From orientation to publication: training in genome sequencing and assembly for new graduate students

Anne Nakamoto[1,2], Lisa Keith[3], Qingyi Yu[3], Lionel Sugiyama[3], Xiaohua Wu[3], Blaine Luiz[3], MaryAnn Villalun[3], Jodie Jacobs[1,2], Russ Corbett-Detig[1,2], Ariana Cisneros[1,2], Harrison D. Heath[1,2], Cole Shanks[1,2], Faith Okamoto[1,2], Alexis Abigail Aroma Alburo[1], Kyle Henricson[1], Yi Jun Lan[1], Henry Moore[1], William Seligmann[1], Yulia Zybina[1]

[1]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA; [2]Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA; [3]Tropical Plant Genetic Resources and Disease Research Unit, Daniel K Inouye U.S. Pacific Basin Agricultural Research Center, Agricultural Research Service, U.S. Department of Agriculture, Hilo, Hawai'i, 96720, USA

Corbett-Detig Lab
aanakamo@ucsc.edu

Background

Graduate orientation provides a unique opportunity for genomics training

First-year PhD and Masters students in the Biomolecular Engineering & Bioinformatics Department at UC Santa Cruz participated in a **two-week orientation "Bootcamp"** during which instructors led a hands-on genome sequencing and assembly exercise. This culminated in a final project developed by each individual student.(Fig. 1)

The Bootcamp was designed to provide the following:

- Training in **wet-lab** and **computational skills**
- Experience with the **research process**
- Practice **developing questions** and **presenting results**
- Contributing meaningful results through **publication**
- **Cohort-building** through shared activities

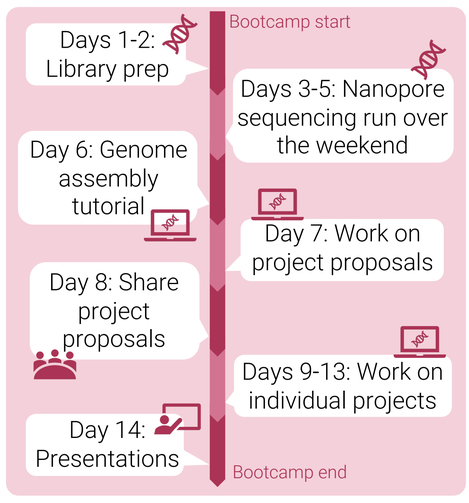

Figure 1: Timeline of two-week "Bootcamp".

Genome sequence of a fungal pathogen of the native 'ōhi'a tree in Hawai'i

For the Bootcamp, we produced a **long-read genome assembly** of an important fungal plant pathogen in Hawai'i. The Rapid 'Ōhi'a Death (ROD) disease affects the keystone and culturally significant native ohia tree (*Metrosideros polymorpha*) species in Hawai'i.[1,2] (Fig. 2) ROD was first characterized in 2014, and is caused by two novel fungal pathogens, *Ceratocystis lukuohia* and *huliohia*.[3] Genomic resources for these pathogens remain limited, and while *C. lukuohia* has a high quality, long-read reference genome, no such genome had been available for *C. huliohia*.

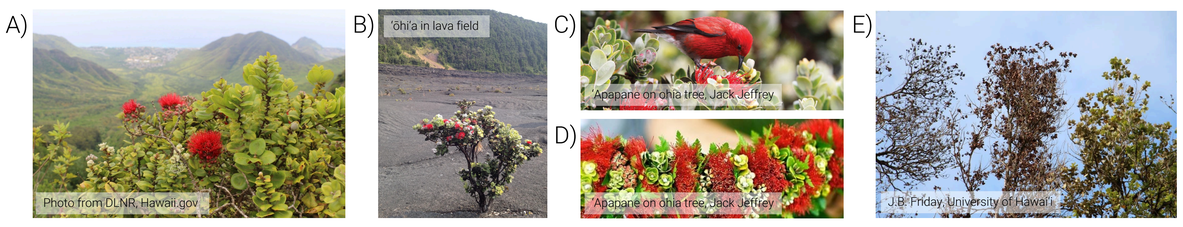

Figure 2: Motivation for sequencing *C. huliohia*. The 'ōhi'a is A) the dominant native forest tree, B) founder species, C) habitat to many native species, and D) important in Hawaiian culture (eg. lei po'o). E) ROD disease progression.

