Understanding Bacterial Evolution and Adaptation in Natural Microbial Environments

Sidaswar Krishnan

Submitted in fulfilment of the requirements for the degree of:
Doctor of Philosophy in Science

Principal Supervisor - Prof. Justin Seymour
Co-supervisor - Prof. Aaron Darling
Co-supervisor - Assoc.Prof. Dominik Beck
Co-supervisor - Dr. Matthew DeMaere

University of Technology Sydney
School of Life Sciences
Climate Change Cluster (C3)

February 2025

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Sidaswar Krishnan declare that this thesis, is submitted in fulfilment of the
requirements for the award of Doctor of Philosophy, in the School of Life Sciences, Faculty of Science at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: 20/02/2025

Acknowledgements

My PhD journey has been one of growth, both personally and as a scientist. I'm grateful to have had the company of many wonderful people who supported me through it. First and foremost, I would like to thank Justin Seymour. Thank you Justin for welcoming me to your group and guiding me to adopt a scientific mindset. Coming from a computer science and bioinformatics software development background, your mentorship helped me apply those skills to effectively ask and answer fundamental biological questions. I'm also grateful for your guidance in developing my scientific writing skills and your patient support during challenging times.

I would also like to thank my co-supervisors, Aaron Darling, Matthew DeMaere, Dominik Beck and Martin Ostrowski. Aaron and Matt, thank you for helping me develop my technical skills and helping me reason out and address challenging problems. Thank you Martin for helping me get familiar with ocean microbiology and helping set up my projects. Thank you Dominik for being my friend and supervisor from the final year of my bachelor's through the honours and PhD, and helping me become a scientist. I would also like to thank John Pimanda who helped me get started on this journey. Thank you all also for our talks and personal support.

During the course of the PhD, I was able to make wonderful friendships. Lilian Hoch, Justin Tierney, Kira Picknell, Bhuwan Ghimire, Chris Songsomboon, Fredrick Jaya, Pierre Bodenes, thank you for your kindness and friendship. Thank you also to all the members of my lab groups, Ocean Microbiology Group and Darling Lab, and the members of C3 for our pleasant chats and exchanges over the years.

I'm deeply grateful to my great uncle Lakshmanan Arunachalam, Vasudevacharya, John Adamedes and my grandparents for their mentorship and support throughout this journey. A special thanks to my longtime friend Awais Nisar for being there for me throughout these years.

None of this would have been possible without my parents. Thank you for your love, support and sacrifices. Thank you for always being there for me.

Thesis Structure

This thesis is presented as a thesis by compilation consisting of published and publishable chapters. It comprises five chapters that cover the rationale, the design of a new bioinformatic platform, its application to environmentally derived samples, and interpretations of the research findings.

Chapter 1 serves as a general introduction, providing the necessary context and background for the three research chapters (Chapters 2–4).

Chapter 2 introduces *Rhometa*, a newly developed bioinformatics pipeline designed to estimate mutation and recombination rates from modern metagenomic datasets. The pipeline leverages a composite likelihood estimator for recombination rate estimation. The work associated with *Rhometa* has been published in *PLoS Genetics* (Krishnan *et al.*, 2023).

Chapter 3 presents the application of *Rhometa* to oceanographic time-series data to investigate inter-species, spatial and temporal patterns in mutation and recombination rates in the ocean.

Chapter 4 expands the analysis to a global scale, examining mutation and recombination rates using data from the Tara Oceans project. This chapter further explores the relationships between evolutionary parameters and environmental variables.

Chapter 5 concludes the thesis by summarising the key findings, discussing the implications of *Rhometa*'s development and its application to ocean bacteria, and provides recommendations for future research.

A combined bibliography for all chapters is included at the end of the document.

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Values to the right of the line represent higher recombination rates relative to mutation. Within each sample, multiple analyses were conducted for different species, and cumulative results were reported per genus. (**A**) ρ/θ values for the genera studied at Port Hacking. (**B**) ρ/θ values for for the genera studied at Maria Island.

Fig 3.2. Port Hacking and Maria Island ρ/θ values for genera by seasons. The vertical dashed line at ρ/θ=1 indicates equal rates of recombination and mutation. Values to the right of the line represent higher recombination rates relative to mutation. Within each sample, multiple analyses were conducted for different species, and cumulative results were reported per genus. (**A**) ρ/θ values for the genera at Port Hacking across seasons. (**B**) ρ/θ values for the genera at Maria Island across seasons.

Fig 3.3. Correlation matrix showing relationships between ρ/θ and environmental variables using combined data from Port Hacking and Mariah Island studies. Analysis includes genera common to both locations. Correlation coefficients are displayed only where statistically significant ($p<0.05$); non-significant correlations appear as black cells.

Fig 3.4. Correlation matrix showing relationships between ρ/θ and environmental variables for Port Hacking and Maria Island. Correlation coefficients are displayed only where statistically significant ($p<0.05$); non-significant correlations appear as black cells. (**A**) Port Hacking correlation matrix. (**B**) Maria Island correlation matrix.

Chapter 4

Fig 4.1. Recombination-to-mutation event rate ratios (ρ/θ) for three microbial genera for global samples. The vertical dashed line at ρ/θ=1 indicates equal rates of recombination and mutation. Values to the right of the line represent higher recombination rates relative to mutation. Each point represents a unique sample. Each point represents a unique sample per genus; points are cumulative across species within each genus.

Fig 4.2. Recombination-to-mutation rate ratios (ρ/θ) for the Pelagibacter genus across global oceanic sites. Each circle represents a sampling location, coloured by ρ/θ: yellow shades indicate higher recombination relative to mutation (ρ/θ >= 1), and blue shades indicate higher mutation relative to recombination (ρ/θ < 1). Each point

represents a unique sample per genus; points are cumulative across species within each genus.

Fig 4.3. Recombination-to-mutation rate ratios (ρ/θ) for the Procholorococcus_A genus across global oceanic sites. Each circle represents a sampling location, coloured by ρ/θ: yellow shades indicate higher recombination relative to mutation (ρ/θ >= 1), and blue shades indicate higher mutation relative to recombination (ρ/θ < 1). Each point represents a unique sample per genus; points are cumulative across species within each genus.

Fig 4.4. Recombination-to-mutation event rate ratios (ρ/θ) for the Synechococcus_C genus across global oceanic sites. Each circle represents a sampling location, coloured by ρ/θ: yellow shades indicate higher recombination relative to mutation (ρ/θ >= 1), and blue shades indicate higher mutation relative to recombination (ρ/θ < 1). Each point represents a unique sample per genus; points are cumulative across species within each genus.

Fig 4.5. Recombination-to-mutation event rate ratios (ρ/θ) for three microbial genera across different ocean regions. Each boxplot represents the distribution of ρ/θ values for a genus in a given region. The horizontal dashed line at ρ/θ = 1 indicates equal rates of recombination and mutation. Values above this line represent a greater contribution of recombination relative to mutation. Brackets with p-values denote statistically significant differences between ocean regions. Each point represents a unique sample per genus; points are cumulative across species within each genus.

Fig 4.6. Correlations between recombination-to-mutation rate ratios (ρ/θ) and environmental variables for three genera. Each coloured tile represents a significant Spearman correlation ($p < 0.05$), with the correlation coefficient indicated numerically. Positive correlations are shown in shades of red, and negative correlations in blue, as indicated by the color scale.

Chapter 5

Fig 5.1. Simplified visualisation of method used to handle varying depths and determine recombination rates in Rhometa.

List of Supplementary Figures

Chapter 2

S2.1 Fig. Results of varying simulated genome lengths for testing LDhat (number of genomes fixed at 10, tract length 500).
https://doi.org/10.1371/journal.pgen.1010683.s001
(TIF)

S2.2 Fig. Comparing simulated single end and paired end read datasets in Rhometa.
(**A**) Single end results. (**B**) Paired end results
https://doi.org/10.1371/journal.pgen.1010683.s002
(TIF)

S2.3 Fig. Simulated vs Estimated population recombination rate (ρ) results for Rhometa. Results for varying numbers of simulated genomes and fold coverage values for population recombination rates 0.0, 0.1, 0.2, 0.3, 0.4, 0.5. The simulation parameters used are the same as for population recombination rates [10.0, 20.0, 30.0, 40.0, 50.0], except lookup tables for population recombination rates 0-2 were used (0-2 in 201 steps) for depths of 3-200.
https://doi.org/10.1371/journal.pgen.1010683.s003
(TIF)

S2.4 Fig. Simulated vs Estimated theta per site (θ) results for LDhat and Rhometa. (**A**) LDhat. Simulated vs Estimated theta per site (θ) for varying number of simulated bacterial genomes. (**B**) Rhometa. Simulated vs Estimated theta per site (θ) for varying number of simulated bacterial genomes.
https://doi.org/10.1371/journal.pgen.1010683.s004
(TIF)

S2.5 Fig. Results of analyzing simulated datasets with lookup tables generated under misspecified θ (genome length fixed at 100,000) (**A**) Simulated datasets analyzed with true θ (0.01) lookup tables (**B**) Simulated datasets analyzed with misspecified θ (0.005) lookup tables (**C**) Simulated datasets analyzed with misspecified θ (0.02) lookup tables (**D**) Simulated datasets analyzed with misspecified θ (0.1) lookup tables. https://doi.org/10.1371/journal.pgen.1010683.s005 **(TIF)**

S2.6 Fig. Results of analyzing simulated datasets with misspecified tract lengths (genome length fixed at 100,000. θ 0.01 lookup tables used**)** (**A**) Simulated datasets analyzed with true tract length 1000 bp (**B**) Simulated datasets analyzed with misspecified tract length 500 bp (**C**) Simulated datasets analyzed with misspecified tract length 2000 bp.
https://doi.org/10.1371/journal.pgen.1010683.s006
(TIF)

S2.7 Fig. Mcorr tested using the simulated datasets used for rhometa (Figs 2.3,2.4). Mcorr uses bam (aligned reads) and gff (gene annotations) files as inputs. For the gene annotations, coding regions need to be provided for mcorr. With the simulated data every 1000 bp was defined as a coding region (CDS). Out of 4000 simulations, some could not be completed and many extremely large unrealistic realistic values. A filter was used where such instances were discarded and only phi_pool values < 100 were kept. After filtering there were only 2386/4000 values.
https://doi.org/10.1371/journal.pgen.1010683.s007
(TIF)

Chapter 3

S3.1 Fig. Port Hacking, pileup mean depth vs ρ and θ. After all filters applied. (**A**) Pileup mean depth vs ρ (**B**) Pileup mean depth vs θ.

S3.2 Fig. Maria Island, pileup mean depth vs ρ and θ. After all filters applied. (**A**) Pileup mean depth vs ρ (**B**) Pileup mean depth vs θ.

S3.3 Fig. Port Hacking and Maria Island variable counts. (**A**) Variable counts for Port Hacking (**B**) Variable counts for Maria Island.

S3.4 Fig. Port Hacking ρ and θ values. (A) ρ values for Port Hacking Genera. **(B)** θ values for Port Hacking Genera.

S3.5 Fig. Maria Island ρ and θ values. (A) ρ values for Maria Island Genera. **(B)** θ values for Maria Island Genera.

S3.6 Fig. Port Hacking ρ and θ values for genera by seasons. (**A**) ρ values for Port Hacking genera by seasons. (**B**) θ values for Port Hacking genera by seasons.

S3.7 Fig. Maria Island ρ and θ values for genera by seasons. (**A**) ρ values for Maria Island genera by seasons. (**B**) θ values for Maria Island genera by seasons.

S3.8 Fig. Port Hacking correlations against environmental variables, only significant correlations (p<0.05) are displayed. (**A**) ρ correlations. (**B**) θ correlations.

S3.9 Fig. Maria Island correlations against environmental variables, only significant correlations (p<0.05) are displayed. (**A**) ρ correlations. (**B**) θ correlations.

Chapter 4

S4.1 Fig. Genus Pelagibacter, pileup mean depth vs ρ and θ. After all filters applied. (**A**) Pileup mean depth vs ρ (**B**) Pileup mean depth vs θ.

S4.2 Fig. Genus Prochlorococcus_A, pileup mean depth vs ρ and θ. After all filters applied. (**A**) Pileup mean depth vs ρ (**B**) Pileup mean depth vs θ.

S4.3 Fig. Genus Synechococcus_C, pileup mean depth vs ρ and θ. After all filters applied. (**A**) Pileup mean depth vs ρ (**B**) Pileup mean depth vs θ.

List of Tables

Chapter 2

Chapter 3

List of Supplementary Tables

Chapter 2

S2.1 Table. Analysis results of s_pnemoniae transformation. Results for all seed values
https://doi.org/10.1371/journal.pgen.1010683.s008 **(DOCX)**

S2.2 Table. *S.pneumoniae* experiment accession codes
https://doi.org/10.1371/journal.pgen.1010683.s009
(DOCX)

S2.3 Table. Accession codes for Port Hacking, Sydney datasets (from SRA run table)
https://doi.org/10.1371/journal.pgen.1010683.s010
(DOCX)

Chapter 3

S3.1 Table. ρ / θ for common genera by location comparisons using Pairwise Wilcoxon test with Bonferroni correction.

S3.2 Table. Port Hacking and Maria Island, ρ/θ for genus by season comparisons using Kruskal-Wallis test. p values for significance are reported.

S3.3 Table. Port Hacking ρ/θ mean values for genera by seasons

S3.4 Table. Maria Island ρ/θ mean values for genera by seasons, standard deviation in brackets.

S3.5 Table. Port Hacking mean values for environmental parameters by season and genera, standard deviation in brackets

S3.6 Table. Maria Island mean values for environmental parameters by season and genera, standard deviation in brackets

S3.7 Table. Port Hacking and Maria Island Species Lists

https://doi.org/10.5281/zenodo.14892007
(CSVS)

Chapter 4

S4.1 Table. Mean ρ/θ values for genus, region groups

S4.2 Table. Species lists for the genera targeted
https://doi.org/10.5281/zenodo.14892031
(CSVS)

List of Additional Files

Chapter 2

S2.1 Appendix Lookup configuration
Matching against the lookup table
https://doi.org/10.1371/journal.pgen.1010683.s011
(DOCX)

Data Availability: All supporting information can be accessed here: https://doi.org/10.5281/zenodo.7634208. The Rhometa software package is available at https://github.com/sid-krish/Rhometa. The metagenomic dataset simulation pipeline is available at https://github.com/sid-krish/rhometa_sim, the LDhat Nextflow Pipeline is available at: https://github.com/sid-krish/Nextflow_LDhat, the full genome version of Rhometa, developed for testing purposes, is available at https://github.com/sid-krish/Rhometa_Full_Genome and the Nextflow_LDhat_sim simulation pipeline (used for simulating full sequences for both Rhometa Full Genome and LDhat Nextflow Pipeline) is available at: https://github.com/sid-krish/Nextflow_LDhat_Sim.

Chapter 3

Data Availability:
The AusMicrobiome Microbial Ocean Atlas project metadata can be accessed via: https://github.com/AusMicrobiome/microbial_ocean_atlas/blob/57dfad6e30359a406ae86208fe554fac429269e6/data/oceanViz_AM_data.csv .

All metadata, relevant scripts and results generated, raw analysis files with metadata merged, are separated and accessible via: https://doi.org/10.5281/zenodo.15907061

Chapter 4

Data Availability:
All metadata, relevant scripts and results generated, raw analysis files with metadata merged, are separated and accessible via: https://doi.org/10.5281/zenodo.16408197

Thesis Abstract

Microbial evolution is underpinned by the processes of mutation and recombination, yet the dynamics of these processes within natural microbiomes remain poorly characterised. The growing application of metagenomics to study prokaryotes in their native habitats presents an excellent opportunity for learning more about recombination and mutation. However, a major bottleneck relates to the paucity of tools that can precisely measure recombination and mutation rates using next-generation sequencing datasets. This thesis aims to explore evolution and adaptation in natural microbial environments by developing and implementing a robust and computationally efficient method for estimating mutation and recombination rates in prokaryotic populations using metagenomic datasets.

In this thesis, I present Rhometa, a new software package for estimating recombination rates from shotgun sequencing reads of metagenomes. Rhometa builds upon the composite likelihood approach, recognised as one of the most accurate methods for estimating population recombination rates, and adapts it to support the analysis of modern short-read datasets. It can be directly operated on reads aligned to representative genomes, and supports the widely used Binary Alignment Map (BAM) format. Although the composite likelihood estimator has demonstrated strong performance with other types of sequence data, its application to metagenomic datasets has been limited. Rhometa represents a new and reliable tool for analysing metagenomic data to calculate recombination rates. In addition, Rhometa is also able to perform mutation rate estimates, making it a comprehensive tool for studying prokaryotic evolution.

Rhometa was validated using simulated short-read datasets and data from a previously published *S. pneumoniae* transformation experiment, and it was further tested on metagenomic data from ocean surface water samples. The results confirmed that Rhometa delivers accurate and reliable recombination estimates for metagenomic datasets. On simulated datasets, it consistently demonstrated strong performance, with accuracy improving as the number of genomes increased. For the transformation experiment dataset, Rhometa produced estimates within the expected range, whereas other tools generated highly improbable values.

After validating Rhometa, I applied it to investigate recombination and mutation dynamics within natural marine microbial communities. The ocean, which covers

approximately 70% of the Earth's surface, contributes to half of the global primary production (Azam and Malfatti, 2007). In addition, the microbes within the ocean account for 70% of the total biomass (Logares, 2024). Despite their ecological importance, our understanding of mutation and recombination in marine microbes remains limited. However, the availability of extensive metagenomic datasets for ocean bacteria, combined with Rhometa's capabilities, provides an opportunity to gain valuable insights into their evolutionary processes.

I first used Rhometa to interrogate metagenomic datasets from two oceanographic time-series sites in eastern Australia, where I assessed differences in the recombination-to-mutation event ratios (ρ/θ) among ecologically important marine bacterial lineages, namely Cyanobiaceae, Pelagibacteraceae, and Rhodobacteraceae. Through this analysis, I evaluated seasonal variation and examined environmental influences on ρ/θ. I found that some genera belonging to these lineages, such as *Pelgibacter,* tend to have high ρ/θ values, which remained relatively stable between seasons and locations. In contrast, other bacteria, such as *Synechococcus,* are characterised by ρ/θ values that are highly variable across seasons and locations, sometimes to such an extent that the dominant force of evolution switches between recombination and mutation. I was also able to identify a suite of environmental parameters, including temperature, nutrient levels, and daylight, which exhibited significant correlations with the ρ/θ values of specific bacteria, highlighting the role of the environment in shaping prokaryote evolutionary processes.

I then used Rhometa to perform a global scale analysis of marine bacterial recombination and mutation dynamics, using metagenomic datasets derived from the Tara Oceans project. I evaluated global recombination-to-mutation event ratios (ρ/θ) for important genera of marine bacteria, including *Pelagibacter*, *Prochlorococcus*, and *Synechococcus* and identified environmental factors that influence these processes on a global scale. This analysis revealed significant differences in ρ/θ among the three genera, with *Pelagibacter* exhibiting the highest ρ/θ, followed by *Prochlorococcus* and *Synechococcus*. Again our results demonstrated that recombination has a high relative importance in the highly abundant heterotrophic bacteria *Pelagibacter.* I also noticed substantial variation in ρ/θ across different ocean regions. The highest ρ/θ values for *Pelagibacter* and *Synechococcus* were found in productive waters, which experience significant mixing and support larger bacterial populations, increasing the potential for cell-to-cell interactions and recombination. The influence of environmental factors on ρ/θ was highly genus-specific, with ρ/θ in *Pelagibacter* positively correlated with by

chlorophyll a, while in Prochlorococcus ρ/θ was most strongly correlated with temperature.

This thesis has delivered a new capacity for quantifying mutation and recombination rates from metagenomic datasets, which has opened the door to understanding these important evolutionary forces in natural microbiomes. Using this new tool, I was able to demonstrate the heterogeneous dynamics of recombination and mutation within the marine microbiome, and highlighted the complex interplay between evolutionary processes and environmental conditions that shape marine bacterioplankton communities. Taken together, the outcomes of this thesis provide an unprecedented view into the dynamics of fundamental evolutionary processes on local and global scales.

Chapter 1: General Introduction

1.1 Prokaryotes and their key evolutionary processes

Prokaryotes comprise a significant fraction of the biomass on Earth (Bar-On, Phillips and Milo, 2018) and play key roles in defining the productivity and biogeochemistry of all natural ecosystems (Whitman, Coleman and Wiebe, 1998; Falkowski, Fenchel and Delong, 2008). They also establish critical ecological relationships with all plants and animals, including humans, whereby host-associated microbiomes fundamentally influence the fitness and health of host organisms (McFall-Ngai *et al.*, 2013; Zhang *et al.*, 2015; Qadir *et al.*, 2024). Therefore, understanding the ecological and evolutionary dynamics of these microorganisms has importance across diverse fields, spanning medicine, agriculture and environmental health (Hayat *et al.*, 2010; Akoijam, Kalita and Joshi, 2022). Within prokaryotes, evolution is strongly governed by the processes of mutation and recombination (Hanage, 2016), factors which play a crucial role in adaptation leading to speciation (Lawrence, 1999). However, our understanding is limited by the lack of tools which can quantify these factors as they occur in natural environments. Within this context, the goal of this thesis is to develop a tool that effectively meets this need and use it to study bacteria from the natural environment.

1.1.1 Mutation and its role in bacterial evolution

Mutations are changes in the nucleotide sequence within a cell's DNA, resulting from molecular changes that are not repaired by the cellular repair systems (Hershberg, 2015). Mutation is fundamental to evolution in prokaryotes, as evolutionary processes are driven by genomic variation originating from mutations (Horton and Taylor, 2023). Mutations can result from errors in cellular processes such as DNA replication, transcription, recombination, and repair, as well as from external influences, including environmental stressors such as nutrient limitation, temperature extremes or exposure to antibiotics (Horton and Taylor, 2023). Based on their impact on genomic fitness, mutations are categorised as deleterious, neutral, or advantageous (Eyre-Walker and Keightley, 2007).

A comprehensive understanding of the processes governing mutation and its effects is essential for comprehending evolution within prokaryotes. Moreover, it has practical implications, for example, antibiotic resistance often arises from mutations, reducing the efficacy of antibiotic-based treatments (Woodford and Ellington, 2007). By studying

how mutations arise and are maintained, we can better predict evolutionary outcomes and develop strategies to mitigate harmful consequences, such as the spread of drug-resistant infections (Woodford and Ellington, 2007). Additionally, understanding the role of mutations can provide insight into how species adapt to environmental changes (Fitzgerald and Rosenberg, 2019), such as climate shifts, thereby supporting ecosystem resilience in the face of global challenges. The rate at which mutations occur in natural environments remains poorly understood, a knowledge gap this thesis addresses by introducing a method to quantify mutation rates in microbial populations within their natural habitat.

1.1.2 Recombination and its role in bacterial evolution

Prokaryotes can exchange genetic material between and within populations through a process known as recombination (Didelot and Maiden, 2010). Bacteria can exchange DNA with distantly related species, enabling the acquisition of beneficial genetic material that can aid in adapting to changing environments (Sun and Luo, 2018). Along with mutation, recombination is a major driving force in bacterial evolution. They work in tandem with recombination helping spread beneficial mutations within a population (Thomas and Nielsen, 2005) (Fig 1.1).

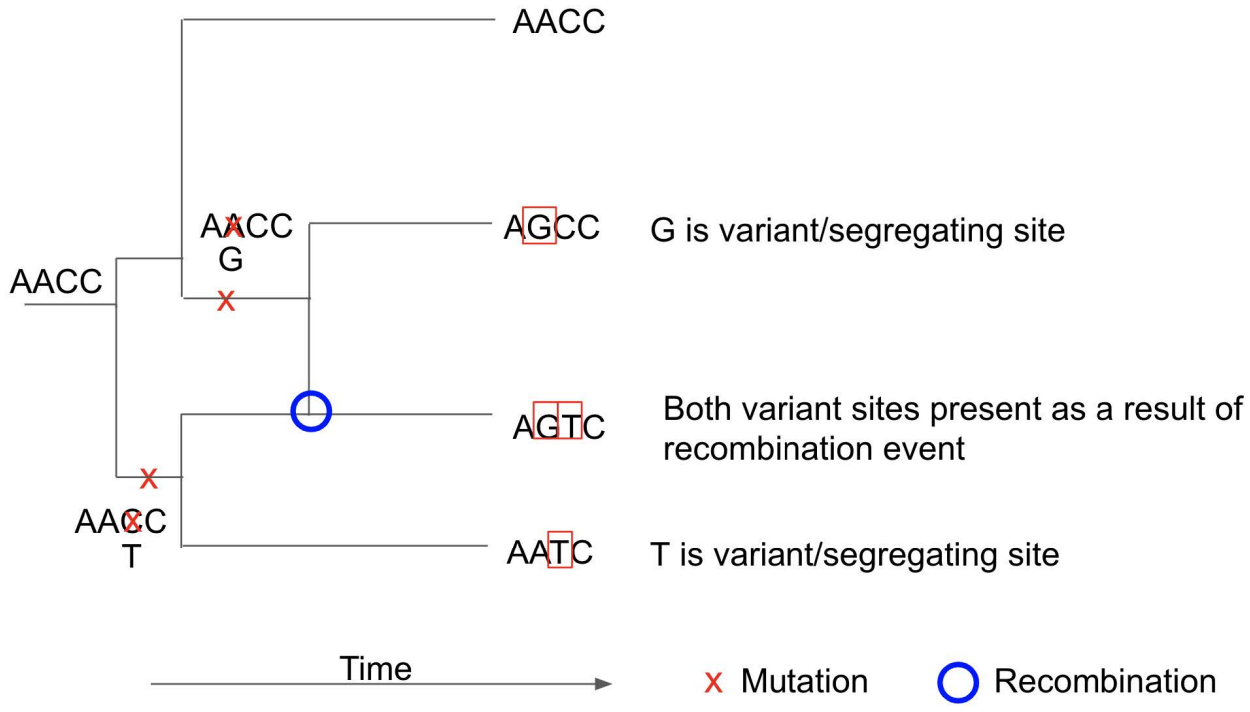


Fig 1.1. Recombination example. Species divergence as depicted through a phylogenetic tree. A sequence belonging to an initial species undergoes mutations as it

evolves and diverges. Over time species with different mutations exchange DNA through the process of recombination, as a result a species now has two different mutations that occurred in two different species.

Recombination among bacteria can occur through three discrete processes, namely transformation, transduction, and conjugation (Didelot and Maiden, 2010). Unlike in sexual eukaryotes, where recombination is of the form of crossing over which occurs during meiosis, bacterial recombination has greater similarities to gene conversion (Hanage, 2016). The key difference between these being that gene conversion involves non-reciprocal exchange of genetic material, while crossing over involves the reciprocal exchange of it (Guirouilh-Barbat *et al.*, 2014).

Transformation entails the uptake and integration of naked DNA from the environment (Thomas and Nielsen, 2005). It requires a specialised membrane system with proteins that help bring DNA into the cell (Claverys, Martin and Polard, 2009). A physiological state of competence, involving 20-50 proteins, is required and is limited and dependent upon various environmental conditions such as nutrient access and cell density (Thomas and Nielsen, 2005; Popa and Dagan, 2011).

Transduction involves the transfer of genetic material from a previous bacterial host into a recipient via bacteriophages and subsequent integration (Arnold, Huang and Hanage, 2022). This process necessitates the infection of a bacterial cell by a virus, it is the longest range transfer mechanism as a result of protection and transport of DNA by the virus (Majewski, 2001; Popa and Dagan, 2011). This method protects the DNA from environmental degradation before it is transferred and does not require direct proximity.

Finally, conjugation is the predominant mechanism of DNA transfer between bacteria, and relies on the direct transfer of DNA between cells (Popa and Dagan, 2011). This process involves the transfer of plasmids through conjugative pili and the subsequent uptake by a recipient cell (Brito, 2021), it requires the donor and recipient to be close enough for direct contact to form a conjugative tunnel (Popa and Dagan, 2011; Brito, 2021; Arnold, Huang and Hanage, 2022).

Recombination is now recognised as a widespread phenomenon among bacteria, with evidence that this form of genetic exchange often plays a more significant role than mutation in the evolution of many bacterial species (Vos and Didelot, 2009; Didelot and

Maiden, 2010). While mutation serves as the source of evolutionary change, recombination facilitates the dissemination of these changes within a population, including the spread of adaptive mutations (Wiedenbeck and Cohan, 2011). This dynamic is particularly relevant in the context of antibiotic resistance, where recombination aids in propagating resistance adaptations across bacterial populations (Wiedenbeck and Cohan, 2011).

Recombination, like mutation, can be influenced by various environmental and ecological factors. These include donor-recipient similarity barriers, where gene transfer is more common among closely related species, ecological barriers, which depend on the physical proximity of donor and recipient species and functional barriers, which affect the retention and integration of exchanged DNA (Popa and Dagan, 2011). Additionally, environmental conditions such as nutrient availability and abiotic factors like temperature, pH, and oxygen levels can also impact recombination frequency, for example it has been shown that the rate of conjugal plasmid transfer in soil varies depending on temperature, further stress conditions may also drive recombination (Aminov, 2011; Le *et al.*, 2020).

1.2 Studying mutation and recombination rates in microbes

Much of our current understanding of microbiology stems from culture-based studies (Handelsman, 2004), including our understanding of mutation and recombination. A landmark comparative analysis (Vos and Didelot, 2009) examined a range of papers that utilised Multi Locus Sequence Typing (MLST) (Maiden *et al.*, 1998) to investigate recombination and mutation in bacteria and archaea by sequencing specific genomic regions. Through these MLST-based studies, researchers inferred patterns of recombination and mutation rates, yielding a nuanced perspective on how microbial populations evolve and diversify. However, such studies are not without limitations.

Due to the inherent limitations of traditional culture based studies, such as MLST (Maiden *et al.*, 1998), they can often fail to capture the ecological interactions and environmental pressures that influence bacterial populations in situ (Wooley, Godzik and Friedberg, 2010). Laboratory conditions rarely replicate the intricate interplay of factors, such as nutrient availability, community interactions, and phage predation, that shape microbial ecosystems in nature. Further most microbes are not readily culturable (Handelsman, 2004). Therefore, without direct observation of microbes in their natural

habitats, the extent to which mutation and recombination contribute to microbial diversity and adaptation remains largely speculative.

1.3 Metagenomics and the uncultured majority

Given the many factors influencing mutation and recombination, studying these processes within their natural context is crucial. A significant, and relatively recent advancement in the field of microbiology has been the development of metagenomics. A discipline that involves analysing sequence data derived directly from environmental samples, with the term "metagenome" referring to the collective environmental genetic material from a microbiome (Handelsman *et al.*, 1998; Wooley, Godzik and Friedberg, 2010; Thomas, Gilbert and Meyer, 2012). Metagenomics overcomes the limitations of traditional culture-based methods, most notably the fact that most microbes within the environment are unculturable (Handelsman, 2004; Singh *et al.*, 2009). Techniques such as shotgun sequencing involve the extraction of DNA from all cells within a community, fragmenting it and directly sequencing these fragments (Sharpton, 2014). The resultant sequence reads, representing the genetic material of numerous community members, can then be processed and studied in detail (Sharpton, 2014). For example, large-scale initiatives such as the Sorcerer II Global Ocean Sampling (Rusch *et al.*, 2007) and Tara Oceans (Sunagawa *et al.*, 2020), have applied metagenomic techniques that lead to the discovery of microbes and their functions across the oceans, similarly among many other applications, it is also helping us understand the human gut microbiome and its function (Armour *et al.*, 2019). Within these metagenomic reads, which represent the genomes of the microbes in the community, lies the evidence of recombination and mutation as it naturally occurred in the environment. The ability to decipher these patterns offers a powerful means to understand how microbes evolve under varying natural environmental conditions. Publicly there are over 200,000+ metagenomics datasets that are currently available (Martiny *et al.*, 2022), a number that is ever growing, that offer the potential to study recombination and mutation rates in microbes.

1.4 Methods used to measure recombination and mutation

Quantifying mutation and recombination rates within prokaryotic populations provides essential insights into the biological processes underpinning their adaptation and survival. Coalescent theory provides a powerful framework for inferring past evolutionary events by analysing a sample from the present population. This model incorporates the stochastic processes of recombination, mutation, and coalescence of

lineages, offering a coherent statistical approach to studying genetic polymorphism (Fu and Li, 1999; Rosenberg and Nordborg, 2002). From this theoretical foundation the population rates of mutation and recombination are derived as $\theta = 4N_e u$ and $\rho = 4N_e r$ respectively, where *u* is the per-generation mutation rate, *r* represents the per-generation recombination rate, and N_e the effective population size (Fu and Li, 1999; McVean, Awadalla and Fearnhead, 2002; Stumpf and McVean, 2003). These parameters provide a quantitative means to link evolutionary processes with genetic diversity and population structure.

The population mutation rate θ can be estimated using statistical methods based on the coalescent theory. These include widely used approaches such as Watterson's estimator (Watterson, 1975), as well as more advanced methods, including those employing maximum likelihood (Fu and Li, 1999). The Watterson estimator is a well known method that can be applied to sequence data to get an estimate of the rate of mutation. It works by considering variant sites which can occur along the DNA sequence of an organism due to mutations and is evaluated against the number of individuals in a sample to get a population mutation rate estimate, that is the rate at which mutation occurs across the genome within a population (Watterson, 1975; McVean, Awadalla and Fearnhead, 2002). While the Watterson estimator uses just the number of variant sites, alternative approaches like likelihood methods use more of the genetic data and estimate the mutation rate that makes the observed data most likely.

Several approaches have been used to estimate the population recombination rate, ρ, including moment estimators, full-likelihood estimators, and composite likelihood estimators (Stumpf and McVean, 2003). The primary methods include, moment estimators which use summary statistics to estimate ρ but are limited in accuracy as they cannot utilise all available genetic data (Fearnhead and Donnelly, 2001, 2002). In simple terms, moment estimators use summary statistics calculated from the data to estimate unknown parameters like the Watterson estimate for θ described earlier. Modern approaches embed these summary statistics within more sophisticated frameworks, including Bayesian and regression-based methods for parameter estimation (Arenas *et al.*, 2015; Hermann *et al.*, 2019).

Full-likelihood estimators, on the other hand, leverage the entire genetic information but are computationally impractical for large datasets (Fearnhead and Donnelly, 2002). To address these limitations, composite likelihood estimators were developed, these

methods analyse reduced subsets of data, such as pairs of alleles, to decrease computational intensity (Stumpf and McVean, 2003). These techniques are dataset agnostic but the programs applying them were initially implemented for aligned sequences. It is possible to adapt these approaches such as the composite likelihood estimator which strikes a balance between accuracy and computational tractability, to work with metagenomic reads. This would require careful consideration of the complexities associated with such datasets, such as the vast amount of short read data from the many microbes present in the environment and the varying coverage of these reads. A central objective of this thesis is to adapt the composite likelihood estimator for metagenomic datasets.

Methods that implement the composite likelihood estimator for recombination rate estimation include LDhat (McVean, Awadalla and Fearnhead, 2002; Auton and McVean, 2007), which is the most well known and widely used pipeline. LDhat works with sequence data from both eukaryotic and prokaryotic organisms and uses the Watterson estimator to estimate the mutation rate. For recombination, LDhat examines configurations of pairs of variant sites to determine the recombination rate that best explains the observed patterns. It uses precomputed likelihood tables for each possible pairwise configuration and a range of recombination rates, allowing for efficient comparison and identification of the most likely recombination rate. For both the mutation and recombination rates it provides an overall estimate calculated across all genomic sites.

Further iterations of this approach have subsequently been produced, including LDhelmet (Chan, Jenkins and Song, 2012), which includes additions and improvements to the model introduced by LDhat and allows for the detection of recombination hotspots but is only applicable to eukaryotic sequences. LDhot (Auton, Myers and McVean, 2014), from the original developers of LDhat includes the ability to identify hotspots but is again limited to eukaryotes. A more recently introduced application, pyrho, (Spence and Song, 2019) pushes the frontier of composite likelihood estimators. Unlike LDhat, it adopts a penalised likelihood approach, which improves performance and accuracy. However, like LDhat, it remains limited to eukaryotic genomes, making it unsuitable for metagenomic analysis.

PIIM (Johnson and Slatkin, 2006, 2009), was a groundbreaking tool for handling metagenomic read data, using the composite likelihood estimator. It is largely based on the LDhat model but adapted for read data, incorporating methods to account for

missing data resulting from variable read coverage. The program was introduced at a time when data was scarce and expensive and includes computationally intensive methods to account for such scenarios. However, today, with affordable and accurate sequencing, it's easier to discard low-quality data than to process it. PIIM is also outdated regarding the data formats it handles, relying on the obsolete ACE format and lacking support for modern formats like BAM. Interestingly, PIIM also includes a maximum likelihood estimator for mutation rate estimation but this is also inapplicable for modern datasets.

Conversely, mcorr (Lin and Kussell, 2019) is a more recent program that differs fundamentally from LDhat-style approaches, it is not likelihood-based, does not directly estimate the conventional population genetic parameters ρ (recombination rate) and θ (mutation rate). It relies on a correlation-based summary statistic approach and derives the relative rates of recombination and mutation, as well as the recombinational and mutational divergence, from metagenomic datasets. The program is limited to coding regions, meaning it does not consider all the regions of the genome, as performed with established composite likelihood estimators, even if it is only for variant sites. Further, it has been demonstrated that mcorr produces unreliable results on some datasets, with unrealistic estimates of the ratio of recombination to mutation sometimes resulting (Krishnan *et al.*, 2023).

It is possible to assemble shotgun metagenomic reads into representative genomes, known as metagenome-assembled genomes (MAGs) (Bowers *et al.*, 2017). These MAGs can then be directly used with tools such as LDhat. However, it is known that MAGs can contain chimeric sequences (Orakov *et al.*, 2021). Since identifying recombination involves examining subtle variations in the genome, chimeric sequences will be a source of error. Further, aligning complete sequences for analysis is a complex and computationally intensive process which limits its applicability. Therefore, working with metagenomic reads, rather than MAGs, is arguably the best approach to study recombination and mutation of microbes in natural environments. This thesis develops and applies a mutation and recombination rate estimator for natural microbial assemblages from metagenomic read datasets.

1.5 Examining recombination and mutation dynamics in natural ecosystems

Accurate quantification of recombination and mutation rates from metagenomic datasets would open the door to studying these important processes within complex microbiomes within natural ecosystems. This in turn will provide new capacity for understanding how the abiotic and biotic features of natural environments shape these processes. Environmental factors play an important role in influencing mutation and recombination (Aminov, 2011; Le *et al.*, 2020; Horton and Taylor, 2023) which in turn play a crucial role in adaptation (Lawrence, 1999). Importantly these processes are deeply interconnected, shaping microbial evolution through a dynamic interplay (Wiedenbeck and Cohan, 2011). If bacteria relied solely on random mutation, their populations would be largely clonal, however, this is rarely the case, and is not common in nature (Didelot and Maiden, 2010; Shapiro, 2016). The balance between mutation and recombination, driven by environmental pressures and ecological contexts, making their interplay a fundamental aspect of microbial evolution (Guttman and Dykhuizen, 1994; Gogarten and Townsend, 2005; Didelot and Maiden, 2010; Wani *et al.*, 2022).

Given the importance of this dynamic, the ratios ρ/θ and r/m are often considered. Where the ratio ρ/θ represents the probability of a recombination event relative to a mutation event and r/m represents the probability that a nucleotide was substituted due to either recombination or mutation (McVean, Awadalla and Fearnhead, 2002; Vos and Didelot, 2009; Didelot and Wilson, 2015). While mutation and recombination are distinct processes, recombination can lead to high r/m values by introducing alleles that are new to a genome through the transfer of existing variants from another chromosome or cell. High r/m values may occur even when ρ/θ is low if recombination events bring in many such variants. Conversely, ρ/θ can be high even when r/m is low if recombination events are frequent but cause only minor changes. The ratio ρ/θ can be interpreted as follows: values less than 1 ($\rho/\theta < 1$) indicate a higher frequency of mutation events compared to recombination events, whereas values greater than 1 ($\rho/\theta > 1$) suggest that recombination is relatively more significant than mutation.

Studies examining the ratio of recombination to mutation in bacteria and archaea across different environments and species have revealed examples of populations that are nearly clonal to those that are highly recombinant (Vos and Didelot, 2009; González-Torres *et al.*, 2019). Notably, research has shown that microbes in extreme

environments often experience a higher frequency of recombination events compared to those in more stable conditions (Li *et al.*, 2014), suggesting that recombination may serve as a crucial mechanism for adaptation in harsh habitats. Another study has shown that influence of recombination to mutation was higher in marine ecosystem for SAR11, one of the most abundant marine bacteria (Morris *et al.*, 2002; Giovannoni, 2017), but lower in freshwater in ecosystems (Zaremba-Niedzwiedzka *et al.*, 2013), indicating environmental and ecological impact on the impact of recombination to mutation.

