EXAMINING THE ROLE OF HOST DEPENDENCE ON THE ECOLOGY AND EVOLUTION OF *EPULOSCIUM* SPP. AND THEIR RELATIVES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Francine Arroyo

August 2019

EXAMINING THE ROLE OF HOST-DEPENDENCE ON THE EVOLUTION AND ECOLOGY OF *EPULOPISCIUM* SPP. AND THEIR RELATIVES

Francine Arroyo, Ph. D.

Cornell University 2019

Studies of the gut microbiome have enlightened our perspective on the contributions of microbes to animal health. Commensal bacteria have evolved strategies to maintain associations with their host and gain a foothold in this competitive environment through colonization, maintenance and transmission. Here I studied how dominant commensals of herbivorous surgeonfish have evolved strategies to overcome the obstacles to host association and as a result of this interaction, further explain the unusual biology of these commensals. *Epulopiscium* spp. and relatives (known as epulos) are morphologically diverse and form a monophyletic clade within the Lachnospiraceae XIVb cluster (Order: Clostridiales). They are renowned for their large size, some reaching lengths up to 0.6 mm, and their extreme polyploidy, containing 100,000s of copies of a ~3.2 Mb genome. The reproductive cycle of epulos follows a predictable diurnal pattern. Establishing a single-cell sequencing technique, I explored questions that were once unattainable for this uncultured bacterium. First, I conducted a population analysis of the mutually exclusive interaction between *Epulopiscium* sp. type B and *Naso tonganus* in the Great Barrier Reef. I found evidence of allopatric speciation, governed by host migratory preferences between reef and island habitats. Maintaining genomic diversity for the symbiont depends on horizontal transmission by the host and rampant recombination to overcome genome

purification in this polyploid bacteria. Next, I explored the distribution and diversity of co-resident symbiont populations across three *Acanthurus* species. Genomic insights into their metabolic potential confirmed that the two cohabitating epulos convert ammonia/urea to amino acids. However, differences among these epulo lineages for the ability to degrade dietary complex polysaccharides and diverse strategies for conserving energy suggest niche differentiation in this competitive environment. Lastly, by examining the transcriptomic profiles of an endospore-forming epulo in *Naso unicornis*, I have shown that the metabolic and reproductive life cycle of this epulo population is synchronized with the feeding/fasting cycles of the host. In conclusion, I suggest that the evolution of epulos is largely influenced by their dependency on surgeonfish hosts. A possible reason for successful transmission and maintenance of epulos in herbivorous surgeonfish worldwide is their ability to anticipate the circadian cycles of their host.

BIOGRAPHICAL SKETCH

Francine Arroyo was born on May 11 in Torrance CA and raised in sunny San Diego. She is the daughter of Central American immigrants and the only sister of six brothers. Francine was the first in her family to attend college. She enrolled at the University of Arizona (UofA) to study architecture, but switched her major to Biology after falling in love with the beautiful complexity of biological systems. She transferred to Grossmont Community College in San Diego and then to UCSD where she received her B.S. in General Biology. Francine worked full-time as a Starbucks barista during college to help pay for her tuition. Her degree eventually helped her leave the service industry and gain experience in biotech and clinical research. She worked as a Research Associate at Pacira Pharmaceuticals, Inc., and then as a Research Administrative Assistant at the Pediatric Dermatology clinic at Rady Children's Hospital. At Rady's, Francine helped physicians and resident fellows design research studies and submit them for IRB approval. She was inspired to pursue her own research and returned to school, where she earned her M.S. in Biology from Humboldt State University. She investigated the physiology and ecological role of iron-oxidizing bacteria from volcanic hot springs in Lassen National Volcanic Park. Her joyful and inquisitive advisor, Prof. Patricia L. Seiring, empowered her to continue her education in microbiology. Francine pursues a PhD in Microbiology at Cornell University (Patty's graduate school alma mater), studying unusual microbial physiology and ecology at Cornell University. At Cornell, she studies the giant intestinal bacteria of coral reef-dwelling surgeonfish with Prof. Esther Angert. When she's not in the field or the lab, she spends her time with her best friend and husband Max.

To Patty, who taught me how to love, play, and study all at the same time.
Even after her passing, her memory still inspires:

"A ripple in still water
when there is no pebble tossed
nor wind to blow" – Grateful Dead

To the world's oceans.

During the course of this study, the Great Barrier Reef suffered consecutive bleaching events that affected half of the world's coral reefs. This dissertation documents the hidden wonders of our seas and the diverse life that depends on them.

ACKNOWLEDGMENTS

Thanks to my patient and supportive advisor, Esther Angert, for giving me the opportunity to study the amazing *Epulopiscium*-surgeonfish system. Over the past six years, you have helped me learn to become a better researcher, communicator, mentor, teacher, advocate, and person. You supported and embraced my wild ideas, from going all out on lab Halloween costumes, to sampling all kinds of host tissue for preservation in order to improve future metadata analyses. I'll always cherish these memories. Thanks to our Lizard Island collaborators: Howard Choat, Kendall Clements, Lindsey White, Paul Caiger, and Will Robbins. Your expertise in fish biology and ecology filled a significant void in my studies. I appreciate all the time and risks you encountered to collect all the fish we needed for my work. I also commemorate the fish that paid the ultimate price. Thanks to all of the members of the Angert lab throughout the years. My generous lab mate Dave Sannino for always making the time to discuss ideas and help me analyze my work. Verena Carvalho-Salman for training me on the microsensors and sharing your life with me. Chuck Pepey-Rainey for giving me a master class on bioinformatics approaches and metagenome assembly. Jen Grenier from the RNA Sequencing Core for her diligent sequencing efforts and advice on RNA analyses. Katie Edwards for reviewing my sulfide calculations. Stephen Zinder for assisting me with methane analysis. To Wenbo Chen for help with preliminary analyses of RNA data. To Eileen Randall, for being my running buddy and providing me with emotional support. To all the undergrads mentees: Zanah Francis, Josh Heller, Alejandro Schmeider, Mike Wilson, Crystal Gill, Karen Giraldo Franco, Isaiah Kela-Pacheco, Selena Lu and Sydney Luitgaarden. You provided me constant reminders of why I love teaching. Your enthusiasm and questions helped keep my work and attitude fresh. To my professional and personal advisors: Ahmed Gaballa, Teresa Pawlowska, Ankita Sachla, Tory

Hendry, and Melanie Smee. I always felt I could call on you to discuss methodology and theory or receive advice on career and life management. To my committee members, Ian Hewson and Joe Peters, for their advice on this dissertation, career development, and advocating for my success. To Dan Buckley for use of his server. To Lynn Johnson from the CBSU for providing me with help on using and understanding statistics on R. To Kalia Bistolas, my sister labmate from another mister, for all your support troubleshooting bioinformatic analyses and sharing your experiences of your grad school life. To the Graduate school, particularly Jan Allen, Colleen McLinn, Janna Lamey, and Christine Holmes, for providing me with career and professional advice as well as a number of valuable resources (BRB Dissertation Write-in, career and writing workshops). To the Microbiology teaching staff: Barb, Rania, and Sue. You make teaching fun and approachable. To the Microbiology Admin staff (Shirley, Lillian, Patti, and Tina) for all your support and friendly attitudes. To my family, for you loving support while holding the fort back at home. We lost many family members over the years, reminding us how precious our family bonds are and how important it is to keep communicating. Finally, to my best friend and husband Maximillian Cannon. You witnessed all the highs and lows of grad school from me. Your calm and loving support helped keep me moving forward through all the obstacles. To all, thank you!

TABLE OF CONTENTS

CHAPTER 3

CHARACTERIZING THE DAILY TRANSCRIPTIONAL OSCILLATIONS OF AN ENDOSPORE-FORMING GUT SYMBIONT OF *NASO UNICORNIS*

APPENDIX A

PHYSIOCHEMICAL ANALYSIS OF THE HINDGUT CHAMBER FROM *POMACANTHUS SEXSTRIATUS*

LIST OF FIGURES

CHAPTER 1

CHAPTER 2

CHAPTER 3

APPENDIX A

LIST OF TABLES

CHAPTER 1

CHAPTER 2

CHAPTER 3

APPENDIX A

CHAPTER 1

RECOMBINATION CONTRIBUTES TO POPULATION DIVERSIFICATION IN THE POLYPLOID INTESTINAL SYMBIONT *EPULOPISCIUM* SP. TYPE B[1]

ABSTRACT

Epulopiscium sp. type B (Lachnospiraceae) is an exceptionally large, highly polyploid, intestinal symbiont of the coral reef dwelling surgeonfish *Naso tonganus*. These obligate anaerobes do not form mature endospores and reproduce solely through the production of multiple intracellular offspring. This likely makes them dependent on immediate transfer to a receptive host for dispersal. During reproduction, only a small proportion of *Epulopiscium* mother-cell DNA is inherited. To explore the impact of this unusual viviparous lifestyle on symbiont population dynamics, we investigated *Epulopiscium* sp. type B and their fish hosts collected over the course of two decades, at island and reef habitats near Lizard Island, Australia. Using multi-locus sequence analysis, we found that recombination plays an important role in the life history of these symbionts. We suggest that congression of unlinked markers may contribute to linkage disequilibrium observed in this and other recombinant populations of bacteria. Codiversification analyses and traits of this partnership suggest that while symbionts are obligately dependent on their host, the host has a facultative association with *Epulopiscium*. Symbiont populations showed spatial but not temporal partitioning. Surgeonfish are long-lived and capable of traveling long distances, yet the population

[1]Presented with minor modification from the original published article: Arroyo FA, Pawlowska TE, Choat JH, Clements KD, Angert ER. Recombination contributes to population diversification in the polyploid intestinal symbiont *Epulopiscium* sp. type B. ISME J. 2019; 13:1084-1097.
All supplementary information can be found online at: https://www.nature.com/articles/s41396-018-0339-y#Sec22
Pawlowska TE contributed to manuscript review and discussions about population biology; Choat JH and Clements KD collected fish and shared their expertise on surgeonfish ecology.

structures of *Epulopiscium* suggest that adult fish tend to not roam beyond a limited locale. The findings here inform our understanding of evolutionary processes within intestinal Lachnospiraceae populations.

INTRODUCTION

Epulopiscium spp. and related bacteria, known as epulos, are a morphologically diverse group of intestinal symbionts renowned for their large size and disparate reproductive strategies (1–4). Phylogenetically, this group is affiliated with the Lachnospiraceae, a family which comprises functionally crucial intestinal symbionts of vertebrates (5, 6). Abundant populations of epulos are found in surgeonfish species that regularly consume algae or detritus (2). Epulos likely contribute to the breakdown of refractory algal polysaccharides ingested by their hosts and serve as important mediators of carbon flow in coral reef systems (7–9). The model for these unusual bacteria is referred to as *Epulopiscium* sp. type B, which displays many useful traits that have facilitated studies of *Epulopiscium* biology (10–14). In addition to their distinct morphology, *Epulopiscium* sp. type B is often the only epulo clade found in adult *Naso tonganus* (15, 16). Other hosts harbor multiple morphologically and phylogenetically distinct epulo lineages (2, 4, 17). Based on draft metagenome sequences from *Epulopiscium* sp. type B populations (14) (NZ_ABEQ01000000), these heterotrophic symbionts are predicted to be obligate fermenters, as neither genes for cytochromes nor genes for enzymes to relieve oxidative stress (e.g. catalase or superoxide dismutase) have been found. Thus far no free-living *Epulopiscium* sp. type B have been detected in environmental surveys. These observations along with its affilation with *Clostridium* cluster XIVb (18) suggest that the intestinal symbiont *Epulopiscium* sp. type B requires a host that regularly consumes algae for a stable anoxic environment rich in nutrients.

Many epulos produce endospores that might aid in dispersal and maintenance of populations in their surgeonfish host (4). However, *Epulopiscium* sp. type B has lost the ability to produce mature endospores and no longer uses binary fission for reproduction (14). Instead these bacteria produce multiple intracellular offspring using a derived form of endosporulation (Fig. 1). The ecological and evolutionary processes leading to this unique form of bacterial viviparity are not well understood, but may have been determined in part by the symbiotic association of *Epulopiscium* sp. type B with its particular surgeonfish host, *Naso tonganus*. In a previous study, we suggested that the manner in which symbionts enter a host intestinal tract can influence the reproductive biology of endospore forming symbionts (19). *Metabacterium polyspora*, a close relative of *Epulopiscium* spp., uses the production of multiple endospores as its primary means of reproduction in its guinea pig host. This pivot away from binary fission and toward multiple endospore formation likely reinforces the symbiotic association, as these bacteria regularly cycle out of and back into the host intestinal tract. Guinea pigs are coprophagous and reproduction by sporulation allows the offspring of *M. polyspora* to survive the oxic environment outside of the host, as well as transit through the mouth and stomach of the host. In epulos associated with fish, codiversification studies suggest specific relationships between distinct epulo clades and particular surgeonfish species that correlate with host feeding preference (17). We hypothesize that these associations impact symbiont reproductive strategies as well. Most surgeonfish that host epulos contain multiple morphotypes and multiple phylotypes (2, 4, 17). Adult *N. tonganus* are often observed to harbor only one type of large epulo, *Epulopiscium* sp. type B (15, 16). This simple system provides a unique opportunity to explore how the symbiotic alliance impacts the population structure and reproductive strategy of epulos and other gut-associated Lachnospiraceae.

Epulos, like most gut microbiota of vertebrates, are likely acquired from the

environment. However, surveys of near-complete 16S rRNA gene sequences of *Epulopiscium* sp. type B cells from many host individuals recovered minimal genetic diversity (> 99% sequence similarity), suggesting that these symbiont populations may be clonal and experience dispersal limitation (15, 20). Yet, given the dynamic life history of surgeonfish (e.g. aggregate spawning, dispersal of eggs and larvae, and pulses of juvenile settlement on a reef (21)), direct transmission of epulos from a parent to its offspring seems unlikely. A recent phylogeographic study of a broadly distributed surgeonfish species suggested that larval fish are dispersed widely (22). Furthermore, surveys of gut symbionts in fish at different life stages revealed no epulos in newly settled juvenile surgeonfish and distinct epulo communities between juveniles and adults (23). Surgeonfish most likely acquire epulos through conspecific coprophagy (23). Unlike vertically transmitted insect endosymbionts (24), populations of horizontally transmitted obligate symbionts do not appear to be as genetically constrained by dispersal bottlenecks (25, 26). The apparent clonality of *Epulopiscium* sp. type B suggests that other factors peculiar to the symbiont (e.g. physiology and cell cycle) or its host (e.g. *N. tonganus* genetics, anatomy, diet or behavior) may be influencing *Epulopiscium* sp. type B diversity.

Epulopiscium sp. type B display several remarkable features. These large cigar-shaped cells range from ~100 – 300 μm in length. They are extremely polyploid and harbor tens of thousands of genome copies (11, 13). Genomes within *Epulopiscium* sp. type B cells are hypothesized to be nearly identical (13), although some variation may exist. The daily reproductive cycle of *Epulopiscium* sp. type B (Fig. 1) is synchronized within an individual host fish (11, 12). Normally two offspring are produced each day per mother cell but as many as 12 intracellular offspring have been observed (27). Importantly, it has been estimated that during reproduction, ~1% of *Epulopiscium* sp. type B mother-cell DNA is passed on to

daughter cells (12). Yet even at late stages of offspring development, mother-cell chromosomes continue to replicate (10). Consequently, genome copies take on either somatic or germline roles (10); some intact copies are inherited by the next generation but most of the DNA appears to be required only to support mother-cell metabolism. Eventually, the somatic copies are dismantled or released into the environment by lysis of the dying mother cell (Fig. 1).

The limited inheritance of mother-cell DNA could restrict genetic diversity of the *Epulopiscium* sp. type B population by loss of novel genes or mutations accrued in somatic but not germline genome copies. Additionally, retention of daughter cells within a mother cell during much of the offspring growth cycle, which occurs separately from other developing cells, poses a physical obstacle to acquiring new genetic material through horizontal gene transfer (HGT). Furthermore, host-dependent transmission may limit the genetic material available for exchange in a confined population within an individual host. As a consequence, the overall fitness of the symbiont population is expected to decline due to clonal interference; a process where competition between individual clones with different beneficial mutations slows down the accumulation of beneficial mutations for the population as a whole (28).

We hypothesize that host-dependent interactions and the unusual reproductive biology of the symbionts are driving the evolution of *Epulopiscium* sp. type B by limiting the diversity of the symbiont populations. Epulos have not yet been cultivated in the lab and surgeonfish lose these symbionts when held in captivity (1). Therefore we used cultivation-independent techniques to explore the population structure of *Epulopiscium* sp. type B collected from wild-caught hosts. Taking advantage of the large size, polyploid genome and distinct morphology of the symbionts, we collected individual cells from archived *N. tonganus* gut contents and subjected each to single-cell, whole-genome amplification. We used a multi-locus sequence analysis (MLSA)

approach for the first fine scale population survey of *Epulopiscium* sp. type B, which allowed us to identify genetic differences between individual cells. We also investigated the *N. tonganus* population using MLSA and analyzed codiversification between partners of this host-symbiont association. Our findings reveal the importance of recombination for maintaining population diversity in this gut symbiont. We predict that the patterns of inheritance and genotype mixing we observed are driven by mechanisms that likely impact the population structure of other intestinal Lachnospiraceae. Furthermore, we suggest that population analyses of intestinal symbionts like these can provide insight into the distribution, movement and life histories of long-lived reef fish.

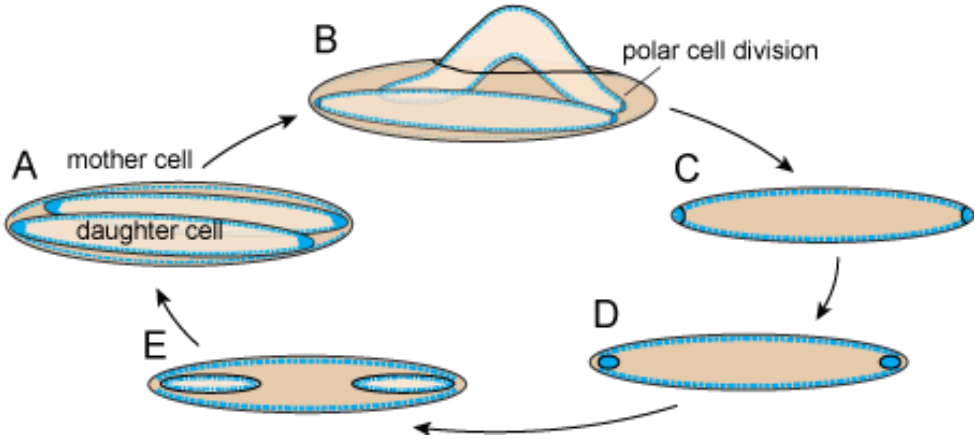


Figure 1. *Epulopiscium* sp. type B daily life cycle. These populations are maintained by an unusual reproductive strategy, which employs the formation of multiple intracellular offspring. Binary fission has never been observed in *Epulopiscium* sp. type B. (A) Accumulation of DNA at the poles of intracellular daughter cells marks the beginning of formation of the next generation (granddaughter cells). (B) Bipolar division often occurs just prior to emergence of daughter cells from their mother cell. Note the DNA in the mother cell is less pronounced at this stage and mother cells do not survive daughter-cell release. (C) Polar cells are (D) fully engulfed by their mother cell. (E) Internal offspring cells continue to grow until (A) they nearly fill the mother-cell cytoplasm. Figure modified from [20].