Methods

Genome sequencing and assembly workflow design for genomics training

In advance of the bootcamp, instructors cultured the C25-5 *C. huliohia* isolate,[3] performed DNA extraction,[4,5] and generated preliminary sequence data. Incoming students were then led in **library preparation and ONT MinION sequencing** during the first part of the Bootcamp.(Fig. 2A) Sequencing reads generated before and during the bootcamp were combined and basecalled by instructors. Students were then guided through a **computational tutorial to assemble the genome.**(Fig. 2B) Finally, students were asked to develop a question based on their own interests that could be answered through computational analysis of the resulting genome and sequencing data, and presented their work on the final day of the Bootcamp.(Fig. 2B)

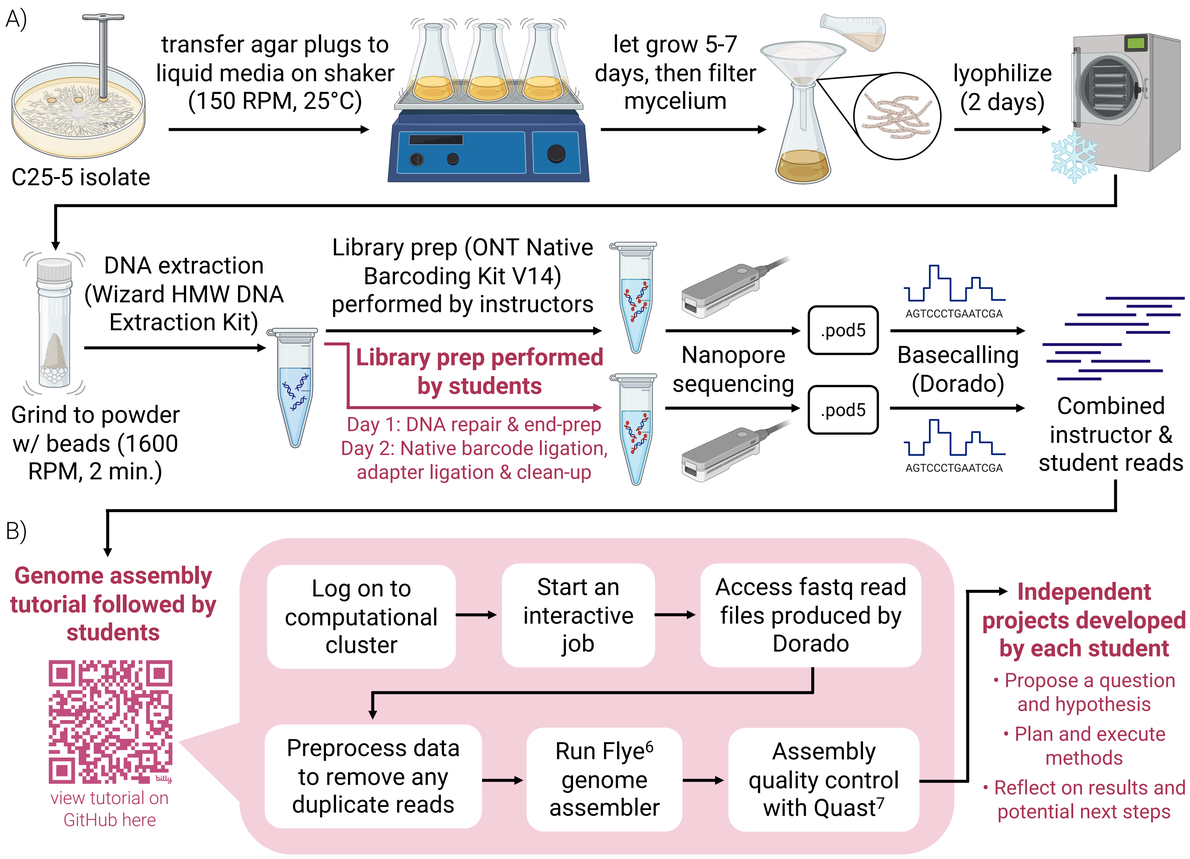

Figure 3: Genome sequencing and assembly workflow in preparation for and during the Bootcamp. A) Wet-lab preparation for sequencing by instructors and library preparation activities performed by students during the Bootcamp. B) Computational portion of the bootcamp, consisting of a genome assembly tutorial and independent student-led analysis. Steps performed by students are shown in pink. Icons from Biorender.com.