Our understanding of the dynamics of the evolutionary processes shaping microbial communities in natural environments and how they vary across space and time remains in its early stages. Metagenomics, along with tools that estimate mutation and recombination rates, provides a potentially powerful new means to unravel these dynamics. Nowhere is this more crucial than in the ocean, which covers 70% of Earth's surface and drives half of global primary production (Azam and Malfatti, 2007). Oceanic microbes, comprising 70% of total biomass (Logares, 2024), are fundamental to biogeochemical cycles and carbon sequestration. Yet, despite their vast influence, we know remarkably little about their diversity, evolution, and ecological functions. Advancing our understanding of the evolutionary drivers shaping natural microbial communities is essential to grasping the broader mechanisms that govern life on Earth.

1.6 Aims of Thesis

The main aim of this thesis was to quantify recombination and mutation rates in natural microbiomes through the development and application of a new bioinformatic pipeline capable of calculating these values from shotgun metagenomic data. With this new tool in hand, the next goal was to deliver a heightened understanding of microbial evolution and adaptation in complex natural environments such as the ocean.

Aim 1

Develop a robust and computationally efficient method for estimating mutation and recombination rates that can be applied to metagenomic datasets for studying prokaryotic populations.

Aim 2

Apply my newly developed program to investigate mutation and recombination rates in major groups of marine bacteria and identify inter-organism, geographic and seasonal

variations in these parameters, to ultimately assess the impact of environmental factors on the evolution of the marine microbiome.

Aim 3

Quantify global ocean-scale patterns in recombination and mutation rates among key marine bacteria and identify environmental features influencing these processes.

Chapter 2:
Rhometa: Population recombination rate estimation from metagenomic read datasets

Published in: PLOS Genetics

Sidaswar Krishnan[1], Matthew Z. DeMaere[3], Dominik Beck[2], Martin Ostrowski[1], Justin R. Seymour[1], Aaron E. Darling[3,4]

[1]Climate Change Cluster, Faculty of Science, University of Technology Sydney, Sydney, NSW, Australia.

[2]Centre for Health Technologies and the School of Biomedical Engineering, University of Technology Sydney, Sydney, NSW, Australia.

[3]Australian Institute for Microbiology & Infection, University of Technology Sydney, Sydney, NSW, Australia

[4]Illumina Australia Pty Ltd, Ultimo, NSW, Australia

Corresponding Author: Matthew Z. DeMaere, email: Matthew.DeMaere@uts.edu.au

Author Contributions

Conceptualization: Matthew Z. DeMaere, Aaron E. Darling.

Data curation: Sidaswar Krishnan.

Formal analysis: Sidaswar Krishnan, Matthew Z. DeMaere.

Funding acquisition: Aaron E. Darling.

Methodology: Sidaswar Krishnan, Matthew Z. DeMaere, Aaron E. Darling.

Project administration: Justin R. Seymour, Aaron E. Darling.

Resources: Aaron E. Darling.

Software: Sidaswar Krishnan, Matthew Z. DeMaere, Aaron E. Darling.

Supervision: Matthew Z. DeMaere, Dominik Beck, Justin R. Seymour, Aaron E. Darling.

Validation: Sidaswar Krishnan, Matthew Z. DeMaere, Aaron E. Darling.

Visualization: Sidaswar Krishnan, Martin Ostrowski.

Writing – original draft: Sidaswar Krishnan.

Writing – review & editing: Matthew Z. DeMaere, Dominik Beck, Martin Ostrowski, Justin R. Seymour, Aaron E. Darling.

Author Signatures

By signing here the authors agree that the contributions are listed correctly

Sidaswar Krishnan: Production Note:
Signature removed prior to publication.

Matthew Z. DeMaere: Production Note:
Signature removed prior to publication.

Dominik Beck: Production Note:
Signature removed prior to publication.

Martin Ostrowski: Production Note:
Signature removed prior to publication.

Justin R. Seymour: Production Note:
Signature removed prior to publication.

Aaron E. Darling: Production Note:
Signature removed prior to publication.

2.1 Abstract

Prokaryotic evolution is influenced by the exchange of genetic information between species through a process referred to as recombination. The rate of recombination is a useful measure for the adaptive capacity of a prokaryotic population. We introduce Rhometa (https://github.com/sid-krish/Rhometa), a new software package to determine recombination rates from shotgun sequencing reads of metagenomes. It extends the composite likelihood approach for population recombination rate estimation and enables the analysis of modern short-read datasets. We evaluated Rhometa over a broad range of sequencing depths and complexities, using simulated and real experimental short-read data aligned to external reference genomes.

Rhometa offers a comprehensive solution for determining population recombination rates from contemporary metagenomic read datasets. Rhometa extends the capabilities of conventional sequence-based composite likelihood population recombination rate estimators to include modern aligned metagenomic read datasets with diverse sequencing depths, thereby enabling the effective application of these techniques and their high accuracy rates to the field of metagenomics. Using simulated datasets, we show that our method performs well, with its accuracy improving with increasing numbers of genomes. Rhometa was validated on a real *S. pneumoniae* transformation experiment, where we show that it obtains plausible estimates of the rate of recombination. Finally, the program was also run on ocean surface water metagenomic datasets, through which we demonstrate that the program works on uncultured metagenomic datasets.

2.2 Author Summary

Microbes, specifically prokaryotes, are able to exchange DNA between them through a process called recombination that takes the form of gene-conversion. Recombination plays a fundamentally important role in microbial speciation and evolution. Metagenomics allows us to study microbes in their natural environment as they are via direct sequencing and analysis of environmental DNA. Indeed, most microbes cannot be cultured and can only be studied in this manner. Metagenomic datasets represent an excellent resource for measuring prokaryotic recombination. Rhometa is a new software program that we have designed to be used to interrogate modern

metagenomic shotgun sequencing read datasets to estimate the population recombination rates. It extends the composite likelihood approach for population recombination rate estimation, and makes it applicable to modern aligned metagenomic datasets of various depths. The input for the program requires little pre-processing and only an aligned BAM and reference FASTA in the form of a complete sequence or MAG are needed. The program performs well on large BAM files and the included subsampling functionality ensures that files of arbitrarily large size are subsampled and analysed. The program has been validated on simulated datasets and further on experimental and environmental datasets where the amount of recombination is quantified. Through the validation we show that the program is able to reliably estimate the recombination rates in simulated and experimental datasets.

2.3 Introduction

A primary question in the field of microbial ecology is to understand the rate at which prokaryotes evolve and form species in nature. A major driving factor of prokaryotic evolution is recombination (Iranzo *et al.*, 2019). Within prokaryotes, recombination often takes the form of gene-conversion where homologous sequences of DNA are non-reciprocally transferred and replaced with another (Paulsson *et al.*, 2017; Bobay, 2020). This process can occur between repeated sequences within the same chromosome and between homologous chromosomal sequences (Paulsson *et al.*, 2017). Recombination often plays a greater role than *de novo* mutation for evolution in prokaryotes (Vos and Didelot, 2009). Furthermore, it is thought that recombination plays an important role in counteracting the effects of Muller's ratchet, the theorised process where deleterious mutations inevitably accumulate over time leading to the irrevocable loss of most mutation free genotypes in a population (Muller, 1964; Andersson and Hughes, 1996). Therefore, understanding the rate at which recombination occurs within prokaryotic populations can provide us insight into a crucial biological process that is necessary for their adaptation and survival.

Currently the most viable way to study microbial populations is via metagenomics, which allows us to study microbes in their natural environment via direct sequencing and analysis of environmental DNA (Thomas, Gilbert and Meyer, 2012; Escobar-Zepeda, Vera-Ponce de León and Sanchez-Flores, 2015). Shotgun metagenomic sequencing yields fragments of DNA sequences, referred to as reads, which taken together represent a random sampling of genome fragments from all the

microbes in the environmental sample (Sharpton, 2014). These reads can then be used to estimate the rates of recombination.

The rate of recombination within a population can be inferred using population genetic models for evolution. The Wright-Fisher model provides an analytical framework that quantifies various forces that can impact the evolution of a population such as random genetic drift and mutation (Tataru *et al.*, 2017). Coalescent theory, building on the Wright-Fisher population model, provides an analytical framework for DNA polymorphism data and can be used to obtain quantitative estimates for recombination and mutation rates (Fu and Li, 1999; McVean, Awadalla and Fearnhead, 2002).

Coalescent theory provides the population scaled recombination rate for the gene-conversion model of recombination. It is formulated as $\rho = 2N_e r$, or *2 × "effective population size" × "per individual" "per generation" "recombination rate"*, respectively (McVean, Awadalla and Fearnhead, 2002) as well as the haploid population scaled mutation rate equation $\theta = 2N_e u$, or *2 × "effective population size" × "per individual" "per generation" "mutation rate"*, respectively (McVean, Awadalla and Fearnhead, 2002). It is difficult to estimate r or u directly without additional prior information, so recombination and mutation rates are typically computed as the population scaled statistics ρ and θ or simultaneously as the ratio r/u also denoted as r/m (per site recombination to per site mutation rate) (McVean, Awadalla and Fearnhead, 2002; Melendrez *et al.*, 2016). An important point of note is that r and u are per site rates. ρ applies to the entire genome, θ on the other hand is also per site.

Several approaches have been used to estimate the recombination rate ρ. These include moment estimators, full-likelihood estimators and composite likelihood estimators. Moment estimators use summary statistics to estimate ρ, but their accuracy is limited by the fact that they cannot use all the genetic information available (Fearnhead and Donnelly, 2001, 2002; Stumpf and McVean, 2003). Full likelihood estimators are able to utilise all the genetic information available to them, but are so computationally intensive that their usage is impractical. To mitigate these issues and to make the approach more computationally tractable, composite likelihood estimators were developed (Hudson, 2001; McVean, Awadalla and Fearnhead, 2002; Stumpf and McVean, 2003). With composite likelihood estimators, the scope of data that is analysed is reduced e.g. to only consider pairs of alleles, this approach is less computationally intensive with only a slight loss in accuracy compared to the

full-likelihood approach (Hudson, 2001; McVean, Awadalla and Fearnhead, 2002; Stumpf and McVean, 2003; Hermann *et al.*, 2019).

There are several programs available that implement the composite likelihood approach for estimating the recombination rate, including LDhat (McVean, Awadalla and Fearnhead, 2002; Auton and McVean, 2007), LDhelmet (Chan, Jenkins and Song, 2012), LDhot (Auton, Myers and McVean, 2014), PIIM (Johnson and Slatkin, 2009) and pyrho (Spence and Song, 2019). Each are excellent for their respective use cases, but have limitations that make them unsuitable for modern read-based metagenomic datasets.

More specifically, LDhat (McVean, Awadalla and Fearnhead, 2002; Auton and McVean, 2007), LDhelmet (Chan, Jenkins and Song, 2012) and LDhot (Auton, Myers and McVean, 2014) were designed for genome sequence analysis, not metagenomes. PIIM (Johnson and Slatkin, 2009) was a pioneering attempt at a metagenomic read-based population recombination rate estimator which deals with the complexities of such datasets such as varying depths across loci. While innovative at the time its application is impractical today. PIIM's approach included computationally expensive techniques to integrate out uncertainty in low quality base-calls so as to retain as much information as possible from the scarce data available at the time. Today, deep sequencing is affordable and highly accurate, such that it's often more practical to simply discard low quality sequence data rather than account for it computationally using complex algorithms. As such PIIM's approach is impractical for the ever-larger datasets that are generated via modern sequencing techniques. Furthermore, it lacks support for modern sequence data formats (e.g., BAM), being limited to the obsolete ACE assembly format that is rarely used today.

pyrho (Spence and Song, 2019) is a recent composite likelihood estimator supporting read based data, however it is limited to the analysis of diploid genomes, making it unsuitable for the analysis of haploid genomes (prokaryotes) which commonly dominate metagenomic datasets. Still other programs exist that calculate the population recombination rate through different approaches such as LDjump (Hermann *et al.*, 2019) and CodABC (Arenas *et al.*, 2015) which utilise summary statistics (Hermann *et al.*, 2019), and programs such as ClonalFrameML (Didelot and Falush, 2007; Didelot and Wilson, 2015) which can provide an estimate of recombination rate relative to the mutation rate, but is designed around bacterial isolate genomes.

mcorr (Lin and Kussell, 2019) is a program that can work with metagenomic reads and estimate the relative rate of recombination to mutation as well as the recombinational divergence. mcorr uses an alternative mathematical formulation to parameterise the recombination process, with $\phi_{\text{pool}} \equiv 2\overline{T}\gamma$, where $\overline{T}$ "is the mean pairwise coalescence time across all loci in the bulk pool" and γ the per base pair (bp) per generation recombination rate, equivalent to r in $\rho = 2N_e r$. Furthermore, the program is limited to coding regions and requires a gene annotation file. It is our aim to build on methods established in previous composite likelihood estimators for population recombination rate estimation to create a tailored solution that is applicable to modern aligned read-based metagenomic datasets.

Here, we present Rhometa, a software implementation of the composite likelihood based population recombination rate ($\rho = 2N_e r$) estimation method, which builds upon the methods introduced in the LDhat pairwise program (McVean, Awadalla and Fearnhead, 2002) and can be applied directly to modern aligned shotgun metagenomic read datasets for prokaryotes. Details of its implementation are presented in the methods, while an evaluation of its accuracy on simulated and real data and comparison to existing tools are presented in the results.

2.4 Description of the Method

Our approach focuses on advancing the composite likelihood recombination rate estimator for use with aligned metagenomic read datasets. We have built our metagenomic population recombination rate estimator program upon the approach introduced in the LDhat program, specifically the LDhat pairwise module (McVean, Awadalla and Fearnhead, 2002). LDhat is a well-known and used program with support for microbial datasets, specifically for the gene-conversion type recombination which occurs in microbes, however, it is limited to aligned genome sequences. We have designed the program to work with aligned read based metagenomic datasets, where the complication of varying depths and short reads needs to be addressed. For our implementation, we have also subsumed features from pyrho (Spence and Song, 2019). pyrho, while lacking support for microbial (haploid) datasets, is a modern composite likelihood estimator implemented in python. Like pyrho, our program is also implemented in python and aims to make use of modern libraries and their features. As a result of this shared implementation approach, we were able to call applicable functionalities from pyrho, helping avoid unnecessary code rewrites.

2.4.1 User Input

We have endeavoured to make the process of preparing a metagenomic dataset for analysis with Rhometa straight forward with few pre-processing steps. Short reads can be aligned to existing reference genomes representing the metagenomic dataset, to reference MAGs (Metagenome-Assembled Genome) or pangenomes representative of the microbial community. Multi sequence references are also supported. The reference genomes provide a scaffold to align the reads, from which the rates of recombination and mutation are determined. For input files, the Rhometa pipeline requires a FASTA format reference sequence and a BAM file of metagenomic reads of interest aligned to the reference. For our pipeline evaluation, we have used BWA MEM (default parameters) to produce the input BAM file (Li, 2013) where necessary.

2.4.2 Variant site pairs

The first step of the pipeline involves identifying variant sites (also known as segregating sites). Our program first filters the user supplied BAM for mapping quality and relative alignment score and subsequently performs variant calling against the user supplied reference FASTA using the program freebayes (default parameters with -p (ploidy) = 1) (Garrison and Marth, 2012). The resulting VCF file, containing information on all predicted variant sites, is reduced to only single nucleotide polymorphisms (SNPs) using bcftools (Danecek *et al.*, 2021).

Rather than individual variant sites, the composite likelihood estimator as implemented in LDhat considers variant site pairs, tracking count and position within the reference genome's coordinate space to estimate the recombination rate. For instance, if variant sites are found at reference positions 1, 3, and 5 the set of variant site pairs would then be (1, 3), (1, 5), and (3, 5).

2.4.3 Pairwise table

The LDhat pairwise module was designed for genome sequences and considers all possible variant site pair combinations across the sequences being analysed. Rhometa restricts its consideration to the set of variant site pairs linked by individual reads or read-pairs. For single-end reads, both sites within a variant pair must fall within the extent of an individual read, while for paired-end reads variants can fall within the insert length. A separation limit of 1000 bp is imposed on paired end variant site pairs reflecting a practical upper limit on insert size for current Illumina short-read

sequencing technology (Tan *et al.*, 2019). Rhometa performs well with both single and pair-end reads, with very little difference in the results between the two (S2.2 Fig).

For all accepted variant site pairs, we construct a pairwise table of observational frequency (Table 2.1). The pairwise table allows for the possibility of all 16 combinations for any variant site pair. The table also captures the fact that multiple reads can align at a position. Instances where variant site pairs contain an ambiguous base (eg. N) are ignored.

Table 2.1. Pairwise table example

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
(130, 136)	0	13	0	21	0	0	0	10	0	0	0	0	0	0	0	0
(130, 143)	13	0	21	0	0	0	10	0	0	0	0	0	0	0	0	0
(130, 169)	0	29	1	3	0	8	0	0	0	0	0	0	0	0	0	0
(130, 311)	2	0	26	0	0	0	9	0	0	0	0	0	0	0	0	0
(130, 358)	0	0	0	19	0	2	0	6	0	0	0	0	0	0	0	0

2.4.4 Splitting the pairwise table by depth

In the pairwise table, variant site pair total alignment depth is calculated by row summation (e.g., For the pair (130, 136) from (Table 2.1), total depth is 13 + 21 + 10 = 44. For the whole genome approach of LDhat, this marginal value is a constant, while for metagenomic data depth of coverage can vary greatly across sites. As such it is necessary to split the pairwise table into constant depth subtables so that the depth can be taken into account and handled in downstream processing.

2.4.5 Bi-allelic pairwise table

For each constant depth subtable, sites that do not contain two alleles are excluded (only biallelic sites should be in the pairwise table).

2.4.6 Lookup tables

Lookup tables improve the computational efficiency of the composite likelihood approach by precomputing the likelihoods for different configurations of sets of allele

pairs. Lookup tables are generated under a fixed population mutation rate and a range of population recombination rates, typically between 0 - 100 (McVean, Awadalla and Fearnhead, 2002; Auton and McVean, 2007). We use the program LDpop (Kamm *et al.*, 2016) for generating lookup tables as it is the most feature rich and most efficient program of its kind currently. Details on how the lookup tables are used can be found in Appendix A. It is a standard process for which we have made use of some functions from pyrho to avoid reimplementation. Generation of lookup tables with LDpop are parameterised by the number of genomes, range of population recombination rates, and θ per site. Further, the "approx" option is used which is significantly faster but still quite accurate when compared to the LDpop's exact algorithm. This also makes generation of large lookup tables, as is necessary with Rhometa, more tractable. Generally, the generation of large lookup tables is a computationally intensive process, more information on which can be found in the paper associated with LDpop (Kamm *et al.*, 2016).

2.4.7 Watterson's theta estimate

A subprogram is provided to estimate the population mutation rate ($\theta = 2N_e u$), per site. The formulation is directly based on Watterson's theta estimate as implemented in LDhat (McVean, Awadalla and Fearnhead, 2002), with some changes on the information used for the input parameter. The program requires the aligned BAM file and the reference FASTA file and makes use of freebayes to identify variant sites which is required for the Watterson estimate. θ is a required parameter for lookup table generation adjusted for read based datasets, the θ estimate is calculated based on dataset depth – specifically mean and median depth – in place of the number of genomes, as in the original formation (S2.4 Fig). For metagenomic samples where the exact number of genomes is unknown, the true number of different genomes in a metagenomic sample is likely to be a large value, as such using the depth value is reasonable and in many cases is likely to be a conservative estimate for the number of genomes.

2.4.8 Lookup table and depth

The size of each constant depth pairwise sub-table determines the lookup table to match against for precomputed negative log-likelihoods. In the context of Rhometa and aligned reads, the number of genomes parameter used to generate lookup tables is instead taken as depth. Where the BAM file is of a high depth, this can result in

pairwise sub-tables of high depth. In such cases, available lookup tables may not be of high enough depth to cover them. A subsampling feature is included that is able to automatically downsample the BAM to a given depth. This ensures that positions with a depth exceeding that of the highest generated lookup table are still evaluated and are not omitted from consideration. BAM subsampling uses a random sampling process and permits a list of seed values for testing and identifying any variance that can stem from the subsampling. In general, if the depth of the largest available lookup table is small, an increased need for downsampling could result in a decrease in estimation accuracy.

2.4.9 Calculating r_{ij}

The next step is to calculate recombination rate values for each variant site pairs, these values are denoted by r_{ij}, with i and j referring to the variant sites. The method used for calculation differs for crossing-over and gene-conversion modes of recombination. prokaryotes undergo recombination via gene-conversion and the equation used to calculate r_{ij} is as follows (McVean, Awadalla and Fearnhead, 2002):

$$r_{ij} = 2ct(1 - e^{-d_{ij}/t}) \qquad \text{Equation 1}$$

For equation 1, c represents the per base recombination rate, t the mean gene conversion tract length and d_{ij} the distance between a variant site pair. ct is taken together and represents the range of population recombination rates being evaluated, this is typically between 0 - 100 and is the same as the range of ρ values used when generating the lookup tables. The process essentially involves computing r_{ij} for each variant site pair for the range of population recombination rate values.

2.4.10 Final pairwise likelihoods

Next, we bring together the information we have generated thus far: the output of the lookup table and depth step and calculating r_{ij} step. For each variant site pair, we use the corresponding negative log-likelihood values from the lookup table and depth step, and on these apply linear interpolation to determine the negative log-likelihood values

for r_{ij} value for that variant pair configuration. The interpolation is performed against the range of population recombination rates used to generate the lookup tables. This process is done for all the variant site pairs and the results are the final negative log-likelihoods for a given range of population recombination rates being evaluated, which again is typically between 0 - 100.

2.4.11 Population recombination rate

Prior to computing the final log-likelihood sums we weight adjust the negative log-likelihoods. As the number of observations provided at a given depth represents the degree of evidential support towards the final ρ estimate, we introduce a novel weighting algorithm that accounts for the additional information in high depth, high observation count site pairs relative to low depth / low count site pairs. The weighting algorithm is a simple solution to reliably add more weight to higher depth regions and the number of observations at that depth (there can be many regions of same depth). The weighting and final negative log-likelihood summation algorithms are as follows:

$$w_d(\rho) = \ln\left(\frac{\exp T_d}{\sum_\rho \exp T_d} dn_d\right) \quad \text{Equation 2}$$

$$\rho_{\max} = \operatorname{argmax} \sum_d w_d(\rho) \quad \text{Equation 3}$$

Here, d represents depths observed in the dataset. Weighting is performed on each per-depth table denoted by $w_d(\rho)$ (equation 2). In the right side of the equation 2, T_d is the unweighted per-depth table and n_d the number of unique variant site pairs. The reweighted negative log-likelihoods $w_d(\rho)$ are collected across depths and summed with respect to the range of population recombination rates being evaluated. The maximum negative log-likelihood value (closest to 0) then corresponds to the final population recombination rate ρ_{max} (equation 3) estimate.

2.4.12 Program Structure

The program is organised into 4 pipelines, each dedicated to a specific task. These pipelines are written using nextflow, a framework for pipeline management (Di Tommaso *et al.*, 2017). All the scripts used in the individual pipeline steps were written using the python programming language and various python libraries. Some python scripts were adapted or used as is from the programs LDpop (Kamm *et al.*, 2016) and pyrho (Spence and Song, 2019). Additional programs used in the pipelines include msprime (Kelleher, Etheridge and McVean, 2016), ART (Huang *et al.*, 2012), BWA MEM (Li, 2013), seaborn (Waskom, 2021) and samtools (Danecek *et al.*, 2021).

The four pipelines, sim_gen, theta_est, lookup_table_gen and rho_est (Fig 2.1), correspond to the nextflow pipeline names, e.g. sim_gen.nf within Rhometa, and perform the functions defined in the following paragraphs.

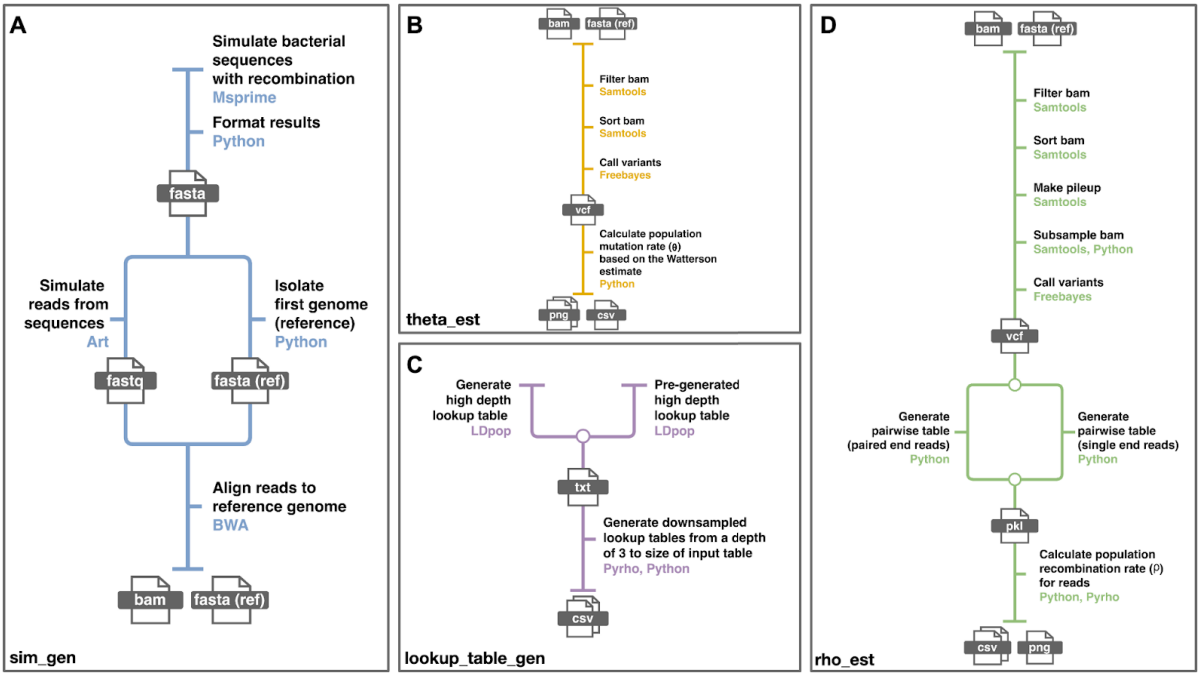


Fig 2.1. The pipelines that together make up Rhometa. (**A**) Pipeline for generating simulated metagenomic read datasets. (**B**) Pipeline for estimating the population mutation rate (**C**) Pipeline for generating the lookup tables required for the recombination rate estimator (**D**) Pipeline for estimating the population recombination rate.

Sim_gen (Fig 2.1A) is used to generate BAM files and FASTA reference files with simulated reads from bacterial genomes with recombination. The bacterial genomes are simulated using msprime. This pipeline is primarily included so that the simulated

datasets used for this paper can be reproduced, but is not required to analyse real datasets. It is in a separate repository and can be accessed at: https://github.com/sid-krish/Rhometa_sim

Theta_est (Fig 2.1B) is used to determine the population mutation rate per site (θ) based on the Watterson estimate as implemented in LDhat, details in methods. This pipeline estimates θ on the dataset of interest, furthermore, θ per site is one of the required parameters for generating lookup tables. The user has the option to use the estimated θ or a different value when generating lookup tables.

The Lookup_table_gen (Fig 2.1C) component of the pipeline makes use of LDpop and pyrho to generate the lookup tables required for the recombination rate estimator and can be launched in one of 2 ways. It can either use a pre-generated lookup table for high depth, which then will be downsampled for each depth from 3 to the depth of the lookup table or the pipeline can generate a high depth lookup table from scratch and then perform the downsampling step. The downsampling algorithm is a part of pyrho, it is significantly faster to generate the required smaller lookup tables from a larger table via downsampling and the results are essentially identical.

The rho_est pipeline (Fig 2.1D) is used to estimate the population recombination rate of metagenomic read based datasets provided in the form of BAM and reference FASTA files. It makes use of the lookup tables generated by the lookup_table_gen pipeline. Rhometa is available at: https://github.com/sid-krish/Rhometa

Our pipelines for evaluating LDHat, the Rhometa_full_genome pipeline and the simulated dataset generator for these pipelines can be accessed here:

- LDhat Nextflow Pipeline: https://github.com/sid-krish/Nextflow_LDhat
- Rhometa Full Genome Pipeline: https://github.com/sid-krish/Rhometa_Full_Genome
- Nextflow_LDhat_sim (used for both Rhometa Full Genome and LDhat Nextflow Pipeline) : https://github.com/sid-krish/Nextflow_LDhat_Sim

2.4.13 Simulated datasets

The development of our program was performed in two major phases, for the first phase we endeavoured to create a full genome recombination rate estimation pipeline

for bacterial sequences based on the LDhat methodology (Rhometa_full_genome), once we were certain that we were able to replicate LDhat's results exactly we then carefully adapted the program to work with read based datasets (Rhometa).

To evaluate LDhat and Rhometa_full_genome, we utilised msprime (Kelleher, Etheridge and McVean, 2016) to simulate bacterial sequences with recombination. Our simulations included multiple genomes (5-100 genomes) of size 25KB, under population recombination rates [5, 12.5, 25, 37.5, 50], mean recombination tract length 500bp, with 10 replicates (seed values 1-10) and population mutation rate 0.01. Lookup tables for population mutation rate 0.01 and population recombination rates 0-100 (101 steps) were used.

The LDhat pipeline configured for gene-conversion is available at: https://github.com/sid-krish/Nextflow_LDhat. Rhometa_full_genome pipeline is available at: https://github.com/sid-krish/Rhometa_Full_Genome. The full genome simulation pipeline is available at https://github.com/sid-krish/Nextflow_LDhat_Sim. A point of note is that the θ estimator is implemented separately by us as per equation 1 (McVean, Awadalla and Fearnhead, 2002) in both our LDhat pipeline and Rhometa_full_genome pipeline. Additionally, all variant sites are used for θ estimation, not just bi-allelic ones.

When simulating the population recombination rate with msprime, the number of samples (genomes), sequence length, recombination rate (r), mean gene conversion tract length (t), seed value and mutation rate (u) are provided, the effective population size (N_e) was 1 (default) and the ploidy (i) was fixed to 1. Default options are used in all other cases. Within msprime, the population recombination rate ρ is calculated as such: $2 * i * N_e * r * t$. The per site population mutation rate θ was calculated as such: $2 * i * N_e * u$.

Initially the number of genomes was fixed and we varied the length of the genomes, but this analysis revealed that varying the genome length does not have a significant impact on the final population recombination rate estimations (S2.1 Fig). We therefore fixed the genome length and varied the number of genomes and in doing so we found that as the number of genomes increased the accuracy and variance of the final estimations also improved (Fig 2.2).

We took a similar approach to evaluating the read based pipeline Rhometa to that used with LDhat and Rhometa_full_genome. For the read-based pipeline, the simulated full bacterial sequences, simulated via msprime, are further processed to be in the form of reads using the read simulator ART (Huang *et al.*, 2012), which simulates sequencing reads by mimicking real sequencing processes with empirical error models. The reads were simulated based on the Illumina HiSeqX TruSeq system. These reads are then aligned to one of the bacterial sequences which represents the reference FASTA file, the first of the simulated sequences is used for this (Fig 2.1A). The aligned BAM and reference FASTA are then used for recombination rate estimation.

The simulation parameters were as follows: the effective population size was 1 (default) and the ploidy was set to 1, number of genomes 20-200, genome length 100KB, population recombination rates [10.0, 20.0, 30.0, 40.0, 50.0], mean recombination tract length 1000bp, with 20 replicates (seed values 1-20) and population mutation rate of 0.01. Each seed value used applies to all aspects of the pipeline where a seed is required. The reads were paired-end of length 150bp, insert length 300bp, standard deviation of 25bp, with window size set to 1000 during analysis and the fold coverage values were [1, 4, 8, 16]. The lookup tables were generated and used for 3-250 (genomes), generated under population mutation rate 0.01 for population recombination rates 0-100 (0-1 in 101 steps plus 1-100 in 100 steps). Bam subsampling was also automatically applied by Rhometa during analysis if needed.

Additionally for the read-based pipeline, we evaluated the deviation of the estimated results from the simulated values. The formula used to calculate the deviation is (Estimated ρ (mean) - Simulated ρ) / Simulated ρ. This makes it easier to gauge the magnitude of deviation from the expected.

2.4.14 Real Datasets - Transformation experiment

To further evaluate Rhometa we applied our pipeline on the data derived from a previously published laboratory transformation experiment, where the extent and distribution of recombination events were quantified. In the experiment (Croucher *et al.*, 2012), *in vitro* recombination through transformation was performed on a *S. pneumoniae* strain. Transformed isolates were then sequenced and recombination events were identified. This dataset was also used to evaluate the mcorr method by its authors and as such it provides us with the opportunity to compare the results of our pipeline against those published in the mcorr paper.

The transformation experiments were performed with different concentrations of donor DNA, 5 ng mL-1 and 500 ng mL-1, 5 ng mL-1 and 500 ng mL-1 experiments (S2.2 Table) had a similar number of recombination events, with the 5 ng mL-1 having a slightly larger number of events, the authors state that this indicates a single piece of DNA can act as the origin for multiple recombination events. The dataset is available in the form of reads, which Rhometa was designed to analyse. Each 5 ng mL-1 sample from experiment 1 was aligned to *S.pneumoniae* reference sequence ATCC 700669, NCBI accession NC_011900.1 the resulting BAM files were then merged and analysed with Rhometa.

To analyse the datasets, we first estimated the θ for median depth using the θ estimation pipeline, from which we obtained θ, that is per site by default. We then generated lookup tables, based on the θ, for population recombination rates 0-20 in 201 steps for 3-200 genomes and used the lookup tables for the recombination rate estimation pipeline. Subsampling was enabled, with a window size of 1000 for paired end reads. As given in the Croucher et al paper, we used the value of t=2300 bp as the mean tract length for analysis. Additionally, we used 5 different seed values [0, 1, 2, 3, 4] for the subsampling step to account for any variance and then took the average of the values for recombination rate estimations.

Using the ρ and θ estimates along with information from the experiment we also calculated the ρ per site and r/m values. The default ρ estimate, by Rhometa, is a whole-genome estimate. To obtain ρ per site, the estimated ρ value was divided by the mean tract length of 2300 bp. To get the r/m value, we used the conversion formula (Didelot and Wilson, 2015): ρ (per site)/θ (per site) * tract length * substitution probability. We estimated the substitution probability between the donor and recipient and found it to be $(17534 - 385)/2221315 = 0.00772$, where 2221315 is the recipient genome length and based on the information provided by Croucher et. al (2012) "... the donor DNA identified 17,534 SNPs when aligned to the recipient sequence, of which 385 appeared to be false positives … These positions were excluded from subsequent analyses." (Croucher *et al.*, 2012).

We repeated the process above for each 500 ng mL-1 sample from experiment 1 and the final merged BAM was analysed with Rhometa. Furthermore, the 5 ng mL-1 samples and 500 ng mL-1 samples in experiment 1 were analysed together, corresponding to 84 sequences. We prepared this dataset by merging the final 5 ng

and 500 ng BAMs. We performed this analysis, to enable comparison with mcorr's published results. The analysis was performed using the same process as with the 5 ng mL-1 samples and 500 ng mL-1 samples.

2.4.15 Real Datasets - Ocean Metagenomic Dataset

To evaluate our pipeline on typical metagenomic data we selected 15 publicly available short-read metagenomic reads-sets (S2.3 Table) sampled from waters at the Australian Integrated Marine Observing System (IMOS) Port Hacking National Reference Station (NRS) (Rinke *et al.*, 2019) (S2.3 Table). These 15 datasets represent 15 consecutive monthly time points (July 2012 - September 2013). Next, we assembled a long-read metagenomic read-set (accession: SRR13002033) collected in March 2018 from the location using metaFlye (v2.8.3-b1705) (default options) (Kolmogorov *et al.*, 2020), Three long contigs (lengths: 0.8 – 1.0 Mbp) were selected from the assembly to be used as references for read-mapping (table 3). Confident taxonomic assignments were determined for each contig using MMseqs2 (v14.7e284) (default easy-taxonomy options) (Mirdita, Steinegger and Söding, 2019) against the latest GTDB taxonomic database (release: R07-RS207) (Parks *et al.*, 2022). The 15 read-sets were mapped to each reference contig using BWA MEM (v0.7.17-r1188) (default options) (Li, 2013) and the resulting BAMs merged respectively with samtools (v1.16.1) (Danecek *et al.*, 2021). CoverM (v.0.6.1) (Woodcroft, 2023) was then used with the metabat method to determine the coverage for the aligned BAMs. Lookup tables for 3-250 (genomes), generated under population mutation rate 0.01 for population recombination rates 0-100 (0-1 in 101 steps plus 1-100 in 100 steps) were used. The tract was fixed at 1000bp as was the window size. As with the transformation experiment analysis we used 5 different seed values [0, 1, 2, 3, 4] for the subsampling step to account for variance and then took the average of the values for ρ estimations.

2.5 Verification and Comparison

2.5.1 Evaluation on Simulated Datasets

We first validated the full genome version of Rhometa (Rhometa_full_genome), which reimplements the core LDhat pairwise method to estimate ρ. This was done to ensure accuracy in reimplantation of core LDhat algorithms which forms the basis for the read based (Rhometa) implementation. Comparison of estimated population recombination rate (ρ) between LDhat and our reimplementation (Rhometa_full_genome), using our sweep of simulated genomes, shows identical results between LDhat (Fig 2A) and our

reimplementation (Fig 2B), thus ensuring that we have captured LDhat's algorithms accurately. With LDhat and our reimplementation, the number of genomes simulated has a large impact on the accuracy of the estimates, with results improving with higher numbers of genomes, especially at higher recombination rates.

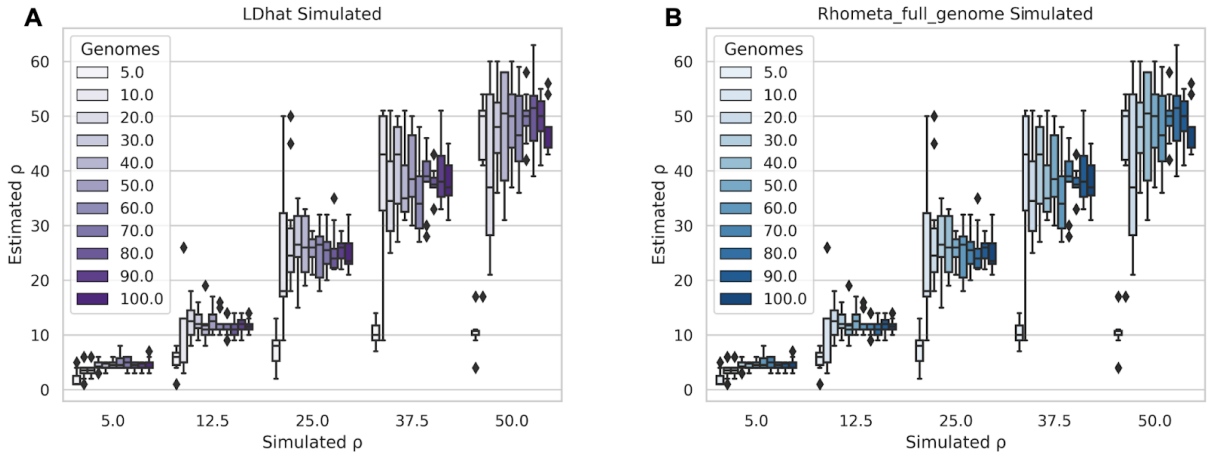


Fig 2.2. Comparison of LDhat and Rhometa_full_genome when running on simulated full genomes (**A**) LDhat. Simulated vs Estimated population recombination rate (ρ) for varying number of simulated full bacterial genome sequences. (**B**) Rhometa_full_genome. Simulated vs Estimated population recombination rate (ρ) for varying number of simulated full bacterial genome sequences.

We next evaluated Rhometa's performance using our sweep of simulated read-sets. The number of simulated genomes had a large bearing on estimation accuracy (Fig 2.3), as also observed with LDhat, accuracy improved as the number of genomes increased and inter-replicate variance decreased as the coverage (fold_coverage) improved. This is especially evident for higher recombination rates. Larger population recombination rate values appear to require a relatively higher number of genomes for accurate estimation. For very low recombination rates between 0-1 (S2.3 Fig), the improvement in accuracy was not seen and a tendency to overestimate was observed.