METHODS

Sample collection

Naso tonganus were collected by spear from island or outer barrier reef habitats around Lizard Island, Australia, at three different 10-year intervals: circa 1990, 2003, and 2013 (Fig. 2). Surgeonfish that regularly consume algae have a long, coiled intestinal tract (2). Other than the coiling pattern, the intestine of *N. tonganus* has no distinct morphological features. To collect samples containing *Epulopiscium* sp. type B, the *N. tonganus* intestinal tract was removed and uncoiled. The unraveled intestine was laid out in 4 equal-length segments, as previously described (7). In this scheme, the stomach is defined as segment I, and from anterior to posterior, the intestinal segments are referred to as segments II – V. Samples of *Epulopiscium* sp. type B were taken from segment IV, where large numbers of symbionts are located. Intestinal contents from each individual fish were fixed in 80% ethanol and stored at -20°C. Sample information, a description of *Epulopiscium* cells present, and demographic data for each host were recorded (Table 1).

Single-cell, whole-genome amplification and MLSA of *Epulopiscium* sp. type B

Briefly, individual *Epulopiscium* cells were manually collected from fixed intestinal contents and subjected to whole-genome amplification (WGA) using the REPLI-g Kit (Qiagen) for 16 hr at 30°C, following the manufacturer's protocol. Seven housekeeping genes (*dnaC, ftsZ, mreB, radA, recA, rpoB,* and *secA*) and the 16S rRNA gene were PCR amplified using primers and conditions described in Table S1. See Supplementary Information for details of single-cell processing, WGA, MLSA scheme design, and sequence analysis.

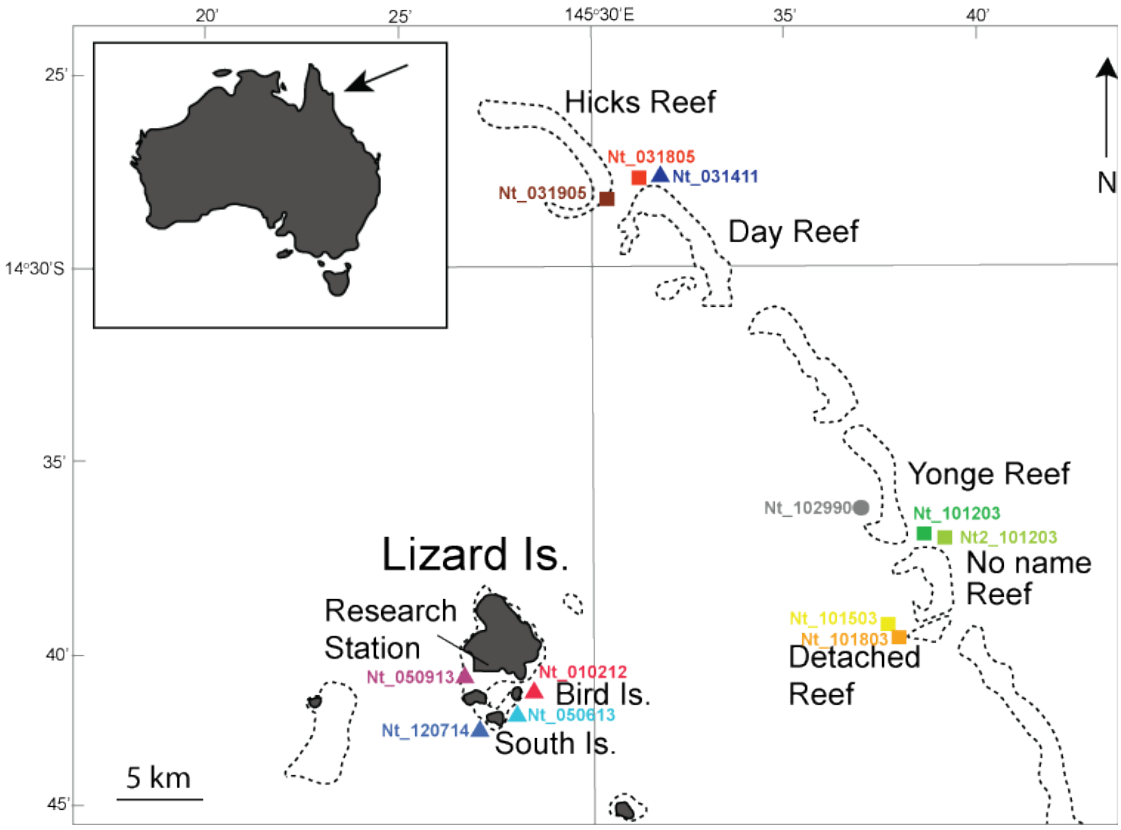


Figure 2. Collection sites for each *Naso tonganus* used in this study. Map of Lizard Island and nearby outer barrier reef. Arrow in the inset indicates location of the collection site off the northeast coast of Australia. Islands (including Lizard Is., Bird Is. and South Is.) are shaded dark grey while dashed lines outline reefs. Approximate sampling locations for each *N. tonganus* individual are indicated by a unique color and sample ID. Details are provided in Table 1. Sampling time is indicated by either a circle, square, or triangle for year intervals circa 1990, 2003 and 2013, respectively.

Table 1. *Naso tonganus* sampling data and corresponding symbiont descriptions.

Sample ID	*N. tonganus* collection data					*Epulopiscium* sp. type B cells[a]		
	Date	Time	Location[b]	Sex	SL[c] (mm)	Avg. mother cell length (µm)	No. of offspring	Relative offspring size[d]
Nt_102990	10/29/1990	14:10	Yonge Reef	n.d.	n.d.	191	2-3	0.73
Nt_101203	10/12/2003	9:45	No name Reef	f	400	173	2	p
Nt2_101203	10/12/2003	10:00	No name Reef	f	412	112	2	0.17
Nt_101503	10/15/2003	11:00	Detached Reef	n.d.	391	135	3	0.14
Nt_101803	10/18/2003	16:45	Detached Reef	f	441	124	2-3	0.76
Nt_031805	03/18/2005	15:50	North Day Reef	f	410	161	2	0.74
Nt_031905	03/19/2005	13:30	Hick's Reef	f	405	123	2-3	0.38
Nt_031411 (M100)	03/14/2011	9:26	North Day Reef	f	450	140	2	0.75
Nt_010212 (M263)	01/02/2012	15:10	Bird Island	im	346	125	2-4	0.47
Nt_050613 (M450)	05/06/2013	15:24	South Island	f	240	139	2	0.71
Nt_050913 (M485)	05/09/2013	16:50	Research Station Beach	f	431	155	2-5	0.13
Nt_120714 (M522)	12/07/2014	16:00	South Island	f	305	205	8-12	0.35

n.d. – not determined; f – female; im – immature.

[a] Morphotype B designation was confirmed by 16S rRNA gene sequence comparisons (≥ 99% identity and coverage).

[b] All fish were collected in proximity to Lizard Island, Australia.

[c] Standard length.

[d] Proportion of daughter-cell length compared to mother-cell length. "p" indicates offspring primordia, condensed DNA at the poles but no evidence of polar division.

DNA extraction and MLSA of *Naso tonganus*

Host DNA was extracted from fixed intestinal contents. Fish haplotypes were based on mitochondrial (*cox1* and *Cytb)* and nuclear genes (*ENC1*, *plagl2*, and *zic1).* See Supplementary Information for details of DNA extraction and sequence analysis. Primer sets and amplification conditions are described in Table S1. Haplotypes were reconstructed from the unphased sequence data using the phase option in DnaSP version 5.10 (29).

Phylogenetic analyses

Concatenated sequences were joined head-to-tail in-frame and trimmed for seven *Epulopiscium* loci and five *N. tonganus* loci (unphased). 16S rRNA gene sequences were not included because *Epulopiscium* has multiple rRNA operons (13). Bayesian and maximum likelihood (ML) phylogenies were constructed in MrBayes version 3.2.4 (30) and PhyML version 3.0 (31), respectively, each using the generalized Time Reversible nucleotide substitution model (32) plus invariant sites (I) and Γ rate heterogeneity. Markov chain Monte Carlo was run for one million generations, sampling every 1000 generations. The first 25% of trees were discarded as burnin. A total of 1000 bootstrap replicates were performed for ML phylogenies.

Nucleotide diversity, recombination and gene flow analyses

For both *Epulopiscium* and host datasets, statistics for single genes and populations were calculated using DnaSP (29). These tests included the number of polymorphic sites, nucleotide diversity (π) corrected using the Jukes Cantor method (33), haplotype diversity (Hd), and neutrality tests (Tajima's *D* (34), Fu and Li's *D** (35), and Fu's *F* (36)). Individual markers were analyzed for signatures of selection by calculating the ratios of nonsynonymous to synonymous substitutions (*dN/dS*) using the default NG86 nucleotide substitution model (37) in START2 (version 2) (38). Markers were further analyzed with PAML 4.9h (39), using multiple codon frequency

models as described in the Supplementary Information.

For the *Epulopiscium* sp. type B dataset, recombination breakpoints were identified using the Genetic Algorithm for Recombination Detection (GARD) (40) available through the Datamonkey webserver (41). Results were visualized using R (42) and package "ggplot2" (43). Population structure was modelled using STRUCTURE 2.3.4 (44) as described in Supplementary Information, to examine population subdivisions and presence of admixture. Admixture within and between subpopulations is indicative of recombination. Additionally, a pairwise homoplasy index test (Φ_W) (45) was performed using SplitsTree 4.13 (46) to detect recombination by examining the genealogical history of pairs of sites. This approach is useful for data with complex population structure and demographic histories, differentiating between population growth and recombination. Estimates for linkage disequilibrium were performed using the standardized index of association (I^S_A) (47) from START2.

To determine the level of gene flow between symbiont groups at different subdivisions, an analysis of molecular variance (AMOVA) (48) was implemented and tested for significance against 1000 permutations with Arlequin 3.5 (49). A Mantel test was performed to examine whether genetic distance correlated with geographic distance and those results were plotted using the R package "ade4" (50). Two distinct hierarchical subdivisions were tested independently: time intervals and habitat location (reef vs. island). Variability was assessed among *Epulopiscium* individuals within time intervals (Φ_{GT}), among *Epulopiscium* populations within each host (Φ_{ST}), and among populations within time intervals (Φ_{SG}). Multiple linear regressions were performed in R to examine whether π or Hd are correlated with demographic parameters listed in Table 1.

Codiversification analysis of *N. tonganus* and *Epulopiscium* sp. type B

The global signal of codiversification and contribution of individual host-

symbiont associations were analyzed using Parafit (51) and PACo (52), implemented in R packages "ape" (53) and "vegan" (54). The null hypothesis differs between the two tests; Parafit tests independent host and symbiont evolution whereas PACo explicitly tests independence of symbiont phylogeny on host phylogeny. The input for both analyses were aligned concatenated sequences from individual hosts and symbiont sequence types (sSTs), converted to distance matrices using the K80 model (55). The significance of both tests were established from 100,000 permutations. The *P*-value was determined for each host-symbiont pair by ParaFitLink1 tests. A tanglegram was generated with iTol version 4.2 (56).

RESULTS

MLSA of *Epulopiscium* sp. type B cells revealed high variability

This study examined symbiont diversity and distribution among and within individual hosts collected over the course of approximately 20 years (1990 to 2014) in a sampling area that covers discontinuous island and reef habitats within a 17 km radius circle (Fig. 2). Comparisons of *Epulopiscium* populations collected over time were used to improve the detection of population structure. A total of 113 individual *Epulopiscium* sp. type B cells were collected and analyzed from 12 different *N. tonganus* (~10 cells/fish) (Table S2). Cells were subjected to WGA and genes were PCR amplified from these DNA samples. The sequence of each PCR product was determined using the Sanger method. Amplified 16S rRNA gene sequences were 99-100% similar to published 16S rRNA gene sequences from *Epulopiscium* sp. type B (Fig. S1). This supported previous observations of low population diversity. However, using a higher resolution approach with additional markers (*dnaC, ftsZ, mreB, radA, recA, rpoB,* and *secA*), we observed high sequence variability and identified 88 symbiont sequence types (sSTs) (Fig. 3, Fig. S2). All housekeeping gene sequences

shared 99-100% sequence identity with genes in the *Epulopiscium* sp. type B draft genome (Table S3). Comparison of the concatenated sequences revealed 64 polymorphic sites across the 4005-base length alignment (Table S4). For the entire dataset, nucleotide diversity (π) was 0.00287 and was similar across time and space (ranging 0.00217-0.00287) (Table 2). Nucleotide diversity varied across individual markers (0.00543-0.00109), except *secA* had no polymorphisms (Table S4).

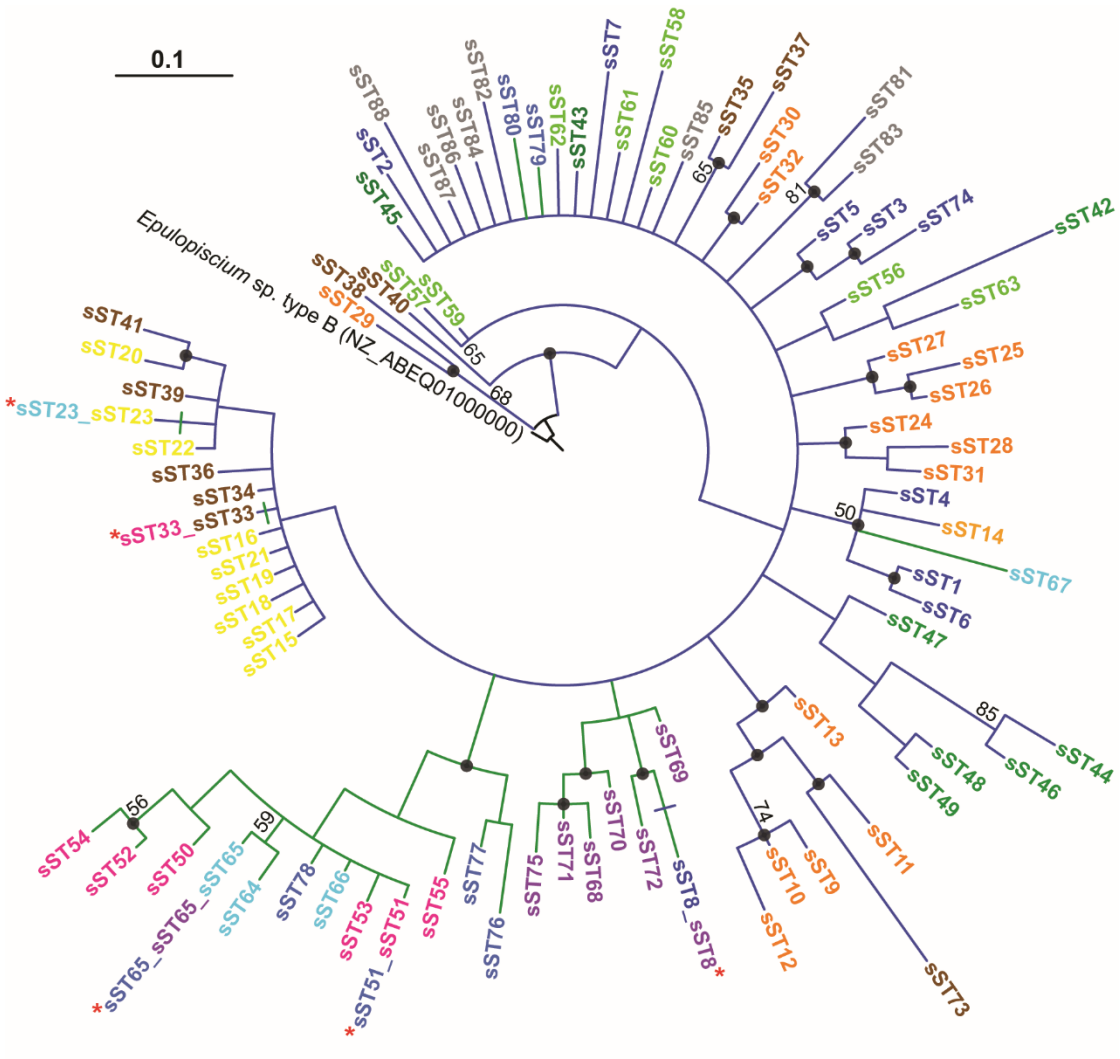


Figure 3. Bayesian phylogenetic tree of *Epulopiscium* sp. type B cells. Tree was constructed using concatenated sequences of seven housekeeping markers (*dnaC, ftsZ, mreB, radA, recA, rpoB, and secA*). Nodes with Bayesian posterior probabilities ≥ 0.70 are indicated by solid circles and nodes with ML bootstrap values ≥ 50% from 1,000 replicates are indicated. Each symbiont subtype (sST) is color coded by host source as in Fig. 2, and sSTs found in multiple hosts are indicated by red asterisks. Colored branches represent habitat type (green = island; blue = reef) and sSTs found in both island and outer barrier reef habitats are indicated with a crossed branch. Scale represents nucleotide changes per position.

Table 2. *Epulopiscium* sp. type B population parameters compared across time and habitats for concatenated MLSA sequences.

Populations	N^a	S^b	sST^c	Hd^d	π^e	Tajima's D^f	Fu an Li's D^{*g}	Fu's F^h	I^S_A [i]	r/m^j (95% CI)	ρ/Θ^k (95% CI)	*Mean* Φ_W [l]
Reef	75	61	65	0.995	0.00276	-0.422	-0.024	**-34.323****	**0.0804****	22.2 (0.161-162)	49.2 (0.334-377)	**0.4140****
Island	38	33	26	0.942	0.00217	0.257	0.503	**-9.314****	**0.2391****	51.1 (0.041-334)	118 (0.0795- 799)	**0.4558****
2003	57	52	49	0.993	0.00268	-0.230	-0.574	**-38.182****	**0.1098****	548.7 (0.066-3,468)	1,286 (0.1291-7,944)	**0.4880****
2013	47	37	33	0.960	0.00247	0.501	1.029	**-13.867****	**0.2516****	600.8 (1.619-5,176)	1,338 (3.6754-13,177)	**0.4091****
Total	113	64	88	0.991	0.00287	-0.206	0.070	**-86.602****	**0.1023****	5,344 (10.2-30,326)	12,506 (27.9-73,487)	**0.4246****
Total no *dnaC*	113	48	81	0.988	0.00239	-0.352	-0.446	**-88.585****	**0.8170****	417 (0.825-3,117)	1,440 (2.395-10,877)	**0.3649****

**Significant values bolded: $P < 0.05$.

[a]Number of individual *Epulopiscium* cells.

