# The Genome Sequence of the Marine Diatom *Pseudo-nitzschia australis*

Tu Do (‘16), Cyrus Fenderson (‘16), Melissa Graham (Ph.D. candidate), Sarah Lach (‘17), Alecsander Larin (‘17), Alice Lee (‘16), Laura Moran (‘17), Emily Seibring (‘16), Jacob Steenwyk (Master’s student), Ethan Wainblat (‘16), & William Walker (‘16) (Sponsors: Professors John Gibbons and Deborah Robertson)

CLARK UNIVERSITY

## *Pseudo-nitzschia australis*

- **Marine diatom**
- **Asexual and sexual reproduction**
- **Globally distributed**
- **Domoic acid producer**
- **Amnesic shellfish poisoning**

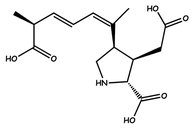




**Figure 1.** ***Pseudo-nitzschia australis***
(Image Credit: Northwest Fisheries Science Center)

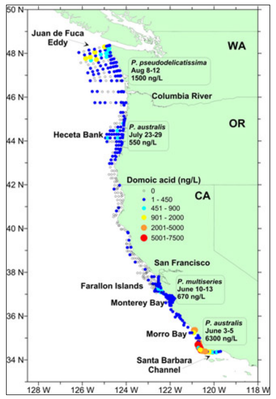



**Figure 2**. Domoic Acid concentrations along US West Coast during 1998.

## *De Novo* Genome Assembly

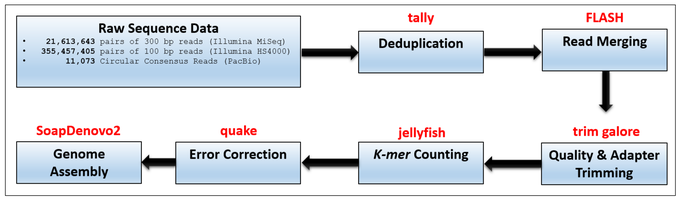



**Figure 3. Genome assembly pipeline.** Raw data was pre-processed to improve the quality of *de novo* genome assembly. We (1) collapsed duplicate paired-end reads, (2) merged overlapping pair-end reads, (3) quality and adapter trimmed reads, (4) quantified the occurrence of 17-mers for genome size estimation and error correction, and (5) corrected or discarded reads with low frequency 17-mers.

| Statistic | K61-81 |
|---|---|
| # Scaffolds | 1,990,571 |
| N50 (bp) | 4,085 |
| N90 (bp) | 120 |
| Longest Scaffold (bp) | 1,517,930 |
| Mean Scaffold Coverage | 27 |
| Cumulative Assembly Size (bp) | 767,504,633 |
| # Scaffolds ≥ 5,000 bp | 26,903 |
| Assembly Size Scaffolds ≥ 5,000 bp | 362,787,643 |

**Table 1.** ***De novo*** **Assembly Statistics.** We used **SOAPdenovo2** for genome assembly, using a *k-mer* range of 61-81. Overall, the assembly was highly fragmented, suggesting high rates of heterozygosity, and/or high levels of repetitive DNA.

## *Pseudo-nitzschia australis* Gene Prediction

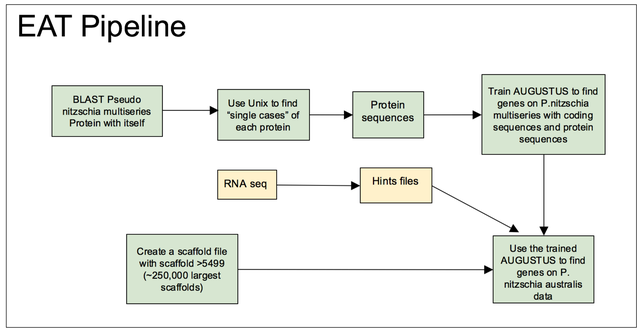



**Figure 4.** ***in silico*** **prediction of** ***Pseudo-nitzschia australis*** **genes.** We used AUGUSTUS to computationally predict genes. The *Pseudo-nitzschia multiseries* genome, proteome, and coding sequence files were used to train AUGUSTUS. Additionally, exonic “hints” were generated by mapping *Pseudo-nitzschia australis* RNA-seq data against the genome assembly. Together, this data was then used to predict protein-coding genes in the *Pseudo-nitzschia australis* genome.

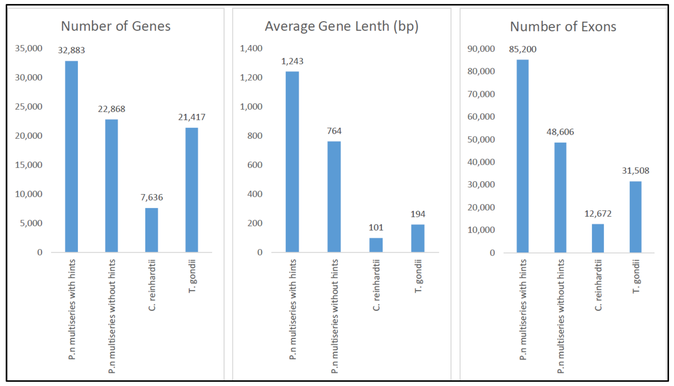



**Figure 5.** Results assessment of AUGUSTUS gene prediction on the largest 250,000 *Pseudo-nitzschia australis* scaffolds using different species-specific training sets (*Pseudo-nitzschia multiseries* (with and without hints), *Chlamydomonas reinhardtii* and *Toxoplasma gondii*).