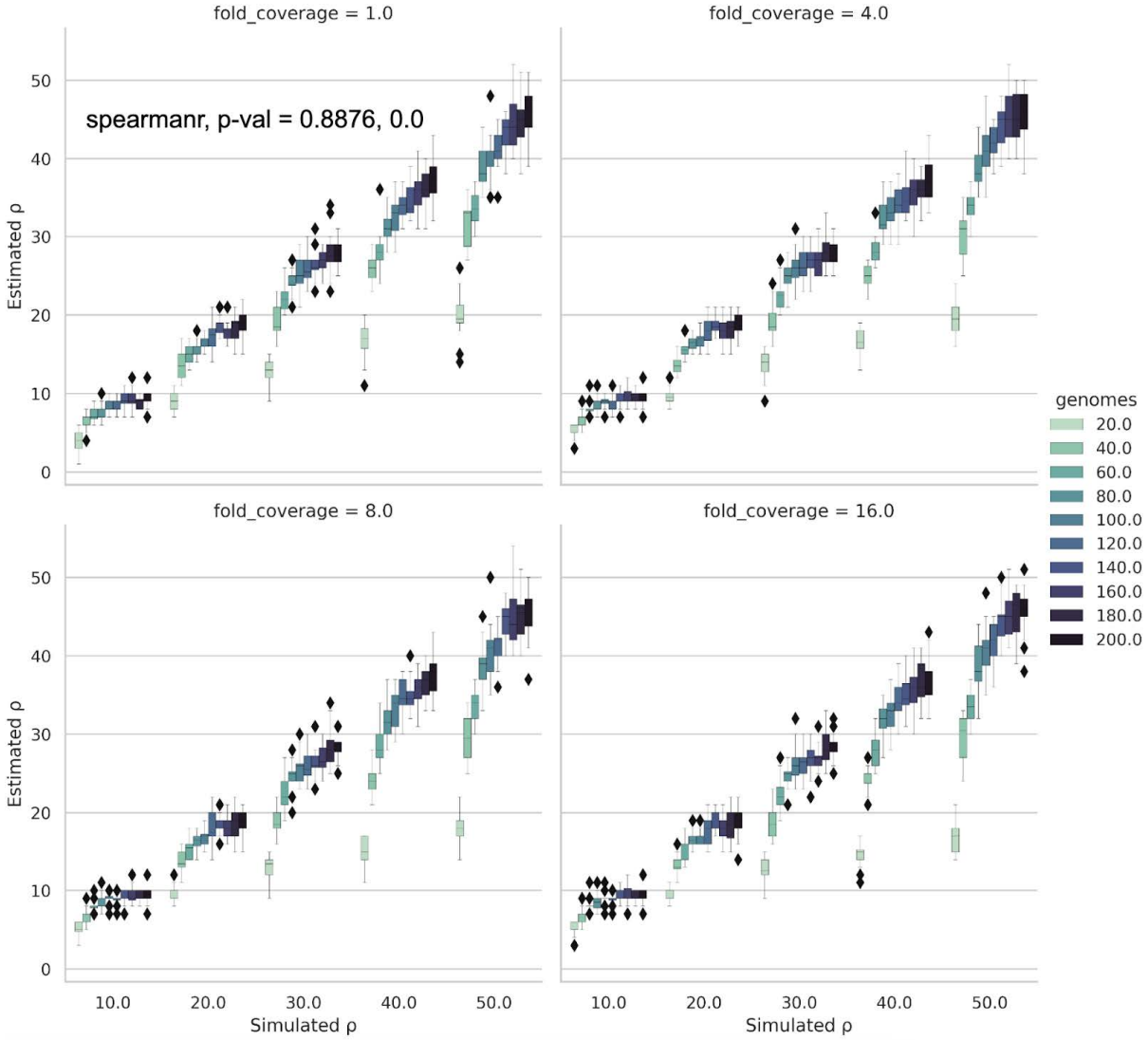


Fig 2.3. Simulated vs Estimated population recombination rate (ρ) results for Rhometa. Results for varying numbers of simulated genomes and fold coverage values for population recombination rates 10.0, 20.0, 30.0, 40.0, 50.0.

To get a clearer picture of the deviation of the estimated population recombination rate from the expected result, we calculated the deviation for the read based results (Fig 2.4). Here values closer to 0 indicated better performance, while values above 0 are overestimations, and values below 0 are underestimations (i.e. a deviation value of +/- 0.1 would indicate that the final result is off by 10%). As the number of simulated genomes increased, the deviation of estimated to expected tended to decrease, achieving a deviation of less than 5-10% in most cases for a simulated ρ of 50 with 200 genomes and 16x coverage. Such improvement is consistent with the patterns observed in LDhat. For simulated population recombination rates between 10-50, having greater than 80 genomes produced the least amount of deviation (generally

within 20-30%), with the results significantly improving when more genomes are present. Our pipeline appears to be robust to variance in fold coverage. The differences between 16x coverage and 1x coverage being minor (Fig 2.3).

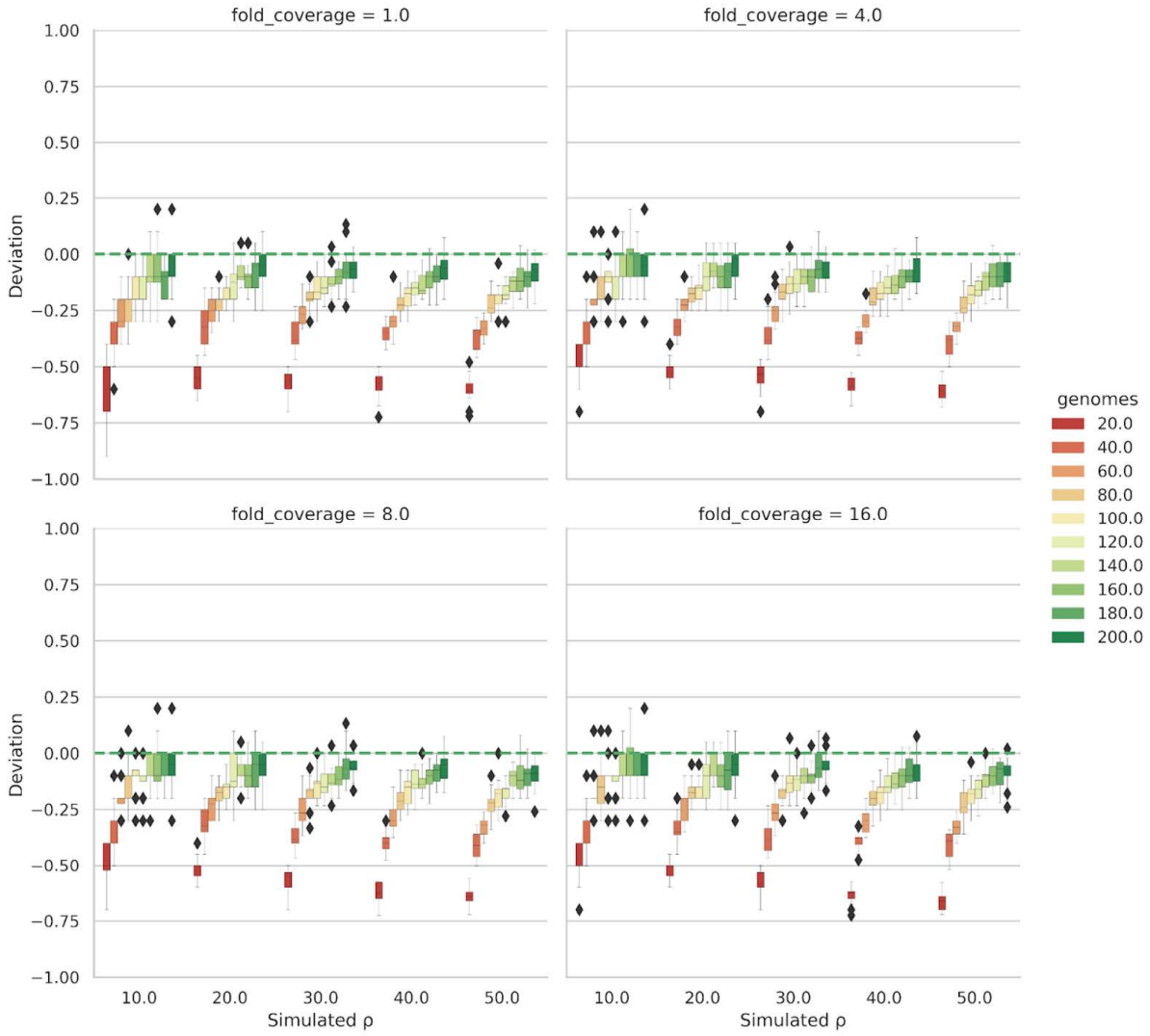


Fig 2.4. Deviation plot for results in Figure 2.3. Deviation is calculated as (Estimated ρ (median) - Simulated ρ) / Simulated ρ. Deviation results corresponding to Figure 2.3 for population recombination rates 10.0, 20.0, 30.0, 40.0, 50.0.

2.5.2 Evaluation on Real Datasets - Transformation Experiment

After establishing the performance of Rhometa on simulated datasets, we interrogated its performance on real short-read sequence data from a lab-based experiment designed to track and study recombination in *S. pneumoniae* (Table 2.2), seed

averaged values were taken for the final results, refer to S3.1 Table for individual seed results.

Table 2.2. *S. pneumoniae* transformation experiment analysis

	ρ	ρ (per site)	θ (per site)	ρ/θ (per site)	r/m
5 ng mL^{-1}	5.56	0.00242	1.8e-5	134.4	2386.4
500 ng mL^{-1}	5.22	0.00227	2e-5	113.5	2015.3
84 sequences	4.48	0.00195	1.8e-5	108.3	1922.97

To compare the results for the same transformation experiment with those of mcorr, the estimated r/m values were used. The authors of mcorr provide γ/μ (similar to r/m) for the evolved strain (reads) representing combined 5 ng mL^{-1} samples and 500 ng mL^{-1} samples in experiment 1 (84 sequences) where they estimate a γ/μ value of 0.93. Due to the nature of the experiment, it was expected that the rate of recombination would be far higher than the rate of mutation. As an experiment designed to induce transformation over short timescales, this should lead to a large excess of substitutions derived from recombination events, relative to *de novo* mutation. As such, mcorr potentially significantly underestimated the true value and Rhometa better reflects our *a priori* expectation for the 84 sequence dataset with an estimated r/m ratio of 1922.97.

The 84 sequence dataset was also used to benchmark Rhometa. The BAM file for this dataset is 11.42 GB in size. On a system with 4 CPUs and 32 GB RAM the rho_est pipeline took 1h 45m to complete and the theta_est pipeline took 2h 45m to complete.

We can calculate the expected average r/m for a genome, as follows. In the Croucher *et al* paper (Croucher *et al.*, 2012), they state that the mean proportion in the recipient genome changed due to recombination was 1.4% using which we can estimate the average number of bases changed by recombination in a single genome as: 2221315 (recipient genome length) * 0.014 * 0.00772 (substitution probability as calculated in methods above) = 240.08. Also, in the paper it is stated that there were 2,312 polymorphic sites, 59 of which not coming from the donor, 6 of these sites were false positives, with the others likely being *de novo* point mutations or intragenomic recombinations. So we can take the upper bound for *de novo* mutations to be 53, then to get the average number of *de novo* mutations per genome we can divide by 84,

which is the total number of sequences in the combined 5 and 500 ng dataset, 53/84=0.631.

Using this information derived for the average number of bases changed by recombination and mutation in a single genome, we can calculate the expected average r/m for a genome as 240/0.631 = 380.3, for the combined 84 sequence dataset. The actual value should be higher due to the fact that a region can experience multiple recombination events, so this would be a lower bound estimate. With Rhometa we do observe a value greater than 380.3 of 1922.97 for the 84 sequence dataset.

2.5.4 Evaluation on Real Datasets - Ocean Metagenomic Dataset

Table 2.3. Port Hacking analysis

Contig_ID	**edge_21626**	**edge_14330**	**edge_29441**
Lineage	**Cyanobacteria, Cyanobacteria, PCC-6307, Cyanobiaceae (A)**	**Cyanobacteria, Cyanobacteria, PCC-6307, Cyanobiaceae (B)**	**Actinobacteriota, Acidimicrobiia, Actinomarinales, Actinomarinaceae**
Coverage depth	40	25	729
Contig length	1029283 bp	926756 bp	797490 bp
ρ (Seed 0)	2.00	0.38	0.18
ρ (Seed 1)	0.96	1.00	0.21
ρ (Seed 2)	0.94	0.41	0.22
ρ (Seed 3)	0.90	0.28	0.17
ρ (Seed 4)	1.00	0.69	0.13
Mean	1.16	0.55	0.18
Stddev	0.47	0.29	0.04

Rhometa was further tested on real datasets, using samples from Port Hacking, Sydney. We produced estimates using reads mapped to 3 different assembly contigs (Table 2.3). A point of note is that within the family Cyanobiaceae, there is a difference in rates of recombination. It is in such comparisons that Rhometa could prove to be of great value.

2.6 Discussion

Recombination plays a crucial role in microbial evolution and speciation (Levin and Cornejo, 2009; Didelot and Maiden, 2010; Schmutzer and Barraclough, 2019). Understanding the rate at which recombination occurs provides us an insight into the impact of this process. Metagenomics is the only method that allows us to study recombination in real-world natural microbial communities without culture bias (Wooley, Godzik and Friedberg, 2010). However, there are currently no software tools to accurately estimate population recombination rates on large metagenomic datasets. To fill this gap, we have developed Rhometa, a software implementation that builds on the composite likelihood estimator for population recombination rate estimation method and enables interrogation of next generation sequencing reads from shotgun metagenomic experiments. It is adjusted for gene-conversion type recombination as experienced by prokaryotic populations.

Composite likelihood population recombination rate estimators are among the most accurate methods known, and our implementation makes these methods available to the wider metagenomics community. This is significant as most microbes cannot readily be cultured, by some estimates only 1-15% are readily cultivable in laboratories (Singh *et al.*, 2009).

Shotgun metagenomics yields reads from microbes taken directly from the natural environment and mitigates issues related to culture dependent studies. However, there has ,until now, not been a viable approach for quantifying the population recombination rate from these reads. PIIM (Johnson and Slatkin, 2009) and mcorr (Lin and Kussell, 2019) come closest to being applicable to shotgun metagenomic datasets, being designed for this use case. However, in the case of PIIM its statistical model uses a very compute intensive approach to account for low quality data, and missing data, including missing coverage, making the compute requirements impractical for large modern datasets.

Meanwhile, with mcorr, the mathematical formulation is distinct from the well-known population recombination rate ($\rho = 2N_e r$), which may represent challenges for interpretation. On an experimental dataset where transformation was used to produce a population of recombinants for sequencing, the approach implemented in mcorr appears to significantly underestimate the recombination rate. Furthermore, mcorr was evaluated on simulated datasets used to test for Rhometa (S2.7 Fig) where we

expected to see an increase in the ϕ_pool estimated as the simulated ρ increases. However, the results showed no correlation between the recombination rate inferred by mcorr and the true simulated rate. Additionally, ϕ_pool, the value for recombination estimated by mcorr complicates interpretability by not using the standard parameter for population recombination (ρ). mcorr's limitation of looking at coding regions provided in a gene annotation file again complicates analysis as this information is not always available. Rhometa builds on established methods and theoretical foundations, without the need for gene annotation files and based on our testing, produces accurate results.

To build our program, our approach was to first reimplement the LDhat pairwise program for gene conversion. Doing so we were able verify that we accurately captured the core algorithms of LDhat while having a modern and adaptable implementation around it. The simulation results for LDhat and the Rhometa_full_genome (Figs 2.2A and 2.2B) show that we were able to reproduce the LDhat results 1:1. Having validated the reimplementation we then adapted it for read based datasets. LDhat is effective at detecting changes in the magnitude of the simulated population recombination rates, and produces accurate estimates for cases with large numbers of genomes (Fig 2.2A). Our analysis showed a trend where the accuracy and variance of the estimates improved as the number of genomes increased.

We then evaluated the performance of Rhometa on simulated datasets and the results (Figs 2.3 and 2.4) demonstrated that the read-based pipeline performs well and consequently represents a successful implementation of the composite likelihood population recombination rate estimator for metagenomic read-based datasets. As with LDhat and Rhometa_full_genome, the performance of the read based pipeline improves with the number of genomes present, having 80 genomes or more produces the best results. However, it should be noted regardless of the number of genomes there is a bias towards underestimation of ρ, though this is small when there are many genomes present (Fig 2.4) . Very small ρ values, those between 0-1 (S2.3 Fig), are an exception as the implementation has a tendency towards over estimation.

Rhometa was further applied to a *S.pneumoniae* transformation experiment (Croucher *et al.*, 2012)**,** where the extent of recombination could be directly quantified. This dataset was also analysed by the authors of mcorr for their paper, enabling a cross comparison between the two programs. Preparing and analysing the dataset (Table 2.2, 84 sequences) in a manner performed by mcorr, it was found that Rhometa was able to accurately detect the higher rate of recombination relative to mutation as was

expected for the dataset. From direct evidence, based on the information provided in the transformation experiment paper, we calculate a conservative lower bound for the ratio of recombination to mutation as r/m > 380.3. Rhometa was able to meet this condition by estimating a r/m value of 1922.97, while mcorr estimated a r/m value of 0.93, which suggests that change due to mutation was greater than recombination, which is extremely unlikely.

Evaluation on real environmental metagenomic samples (Table 2.3) showed that the program works on real datasets. While the analysis performed here was purely to test the applicability of the pipeline, it demonstrates potential for many comprehensive studies. Such as looking at the rate of recombination of a given species over time or across locations.

Rhometa is well positioned to exploit the abundance of preexisting metagenomic datasets to enable a thorough first-pass study of recombination rates in microbial communities.

The main difference between the LDhat approach and the read based approach is as follows. In LDhat the final population recombination rate estimate “is obtained by combining the likelihoods from all pairwise comparisons” (McVean, Awadalla and Fearnhead, 2002), the likelihoods (negative log-likelihoods) come from the pregenerated lookup tables as mentioned in methods. For genome sequence datasets this means we can use a likelihood table generated for the exact number of genomes/depth in a dataset, the number of genomes represents the depth which is fixed and all pairs of sites we look at will have this depth. Additionally, LDhat considers all possible variant site pair combinations across the sequences being analysed.

A complication associated with read-based datasets is that the depth can vary greatly from site to site. We addressed this issue by using an appropriate depth lookup table for the variant site pairs being considered, the rationale for which is that the negative log-likelihoods for the variant site pairs considered is obtained individually and then combined for a final composite negative log-likelihood. Taking into account that the negative log-likelihoods are obtained individually for each pair of variant sites, the variant pair combinations considered are limited to the extent of a read for single-end reads, for the insert length in the case of pair-end reads.

This provides a way for us to group the variant sites by depth, following which we use the appropriate depth negative log-likelihood table to obtain the negative log-likelihoods for each pair and then finally combine the negative log-likelihoods to get a result for the entire dataset. Additionally, we have introduced a novel weighted sum when calculating the composite negative log-likelihood across coverage depth (equation 2). Rhometa thus enables the application of the composite likelihood estimator approach for current shotgun metagenomic datasets.

An important advantage of Rhometa and its use of raw reads over a consensus assembly from each sample, is that the potential microdiversity within each dataset is preserved for analysis.

As discussed, Rhometa performs better the more genomes there are, it is possible to get a minimum count for the number of genomes present when simulated under the coalescent model with recombination. In real metagenomic samples, any single sample may have millions of genomes of the same species, and across samples there may be significant population structure that is not captured by the standard coalescent model with recombination. The relationship between the number of metagenomic samples, the depth of sequencing of each sample, and the genome count in our simulation study is therefore not straightforward.

2.6.1 Limitations and Future Directions

While we have endeavoured to make a complete package with Rhometa that addresses all aspects of population recombination rate, there are a few limitations. One such limitation is the automatic inference of the tract length, which is also not possible with LDhat (McVean, Awadalla and Fearnhead, 2002) or PIIM (Johnson and Slatkin, 2009). In the context of the composite likelihood approach, the authors of both LDhat and PIIM suggest that while it may be theoretically possible to co-estimate the population recombination rate and tract length, in practice it is challenging. Instead, following the examples of LDhat and PIIM, Rhometa fixes the mean tract length for population recombination rate estimation. As observed by the authors of PIIM, tract length tends to rescale the population recombination rate estimate and large misspecifications can cause further deviations (Johnson and Slatkin, 2009). In our tests with simulated datasets where the tract length was varied, it was found that misspecifying the tract length tends to linearly scale the final estimates in a predictable manner (S2.6 Fig). If a fixed value of tract length is given by the user for all analyses

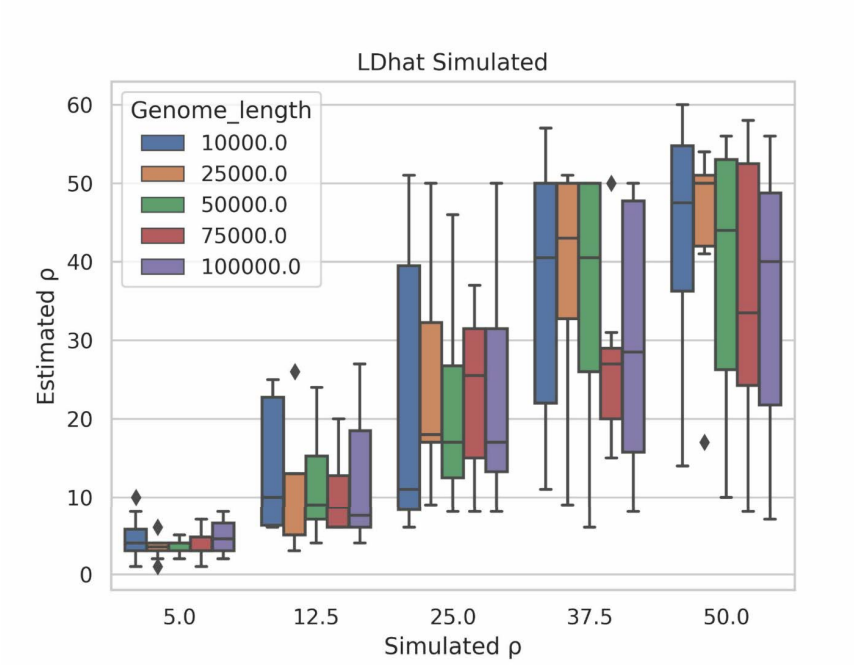
(i.e. an accurate sample-specific tract length is not provided by the user), then differences in inferred rho among samples could be due to either a change in the recombination rate or a change in tract length. The user will need to take this into account when interpreting the results. For instance, if performing same species comparisons across time and location, the tract length can be fixed to determine relative changes in the magnitude of recombination.

Furthermore, the nature of our method is not sensitive to very low rates of recombination as observed when attempting to evaluate rates between 0-1 and we suggest exercising caution for such fine scale analysis.

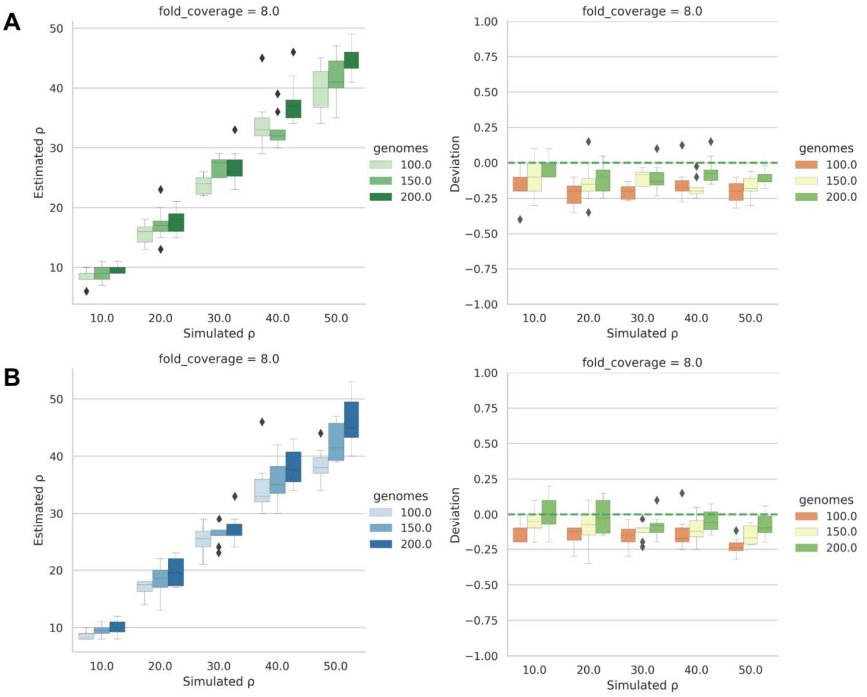
Another point of note concerns the generation of the lookup tables for the program. While it is relatively fast to generate lookup tables due to the incorporation of LDpop, it can still require substantial time and resources for a high-resolution table with a large number of genomes. Generation of lookup tables require specification of θ , in our tests we have found that Rhometa is fairly robust to misspecifications of θ. It is only extreme misspecifications, such as misspecifications by a factor of 10, that seem to have a large impact (S2.5 Fig).

We believe the availability of a tool such as Rhometa, which can be easily applied to current metagenomic datasets is timely and significantly expands the range of habitats and therefore microbial communities that can be studied for recombination, giving us an insight into the extent to which they can adapt and speciate. How ρ varies within environments and between taxa is unknown, Rhometa can help investigate many such fundamental questions related to the evolution and survival capacity of microbes. With the aid of data analysis techniques, metagenomic datasets can be further combined with environmental and sequencing metadata to help study the intricacies of recombination. Many ecological factors can modulate and effect recombination (González-Torres *et al.*, 2019). Synthesis of other data types with the results of our program may yield a clearer understanding of such relationships. We have built our approach in a modular and easy to adapt manner making this and similar applications easy to explore in the future.

2.7 Supporting Information



S2.1 Fig. Results of varying simulated genome lengths for testing LDhat (number of genomes fixed at 10, tract length 500).
https://doi.org/10.1371/journal.pgen.1010683.s001
(TIF)

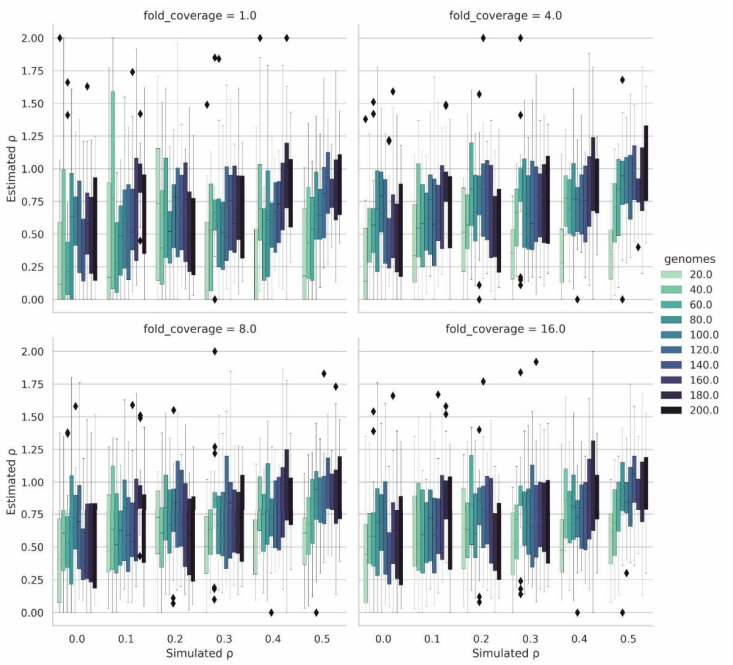


S2.2 Fig. Comparing simulated single end and paired end read datasets in Rhometa.

(**A**) Single end results. (**B**) Paired end results

https://doi.org/10.1371/journal.pgen.1010683.s002

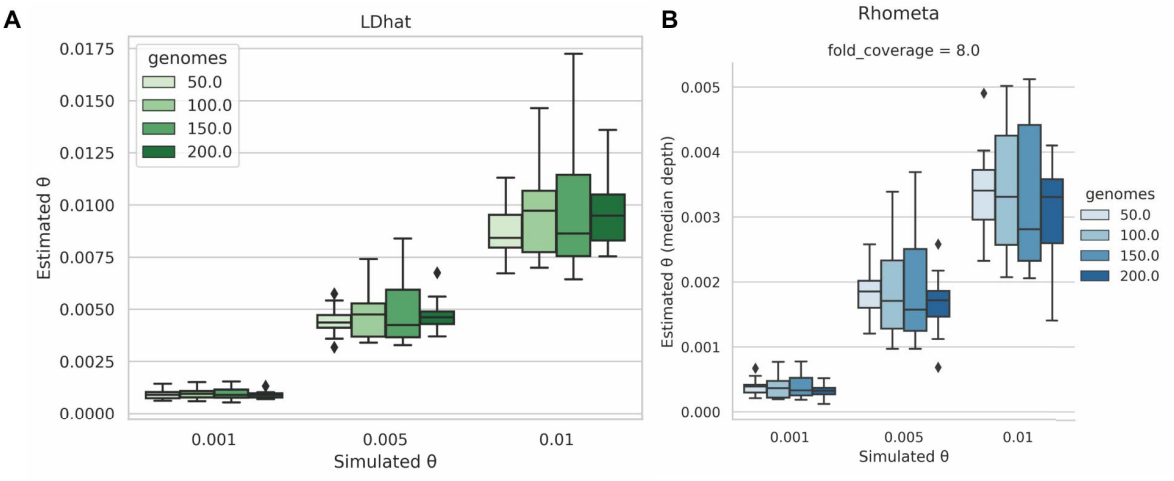
(TIF)



S2.3 Fig. Simulated vs Estimated population recombination rate (ρ) results for Rhometa. Results for varying numbers of simulated genomes and fold coverage values for population recombination rates 0.0, 0.1, 0.2, 0.3, 0.4, 0.5. The simulation parameters used are the same as for population recombination rates [10.0, 20.0, 30.0, 40.0, 50.0], except lookup tables for population recombination rates 0-2 were used (0-2 in 201 steps) for depths of 3-200.

https://doi.org/10.1371/journal.pgen.1010683.s003

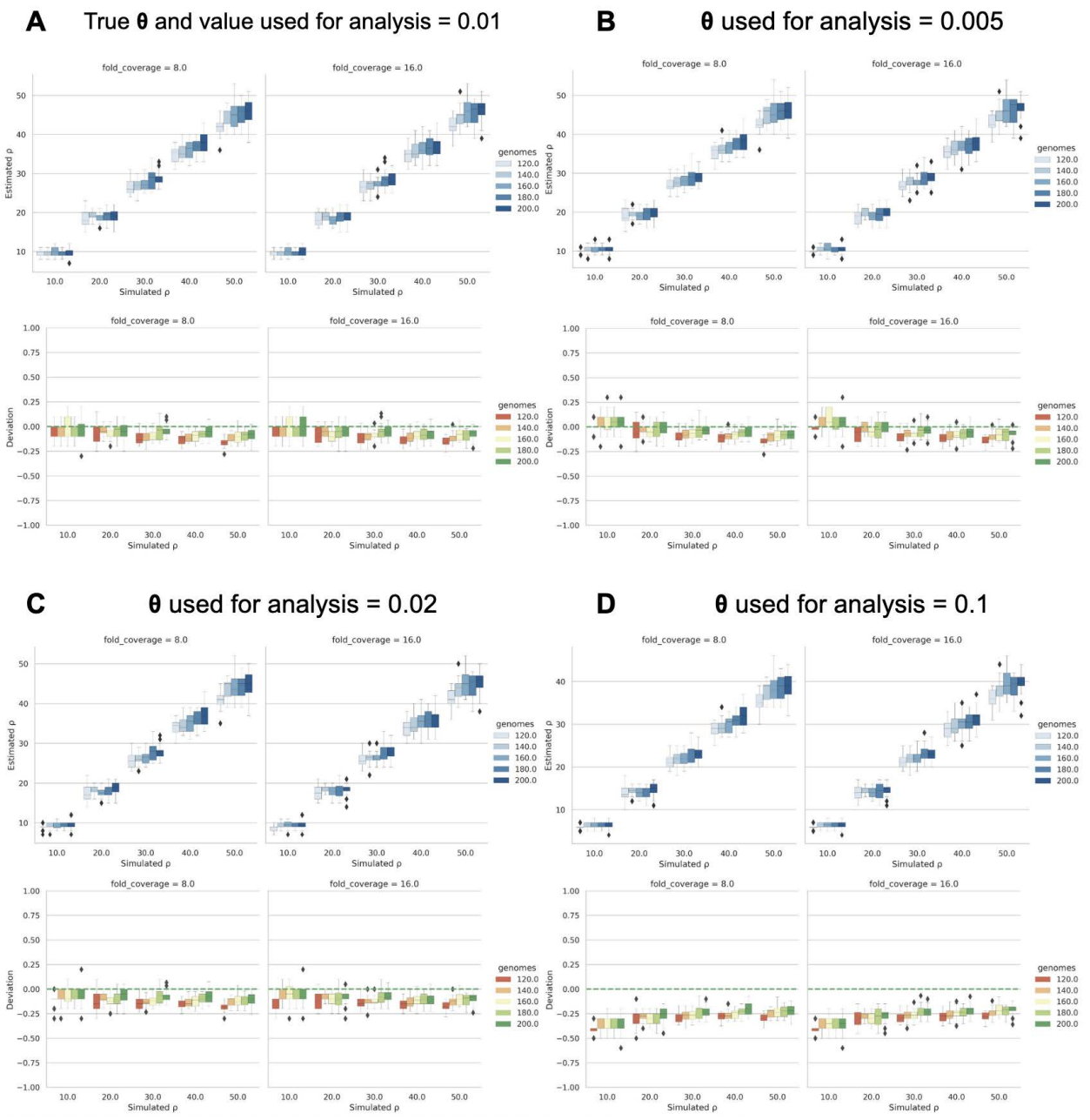
(TIF)



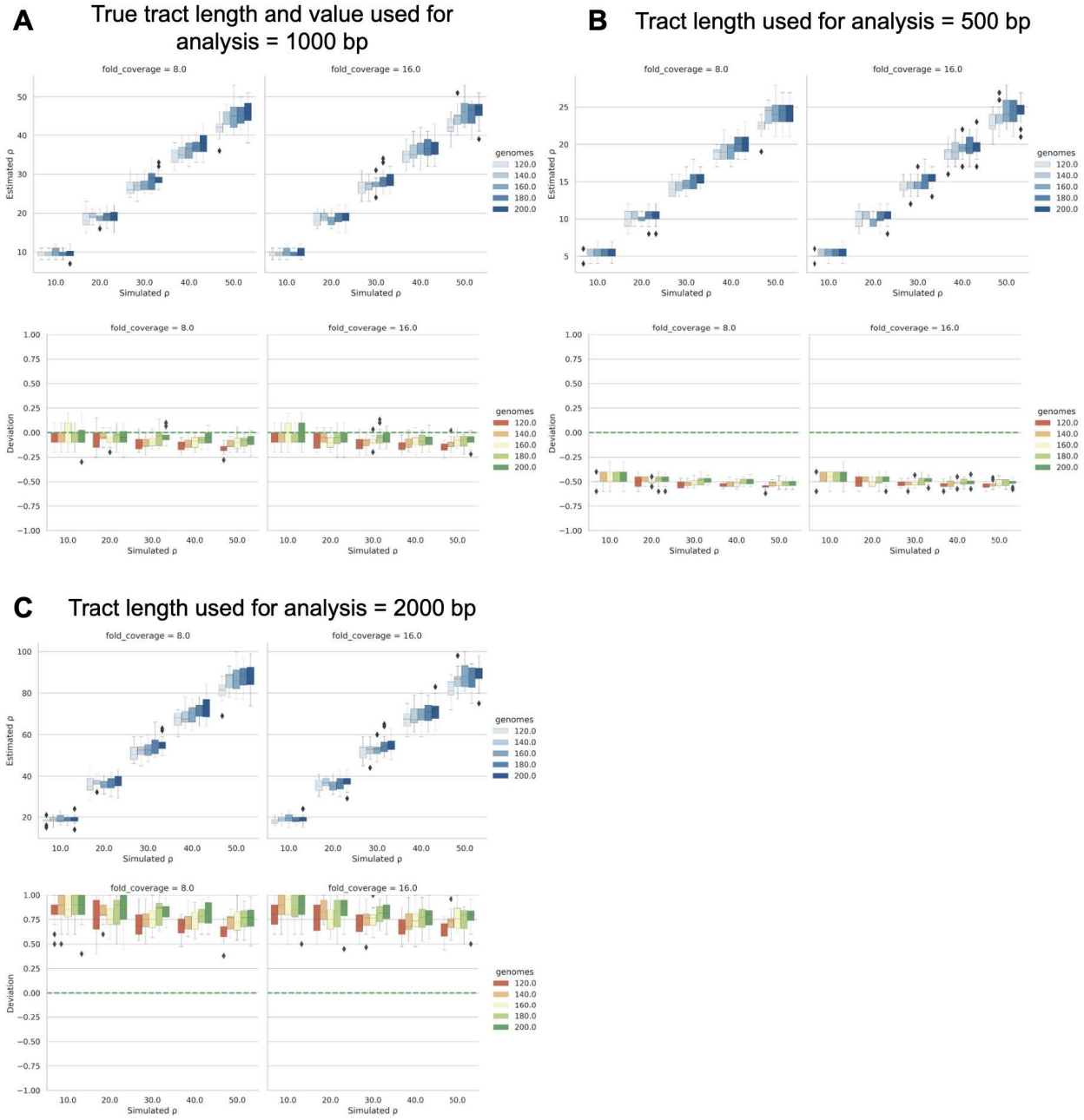
S2.4 Fig. Simulated vs Estimated theta per site (θ) results for LDhat and Rhometa. (**A**) LDhat. Simulated vs Estimated theta per site (θ) for varying number of simulated bacterial genomes. (**B**) Rhometa. Simulated vs Estimated theta per site (θ) for varying number of simulated bacterial genomes.

https://doi.org/10.1371/journal.pgen.1010683.s004

(TIF)



S2.5 Fig. Results of analyzing simulated datasets with lookup tables generated under misspecified θ (genome length fixed at 100,000) (**A**) Simulated datasets analyzed with true θ (0.01) lookup tables (**B**) Simulated datasets analyzed with misspecified θ (0.005) lookup tables (**C**) Simulated datasets analyzed with misspecified θ (0.02) lookup tables (**D**) Simulated datasets analyzed with misspecified θ (0.1) lookup tables. https://doi.org/10.1371/journal.pgen.1010683.s005 **(TIF)**

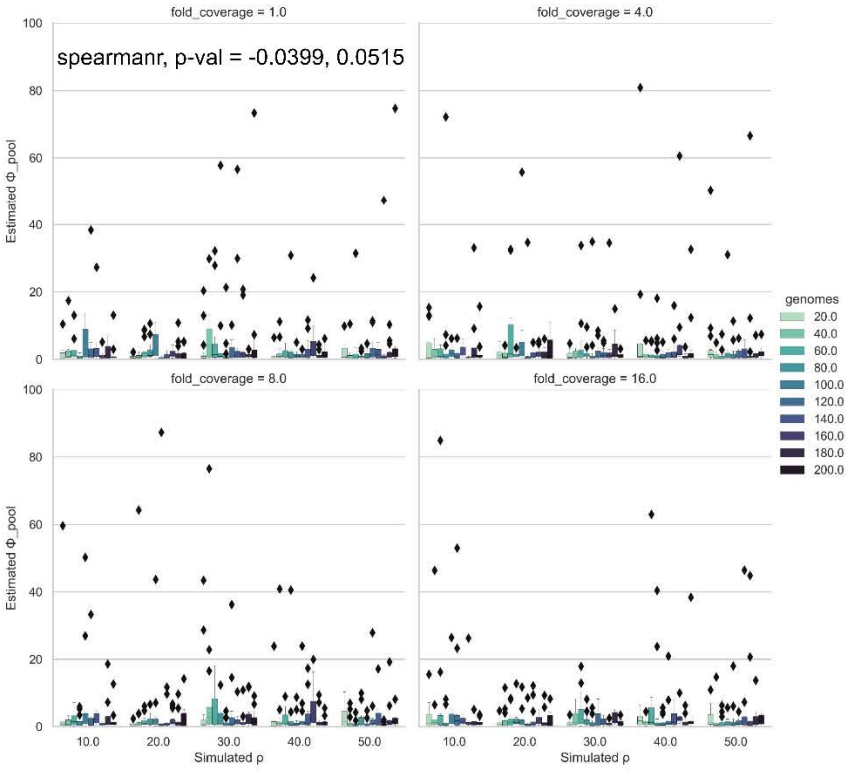


S2.6 Fig. Results of analyzing simulated datasets with misspecified tract lengths (genome length fixed at 100,000. θ 0.01 lookup tables used**)** (**A**) Simulated datasets analysed with true tract length 1000 bp (**B**) Simulated datasets analyzed with misspecified tract length 500 bp (**C**) Simulated datasets analysed with misspecified tract length 2000 bp.