[b]Number of total polymorphic sites.

[c]Symbiont sequence type identifies a unique set of alleles for all seven loci.

[d]Haplotype diversity represents the uniqueness of a haplotype in the population.

[e]Nucleotide diversity corrected using the Jukes Cantor method [32]. Measures the degree of polymorphism per site within a population.

[f]Class I statistical test for neutral evolution based on frequency of polymorphic sites [33].

[g]Class I statistical test for neutrality based on intraspecies diversity [34].

[h]Class II statistical test against population growth and genetic hitchhiking [35].

[i]Standardized index of association [45] calculated using START2 [37]. Freely recombining populations are predicted to have independent assortment of alleles (linkage equilibrium) and contain an I^S_A value of zero. Whereas values significant from zero reflect clonal populations that harbor a high abundance of alleles with linked loci (linkage disequilibrium).

[j]Clonalframe calculations for the relative rate of recombination and mutation that contributes to the nucleotide substitutions observed [42].

[k]Clonalframe calculations for the ratio of absolute numbers of recombination (ρ) and mutation (Θ) events.

[l]Pairwise homoplasy index test [43] examines genealogic histories and measures whether pairs of sites were a result from direct inheritance of an ancestor through mutation rather than homoplasy (convergence). Populations with an excess of homoplasies are significantly different from zero and reflect a history of recombination. Calculations were performed using SplitsTree [44].

Neutrality tests predict either a recent symbiont population expansion or recombination

Next, neutral processes were explored to determine whether stochastic processes of dispersal and genetic drift explained the level of diversity observed in the symbiont population. Subpopulations within different habitats as well as the entire population sampled across space and time were examined; Class I neutrality tests (Tajima's *D* and Fu and Li's *D**) were not significant and the Class II neutrality test (Fu's *F*) was significantly negative (Table 2). Genes that likely contribute to the significant Fu's *F* were *dnaC*, *radA*, and *recA* (Table S4). Although a significantly negative Fu's *F* would suggest an excess number of alleles due to a recent population expansion or genetic hitchhiking, this test is strongly affected by recombination which may produce a false positive result (57). Therefore, the contribution of recombination was analyzed to determine which demographic parameters were influencing the diversity of the symbiont populations.

Host and location influence symbiont population structure

Deviations from the neutrality model prompted the exploration of parameters that might be influencing symbiont allele frequencies. Most sSTs were unique, but sST65 was more frequently encountered and observed in hosts collected at the Lizard Island sites (Fig. 3, Fig. S2). *Epulopiscium* populations showed high haplotype diversity, averaging 0.991 across all samples (Table 2). Generally, individual hosts harbored diverse symbiont populations (Table S2). However, host Nt_050613 had the least diverse population (Hd 0.667). No host demographic information (fish size, and location collected) correlated with low symbiont Hd.

Time and location were discernible despite the notable location bias in the samples used here, in which fish collected prior to 2011 were predominantly from the Outer Barrier Reef and after 2011 most were taken near Lizard Island (Fig. 2, Table

1). AMOVA indicated time intervals contributed ~8% of the variation ($P = 0.0499$), whereas habitat location (reef vs. island) contributed 16% ($P < 0.0001$) (Table S5). In both subdivisions, the majority of the contribution was found within populations ($P < 0.0001$). Thus, factors within hosts had the greatest influence on the diversity of symbiont populations. These may include host genotype, social behavior or dietary differences.

Despite the statistical noise contributed by variation within host populations, there was a significant correlation between geographic distance and pairwise F_{ST} (Mantel test: $R^2 = 0.136$, $P = 0.0014$) (Fig. S3). A Principle Coordinate Analysis (PCoA) of pairwise F_{ST} values further supported clustering based on location (Fig. 4). Island populations clustered tightly together with the exception of Nt_050913 from Research Station Beach, which grouped with Nt_031411 from North Day Reef. Other reef populations clustered together with the exception of the two populations from the southernmost outer reef collection location within our study area, Detached Reef. Both of these divergent symbiont samples were collected in the same location and year. These observations of population subdivisions and admixture were confirmed by simulations using STRUCTURE (see Supplementary Information) (Fig. S6). Altogether, these data suggest that symbiont diversity within an individual fish is dependent on host feeding and location.

Recombination contributes to diversity in *Epulopiscium* sp. type B populations

Symbiont population structure and a significantly negative Fu's *F* suggested symbiont gene flow between hosts. Therefore we investigated whether recombination or spontaneous mutation was the mechanism contributing to the variation observed. Multiple tests found evidence of recombination among the *Epulopiscium* sp. type B populations. GARD identified at least one recombination breakpoint at position 615 of the concatenated sequences (Fig. S4). Topological incongruence was supported by the

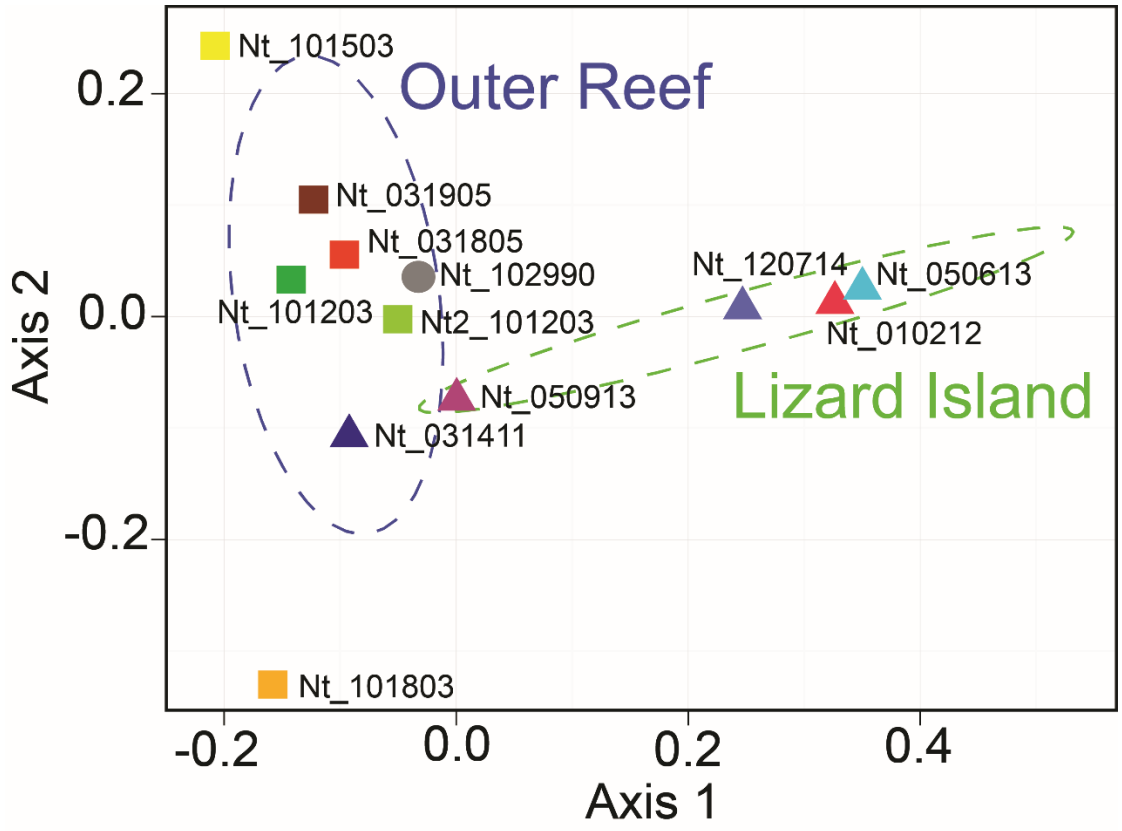


Figure 4. PCoA of the genetic differentiation (F_{ST}) of *Epulopiscium* sp. type B populations. The 95% confidence interval for each habitat type (island or outer barrier reef) is indicated by ellipses. Sampling time is indicated by either a circle, square, or triangle for year intervals circa 1990, 2003 and 2013, respectively.

KH test ($P < 0.001$) (Fig. S5). The homoplasy index test (Φ_W) had a significant level of convergence/recombination among the total population, within habitats, and even within certain hosts (Nt_101203, Nt_101803, Nt_031905, Nt_031411) (Table 2, Table S2). These data suggest that HGT among *Epulopiscium* sp. type B cells helps maintain allelic diversity in the symbiont population and may diminish clonal interference.

Preliminary MLSA of *Naso tonganus* suggests deviation from neutrality

To examine the possible contribution of host genetics to symbiont population structure, host allelic frequencies were characterized. Previously reported acanthurid markers and four new loci that have not been reported for *N. tonganus* were used

(Table S1). Five loci were successfully amplified and analyzed from 11 out of 12 fish used in this study (Table S6). Analysis of the concatenated sequences revealed 19 polymorphic sites across 4,125 bp, with π ranging from 0.00041-0.00211 per locus. All markers were under stabilizing or neutral selection. Neutrality tests for individual markers were not significant. However, Fu and Li's D^* were significantly positive (1.593, $P < 0.05$) across the length of the concatenated sequences, reflecting an excess of intermediate-frequency alleles which can result from population bottlenecks, structure and/or balancing selection.

Fourteen haplotypes were identified in which 7 individuals were homozygous (hap1-3, hap6, hap13-14) and 4 were heterozygous (Fig. S7). Since the heterozygous haplotypes clustered tightly within each individual, the multilocus phylogeny was constructed using unphased sequences and each host is referred herein per their sample ID (Fig. S8). Although the genealogy is not well resolved, two main clades are distinguishable. One clade contains 3 out of the 4 Lizard Island samples and one host collected from the outer barrier reef. The other clade contains mostly hosts from the outer barrier reef and the other Lizard Island associated host.

Codiversification analysis reveals symbiont dependence on host phylogeny

Significant global codiversification links between sSTs and individual hosts were detected from both Parafit (3.268×10^{-9}, $P = 0.0002$) and PACo ($m^2 = 8.644 \times 10^{-5}$, $P < 0.00001$). However, only a few significant ParaFit1 links ($P < 0.05$) contributed to the codiversification signal (Fig. 5). Yet, the significant PACo results indicate that the symbiont phylogeny is dependent on the host. The Procrustean superimposition plot confirmed the influence of some host individuals (Nt_050913, Nt_120714, Nt_050613, Nt_102990, and Nt2_101203) on the symbiont phylogeny (Fig. S9). Host Nt_050913 had the greatest number of significant links (9 links) with its symbionts. Including marginally significant links ($P < 0.1$), better resolved the

symbionts/host relationship. These data suggest that *Epulopiscium* sp. type B and their surgeonfish host share a facultative relationship where symbionts are more dependent on their host than vice versa. This model is further supported by our observations of apparently healthy adult *N. tonganus* which were collected within our study area but appeared to harbor no *Epulopiscium* sp. type B. Notably, sSTs that occur in multiple hosts (asterisks, Fig. 5) often had significant links with hosts that were more phylogenetically related and collected in proximal locations, thus suggesting that host genetics or habitat sharing influences symbiont populations. For example, sST65 occurred in three hosts and was significantly linked to Nt_120714 and Nt_050913 but marginally linked to Nt_050613. Host Nt_050613 was more distantly related to the sister pair Nt_120714 and Nt_050913 but all three were collected from Island locations. sST51 occurred in hosts Nt_120714 and Nt_010212, but had a significant link with only Nt_120714. Likewise, sST8 was significantly linked with host Nt_050913 and not Nt_031411.

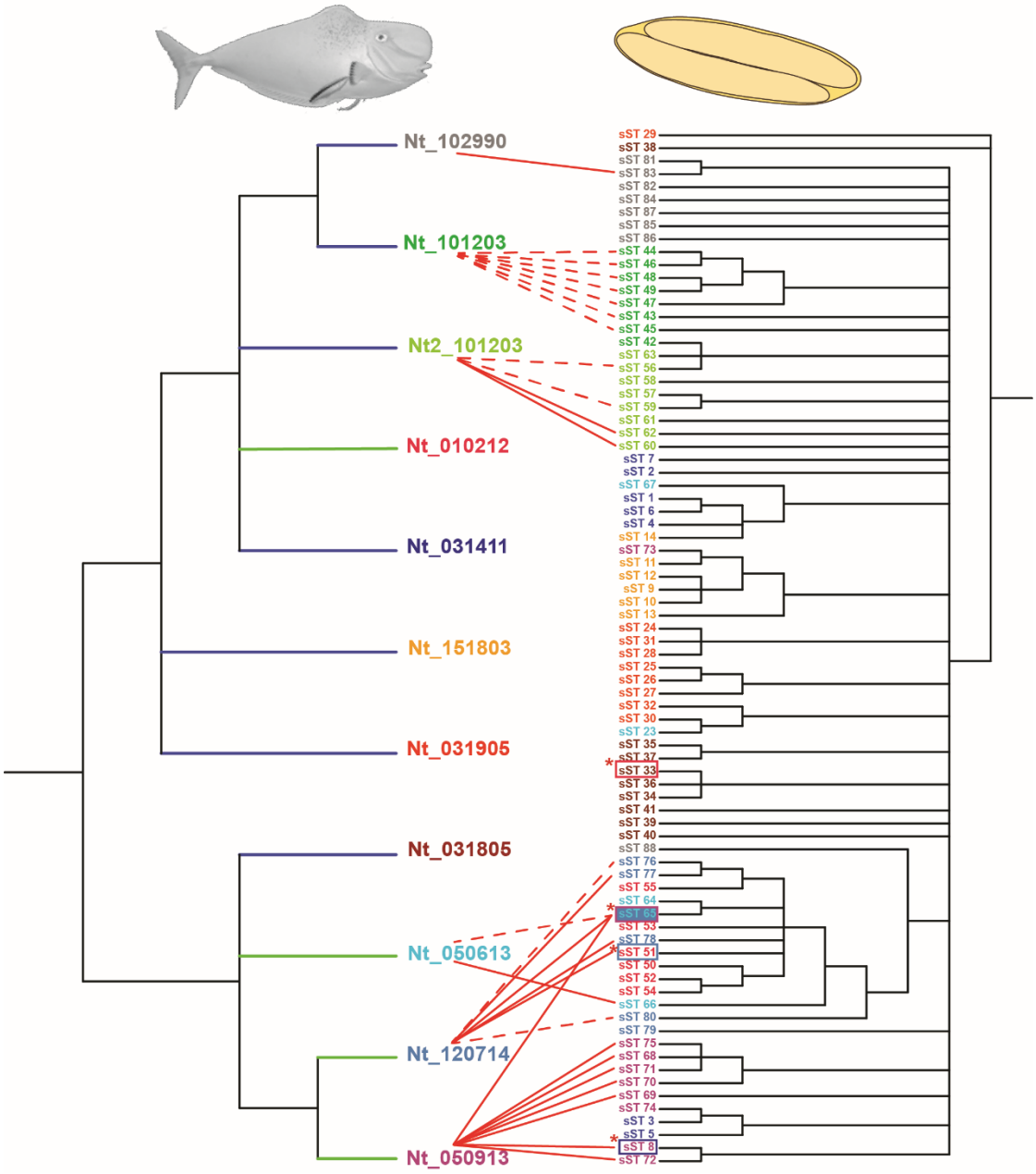


Figure 5. Tanglegram of host *N. tonganus* (left) and symbiont *Epulopiscium* sp. type B sSTs (right). Parafit Global test = 3.268×10^{-9}, $P = 0.0002$ (10,000 permutations). 85 host-sST links were detected and are indicated by matching colors. sSTs found in multiple hosts are indicated with an asterisk; an outlined box specifies two hosts while a shaded box specifies three hosts (e.g. sST65). Significant ParaFitLink1 tests are shown by red dashed lines ($P < 0.1$) and red solid lines ($P < 0.05$). Host collection habitats are indicated with green (island) or blue (outer barrier reef) cladogram branches.

DISCUSSION

The Lachnospiraceae are recognized as influential members of the gastrointestinal tract microbiota of terrestrial vertebrates and some marine vertebrates, including surgeonfish. One member of the Lachnospiraceae family, *Epulopiscium* sp. type B, can form a specific relationship with the surgeonfish *N. tonganus*. Moreover, these intestinal symbionts display an unusual daily reproductive cycle. Both traits likely impact the evolutionary trajectory of the bacteria. Using a fine scale population survey of this intestinal symbiont and its host, we explored their codiversification and identified factors that contribute to the evolution of *Epulopiscium* sp. type B. Remarkably, we discovered high genetic diversity among *Epulopiscium* sp. type B populations within individual fish as well as evidence of extensive symbiont genotype mixing between fish. Despite this, widespread linkage disequilibrium (LD) in the symbiont population and the identification of one host-associated subpopulation with low haplotype diversity support the hypothesis that transmission bottlenecks are occurring. These data suggest that recombination contributes to *Epulopiscium* sp. type B genetic diversity and compensates for deleterious effects imposed by its lifestyle. Furthermore, our analyses suggest that *Epulopiscium* sp. type B is not only dependent on its host for transmission, but horizontal transmission of the symbiont is also necessary for the incursion of new genetic material.

Evolution of horizontally acquired symbionts depends on host

The limited number of significant codiversification links and structure of symbiont populations support the model that *Epulopiscium* sp. type B cells are horizontally transmitted. Samples collected over the course of more than two decades were used in the study to try to improve the likelihood of detecting changes in symbiont population structure over time. Notably, acanthurids collected from the Great Barrier Reef can live 30 to 45 years (58). This suggests that sampling beyond

the life span of *N. tonganus* may be needed to illuminate time as a more significant contributor to symbiont population variation. However, the presence of identical alleles and similar population structures between fish collected at different time points implies that the introduction of new symbionts to an established population may occur throughout the life span of an individual host. Since *Epulopiscium* populations are easily lost when surgeonfish are brought into captivity, we assert that in the wild, dietary changes or stress may alter symbiont populations leading to sweeps or facilitating introgression.

Codiversification analyses suggest that host genetics contributes to *Epulopiscium* sp. type B population structure. There is evidence that host genetics refines the composition of microbiota associated with animals, including some fish (59, 60). Interactions between gut microbiota and the host immune system likely contribute to observed variation (61, 62). A study of the threespine stickleback (*Gasterosteus aculeatus*) found that the composition of gut microbiota depended more on host genotype than on any other transient environmental factors (63), suggesting that the fish host filters through countless environmental microorganisms to establish its gut microbiome.

Selection of symbionts may also arise through host ecology (e.g. diet) and/or composition of the resident microbiota. In herbivorous acanthurids, *Epulopiscium* spp. can be the most dominant taxa in the hindgut (2, 60). Compared to zooplanktivores, these fish have longer intestinal tracts, which contain higher levels of fermentation products (e.g. acetate) (7, 8). Although *N. tonganus* has been referred to as an herbivore (64, 65), it is generally considered an omnivore (9). Thus, individual feeding preferences or the availability of suitable food may be wide-ranging and may impose a strong selection for particular microbiota and compatible *Epulopiscium* strains. Evidence of codiversification of gut symbionts with their surgeonfish hosts suggests

that there is pressure to retain specific phylotypes (4, 17). Here, we have increased the resolution of codiversification and extend it to the population level for *Epulopiscium* sp. type B.

