W. R., Falk, R., Hem, V., Tsuchiya, M. T. N., Schuler, G. D., Zhang, X., Torcivia, J., Ketter, A., Breen, L., Cothran, J., Bajwa, H., Tinne, J., Meric, P. A., Hlavina, W. & Schneider, V. A. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Sci. Data* **11,** 732 (2024).
16. Schwengers, O., Jelonek, L., Dieckmann, M. A., Beyvers, S., Blom, J. & Goesmann, A. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb. Genomics* **7,** 000685 (2021).

17. Tegenfeldt, F., Kuznetsov, D., Manni, M., Berkeley, M., Zdobnov, E. M. & Kriventseva, E. V. OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. *Nucleic Acids Res.* **53,** D516–D522 (2025).
18. Le, D. Q., Nguyen, T. A., Nguyen, S. H., Nguyen, T. T., Nguyen, C. H., Phung, H. T., Ho, T. H., Vo, N. S., Nguyen, T., Nguyen, H. A. & Cao, M. D. Efficient inference of large prokaryotic pangenomes with PanTA. *Genome Biol.* **25,** 209 (2024).
19. Melville, J. uwot: The Uniform Manifold Approximation and Projection (UMAP) Method for Dimensionality Reduction. 0.2.3 (2019). doi:10.32614/CRAN.package.uwot
20. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2024). at <https://www.R-project.org/>
21. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9,** 5114 (2018).
22. Wright, M. N. ranger: A Fast Implementation of Random Forests. 0.17.0 (2015). doi:10.32614/CRAN.package.ranger
23. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33,** 1–22 (2010).
24. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J. Stat. Softw.* **39,** (2011).
25. Tay, J. K., Narasimhan, B. & Hastie, T. Elastic Net Regularization Paths for All Generalized Linear Models. *J. Stat. Softw.* **106,** (2023).
26. Kuhn, M. & Wickham, H. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. (2020). at <https://www.tidymodels.org>
27. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21,** 6 (2020).
28. Greenwell, B. M. & Boehmke, B. C. Variable Importance Plots—An Introduction to the vip Package. *R J.* **12,** 343–366 (2020).
29. Krol, J. D., Burke, J. T., Chen, S. Z., Sosinski, L., Alquaddoomi, F. S., Brenner, E. P., Wolfe, E. P., Rubinetti, V. P., Amolitos, S., Reason, K. M., Johnston, J. B. & Ravi, J. MolEvolvR: A web-app for characterizing proteins using molecular evolution and phylogeny. Preprint at https://doi.org/10.1101/2022.02.18.461833 (2022)
30. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).
31. Hualing, M. Cyanobacterial NDH-1 Complexes. *Front. Microbiol.* **13,** 933160 (2022).
32. Rubin, B. E., Wetmore, K. M., Price, M. N., Diamond, S., Shultzaberger, R. K., Lowe, L. C., Curtin, G., Arkin, A. P., Deutschbauer, A. & Golden, S. S. The essential gene set of a photosynthetic organism. *Proc. Natl. Acad. Sci.* **112,** (2015).
33. Johnson, C. G. M., Johnson, Z., Mackey, L., Li, X., Sadler, N. C., Zhang, T., Qian, W., Bohutskyi, P., Fend, S., Cheung, M. S. Multi-Omics Reveals Temporal Scales of Carbon Metabolism in *Synechococcus Elongatus* PCC 7942 Under Light Disturbance. *PRX Life* **3**, 3 (2025).
34. Gilliam, A., Sadler, N. C., Li, X., Garcia, M., Johnson, Z., Veličković, M., Kim, Y., Qian, W., Cheung, M. S., Bohutskyi, P. Cyanobacterial circadian regulation enhances bioproduction under subjective nighttime through rewiring of carbon partitioning dynamics, redox balance orchestration, and cell cycle modulation. *Microbial Cell Factories* **24**, 56 (2025).