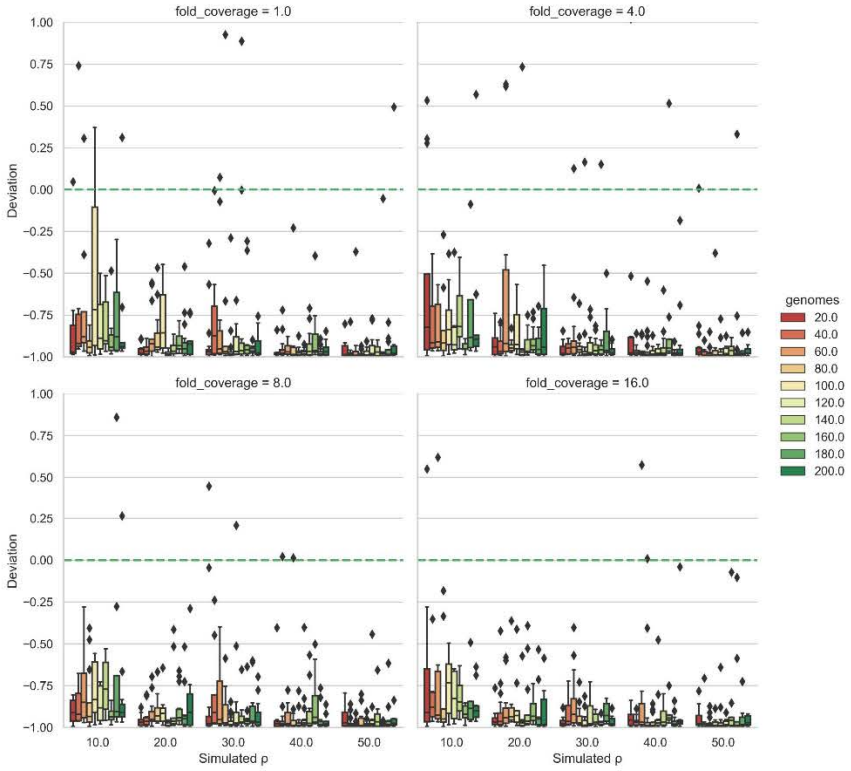
https://doi.org/10.1371/journal.pgen.1010683.s006

(TIF)

A


fold_coverage = 1.0
fold_coverage = 4.0
fold_coverage = 8.0
fold_coverage = 16.0
spearmanr, p-val = -0.0399, 0.0515
Estimated Φ_pool
Simulated ρ
genomes
20.0
40.0
60.0
80.0
100.0
120.0
140.0
160.0
180.0
200.0

B


fold_coverage = 1.0
fold_coverage = 4.0
fold_coverage = 8.0
fold_coverage = 16.0
Deviation
Simulated ρ
genomes
20.0
40.0
60.0
80.0
100.0
120.0
140.0
160.0
180.0
200.0

S2.7 Fig. Mcorr tested using the simulated datasets used for rhometa (Figs 2.3, 2.4). mcorr uses BAM (aligned reads) and gff (gene annotations) files as inputs. For the gene annotations, coding regions need to be provided for mcorr. With the simulated data every 1000 bp was defined as a coding region (CDS). Out of 4000 simulations, some could not be completed and many extremely large unrealistic realistic values. A filter was used where such instances were discarded and only phi_pool values < 100 were kept. After filtering there were only 2386/4000 values.

https://doi.org/10.1371/journal.pgen.1010683.s007

(TIF)

S2.1 Table. Analysis results of s_pnemoniae transformation. Results for all seed values

https://doi.org/10.1371/journal.pgen.1010683.s008 **(DOCX)**

Full sequence population recombination rate			
Seed value	**exp-1 5ng**	**exp-1 500ng**	**84 sequences**
0	**6.8**	**4.4**	**4.7**
1	**7.7**	**6.4**	**5.8**
2	**3.4**	**4.9**	**3.7**
3	**5.7**	**7.7**	**4.9**
4	**4.2**	**2.7**	**3.3**
Mean	**5.56**	**5.22**	**4.48**
Per-site population recombination rate (Full seq/tract length)			
Seed value	**exp-1 5ng**	**exp-1 500ng**	**84 sequences**
0	**0.00296**	**0.00191**	**0.00204**
1	**0.00335**	**0.00278**	**0.00252**
2	**0.00148**	**0.00213**	**0.00161**
3	**0.00248**	**0.00335**	**0.00213**
4	**0.00183**	**0.00117**	**0.00143**
Mean	**0.00242**	**0.00227**	**0.00195**
Tract length	**2300**		

S2.2 Table. *S.pneumoniae* experiment accession codes

https://doi.org/10.1371/journal.pgen.1010683.s009

(DOCX)

S2.3 Table. Accession codes for Port Hacking, Sydney datasets (from SRA run table)

https://doi.org/10.1371/journal.pgen.1010683.s010

(DOCX)

S2.1 Appendix Lookup configuration

Matching against the lookup table

https://doi.org/10.1371/journal.pgen.1010683.s011

(DOCX)

Data Availability: All supporting information can be accessed here: https://doi.org/10.5281/zenodo.7634208. The Rhometa software package is available at https://github.com/sid-krish/Rhometa. The metagenomic dataset simulation pipeline is available at https://github.com/sid-krish/rhometa_sim, the LDhat Nextflow Pipeline is available at: https://github.com/sid-krish/Nextflow_LDhat, the full genome version of Rhometa, developed for testing purposes, is available at https://github.com/sid-krish/Rhometa_Full_Genome and the Nextflow_LDhat_sim simulation pipeline (used for simulating full sequences for both Rhometa Full Genome and LDhat Nextflow Pipeline) is available at: https://github.com/sid-krish/Nextflow_LDhat_Sim.

Chapter 3:
Spatiotemporal patterns in mutation and recombination rates among major marine bacterioplankton

Sidaswar Krishnan[a], Martin Ostrowski[a], Matthew Z. DeMaere[b], Aaron E. Darling[b,c], Kittikun Songsomboon[a], Dominik Beck[d], Justin R. Seymour[a]

[a]Climate Change Cluster, University of Technology Sydney, Sydney, NSW, Australia

[b]Australian Institute for Microbiology & Infection, University of Technology Sydney, Sydney, NSW, Australia

[c]Lamnoo Inc., Iowa City, IA, USA

[d]Centre for Health Technologies and the School of Biomedical Engineering, University of Technology Sydney, Sydney, NSW, Australia

The order of authors reflects the level of their contributions, with the final author serving as the project lead.

3.1 ABSTRACT

Understanding the frequency of mutation and recombination in natural microbial communities is critical to uncovering the mechanisms of microbial evolution. Yet, these rates remain poorly characterised in environmental settings. To address this gap, we applied *Rhometa*, a novel bioinformatic tool for detecting mutation and recombination events, to metagenomic datasets collected over seven years from two long-term oceanographic monitoring sites in the Tasman Sea, eastern Australia. Our objectives were to (i) compare recombination-to-mutation ratios (ρ/θ) across key marine bacterial genera, (ii) examine seasonal variation in ρ/θ, and (iii) investigate environmental drivers influencing ρ/θ within pelagic bacterial communities.

We quantified ρ/θ in bacterial genera belonging to the families Cyanobiaceae, Pelagibacteraceae, and Rhodobacteraceae. Our analysis revealed significant variation in ρ/θ patterns among genera and between sites. Pelagibacter consistently exhibited high ρ/θ values across both seasons and locations, whereas Synechococcus showed considerable spatial and temporal variability. Moreover, Pelagibacter and Synechococcus exhibited distinct correlation profiles between ρ/θ and environmental parameters, indicating that the environmental drivers of recombination and mutation differ between taxa and locations. These findings highlight the complex interplay between genetic exchange processes and environmental factors in marine bacterioplankton, and provide new insights into the evolutionary forces shaping the marine microbiome.

3.2 IMPORTANCE

Marine bacteria are fundamental to ocean productivity and global biogeochemical cycles, yet the evolutionary dynamics that shape their populations remain poorly understood. In particular, the relative roles of mutation and recombination in natural marine environments have not been well characterised. Here, we estimated recombination-to-mutation ratios (ρ/θ) in several dominant marine bacterial genera using long-term metagenomic datasets from two sites in the Tasman Sea. Our analysis revealed marked variation in ρ/θ across genera and between locations. Pelagibacter consistently exhibited higher recombination relative to mutation, while Synechococcus displayed substantial spatial and temporal variability. These patterns suggest that evolutionary processes in marine bacterioplankton are not uniform, but instead reflect lineage-specific and environmentally driven dynamics. By uncovering these patterns,

our findings contribute new insight into the mechanisms that govern microbial evolution in ocean ecosystems.

3.3 INTRODUCTION

Bacteria play key roles in shaping ecosystem structure and function, yet the evolutionary forces that drive their diversification in natural environments remain poorly understood. Mutation and recombination are central to prokaryotic evolution, with mutation introducing changes to a cell's DNA sequence and recombination involving the exchange and incorporation of genetic material between cells (Hanage, 2016). Mutation rates in bacteria may be influenced by the surrounding environment, with stressors such as nutrient deprivation triggering stress-induced mutagenesis (SIM), leading to elevated mutation rates (Ferenci, 2019). Recombination is also shaped by environmental context, with rates affected by donor-recipient similarity, spatial proximity between cells, and the functional compatibility of exchanged DNA (Popa and Dagan, 2011). Abiotic factors such as temperature, pH, nutrient availability, and oxygen levels can also modulate recombination frequency (Aminov, 2011; Le *et al.*, 2020). Given the pronounced spatial and temporal variability in environmental conditions and microbial community composition across natural ecosystems (Fuhrman *et al.*, 2006; Wang *et al.*, 2023), the relative importance of mutation and recombination is likely to be heterogeneous and context-dependent.

Given the complex and variable environmental conditions that shape mutation and recombination in natural systems, microbes are best studied as they are found in the natural environment. However, much of our current understanding of microbiology, including the dynamics of mutation and recombination, derives from culture-based approaches (Handelsman, 2004). While isolate-based studies have shown that the relative importance of these processes can differ across bacterial species (Vos and Didelot, 2009), they are inherently limited in their ability to capture the ecological interactions and environmental pressures that influence microbial evolution (Handelsman, 2004; Singh *et al.*, 2009). In contrast, metagenomic approaches allow researchers to study microbial communities directly in situ, through the collection and sequencing of environmental DNA. Shotgun metagenomic sequencing involves extracting DNA from all cells within a community, fragmenting it, and sequencing it without the need for culturing (Sharpton, 2014). Recent advances in bioinformatics now enable the reliable estimation of population recombination rates (ρ) and mutation rates (θ) directly from such datasets (Krishnan *et al.*, 2023), creating new opportunities to

investigate these evolutionary processes in natural microbiomes and to explore how they respond to environmental variation.

Building on these advances, measuring the relative importance of recombination and mutation within natural microbial communities, such as the pelagic marine microbiome, offers a powerful means to uncover the evolutionary forces shaping microbial diversity in the ocean. In this study, we used *Rhometa* to analyse shotgun metagenomic data from two long-term oceanographic time-series stations in the southwestern Pacific Ocean, separated by eight degrees of latitude.

Our goal was to estimate the relative recombination-to-mutation event probability (ρ/θ), a metric that captures the balance between these two evolutionary processes, for marine bacterial genera belonging to the families Cyanobiaceae, Pelagibacteraceae, and Rhodobacteraceae. These families are critically important because they represent some of the most abundant and influential bacterial lineages in the ocean: Pelagibacteraceae dominates marine prokaryotic communities and is central to carbon cycling (Morris *et al.*, 2002; Giovannoni, 2017), Cyanobiaceae such as *Synechococcus* (Flombaum *et al.*, 2013) are key drivers of oceanic primary production, and Rhodobacteraceae, including Roseobacter, play major roles in both carbon and sulfur cycling (Luo and Moran, 2014).

Here we investigate: (i) how ρ/θ varies across major bacterial genera, (ii) whether these ratios differ by location and season, and (iii) which environmental factors are associated with variation in ρ/θ within pelagic microbial communities.

3.4 RESULTS

3.4.1 Recombination to mutation ratios differ significantly across marine bacterial genera and locations

We analysed shotgun metagenomic data from two long-term oceanographic time-series sites in the eastern Australian marine environment: Port Hacking (34°05′ S, 151°15′ E) and Maria Island (42°35′ S, 148°14′ E), with samples collected between 2013 and 2020. These sites are located within the Tasman Sea and are characterised by high temporal variability in physical, chemical, and microbial conditions (Messer *et al.*, 2020; Doane *et al.*, 2023). Port Hacking tends to exhibit more dynamic and less predictable seasonal patterns, whereas Maria Island is generally characterised by more stable and repeatable seasonal variation.

We first performed a between-site comparison of ρ/θ values for bacterial genera present at both Port Hacking and Maria Island to assess spatial differences in recombination and mutation dynamics. Three genera, *Pelagibacter*, *Synechococcus_E*, and Rhodobacteraceae member *LGRT01*, were detected at both sites and allowed for direct comparison. ρ/θ values for *Synechococcus_E* and *Pelagibacter* differed significantly between locations (Wilcoxon test, p = 2.4 × e-4, r = 0.38 and p = 1.7 × e-4, r = 0.39, respectively), while no significant difference was observed for *LGRT01* (Fig 3.1; S3.1 Table). Additional taxa analysed at Port Hacking included *Roseobacter HIMB11*, *Pelagibacter_A*, and *Synechococcus RCC307*, while Maria Island samples included Rhodobacteraceae member *Planktomarina*. These genera were selected based on genome abundance within each site.

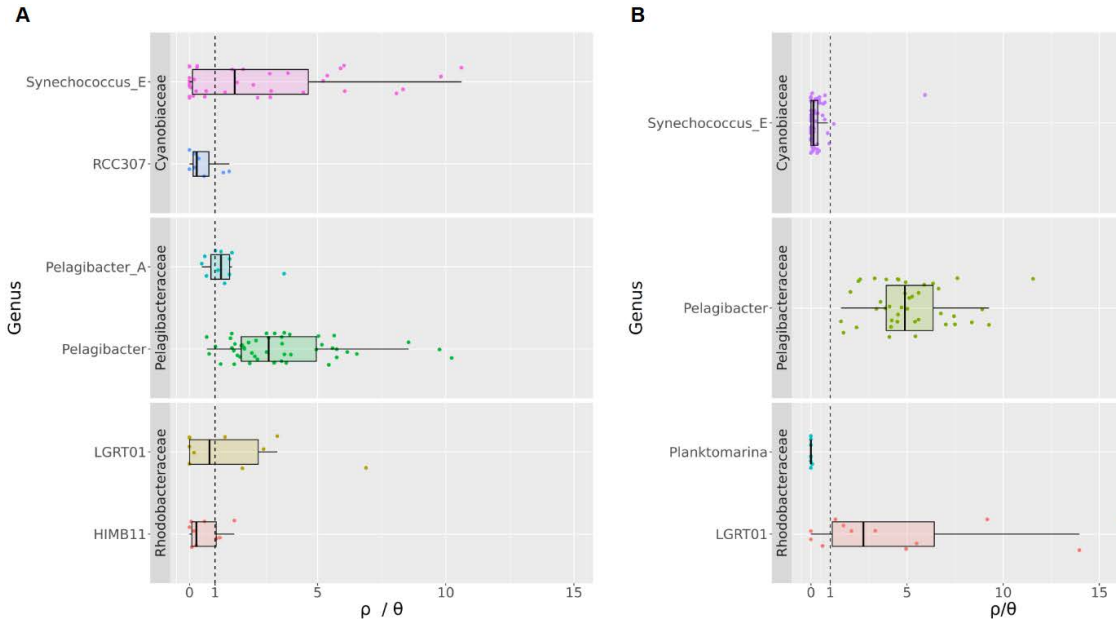


Fig 3.1. Recombination to mutation event ratios at Port Hacking and Maria Island. The vertical dashed line at ρ/θ=1 indicates equal rates of recombination and mutation. Values to the right of the line represent higher recombination rates relative to mutation. Within each sample, multiple analyses were conducted for different species, and cumulative results were reported per genus. (**A**) ρ/θ values for the genera studied at Port Hacking. (**B**) ρ/θ values for the genera studied at Maria Island.

We next conducted a within-site comparison of ρ/θ values to examine how recombination and mutation dynamics vary across bacterial taxa at each location. At Port Hacking, there were statistically significant differences in ρ/θ values between the

targeted bacterial genera (Kruskal–Wallis test, p = 1.32e-06, η^2 = 0.26). *Pelagibacter* exhibited the highest mean ρ/θ of 3.6 (SD ± 2.2), followed by Rhodobacteraceae member *LGRT01* at 1.7 (SD ± 2.2). In contrast, *Synechococcus RCC307* had the lowest mean ρ/θ at 0.5 (SD ± 0.6), while *Roseobacter HIMB11* also displayed a mean ρ/θ below 1 (0.6, SD ± 0.6), suggesting a greater relative importance of mutation over recombination in these taxa.

Similarly, at Maria Island, there were significant differences in ρ/θ values between genera (Kruskal–Wallis test, p < 2.2e-16, η^2 = 0.67). *Pelagibacter* again showed the highest mean ρ/θ at 5.1 (SD ± 2.1), and *LGRT01* also exhibited a relatively high mean of 4.9 (SD ± 5.5). In contrast, *Planktomarina*, another member of the Rhodobacteraceae, had the lowest mean ρ/θ of 0.008 (SD ± 0.023). Notably, *Synechococcus_E* displayed a mean ρ/θ of 0.3 (SD ± 0.8) at Maria Island, in contrast to 2.7 (SD ± 3.1) at Port Hacking. This marked difference suggests a shift in the relative importance of recombination and mutation between the two environments.

Table 3.1. Port Hacking and Maria Island ρ/θ mean values per genus, with standard deviations in brackets.

Port Hacking		**Maria Island**	
Genus	**Mean ρ/θ (std)**	**Genus**	**Mean ρ/θ (std)**
Roseobacter HIMB11	0.576 (0.619)	LGRT01 (Rhodobacteraceae)	4.919 (5.505)
LGRT01 (Rhodobacteraceae)	1.685 (2.247)	Pelagibacter	5.142 (2.145)
Pelagibacter	3.559 (2.156)	Planktomarina (Rhodobacteraceae)	0.008 (0.023)
Pelagibacter_A	1.36 (0.874)	Synechococcus_E	0.323 (0.786)
Synechococcus RCC307	0.526 (0.595)		
Synechococcus_E	2.718 (3.085)		

3.4.2 Seasonal trends in recombination to mutation ratios

Seasonal averages of ρ/θ values differed in magnitude and variability between sites. At Port Hacking, mean ρ/θ values averaged across all targeted genera for each season were highest in autumn (3.5, SD ± 3.3) and winter (3.4, SD ± 2.1), and lowest in spring (1.2, SD ± 1.2). In contrast, Maria Island exhibited more stable seasonal trends, with overall means fluctuating only slightly between 2.2 (SD ± 2.9) in autumn and 2.7 (SD ± 2.8) in spring. These results indicate more pronounced seasonal dynamics at Port Hacking relative to Maria Island.

Further variation was observed when ρ/θ values were examined by genus and season. At Port Hacking, genus-by-season mean ρ/θ values ranged from 0.09 (SD ± 0.09) to 5.2 (SD ± 3.3) (S3.3 Table), while at Maria Island, the corresponding values ranged from 0.02 (SD ± 0.04) to 7.09 (SD ± 6.6) (S3.4 Table). This broad range reflects substantial variability in evolutionary dynamics across bacterial taxa and seasons.

At Port Hacking, Synechococcus_E was the only genus to exhibit significant seasonal variation (Kruskal–Wallis test, $p = 0.004$, $\eta^2 = 0.33$). This genus displayed elevated ρ/θ values in autumn (5.2, SD ± 3.3) and winter (4.0, SD ± 2.4), and lower values in spring (0.8, SD ± 1.3) and summer (1.6, SD ± 2.9), with a significant difference detected between summer and autumn (Wilcoxon test, $p = 0.036$, $r = 0.58$) (Fig 3.2; S3.2 and S3.3 Tables). Other genera at Port Hacking, including Pelagibacter, Pelagibacter_A, Roseobacter HIMB11, Rhodobacteraceae *LGRT01* and Synechococcus RCC307, showed relatively stable ρ/θ values across seasons. However, clear differences between taxa were observed. Pelagibacter consistently exhibited high ρ/θ values in all seasons ($\rho/\theta > 2$), suggesting a dominant role for recombination. Pelagibacter_A also maintained values above 1 in all seasons except winter, whereas Synechococcus RCC307 and Roseobacter HIMB11 had ρ/θ values below 1, indicating a greater influence of mutation.

At Maria Island, seasonal variation in ρ/θ was limited across all genera. No significant seasonal differences were observed for any taxon (Fig 3.2; S3.2 and S3.4 Tables). Pelagibacter again maintained high ρ/θ values (>4) throughout the year, consistent with the pattern at Port Hacking. In contrast, Synechococcus_E at Maria Island exhibited mean values below 1 in all seasons, suggesting a stronger role for mutation at this site relative to Port Hacking.

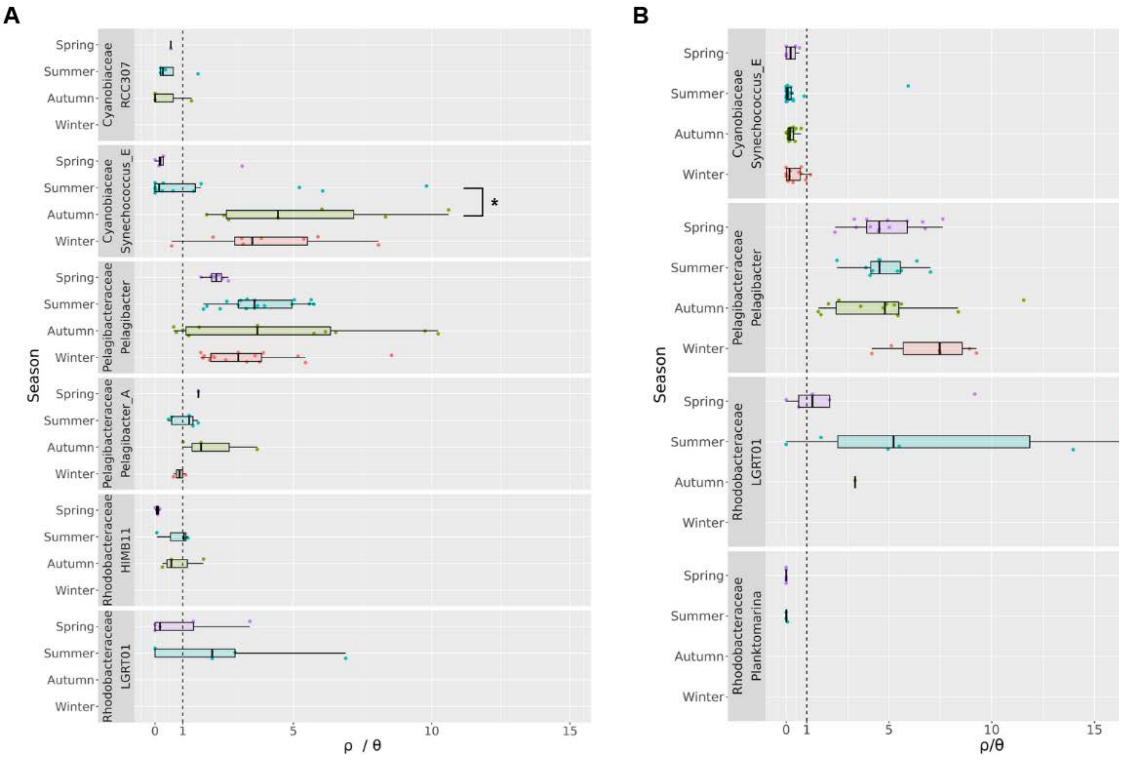


Fig 3.2. Port Hacking and Maria Island ρ/θ values for genera by seasons. The vertical dashed line at ρ/θ=1 indicates equal rates of recombination and mutation. Values to the right of the line represent higher recombination rates relative to mutation. Within each sample, multiple analyses were conducted for different species, and cumulative results were reported per genus. (**A**) ρ/θ values for the genera at Port Hacking across seasons. (**B**) ρ/θ values for the genera at Maria Island across seasons.

3.4.3 Relationships between ρ/θ and environmental parameters

Several significant correlations were detected between ρ/θ values and environmental parameters across the full dataset, which combined samples from both locations (Fig 3.3). Among the genera common to both sites, *Pelagibacter* ρ/θ values were positively correlated with dissolved inorganic carbon (DIC) and negatively correlated with ammonium, temperature, salinity, and alkalinity. In contrast, *Synechococcus_E* ρ/θ values showed positive correlations with ammonium, and Secchi depth, and a negative correlation with daylight. *LGRT01* (Rhodobacteraceae) did not exhibit any significant correlations in the combined analysis.

When analysed separately by site, distinct correlation patterns emerged. At Port Hacking (Fig 3.4A), *Synechococcus_E* ρ/θ values were positively correlated with

silicate, and Secchi depth, and negatively correlated with daylight. For *Pelagibacter*, ρ/θ was negatively correlated only with Secchi depth. Among the Rhodobacteraceae genera, *Roseobacter HIMB11* was the only taxon to exhibit significant correlations, with ρ/θ values positively associated with Secchi depth.

At Maria Island (Fig 3.4B), *Pelagibacter*, was the only genus to exhibit correlations between ρ/θ values and environmental factors, which showed a negative correlation with ammonium.

Lastly, analysis of mean environmental values grouped by species, season, and location revealed substantial variation in physicochemical conditions (S3.5 and S3.6 Tables). This heterogeneity highlights the distinct environmental profiles encountered by each genus across space and time, and underscores the potential for ecological context to influence recombination and mutation dynamics.

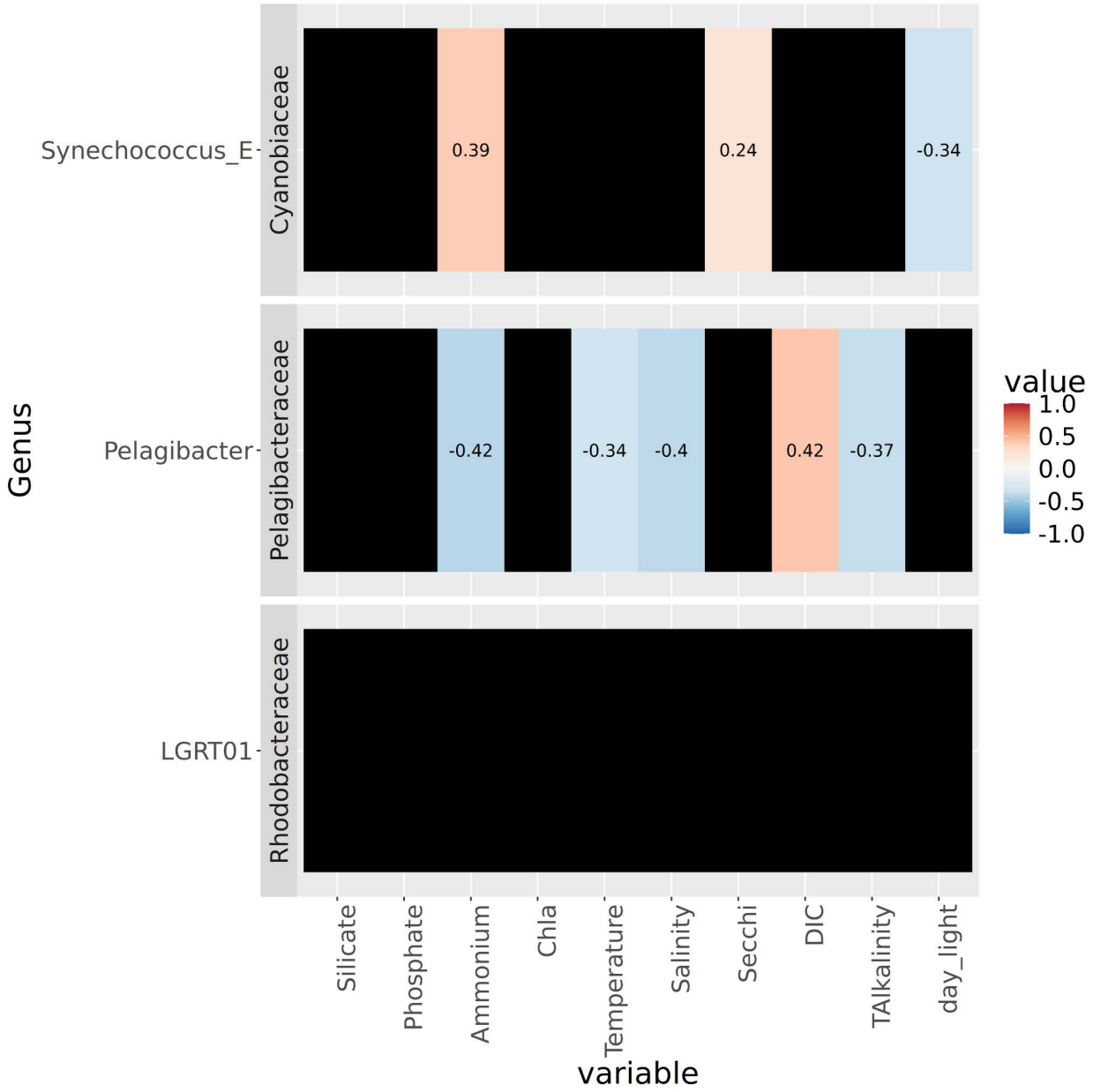


Fig 3.3. Correlation matrix showing relationships between ρ/θ and environmental variables using combined data from Port Hacking and Mariah Island studies. Analysis includes genera common to both locations. Correlation coefficients are displayed only where statistically significant ($p<0.05$); non-significant correlations appear as black cells.

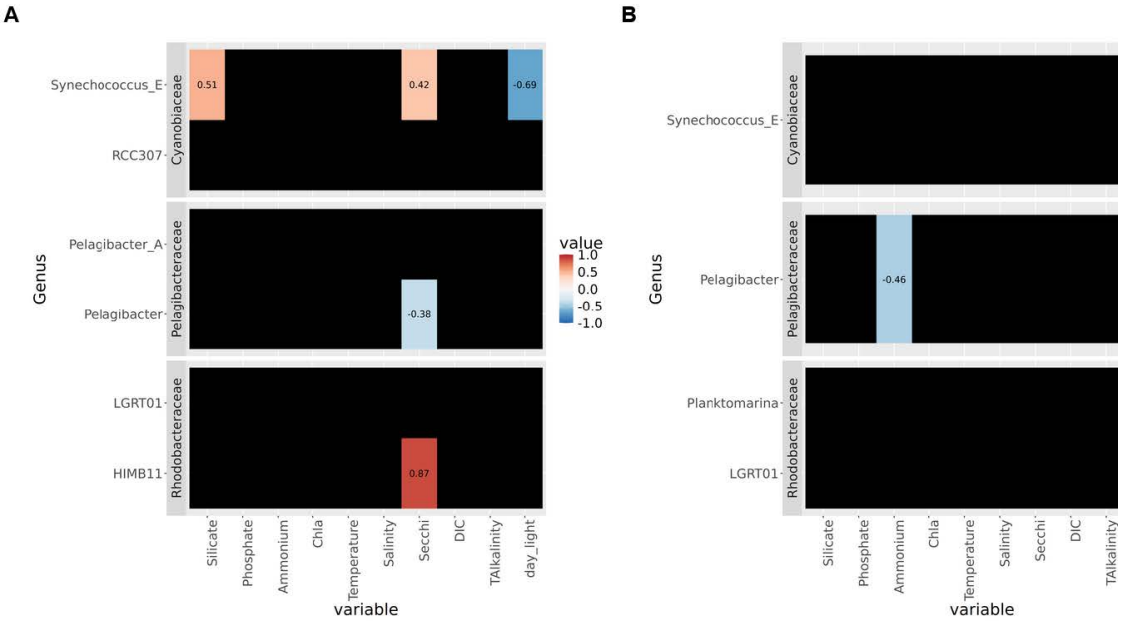


Fig 3.4. Correlation matrix showing relationships between ρ/θ and environmental variables for Port Hacking and Maria Island. Correlation coefficients are displayed only where statistically significant ($p<0.05$); non-significant correlations appear as black cells. (**A**) Port Hacking correlation matrix. (**B**) Maria Island correlation matrix.

3.5 DISCUSSION

This study demonstrates that the relative contributions of recombination and mutation to microbial evolution are both taxon-specific and environmentally responsive in marine systems. While recombination and mutation are widely recognised as core drivers of prokaryotic evolution, their dynamics in natural microbial communities, especially across time and space, remain poorly understood. Our analysis of long-term metagenomic data from two oceanographic time-series sites revealed that ρ/θ values vary substantially across bacterial genera and are shaped by seasonal and site-specific environmental conditions. These findings highlight the importance of measuring evolutionary processes in situ and suggest that ecological context plays a key role in modulating the balance between recombination and mutation.

3.5.1 Within location variation in recombination to mutation ratios across genera

Significant variation in ρ/θ values was observed among bacterial genera at both Port Hacking and Maria Island, suggesting that different taxa within the same environment may be subject to distinct evolutionary pressures. At Port Hacking, *Pelagibacter* exhibited the highest mean ρ/θ (3.6), while *Synechococcus RCC307* had the lowest (0.5). A similar pattern was observed at Maria Island, with *Pelagibacter* again displaying the highest value (5.1), and *Planktomarina* (Rhodobacteraceae) the lowest (0.008). These findings indicate that even under shared environmental conditions, the balance between recombination and mutation can differ markedly across taxa, ranging from highly recombinant to highly clonal lineages.

Pelagibacter a member of the Pelagibacteraceae family, is the most numerically abundant marine bacterium, accounting for up to 25% of all prokaryotic cells in the ocean, and plays a central role in marine carbon cycling (Morris *et al.*, 2002; Giovannoni, 2017). Given its ecological dominance, understanding the evolutionary forces acting on this lineage is essential. Previous studies using isolates have highlighted a strong influence of recombination in *Pelagibacter* (Vergin *et al.*, 2007; Vos and Didelot, 2009), a finding supported by recent metagenomic work showing consistently high ρ/θ values in this genus compared to other taxa (López-Pérez *et al.*, 2020). That study also reported relatively uniform recombination rates across surface-ocean lineages of *Pelagibacter*, with lower values observed in freshwater and marine bathytype clades. Our findings align with and extend this pattern by showing

consistently high ρ/θ values at both Port Hacking and Maria Island, supporting the idea that recombination is a dominant force in the evolution and ecological success of *Pelagibacter* in surface marine environments.

The cyanobacterium *Synechococcus* is a dominant photoautotrophic picoplankton and plays a major role in oceanic primary production (Flombaum *et al.*, 2013). In contrast to *Pelagibacter*, we observed generally lower ρ/θ values in *Synechococcus*, indicating a comparatively weaker influence of recombination. This pattern is consistent with previous findings (López-Pérez *et al.*, 2020), although the species analysed in that study differ from those examined here. These patterns suggest that recombination plays a comparatively smaller role in *Synechococcus* evolution relative to *Pelagibacter*.

Roseobacter, a member of the Rhodobacteraceae family, can account for up to 20% of coastal marine microbial communities and plays a critical role in the global carbon and sulfur cycles (Luo and Moran, 2014). At Port Hacking, *Roseobacter HIMB11* exhibited a mean ρ/θ below 1, indicating a stronger relative influence of mutation than recombination. This finding is consistent with (Wang *et al.*, 2020), who reported similarly low ρ/θ values (0.076) in *Roseobacter* isolates. In contrast, we observed higher ρ/θ values (>1) for another Rhodobacteraceae member, *LGRT01*, at both Port Hacking and Maria Island. The divergent ρ/θ values between *HIMB11* and *LGRT01* challenge the assumption that closely related taxa exhibit similar evolutionary dynamics, highlighting the need for lineage-specific resolution in microbial evolution studies.

3.5.2 Between location variation in recombination to mutation ratios

Of the bacterial genera found at both Port Hacking and Maria Island, *Pelagibacter* and *Synechococcus_E* exhibited significant differences in ρ/θ values between sites. This suggests that the relative importance of recombination and mutation in these lineages is influenced by environmental context. Port Hacking is a sub-tropical site that experiences substantial influence from the East Australian Current (EAC) through its dynamic flow and eddy field. In contrast, Maria Island is a temperate site with cooler waters and more predictable seasonal cycles, and is only intermittently affected by the EAC (Brown *et al.*, 2018). These environmental differences are reflected in previous studies showing site-specific microbial community structure and seasonal dynamics (Messer *et al.*, 2020; O'Brien *et al.*, 2022). This variation in ρ/θ across sites highlights how local environmental conditions can influence microbial evolutionary processes in key marine taxa such as *Pelagibacter* and *Synechococcus_E*.

3.5.3 Environmental impact and seasonal trends in recombination to mutation ratios

To understand how variability in time and environment may shape microbial evolution, we analysed seasonal patterns in ρ/θ values across sites and taxa, alongside correlations with key environmental parameters.

At Port Hacking, ρ/θ values averaged across all targeted bacterial genera were highest in autumn and winter, and lowest in spring. This seasonal pattern aligns with periods of intensified influence from the East Australian Current (EAC), which has been shown to shape microbial community composition during late summer and autumn at this site (Messer *et al.*, 2020). The overlap suggests that EAC-driven shifts in microbial diversity may contribute to increased recombination during these periods. However, when ρ/θ was examined at the level of individual genera, only *Synechococcus_E* exhibited significant seasonal variation. This pattern, limited to Port Hacking and occurring between summer and autumn, was not observed at Maria Island, despite similarly pronounced seasonal environmental variability.

Seasonal variation in *Synechococcus_E* ρ/θ at Port Hacking may reflect nutrient-driven shifts in recombination dynamics linked to EAC influence. The highest ρ/θ values were observed in autumn and winter, coinciding with periods of elevated silicate, phosphate, and ammonium concentrations. These nutrient increases occur during the well-mixed season between May and September (Brown *et al.*, 2018), and may promote microbial growth and cell–cell interactions, which are known to facilitate recombination (Popa and Dagan, 2011). Autumn is also among the seasons most strongly impacted by EAC-driven turnover in microbial composition (Messer *et al.*, 2020), further supporting a link between environmental shifts and elevated recombination. In contrast, ρ/θ values for *Synechococcus_E* at Maria Island remained consistently below 1 across all seasons. The reason for this is unclear, but higher relative mutation rates at this site may reflect greater environmental stress or ecological constraints on recombination.

In contrast to the variability observed in *Synechococcus_E*, *Pelagibacter* maintained consistently high ρ/θ values across seasons and sites, indicating a stable recombination-dominated evolutionary mode. Values consistently exceeded 2 at Port Hacking and 4 at Maria Island, regardless of season. This pattern suggests that recombination plays a dominant role in *Pelagibacter* evolution, possibly due to its high population abundance, which may enhance opportunities for gene exchange through

frequent cell–cell interactions, a notion also postulated by previous studies (Vergin *et al.*, 2007; López-Pérez *et al.*, 2020).

In the combined-site analysis, *Pelagibacter* exhibited a greater number of significant correlations with environmental parameters than in either individual site analysis. This suggests that *Pelagibacter*'s genomic responses are shaped by broader environmental gradients that only emerge when both locations are examined together. Further, a correlation with temperature is observed.While the mechanisms behind this correlation is not currently clear, previous studies have shown strong links between temperature and SAR11 niche divergence (Brown *et al.*, 2012).

In contrast, LGRT01 (*Rhodobacteraceae*) and other *Rhodobacteraceae* genera showed few, if any, significant correlations across analyses. These taxon-specific patterns underscore the complexity of microbial–environment interactions and highlight the importance of analyzing community dynamics at multiple spatial scales.

Taken together, these findings show that environmental factors such as seasonality, nutrient availability, and temperature may influence the balance between recombination and mutation, but their effects vary across taxa and locations. Some genera, like *Pelagibacter*, maintain stable ρ/θ values despite environmental change, while others, like *Synechococcus_E*, show marked temporal and spatial variability. This highlights the importance of considering both taxonomic identity and environmental context when interpreting microbial evolutionary dynamics in the ocean, while recognising that these patterns represent a snapshot of the community at the time of sampling and are subject to methodological biases and constraints inherent to metagenomic approaches.

3.6 CONCLUSION

This study advances our understanding of how mutation and recombination contribute to microbial evolution in natural marine ecosystems. By applying Rhometa, a metagenomic tool for quantifying recombination-to-mutation ratios (ρ/θ), we assessed evolutionary dynamics in key bacterial genera across two long-term oceanographic time-series. Our results revealed substantial taxonomic variation in ρ/θ, as well as distinct spatial and seasonal patterns. *Pelagibacter* exhibited consistently high ρ/θ values across time and location, suggesting a recombination-dominated evolutionary strategy, while *Synechococcus* showed marked variability, with recombination rates more sensitive to environmental context. Correlation analyses further highlighted taxon-specific associations between ρ/θ and physicochemical variables, including

temperature, nutrients, and water clarity. Together, these findings demonstrate that recombination and mutation are not uniform across marine bacteria, but instead respond to lineage-specific and environmental factors, with implications for how microbial populations evolve under changing ocean conditions.

3.7 MATERIALS AND METHODS

Shotgun metagenomic data from two long-term oceanographic time-series sites in the eastern Australian marine environment were analysed to investigate mutation and recombination patterns within marine bacterial assemblages. Samples were collected between 2013 and 2020 from the Port Hacking (34°05.00' S, 151°15.00' E) and Maria Island (42°35.80' S, 148°14.00' E) National Reference Stations (NRS). Port Hacking is a subtropical site substantially influenced by the East Australian Current (EAC) and its associated eddy field, whereas Maria Island, is a temperate site located in Tasmania, which experiences seasonal influence from the southern extent of the EAC (Brown *et al.*, 2018).

Seawater samples were collected and microbial DNA was extracted following protocols described in (Brown *et al.*, 2018). Briefly, seawater was collected from multiple depths using Niskin bottles, and microbial cells were retrieved by filtering 2 L of seawater through 0.2 µm pore Sterivex GP filters (Millipore, Massachusetts; Cat. # SVGPL10RC) using a Masterflex L/S Compact peristaltic pump fitted with an L/S 8 channel head (Cole Parmer). DNA was extracted and purified using the DNeasy® PowerWater® Sterivex™ DNA Isolation Kit (Qiagen, Germany), following a slightly modified version of the manufacturer's instructions. DNA quality and quantity were assessed using a NanoDrop™ 8000 Spectrophotometer (Thermo Scientific™).