The phylogenetic diversity of the *Epulopiscium* sp. type B parallels gut symbiont profiles described for social bees and mammals (66–68). In these systems, communities tend to have low species richness, and specific lineages exhibit shallow fan-like branching patterns which suggests that hosts are inoculated with a few founder species that later diversify *in situ*. Diversification from founder species may enable niche partitioning among strains in a nutrient-rich gut environment. Recent studies in an experimental model using gut inoculations with an auxotrophic *E. coli* strain demonstrated that niche partitioning reduced clonal interference within a mouse host (69). *Epulopiscium* sp. type B likely plays a major role in host nutrition and our data suggest niche partitioning may also be occurring here. The functional significance of these diversified populations is worth further study.

Gut microbiota footprint reveals limited patterns of movement of *N. tonganus*

Some coral reef fish, including some acanthurids, exhibit high genetic connectivity across large oceanic distances (70, 71). This may involve dispersal of pelagic larvae over distances exceeding 10,000 km (72), whereas small ranging fish exhibit spatial structure influenced by suitable reef habitats and seascape discontinuity (73). Recent studies of large, coral reef associated fish species have revealed high levels of larval retention to parental habitats (74), suggesting that larval dispersal is not as extensive as previously proposed. The small sample size and markers used in this study were insufficient to address genetic connectivity. However, the divergence of host-specific symbiont populations suggests host foraging and movement patterns are governed in part by seascape discontinuity. This observation is further supported by the host's positive Fu and Li's D^* that suggest these fish may have experienced a

population bottleneck or more likely have undergone population subdivision.

Spatial structuring at our study site was surprising given that the large *N. tonganus* would be expected to easily traverse this distance. However, traveling from Lizard Island to the outer reef would require a fish to cross 25 km of open water up to 50 m deep. Surgeonfish may be reluctant to venture far from a reef due to increased vulnerability to predators. With a few notable exceptions, symbionts from fish collected near Lizard Island were more closely related to one another than to symbionts of hosts from outer barrier reef, and vice versa. For example, host Nt_050913 is genetically more similar to Nt_120714 from Bird Island but its *Epulopiscium* population is more closely related to symbionts of Nt_031411 from North Day reef. Another form of symbiont population discontinuity was observed at the southern end of the outer barrier reef within our study area. Both symbiont populations from fish collected near Detached Reef are unique and suggest that these fish came from a more distant location, perhaps south of our sampling area. Clearly more detailed surveys are needed to test the hypothesis, but data collected in this study indicate that *Epulopiscium* sp. type B populations provide a record of movement of individual fish among groups of fish associated with the Great Barrier Reef. These suggested patterns of movement are consistent with the 'commuting' and 'foraying' patterns observed in *Naso unicornis* studies using radio telemetry-based tracking (75, 76).

Genetic exchange within the *Epulopiscium* sp. type B populations

The results provided here support the hypothesis that *Epulopiscium* sp. type B populations depend on environmentally acquired alleles to conserve genetic diversity. Based on the draft genome, *Epulopiscium* sp. type B has the recombination and DNA maintenance genes needed to support this mechanism (11).

Adaptive rate studies highlight the advantage that recombination has over

spontaneous mutations, especially in systems where population structures exist. Experimental studies of *E. coli* and *Saccharomyces cerevisiae* demonstrated that beneficial mutations became fixed sooner in strains with high recombination rates than in strains with high mutation rates (77, 78). Model simulations have shown that higher rates of HGT in small, structured populations made these populations more resistant to Muller's ratchet than larger, mixed populations (79). This suggests that "cross-referencing" between subdivided populations, facilitated by HGT, enhance genetic diversity. The gut ecosystem provides a structural framework in which recombination could be highly impactful as we observed in *Epulopiscium* sp. type B populations.

Despite strong evidence for recombination, the *Epulopiscium* populations studied exhibited LD, suggesting that there is not enough recombination occurring to observe random assortment of markers. Previous reports estimated that at least a 20-fold relative contribution of recombination per point mutation (*r/m*) is needed for loci to assort independently (47, 80). However, Spratt *et al.* (81) cautions against using LD as a proxy for relative recombination rates, recognizing that highly recombinant bacterial populations may still appear to be in LD. We suggest that LD may be common in naturally competent populations of bacteria. The ability to take up DNA from the environment and stably integrate that DNA into the genome is widespread in the bacteria and archaea (82). Congression, the phenomenon by which competent cells are co-transformed with unlinked DNA molecules at high frequencies (83, 84), has been used by geneticists for decades to introduce specific genetic changes without the need for selection of both markers. However, the frequency at which multiple unlinked pieces of DNA are incorporated in a single competent cell has only recently been analyzed systematically (85). For both *Vibrio cholerae* and *Streptococcus mutans*, a single cell can take up two unlinked markers at surprisingly high co-transformation frequencies of 50 – 60%. This tendency for some members of a

naturally competent population to be transformed by numerous unlinked, unselected genes would contribute to LD in natural populations.

The detection of at least one recombination site, statistical support of topology incongruence, and admixture within and between subpopulations provides additional support for recombination *Epulopiscium* sp. type B. Evidence of a transmission bottleneck (low *Epulopiscium* Hd in host Nt_050613) suggests that the detected recombination events were relatively recent in *Epulopiscium* evolutionary history. Even in the obligate symbiosis between the bivalve *Solemya velum* and its bacterial gill endosymbionts, mixed infections and recombination occurs at a high enough frequency to maintain symbiont diversity (25). This further suggests that symbiont allele frequencies reflect a dynamic state in which populations may not reach equilibrium. Therefore, we speculate that microbial populations in gut systems might be predominantly in a state of flux.

The extreme polyploidy of *Epulopiscium* sp. type B may be confounding the typically observed relationship between LD and high recombination rate as well. Intracellular genetic diversity appears to be low, suggesting a strong pressure for gene conversion. Genome redundancy may mask the effects of deleterious mutations by purifying genes through gene conversion as observed in asexual amoeba (86). High conversion rates may also limit genetic diversity and thus contribute to LD through clonal interference.

Model for the genetic inheritance of externally acquired DNA in *Epulopiscium* sp. type B

Multiple Firmicutes species coordinate DNA release with competence, either by regulating cell lysis (87, 88) or secretion of DNA (89). The circadian cell cycle of *Epulopiscium* sp. type B provides an opportunity to coordinate competence induction with the daily release of genomic DNA from the population. This scenario would

allow a newly independent offspring cell to take up and incorporate DNA from both its own mother cell and others in the population (Fig. 6). We hypothesize that some somatic genomic DNA is released when mother cells lyse (as in Fig. 1 stage B). If competence complexes are located near the poles of emerging daughter cells, where the next generation (granddaughters) have been initiated but are not yet fully engulfed (see Fig. 1 stages B & C), extracellular DNA could be incorporated into pole-associated chromosomes, thus increasing the likelihood of vertical transmission to future generations. Recombination would likely take place during replication in the offspring (Fig. 1 stage D).

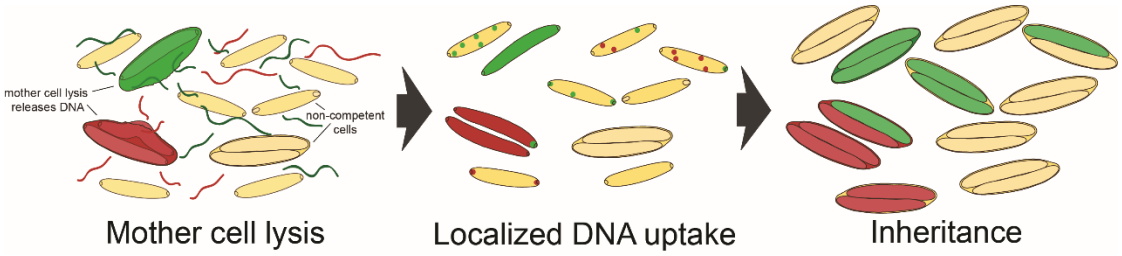


Figure 6. Model for coordinated mother-cell lysis and offspring competence induction in *Epulopiscium* sp. type B. Within each host, *Epulopiscium* sp. populations are synchronized with respect to development. The nearly simultaneous lysis of mother cells could provide a diverse pool of genomic DNA available for uptake by competent cells. We hypothesize that daughter cells, emerging at the time of mother-cell lysis, are competent. The newly released cells would be in the earliest stages of offspring development. Some may have divided asymmetrically but not yet engulfed the polar cells, others may be more advanced in their development. Those cells at later stages of development, after polar-cell engulfment is complete, would have physical barriers to the uptake of DNA that could be inherited. We suggest that the uptake of DNA at the poles of a newly emerged *Epulopiscium* cells would increase the chances of inheritance of DNA acquired by HGT whereas uptake away from polar cells would not be inherited. Transformation of cells by the uptake of DNA from a different sequence type is indicated by color changes.

CONCLUSIONS

The size and extreme polyploidy of *Epulopiscium* sp. type B make it an ideal model for using the single-cell genome-amplification approaches which facilitated this population study. We found that population bottlenecks imposed by unusual life history strategies, which are closely tied to maintaining a symbiotic association, can be overcome by simple changes to widely available mechanisms: increasing ploidy, and allele exchange using HGT and homologous recombination. There is a growing appreciation for the impact of polyploidy and recombination on the evolution of bacterial populations (90). Given the broad range of bacterial life histories of many gut microbes (including non-endospore-forming Lachnospiraceae) (91), we suggest that these populations may be under similar pressures as we have observed for *Epulopiscium* sp. type B. Furthermore, the close association of microbes in a densely populated gut ecosystem is ideal for HGT-based mechanisms to develop diverse populations.

ACKNOWLEDGEMENTS

We thank Rick Harrison and Lynn Johnson for their advice about study design and statistical analysis. We also thank Amy Dapper, Xavier Didelot, and Melanie Smee for advice on our population analyses. We thank Zanah Francis for help screening samples and David Sannino for insightful discussions. We are grateful to Daniel Buckley for the use of his server to run analyses. Fish were collected under the James Cook University Ethics approvals A503 and A1641 and with approval of the Great Barrier Reef Marine Park Authority permits G01-356 and G10-33239.1. This study was supported by National Science Foundation grants MCB1244378 and IOS1354911.

DATA ACCESSIBILITY

Sequences obtained in this study have been deposited in GenBank under Accession No. MH259595-MH259695 (16S rRNA genes), MH268406-MH269196 (*Epulopiscium* sp. type B MLSA genes), and MH282951-MH283005 (*N. tonganus* MLSA genes).

REFERENCES

1. Fishelson L, Montgomery WL, Myrberg Jr. A (1985) A unique symbiosis in the gut of tropical herbivorous surgeonfish (Acanthuridae: Teleostei) from the Red Sea. *Science* 229(4708):49–51.

2. Clements KD, Sutton DC, Choat JH (1989) Occurrence and characteristics of unusual protistan symbionts from surgeonfishes (Acanthuridae) of the Great Barrier Reef, Australia. *Mar Biol* 102(3):403–412.

3. Angert ER (2005) Alternatives to binary fission in bacteria. *Nat Rev Microbiol* 3(3):214–224.

4. Flint JF, Drzymalski D, Montgomery WL, Southam G, Angert ER (2005) Nocturnal production of endospores in natural populations of *Epulopiscium*-like surgeonfish symbionts. *J Bacteriol* 187(21):7460–70.

5. Meehan CJ, Beiko RG (2014) A phylogenomic view of ecological specialization in the Lachnospiraceae, a family of digestive tract-associated bacteria. *Genome Biol Evol* 6(3):703–13.

6. Thompson CL, Vier R, Mikaelyan A, Wienemann T, Brune A (2012) '*Candidatus* Arthromitus' revised: segmented filamentous bacteria in arthropod guts are members of Lachnospiraceae. *Environ Microbiol* 14(6):1454–1465.

7. Clements KD, Choat JH (1995) Fermentation in tropical marine herbivorous fishes. *Physiol Zool* 68(3):355–378.

8. Choat JH, Clements KD, Robbins W. (2002) The trophic status of herbivorous fishes on coral reefs I. Dietary analysis. *Mar Biol* 140(3):613–623.

9. Choat JH, Robbins WD, Clements KD (2004) The trophic status of herbivorous fishes on coral reefs: II. Food processing modes and trophodynamics. *Mar Biol* 145(3):445–454.

10. Ward RJ, Clements KD, Choat JH, Angert ER (2009) Cytology of terminally differentiated *Epulopiscium* mother cells. *DNA Cell Biol* 28(2):57–64.

11. Hutchison E, et al. (2018) Developmental stage influences chromosome segregation patterns and arrangement in the extremely polyploid, giant bacterium *Epulopiscium* sp. type B. *Mol Microbiol* 107(1):68–80.

12. Angert ER, Clements KD (2004) Initiation of intracellular offspring in *Epulopiscium*. *Mol Microbiol* 51(3):827–35.

13. Mendell JE, Clements KD, Choat JH, Angert ER (2008) Extreme polyploidy in a large bacterium. *Proc Natl Acad Sci U S A* 105(18):6730–6734.

14. Miller DA, Suen G, Clements KD, Angert ER (2012) The genomic basis for the evolution of a novel form of cellular reproduction in the bacterium *Epulopiscium*. *BMC Genomics* 13:265.

15. Angert ER (2006) The enigmatic cytoarchitecture of Epulopiscium spp. *Microbiology Monographs: Complex Intracelullar Structures in Prokaryotes*, ed Shively JM (Springer-Verlag, Berlin Heidelberg), pp 285–301. vol 2.

16. Angert ER, Clements KD, Pace NR (1993) The largest bacterium. *Nature* 362(6417):239–41.

17. Miyake S, Ngugi DK, Stingl U (2016) Phylogenetic diversity, distribution, and cophylogeny of giant bacteria (*Epulopiscium*) with their surgeonfish hosts in the Red Sea. *Front Microbiol* 7:285.

18. Collins MD, et al. (1994) The phylogeny of the genus *Clostridium*: proposal of five new genera and eleven new species combinations. 44(4):812–826.

19. Angert ER, Losick RM (1998) Propagation by sporulation in the guinea pig symbiont *Metabacterium polyspora*. *Proc Natl Acad Sci U S A* 95(17):10218–23.

20. Angert ER (2012) DNA replication and genomic architecture of very large bacteria. *Annu Rev Microbiol* 66(1):197–212.

21. Randall JE (2001) *Surgeonfishes of the world* (Mutual Publishing, Honolulu).

22. Eble J a., Rocha L a., Craig MT, Bowen BW (2011) Not all larvae stay close to home: insights into marine population connectivity with a focus on the brown surgeonfish (*Acanthurus nigrofuscus*). *J Mar Biol* 2011:1–12.

23. Clements KD (1997) Fermentation and gastrointestinal microorganisms in fishes. *Gastrointestinal Microbiology*, eds Mackie RI, White BA (Springer US, Boston, MA), pp 156–198. Volume 1.

24. Wernegreen JJ (2002) Genome evolution in bacterial endosymbionts of insects. *Nat Rev Genet* 3(11):850–861.

25. Russell SL, Corbett-Detig RB, Cavanaugh CM (2017) Mixed transmission modes and dynamic genome evolution in an obligate animal–bacterial symbiosis. *ISME J* 11(6):1359–1371.

26. Bright M, Bulgheresi S (2010) A complex journey: transmission of microbial symbionts. *Nat Rev Microbiol* 8(3):218–30.

27. Hutchison EA, Miller DA, Angert ER (2014) Sporulation in bacteria: beyond the standard model. *Microbiol Spectr* 2(5):87–102.

28. Vos M (2009) Why do bacteria engage in homologous recombination? *Trends Microbiol* 17(6):226–32.

29. Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25(11):1451–2.

30. Ronquist F, et al. (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* 61(3):539–42.

31. Guindon S, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59(3):307–21.

32. Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA Sequences. *Am Math Soc Lect Math Life Sci* 17:57–86.

33. Jukes T, Cantor C (1969) Evolution of protein molecules. *Mammalian Protein Metabolism*, ed Munro H (Academic Press, New York), pp 21–132.

34. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585–595.

35. Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. *Genetics* 133(3):693–709.

36. Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147(2):915–25.

37. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 3(5):418–26.

38. Jolley KA, Feil EJ, Chan MS, Maiden MC (2001) Sequence type analysis and recombinational tests (START). *Bioinformatics* 17(12):1230–1.

39. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24(8):1586–1591.

40. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW (2006) Automated phylogenetic detection of recombination using a genetic algorithm. *Mol Biol Evol* 23(10):1891–901.

41. Delport W, Poon AFY, Frost SDW, Kosakovsky Pond SL (2010) Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* 26(19):2455–7.

42. R Core Team (2014) R: A language and environment for statistical computing. Available at: http://www.r-project.org [Accessed November 2, 2015].

43. Wickham H (2009) *Ggplot2 : elegant graphics for data analysis* (Springer-Verlag, New York).

44. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure

using multilocus genotype data. *Genetics* 155(2):945–59.

45. Bruen TC, Philippe H, Bryant D (2006) A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172(4):2665–81.

46. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23(2):254–67.

47. Smith JM, Smith NH, O'Rourke M, Spratt BG (1993) How clonal are bacteria? *Proc Natl Acad Sci* 90(10):4384–4388.

48. Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131(2):479–491.

49. Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* 10(3):564–7.

50. Dray S, Dufour A-B (2007) The ade4 package: implementing the duality diagram for ecologists. *J Stat Softw* 22(4):1–20.

51. Legendre P, Desdevises Y, Bazin E (2002) A statistical test for host-parasite coevolution. *Syst Biol* 51(2):217–34.

52. Balbuena JA, Míguez-Lozano R, Blasco-Costa I (2013) PACo: a novel procrustes application to cophylogenetic analysis. *PLoS One* 8(4):e61048.

53. Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20(2):289–290.

54. Oksanen J, et al. (2016) vegan: Community Ecology Package. R package version 2.3-5. Available at: http://cran.r-project.org/package=vegan.

55. Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 16(2):111–120.

56. Letunic I, Bork P (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res*. doi:10.1093/nar/gkw290.

57. Ramirez-Soriano A, Ramos-Onsins SE, Rozas J, Calafell F, Navarro A (2008) Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. *Genetics* 179(1):555–567.

58. Choat JH, Axe LM (1996) Growth and longevity in acanthurid fishes: an analysis of otolith increments. *Mar Ecol Prog Ser* 134:15–26.

59. McKnite AM, et al. (2012) Murine gut microbiota is defined by host genetics and modulates variation of metabolic traits. *PLoS One* 7(6):e39191.

60. Miyake S, Ngugi DK, Stingl U (2015) Diet strongly influences the gut microbiota of surgeonfishes. *Mol Ecol* 24(3):656–72.

61. Hooper L V, Littman DR, Macpherson AJ (2012) Interactions between the microbiota and the immune system. *Science* 336(6086):1268–73.