Results

Students developed diverse analyses based on the sequencing data

The Bootcamp students used their creativity and individual interests to develop a question that could be answered through computational analysis of the genome assembly. Projects ranged from investigating structural variation between *C. huliohia* and *C. lukuohia*, to protein structural prediction of potential fungicide targets,(Fig. 4) and were presented on the last day of Bootcamp.

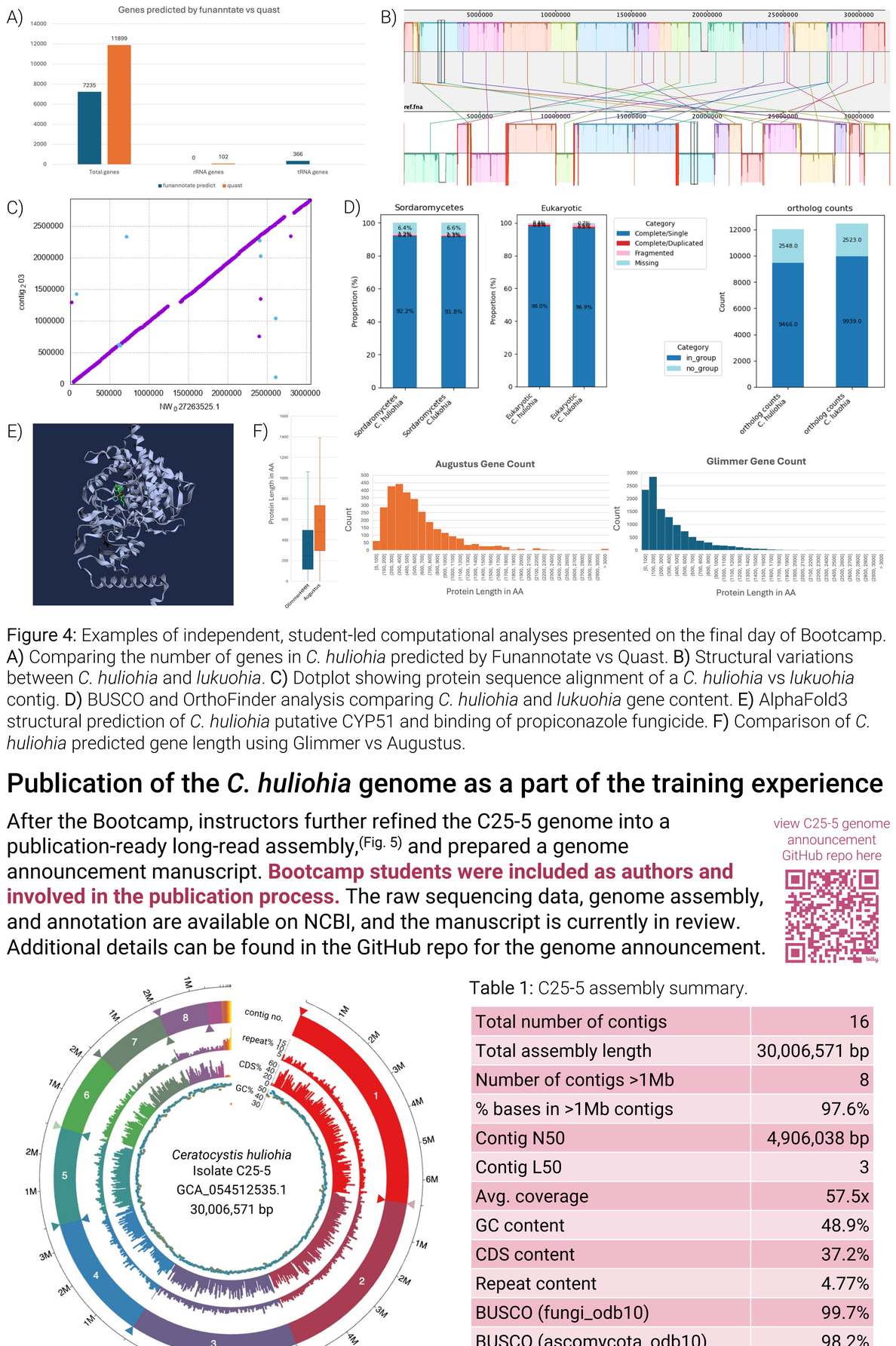

Figure 4: Examples of independent, student-led computational analyses presented on the final day of Bootcamp. A) Comparing the number of genes in *C. huliohia* predicted by Funannotate vs Quast. B) Structural variations between *C. huliohia* and *lukuohia*. C) Dotplot showing protein sequence alignment of a *C. huliohia* vs *lukuohia* contig. D) BUSCO and OrthoFinder analysis comparing *C. huliohia* and *lukuohia* gene content. E) AlphaFold3 structural prediction of *C. huliohia* putative CYP51 and binding of propiconazole fungicide. F) Comparison of *C. huliohia* predicted gene length using Glimmer vs Augustus.

Publication of the *C. huliohia* genome as a part of the training experience

After the Bootcamp, instructors further refined the C25-5 genome into a publication-ready long-read assembly,(Fig. 5) and prepared a genome announcement manuscript. **Bootcamp students were included as authors and involved in the publication process.** The raw sequencing data, genome assembly, and annotation are available on NCBI, and the manuscript is currently in review. Additional details can be found in the GitHub repo for the genome announcement.

view C25-5 genome announcement GitHub repo here

Table 1: C25-5 assembly summary.

Total number of contigs	16
Total assembly length	30,006,571 bp
Number of contigs >1Mb	8
% bases in >1Mb contigs	97.6%
Contig N50	4,906,038 bp
Contig L50	3
Avg. coverage	57.5x
GC content	48.9%
CDS content	37.2%
Repeat content	4.77%
BUSCO (fungi_odb10)	99.7%
BUSCO (ascomycota_odb10)	98.2%
BUSCO (sordariomycetes_odb10)	93.5%

Figure 5: Genome map of the final, publicly available C25-5 assembly. From outermost track inwards: contig size, contig number (for the eight largest) with flanking telomeric repeats indicated by triangles (outer=5' end, inner=3' end), % repeat content, % CDS content, and % GC content. Plotted using circa.omgenomics.com.