Metagenomic libraries were prepared using the Illumina Nextera XT protocol and sequenced on the Illumina HiSeq 2500 platform following Australian Microbiome procedures (https://research.csiro.au/ambsm/). Library preparation and sequencing were performed at the Ramaciotti Centre for Genomics, University of New South Wales, Australia.

In parallel with microbial sampling, environmental parameters were measured and compiled following the Integrated Marine Observing System National Reference Station sampling protocols (Lynch *et al.*, 2008, 2014). Environmental metadata used in this study include, silicate, phosphate, ammonium, chlorophyll a (Chl-a), temperature,

salinity, depth, Secchi depth, dissolved inorganic carbon (DIC), and total alkalinity (TAlkalinity), and were sourced from the AusMicrobiome Microbial Ocean Atlas project (data provided in Supplementary Files). Day length at each sampling event was calculated using the Python package Astral (https://astral.readthedocs.io/en/latest/).

3.7.1 Reference genome selection

Analysis was performed on FASTQ files from 131 metagenomic datasets were extracted from the Australian Microbiome Data Portal (data.bioplatforms.com). To determine the species to investigate, the 50 most abundant reference genomes across the Port Hacking and Maria Island samples were identified, of which 47 were available for download and subsequent analysis. Reference genome assemblies were selected based on their recruitment of the highest proportion of reads mapping to the seven conserved ribosomal protein genes in the SingleM reference collection, accounting for over 40% of all short reads across the 131 metagenomes.

From these, reference genomes corresponding to the numerically and ecologically important marine bacterial families Cyanobiaceae, Pelagibacteraceae, and Rhodobacteraceae were prioritised. Notably, 84% of the top 50 abundant reference genomes represented members of these families. Reference genome metadata was obtained from GTDB release v207 (Parks *et al.*, 2022) and merged with accessions for analysis, and all representative genome assemblies were downloaded from NCBI.

To address taxonomic ambiguities arising from closely related species, metagenomic reads were grouped at higher taxonomic levels, specifically genus and family. This approach provided a practical solution for managing uncertainties in species-level relationships. Within each sample, multiple analyses were conducted for different species, and cumulative results were reported per genus. The complete lists of families, genera, and species included in the analysis are provided in the Supplementary Files.

In this study, some genera include an alphabetic suffix introduced by the GTDB classification system. For example, within the Pelagibacteraceae family at Port Hacking, both *Pelagibacter* and *Pelagibacter_A* were identified. According to the GTDB FAQ, such suffixes (e.g., '_A') indicate either (i) polyphyletic genera, based on the current GTDB reference tree, or (ii) subdivisions introduced to maintain taxonomic rank normalisation.

3.7.2 Mutation and recombination rate estimation

Population mutation rates (θ) and recombination rates (ρ) were estimated for the targeted genera using Rhometa version 1.0.8. Reads from the Port Hacking and Maria Island datasets were aligned to the selected reference genomes using the align_reads_input_list.nf script with default settings. Mean read coverage of the aligned binary alignment map (BAM) files was then assessed using CoverM (v0.7.0) (Aroney *et al.*, 2024). To account for differences in sequencing depth and to enable meaningful comparisons across samples, coverage was normalised to a mean of 25× using Samtools (v1.18) (Danecek *et al.*, 2021), and BAM files below this threshold were discarded.

To further ensure sufficient genome representation, the percentage of the reference genome covered by reads was calculated using Samtools coverage, and the weighted mean coverage was determined for each BAM file. Only files with a weighted coverage of 80% or greater were retained for downstream analysis.

Values for ρ and θ were calculated from the processed BAM files. Recombination rates (ρ) were estimated using lookup tables generated with depths ranging from 3 to 300 and ρ values spanning 0 to 100, increasing in increments of 0.25 (from 0 to 25), 0.5 (from 25 to 50), and 1 (from 50 to 100). Lookup tables were generated with a constant effective population size of 1, and all other parameters were set to Rhometa defaults for both ρ and θ calculations.

To account for stochastic variation during subsampling, BAM file subsampling was performed using seed values of 1, 2, and 3, and final mean ρ values were calculated by averaging across the three seeds.

Mean ρ per site was calculated by dividing the mean ρ value by the tract length and then by 2, to account for the two breakpoints characteristic of gene conversion-style recombination. The tract length was fixed at 1000 base pairs for all ρ calculations. For mutation rate estimation, θ per site was calculated based on the mean aligned read depth across genomic positions.

Rather than analysing ρ and θ individually, we examined their ratio, ρ/θ. Values less than 1 ($\rho/\theta < 1$) indicate a higher relative frequency of mutation events compared to recombination events, whereas values greater than 1 ($\rho/\theta > 1$) suggest recombination is relatively more significant than mutation. The rationale for using the ratio ρ/θ is that both ρ and θ are compound parameters that implicitly incorporate the effective

population size (N_e) in both the Rhometa and LDhat models (McVean, Awadalla and Fearnhead, 2002; Krishnan *et al.*, 2023). Since N_e varies between samples and is difficult to estimate, interpreting ρ and θ directly can complicate cross-sample comparisons. In contrast, the ratio ρ/θ cancels out N_e and provides a robust measure of the relative contribution of recombination events versus mutation events, per site.

3.7.3 Statistical analysis

To ensure ecological consistency, all statistical analyses were conducted using surface metagenomic samples (0–10 m depth), as microbial communities in surface waters experience distinct environmental conditions compared to those in deeper layers. We retained only bacterial genera with at least five ρ/θ observations per location to ensure robust statistical inference. The final dataset included 6 genera from Port Hacking and 4 genera from Maria Island. Prior to statistical testing, it was confirmed that ρ and θ estimates were centred around the normalised sequencing depth of 25× (S3.1 and S3.2 Figs).

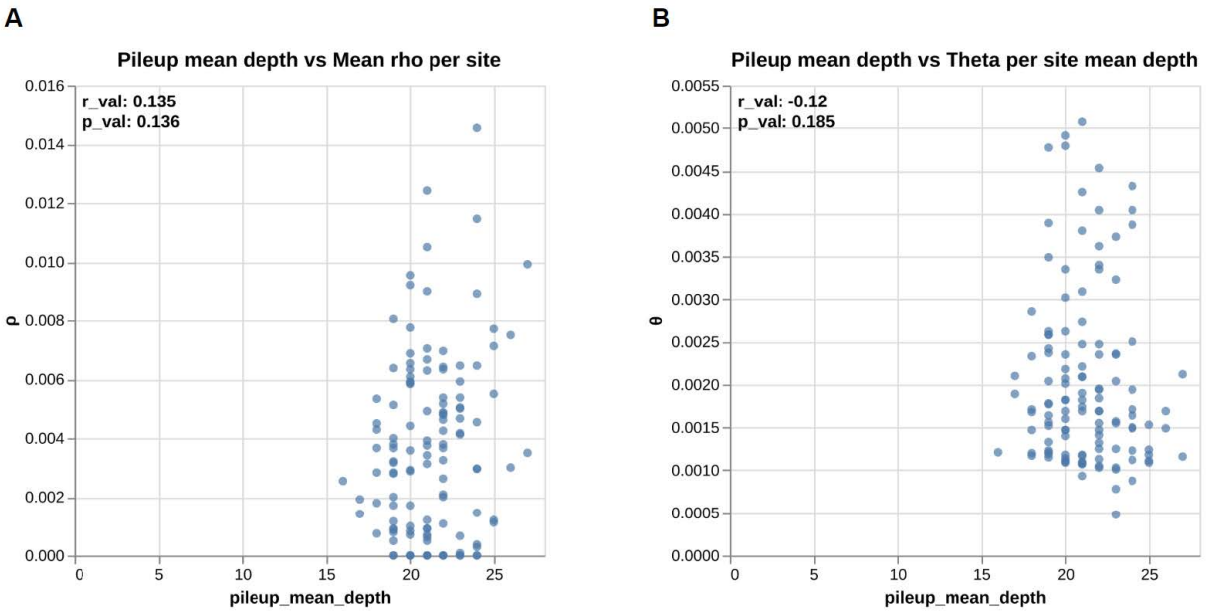
All statistical analyses were performed in R (v4.4) unless otherwise specified. To determine the suitability of parametric versus non-parametric tests for downstream analyses, overall normality of the ρ/θ values at each location was assessed using the Shapiro–Wilk test. As the ρ/θ data were not normally distributed (Shapiro–Wilk test, $p < 0.05$), non-parametric tests were subsequently applied.

To examine whether different marine bacterial genera exhibited differences in recombination-to-mutation ratios, Kruskal–Wallis (KW) tests were conducted within each location and across seasons (significance threshold: $p < 0.05$). Pairwise comparisons of ρ/θ values between locations for each genus were performed using Wilcoxon rank-sum tests with Bonferroni correction (significance threshold: $p < 0.05$). For genera showing significant seasonal variation based on KW results, seasonal differences were further examined using pairwise Wilcoxon tests with Bonferroni correction. Additionally, effect sizes (η^2 for Kruskal–Wallis, r for Wilcoxon) were calculated using the rstatix package in R and are reported alongside p-values.

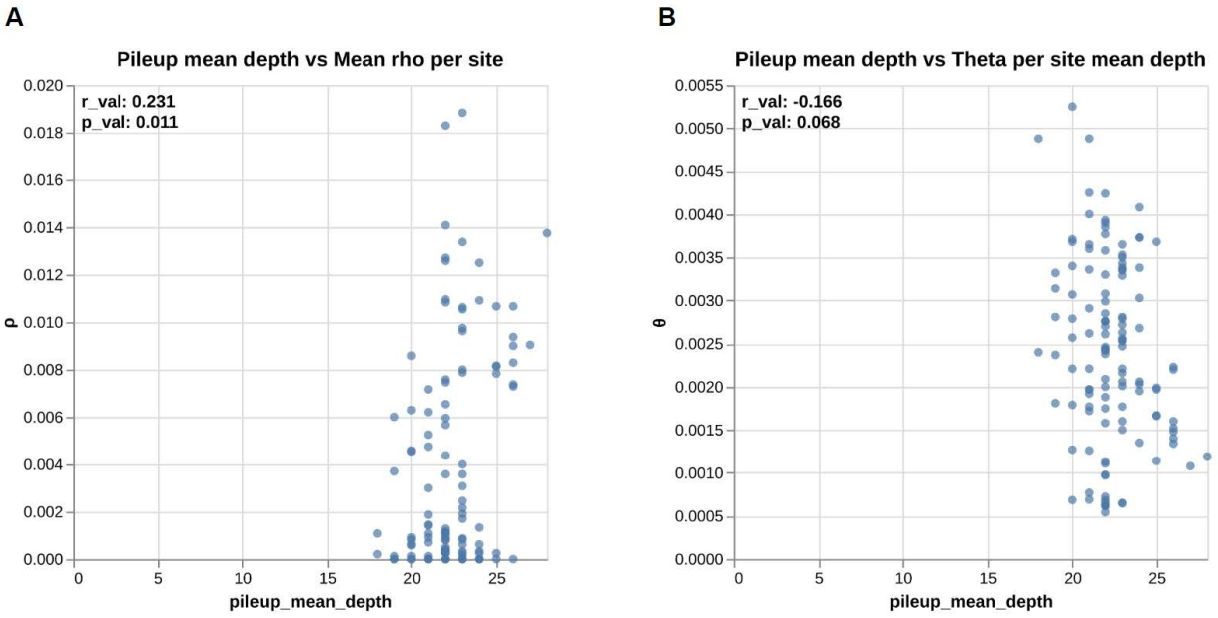
To assess relationships between ρ/θ values and environmental variables, Spearman correlation analyses were conducted using the spearmanr function from the scipy.stats library in Python (v3.12) (Virtanen *et al.*, 2020). P-values for correlations were corrected for type 1 error with false discovery rate (FDR). Mean values of environmental parameters across genera, seasons, and locations were also calculated to support

seasonal comparisons. These tables summarise environmental profiles per genus based on the samples in which they occur. It highlights seasonal fluctuations in environmental conditions relative to each genus in the samples in which they are observed.

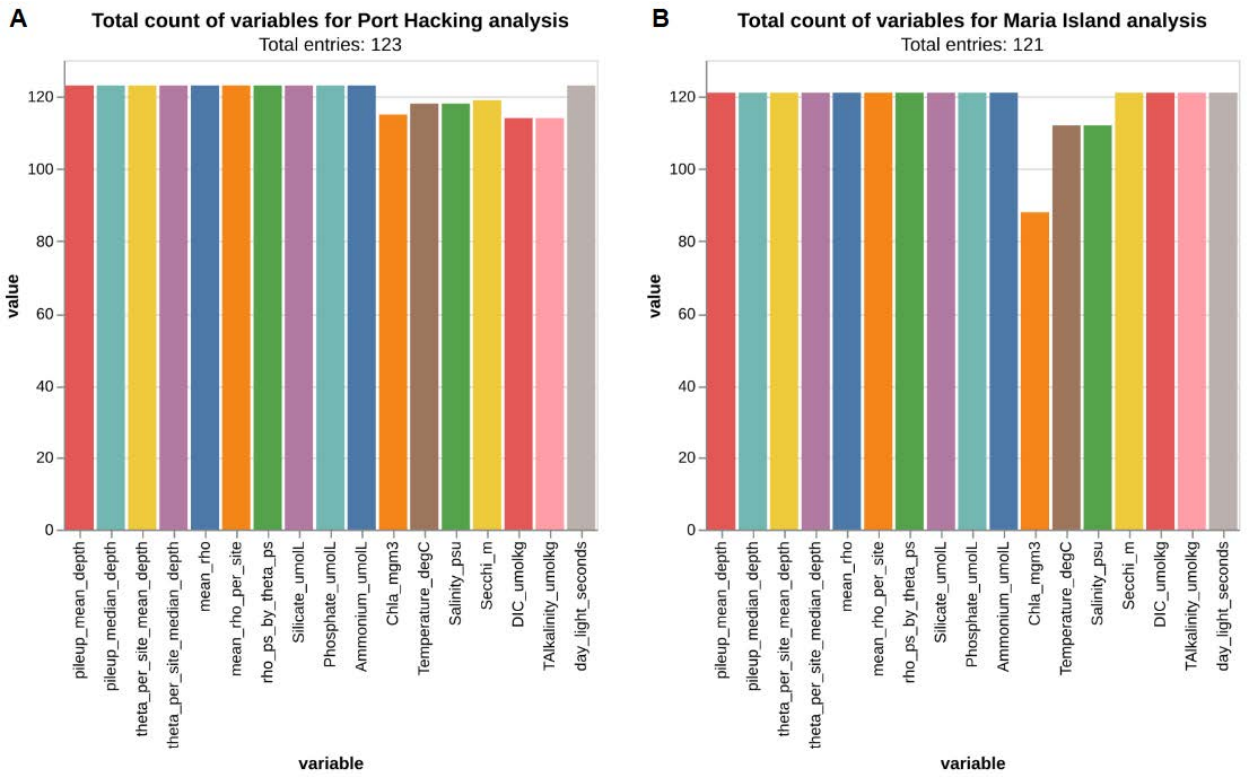
3.8 Supporting Information



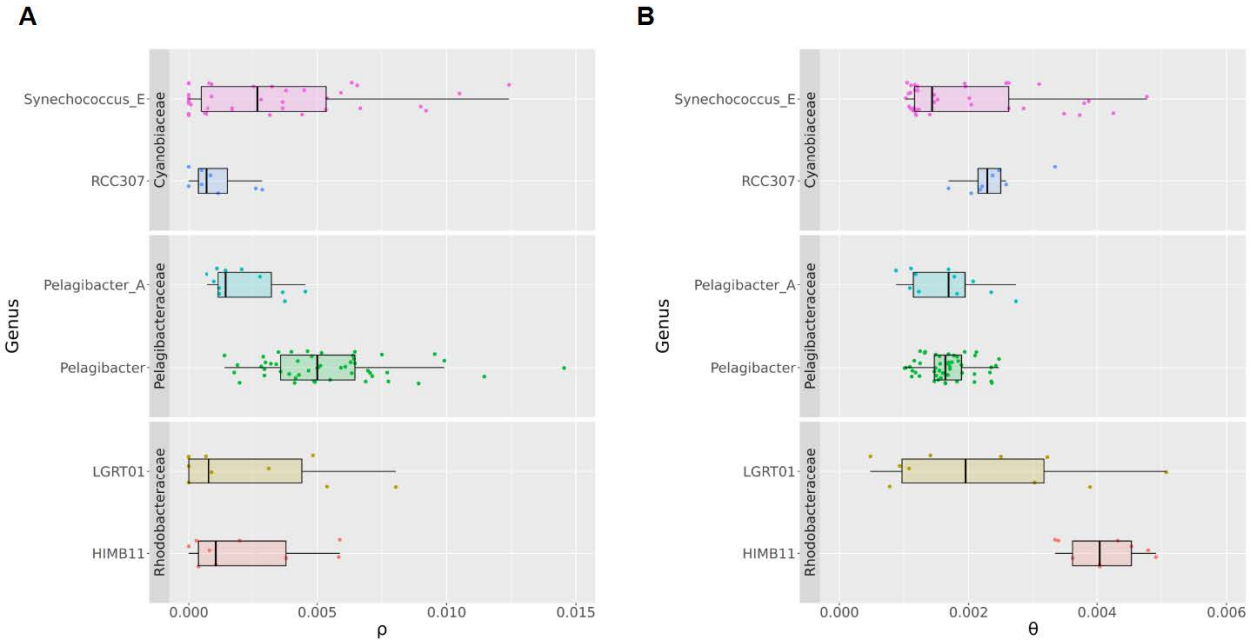
S3.1 Fig. Port Hacking, pileup mean depth vs ρ and θ. After all filters were applied. (**A**) Pileup mean depth vs ρ (**B**) Pileup mean depth vs θ.



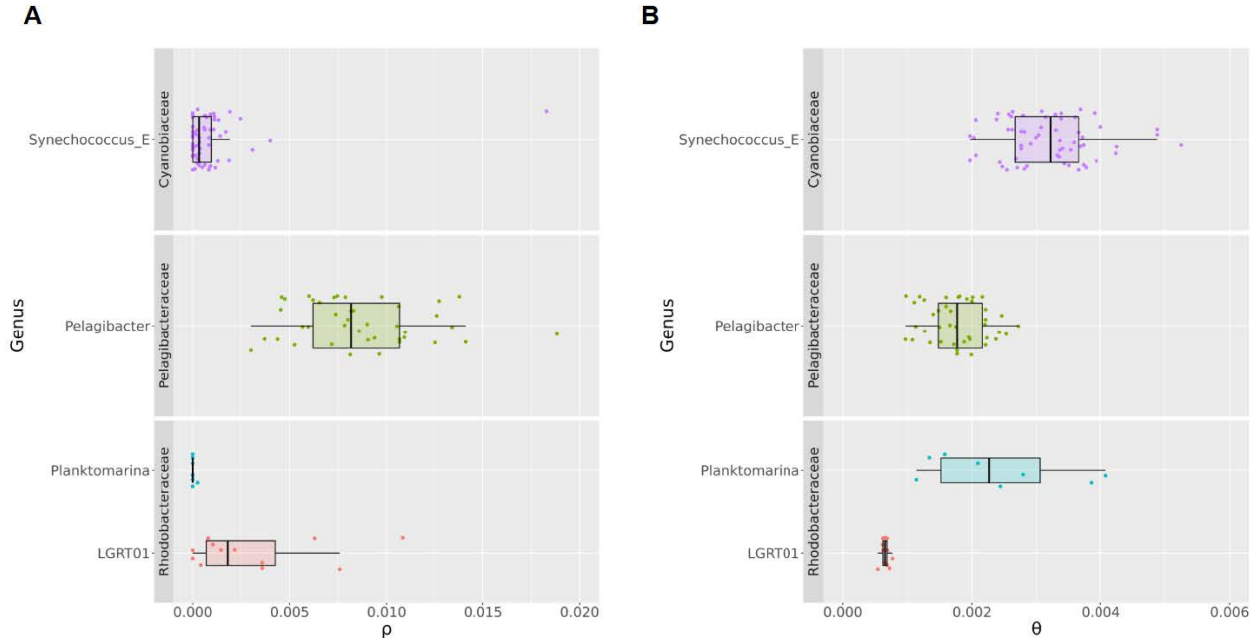
S3.2 Fig. Maria Island, pileup mean depth vs ρ and θ. After all filters were applied. (**A**) Pileup mean depth vs ρ (**B**) Pileup mean depth vs θ.



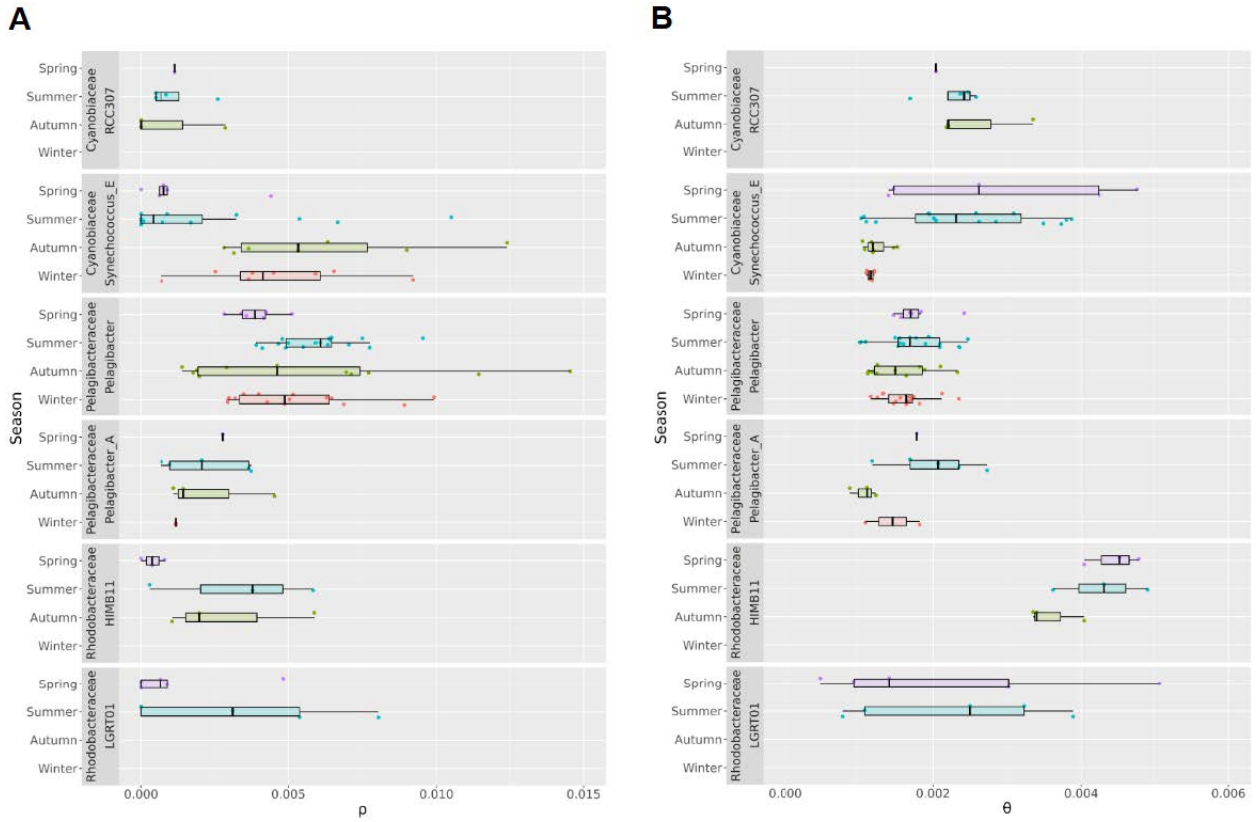
S3.3 Fig. Port Hacking and Maria Island variable counts. (**A**) Variable counts for Port Hacking (**B**) Variable counts for Maria Island.



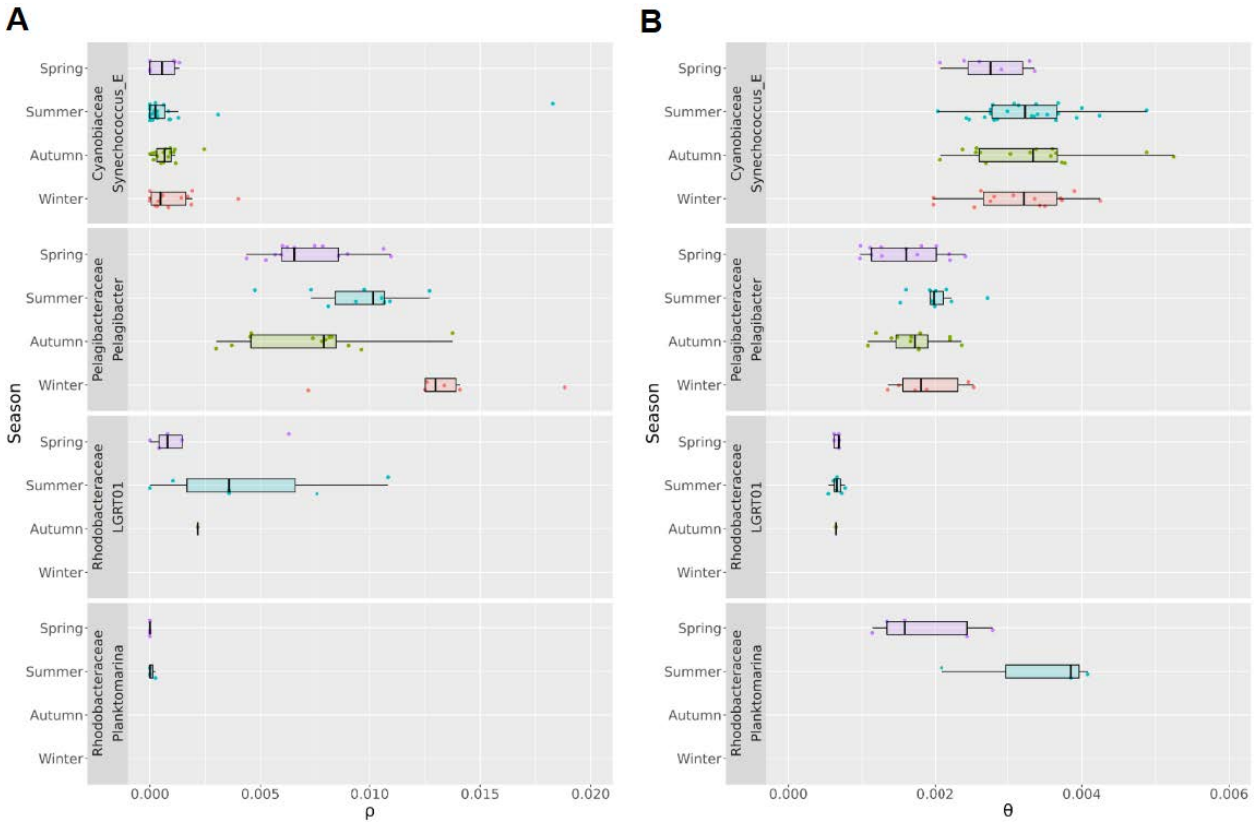
S3.4 Fig. Port Hacking ρ and θ values. (A) ρ values for Port Hacking Genera. **(B)** θ values for Port Hacking Genera.



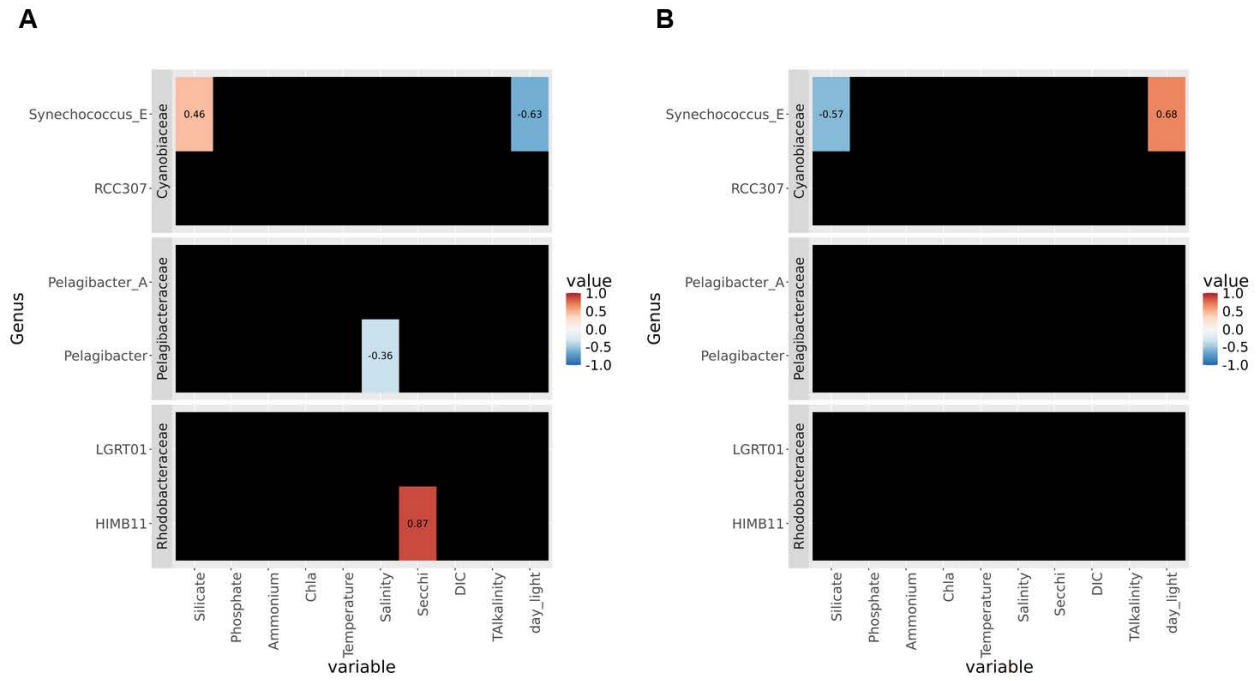
S3.5 Fig. Maria Island ρ and θ values. (A) ρ values for Maria Island Genera. **(B)** θ values for Maria Island Genera.



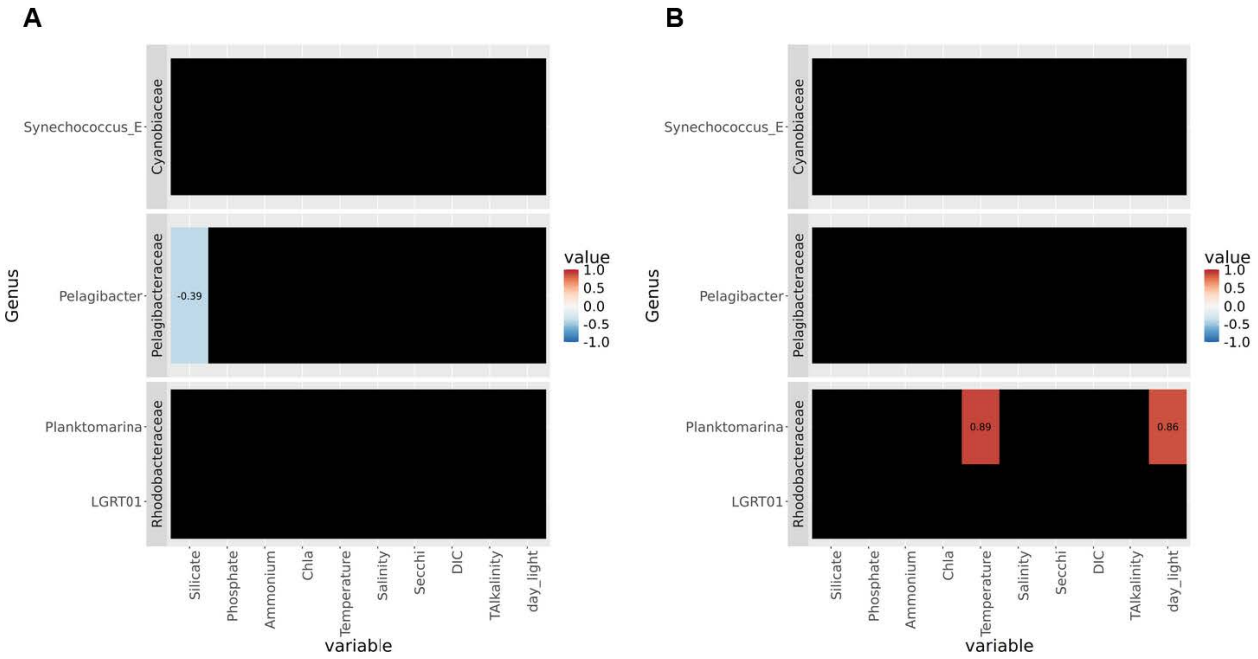
S3.6 Fig. Port Hacking ρ and θ values for genera by seasons. (**A**) ρ values for Port Hacking genera by seasons. (**B**) θ values for Port Hacking genera by seasons.



S3.7 Fig. Maria Island ρ and θ values for genera by seasons. (**A**) ρ values for Maria Island genera by seasons. (**B**) θ values for Maria Island genera by seasons.



S3.8 Fig. Port Hacking correlations against environmental variables, only significant correlations (p<0.05) are displayed. (**A**) ρ correlations. (**B**) θ correlations.



S3.9 Fig. Maria Island correlations against environmental variables, only significant correlations (p<0.05) are displayed. (**A**) ρ correlations. (**B**) θ correlations.

S3.1 Table. ρ / θ for common genera by location comparisons using Pairwise Wilcoxon test with Bonferroni correction.

Genus	p-value	Effect size
Synechococcus_E	2.4e-4	0.38
LGRT01	0.134	0.33
Pelagibacter	1.71e-4	0.39

S3.2 Table. Port Hacking and Maria Island, ρ/θ for genus by season comparisons using Kruskal-Wallis test. p values for significance are reported.

Port Hacking			Maria Island		
Genus	**p-value**	**Effect Size**	**Genus**	**p-value**	**Effect Size**
Synechococcus_E	0.004	0.33	Synechococcus_E	0.322	0.01
LGRT01	0.59	-0.09	LGRT01	0.421	-0.03
RCC307	0.539	-0.15	Pelagibacter	0.151	0.06
Pelagibacter	0.265	0.02	Planktomarina	0.197	0.11
Pelagibacter_A	0.298	0.1			
HIMB11	0.177	0.24			

S3.3 Table. Port Hacking ρ/θ mean values for genera by seasons		
Genus	**Season**	**Mean ρ/θ (std)**
HIMB11	Autumn	0.87 (0.782)
HIMB11	Spring	0.089 (0.087)
HIMB11	Summer	0.768 (0.61)
LGRT01	Spring	1.0 (1.476)
LGRT01	Summer	2.37 (2.83)
Pelagibacter	Autumn	4.31 (3.592)
Pelagibacter	Spring	2.2 (0.349)
Pelagibacter	Summer	3.713 (1.281)
Pelagibacter	Winter	3.378 (1.871)
Pelagibacter_A	Autumn	2.123 (1.397)
Pelagibacter_A	Spring	1.568 (NA)
Pelagibacter_A	Summer	1.05 (0.48)
Pelagibacter_A	Winter	0.886 (0.314)
RCC307	Autumn	0.44 (0.761)
RCC307	Spring	0.572 (NA)
RCC307	Summer	0.58 (0.654)
Synechococcus_E	Autumn	5.206 (3.298)
Synechococcus_E	Spring	0.759 (1.344)
Synechococcus_E	Summer	1.588 (2.889)
Synechococcus_E	Winter	4.027 (2.351)

S3.4 Table. Maria Island ρ/θ mean values for genera by seasons, standard deviation in brackets		
Genus	**Season**	**Mean ρ/θ (std)**
LGRT01	Autumn	3.349 (NA)
LGRT01	Spring	2.633 (3.738)
LGRT01	Summer	7.086 (6.644)
Pelagibacter	Autumn	4.778 (2.919)
Pelagibacter	Spring	4.845 (1.524)
Pelagibacter	Summer	4.809 (1.317)
Pelagibacter	Winter	7.065 (2.038)
Planktomarina	Spring	0.0 (0.0)
Planktomarina	Summer	0.022 (0.037)
Synechococcus_E	Autumn	0.234 (0.192)
Synechococcus_E	Spring	0.255 (0.289)
Synechococcus_E	Summer	0.378 (1.2)
Synechococcus_E	Winter	0.359 (0.404)

S3.5 Table. Port Hacking mean values for environmental parameters by season and genera, standard deviation in brackets

Season	Genus	Silicate_umolL	Phosphate_umolL	Ammonium_umolL	Chla_mgm3	Temperature_degC	Salinity_psu	Secchi_m	DIC_umolkg	TAlkalinity_umolkg	day_light_seconds
Autumn	HIMB11	1.07 (0.6)	0.16 (0.08)	0.43 (0.56)	0.47 (0.14)	21.09 (2.24)	35.41 (0.32)	15.75 (3.18)	2029.59 (33.8)	2323.64 (18)	40407.67 (3717.49)
Spring	HIMB11	0.3 (0.36)	0.13 (0.04)	0.4 (0.47)	0.77 (0.31)	19.49 (0.55)	35.52 (0.09)	12.5 (0.5)	2039.16 (18.84)	2331.4 (12.64)	50195.33 (245.26)
Summer	HIMB11	0.2 (0.35)	0.08 (0.03)	0.08 (0.07)	0.45 (0.27)	21.6 (0.59)	35.54 (0.09)	13 (3.46)	2036.33 (8.8)	2333.36 (3.61)	50948 (842.3)
Spring	LGRT01	0.48 (0.37)	0.15 (0.04)	0.34 (0.34)	0.88 (0.27)	19.02 (1.04)	35.55 (0.1)	13.3 (3.38)	2048.83 (15.76)	2332.16 (9.07)	47670.8 (3466.43)
Summer	LGRT01	0.22 (0.25)	0.1 (0.05)	0.16 (0.19)	0.34 (0.24)	21.38 (0.52)	35.55 (0.07)	14.8 (3.63)	2036.71 (7.22)	2334.37 (3.58)	51291.6 (759.8)
Autumn	Pelagibacter	0.53 (0.57)	0.13 (0.14)	0.59 (0.78)	0.33 (0.19)	22.48 (1.98)	35.54 (0.2)	18 (3.58)	2028.06 (24.67)	2333.16 (11.15)	39205.09 (2778.67)
Spring	Pelagibacter	0.6 (0)	0.13 (0)	0.08 (0)	0.14 (0)	21 (0)	35.66 (0)	19 (0)	2033.09 (0)	2338.4 (0)	46143 (0)
Summer	Pelagibacter	0.5 (0.34)	0.07 (0.04)	0.14 (0.1)	0.23 (0.11)	23.17 (1.72)	35.48 (0.14)	16.65 (3.06)	2026.12 (23.57)	2330.41 (8.49)	48550.94 (1880.71)
Winter	Pelagibacter	0.75 (0.28)	0.19 (0.09)	0.35 (0.4)	0.7 (0.12)	18.68 (0.77)	35.64 (0.07)	15.43 (0.92)	2051.09 (12.52)	2336.06 (3.14)	37941.47 (2020.77)
Autumn	Pelagibacter_A	0.9 (0.85)	0.14 (0.1)	0.44 (0.56)	0.5 (0.09)	21.49 (2.54)	35.39 (0.32)	15.75 (3.18)	2042.2 (36.47)	2325.07 (25.22)	38870.67 (4070.18)
Spring	Pelagibacter_A	0.6 (NA)	0.13 (NA)	0.08 (NA)	0.14 (NA)	21 (NA)	35.66 (NA)	19 (NA)	2033.09 (NA)	2338.4 (NA)	46143 (NA)
Summer	Pelagibacter_A	0.44 (0.34)	0.07 (0.04)	0.13 (0.09)	0.27 (0.14)	23.08 (1.73)	35.4 (0.21)	16 (3.46)	2015.83 (31.82)	2321.52 (20.69)	48155.8 (2137.62)
Winter	Pelagibacter_A	0.7 (0.57)	0.21 (0.15)	0.54 (0.64)	0.68 (0.25)	18.54 (1.45)	35.57 (0.01)	14.5 (0.71)	2049.11 (24.98)	2332.97 (1.11)	38786 (2134.05)
Autumn	RCC307	0.5 (0.5)	0.13 (0.1)	0.43 (0.57)	0.42 (0.13)	21.17 (2.33)	35.52 (0.17)	16.5 (2.6)	2036.18 (44.99)	2331.84 (15.64)	38132.33 (2793.92)
Spring	RCC307	0 (NA)	0.1	0 (NA)	0.45	19.02	35.63	12	2052.48	2340.33	50224

			(NA)		(NA)	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)
Summer	RCC307	0.38 (0.45)	0.08 (0.02)	0.12 (0.1)	0.45 (0.27)	21.9 (0.78)	35.54 (0.07)	13.5 (3)	2036.38 (7.18)	2334.28 (3.48)	50003.5 (2010.3)
Autumn	Synechococcus_E	0.71 (0.61)	0.16 (0.1)	0.5 (0.73)	0.45 (0.23)	21.9 (1.94)	35.45 (0.25)	17.75 (3.82)	2029.66 (22.78)	2331.03 (14.62)	39432.86 (2982.78)
Spring	Synechococcus_E	0.16 (0.26)	0.11 (0.01)	0.07 (0.13)	0.46 (0.23)	19.63 (0.9)	35.6 (0.08)	13.5 (3.08)	2043.27 (12.86)	2336.37 (7.82)	49350.4 (1797.29)
Summer	Synechococcus_E	0.26 (0.28)	0.09 (0.04)	0.11 (0.08)	0.35 (0.23)	21.97 (1.29)	35.51 (0.15)	15 (3.74)	2028.37 (20.01)	2329.66 (12.79)	50193.19 (1935.1)
Winter	Synechococcus_E	0.88 (0.34)	0.2 (0.07)	0.29 (0.25)	0.66 (0.19)	18.68 (0.79)	35.63 (0.07)	17.19 (2.05)	2053.28 (13.25)	2336.69 (4.01)	37198.88 (1588.32)

S3.6 Table. Maria Island mean values for environmental parameters by season and genera, standard deviation in brackets

Season	Genus	Silicate_umolL	Phosphate_umolL	Ammonium_umolL	Chla_mgm3	Temperature_degC	Salinity_psu	Secchi_m	DIC_umolkg	TAlkalinity_umolkg	day_light_seconds
Autumn	LGRT01	0 (NA)	0.07 (NA)	0 (NA)	0.38 (NA)	18.09 (NA)	35.54 (NA)	15 (NA)	2055.2 (NA)	2337.88 (NA)	47122 (NA)
Spring	LGRT01	0.66 (0.55)	0.23 (0.11)	0.23 (0.21)	1.26 (0.55)	14.17 (1.53)	35.19 (0.17)	12.6 (2.97)	2072.83 (13.22)	2324.11 (5.87)	49306 (3585.7)
Summer	LGRT01	0.25 (0.08)	0.11 (0.05)	0.13 (0.15)	0.45 (0.21)	16.25 (1.7)	35.29 (0.21)	17.25 (3.16)	2065.61 (8.24)	2322 (12.6)	53721 (2695.02)
Autumn	Pelagibacter	1.1 (0.67)	0.26 (0.17)	0.17 (0.19)	0.49 (0.1)	16.42 (1.14)	35.42 (0.06)	15.5 (2.68)	2074.85 (7.05)	2329.1 (4.62)	39308.25 (3626.02)
Spring	Pelagibacter	0.79 (0.43)	0.2 (0.03)	0.06 (0.13)	1.87 (1.14)	13.94 (1.16)	35.23 (0.1)	13.31 (0.85)	2082.33 (10.65)	2321.86 (4.4)	47547.62 (4694.74)
Summer	Pelagibacter	0.39 (0.17)	0.11 (0.03)	0.05 (0.04)	0.5 (0.14)	17.79 (1.14)	35.4 (0.1)	16.45 (1.89)	2062.44 (8.94)	2327.96 (4.65)	51677.5 (2361.4)
Winter	Pelagibacter	0.98 (0.38)	0.34 (0.05)	0.04 (0.05)	0.5 (0.16)	13.16 (0.7)	35.31 (0.07)	21.67 (2.58)	2102.44 (3.51)	2325.86 (3.98)	33995.5 (2692.44)
Spring	Planktomarina	0.66 (0.51)	0.32 (0.11)	0.26 (0.18)	1.23 (0.74)	12.68 (0.77)	35.22 (0.1)	13.8 (3.03)	2085.01 (12.51)	2321.9 (4.64)	45498.6 (2154.77)
Summer	Planktomarina	0.23 (0.06)	0.15 (0.04)	0.24 (0.14)	0.53 (0.25)	15.05 (2.17)	35.12 (0.03)	17.67 (4.04)	2065.43 (11.65)	2315.58 (13.72)	54658 (103.41)
Autumn	Synechococcus_E	0.59 (0.24)	0.15 (0.08)	0.04 (0.03)	0.32 (0.06)	17.32 (1.47)	35.5 (0.12)	19.06 (5.43)	2067.87 (9.99)	2333.93 (5.82)	41356 (3366.41)
Spring	Synechococcus_E	0.88 (0.18)	0.26 (0.11)	0.07 (0.05)	0.89 (0.43)	13.61 (1.9)	35.22 (0.07)	15.17 (2.04)	2090.26 (13.99)	2324 (3.68)	46744.83 (5135.84)
Summer	Synechococcus_E	0.26 (0.21)	0.09 (0.03)	0.03 (0.04)	0.44 (0.12)	17.75 (1.23)	34.77 (1.86)	16.02 (3.3)	2060.54 (9.64)	2328.01 (6.36)	52268.38 (2891.71)
Winter	Synechococcus_E	1.01 (0.33)	0.35 (0.04)	0.05 (0.04)	0.44 (0.19)	13.34 (0.82)	35.34 (0.1)	19.43 (2.14)	2101.38 (4.92)	2326.52 (4.59)	33947.29 (2349.03)

S3.7 Table. Port Hacking and Maria Island Species Lists

https://doi.org/10.5281/zenodo.14892007

(CSVS)

Data Availability:

The AusMicrobiome Microbial Ocean Atlas project metadata can be accessed via: https://github.com/AusMicrobiome/microbial_ocean_atlas/blob/57dfad6e30359a406ae86208fe554fac429269e6/data/oceanViz_AM_data.csv .