62. Gómez GD, Balcázar JL (2008) A review on the interactions between gut microbiota and innate immunity of fish. *FEMS Immunol Med Microbiol* 52(2):145–54.

63. Smith CCR, Snowberg LK, Gregory Caporaso J, Knight R, Bolnick DI (2015) Dietary input of microbes and host genetic variation shape among-population differences in stickleback gut microbiota. *ISME J* 9(11):2515–26.

64. Sullam KE, et al. (2012) Environmental and ecological factors that shape the gut bacterial communities of fish: a meta-analysis. *Mol Ecol* 21(13):3363–78.

65. Wong S, Rawls JF (2012) Intestinal microbiota composition in fishes is influenced by host ecology and environment. *Mol Ecol* 21(13):3100–2.

66. Kwong WK, Moran NA (2016) Gut microbial communities of social bees. *Nat Rev Microbiol* 14(6):374–384.

67. Engel P, Stepanauskas R, Moran NA (2014) Hidden diversity in honey bee gut symbionts detected by single-cell genomics. *PLoS Genet* 10(9):e1004596.

68. Ley RE, Peterson DA, Gordon JI (2006) Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* 124(4):837–48.

69. Sousa A, et al. (2017) Recurrent reverse evolution maintains polymorphism after strong bottlenecks in commensal gut bacteria. *Mol Biol Evol* 34(11):2879–2892.

70. Horne JB, van Herwerden L, Choat JH, Robertson DR (2008) High population connectivity across the Indo-Pacific: congruent lack of phylogeographic structure in three reef fish congeners. *Mol Phylogenet Evol* 49(2):629–38.

71. Klanten OS, Choat JH, van Herwerden L (2007) Extreme genetic diversity and temporal rather than spatial partitioning in a widely distributed coral reef fish. *Mar Biol* 150(4):659–670.

72. Lester SE, Ruttenberg BI (2005) The relationship between pelagic larval duration and range size in tropical reef fishes: a synthetic analysis. *Proc Biol Sci* 272(1563):585–91.

73. D'Aloia CC, Bogdanowicz SM, Harrison RG, Buston PM (2014) Seascape continuity plays an important role in determining patterns of spatial genetic structure in a coral reef fish. *Mol Ecol* 23(12):2902–13.

74. Harrison HB, et al. (2012) Larval export from marine reserves and the recruitment benefit for fish and fisheries. *Curr Biol* 22(11):1023–1028.

75. Meyer CG, Holland KN (2005) Movement patterns, home range size and habitat utilization of the bluespine unicornfish, *Naso unicornis* (Acanthuridae) in a Hawaiian marine reserve. *Environ Biol Fishes* 73(2):201–210.

76. Hardman E, Green JM, Sabrina Desiré M, Perrine S (2010) Movement of sonically tagged bluespine unicornfish, *Naso unicornis*, in relation to marine reserve boundaries in Rodrigues, western Indian Ocean. *Aquat Conserv Mar Freshw Ecosyst* 20(3):357–361.

77. Chaguza C, Cornick JE, Everett DB (2015) Mechanisms and impact of genetic recombination in the evolution of *Streptococcus pneumoniae*. *Comput Struct Biotechnol J* 13:241–7.

78. Cooper TF (2007) Recombination speeds adaptation by reducing competition between beneficial mutations in populations of *Escherichia coli*. *PLoS Biol* 5(9):e225.

79. Takeuchi N, Kaneko K, Koonin E V. (2014) Horizontal gene transfer can rescue prokaryotes from Muller's ratchet: benefit of DNA from dead cells and population subdivision. *G3* 4(2):325–339.

80. Hudson RR (1994) Analytical results concerning linkage disequilibrium in models with genetic transformation and conjugation. *J Evol Biol* 7(5):535–548.

81. Spratt BG, Hanage WP, Feil EJ (2001) The relative contributions of recombination and point mutation to the diversification of bacterial clones. *Curr Opin Microbiol* 4(5):602–606.

82. Lorenz MG, Wackernagel W (1994) Bacterial gene transfer by natural genetic transformation in the environment. *Microbiol Rev* 58(3):563–602.

83. Erickson RJ, Copeland JC (1973) Congression of unlinked markers and genetic mapping in the transformation of *Bacillus subtilis* 168. *Genetics* 73(1):13–21.

84. Mell JC, Lee JY, Firme M, Sinha S, Redfield RJ (2014) Extensive cotransformation of natural variation into chromosomes of naturally competent *Haemophilus influenzae*. *G3* 4(4):717–731.

85. Dalia AB, McDonough E, Camilli A (2014) Multiplex genome editing by natural transformation. *Proc Natl Acad Sci* 111(24):8937–8942.

86. Maciver SK (2016) Asexual amoebae escape Muller's ratchet through polyploidy. *Trends Parasitol* 32(11):855–862.

87. Steinmoen H, Knutsen E, Håvarstein LS (2002) Induction of natural competence in *Streptococcus pneumoniae* triggers lysis and DNA release from a subfraction of the cell population. *Proc Natl Acad Sci U S A* 99(11):7681–6.

88. Steinmoen H, Teigen A, Håvarstein LS (2003) Competence-induced cells of *Streptococcus pneumoniae* lyse competence-deficient cells of the same strain during cocultivation. *J Bacteriol* 185(24):7176–83.

89. Zafra O, Lamprecht-Grandío M, de Figueras CG, González-Pastor JE (2012) Extracellular DNA release by undomesticated *Bacillus subtilis* is regulated by early competence. *PLoS One* 7(11):e48716.

90. Soppa J (2017) Polyploidy and community structure. *Nat Microbiol* 2(2):16261.

91. Browne HP, et al. (2016) Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature* 533(7604):543–546.

CHAPTER 2

NICHE DIFFERENTIATION BETWEEN TWO CO-OCCURRING GIANT BACTERIAL INTESTINAL SYMBIONTS OF SURGEONFISH *ACANTHURUS LINEATUS*[2]

ABSTRACT

The relationships that gut microbes have with each other and their hosts are complex and dynamic. *Epulopiscium* spp. and related bacteria, known as 'epulos', undergo a unique reproductive life cycle that is temporally synchronized within wild-caught surgeonfish. To understand the interactions between two co-occurring epulos we explored the distribution of distinct clades (A1 and A2) within three host species (*Acanthurus nigrofuscus*, *Acanthurus lineatus*, and *Acanthurus triostegus*) collected from the Great Barrier Reef. The two epulo clades were found in the three fish species sampled and were not phylogenetically differentiated between fish, indicative of symbiont transmission between host species. However, morphological characteristics of the epulos differed significantly between fish species, suggesting that the host environment may be influencing symbiont morphology. Single-cell genomes from the co-occurring symbionts of *A. lineatus* represent new genera and species: "*Ca.* Epulopiscium bomborense" clade A2 and "*Ca.* Industriepulo choatii" clade A1. Metabolic reconstructions revealed that the symbionts have distinct physiologies. A2 is predicted to degrade a wide range of complex polysaccharides found in red and green algae, providing acetate for the host. A1 has the potential to take up simple sugars and carboxylate intermediates to fuel a robust energy-generating and redox-balancing physique to maintain its large size. Transcriptomic profiles of natural

[2]David R. Sannino provided insight on *Epulopiscium viviparus* genome. Charles Pepe-Ranney provided bioinformatics expertise and collaborated on genome assembly pipeline.

symbiont populations provided further evidence for time-dependent synchrony of metabolic processes. A switch from active metabolism to a solventogenic/fasting state correlated with the diurnal feeding pattern of the host. Altogether, these data provide insight on the evolution of large bacteria and their ecological significance in animal nutrition.

INTRODUCTION

Host-microbiome studies have improved our understanding of the major taxa influencing animal health. For many vertebrates, gut microbiome diversity can be explained by factors such as host diet and phylogenetic history (1). Diet is a predominant driver for the selection of functional guilds of bacteria, particularly with herbivores. In humans, Bacteroidetes and Firmicutes dominate the community and are major contributors to polysaccharide degradation of recalcitrant cellulose and mucin into beneficial fermentative products (such as short chain fatty acids, SCFAs)(2). Comparing phylogenetic histories between host and microbiota may provide insight into how microbiota are acquired by the host. Strong cophylogeny posits a high frequency of vertical transmission (parent to offspring) of intestinal symbionts versus horizontal acquisition from environmental sources. Routes of vertical transmission can occur during the birthing process (3), or feeding behaviors such as coprophagy (4, 5). Trends of cophylogeny of particular vertebrate taxa are observed most often in mammals whereas non-mammalian hosts show mixed modes of transmission (vertical and horizontal)(1, 6). However, the interplay between microbial community members within their hosts and their effects on the ecology and evolution of these host-microbiome partnerships are not well understood.

The direction of human microbiome research has begun to pivot away from generating taxonomic surveys to understanding ecological and evolutionary dynamics

that shape microbiome communities (7). Commensal bacteria play a major role in preserving gut health as the mediators of community dynamics. They can directly compete with opportunistic pathogens for nutrients (8) or provide microbiota-nourishing immunity (9) thereby establishing balanced communities that resist pathogen colonization. Members of one of the most abundant orders of gut commensals, the Bacteroidales, degrade complex polysaccharides extracellularly and in the process feed other species. In exchange, other *Bacteroidetes* spp. detoxify the environment and provide additional growth factors to maintain a stable environment (10). Therefore, uncovering the principles of competition and cooperation would improve our understanding of how the microbiota influence dynamism and stability among the gut community of the host. This ecological framework will improve treatment of gut microbiota-related diseases/developmental disorders and provide insights for more effective conservation efforts of protected wildlife.

Epulopiscium spp. and related bacteria (herein after referred to as epulos), provide an exceptional model to study the ecological and evolutionary forces shaping the gut bacterial community of herbivorous marine fish. Epulos dominate the gut communities of herbivorous and detritivorous surgeonfish from coral-reef habitats in the Pacific Ocean and Red Sea (11, 12). Population surveys suggest that specific epulo clades represent functional guilds based on host feeding preferences (13, 14); several surgeonfish species that predominantly feed on red and green algae harbor epulo clades A and B whereas surgeonfish that feed on brown algae harbor epulo clades C and J. These morphologies correlate with distinct phylogenetic lineages that can be distinguished using 16S rRNA gene markers (15, 16). Further, epulos display disparate reproductive strategies that are hypothesized to influence transmission between hosts (16, 17). Reproductive strategies vary widely and may include binary fission, endospore formation, and the formation of multiple intracellular offspring.

Cell development is often synchronized within individual fish (16, 18, 19) and follows a predictable daily cycle.

Metabolic reconstructions of the model type symbiont "*Ca.* Epulopiscium viviparus clade B", suggest that epulos are nutritional symbionts that degrade complex polysaccharides from the host diet. Metabolism and energy of "*Ca.* E. viviparus" is fermentation-based but these cells appear to have the capacity of form and use a sodium motive force (SMF) to generate ATP (Sannino, unpublished). Although most epulos observed to date occur as mixed populations, "*Ca.* E. viviparus" are often exclusively found in their host *Naso tonganus* (15, 20). The association of type B symbiont populations with their host is only obligate for the symbiont, and symbiont populations are diverse suggesting that this host environment is conducive for niche differentiation (17).

To further explore the functional roles and potential dynamics between gut community members, we compared the single-cell genomes and transcriptional profiles from two co-occurring epulos (A1 and A2) found in *Acanthurus lineatus*. Both A1 and A2 morphotypes are similar in that they are large, cigar shaped and reproduce by intracellular offspring formation (Fig. 1), reminiscent of "*Ca.* E. viviparus"(15, 20). Yet, the A1 and A2 clades are likely different genera sharing only 91% 16S rRNA gene sequence identity. Clade A2 is phylogenetically similar to "*Ca.* E. viviparus" (97% 16S rRNA gene sequence identity) while clade A1 forms a more distant group. The co-occurring symbionts have been observed in acanthurid species that feed predominantly on red algae in both the Red Sea and the Great Barrier Reef (GBR): *A. nigrofuscus* (Red Sea and GBR) and *A. sohal* (Red Sea only). We surveyed additional acanthurid species (*A. lineatus* and *A. triostegus*) to verify that co-residence of A1 and A2 is consistent among GBR red-algae feeding acanthurids and characterize the symbiont diversity within and between hosts.

METHODS

Sample summary

Acanthurus lineatus, *A. nigrofuscus*, and *A. triostegus* were collected by spear from reef habitats near Lizard Island, Australia from 2014-2016. Fish were collected under the James Cook University Ethics approvals A1641 and A2345 with approval of the Great Barrier Reef Marine Park Authority permit G13-35909.1. The entire gastrointestinal tract was removed from each fish and divided into five sections as described in (21). Gut fluid samples from section IV, containing large numbers of epulos, were either fixed in 80% ethanol, and stored at -20 ˚C or processed immediately to extract RNA. A single *A. lineatus* host served as a sample source for genomic analysis. For transcriptome studies, multiple fish were caught within time intervals 0900-1200 hours (MM), 1200-1500 (AF), and 1500-1800 (LAF). Sampling information, demographic data for each fish host, and descriptions of corresponding epulo morphotypes can be found in Table 1.

Single-cell genome extraction and amplification

Epulopiscium A1 and A2 cells were manually selected from fixed intestinal contents of *A. lineatus*, using a standard Gilson pipettor and a Nikon SMZ-U dissecting microscope. Each cell was processed separately, washed with Agencourt AMPure magnetic beads to remove exogenous DNA, then lysed and subjected to whole-genome, single-cell amplification (Qiagen REPLI-g Kit) as described previously (17).

Amplified genomic DNA was visualized on a 1% agarose gel. Samples with high molecular weight genomic DNA (gDNA) were purified using QIAmp DNA mini kit (Qiagen). Purified gDNA was screened via PCR (HotStarTaq Master Mix, Qiagen) using Epulo-specific primer sets 515F/1423R (15, 22) and screened for purity using bacterial universal primers set 8F/1492R (22). After extraction with QIAquick PCR

purification system kit (Qiagen), amplicons were sequenced (in both directions) on an ABI 3730 automated sequencer (Biotechnology Resource Center, Cornell University). Sequences were trimmed and analyzed using Geneious version 6.0.6 (*http://www.geneious.com***,** Kearse *et al.* 2012) and compared to nucleotide sequences in Genbank using BLAST (24). Samples that shared >98% 16S rRNA gene sequence identity with *Epulopiscium* clade targets were used for sequencing. The amplified DNA of these candidate samples were de-branched using Nuclease S1 (Promega) and purified by phenol/chloroform extraction prior to library preparation.

Whole genome sequencing and assembly

Single-cell amplified DNA was sequenced at the Biotechnology Resource Center (Cornell University). All samples were sequenced in one run and the same lane using the Illumina MiSeq platform as 2x250 paired-end reads.

All raw reads were processed by the open source BBtools software package (*https://sourceforge.net/projects/bbmap/*). Pre-assembly processing of reads resulted in either paired-end, interleaved reads or merged, error-corrected and normalized reads as inputs for assembly. The paired-end, interleaved reads were assembled using the metaSpades mode of SPAdes 3.7. The normalized-merged-error-corrected reads were assembled using the single-cell mode and "careful" parameter of SPAdes 3.6. All samples were processed using both approaches. Trimmed reads were mapped back to their concurrent assembly to determine read coverage using bbmap and samtools. Genome statistics were calculated by QUAST 4.5 (25) using either default parameters for contig assemblies or the scaffold option for scaffold assemblies.

Table 1. Sample description of acanthurid hosts, corresponding epulo A1 and A2, and analyses performed.

Acanthurid host						*Epulopiscium* symbiont			Analyses
Species	ID	Sample time*	Sample date	Sample location	Sex	Avg. mother-cell length (μm)	Number of offspring per mother-cell	Relative offspring size[#]	
A. lineatus	M556	1610	12/10/2014	Bommie Bay	f	A1: 194 A2: 186	A1: 2 A2: 2	A1: 0.351 A2: 0.343	Genome assembly, FISH microscopy, 16S clone libraries
A. lineatus	G155	955	12/8/2017	South Is. Front	f	A1: 157 A2: 173	A1: 2 A2: 2	A1: 0.084 A2: 0.095	RNAseq
A. lineatus	G156	1005	12/8/2017	South Is. Front	f	A1: 118 A2: 90	A1: 1-2 A2: 1-2	A1: 0.081 A2: 0.143	RNAseq
A. lineatus	G182	1205	12/11/2017	North Reef	m	A1: 98 A2: 114	A1: 1-2 A2: 1-2	A1: 0.315 A2: 0.348	RNAseq
A. lineatus	G186	1202	12/11/2017	North Reef	f	A1: 196 A2: 201	A1: 1-3 A2: 1-2	A1: 0.397 A2: 0.386	RNAseq
A. lineatus	G174	1625	12/10/2017	North Reef	-	A1: 146 A2: 205	A1: 1-3 A2: 1-2	A1: 0.685 A2: 0.685	RNAseq
A. lineatus	G187	1620	12/11/2017	Washing Machine	f	A1: 165 A2: 214	A1: 2 A2: 1-2	A1: 0.576 A2: 0.674	RNAseq
A. triostegus	G112	1330	12/10/2016	Osprey Island	f	A1: 93.3 A2: 76.4	A1: 2 A2: 2	A1: 0.288 A2: 0.269	FISH microscopy, 16S clone libraries
A. triostegus	G113	1330	12/10/2016	Osprey Island	f	A1: 175 A2: 226	A1: 2 A2: 2	A1: 0.351 A2: 0.151	FISH microscopy, 16S clone libraries
A. nigrofuscus	G114	1330	12/10/2016	Osprey Island	m	A1: 338 A2: 336	A1: 2 A2: 2	A1: 0.207 A2: 0.159	FISH microscopy, 16S clone libraries

*RNAseq samples were collected at time intervals (0900-1200), middle morning (MM); (1200-1500), afternoon (AF); and (1500-1800), late afternoon (LAF).

Offspring size shown as the average offspring cell length divided by the mother-cell length.

Contamination control, genome completeness and genome finishing

Metagenome binning approaches were implemented to enrich for epulo-specific contigs and remove residual contamination from amplified DNA. Contigs were binned based on percent G-C, coverage, length, taxon assignment, and by linkage in an assembly network. A taxon assignment was produced for each contig by first comparing nucleotide sequences to the protein database of NCBI (blastx) and then assigning the lowest common taxon to each contig using the BLAST2LCA program (*https://github.com/emepyc/Blast2lca*). An assembly network was constructed by linking the ends of contigs (100 bp from end) based on overlapping read coverage using BamM. All analyses were done in R 3.3.1 (26).

Assemblies were assessed for completeness using checkM at the class level Clostridia (27) and using a phylogenetically-targeted conserved single-copy gene list (334 genes). *Epulopiscium* spp. belong to an uncharacterized family (Lachnospiraceae cluster XIVb) that is not specified in checkM. Instead, a conserved gene list was constructed using genes that were experimentally determined to be essential in *Bacillus subtilis* (28, 29), considered single-copy in 95% bacterial genomes according to the Comprehensive Microbial Resource database (30), and predicted as the core minimal bacterial gene set required for life in the absence of environmental stress (31). This conserved list was normalized against the closest epulo relative with a complete genome, *Cellulolyticum lentocellum* DSM 5247 (CP002582).