Conclusion

Integrating end-to-end genomics training into graduate orientation

This training framework demonstrates that authentic research experiences can be incorporated into graduate training from day one. By combining instructor-prepared and student-generated sequencing data, **students complete a full genome sequencing and assembly workflow within two weeks**, and develop independent research projects in a low-stakes environment. This approach not only teaches practical skills, but allows students to **contribute meaningful results** and **gain authorship on a research publication** early in their graduate careers.

References

1) Jacobi JD, et al. *Pac Sci*, 2024. **2)** Keith LM, et al. *Plant Dis*, 2015. **3)** Barnes I, et al. *Mol Phylogeny Evol Fungi*, 2018. **4)** Petersen C, et al. *Microb Genomics*, 2022. **5)** Maguvu TE, et al. *Sci Rep*, 2023. **6)** Kolmogorov M, et al. *Nat Biotechnol*, 2019. **7)** Gurevich A, et al. *Bioinformatics*, 2013.

Acknowledgements

We acknowledge the University of California Santa Cruz Genomics Institute for providing support for this project, including use of the Phoenix computational cluster. We also thank Rion Parsons for supporting our use of the Hummingbird computational cluster provided by the University of California Santa Cruz. We acknowledge the USDA-ARS, Daniel K. Inouye U.S. Pacific Basin Agricultural Research Center in Hilo, HI for use of laboratory space, and thank Jon Suzuki, Katelin Branco, and Andrew Paresa for their support. We thank the University of California Santa Cruz Baskin Engineering Lab Support team for providing laboratory space, equipment, and support. We also thank Christopher Vollmers and James Letchinger each for the use of their computers to perform Nanopore sequencing, and Honey Mekonen for technical support. Mention of trademark, proprietary product or vendor does not constitute a guarantee or warranty of the product by the U.S. Department of Agriculture and does not imply its approval to the exclusion of other products or vendors that also may be suitable. **Funding:** UC Santa Cruz, NSF-GRFP, NIH (T32HG012344)

UC SANTA CRUZ | Genomics Institute
NSF NIH Baskin Engineering