All metadata, relevant scripts and results generated, raw analysis files with metadata merged, are separated and accessible via: https://doi.org/10.5281/zenodo.15907061

Chapter 4:
Global patterns in recombination and mutation among marine bacterioplankton

Sidaswar Krishnan[a], Kittikun Songsomboon[a] , Aaron E. Darling[b,c], Matthew Z. DeMaere[b], Martin Ostrowski[a], Dominik Beck[d], Justin R. Seymour[a]

[a]Climate Change Cluster, University of Technology Sydney, Sydney, NSW, Australia

[b]Australian Institute for Microbiology & Infection, University of Technology Sydney, Sydney, NSW, Australia

[c]Lamnoo Inc., Iowa City, IA, USA

[d]Centre for Health Technologies and the School of Biomedical Engineering, University of Technology Sydney, Sydney, NSW, Australia

The order of authors reflects the level of their contributions, with the final author serving as the project lead.

4.1 Abstract

Microbial evolution is strongly influenced by mutation and recombination events, yet understanding the biotic and abiotic mechanisms governing these processes within natural ecosystems has been challenging. Recent bioinformatic advances that allow for mutation and recombination rates to be estimated from metagenomic data have opened the door to studying these phenomena as they occur in natural environments. In this study, we used the Rhometa pipeline to analyse metagenomic datasets derived from the Tara Oceans expedition with two primary objectives: (i) to determine the global relative mutation and recombination rates for important groups of marine bacteria, including *Pelagibacter*, *Prochlorococcus*, and *Synechococcus*, and (ii) to identify environmental factors that influence these processes on a global scale. We identified significant differences in the ratio of recombination to mutation events (ρ/θ) among the three genera, with *Pelagibacter* exhibiting the highest ρ/θ, followed by *Prochlorococcus* and *Synechococcus*. Notably, ρ/θ values for *Pelagibacter* were always greater than one, suggesting that recombination plays a more significant role than mutation in bacterial adaptation. Each genus exhibited unique ρ/θ patterns, reflecting adaptations to specific environmental conditions, with clear differences in the types of locations where the highest and lowest values occurred. *Pelagibacter* displayed highest ρ/θ values at a site in the South Atlantic Ocean off the coast of Argentina in a highly productive region, while the lowest ρ/θ values were observed in the South Pacific Ocean, in a region characterised by extreme oligotrophic conditions. This phenomenon of lower values in more oligotrophic waters was also observed for *Synechococcus* and *Prochlorococcus* to varying extents. Correlations between ρ/θ and environmental factors also identified different environmental determinants of recombination and mutation within each genus, whereby *Pelagibacter* ρ/θ was positively correlated with chlorophyll *a* and negatively correlated with temperature, while in contrast *Prochlorococcus* ρ/θ was positively correlated with temperature. These findings provide the first global-scale view into the dynamics of mutation and recombination within ecologically important groups of marine bacteria, advancing our understanding of how environmental factors shape microbial evolution in marine ecosystems.

4.2 Introduction

Prokaryotes play essential roles in shaping the productivity and biogeochemistry of natural ecosystems (Whitman, Coleman and Wiebe, 1998; Falkowski, Fenchel and Delong, 2008). Their evolution is primarily driven by two key processes: recombination and mutation (Hanage, 2016). Recombination allows bacteria to exchange genetic material both within and between populations through the mechanisms of transformation, transduction and conjugation (Didelot and Maiden, 2010). In contrast, mutations arise from changes in a cell's nucleotide sequence that escape cellular repair mechanisms (Hershberg, 2015). To understand prokaryotic evolution, it is crucial to examine the relative contributions of these two processes within natural microbiomes (Didelot and Maiden, 2010; Nei and Nozawa, 2011).

Mutation and recombination play important roles in helping prokaryotes adapt to environmental variability and stress (Arnold, Huang and Hanage, 2022). Environmental stressors, such as nutrient deprivation, influence mutation rates through stress-induced mutagenesis (SIM) (Ferenci, 2019). On the other hand, recombination rates in bacteria are affected by factors such as DNA donor-recipient similarity, proximity of cells and functional compatibility within the environment (Popa and Dagan, 2011). Given the importance of mutation and recombination in shaping microbial evolution and diversity, understanding their dynamics within natural ecosystems is essential for defining the factors structuring natural microbiomes.

Until recently, much of our understanding of bacteria have come from culture based studies (Handelsman, 2004). Such studies have revealed significant variability in recombination and mutation rates among bacterial species (Vos and Didelot, 2009), but their scope is often constrained by simplified laboratory settings, which generally can not fully reflect the complexities of natural environments and populations. However, recent advances in bioinformatic approaches now enable reliable calculation of population recombination rates (ρ) and mutation rates (θ) directly from metagenomic datasets derived from natural environments (Krishnan *et al.*, 2023). This has opened new avenues for studying these processes in natural microbiomes and exploring the environmental factors shaping their dynamics.

We recently applied the new bioinformatic pipeline called Rhometa (Krishnan *et al.*, 2023) to demonstrate that mutation and recombination rates vary significantly between different marine bacteria, and across seasons and geographic locations (Chapter 3;

Krishnan et al., 2025, *under review*). However, this analysis was constrained to two locations in southeastern Australia, limiting our ability to provide insights into broader biogeographical patterns. Expanding this perspective is essential because the global ocean is highly heterogeneous in abiotic and biotic characteristics across a continuum of spatial (e.g., mesoscale oceanographic features such as eddies, currents, and fronts) and temporal scales (e.g., diel to interannual cycles). Temporal patterns, including seasonal and annual reoccurrences in microbial community composition, are strongly linked to environmental factors such as temperature, salinity, and nutrients (Fuhrman *et al.*, 2006). Mesoscale phenomena, spanning tens to hundreds of kilometers, represent a particularly strong manifestation of physical-biological-biogeochemical interactions and play a fundamental role in ocean dynamics (McGillicuddy Jr, 2016). This environmental variability profoundly structures marine microbiomes and drives ecological and evolutionary processes within ocean systems. However, to what extent this variability in ocean physicochemical conditions impacts the important microbiological processes of mutation and recombination is unknown.

Global ocean-scale initiatives such as the Sorcerer II Global Ocean Survey (Rusch *et al.*, 2007) and Tara Oceans Expedition (Sunagawa *et al.*, 2020) have delivered unprecedented insights into the microbial ecology of marine ecosystems. The Sorcerer II survey pioneered the use of metagenomics to explore marine microbial communities, uncovering extensive taxonomic diversity (Rusch *et al.*, 2007). Tara Oceans built upon this foundation by greatly expanding the known genetic diversity of ocean microbes, cataloging over 40 million non-redundant protein-coding sequences (Sunagawa *et al.*, 2020). The Tara Oceans dataset offers an unparalleled resource for analysing microbial community structure, functional diversity, and environmental interactions on a global scale. Its comprehensive scope, encompassing samples from diverse oceanic regions, depths, and environmental conditions, allows for comparisons across spatial and ecological gradients. The dataset's standardised sampling and metadata integration provide a robust framework for linking microbial diversity to environmental factors such as temperature, salinity, and nutrients.

Applying Rhometa to the global Tara Ocean dataset allowed us to identify global patterns in recombination and mutation rates among key marine bacterial genera, including *Pelagibacter*, *Prochlorococcus*, and *Synechococcus*. Additionally, we explored relationships in this data with environmental factors. This work represents the first comprehensive global analysis of recombination and mutation rates in marine

bacteria and provides novel insights into how global-scale environmental variability shapes microbial evolution.

4.3 Methods

4.3.1 Sample selection

Metagenomic samples corresponding to prokaryotes from Tara Oceans project were chosen according to Alberti *et al.*, (2017) yielding a final set of 136 samples. The Kingfisher tool (https://doi.org/10.5281/zenodo.10525139) was then used to download the corresponding datasets.

To determine the taxonomic composition of the selected samples, we utilised Sandpiper (Woodcroft *et al.*, 2024), a platform built on the SingleM pipeline to analyse public metagenomic datasets from the NCBI SRA. Sandpiper identifies and quantifies operational taxonomic units (OTUs) in metagenomes. By using a combination of GTDB taxonomy tree (Release 220) (Parks *et al.*, 2022) with the Sandpiper platform, we identified and selected the genera of cyanobacteria and Pelagibacteria with the highest number of samples. This led to the selection of the Pelagibacter, Prochlorococcus_A and Synechococcus_C genera for downstream analysis. The suffixes (e.g., _A) in genus names, introduced by the GTDB classification system, indicate either polyphyletic genera or subdivisions for taxonomic rank normalisation. The exact species evaluated is provided as a table for each genera in the supplementary.

To obtain taxonomic information specific to our samples and genera of interest, we downloaded the complete Sandpiper database (v0.3.0) for GTDB release 220 (https://doi.org/10.5281/zenodo.11516218) and filtered the results for the 136 samples (246 runs). Building on this, we selected the most suitable reference genomes for the identified genera and species using the latest GTDB metadata file (release 220) (https://data.gtdb.ecogenomic.org/releases/release220/220.0/bac120_metadata_r220.tsv.gz). By filtering the metadata file to retain entries with “gtdb_representative” set to 't', we retrieved reference genome accessions from the “gtdb_genome_representative” column. These accessions were merged with the filtered Sandpiper dataset, limited to species-level entries, to link representative genomes with the taxonomy of the 136 samples.

Using the merged file, entries with a coverage (singlem read coverage) less than 10 were filtered out. For each genus, the sequencing run with the highest singlem read coverage corresponding to a species was retained per sample, resulting in unique samples per genus, each potentially mapping to a different species; sample points thus represent a cumulative distribution across species within each genus. The final list of representative genomes for each sample was then downloaded using the NCBI datasets program for read alignment.

4.3.2 Rhometa analysis

Rhometa version 1.1.0 was used to calculate recombination and mutation rates in the targeted bacteria, starting with the alignment of samples to the appropriate reference using Rhometa's `align_reads_input_list.nf` script under default settings.

Steps to normalise the bam files were then performed. First CoverM (v0.7.0) (Aroney *et al.*, 2024) was used to identify the mean read coverage of the aligned binary alignment map (BAM) field. Coverage was normalised to a mean of 20x using Samtools (v1.20) (Danecek *et al.*, 2021), with files below this threshold being discarded. We further ensured that only the bams with sufficient genome coverage (percentage of genome covered by reads) were used. To do this we used Samtools coverage on remaining bams and calculated the weighted coverage. Then the bams with a weighted coverage of greater than or equal to 80% were used for further analysis.

The population recombination rate (ρ) and population mutation rate (θ) were computed from the normalised BAM files. The ρ estimates were made using lookup tables of depth 3 to 300 with ρ values ranging from 0-25 going in steps of 0.25, 25 to 50 in steps of 0.5 and 50 to 100 in steps of 1, with default parameters used otherwise. A default constant effective population size of 1 was used for lookup table generation. The Rhometa pipeline used seed values 1, 2, and 3 for BAM file subsampling, and mean ρ values were calculated based on the ρ values obtained from each of the three seeds. The mean ρ per site was then obtained by dividing this mean ρ by the tract length and subsequently by 2, which accounts for the two break points characteristic of gene conversion-style recombination. This gives a final ρ (per site), for simplicity we refer to the value as ρ moving forward. The tract length was fixed to 1000 for all the ρ value calculations.

For population mutation rate (θ) per site, the mean depth measure was used. In Rhometa, θ is calculated per site by default, based on the mean aligned read depth across genomic positions. For simplicity, we will refer to this value as θ from this point forward. Both ρ and θ values implicitly incorporate the effective population size (N_e) as part of the calculation in Rhometa and the LDhat model on which it is based (McVean, Awadalla and Fearnhead, 2002; Krishnan *et al.*, 2023). Including N_e complicates interpretation and comparisons, as N_e varies between samples and is challenging to estimate. Therefore, in this work we have applied the the ratio ρ/θ, which effectively cancels out this factor and provides a robust measure of the recombination event probability relative to the mutation probability.

4.3.3 Statistical analysis

Environmental metadata from the Tara Oceans Project was downloaded from https://doi.pangaea.de/10.1594/PANGAEA.875582. Ammonium, iron, nitrite, nitrate, particulate organic carbon, particulate inorganic carbon, pH, carbon dioxide, total carbon, total alkalinity, and depth (top/min), phosphate and silicate, chlorophyll *a*, density (*sigma-theta*), temperature, salinity, oxygen and sunshine minutes were included. The metadata was merged with the rhometa results for the organisms for downstream analysis. Statistical analysis was then performed using R. The Shapiro-Wilk normality test was conducted which revealed that two out of the three targeted genera did not meet the assumption of normality. A Kruskal-Wallis test was then performed to compare differences between the three genera.

For Prochlorococcus_A, which was normally distributed, the one-way ANOVA (significance threshold: $p < 0.05$) was performed to test for significant differences between ocean regions. Then follow up pairwise comparisons between ocean regions were performed using Tukey's HSD post hoc test, with adjusted p-values reported for all pairs. For each pairwise comparison, Cohen's d was calculated as a measure of effect size. For Pelagibacter and Synechococcus_C, the Kruskal-Wallis test and Wilcoxon test with Bonferroni correction for multiple comparisons was likewise performed for ocean regions. Effect sizes (η^2 for Kruskal–Wallis and ANOVA) were calculated using R and are reported alongside p-values.

Lastly, to investigate relationships between ρ/θ values and environmental variables, Spearman correlation analyses were conducted using the spearmanr function from the

scipy.stats library in Python (v3.12) (Virtanen *et al.*, 2020). P-values for correlations were corrected for type 1 error with false discovery rate (FDR).

4.4 Results

4.4.1 The recombination to mutation ratio varied between marine bacterial genera

The recombination to mutation ratio, ρ/θ, differed significantly (Kruskal–Wallis test, p = 8.5e-07, η^2 = 0.2) between the three marine bacteria examined here. *Pelagibacter* displayed the highest mean ρ/θ of 5.6 (SD ± 2.8), followed by *Prochlorococcus_A* (4.1; SD ± 2.4) and *Synechococcus_C* (1.8; SD ±1.6). A ρ/θ value > 1 indicates that there are more recombination events apparent in sequence data relative to mutation events, which is evident in the mean values for each of the genera evaluated here. However, while ρ/θ values for *Pelagibacter* were always above 1 across the entire data set, in contrast 38 and 18% of samples ρ/θ was below 1 for *Synechococcus_C* and *Prochlorococcus_A,* respectively (Fig 4.1).

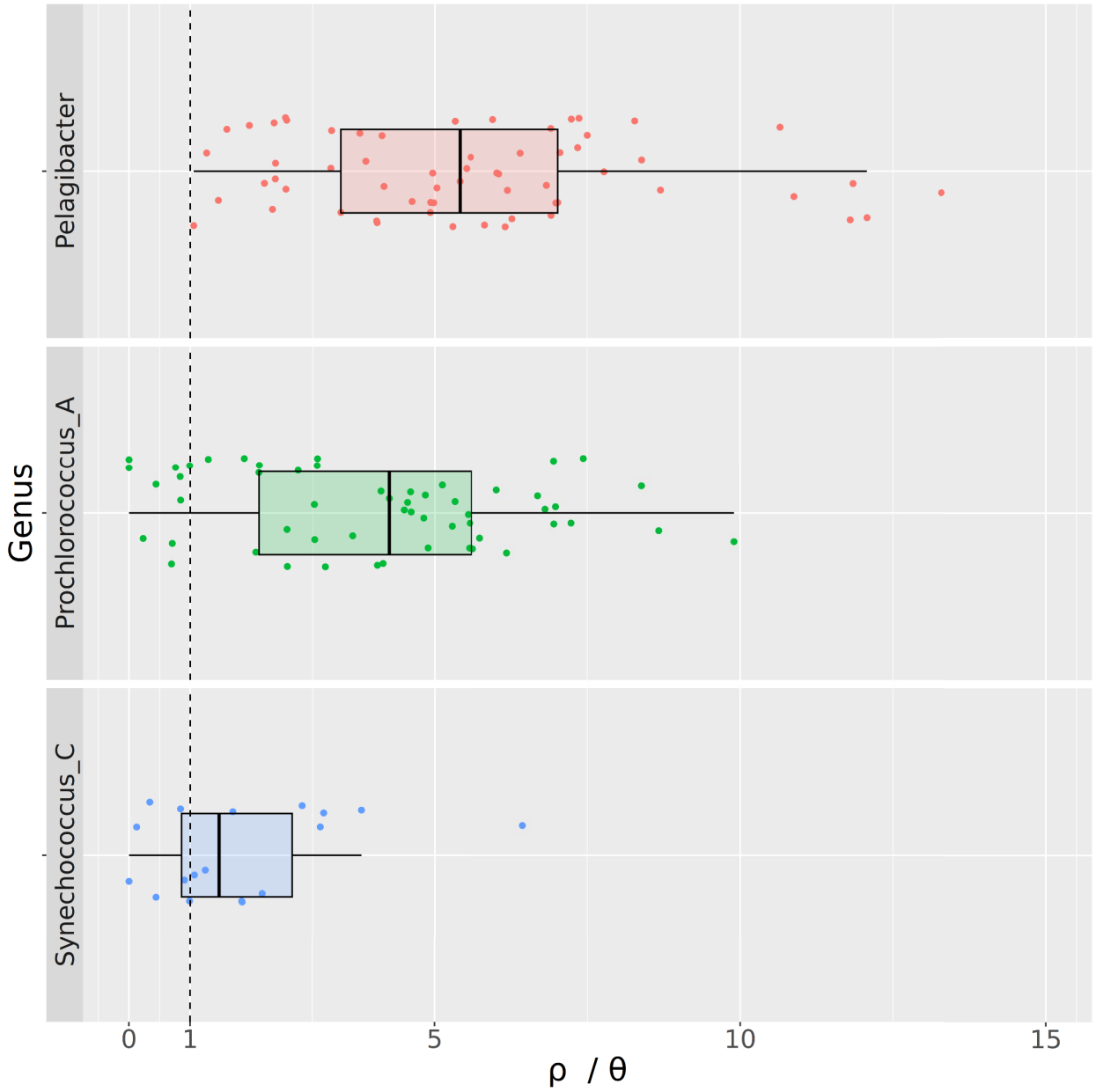


Fig 4.1. Recombination-to-mutation event rate ratios (ρ/θ) for three microbial genera for global samples. The vertical dashed line at ρ/θ=1 indicates equal rates of recombination and mutation. Values to the right of the line represent higher recombination rates relative to mutation. Each point represents a unique sample. Each point represents a unique sample per genus; points are cumulative across species within each genus.

4.4.2 The recombination to mutation ratio varied globally across genera

Pelagibacter exhibited consistently high ρ/θ values, indicative of relative importance of recombination events, though pronounced spatial variability was still apparent (Fig 4.2, 4.5). The highest Pelagibacter ρ/θ levels occurred in the South Atlantic Ocean, where mean values reached 8.3 (SD ± 3.7), while lowest levels were observed in the Mediterranean Sea, where mean ρ/θ levels were 2.1 (SD ± 0.8) (Fig 4.5, S4.1 Table). At a finer spatial resolution, the highest Pelagibacter ρ/θ value of 13.3 occurred in the South Atlantic Ocean, north of the Falkland Islands and off the coast of Argentina (station 082, -47.1863, -58.2902), while the lowest ρ/θ of 1.06 occurred in open ocean waters in the central South Pacific Ocean (station 098; -25.8051, -111.7202) (Fig 4.2, 4.5). The global distribution of Pelagibacter ρ/θ displayed two statistically significant correlations with measured environmental factors (Fig 4.6), with the strongest positive correlation with chlorophyll *a* (r = 0.38), and another correlation with ammonium (r = 0.34).

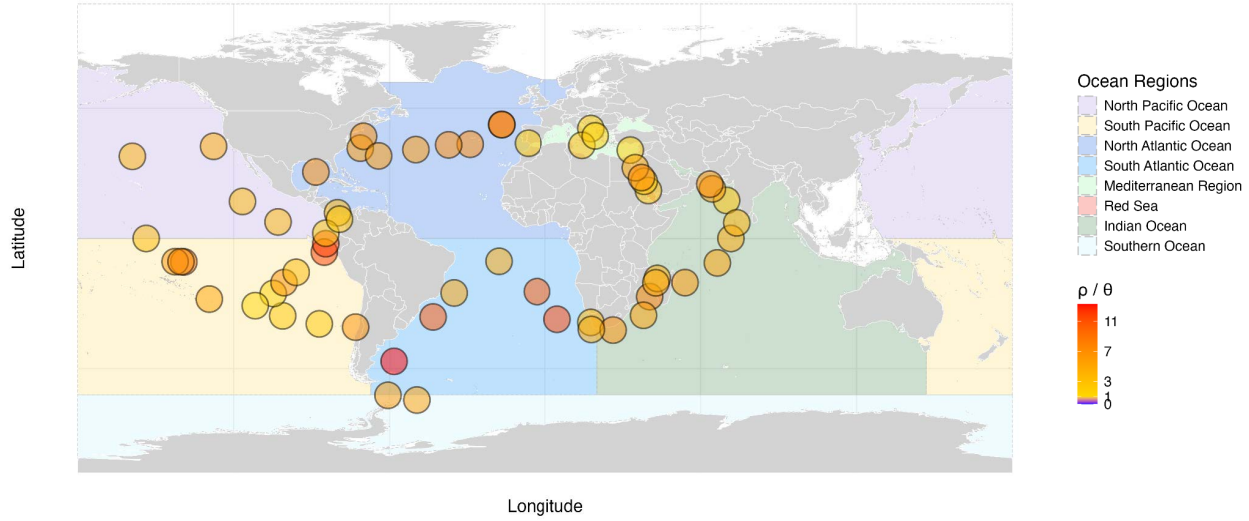


Fig 4.2. Recombination-to-mutation rate ratios (ρ/θ) for the Pelagibacter genus across global oceanic sites. Each circle represents a sampling location, coloured by ρ/θ: yellow shades indicate higher recombination relative to mutation (ρ/θ >= 1), and blue shades indicate higher mutation relative to recombination (ρ/θ < 1). Each point represents a unique sample per genus; points are cumulative across species within each genus.

Prochlorococcus_A exhibited significant regional differences (One-Way ANOVA, p = 0.002, η^2 = 0.34) in ρ/θ values across the global ocean, as well as substantially

different patterns to Pelagibacter (Fig 4.3, 4.5). Significant difference was observed between the South Pacific Ocean and Mediterranean Sea (p=0.003, Cohen's d = 1.63), South Pacific Ocean and North Atlantic Ocean (p = 0.02, Cohen's d = 1.14) and Red Sea and Mediterranean Sea (p = 0.04, Cohen's d = 4.1). The highest mean *Prochlorococcus_A* ρ/θ values of 5.5 (SD ± 3.3) occurred in the South Pacific Ocean, while the lowest values occurred in the Mediterranean sea 0.6 (SD ± 1), where the relative importance of mutation events exceeded recombination (Fig 4.5, S4.1 Table). At a finer resolution, the highest *Prochlorococcus_A* ρ/θ values reached 9.9 in the central South Pacific Ocean, far off the coast of Peru (station 100; -13.0023, -95.9759). The lowest ρ/θ was 0, which occurred at two stations (023; 42.2038, 17.715 and 025; 39.3888, 19.3905), both located in the Mediterranean Sea, where there was no evidence for recombination, leading to a 0/θ scenario. The lowest non-zero value for ρ/θ was 0.23 in the central North Atlantic Ocean (station 149; 34.1132, -49.9181). The global distribution of *Prochlorococcus_A* ρ/θ displayed 5 statistically significant correlations with environmental parameters (Fig 4.6), with the strongest positive correlation between ρ/θ and temperature (r = 0.46).

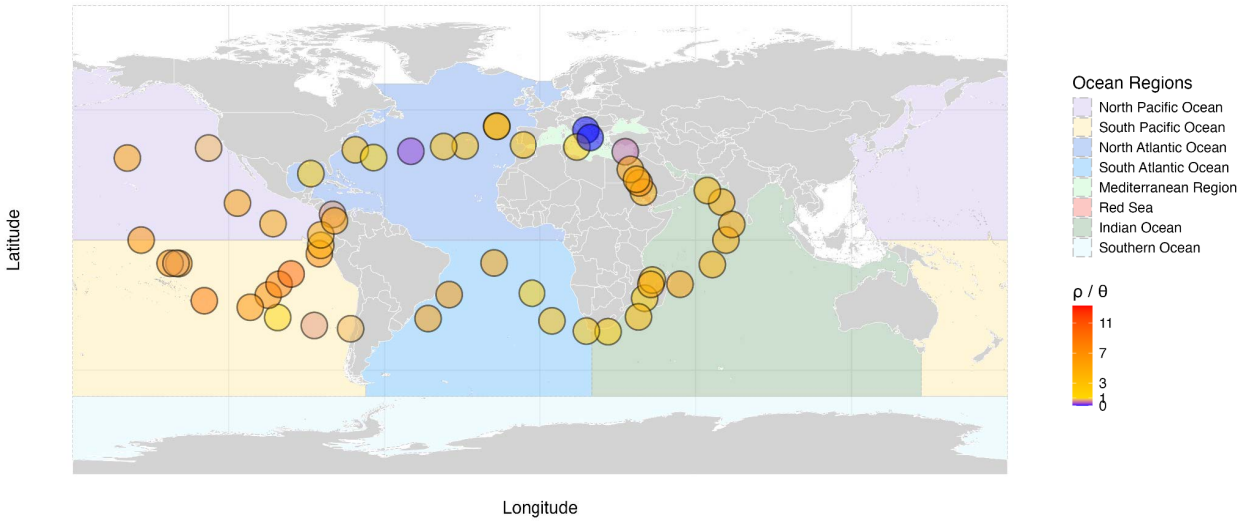


Fig 4.3. Recombination-to-mutation rate ratios (ρ/θ) for the Prochlorococcus_A genus across global oceanic sites. Each circle represents a sampling location, coloured by ρ/θ: yellow shades indicate higher recombination relative to mutation (ρ/θ >= 1), and blue shades indicate higher mutation relative to recombination (ρ/θ < 1). Each point represents a unique sample per genus; points are cumulative across species within each genus.

Among the genera examined in this study, *Synechococcus_C* displayed the lowest mean ρ/θ (1.8; SD ±1.6) value across the global ocean (Figs 4.4, 4.5). The highest mean ρ/θ for *Synechococcus_C* of 3.1 occurred in the South Atlantic Ocean, while the lowest mean ρ/θ value of 0.97 (SD ± 0.1) was recorded in the North Pacific Ocean (Fig 4.5, S4.1 Table). Among the individual sampling sites, the highest *Synechococcus_C* ρ/θ value of 6.4, was observed in the South Pacific Ocean offshore from Peru (station 102; -5.2529, -85.1545), while the lowest value of 0, indicative of no recombination, was also recorded in the South Pacific Ocean, but further south and near to the coastline of Chile (station 093; -34.0614, -73.1066). The lowest non-zero ρ/θ value, 0.12, was also found in the same region of the South Pacific Ocean, (station 094; (-32.7971, -87.0693). *Synechococcus_C* did not exhibit any correlations between ρ/θ and environmental factors (Fig 4.6), which may partly reflect the limited number of samples that this organism was detectable in.

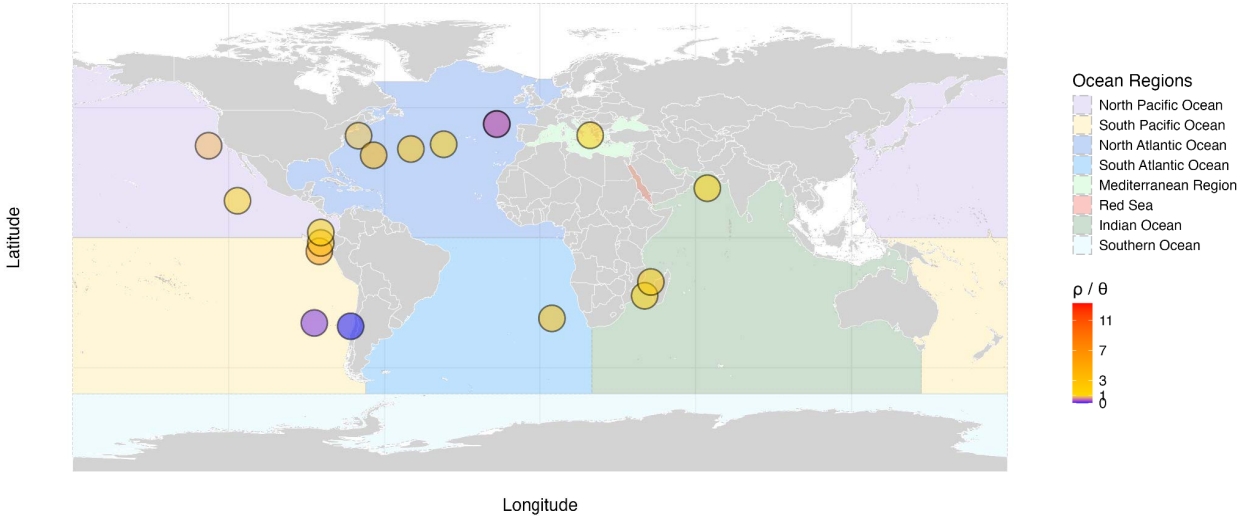


Fig 4.4. Recombination-to-mutation event rate ratios (ρ/θ) for the Synechococcus_C genus across global oceanic sites. Each circle represents a sampling location, coloured by ρ/θ: yellow shades indicate higher recombination relative to mutation (ρ/θ >= 1), and blue shades indicate higher mutation relative to recombination (ρ/θ < 1). Each point represents a unique sample per genus; points are cumulative across species within each genus.

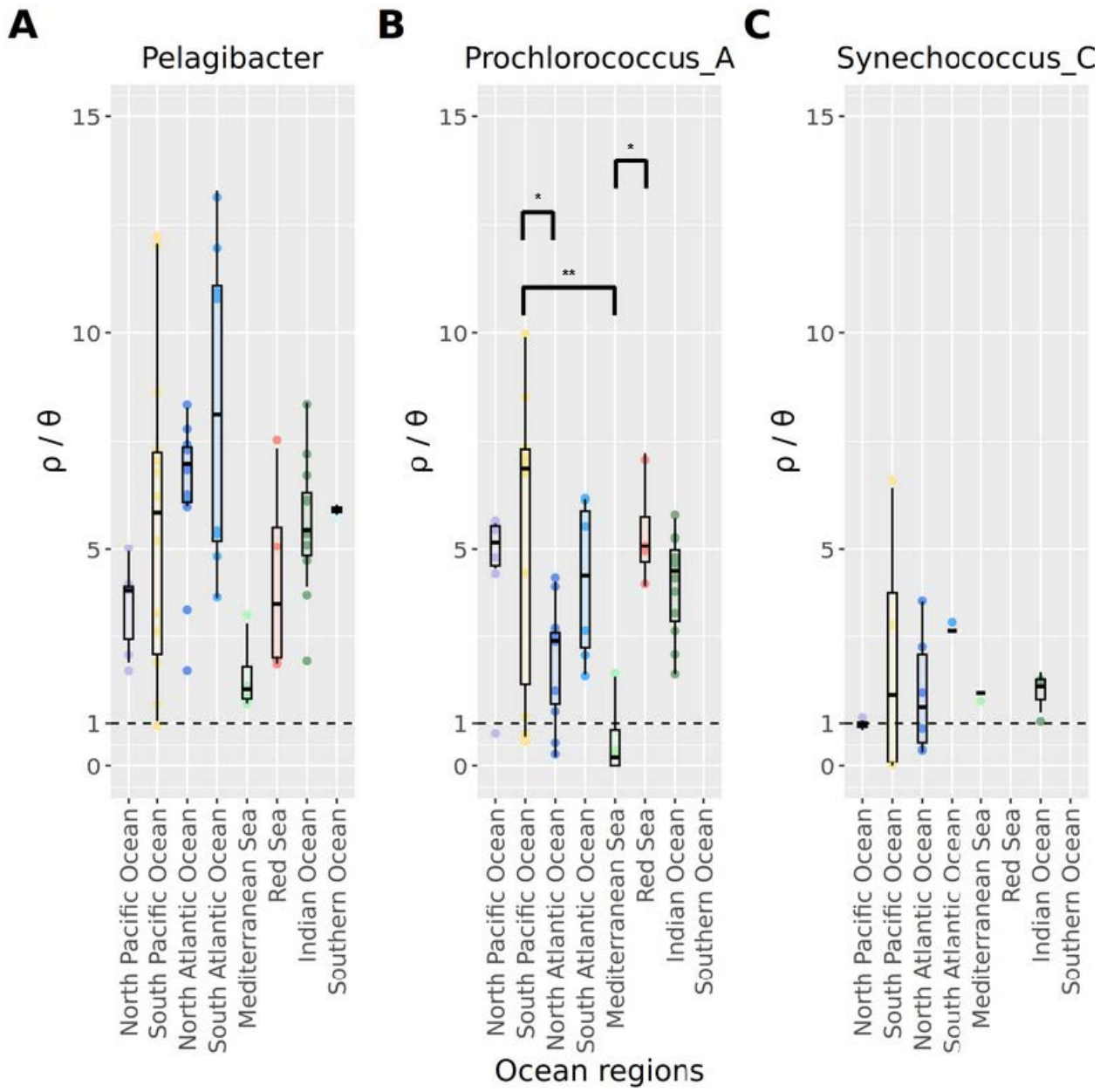


Fig 4.5. Recombination-to-mutation event rate ratios (ρ/θ) for three microbial genera across different ocean regions. Each boxplot represents the distribution of ρ/θ values for a genus in a given region. The horizontal dashed line at ρ/θ = 1 indicates equal rates of recombination and mutation. Values above this line represent a greater contribution of recombination relative to mutation. Brackets with p-values denote statistically significant differences between ocean regions. Each point represents a unique sample per genus; points are cumulative across species within each genus.