Relative contamination and heterogeneity were assessed from assembly networks, taxonomy reports, checkM, and the conserved gene list. Duplicated contigs were removed at a cutoff of 95% nucleotide sequence identity using dedupe from bbtools.

Scaffolds were produced using finishM (*https://github.com/wwood/finishm*) and manual curation from Mauve alignments. Different assemblies from the same

sample were compared using Mauve to investigate potential scaffolds. These scaffolds were assembled *de novo* using Geneious and verified by gene neighborhoods of the draft genome from "*Ca*. Epulopiscium viviparus clade B Nt_PAG450" (Sannino, unpublished).

Gene identification

Near-complete draft genomes were submitted to IMG for annotation using the MGAP v4 pipeline (32). Gene content and gene neighborhood comparisons were initially analyzed on the Integrated Microbial Genomes (IMG) comparative analysis system (33). Searches for polysaccharide metabolic enzymes were performed using the Hmmerscan program in HMMER (3.1) against the CAZyme family-specific HMM (hidden Markov model) database downloaded from dbCAN (34). Initial results were processed by the hmmscan-parser script provided by dbCAN. Briefly, the parser removes overlapping and redundant HMM matches by retaining the lowest e-value, calculates the covered fraction of the HMM alignment, and filters with an e-value cutoff 1e-5 (or 1e-3 if HMM length < 80 aa) and covered HMM fraction > 0.3. To eliminate false positives and redundancies not identified using the parser, we further filtered all HMM hits at an e-value < 1e-5 using R and manually removed duplicated GH74 and GH117 domains that were not overlapping but found in tandem, collectively aligning to the entire length of an HMM domain. Secretory proteins were first detected from IMG annotations and verified/revised using SignalP. Gene specific BLAST searches (e-value less than 1e-5, minimal length larger than 50%) were performed against the Transporter Classification Database (TCDB; *http://www.tcdb.org*)(35) and the RCSB protein databank (PDB; *www.rcsb.org*)(36) for determining substrate specificity of transport proteins and catalytic enzymes, respectively. The metabolic pathways were inferred from KEGG database and MicroCyc.

We performed an independent COG (cluster of orthologous groups) analysis of each of the genomes. We used prodigal (v2.6.2) to generate ORFs and performed a protein BLAST against the COG database. Using R, we filtered for sequences with best hits and e-values greater than 1e-5, assigned unique matches to COG clusters, and calculated the proportion of genes assigned to each COG category. Some gene queries can be classified into multiple categories which overestimates the total gene counts per genome. We also performed a COG analysis using OrthoVenn to compare genomes and visualize the results (37).

Morphological diversity using Fluorescent *in situ* Hybridization (FISH)

The reproductive cell cycle of epulo symbionts are often synchronized within each host and follow a predictable daily cycle. Therefore, hosts caught around the same time of day and at a similar stage of development (Fig. 1, stage IV) were selected for FISH (Table 1) to reduce variation in cell morphologies affected by developmental stage. FISH using small-subunit rRNA probes were performed on ethanol-fixed cells as described in (38) with minor modifications. Briefly, a small aliquot (~40 μl) of fixed intestinal contents were processed separately in centrifuge tubes. Cells were also stained with DNA-specific DAPI (4′,6-diamidino-2-phenylindol) at a final concentration of 2 μg/ml. Processed samples were placed on slides, mounted in Citifluor, and sealed with 22x55mm coverglass. To differentiate between the two morphotypes, an epulo A1-specific probe was designed (A1_132R: 5'-TEX15/CGTTATTCCACAGTATAAGGTAGG) by aligning available 16S rRNA sequences from the Genbank database along with sequences derived in this study. Epulo A2-morphotypes were identified using the previously described probe (A2_105R: 5'-FAM/CCACTAACTTCTTAGAGCAA)(39). Brightfield and fluorescence microscopy were performed using an OlympusBX61 epifluorescence microscope. All epulo mother cells present on each slide that successfully hybridized

to an epulo specific probe were measured using a Cooke SensiCam with a Sony Interline chip and Slidebook software (Intelligent Imaging Inc.). Given the cigar-shape cellular morphology of large epulos, volume was calculated using the formula for a prolate ellipsoid. The width-to-length ratio represents the overall cell shape: thin and needle-like to wide and bulbous.

Statistical comparisons for each morphological parameter (mother-cell length, volume and width-to-length ratio) were conducted separately using R version 3.5 (26). To determine whether symbiont morphotypes differed between hosts, a two-way ANOVA test with a Bonferroni correction for multiple comparisons was conducted using package *car* 3.0-2 (40). A *post hoc* Tukey HSD test was used to identify which hosts contained significantly different symbiont morphotypes using package *multcomp* 1.4-8. A two-sample Wilcox test was used to determine intrahost differences between morphotypes A1 and A2 using the *stats* R package. Boxplots with points were generated using *ggplot* 2.3.0.0 and *cowplot* 0.9.3.

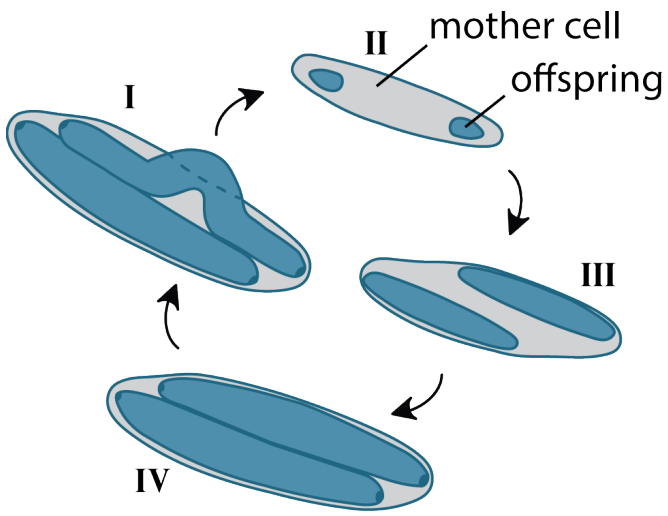


Figure 1. Intracellular offspring formation overview. Fully developed daughter cells emerge from the mother cell in the early morning (stage I) and often have condensed DNA or have divided at both poles of the cell. After division, each polar cell is engulfed (stage II) and continue to grow within the mother cell into the afternoon (stage III). In the evening, daughter cells are near full grown (stage IV) and emerge the following morning.

16S rRNA gene clone library surveys

Approximately 1,000-2,000 *Epulopiscium* cells were manually isolated from fixed intestinal contents of *A. lineatus*, *A. nigrofuscus* and *A. triostegus*, as described above. Cells were lysed using a combination of rounds of rapid freezing and thawing (3 times at -80 °C then 37 °C) in 10 mM Tris buffer followed by alkaline lysis for 30 minutes at 70 °C. Universal primers 8F/1492R were used to amplify the entire small subunit (SSU) rRNA gene from the lysate and screened for epulo-derived 16S rRNA gene sequences using epulo-specific primers 515F/1423R. After confirming presence of *Epulopiscium*-derived DNA, amplicons using the 8F/1492R primer set were cloned into the pCR 2.1-TOPO vector and transformed into One Shot TOP10 competent *E. coli* (Life Technologies, Carlsbad, CA) per manufacturer's recommendations. Plasmids from white clones were subjected to single colony PCR using vector primers

T7F/M13R and confirmed for expected product size by agarose gel electrophoresis. Plasmid DNA was isolated from confirmed clones using QIAprep Spin Miniprep kit (Qiagen, Venlo, Netherlands), according to manufacturer's recommendations. Concentrations of plasmid DNA were determined spectrophotometrically (Thermo Scientific NanoDrop 2000 spectrophotometer, Wilmington, DE). For each putative *Epulopiscium* morphotype per host, at least 5 plasmids from the 8F/1492 library were sequenced in both directions by an ABI 3730 automated sequencer (Biotechnology Resource Center, Cornell University). Sequences were analyzed and vector sequences were removed using Geneious version 6.0.6 (http://www.geneious.com, Kearse *et al.* 2012) and subsequently compared to nucleotide sequences in Genbank using BLAST (24).

Phylogenetic analysis

16S rRNA genes were used to confirm previously identified clades, validate identity of assembled genomes, and determine phylogenetic diversity of A1 and A2 morphotypes. Nucleotide sequences were aligned in Geneious (23) using MUSCLE (41). Maximum likelihood (ML) trees were constructed using PhyML 3.0 (42) and Modelfinder (43). Bootstrap values (44) were generated from 1,000 replicates using the generalized Time Reversible nucleotide substitution model (45) plus invariant sites (I) and Γ rate heterogeneity.

A high resolution, concatenated ribosomal protein phylogeny was constructed with sequences from a few members of the Lachnospiraceae XIVa and all available genomes from XIVb. Also included were the recently published genomes of epulos from the Red Sea if they contained a complete 16S rRNA gene sequence. Ribosomal proteins were aligned individually and trimmed using Geneious. Proteins with less than 50% average length of alignment were removed. Three genomes (*Ce. ruminocola, C. propionicum*, Nuni2H_MBin003) contained duplicate genes that were

identified and one copy removed. In most cases, duplicates were 100% identical (amino acid sequence) and/or shorter than the average protein length. Duplicates that were not identical were selected based on amino acid similarity to the closest relative. For example, *C. propionicum* contains two separately annotated S4 proteins that appear to be non-overlapping fragments of the protein. I selected the longer of the two for analysis. A few genomes contained partial sequences that could be re-assembled to produce a complete sequence which was then included in the alignment.

Maximum likelihood (ML) and a Bayesian phylogenetic trees were constructed from the concatenation of 28 ribosomal protein alignments using PhyML (42) and MrBayes (46), respectively. Bootstrap values were generated from 1,000 replicates using the WAG matrix model. The Markov chain Monte Carlo (MCMC) was run for 20,000 generations, sampling every 1,000 generations, and a burnin fraction of 25%, using the optimal WAG amino acid replacement model. To further characterize the phylogenetic relationships of these clades, we calculated the average nucleotide identity (ANI) of whole genomes using pyani Python3 module ANIb method (47). Results were visualized with the *pheatmap* 1.0.12 R package (*https://cran.r-project.org/web/packages/pheatmap/index.html*).

Transcriptomic profiles of symbionts over time

Total RNA was extracted from gut fluid. Cells were lysed by bead beating in RLT buffer (Qiagen) containing β-mercaptoethanol and Ambion RNase inhibitor. RNA was further purified using the Qiagen RNAeasy kit following manufacturer's instructions, stored in liquid nitrogen for international shipment, and stored at -20°C. DNA was removed using the TURBO DNA-free kit (Invitrogen). The quantity and quality of RNA was determined by Qubit spectrofluorometery and fragment analysis (ABI 3730xl, Cornell University Genomics facility), respectively. Ribosomal RNA from the host and gut community was removed using Illumina Ribozero Gold rRNA

removal kit (epidemiology). Libraries were prepared as directional (stranded) and ~75nt single reads were sequenced using Illumina NextSeq500 by the RNA sequencing core (Cornell).

Reads were trimmed using Trimmomatic v0.36 (48) and residual rRNA sequences were removed by mapping known rRNA sequences with bwa mem (49). Processed reads were mapped to draft genomes *Al*_SAG11, *Al*_SAG7, and *Al*_SAG9 using bwa mem. Since the Alin6 genome is nearly identical to that of *Al*_SAG11, mapped reads were not analyzed for *Al*_SAG6. Raw count summaries were calculated and organized using samtools and the mRNAtool.pl script (*https://github.com/kentnf/KTools/blob/master/mRNAtool.pl*). Differential expression across time intervals (MM, AF, and LAF; see supplemental Table1) was analyzed using R packages *DeSeq2* 1.22 and *vsn* 3.50 and visualized as heatmaps using *pheatmap*.

RESULTS AND DISCUSSION

Phylogenetic and morphological diversity of epulos across acanthurid hosts

Although the largest bacterial symbionts of acanthurids were given the name *Epulopiscium fishelsoni* (50), early sequence-based and FISH-based surveys suggested that these cells comprised two distinct lineages, called A1 and A2 (15). This pattern was found in symbiont populations from *A. nigrofuscus* collected in the Red Sea as well as from *A. nigrofuscus* and *A. lineatus* collected from the Great Barrier Reef (GBR)(39). Bacterial community surveys supported this observation in intestinal samples from *A. sohal* and *A. nigrofuscus* in the Red Sea (13). Features of these co-inhabitant populations were more deeply explored using 16S rRNA gene surveys from gut contents of GBR-dwelling *A. lineatus*, *A. nigrofuscus* and *A. triostegus*, which confirmed the presence of both clades in these fish. Sequences derived from the GBR

form a sister group with sequences derived from the Red Sea which likely represent distinct epulo species (Fig. 2).

Morphological diversity of symbionts within and between acanthurid hosts from the GBR were compared using FISH probes specific to the A1 and A2 clades. Cells were collected at the same time of day to control for developmental stage of symbionts (Fig. 1, stage IV) and yet morphological diversity of symbionts varied widely between hosts. Symbiont length differed significantly in the A1 and A2 clades between hosts (Anova $P < 0.001$; *post hoc* Tukey's HSD $P < 0.001$) (Fig. 3). The largest symbionts, averaging around 320 μm in length and 60,000 μm^3 in volume, were found in *A. nigrofuscus* while *A. triostegus* often contained the smallest, reaching lengths ~30 μm. Bacterial morphological plasticity may occur as an adaptation to environmental changes (51–53). These adaptive features include the ability to readily acquire nutrients, enhance motility and dispersal, tolerate external stresses, and aid in interactions with other microorganisms.

Several population and phylogenetic parameters distinguished A2 and A1. Based on FISH surveys and 16S rRNA gene amplicon libraries, A2 symbionts are more abundant than A1 symbionts in all hosts examined. The 16S rRNA gene sequences for clade A2 were too similar to resolve differences between populations from different hosts (all symbionts sharing >99% sequence identity) (Fig. 2). Clade A1, on the other hand, contained more divergent sequences (sharing ~97% sequence identity). Some clustering by host was observed for A1 but it was not uniform. The admixture in this clade suggests that the morphological and population differences observed may not be due to phylogenetic factors but functional plasticity or differences in roles within the host environment.

Figure 2. Maximum Likelihood 16S rRNA gene phylogeny of epulos and their relatives. Sequences from A1 and A2 genomes characterized in this study are highlighted in bold font. If clone sequences contained additional similar sequences, the total number is shown in parentheses. Geographic locations for A1 and A2 sequences are indicated by solid circle (Great Barrier Reef "GBR") or open circle (Red Sea). Hosts are color coded (pink – *Acanthurus lineat*us; aqua – *Acanthurus triostegus*; orange – *Acanthurus nigrofuscus*; green – *Acanthurus sohal*). Previously described clades are indicated by vertical lines. A member of the Clostridial cluster IV Ruminococcaceae (*Faecalibacterium prausnitzii*) was used as the outgroup. Genus names that are pending re-classification are bracketed. Red arrow indicates original sequence named *Epulopiscium fishelsoni*. Bootstrap values ≥ 70% are indicated at nodes. The scale represents 0.09 nucleotide substitutions.

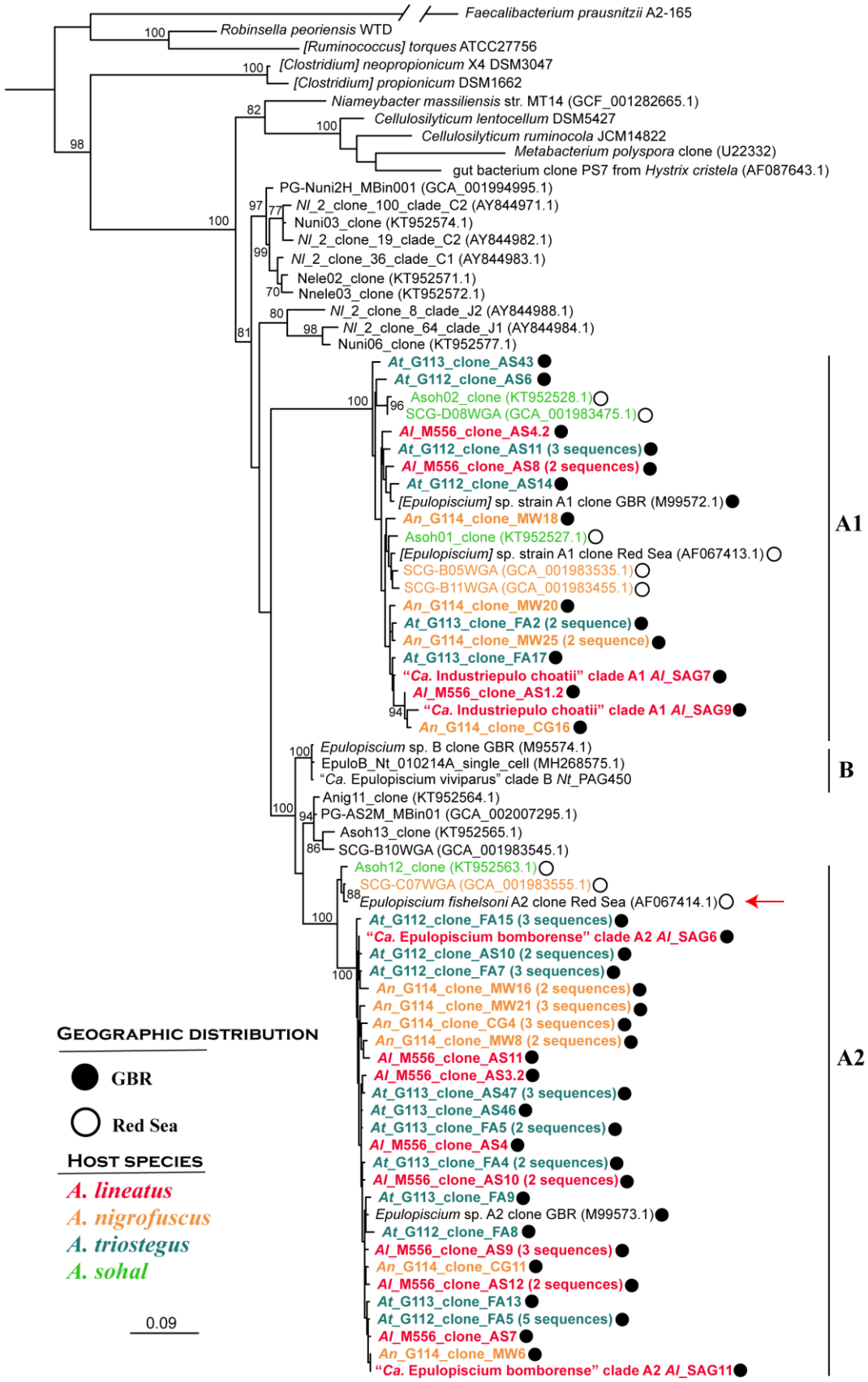
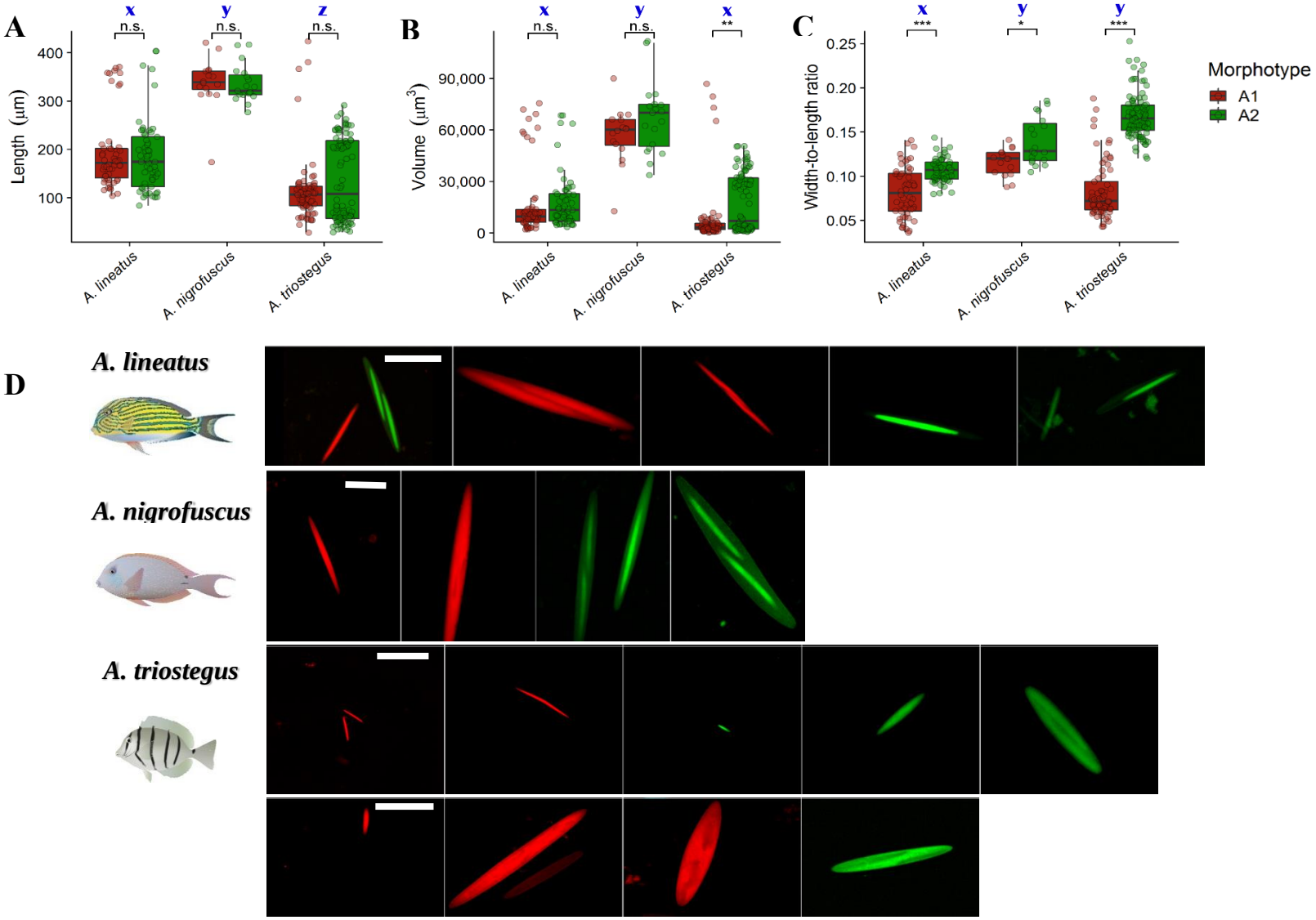