## *Pseudo-nitzschia australis* Genome Size Estimation

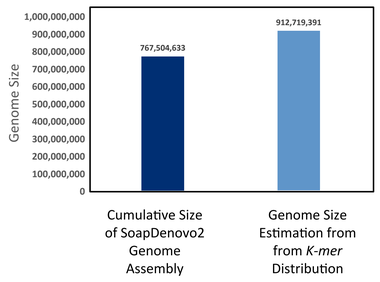



**Figure 6.** The *Pseudo-nitzschia australis* genome size was estimated from the cumulative size of the SoapDenovo2 assembly, and from the frequency distribution of 17-mers from the error-corrected Illumina MiSeq data. Our estimates suggest the *Pseudo-nitzschia australis* genome is between 700-900 Mb.

## Ecological Relevance of Bacterial Sequence Data

- Genes with outlier GC content may represent:
  (1) Contamination from ecologically relevant marine bacterial species
  (2) Genes horizontally transferred from bacteria to *Pseudo-nitzschia australis*

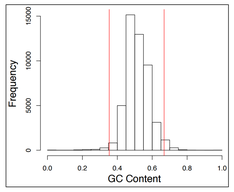



**Fig 7.** GC content distribution of the predicted *Pseudo-nitzschia australis* genes.

| %GC | Query cover | E value | Identity | Accession | Species | Class | Order | Family |
|---|---|---|---|---|---|---|---|---|
| 17.41% | 3% | 2.E-04 | 100% | CP008727.1 | *Burkholderia oklahomensis* | Betaproteobacteria | Burkholderiales | Burkholderiaceae |
| 33.62% | 19% | 8.E-09 | 77% | XM_014048693.1 | *Monorophalium neglectum* | Alphaproteobacteria | Rhizobiales | Methylobacteriaceae |
| 69.30% | 30% | 2.E-53 | 76% | CP000264.1 | *Jannaschia spp.* | Alphaproteobacteria | Rhodobacterales | Rhodobacteraceae |
| 69.35% | 36% | 2.E-64 | 81% | CP002279.1 | *Mesorhizobium opportunistum* | Alphaproteobacteria | Rhizobiales | Phyllobacteriaceae |
| 69.36% | 12% | 4.E-26 | 77% | CP000264.1 | *Jannaschia spp.* | Alphaproteobacteria | Rhodobacterales | Rhodobacteraceae |
| 69.73% | 17% | 4.E-21 | 81% | CP012908.1 | *Ketogulonicigenium vulgare* | Alphaproteobacteria | Rhodobacterales | Rhodobacteraceae |
| 70.11% | 57% | 4.E-61 | 75% | CP014028.1 | *Achromobacter xylosoxidans* | Betaproteobacteria | Burkholderiales | Alcaligenaceae |
| 70.72% | 67% | 4.E-62 | 75% | CP010855.1 | *Marinovum algicola* | Alphaproteobacteria | Rhodobacterales | Rhodobacteraceae |
| 71.28% | 63% | 1.E-106 | 74% | CP000031.2 | *Ruegeria pomeroyi* | Alphaproteobacteria | Rhodobacterales | Rhodobacteraceae |
| 71.49% | 93% | 3.E-83 | 81% | CP012960.1 | *Rhodobacter sphaeroides* | Alphaproteobacteria | Rhodobacterales | Rhodobacteraceae |
| 71.53% | 67% | 8.E-64 | 75% | CP010855.1 | *Marinovum algicola* | Alphaproteobacteria | Rhodobacterales | Rhodobacteraceae |
| 73.01% | 98% | 1.E-116 | 75% | CP000031.2 | *Ruegeria pomeroyi* | Alphaproteobacteria | Rhodobacterales | Rhodobacteraceae |

**Table 2.** Best BLAST hit for the 10 highest and lowest GC content *Pseudo-nitzschia australis* genes against the NCBI non-redundant database. All top hits are of bacterial origin.

## Genome Assembly Improvement

- Long-read PacBio data is being integrated to improve genome assembly through 2 projects:
  (1) The development of a novel scaffolding algorithm
  (2) The evaluation of additional scaffolding software

**Fig 8. The Low Coverage Long Read Scaffolder (LCLRS).** The LCLRS algorithm incorporates a graph-based approach to best path reconstruction and aims to improve scaffolding through the use of low coverage PacBio data.

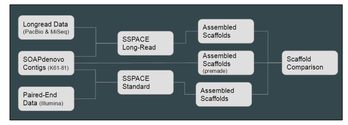
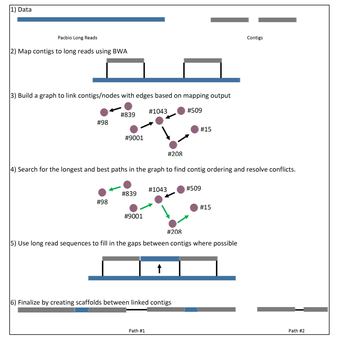
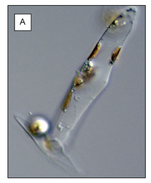



**Fig 9.** Evaluation of SSPACE and SSPACE-Long-Read scaffolding algorithms in comparison to SoapDenovo2.




## Future Directions

- Improvement of assembly quality through optimal scaffolding
- Functional annotation of predicted genes and proteins
- Identification of genomic elements contributing to genome size expansion
- Validation of putative HGT events through the analysis of flanking regions
- Identification of genes of genes involved in domoic acid production




*Trainer et al. 2012*

## Acknowledgments

- This research was conducted during the Spring 2016 semester of The Genome Project (BIOL 209) through the use of the Clark University Supercomputing Cluster. We thank Jason Smith for genomic DNA, Matt Essig and Sarjan Shrestha for computational support, and Antonis Rokas for temporary use of the Vanderbilt University high performance computing cluster (ACCRE). This work was generously supported by the Gordon and Betty Moore Foundation.

CLARK UNIVERSITY

GORDON AND BETTY MOORE FOUNDATION