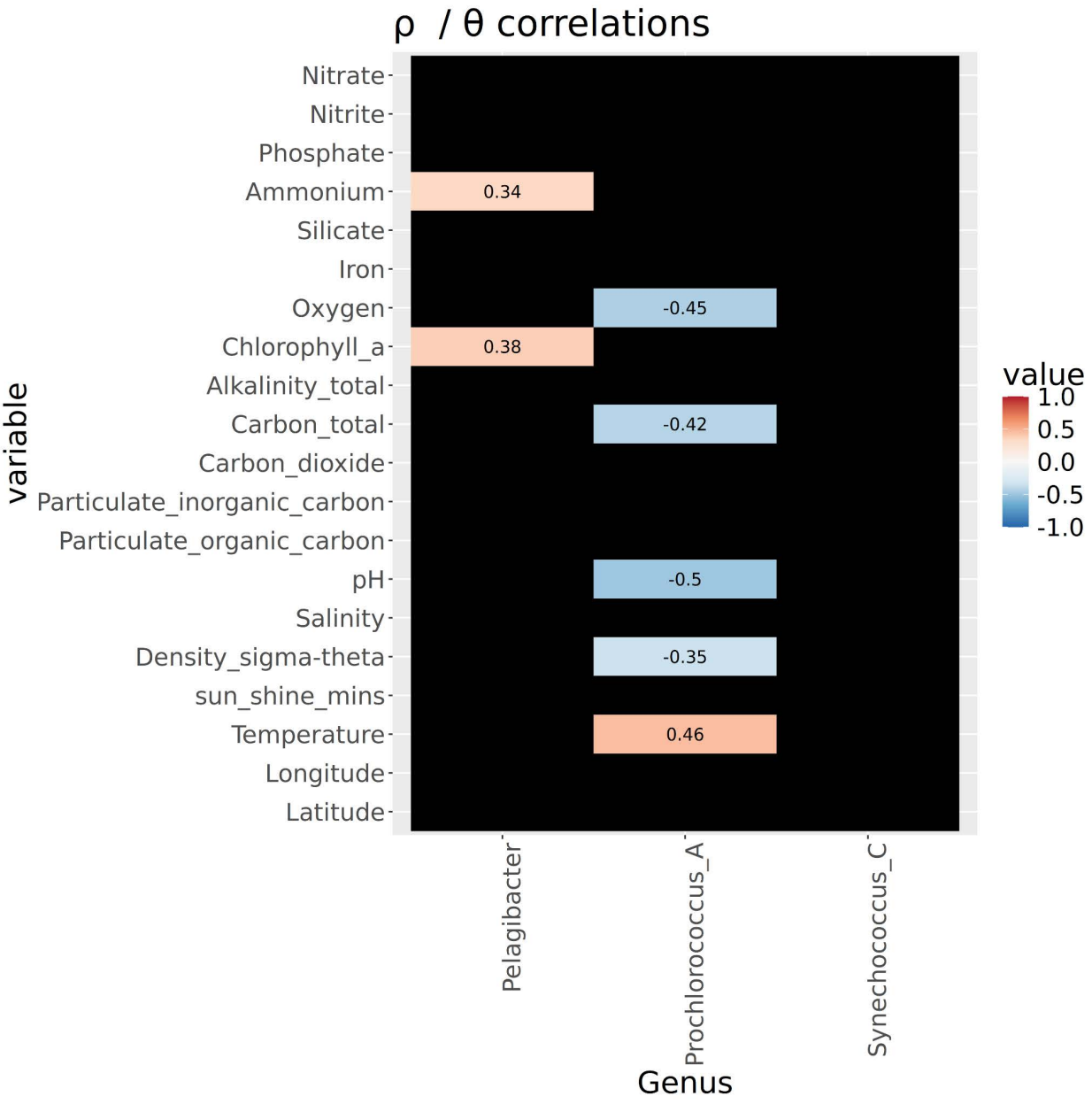


Fig 4.6. Correlations between recombination-to-mutation rate ratios (ρ/θ) and environmental variables for three genera. Each coloured tile represents a significant Spearman correlation ($p < 0.05$), with the correlation coefficient indicated numerically. Positive correlations are shown in shades of red, and negative correlations in blue, as indicated by the color scale.

4.5 Discussion

We aimed to identify global patterns in the relative rates of mutation and recombination in numerically abundant and ecologically important groups of marine bacteria and to determine the environmental factors influencing these processes. To address these aims, we analysed metagenomic data from the Tara Oceans Project (Sunagawa *et al.*, 2020), one of the most comprehensive global datasets on oceanic microbial diversity and function, using Rhometa (Krishnan *et al.*, 2023), a recently developed tool designed to accurately estimate population recombination (ρ) and mutation rates (θ) from metagenomic datasets. This analysis focused on three major marine bacterial genera: *Pelagibacter*, *Prochlorococcus*, and *Synechococcus*.

To investigate the relative rates of mutation and recombination, we utilised the compound ratio ρ/θ, which represents the likelihood of a recombination event relative to a mutation event. A ρ/θ value greater than 1 indicates that recombination events occur more frequently than mutation events, while a value less than 1 suggests that mutation events occur more frequently than recombination. While mutation serves as the source of evolutionary change, recombination facilitates the dissemination of these changes within a population, including the spread of adaptive mutations (Wiedenbeck and Cohan, 2011). Elevated recombination levels can therefore enhance the dissemination of beneficial mutations.

In this study, we focused on the cyanobacteria *Synechococcus* and *Prochlorococcus* and the heterotrophic bacteria *Pelagibacter*. *Synechococcus* and *Prochlorococcus* are the most abundant photoautotrophic picoplankton and are key contributors to oceanic primary production (Flombaum *et al.*, 2013), while *Pelagibacter* (SAR11) is the most numerically abundant organism on the planet, comprising 24–55% of all prokaryotic cells in the ocean (Morris *et al.*, 2002; Giovannoni, 2017). *Pelagibacter* is adapted to oligotrophic (nutrient-poor) waters, as is *Prochlorococcus* (Giovannoni *et al.*, 2005; Fuszard, Wright and Biggs, 2012). In contrast, *Synechococcus* demonstrates broader ecological adaptability capable of thriving across both nutrient-rich and nutrient-poor environments (Ahlgren and Rocap, 2012).

4.5.1 Regional trends in recombination to mutation ratios across marine bacterial genera

Across the global ocean, the three targeted marine bacterial genera displayed significant differences in their ρ/θ values. *Pelagibacter* exhibited the highest mean ρ/θ of 5.6, followed by *Prochlorococcus_A* (4.1), and *Synechococcus_C* (1.8). These findings are consistent with previous studies where recombination has been shown to play an influential role in *Pelagibacter* (Vos and Didelot, 2009; López-Pérez *et al.*, 2020). The high ρ/θ observed for *Pelagibacter* in this study is also aligned with our previous observations from two oceanographic time-series sites along the East Australian coast, where similarly elevated values were reported (Chapter 3; Krishnan et al., 2025, *under review*). Overall, these findings reinforce the notion that *Pelagibacter* undergoes frequent recombination relative to mutation. This may be due to the fact that the high abundance of SAR11 increases the chances of encountering other cells and recombining with DNA, contributing to its high recombination rates, a notion also postulated by previous studies (Vergin *et al.*, 2007; López-Pérez *et al.*, 2020).

In this study, we observed varying results with *Prochlorococcus_A* where 18% of *Prochlorococcus_A* samples had ρ/θ values below 1, these varying results are reflected in previous studies. A study by (López-Pérez *et al.*, 2020) observed that the *Prochlorococcus* species they evaluated experienced a stronger influence of recombination than mutation, relative to both *Pelagibacter* and *Synechococcus*. However, (González-Torres *et al.*, 2019) found the opposite pattern, reporting a much lower impact of recombination on the *Prochlorococcus* species they evaluated compared to those analysed by López-Pérez et al. (2020). A potential explanation for the discrepancy between these previous results could be spatial variability in the relative importance of recombination relative to mutation in *Prochlorococcus*, which is supported by the patterns observed in the present study.

High spatial variability in the ρ/θ levels of *Synechococcus_C* was also observed in this study, whereby 38% of *Synechococcus_C* samples had ρ/θ values below 1. These results are consistent with the patterns we observed in the cyanobacterial genera along the East Australian coast, (Chapter 3; Krishnan et al., 2025, in review) found that Synechococcus_E exhibited a high ρ/θ ratio in one of two oceanographic time-series stations examined (mean 2.7), indicating recombination dominance, while at the other site substantially lower ρ/θ values (mean 0.3) were measured, indicating that mutation was the primary evolutionary force.

Taken together, these findings suggest that, unlike *Pelagibacter*, the cyanobacterial genera *Prochlorococcus* and *Synechococcus* exhibit greater global variability in their dominant evolutionary forces, with the relative importance of recombination or mutation more dynamic across locations.

Interestingly, the highest regional mean ρ/θ values for both *Pelagibacter* and *Synechococcus_C* were observed in the South Atlantic Ocean, suggesting that recombination plays a more significant role than mutation in this region for both genera. The elevated ρ/θ values observed here may be influenced by the area's dynamic environmental conditions, where the influence of the Benguela Upwelling System, one of the most productive marine ecosystems (Martin *et al.*, 2024), and the Brazil–Malvinas Confluence Zone, where the convergence of the Brazil and Falkland currents supports high phytoplankton populations and intense water mixing (Brandini *et al.*, 2000; Willson and Rees, 2000) may have an impact. Factors like high phytoplankton density in the Brazil–Malvinas Confluence Zone and increased water mixing could potentially enhance cell-to-cell interactions, which are crucial for recombination (Popa and Dagan, 2011), these factors could therefore help promote genetic exchange through greater microbial community interactions.

Conversely, the lowest mean ρ/θ values for both *Pelagibacter* and *Prochlorococcus_A* occurred in the Mediterranean Sea. This trend is potentially consistent with the Mediterranean's semi-enclosed geography and highly oligotrophic nature, which restrict nutrient input (Siokou-Frangou *et al.*, 2010) and may lead to stress-induced mutations over recombination. A more detailed analysis of local-scale data from this region could provide further insights into the underlying factors driving these patterns.

4.5.2 Spatial and environmental impact on recombination to mutation ratio

Pelagibacter exhibited the highest ρ/θ values at a site in the South Atlantic Ocean off the coast of Argentina, near the Brazil–Falkland Confluence Zone. This region, where the Brazil and Falkland currents converge, is known for high phytoplankton productivity and intense water mixing (Brandini *et al.*, 2000; Willson and Rees, 2000). Given that *Pelagibacter* relies on phytoplankton nutrients (Giovannoni, 2017), the high primary productivity of this region may support increased *Pelagibacter* populations, leading to potentially more cell-to-cell interactions, an important factor for recombination (Popa and Dagan, 2011). In contrast, the lowest ρ/θ values for *Pelagibacter* were recorded in

a region of the South Pacific Ocean characterised by extreme oligotrophic conditions. These patterns are supported by the strong positive correlation between *Pelagibacter* ρ/θ values and chlorophyll a observed here.

Uniquely amongst the genera, significant regional differences in ρ/θ values were observed for *Prochlorococcus_A* across ocean basins. The greatest contrast was between the South Pacific Ocean, where the mean ρ/θ was 5.5, and the Mediterranean Sea, where it was considerably lower at just 0.63. The highest and lowest individual points were also found in these areas respectively. This indicates that the mode of evolution for *Prochlorococcus_A* is more clonal than recombinant in the Mediterranean. Possible factors for this could be the fact that the mediterranean is very oligotrophic in nature and semi enclosed (Siokou-Frangou *et al.*, 2010). While both locations are oligotrophic in nature, it could be that very low nutrient levels and resultant stress, are tipping the scale towards mutation rather than recombination as the primary driver of genetic variation in the mediterranean, further the semi enclosed nature might mean reduced gene flow.

Notably, Prochlorococcus_A displayed a positive correlation with temperature. Previous studies have shown a positive relationship between light, temperature, and Prochlorococcus abundance (Zinser *et al.*, 2007). The observed positive correlation between ρ/θ and temperature may reflect increased recombination driven by temperature-dependent growth, which could enhance opportunities for cell-to-cell interactions or cause bias towards adaptation rather than genomic stability.

The highest ρ/θ values in *Synechococcus_C*, were recorded in highly productive waters in the South Pacific Ocean near Peru, at a station influenced by the Humboldt Current (Gutiérrez, Akester and Naranjo, 2016). This region's high productivity is known to support diverse biological interactions and increased nutrient flow. In such an environment, increased recombination may enhance adaptability, allowing *Synechococcus* to persist amid changing ecological conditions. Adapting to new niches is an important function of recombination (Gogarten and Townsend, 2005) and such a dynamic and fluctuating environment could drive higher recombination. In a recent study (Chapter 3; Krishnan et al., 2025, in review), we found mean ρ/θ values of 2.7 (SD ± 3.1) and 0.3 (SD ± 0.786) for *Synechococcus_E* at two discrete oceanographic time-series sites in the Tasman Sea, which we argued could be driven by divergent oceanographic conditions. Interestingly, the lowest ρ/θ value for *Synechococcus_C*, were also recorded in the South Pacific Ocean, off the coast of chile, in a region that is less productive compared to waters near Peru, where we observed high ρ/θ,

(Gutiérrez, Akester and Naranjo, 2016). It is currently unclear why mutation rates might be so high in this area, but nutrient-related factors could be involved, as NO_3 has been shown to directly influence the biomass of phytoplankton in these waters (Spilling *et al.*, 2019).

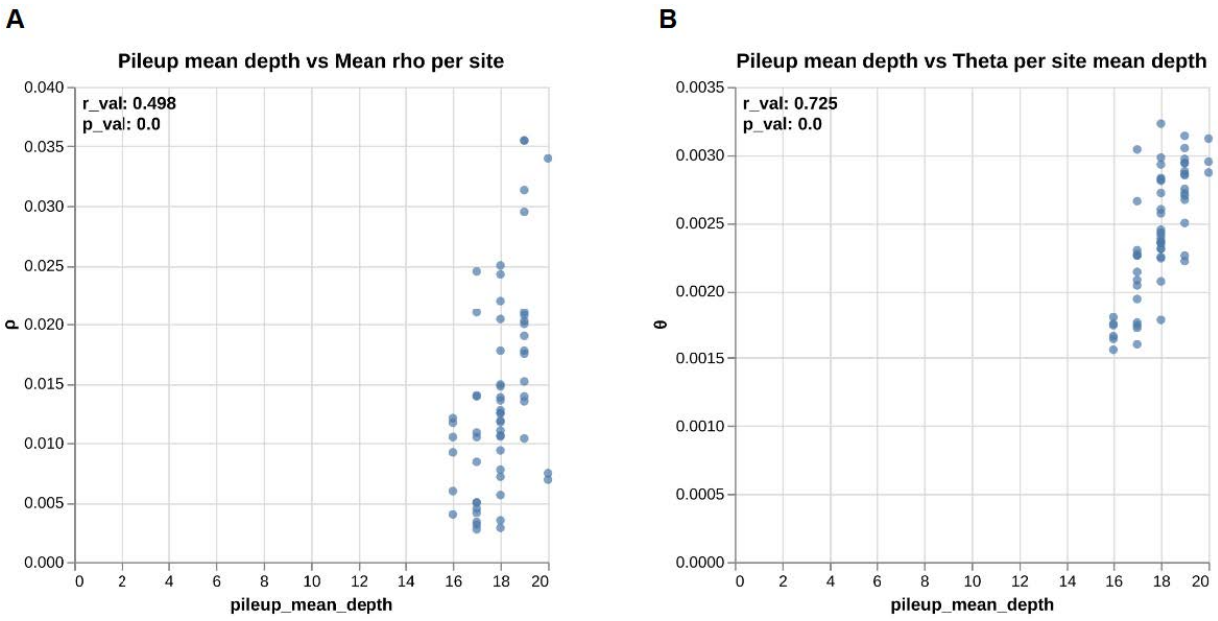
Such large-scale environmental studies investigating recombination and mutation rates in bacteria offer unprecedented glimpses into the global behaviour of these organisms and the environmental pressures that may shape these processes. Nonetheless, it is important to recognise that metagenomic data represent a snapshot of the community at the time of sampling and are subject to methodological biases and constraints inherent to metagenomic approaches.

4.6 Conclusion

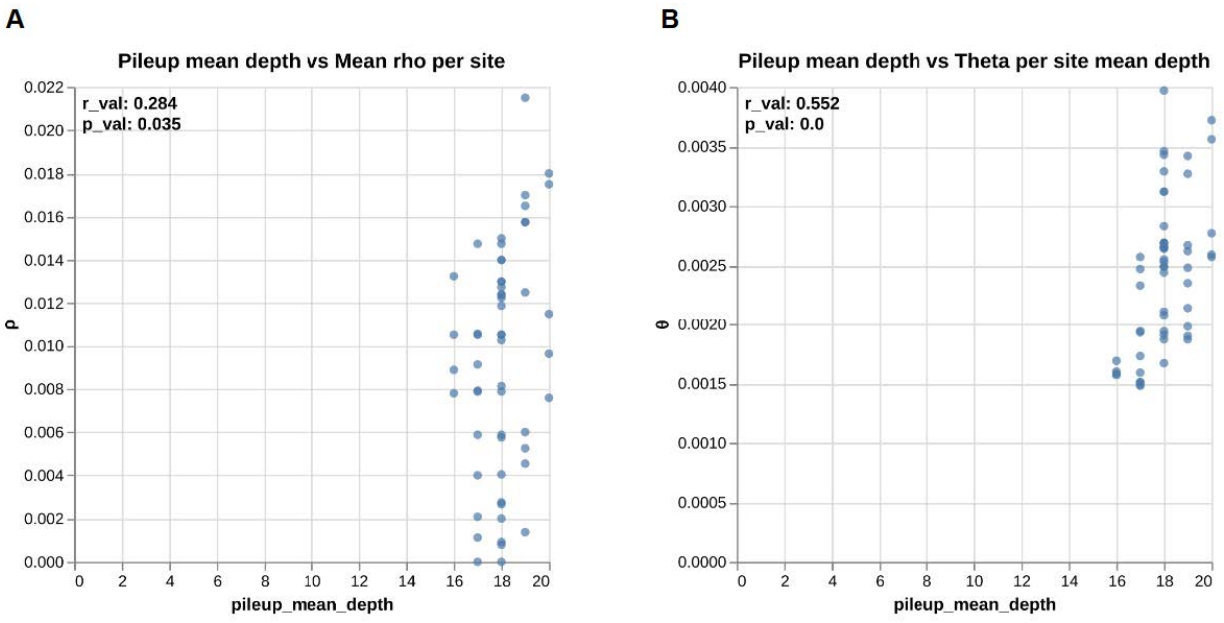
To investigate the relative importance of recombination and mutation in key marine bacterial genera and the influence of environmental factors on these processes, we utilised Rhometa, a program designed to measure mutation and recombination rates from metagenomic data. Leveraging global ocean-scale metagenomic datasets from the Tara Oceans project, we identified distinct evolutionary strategies reflected in the recombination-to-mutation (ρ/θ) ratios of the genera *Pelagibacter, Prochlorococcus,* and *Synechococcus*. Among the three genera, *Pelagibacter* exhibited the highest ρ/θ, followed by *Prochlorococcus* and *Synechococcus*, reinforcing previous findings on the elevated role of recombination in *Pelagibacter* evolution (Vergin *et al.*, 2007; Vos and Didelot, 2009). Interestingly, while only *Prochlorococcus* showed statistically significant regional variation in ρ/θ, all three genera displayed distinct hotspots and coldspots in ρ/θ values at finer spatial scales. Some patterns remained consistent across genera, such as the lowest mean ρ/θ values for both *Pelagibacter* and *Prochlorococcus* occurring in the Mediterranean Sea, potentially due to its oligotrophic conditions. Conversely, the South Atlantic Ocean, which is characterised by highly productive waters exhibited the highest mean ρ/θ values for *Pelagibacter* and *Synechococcus*. At a finer scale, *Pelagibacter* and *Synechococcus* exhibited higher ρ/θ values near nutrient-rich ocean currents, whereas *Prochlorococcus* showed elevated ρ/θ values in oligotrophic open-ocean environments. The relatively high ρ/θ values consistently observed in *Pelagibacter* suggest that recombination events occur more frequently than mutations, making recombination the dominant evolutionary strategy for adaptation. In contrast, *Prochlorococcus* and *Synechococcus* exhibit more spatially variable strategies, adapting through either recombination or mutation depending upon local

environmental conditions. By integrating global-scale metagenomic analyses with a novel bioinformatics approach, this study highlights the dynamic interplay between recombination and mutation in marine microbial evolution, shedding new light on the complex interactions between evolutionary processes and the ocean microbiome.

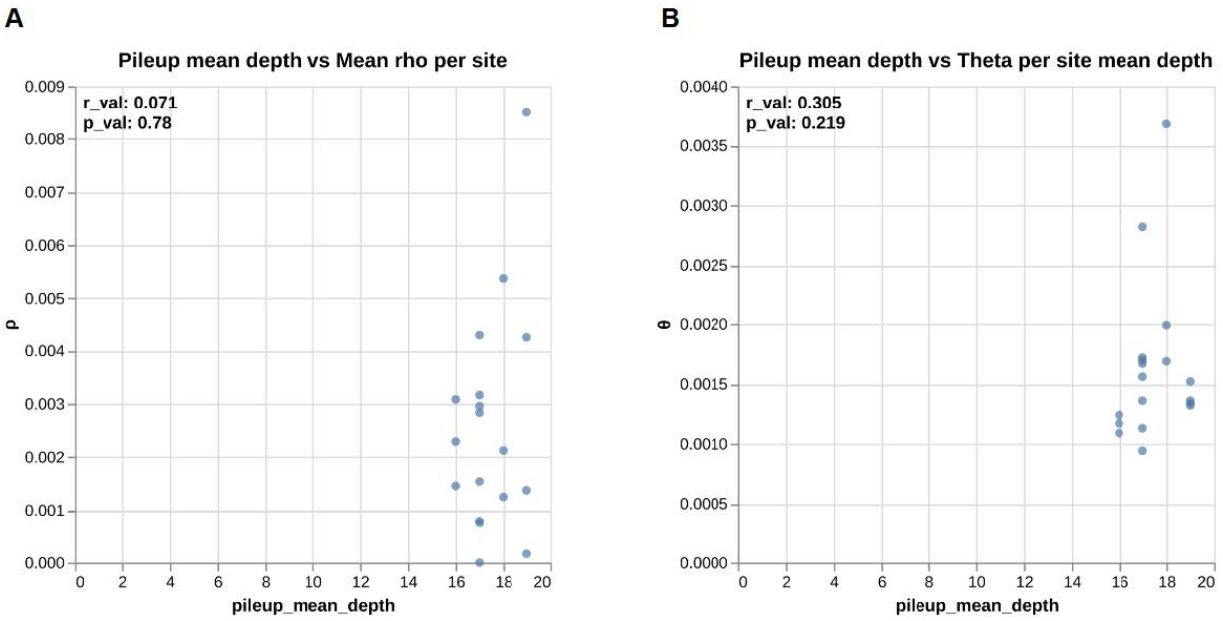
4.7 Supporting Information



S4.1 Fig. Genus Pelagibacter, pileup mean depth vs ρ and θ. After all filters applied. (**A**) Pileup mean depth vs ρ (**B**) Pileup mean depth vs θ.



S4.2 Fig. Genus Prochlorococcus_A, pileup mean depth vs ρ and θ. After all filters applied. (**A**) Pileup mean depth vs ρ (**B**) Pileup mean depth vs θ.



S4.3 Fig. Genus Synechococcus_C, pileup mean depth vs ρ and θ. After all filters applied. (**A**) Pileup mean depth vs ρ (**B**) Pileup mean depth vs θ .

S4.1 Table. Mean ρ/θ values for genus, region groups.

Genus	Ocean regions	Mean ρ/θ (std)
Pelagibacter	(IO) Indian Ocean	5.548 (1.5)
Pelagibacter	(MS) Mediterranean Sea	2.081 (0.839)
Pelagibacter	(NAO) North Atlantic Ocean	6.316 (1.766)
Pelagibacter	(NPO) North Pacific Ocean	3.698 (1.006)
Pelagibacter	(RS) Red Sea	4.298 (2.339)
Pelagibacter	(SAO) South Atlantic Ocean	8.295 (3.712)
Pelagibacter	(SO) Southern Ocean	5.935 (0.164)
Pelagibacter	(SPO) South Pacific Ocean	5.656 (3.606)
Prochlorococcus_A	(IO) Indian Ocean	4.15 (1.169)
Prochlorococcus_A	(MS) Mediterranean Sea	0.628 (0.985)
Prochlorococcus_A	(NAO) North Atlantic Ocean	2.438 (1.361)
Prochlorococcus_A	(NPO) North Pacific Ocean	4.49 (1.845)
Prochlorococcus_A	(RS) Red Sea	5.389 (1.316)
Prochlorococcus_A	(SAO) South Atlantic Ocean	4.28 (1.843)
Prochlorococcus_A	(SPO) South Pacific Ocean	5.477 (3.274)
Synechococcus_C	(IO) Indian Ocean	1.757 (0.471)
Synechococcus_C	(MS) Mediterranean Sea	1.697 (NA)
Synechococcus_C	(NAO) North Atlantic Ocean	1.692 (1.397)
Synechococcus_C	(NPO) North Pacific Ocean	0.966 (0.115)
Synechococcus_C	(SAO) South Atlantic Ocean	3.125 (NA)
Synechococcus_C	(SPO) South Pacific Ocean	2.436 (3.047)

S4.2 Table. Species lists for the genera targeted

https://doi.org/10.5281/zenodo.14892031

(CSVS)

Data Availability:

All metadata, relevant scripts and results generated, raw analysis files with metadata merged, are separated and accessible via: https://doi.org/10.5281/zenodo.16408197

Chapter 5:
General Discussion

5.1 Summary

The primary aim of this thesis was to explore the processes of evolution and adaptation in natural microbial environments by analysing recombination and mutation rates of bacteria. This objective was accomplished through the development of a new program designed to accurately calculate these rates from metagenomic datasets, which I called Rhometa. I then applied this program to measure recombination and mutation in key ocean bacterial lineages, utilising metagenomic data derived from both localised oceanographic time-series datasets and global-scale datasets. Relationships between these evolutionary parameters and environmental factors were also examined to assess the impact of environmental influences on microbial evolution.

Chapter 2 introduced Rhometa, which is a robust and computationally efficient method for estimating mutation and recombination rates, specifically designed for application to metagenomic datasets, to study prokaryotic populations. Validation on both simulated and experimental datasets confirmed its high accuracy and highlighted its unique ability to address the need for such a program.

Chapter 3 leveraged Rhometa and metagenomic datasets from the Australian east coast to reveal significant variations in recombination-to-mutation ratios among major marine bacterial genera, across space and time. Notably, the globally abundant marine bacterium *Pelagibacter* consistently exhibited high recombination-to-mutation ratios across various locations and time points, whereas the abundant marine cyanobacteria *Synechococcus* showed considerable variation. Furthermore, a suite of different environmental factors were shown to influence these rates, highlighting their critical role in shaping microbial evolution.

Chapter 4 utilised Rhometa and metagenomic data generated from the global Tara Oceans expedition to analyse recombination-to-mutation ratios in *Pelagibacter*, *Prochlorococcus*, and *Synechococcus*, uncovering distinct evolutionary strategies among these genera. *Pelagibacter* again displayed the highest ρ/θ ratio, followed by *Prochlorococcus* and *Synechococcus*. Relatively high ρ/θ values were consistently observed in *Pelagibacter* suggesting that recombination events occur more frequently

than mutations, making recombination the dominant evolutionary strategy for adaptation. In contrast, *Prochlorococcus* and *Synechococcus* exhibited more spatially variable strategies, altering between dominance of recombination or mutation depending upon local environmental conditions. Relationships between ρ/θ and environmental factors again revealed distinct environmental influences on recombination and mutation within each genus.

In this final chapter, the findings from the three research chapters are synthesised into cohesive overall insights and contributions. The results are presented in the context of key overarching questions investigated across the chapters. These are:

- Why do we need to measure mutation and recombination in natural microbiomes and how can this be done?
- Do recombination and mutation rates vary between different groups of bacteria?
- Do recombination and mutation rates vary spatially?
- To what extent do environmental and seasonal conditions shape recombination and mutation rates in natural microbiomes?

The chapter concludes with proposed future research directions and a final summary of the thesis's key achievements.

5.2 Synthesis of results

5.2.1 Why do we need to measure mutation and recombination in natural microbiomes and how can this be done?

Within prokaryotes, evolution is governed by the processes of mutation and recombination (Hanage, 2016). Mutations are changes in the nucleotide sequence within a cell’s DNA, resulting from molecular changes that are not repaired by the cellular repair systems (Hershberg, 2015), while recombination allows bacteria to exchange genetic material both within and between populations through the mechanisms of transformation, transduction and conjugation (Didelot and Maiden, 2010). The interplay between these two processes is fundamental to prokaryotic evolution (Hanage, 2016). Mutation and recombination levels can both be influenced by environmental conditions such as nutrient availability and temperature (Aminov, 2011; Le *et al.*, 2020; Horton and Taylor, 2023). Factors such as donor-recipient similarity

barriers, where gene transfer is more common among closely related species, ecological barriers, which depend on the physical proximity of donor and recipient species and functional barriers, which affect the retention and integration of exchanged DNA (Popa and Dagan, 2011), also impact recombination. Importantly this interplay of environmental and ecological factors plays a dynamic role, where mutation serves as the source of evolutionary change, whereby recombination facilitates the dissemination of these changes within a population, including the spread of adaptive mutations (Wiedenbeck and Cohan, 2011). Understanding the dynamics of these evolutionary processes helps us to understand how prokaryotes adapt and speciate (Lawrence, 1999).

Much of our understanding of microbiology stems from culture-based studies (Handelsman, 2004). Such studies, however, are limited by the fact that they do not reflect the natural environment, indeed most microorganisms cannot even be successfully cultured in the laboratory (Handelsman, 2004; Singh *et al.*, 2009). This limitation hinders our ability to fully understand the intricate dynamics of microbial evolution and the influence of environmental factors on genetic processes such as recombination and mutation. Without direct observation of microbes in their natural habitats, the extent to which these mechanisms contribute to microbial diversity and adaptation remains largely speculative.

Metagenomics involves analysing sequence data derived directly from environmental samples (Handelsman *et al.*, 1998; Wooley, Godzik and Friedberg, 2010; Thomas, Gilbert and Meyer, 2012), thereby addressing the culture based limitations. Metagenomic shotgun sequencing yields fragments of DNA sequences, referred to as reads, which taken together represent a random sampling of genome fragments from all the microbes in an environmental sample, which can then be processed and studied in detail (Sharpton, 2014). One of the primary aims of this thesis was to meet the need for a program that could accurately and efficiently estimate mutation and recombination rates from metagenomic read datasets. Thereby, quantifying these processes as they occur in nature.

When considering existing programs, LDhat (McVean, Awadalla and Fearnhead, 2002; Auton and McVean, 2007), holds historical significance as a key methodological advancement in the estimation of mutation and recombination rates. It pioneered the use of the composite likelihood estimator, which has since been recognized as one of the most accurate approaches in this domain. Over time, LDhat has been widely

adopted in population genetics, cementing its role as a foundational tool in the study of genetic variation and recombination. It is capable of analysing both crossing-over and gene conversion types of recombination, but is limited to sequence based datasets. Crossing over occurs during meiosis, while bacterial recombination has greater similarities to gene conversion (Hanage, 2016). The key difference between these being that gene conversion involves non-reciprocal exchange of genetic material, while crossing over involves the reciprocal exchange of it (Guirouilh-Barbat *et al.*, 2014). Later programs, LDhelmet (Chan, Jenkins and Song, 2012) and LDhot (Auton, Myers and McVean, 2014) build on LDhat, but are designed for the crossing over form of recombination. Cutting edge programs like pyrho (Spence and Song, 2019), uses a more sophisticated implementation of the composite likelihood estimator that is much faster, however it is also designed for crossing over type of recombination.

Given LDhat's established status as one of the most powerful methods available, we chose to build on its foundation and enable it for metagenomics datasets under the gene-conversion model of recombination. Our attempt, however, was not the first to take this approach. PIIM (Johnson and Slatkin, 2006, 2009) was a pioneering tool for estimating recombination rates in metagenomic data with a composite likelihood estimator as implemented in LDhat, however it included computationally intensive methods to account for high sequencing error rates and scarce data, and it relies on the obsolete ACE file format, lacking support for modern formats like BAM. Therefore, this pipeline is largely outdated and unsuitable for modern metagenomic data.

mcorr (Lin and Kussell, 2019) is a recent program that estimates the recombination-to-mutation rate and recombinational divergence in metagenomic reads. Which is different from the population scaled recombination rate (ρ) and population scaled mutation rate (θ) estimated by LDhat. It however outputs r/m which represents the probability that a nucleotide was substituted due to either recombination or mutation (Didelot and Wilson, 2015), which can be derived from ρ and θ and compared against. In addition it is also limited to coding regions, meaning it does not consider all the regions of the genome, as performed with established composite likelihood estimators, even if it is only for variant sites. It has since been demonstrated that mcorr produces unreliable results on some datasets, with unrealistic estimates of the ratio of recombination to mutation sometimes resulting (Krishnan *et al.*, 2023).

My new program called Rhometa (Krishnan *et al.*, 2023) (Chapter 2) successfully adapts the composite likelihood estimator as implemented in LDHat for metagenomic

datasets. It is designed to work with aligned short reads. I extensively tested this new program using both simulated based datasets generated using the program msprime (Baumdicker *et al.*, 2022) and available data from an *S. pneumoniae* transformation experiment (Croucher *et al.*, 2012). In both cases, reliable results were obtained, especially in comparison to the mcorr platform that I benchmarked it against. This makes Rhometa currently the best method available for studying mutation and recombination in metagenomic datasets.

Rhometa introduces a novel method for handling varying depths, a challenge with handling aligned reads, which makes it different to the approach taken by PIIM, by using precomputed likelihood for each specific depth and weighting values by depth prior to final estimation of the most likely recombination rate (Fig 5.1). It is also able to downsample and handle reads of arbitrary size. Moving forward it is likely that further improvements could be built on this strong foundation.

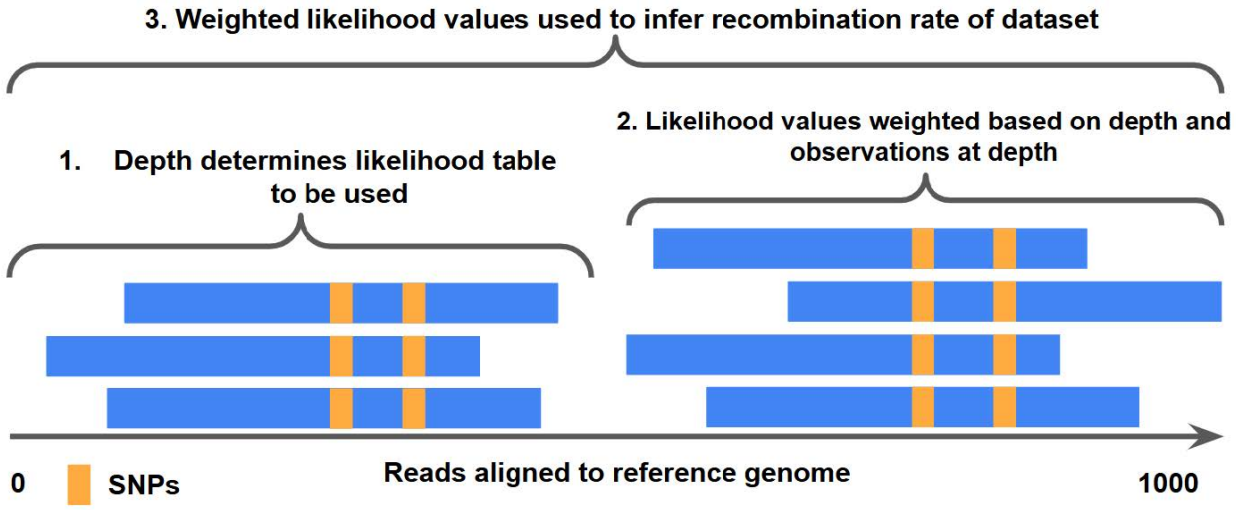


Fig 5.1. Simplified visualisation of method used to handle varying depths and determine recombination rates in Rhometa.

Rhometa's ability to work at the level of reads is crucial. While it is possible that metagenome-assembled genomes (MAG) could have been directly used with LDhat, it is known that they can contain chimeric sequences (Orakov *et al.*, 2021). Reads represent genomic data of microbes as it originally existed, providing a direct window to study the recombination and mutation events, and reducing the potential complications associated with chimeric data

Notably, programs like LDhot extend the composite likelihood estimator to enable fine-scale recombination rate estimation and the detection of recombination hotspots, however, they are specifically designed for eukaryotes. Recombination hotspots are also present in bacteria (Yahara *et al.*, 2016). Moving forward tools like Rhometa could potentially be adapted to address the need for recombination hotspot detection in bacterial metagenomic datasets.

In summary, this thesis has delivered a new tool for accurate quantification of the rates of mutation and recombination from metagenomic datasets. Reflecting the rates at which these processes occur within natural microbiomes.

5.2.2 Do recombination and mutation rates vary between different groups of bacteria?

Following confirmation of the utility of Rhometa, I next applied it to study the dynamics of recombination and mutation among bacteria in the marine microbiome. The ocean covers 70% of Earth's surface and delivers half of global primary production (Azam and Malfatti, 2007). Ocean microbes, comprise 70% of total marine biomass, and control key biogeochemical cycles, which has significance for carbon sequestration, and climate regulation (Logares, 2024). However, very little is known about the extent and influence of mutation and recombination on their evolution.

We focused on studying mutation and recombination rates in the major marine bacterial lineages Cyanobiaceae, Pelagibacteraceae, and Rhodobacteraceae. Pelagibacter (SAR11) is the most numerically abundant organism on the planet, accounting for up to 25% of all prokaryotic cells in the ocean and plays an important role in marine carbon cycling (Morris *et al.*, 2002; Giovannoni, 2017). The cyanobaceria, *Synechococcus* and *Prochlorococcus* are the most abundant photoautotrophic picoplankton and are key contributors to oceanic primary production (Flombaum *et al.*, 2013). Rhodobacteraceae, particularly the copiotrophic genus *Roseobacter*, play a pivotal role in carbon and sulphur cycling. In coastal waters, *Roseobacter* can account for up to 20% of bacterial cells (Luo and Moran, 2014).

We wished to investigate the dynamics of mutation and recombination among these important groups of marine bacteria. While mutation introduces new genetic variation, recombination plays a crucial role in spreading beneficial mutations within bacterial

populations (Wiedenbeck and Cohan, 2011). I employed the ratio ρ/θ, which represents the likelihood of a recombination event relative to a mutation event in order to study their relative contribution.

Chapter 3 analysed metagenomic time series datasets from Port Hacking and Maria Island, focusing on several bacterial genera from the lineages of interest at these sites. Port Hacking is a subtropical site that is influenced by the East Australian Current (EAC) and its eddy field, while Maria Island, located in Tasmania, is a temperate site affected seasonally by the southernmost reach of the EAC (Brown et al., 2018). In contrast, Chapter 4 examined three specific genera, namely *Pelagibacter, Prochlorococcus* and *Synechococcus,* across numerous global locations using the Tara Oceans datasets (Sunagawa *et al.*, 2020).

By focusing on the ρ/θ values for these marine genera, we found significant differences between the genera. This confirms some previous observations that the importance of recombination and mutation can vary substantially between different species of bacteria (Vos and Didelot, 2009; González-Torres *et al.*, 2019). In both Chapters 3 and 4, I found that *Pelagibacter* was consistently highly recombinant relative to mutation, displaying the highest mean ρ/θ. These findings are consistent with previous studies where recombination has been shown to play an influential role in *Pelagibacter* (Vos and Didelot, 2009; López-Pérez *et al.*, 2020). Pelagibacter are the most abundant ocean bacteria, it is likely that the large role of recombination plays a key factor in enabling their adaptation to many niches across oceans.

In contrast to Pelagibacter, the relative levels of recombination to mutation in the marine cyanobacteria *Synechococcus* and *Prochlorococcus* revealed spatial and temporal heterogeneity, with ρ/θ in these genera sometimes switching between positive and negative values, which is indicative of change in the relative dominance of recombination to mutation. Temporal trends (in Synechococcus) and spatial patterns (in both Synechococcus and Prochlorococcus) in ρ/θ, as well as significant correlations between ρ/θ and a suite of environmental parameters highlights the dynamic nature of recombination and mutation in these important marine phototrophs and confirms the critical role of environmental factors in shaping evolutionary processes in these bacteria.

Taken together my findings demonstrate that different groups of marine bacteria display markedly different characteristics in their recombination and mutation dynamics, with

Pelagibacter showing consistently high levels of recombination, while the relative importance of recombination vs mutation in *Synechococcus* and *Prochlorococcus* shifts with space and time, with apparent links to different environmental factors. These findings provide the first evidence of the distinct role of mutation and recombination across space and time, and highlight the fact that evolutionary forces shaping important groups of marine bacteria co-inhabiting the same environments are not uniform.

5.2.3 Do recombination and mutation rates vary spatially?

The global ocean varies greatly in its physical and biological properties, shaping marine microbiomes and influencing ecological processes. However, the impact of this environmental variability on key evolutionary processes like mutation and recombination remains unknown. Given that the relative influence of evolutionary factors can vary spatially, as seen with *Synechococcus* and *Prochlorococcus*, further investigation was performed to understand how environmental dynamics shape microbial evolution across different ocean regions.

In Chapter 3, ρ/θ were compared between two time-series sites in the Tasman Sea. Despite the environmental variance across the sites *Pelagibacter* was highly recombinant between them, while *Synechococcus* on average saw recombination as the dominant force in Port Hacking and mutation as the dominant force in Maria Island.

In chapter 4, a similar trend was observed, whereby consistently high ρ/θ was observed for *Pelagibacter* with recombination always being the dominant force. *Synechococcus* and *Prochlorococcus* often switched between experiencing mutation and recombination as the dominant force across various sites globally. Some spatial trends in *Pelagibacter* ρ/θ levels were still evident, with highest ρ/θ values recorded at a site in the South Atlantic Ocean, near the Brazil–Falkland Confluence Zone. This region, where the Brazil and Falkland currents converge, is known for high phytoplankton productivity and intense water mixing (Brandini *et al.*, 2000; Willson and Rees, 2000) which could potentially enhance cell-to-cell interactions, which are crucial for recombination (Popa and Dagan, 2011). The lowest ρ/θ values for *Pelagibacter* were recorded in a region of the South Pacific Ocean characterised by extreme oligotrophic conditions.