Faecalibacterium prausnitzii A2-165
Robinsella peoriensis WTD
[Ruminococcus] torques ATCC27756
[Clostridium] neopropionicum X4 DSM3047
[Clostridium] propionicum DSM1662
Niameybacter massiliensis str. MT14 (GCF_001282665.1)
Cellulosilyticum lentocellum DSM5427
Cellulosilyticum ruminocola JCM14822
Metabacterium polyspora clone (U22332)
gut bacterium clone PS7 from Hystrix cristela (AF087643.1)
PG-Nuni2H_MBin001 (GCA_001994995.1)
Nl_2_clone_100_clade_C2 (AY844971.1)
Nuni03_clone (KT952574.1)
Nl_2_clone_19_clade_C2 (AY844982.1)
Nl_2_clone_36_clade_C1 (AY844983.1)
Nele02_clone (KT952571.1)
Nnele03_clone (KT952572.1)
Nl_2_clone_8_clade_J2 (AY844988.1)
Nl_2_clone_64_clade_J1 (AY844984.1)
Nuni06_clone (KT952577.1)
At_G113_clone_AS43
At_G112_clone_AS6
Asoh02_clone (KT952528.1)
SCG-D08WGA (GCA_001983475.1)
Al_M556_clone_AS4.2
At_G112_clone_AS11 (3 sequences)
Al_M556_clone_AS8 (2 sequences)
At_G112_clone_AS14
[Epulopiscium] sp. strain A1 clone GBR (M99572.1)
An_G114_clone_MW18
Asoh01_clone (KT952527.1)
[Epulopiscium] sp. strain A1 clone Red Sea (AF067413.1)
SCG-B05WGA (GCA_001983535.1)
SCG-B11WGA (GCA_001983455.1)
An_G114_clone_MW20
At_G113_clone_FA2 (2 sequence)
An_G114_clone_MW25 (2 sequence)
At_G113_clone_FA17
"Ca. Industriepulo choatii" clade A1 Al_SAG7
Al_M556_clone_AS1.2
"Ca. Industriepulo choatii" clade A1 Al_SAG9
An_G114_clone_CG16
A1
Epulopiscium sp. B clone GBR (M95574.1)
EpuloB_Nt_010214A_single_cell (MH268575.1)
"Ca. Epulopiscium viviparus" clade B Nt_PAG450
B
Anig11_clone (KT952564.1)
PG-AS2M_MBin01 (GCA_002007295.1)
Asoh13_clone (KT952565.1)
SCG-B10WGA (GCA_001983545.1)
Asoh12_clone (KT952563.1)
SCG-C07WGA (GCA_001983555.1)
Epulopiscium fishelsoni A2 clone Red Sea (AF067414.1)
At_G112_clone_FA15 (3 sequences)
"Ca. Epulopiscium bomborense" clade A2 Al_SAG6
At_G112_clone_AS10 (2 sequences)
At_G112_clone_FA7 (3 sequences)
An_G114_clone_MW16 (2 sequences)
An_G114 _clone_MW21 (3 sequences)
An_G114_clone_CG4 (3 sequences)
An_G114_clone_MW8 (2 sequences)
Al_M556_clone_AS11
Al_M556_clone_AS3.2
At_G113_clone_AS47 (3 sequences)
At_G113_clone_AS46
At_G113_clone_FA5 (2 sequences)
Al_M556_clone_AS4
At_G113_clone_FA4 (2 sequences)
Al_M556_clone_AS10 (2 sequences)
At_G113_clone_FA9
Epulopiscium sp. A2 clone GBR (M99573.1)
At_G112_clone_FA8
Al_M556_clone_AS9 (3 sequences)
An_G114_clone_CG11
Al_M556_clone_AS12 (2 sequences)
At_G113_clone_FA13
At_G112_clone_FA5 (5 sequences)
Al_M556_clone_AS7
An_G114_clone_MW6
"Ca. Epulopiscium bomborense" clade A2 Al_SAG11
A2
GEOGRAPHIC DISTRIBUTION
GBR
Red Sea
HOST SPECIES
A. lineatus
A. nigrofuscus
A. triostegus
A. sohal
0.09

Figure 3. Morphological diversity of symbiont cells within and between acanthurid hosts. Tukey's boxplots show the (A) length, (B) volume and (C) width-to-length ratio of symbiont mother-cells across hosts. Morphotypes A1 (red) and A2 (green) were counted separately. Reported *P* vales indicate whether morphotypes A1 and A2 were statistically different within each host: n.s. – no significance; * < 0.05; ** < 0.01; ***< 0.001. Morphological differences between hosts that were significantly different from each other are based on *post hoc* Tukey's HSD ($P < 0.001$). Significant groups are represented as letters "x", "y", and "z" above each bar. Sample size for each host are as follows: *A. lineatus* = 54-57, *A. nigrofuscus* = 16-19, and *A. triostegus* = 69-94. FISH images of intestinal symbionts present in each acanthurid host are shown (D). 16S rRNA FISH probes specific to A1-morphotype cells are labeled with Texas 615 (red) and A2-morphotype cells are labeled with Fluorescein (green). The scale bar represents 100 µm.


A
B
C
D
x
y
z
n.s.
**
***
*
Length (µm)
Volume (µm3)
Width-to-length ratio
A. lineatus
A. nigrofuscus
A. triostegus
Morphotype
A1
A2

Genome characteristics of symbiont SAGs

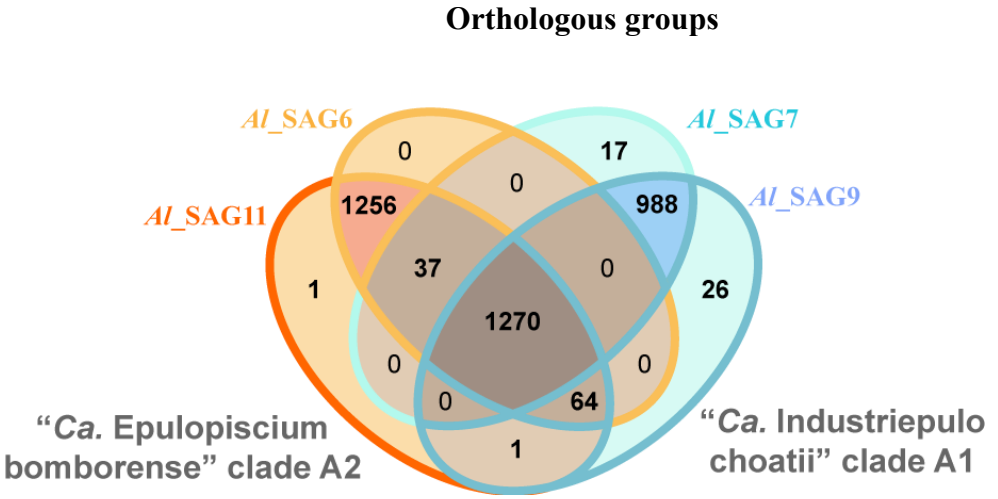
We assembled 4 high quality draft single-cell amplified genomes (SAGs) representing the *Epulopiscium* spp. A1 and A2 clades (Fig. 4) collected from a single *A. lineatus* host. *Al*_SAG7 and *Al*_SAG9 shared 98-99% 16S rRNA gene sequence similarity with *Epulopiscium* spp. clade A1 clones from the Great Barrier Reef and Red Sea (M99572.1 and AF067413.1, respectively). These two clade A1 SAGs are likely the same species, sharing 92% ANI (Fig. 5) and 98% 16S rRNA gene sequence similarity but their gene content differs slightly (Fig. 4). *Al*_SAG6 and *Al*_SAG11 shared 99% 16S rRNA sequence similarity to *Epulopiscium* spp. clade A2 from GBR (M99573.1). These two clade A2 SAGs are nearly identical to each other sharing 99.9% ANI, 99% 16S rRNA gene sequence identity and 99.9% of their orthologous genes (Fig. 4, Fig. 5). The ANI between genomes from clades A1 and A2 is 73%, signifying that clades A1 and A2 are different genera (Fig. 5).

Common characteristics that *Epulopiscium* spp. genomes share are the presence of multiple short tandem repeats, many with putative cell surface proteins and low frequency of mobile elements. The most complete assembly of an *Epulopiscium* spp. genome to date, "*Ca.* Epulopiscium viviparus" with 7 contigs (Sannino unpublished), was able to overcome many of these repetitive features by combining long PacBio reads and short accurate Illumina reads.

The interplay between mobile elements and polyploidy can have major consequences to genome structure and stability. At one extreme, free-living *Thiomargarita* spp. appear to tolerate an excessive number of mobile elements, compared to other gammaproteobacteria, due to genome redundancy. It has been

suggested that these dynamic genomes increase adaptability to new environments via changes facilitated by mobile elements (54). In contrast, the nominal frequency of mobile elements in epulos (Fig. 6) may be attributed to strong genome purification forces associated with the presence of multiple genome copies and reproductive bottlenecks (17).

The A1 and A2 genomes code for multiple surface proteins that range in size from 200-7,000 amino acids. Many of these giant genes (encode for proteins >2,000 amino acids in length) contain S-layer domains and/or signal peptides, and are likely secreted outside of the cell. Some of the large genes contain leucine-rich repeats or Ig-like domains with unknown functions, that might interact with host or other proteins in the gut. Other large genes contain domains related to carbohydrate-specific binding or degradation processes. Very often, these large predicted proteins contain multiple domains. For example, genome *Al*_SAG11 encodes for an 8,103 amino acid long secreted beta-glucanase that also contains an S-layer domain. It is possible that epulos coat their exterior with proteins to aid capture of complex polysaccharides for localized degradation and import of compounds into the cell via its diverse array of transporters for specific organic carbon molecules. Other large proteins lacking signal peptides might aid in post-translational regulation of intracellular proteins. Many of these putative cytoplasmic large proteins contain a HECT-domain that is commonly found in the E3 ubiquitin ligase family.



Genome Stats	*"Ca. Epulopiscium bomborense" Al*_SAG6	*"Ca. Epulopiscium bomborense" Al*_SAG11	*"Ca. Industriepulo choatii" Al*_SAG7	*"Ca. Industriepulo choatii" Al*_SAG9	*"Ca. Epulopiscium viviparus" Nt*_PAG450
Clade/morphotype	A2		A1		B
Assembly quality & analysis type	High quality draft SAG	High quality draft SAG	High quality draft SAG	High quality draft SAG	High quality draft PAG
Genome size (bp)	3,482,333	3,525,435	3,077,541	3,366,594	3,282,201
Scaffolds (Contigs)	69 (74)	49	31	35	7
%GC	30.52	30.58	30.03	30.30	38.08
Contig N50/L50	91,473/16	109,169/11	195,656/6	214,851/6	602,975/3
Max contig length (bp)	250,932	250,934	358,768	435,353	905,387
Protein coding genes (percent of total genes)	2,714 (97.56%)	2,716 (97.42%)	2,663 (97.15%)	2,899 (97.15%)	2,635 (97.09%)
%Completeness (CheckM/Reference)	99/92	99/92	98/91	98/91	99/92
%Contamination/ Heterogeneity	0.91/0	0.91/0	0/0	0.91/0	0.91/0

Figure 4. Genome comparison of epulos from the Great Barrier Reef. (Top) Venn diagram comparison of shared and disparate orthologous clusters between genomes from "*Ca.* Epulopiscium bomborense" clade A2 and "*Ca.* Industriepulo choatii' clade A1. (Bottom) Genome assembly descriptions for single-cell amplified genomes (SAGs) from clades A2 and A1 along with a population assembled genome (PAG) from clade B "*Ca.* Epulopiscium viviparus" *Nt*_PAG450 (Sannino, unpublished). Assembly quality defined using previously described parameters (55). Completeness, contamination and heterogeneity percentages were calculated using CheckM for class Clostridia. Additional completeness calculations were compared to the closest relative with a complete genome, *C. lentocellum* DSM5427.

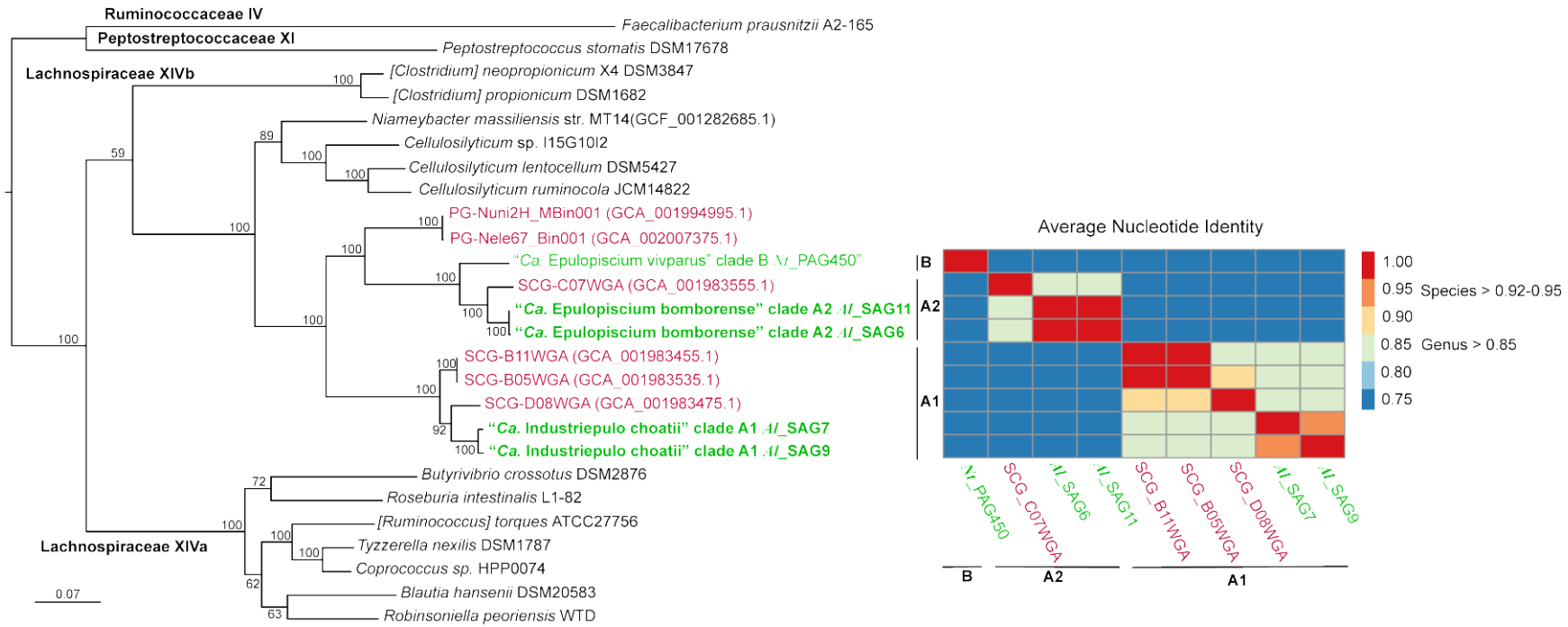


Figure 5. Taxonomic classification of novel epulos and their relatives. Maximum Likelihood phylogeny reconstruction of *Epulopiscium* spp. and their relatives from 26 ribosomal proteins (4,090 aa). Genomes characterized in this study are in bold and sampling site location are color coded (green – Great Barrier Reef; pink – Red Sea). Lachnospiraceae subdivisions are indicated near ancestral clusters. Previously described clades for *Epulopiscium* spp. A1, A2, and B are indicated by vertical lines. Genus names that are pending re-classification are bracketed. Bootstrap values ≥ 50% are indicated at nodes. The scale represents 0.07 nucleotide substitutions. Heatmap shows the average nucleotide identity (ANI) of genomes from clades A1, A2 and B. Taxonomic cutoff values are indicated for genus and species.