Interestingly the pattern of the highest value ρ/θ being observed in productive waters and lowest in oligotrophic waters was common amongst the genera studied. *Prochlorococcus* was the only genus that displayed statistically significant regional differences in ρ/θ values, with the greatest contrast being observed between the South

Pacific Ocean and the Mediterranean Sea. The highest and lowest ρ/θ values were also found in these areas respectively, unlike the South Pacific Ocean, the Mediterranean is characterised by its very oligotrophic and semi enclosed nature (Siokou-Frangou *et al.*, 2010). The highest ρ/θ values in *Synechococcus*, were recorded in highly productive waters in the South Pacific Ocean near Peru in an area influenced by the Humboldt Current (Gutiérrez, Akester and Naranjo, 2016). This region's high productivity is known to support diverse biological interactions and increased nutrient flow. While in contrast the lowest ρ/θ value for *Synechococcus*, were also recorded in the South Pacific Ocean, off the coast of chile, in a region that is characterised by lower productivity (Gutiérrez, Akester and Naranjo, 2016)

5.2.4 To what extent do environmental and seasonal conditions shape recombination and mutation rates in natural microbiomes?

Chapter 3 used metagenomic data derived from two oceanographic time-series, whereby samples were collected approximately monthly over a period of seven years. This enabled me to examine the seasonal dynamics of recombination and mutation among the targeted marine bacteria and to reveal any potential links to environmental factors.

Both Port Hacking and Maria Island have distinct temporal dynamics, The Port Hacking microbiome composition is seasonally impacted by the EAC (Messer *et al.*, 2020) and recurring seasonality in microbial community characteristics has been demonstrated at Maria Island (O'Brien *et al.*, 2022).

Overall distinct correlation profiles between environmental factors and ρ/θ were observed for *Pelagibacter* and *Synechococcus* this indicates that environmental differences between these two locations also drive shifts in the relative importance of recombination and mutation within these bacteria. This was likewise true for the genera studied in chapter 4 and their environmental correlation profiles.

Focusing on temporal variations in environmental factors, despite the significant seasonal fluctuations in environmental conditions at both sites, *Pelagibacter* remained highly recombinant on average, and no statistically different changes in ρ/θ were observed over time, this indicates that the highly recombinant nature of this bacteria is retained even in the face of high environmental heterogeneity, reinforcing the importance of this strategy.

Synechococcus showed significant seasonal variation at Port Hacking, particularly between summer and autumn. The highest ρ/θ values occurred in autumn and winter, aligning with increased nutrients at this time. This along with the EAC's strong microbial influence in autumn (Messer *et al.*, 2020), suggests that nutrient availability and microbial dynamics might contribute to higher recombination. The increase in nutrients overlaps with the well-mixed period (May–September) (Brown *et al.*, 2018), potentially enhancing cell growth and consequently chances for interaction, a key factor for recombination (Popa and Dagan, 2011). Additionally, the dynamic environment itself may promote recombination as an adaptive response. At Maria Island, mutation dominance was observed year-round, though the cause remains unclear. This pattern may indicate a more stressful environment.

In summary, these findings indicate that certain genera, such as *Pelagibacter*, maintain consistent ρ/θ patterns across seasons, with consistently high mean values. In contrast, other genera, like *Synechococcus*, show significant fluctuations in mean ρ/θ values across different seasons.

5.3 Future Directions

Since Rhometa was published, there have been several updated releases of the program, which optimised performance, introduced new pipelines to facilitate ease of use, such as a high level pipeline which can align reads and perform analysis all at once, and many other usability improvements. I have also additionally optimised filters to improve reliability of estimates.

In this thesis, the ratio ρ/θ was calculated when studying recombination and mutation in marine bacteria. However r/m which represents the probability that a nucleotide was substituted due to either recombination or mutation, is often also considered (McVean, Awadalla and Fearnhead, 2002; Vos and Didelot, 2009; Didelot and Wilson, 2015). At present, Rhometa, which is based on composite likelihood estimators as implemented in LDhat (McVean, Awadalla and Fearnhead, 2002) and PIIM (Johnson and Slatkin, 2006, 2009), remains limited to estimating ρ and θ.

A major constraint in deriving r/m lies in the challenge of accurately determining the nucleotide substitution probability as a result of recombination. Logically, if a variant is at abundance x then the substitution probability can be thought of as the probability of choosing a pair of cells for recombination where one has the variant and the other

doesn't. Using the allele frequency value which can be computed by modern variant callers it is possible to estimate the substitution probability. I have written the script to estimate the substitution probability and therefore am able to compute the r/m ratio in Rhometa. I intend to include this feature in an upcoming version 2 release of the program.

A recent evaluation of the program ANGSD (Korneliussen, Albrechtsen and Nielsen, 2014), revealed that its Watterson θ (population mutation rate) estimation method might be viable for incorporation in Rhometa. The ANGSD implementation is a more sophisticated approach than the one implemented in Rhometa and is also able to work on read datasets in the form of BAM files. The tool uses an empirical Bayes method to estimate the Watterson θ (Watterson, 1975). Initial tests show promising results, but further evaluation is needed before it can be incorporated into Rhometa. Incorporating this new tool would further improve Rhometa's ρ/θ and r/m estimates. Making it one of the most powerful programs of its kind.

While this is sufficient for a version 2 release of Rhometa, there is potential for further improvement of the program. The penalised likelihood approach taken in pyrho (Spence and Song, 2019) for eukaryotic genomes, offers improved performance and accuracy for recombination rate estimation when compared to the composite likelihood approach as implemented in LDhat. The improved method could potentially be repurposed to work with prokaryotes and used to improve Rhometa's recombination rate estimation performance.

Rhometa has been applied by external researchers, such as in (Breusing *et al.*, 2023), to investigate patterns of homologous recombination and mutation in bacterial symbionts associated with deep-sea hydrothermal vent mussels and snails. The study estimated recombination-to-mutation rate ratios across the symbiont species, revealing elevated levels of recombination relative to mutation. The upcoming version 2 of Rhometa is expected to facilitate further studies by providing more accurate estimates and a streamlined workflow for calculating r/m.

While the Tara Oceans project (Sunagawa *et al.*, 2020) is unparalleled in its global scale, other efforts such as the Sorcerer II Global Ocean Survey (Rusch *et al.*, 2007) can complement large-scale studies like the one presented in Chapter 4, helping to further validate its findings. For more localised and temporally resolved analyses, such as the East Australian coast study described in Chapter 3, additional time series

datasets are available. For instance, one dataset focuses on Japanese coastal waters and spans the period between March 2012 and May 2016 (Yoshitake *et al.*, 2021). Such time-series datasets can be used to investigate recombination-to-mutation rate ratios over time and assess their relationships with environmental variables in different locations. Taken together, such studies can help validate and extend the findings of Chapters 3 and 4.

As demonstrated in Chapter 4, through the global ocean sample analysis, the Sandpiper platform presents a promising opportunity to systematically analyse the full range of metagenomic datasets available on the NCBI SRA for many bacteria of interest. Nevertheless, challenges related to metadata quality and data accessibility must be taken into account. Despite these limitations, Rhometa can be leveraged for such large-scale investigations and offers an exciting avenue for exploring the dynamics of recombination and mutation across a wide array of bacterial taxa and samples.

While metagenomics combined with environmental metadata provides a valuable avenue for investigating the dynamics of recombination and mutation in bacteria, its results are constrained by methodological biases inherent to the approach and reflect the bacterial community as it exists in a dynamically changing environment at the time of sampling. Even so, the environmental associations observed here offer insights into how the bacteria studied behave in situ and present promising avenues for deeper investigation. The correlation patterns identified may serve as a basis for designing targeted experimental inquiries to test and refine our understanding of the mechanisms driving these relationships.

Finally, beyond its application in understanding evolutionary processes in ocean bacteria, Rhometa can be extended to a broader range of scenarios, including investigations into the relationship between recombination and antimicrobial resistance (AMR). Recombination is known to play a pivotal role in the dissemination of AMR (Bhat *et al.*, 2023). Estimating recombination rates using Rhometa in AMR-related studies could provide valuable insights into the pace and dynamics of AMR gene spread within various environments.

5.4 Conclusion

This thesis involved the development of a novel computational tool, Rhometa, designed to estimate recombination and mutation rates from metagenomic datasets by extending the composite likelihood method, which is among the most accurate methods available to accommodate modern aligned read data across varying sequencing depths.

By applying Rhometa to global and time-series ocean metagenomic datasets, we assessed the evolutionary dynamics of key marine bacteria, namely Pelagibacter, Prochlorococcus, and Synechococcus, and observed that the balance between recombination and mutation varied not only across taxa, but also over temporal and spatial scales.

Some bacterial groups exhibited consistently high recombination-to-mutation ratios, while others shifted between the two forces. These patterns were shaped by environmental factors such as temperature and nutrient availability, and a broader global trend emerged wherein productive waters were associated with higher recombination influence, while oligotrophic regions were dominated by mutation. Together, these findings advance our ability to quantify microbial evolutionary forces in situ and provide insights into the ecological and evolutionary processes shaping natural microbial communities.

Bibliography

Ahlgren, N.A. and Rocap, G. (2012) ‘Diversity and Distribution of Marine Synechococcus: Multiple Gene Phylogenies for Consensus Classification and Development of qPCR Assays for Sensitive Measurement of Clades in the Ocean’, *Frontiers in Microbiology*, 3. Available at: https://doi.org/10.3389/fmicb.2012.00213.

Akoijam, N., Kalita, D. and Joshi, S.R. (2022) ‘Bacteria and Their Industrial Importance’, in P. Verma (ed.) *Industrial Microbiology and Biotechnology*. Singapore: Springer, pp. 63–79. Available at: https://doi.org/10.1007/978-981-16-5214-1_2.

Alberti, A. *et al.* (2017) ‘Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition’, *Scientific Data*, 4(1), p. 170093. Available at: https://doi.org/10.1038/sdata.2017.93.

Aminov, R.I. (2011) ‘Horizontal Gene Exchange in Environmental Microbiota’, *Frontiers in Microbiology*, 2. Available at: https://doi.org/10.3389/fmicb.2011.00158.

Andersson, D.I. and Hughes, D. (1996) ‘Muller’s ratchet decreases fitness of a DNA-based microbe.’, *Proceedings of the National Academy of Sciences of the United States of America*, 93(2), pp. 906–907.

Arenas, M. *et al.* (2015) ‘CodABC: A Computational Framework to Coestimate Recombination, Substitution, and Molecular Adaptation Rates by Approximate Bayesian Computation’, *Molecular Biology and Evolution*, 32(4), pp. 1109–1112. Available at: https://doi.org/10.1093/molbev/msu411.

Armour, C.R. *et al.* (2019) ‘A Metagenomic Meta-analysis Reveals Functional Signatures of Health and Disease in the Human Gut Microbiome’, *mSystems*, 4(4), p. 10.1128/msystems.00332-18. Available at: https://doi.org/10.1128/msystems.00332-18.

Arnold, B.J., Huang, I.-T. and Hanage, W.P. (2022) ‘Horizontal gene transfer and adaptive evolution in bacteria’, *Nature Reviews Microbiology*, 20(4), pp. 206–218. Available at: https://doi.org/10.1038/s41579-021-00650-4.

Aroney, S.T.N. *et al.* (2024) ‘CoverM: Read coverage calculator for metagenomics’. Zenodo. Available at: https://doi.org/10.5281/zenodo.10531254.

Auton, A. and McVean, G. (2007) ‘Recombination rate estimation in the presence of hotspots’, *Genome Research*, 17(8), pp. 1219–1227. Available at: https://doi.org/10.1101/gr.6386707.

Auton, A., Myers, S. and McVean, G. (2014) ‘Identifying recombination hotspots using population genetic data’, *arXiv:1403.4264 [q-bio]* [Preprint]. Available at: http://arxiv.org/abs/1403.4264 (Accessed: 27 September 2021).

Azam, F. and Malfatti, F. (2007) ‘Microbial structuring of marine ecosystems’, *Nature Reviews Microbiology*, 5(10), pp. 782–791. Available at: https://doi.org/10.1038/nrmicro1747.

Bar-On, Y.M., Phillips, R. and Milo, R. (2018) ‘The biomass distribution on Earth’, *Proc Natl Acad Sci U S A*. 2018/05/23 edn, 115(25), pp. 6506–6511. Available at: https://doi.org/10.1073/pnas.1711842115.

Baumdicker, F. *et al.* (2022) ‘Efficient ancestry and mutation simulation with msprime 1.0’, *Genetics*, 220(3), p. iyab229. Available at: https://doi.org/10.1093/genetics/iyab229.

Bhat, B.A. *et al.* (2023) ‘Integrons in the development of antimicrobial resistance: critical review and perspectives’, *Frontiers in Microbiology*, 14. Available at: https://doi.org/10.3389/fmicb.2023.1231938.

Bobay, L.-M. (2020) ‘CoreSimul: a forward-in-time simulator of genome evolution for prokaryotes modeling homologous recombination’, *BMC Bioinformatics*, 21(1), p. 264. Available at: https://doi.org/10.1186/s12859-020-03619-x.

Bowers, R.M. *et al.* (2017) ‘Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea’, *Nature Biotechnology*, 35(8), pp. 725–731. Available at: https://doi.org/10.1038/nbt.3893.

Brandini, F.P. *et al.* (2000) ‘Multiannual trends in fronts and distribution of nutrients and chlorophyll in the southwestern Atlantic (30–62°S)’, *Deep Sea Research Part I: Oceanographic Research Papers*, 47(6), pp. 1015–1033. Available at: https://doi.org/10.1016/S0967-0637(99)00075-8.

Breusing, C. *et al.* (2023) ‘Ecological differences among hydrothermal vent symbioses may drive contrasting patterns of symbiont population differentiation’. bioRxiv, p. 2022.08.30.505939. Available at: https://doi.org/10.1101/2022.08.30.505939.

Brito, I.L. (2021) ‘Examining horizontal gene transfer in microbial communities’, *Nature Reviews Microbiology*, 19(7), pp. 442–453. Available at: https://doi.org/10.1038/s41579-021-00534-7.

Brown, M.V. *et al.* (2012) ‘Global biogeography of SAR11 marine bacteria’, *Molecular Systems Biology*, 8(1), p. 595. Available at: https://doi.org/10.1038/msb.2012.28.

Brown, M.V. *et al.* (2018) ‘Systematic, continental scale temporal monitoring of marine pelagic microbiota by the Australian Marine Microbial Biodiversity Initiative’, *Scientific Data*, 5(1), p. 180130. Available at: https://doi.org/10.1038/sdata.2018.130.

Chan, A.H., Jenkins, P.A. and Song, Y.S. (2012) ‘Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster’, *PLOS Genetics*, 8(12), p. e1003090. Available at: https://doi.org/10.1371/journal.pgen.1003090.

Claverys, J.-P., Martin, B. and Polard, P. (2009) ‘The genetic transformation machinery: composition, localization, and mechanism’, *FEMS Microbiology Reviews*, 33(3), pp. 643–656. Available at: https://doi.org/10.1111/j.1574-6976.2009.00164.x.

Croucher, N.J. *et al.* (2012) ‘A High-Resolution View of Genome-Wide Pneumococcal Transformation’, *PLOS Pathogens*, 8(6), p. e1002745. Available at: https://doi.org/10.1371/journal.ppat.1002745.

Danecek, P. *et al.* (2021) ‘Twelve years of SAMtools and BCFtools’, *GigaScience*, 10(2), p. giab008. Available at: https://doi.org/10.1093/gigascience/giab008.

Di Tommaso, P. *et al.* (2017) ‘Nextflow enables reproducible computational workflows’, *Nature Biotechnology*, 35(4), pp. 316–319. Available at: https://doi.org/10.1038/nbt.3820.

Didelot, X. and Falush, D. (2007) ‘Inference of Bacterial Microevolution Using Multilocus Sequence Data’, *Genetics*, 175(3), pp. 1251–1266. Available at: https://doi.org/10.1534/genetics.106.063305.

Didelot, X. and Maiden, M.C. (2010) ‘Impact of recombination on bacterial evolution’, *Trends Microbiol*. 2010/05/11 edn, 18(7), pp. 315–22. Available at: https://doi.org/10.1016/j.tim.2010.04.002.

Didelot, X. and Wilson, D.J. (2015) ‘ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes’, *PLOS Computational Biology*, 11(2), p. e1004041. Available at: https://doi.org/10.1371/journal.pcbi.1004041.

Doane, M.P. *et al.* (2023) ‘Defining marine bacterioplankton community assembly rules by contrasting the importance of environmental determinants and biotic interactions’, *Environmental Microbiology*, 25(6), pp. 1084–1098. Available at: https://doi.org/10.1111/1462-2920.16341.

Escobar-Zepeda, A., Vera-Ponce de León, A. and Sanchez-Flores, A. (2015) ‘The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics’, *Frontiers in Genetics*, 6, p. 348. Available at: https://doi.org/10.3389/fgene.2015.00348.

Eyre-Walker, A. and Keightley, P.D. (2007) ‘The distribution of fitness effects of new mutations’, *Nature Reviews Genetics*, 8(8), pp. 610–618. Available at: https://doi.org/10.1038/nrg2146.

Falkowski, P.G., Fenchel, T. and Delong, E.F. (2008) ‘The Microbial Engines That Drive Earth’s Biogeochemical Cycles’, *Science*, 320(5879), pp. 1034–1039. Available at: https://doi.org/10.1126/science.1153213.

Fearnhead, P. and Donnelly, P. (2001) ‘Estimating recombination rates from population genetic data.’, *Genetics*, 159(3), pp. 1299–1318.

Fearnhead, P. and Donnelly, P. (2002) ‘Approximate likelihood methods for estimating local recombination rates’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), pp. 657–680. Available at: https://doi.org/10.1111/1467-9868.00355.

Ferenci, T. (2019) ‘Irregularities in genetic variation and mutation rates with environmental stresses’, *Environmental Microbiology*, 21(11), pp. 3979–3988. Available at: https://doi.org/10.1111/1462-2920.14822.

Fitzgerald, D.M. and Rosenberg, S.M. (2019) ‘What is mutation? A chapter in the series: How microbes “jeopardize” the modern synthesis’, *PLOS Genetics*, 15(4), p. e1007995. Available at: https://doi.org/10.1371/journal.pgen.1007995.

Flombaum, P. *et al.* (2013) ‘Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus’, *Proceedings of the National Academy of Sciences*, 110(24), pp. 9824–9829. Available at: https://doi.org/10.1073/pnas.1307701110.

Fu, Y.-X. and Li, W.-H. (1999) ‘Coalescing into the 21st Century: An Overview and Prospects of Coalescent Theory’, *Theoretical Population Biology*, 56(1), pp. 1–10. Available at: https://doi.org/10.1006/tpbi.1999.1421.

Fuhrman, J.A. *et al.* (2006) ‘Annually reoccurring bacterial communities are predictable

from ocean conditions', *Proceedings of the National Academy of Sciences*, 103(35), pp. 13104–13109. Available at: https://doi.org/10.1073/pnas.0602399103.

Fuszard, M.A., Wright, P.C. and Biggs, C.A. (2012) 'Comparative quantitative proteomics of prochlorococcus ecotypes to a decrease in environmental phosphate concentrations', *Aquatic Biosystems*, 8(1), p. 7. Available at: https://doi.org/10.1186/2046-9063-8-7.

Garrison, E. and Marth, G. (2012) 'Haplotype-based variant detection from short-read sequencing', *arXiv:1207.3907 [q-bio]* [Preprint]. Available at: http://arxiv.org/abs/1207.3907 (Accessed: 8 March 2022).

Giovannoni, S.J. *et al.* (2005) 'Genome Streamlining in a Cosmopolitan Oceanic Bacterium', *Science*, 309(5738), pp. 1242–1245. Available at: https://doi.org/10.1126/science.1114057.

Giovannoni, S.J. (2017) 'SAR11 Bacteria: The Most Abundant Plankton in the Oceans', *Annual Review of Marine Science*, 9(1), pp. 231–255. Available at: https://doi.org/10.1146/annurev-marine-010814-015934.

Gogarten, J.P. and Townsend, J.P. (2005) 'Horizontal gene transfer, genome innovation and evolution', *Nature Reviews Microbiology*, 3(9), pp. 679–687. Available at: https://doi.org/10.1038/nrmicro1204.

González-Torres, P. *et al.* (2019) 'Impact of Homologous Recombination on the Evolution of Prokaryotic Core Genomes', *mBio*, 10(1). Available at: https://doi.org/10.1128/mBio.02494-18.

Guirouilh-Barbat, J. *et al.* (2014) 'Is homologous recombination really an error-free process?', *Frontiers in Genetics*, 5. Available at: https://doi.org/10.3389/fgene.2014.00175.

Gutiérrez, D., Akester, M. and Naranjo, L. (2016) 'Productivity and Sustainable Management of the Humboldt Current Large Marine Ecosystem under climate change', *Environmental Development*, 17, pp. 126–144. Available at: https://doi.org/10.1016/j.envdev.2015.11.004.

Guttman, D.S. and Dykhuizen, D.E. (1994) 'Clonal Divergence in Escherichia coli as a Result of Recombination, Not Mutation', *Science*, 266(5189), pp. 1380–1383. Available at: https://doi.org/10.1126/science.7973728.

Hanage, W.P. (2016) 'Not So Simple After All: Bacteria, Their Population Genetics, and Recombination', *Cold Spring Harbor Perspectives in Biology*, 8(7). Available at: https://doi.org/10.1101/cshperspect.a018069.

Handelsman, J. *et al.* (1998) 'Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products', *Chemistry & Biology*, 5(10), pp. R245–R249. Available at: https://doi.org/10.1016/S1074-5521(98)90108-9.

Handelsman, J. (2004) 'Metagenomics: Application of Genomics to Uncultured Microorganisms', *Microbiology and Molecular Biology Reviews*, 68(4), pp. 669–685. Available at: https://doi.org/10.1128/MMBR.68.4.669-685.2004.

Hayat, R. *et al.* (2010) 'Soil beneficial bacteria and their role in plant growth promotion: a review', *Annals of Microbiology*, 60(4), pp. 579–598. Available at: https://doi.org/10.1007/s13213-010-0117-1.

Hermann, P. *et al.* (2019) ‘LDJump: Estimating variable recombination rates from population genetic data’, *Molecular Ecology Resources*, 19(3), pp. 623–638. Available at: https://doi.org/10.1111/1755-0998.12994.

Hershberg, R. (2015) ‘Mutation—The Engine of Evolution: Studying Mutation and Its Role in the Evolution of Bacteria’, *Cold Spring Harbor Perspectives in Biology*, 7(9), p. a018077. Available at: https://doi.org/10.1101/cshperspect.a018077.

Horton, J.S. and Taylor, T.B. (2023) ‘Mutation bias and adaptation in bacteria’, *Microbiology*, 169(11), p. 001404. Available at: https://doi.org/10.1099/mic.0.001404.

Huang, W. *et al.* (2012) ‘ART: a next-generation sequencing read simulator’, *Bioinformatics (Oxford, England)*, 28(4), pp. 593–594. Available at: https://doi.org/10.1093/bioinformatics/btr708.

Hudson, R.R. (2001) ‘Two-Locus Sampling Distributions and Their Application’, *Genetics*, 159(4), pp. 1805–1817.

Iranzo, J. *et al.* (2019) ‘Gene gain and loss push prokaryotes beyond the homologous recombination barrier and accelerate genome sequence divergence’, *Nature Communications*, 10(1), p. 5376. Available at: https://doi.org/10.1038/s41467-019-13429-2.

Johnson, P.L. and Slatkin, M. (2009) ‘Inference of microbial recombination rates from metagenomic data’, *PLoS genetics*, 5(10), p. e1000674.

Johnson, P.L.F. and Slatkin, M. (2006) ‘Inference of population genetic parameters in metagenomics: A clean look at messy data’, *Genome Research*, 16(10), pp. 1320–1327. Available at: https://doi.org/10.1101/gr.5431206.

Kamm, J.A. *et al.* (2016) ‘Two-Locus Likelihoods Under Variable Population Size and Fine-Scale Recombination Rate Estimation’, *Genetics*, 203(3), pp. 1381–1399. Available at: https://doi.org/10.1534/genetics.115.184820.

Kelleher, J., Etheridge, A.M. and McVean, G. (2016) ‘Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes’, *PLOS Computational Biology*, 12(5), p. e1004842. Available at: https://doi.org/10.1371/journal.pcbi.1004842.

Kolmogorov, M. *et al.* (2020) ‘metaFlye: scalable long-read metagenome assembly using repeat graphs’, *Nature Methods*, 17(11), pp. 1103–1110. Available at: https://doi.org/10.1038/s41592-020-00971-x.

Korneliussen, T.S., Albrechtsen, A. and Nielsen, R. (2014) ‘ANGSD: Analysis of Next Generation Sequencing Data’, *BMC Bioinformatics*, 15(1), p. 356. Available at: https://doi.org/10.1186/s12859-014-0356-4.

Krishnan, S. *et al.* (2023) ‘Rhometa: Population recombination rate estimation from metagenomic read datasets’, *PLOS Genetics*, 19(3), p. e1010683. Available at: https://doi.org/10.1371/journal.pgen.1010683.

Lawrence, J.G. (1999) ‘Gene transfer, speciation, and the evolution of bacterial genomes’, *Current Opinion in Microbiology*, 2(5), pp. 519–523. Available at: https://doi.org/10.1016/S1369-5274(99)00010-7.

Le, L.A.T. *et al.* (2020) ‘Nutritional conditions and oxygen concentration affect spontaneous occurrence of homologous recombination events but not spontaneous

mutagenesis in Escherichia coli', *Genes & Genetic Systems*, 95(2), pp. 85–93. Available at: https://doi.org/10.1266/ggs.19-00008.

Levin, B.R. and Cornejo, O.E. (2009) 'The Population and Evolutionary Dynamics of Homologous Gene Recombination in Bacteria', *PLOS Genetics*, 5(8), p. e1000601. Available at: https://doi.org/10.1371/journal.pgen.1000601.

Li, H. (2013) 'Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM'. Available at: https://arxiv.org/abs/1303.3997v2 (Accessed: 17 September 2021).

Li, S.-J. *et al.* (2014) 'Microbial communities evolve faster in extreme environments', *Scientific Reports*, 4(1), p. 6205. Available at: https://doi.org/10.1038/srep06205.

Lin, M. and Kussell, E. (2019) 'Inferring bacterial recombination rates from large-scale sequencing datasets', *Nature Methods*, 16(2), pp. 199–204. Available at: https://doi.org/10.1038/s41592-018-0293-7.

Logares, R. (2024) 'Decoding populations in the ocean microbiome', *Microbiome*, 12(1), p. 67. Available at: https://doi.org/10.1186/s40168-024-01778-0.

López-Pérez, M. *et al.* (2020) 'The Evolutionary Success of the Marine Bacterium SAR11 Analyzed through a Metagenomic Perspective', *mSystems*, 5(5). Available at: https://doi.org/10.1128/mSystems.00605-20.

Luo, H. and Moran, M.A. (2014) 'Evolutionary Ecology of the Marine Roseobacter Clade', *Microbiology and Molecular Biology Reviews : MMBR*, 78(4), pp. 573–587. Available at: https://doi.org/10.1128/MMBR.00020-14.

Lynch, T.P. *et al.* (2008) 'A National Reference Station infrastructure for Australia - Using telemetry and central processing to report multi-disciplinary data streams for monitoring marine ecosystem response to climate change', in *OCEANS 2008*. *OCEANS 2008*, pp. 1–8. Available at: https://doi.org/10.1109/OCEANS.2008.5151856.

Lynch, T.P. *et al.* (2014) 'IMOS National Reference Stations: A Continental-Wide Physical, Chemical and Biological Coastal Observing System', *PLOS ONE*, 9(12), p. e113652. Available at: https://doi.org/10.1371/journal.pone.0113652.

Maiden, M.C.J. *et al.* (1998) 'Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms', *Proceedings of the National Academy of Sciences*, 95(6), pp. 3140–3145. Available at: https://doi.org/10.1073/pnas.95.6.3140.

Majewski, J. (2001) 'Sexual isolation in bacteria', *FEMS Microbiology Letters*, 199(2), pp. 161–169. Available at: https://doi.org/10.1111/j.1574-6968.2001.tb10668.x.

Martin, B. *et al.* (2024) 'Studies of the Ecology of the Benguela Current Upwelling System: The TRAFFIC Approach', in G.P. von Maltitz et al. (eds) *Sustainability of Southern African Ecosystems under Global Change: Science for Management and Policy Interventions*. Cham: Springer International Publishing, pp. 277–312. Available at: https://doi.org/10.1007/978-3-031-10948-5_11.

Martiny, H.-M. *et al.* (2022) 'A curated data resource of 214K metagenomes for characterization of the global antimicrobial resistome', *PLOS Biology*, 20(9), p. e3001792. Available at: https://doi.org/10.1371/journal.pbio.3001792.

McFall-Ngai, M. *et al.* (2013) ‘Animals in a bacterial world, a new imperative for the life sciences’, *Proceedings of the National Academy of Sciences*, 110(9), pp. 3229–3236. Available at: https://doi.org/10.1073/pnas.1218525110.

McGillicuddy Jr, D.J. (2016) ‘Mechanisms of Physical-Biological-Biogeochemical Interaction at the Oceanic Mesoscale’, *Annual Review of Marine Science*, 8(Volume 8, 2016), pp. 125–159. Available at: https://doi.org/10.1146/annurev-marine-010814-015606.

McVean, G., Awadalla, P. and Fearnhead, P. (2002) ‘A Coalescent-Based Method for Detecting and Estimating Recombination From Gene Sequences’, *Genetics*, 160(3), pp. 1231–1241.

Melendrez, M.C. *et al.* (2016) ‘Recombination Does Not Hinder Formation or Detection of Ecological Species of Synechococcus Inhabiting a Hot Spring Cyanobacterial Mat’, *Frontiers in Microbiology*, 6. Available at: https://www.frontiersin.org/articles/10.3389/fmicb.2015.01540 (Accessed: 28 July 2022).

Messer, L.F. *et al.* (2020) ‘Microbial tropicalization driven by a strengthening western ocean boundary current’, *Global Change Biology*, 26(10), pp. 5613–5629. Available at: https://doi.org/10.1111/gcb.15257.

Mirdita, M., Steinegger, M. and Söding, J. (2019) ‘MMseqs2 desktop and local web server app for fast, interactive sequence searches’, *Bioinformatics*, 35(16), pp. 2856–2858. Available at: https://doi.org/10.1093/bioinformatics/bty1057.

Morris, R.M. *et al.* (2002) ‘SAR11 clade dominates ocean surface bacterioplankton communities’, *Nature*, 420(6917), pp. 806–810. Available at: https://doi.org/10.1038/nature01240.

Muller, H.J. (1964) ‘The relation of recombination to mutational advance’, *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 1(1), pp. 2–9. Available at: https://doi.org/10.1016/0027-5107(64)90047-8.

Nei, M. and Nozawa, M. (2011) ‘Roles of Mutation and Selection in Speciation: From Hugo de Vries to the Modern Genomic Era’, *Genome Biology and Evolution*, 3, pp. 812–829. Available at: https://doi.org/10.1093/gbe/evr028.

O’Brien, J. *et al.* (2022) ‘Biogeographical and seasonal dynamics of the marine Roseobacter community and ecological links to DMSP-producing phytoplankton’, *ISME Communications*, 2(1), pp. 1–13. Available at: https://doi.org/10.1038/s43705-022-00099-3.

Orakov, A. *et al.* (2021) ‘GUNC: detection of chimerism and contamination in prokaryotic genomes’, *Genome Biology*, 22(1), p. 178. Available at: https://doi.org/10.1186/s13059-021-02393-0.

Parks, D.H. *et al.* (2022) ‘GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy’, *Nucleic Acids Research*, 50(D1), pp. D785–D794. Available at: https://doi.org/10.1093/nar/gkab776.

Paulsson, J. *et al.* (2017) ‘The processive kinetics of gene conversion in bacteria’, *Molecular Microbiology*, 104(5), pp. 752–760. Available at:

https://doi.org/10.1111/mmi.13661.

Popa, O. and Dagan, T. (2011) ‘Trends and barriers to lateral gene transfer in prokaryotes’, *Current Opinion in Microbiology*, 14(5), pp. 615–623. Available at: https://doi.org/10.1016/j.mib.2011.07.027.

Qadir, M. *et al.* (2024) ‘Exploring Plant–Bacterial Symbiosis for Eco-Friendly Agriculture and Enhanced Resilience’, *International Journal of Molecular Sciences*, 25(22), p. 12198. Available at: https://doi.org/10.3390/ijms252212198.

Rinke, C. *et al.* (2019) ‘A phylogenomic and ecological analysis of the globally abundant Marine Group II archaea (Ca. Poseidoniales ord. nov.)’, *The ISME Journal*, 13(3), pp. 663–675. Available at: https://doi.org/10.1038/s41396-018-0282-y.

Rosenberg, N.A. and Nordborg, M. (2002) ‘Genealogical trees, coalescent theory and the analysis of genetic polymorphisms’, *Nature Reviews Genetics*, 3(5), pp. 380–390. Available at: https://doi.org/10.1038/nrg795.

Rusch, D.B. *et al.* (2007) ‘The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific’, *PLOS Biology*, 5(3), p. e77. Available at: https://doi.org/10.1371/journal.pbio.0050077.

Schmutzer, M. and Barraclough, T.G. (2019) ‘The role of recombination, niche-specific gene pools and flexible genomes in the ecological speciation of bacteria’, *Ecol Evol*. 2019/04/30 edn, 9(8), pp. 4544–4556. Available at: https://doi.org/10.1002/ece3.5052.

Shapiro, B.J. (2016) ‘How clonal are bacteria over time?’, *Current Opinion in Microbiology*, 31, pp. 116–123. Available at: https://doi.org/10.1016/j.mib.2016.03.013.

Sharpton, T.J. (2014) ‘An introduction to the analysis of shotgun metagenomic data’, *Frontiers in Plant Science*, 5, p. 209. Available at: https://doi.org/10.3389/fpls.2014.00209.

Singh, J. *et al.* (2009) ‘Metagenomics: Concept, methodology, ecological inference and recent advances’, *Biotechnology Journal*, 4(4), pp. 480–494. Available at: https://doi.org/10.1002/biot.200800201.

Siokou-Frangou, I. *et al.* (2010) ‘Plankton in the open Mediterranean Sea: a review’, *Biogeosciences*, 7(5), pp. 1543–1586. Available at: https://doi.org/10.5194/bg-7-1543-2010.

Spence, J.P. and Song, Y.S. (2019) ‘Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations’, *Science Advances*, 5(10), p. eaaw9206. Available at: https://doi.org/10.1126/sciadv.aaw9206.

Spilling, K. *et al.* (2019) ‘Impacts of reduced inorganic N:P ratio on three distinct plankton communities in the Humboldt upwelling system’, *Marine Biology*, 166(9), p. 114. Available at: https://doi.org/10.1007/s00227-019-3561-x.

Stumpf, M.P.H. and McVean, G.A.T. (2003) ‘Estimating recombination rates from population-genetic data’, *Nature Reviews Genetics*, 4(12), pp. 959–968. Available at: https://doi.org/10.1038/nrg1227.

Sun, Y. and Luo, H. (2018) ‘Homologous Recombination in Core Genomes Facilitates Marine Bacterial Adaptation’, *Applied and Environmental Microbiology*, 84(11), pp. e02545-17. Available at: https://doi.org/10.1128/AEM.02545-17.

Sunagawa, S. *et al.* (2020) ‘Tara Oceans: towards global ocean ecosystems biology’, *Nature Reviews Microbiology*, 18(8), pp. 428–445. Available at: https://doi.org/10.1038/s41579-020-0364-5.

Tan, G. *et al.* (2019) ‘Long fragments achieve lower base quality in Illumina paired-end sequencing’, *Scientific Reports*, 9(1), p. 2856. Available at: https://doi.org/10.1038/s41598-019-39076-7.

Tataru, P. *et al.* (2017) ‘Statistical Inference in the Wright–Fisher Model Using Allele Frequency Data’, *Systematic Biology*, 66(1), pp. e30–e46. Available at: https://doi.org/10.1093/sysbio/syw056.

Thomas, C.M. and Nielsen, K.M. (2005) ‘Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria’, *Nature Reviews Microbiology*, 3(9), pp. 711–721. Available at: https://doi.org/10.1038/nrmicro1234.

Thomas, T., Gilbert, J. and Meyer, F. (2012) ‘Metagenomics - a guide from sampling to data analysis’, *Microb Inform Exp*. 2012/05/17 edn, 2(1), p. 3. Available at: https://doi.org/10.1186/2042-5783-2-3.

Vergin, K.L. *et al.* (2007) ‘High intraspecific recombination rate in a native population of Candidatus Pelagibacter ubique (SAR11)’, *Environmental Microbiology*, 9(10), pp. 2430–2440. Available at: https://doi.org/10.1111/j.1462-2920.2007.01361.x.

Virtanen, P. *et al.* (2020) ‘SciPy 1.0: fundamental algorithms for scientific computing in Python’, *Nature Methods*, 17(3), pp. 261–272. Available at: https://doi.org/10.1038/s41592-019-0686-2.

Vos, M. and Didelot, X. (2009) ‘A comparison of homologous recombination rates in bacteria and archaea’, *The ISME Journal*, 3(2), pp. 199–208. Available at: https://doi.org/10.1038/ismej.2008.93.

Wang, H. *et al.* (2023) ‘Spatial and temporal dynamics of microbial community composition and factors influencing the surface water and sediments of urban rivers’, *Journal of Environmental Sciences*, 124, pp. 187–197. Available at: https://doi.org/10.1016/j.jes.2021.10.016.

Wang, X. *et al.* (2020) ‘Cryptic speciation of a pelagic Roseobacter population varying at a few thousand nucleotide sites’, *The ISME Journal*, 14(12), pp. 3106–3119. Available at: https://doi.org/10.1038/s41396-020-00743-7.

Wani, A.K. *et al.* (2022) ‘Microbial adaptation to different environmental conditions: molecular perspective of evolved genetic and cellular systems’, *Archives of Microbiology*, 204(2), p. 144. Available at: https://doi.org/10.1007/s00203-022-02757-5.

Waskom, M.L. (2021) ‘seaborn: statistical data visualization’, *Journal of Open Source Software*, 6(60), p. 3021. Available at: https://doi.org/10.21105/joss.03021.

Watterson, G.A. (1975) ‘On the number of segregating sites in genetical models without recombination’, *Theoretical Population Biology*, 7(2), pp. 256–276. Available at: https://doi.org/10.1016/0040-5809(75)90020-9.

Whitman, W.B., Coleman, D.C. and Wiebe, W.J. (1998) ‘Prokaryotes: The unseen majority’, *Proceedings of the National Academy of Sciences*, 95(12), pp. 6578–6583. Available at: https://doi.org/10.1073/pnas.95.12.6578.

Wiedenbeck, J. and Cohan, F.M. (2011) ‘Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches’, *FEMS Microbiology Reviews*, 35(5), pp. 957–976. Available at: https://doi.org/10.1111/j.1574-6976.2011.00292.x.

Willson, H.R. and Rees, N.W. (2000) ‘Classification of mesoscale features in the Brazil-Falkland Current confluence zone’, *Progress in Oceanography*, 45(3), pp. 415–426. Available at: https://doi.org/10.1016/S0079-6611(00)00011-2.

Woodcroft, B.J. (2023) ‘CoverM’. Available at: https://github.com/wwood/CoverM (Accessed: 7 February 2023).

Woodcroft, B.J. *et al.* (2024) ‘SingleM and Sandpiper: Robust microbial taxonomic profiles from metagenomic data’. bioRxiv, p. 2024.01.30.578060. Available at: https://doi.org/10.1101/2024.01.30.578060.

Woodford, N. and Ellington, M.J. (2007) ‘The emergence of antibiotic resistance by mutation’, *Clinical Microbiology and Infection*, 13(1), pp. 5–18. Available at: https://doi.org/10.1111/j.1469-0691.2006.01492.x.

Wooley, J.C., Godzik, A. and Friedberg, I. (2010) ‘A primer on metagenomics’, *PLoS Comput Biol*. 2010/03/03 edn, 6(2), p. e1000667. Available at: https://doi.org/10.1371/journal.pcbi.1000667.

Yahara, K. *et al.* (2016) ‘The Landscape of Realized Homologous Recombination in Pathogenic Bacteria’, *Molecular Biology and Evolution*, 33(2), pp. 456–471. Available at: https://doi.org/10.1093/molbev/msv237.

Yoshitake, K. *et al.* (2021) ‘Development of a time-series shotgun metagenomics database for monitoring microbial communities at the Pacific coast of Japan’, *Scientific Reports*, 11(1), p. 12222. Available at: https://doi.org/10.1038/s41598-021-91615-3.

Zaremba-Niedzwiedzka, K. *et al.* (2013) ‘Single-cell genomics reveal low recombination frequencies in freshwater bacteria of the SAR11 clade’, *Genome Biology*, 14(11), p. R130. Available at: https://doi.org/10.1186/gb-2013-14-11-r130.

Zhang, Y.-J. *et al.* (2015) ‘Impacts of Gut Bacteria on Human Health and Diseases’, *International Journal of Molecular Sciences*, 16(4), pp. 7493–7519. Available at: https://doi.org/10.3390/ijms16047493.

Zinser, E.R. *et al.* (2007) ‘Influence of light and temperature on Prochlorococcus ecotype distributions in the Atlantic Ocean’, *Limnology and Oceanography*, 52(5), pp. 2205–2220. Available at: https://doi.org/10.4319/lo.2007.52.5.2205.