Proposed names for novel genera and species "*Ca*. Epulopiscium bomborense" clade A2 and "*Ca*. Industriepulo choatii" clade A1

We would like to clarify and reclassify *Epulopiscium* spp. based on the improved resolution that these genome data provide. We propose that the A2 clade from the Red Sea retain the name "*Ca*. Epulopiscium fishelsoni" (15)(Fig. 2, red arrow), and the A2 clade from the GBR (that shares ~85% ANI) is a novel species "*Candidatus* Epulopiscium bomborense". Further, the A1 clone from the GBR will be reclassified as a different genus ("*Ca*. Industriepulo", see below).

According to 16S rRNA gene sequence similarity (98%), *Al*_SAG6 and *Al*_SAG11 are closely related to the originally named "*Ca*. Epulopiscium fishelsoni" cloned A2 sequence from the Red Sea (*A. nigrofuscus*). A recently available genome from the Red Sea C07WGA_EpuloA2 (14) shares 99% 16S rRNA gene sequence similarity with "*Ca*. E. fishelsoni". However, genome average nucleotide identity (ANI) comparisons of ~85% between the GBR genomes and the Red Sea genomes indicate that they are different species roughly within the same genus "*Ca.* Epulopiscium" (Figure 5). We propose that the A2 clade from GBR is a novel species "*Candidatus* Epulopiscium bomborense", in reference to the location where the surgeonfish host was caught (Bommie Bay, AUS). Bommie, Australian slang for bombora, is an indigenous australian term that describes a distant area of waves breaking over a reef, rock shelf or sand bar. SAGs *Al*_SAG6 and *Al*_SAG11 are nearly identical, sharing 99% 16S rRNA gene sequence similarity and 99.9% ANI. These two cells will be the first representative genomes of this species: "*Candidatus* Epulopiscium bomborense" clade A2 *Al*_SAG6 and "*Candidatus* Epulopiscium bomborense"clade A2 *Al*_SAG11. Members of this species have been observed in surgeonfish *Acanthurus nigrofuscus*, *A. lineatus*, and *A. triostegus*. The latter two surgeonfish species are not found in the Red Sea.

*Al*_SAG7 and *Al*_SAG9 from clade A1 represents a novel genus and species, "*Candidatus* Industriepulo choatii". Recently sequenced A1 clade symbiont genomes collected from Red Sea-dwelling *A. sohal* and *A. nigrofuscus* (SCG-D08WGA, SCG-B05WGA, and SCG-B11WGA) are distinct from *Al*_SAG7 and *Al*_SAG9; sharing 98% 16S rRNA gene sequence identity and 83-84% ANI (Figure 5). The species "choatii" honors fish ecologist Howard Choat. Members of "*Ca.* I. choatii" co-occur with "*Ca.* E. bomborense" in the same species of surgeonfish (*A. nigrofuscus, A. lineatus,* and *A. triostegus*) from the South Pacific.

Genome functional diversity

SAGs from the phylogenetically distinct A1 and A2 GBR clades share ~50% of their orthologous genes (Fig. 4). Major differences between the two clades are the proportion of COGs contributing to carbohydrate transport and metabolism versus energy production and conversion (Fig. 6). Similar to "*Ca.* Epulopiscium viviparus", genomes from clades A1 and A2 have a high percentage of genes associated with carbon metabolism transport (12.4% and 14.5%, respectively), with A2 more comparable to "*Ca.* Epulopiscium viviparus". Both A1 and A2 clades are predicted to have obligate fermentative lifestyles similar to "*Ca.* Epulopiscium viviparus". Surprisingly, A1 SAGs are twice as enriched for energy processes (7.1% vs 3.6%) than either A2 or type B suggesting that symbiont energy dependencies differ. Analyses of carbohydrate active enzymes (Table 2) showed differences between A1 and A2 SAGs with respect to their carbohydrate use preferences. Similar to type B, A2 was 2-3 times more enriched for glycoside hydrolases and carbohydrate-binding modules than A1. However, A1 genomes are more enriched for glycosyltransferases, carbohydrate esterases and polysaccharide lyases than A2. These data indicate that A1 and A2 epulos have divergent metabolic roles. In general, type A2 is predicted to use a wider range of complex polysaccharides and synthesize all 20 amino acids (Fig. 7),

whereas type A1 is incapable of breaking down complex polysaccharides and is an auxotroph for histidine, arginine, methionine, proline, threonine and tryptophan (Fig. 8).

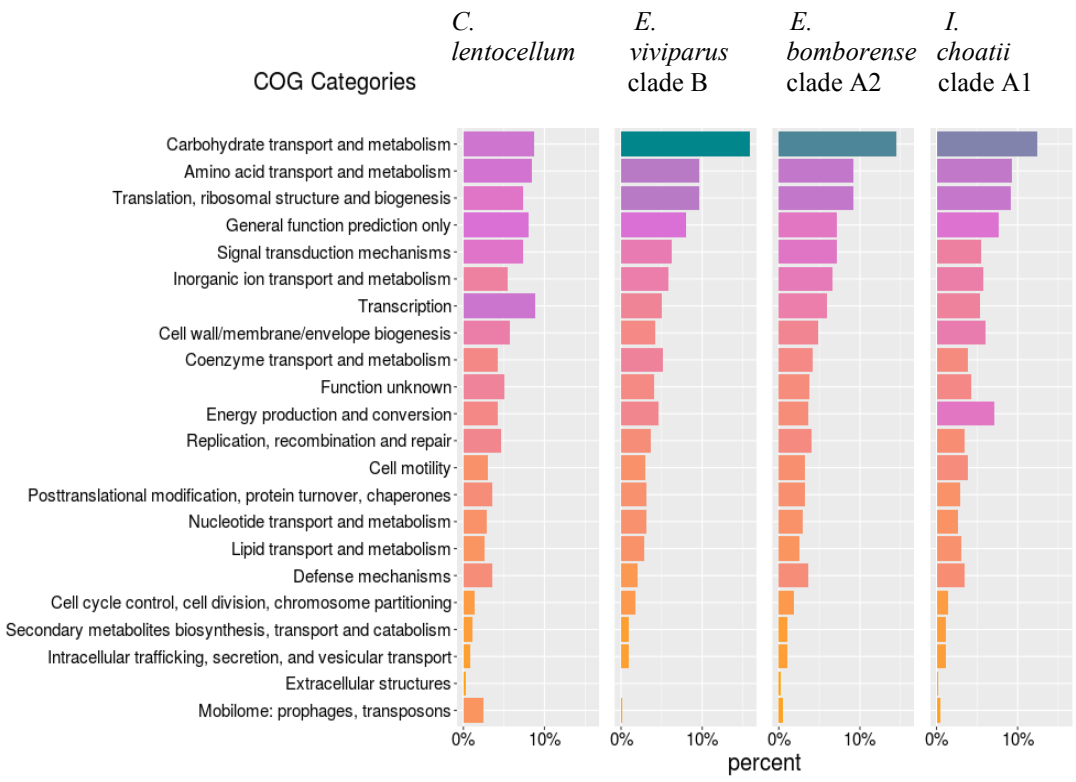


Figure 6. COG categories comparing the three different epulo clades A1, A2 and B along with non epulo relative *C. lentocellum*. Due to the similarity of gene content between clades, representative genomes *Al*_SAG9 and *Al*_SAG11 were used for clades A1 and A2, respectively.

Complex polysaccharide degradation capabilities differ between symbionts

Epulopiscium bomborense is the closest sequenced epulo to the model strain *E. viviparus* clade B and as such has similar metabolic characteristics. Both lack genes for the quinones and cytochromes required for respiration. They have a complete glycolysis pathway but an incomplete tricarboxylaic acid (TCA) cycle which is characteristic of many Clostridiales including *C. lentocellum*. They are predicted to degrade and assimilate a wide variety of complex and simple carbohydrates, eventually producing acetate and ethanol from fermenation (Fig. 7, Table 3).

Most of the complex polysaccharides that *E. bomborense* can potentially use are found in red algae (*Rhodophyta*). This correlates with the observations that its host *A. lineatus* predominantly feeds on red thallate/filamentous algae (56, 57) and brown algae minimally. Preference for red algae over brown may be related to a higher level of digestible carbohydrates and slightly higher levels of protein and lipids (58, 59). Agarose is the major constituent of red algae, taking up to 70% of the cell wall, and is composed of repeating neoagarobiose (3,6-anhydro-L-galactose-α-1,3-D-galactose) units joined by β-1,4-glycosidic bonds. *E. bomborense* is predicted to degrade agarose using a similar mechanism as the proteobacterium *Saccharophagus degradans* (60). Sulfated polysaccharides porphyran and carageenan are also constituents of red algae. The catalytic enzymes that break down these compounds are similar to agarases with the exception that they recongize the sulfate modifications of these polysaccharides. Another complex polysaccharide that can be associated with cell walls of red algae and siphonous green algae are xylans. Xylans are generally composed of β-1,3- or β-1,4-linked xyloses but can also contain acetyl groups, arabinose, arabinogalactans and methylated glucuronate. Enzymes predicted to breakdown components of brown seaweeds, including D-mannitol and laminarin, were also found in the genomes of *E. bomborense*. Although no genes that encode for alginases were identified in the *E.*

bomborense genome, they are predicted to import alginate-like oligosaccharides, derived from brown algae, via at least 19 possible transporters. Lastly, *E. bomborense* possess multiple genes putatively annotated as carbohydrate transport systems for the import of simple sugars. These include two CUT2 monosaccharide-specific ABC transporters 24 CUT1 di- and oligo-saccharide ABC transporters two ribose-specific ABC transporters, and fifteen carbohydrate-specific phosphotransferase (PTS) transporters for ascorbate, cellobiose, fructose, galactitol, mannitol, maltose/glucose, mannosylglycerate, sucrose and phosphoenolpyruvate.

Industriepulo cells are predicted to use the simple sugars xylose, fructose, sucrose, and galactose (Fig. 8). They contain multiple CUT1 and CUT2 ABC transporters with similar substrate specificities as *E. bomborense*. However, *Industriepulo* either lack or contain one aldouronate-specific ABC transporter and fewer PTS transporters than its co-inhabitant. Surprisingly, *I. choatii* lacks the genes required to degrade complex polysaccharides such as agarose, porphyran or carrageenan (Table 2). It has the potential to degrade alginate, laminarin and mannitol although these are brown algae components that would be rarely seen in the intestinal tracts of its host. Futhermore, between the two different genomes (*Al*_SAG7 and *Al*_SAG9), the ability to degrade alginate varies. Alginate is degraded extracellularly by polysaccharide lyases (or alginases) containing domains PL6, PL7, PL14, and PL18. Only genome *Al*_SAG9 contains putative alginases in which three have been identified: secreted endo-alginate lyase (PL7) and two cytoplasmic exo-alginate lyases (PL15 and PL17). Overall, the metabolic capacity of *I. choatii* highlights the functional diversity among epulos and within the *Industriepulo* genus.

Table 2. Genes detected by CAZyDB or protein BLAST comparsions to *Pseudoalteromonas*.

Catalytic domain	Polysaccharide target	Enzyme[1]	Reference match[1]	Other domains	Signal Peptide[2]	Qty
"Candidatus Epulopiscium bomborense" clade A2						
GH16	agarose/porphyranase	beta-porphyranase A	PDB:4ATE	CBM9, CBM16	yes	3
GH86	agarose/porphyranase	beta-porphyranase	PDB:4AW7	-	yes	6
GH50	agarose	exo-β-agarase	PDB:4BQ2	-	yes	3
GH117	agarose	alpha-neoagarobiose hydrolase	PDB:3R4Y	-	no	4
GH16	laminarin	endo-1,3-beta-glucanase	PDB:4CRQ	CBM61, CBM4, CBM4, CBM56, SLH	yes	3
GH31	xylan	xyloglucan, alpha-xylosidase	PDB:5JOU, PDB:1WE5	-	no	3
GH43	xylan	xyloglucan, xylosidase/arabinofuranosidase	PDB:5JOW, PDB:4NOV	-	no	3
GH150	lambda-carrageenan	lambda-carrageenase	3HXJ	-	yes	5
"Candidatus Industriepulo choatii" clade A1						
PL7	alginate	alginate lyase	PDB:2CWS	-	yes	1 (A9)
PL15	alginate	exo-alginate lyase	PDB:3A0O	-	no	1 (A9)
PL17	alginate	oligoalginate lyase	PDB:4NEI	-	no	1 (A9)
GH16	laminarin	laminarinase	PDB:3ILN	-	no	1
GH31	xylan	alpha-glucosidase	PDB:5AED	-	no	1
GH43	xylan	xyloglucan	PDB:5JOW	-	no	1 (A9)

[1]Enzyme identity was confirmed by sequence homology in the PDB database.
[2]Secretion signals were detected using SignalPeptide software.

Figure 7. Model of the carbohydrate and energy metabolism of "*Ca.* Epulopiscium bomborense" clade A2. Pathways were deduced from genome annotations and KEGG/Microcyc references. End products from fermentation are indicated in bold. Incomplete pathways are shown as dotted lines and reversible reactions contain white arrows. Buffering pathways that produce acetoin, butanediol and ethanol are indicated by grey arrows. Amino acid biosynthesis capability shown as filled in circles and incomplete pathways are indicated by amino acid single letter codes. Synthesis of vitamins is indicated by shaded in triangle, whereas incomplete pathways are not shaded. Putative enzymes are abbreviated as follows: Pyk, pyruvate kinase; PFOR, pyruvate synthase; PFL, formate C-acetyltransferase; CS, citrate synthase; MDH, malate dehydrogenase; PpdK, pyruvate orthophosphate dinkinase; PEPCK, phosphoenolpyruvate carboxykinase (ATP); Pta, phosphate acetyltransferase; AckA, acetate kinase; AdhE, acetaldehyde dehydrogenase/alcohol dehydrogenas; ALS, acetolactate synthase; DR, (R)-acetoin dehydrogenase; BDH, (R,R)-butanediol dehydrogenase; GS, glutamine synthetase; GOGAT, glutamate synthase; ANR, assimilatory nitrate reductase; OAD, Na+ - translocating oxaloacetate decarboxylase; GH, glycoside hydrolases; F_1F_0, predicated Na^+ -dependent ATPase; TCT, tricarboxylate transporter (TTT) family; AmT, Ammonia transporter; FNT, formate/nitrate transporter; RfT, riboflavin-H^+ symporter, ThiT, thiamine transporter; WXG100, WXG100 secretion system in the sagEsxA-like subfamily. Pathways are described in Table 3.

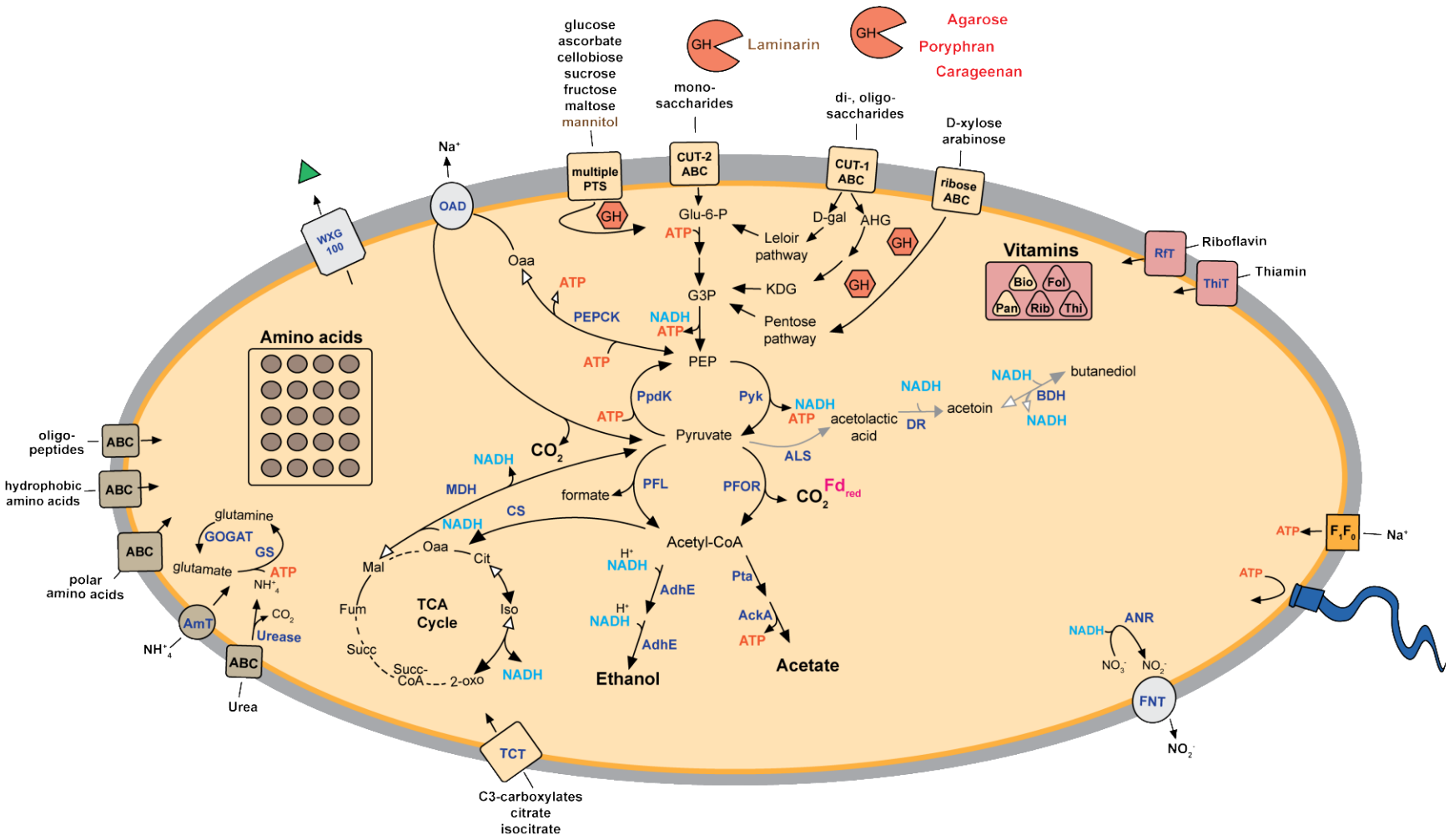

glucose
ascorbate
cellobiose
sucrose
fructose
maltose
mannitol
multiple PTS
GH
Laminarin
mono-saccharides
CUT-2 ABC
Agarose
Poryphran
Carageenan
di-, oligo-saccharides
CUT-1 ABC
D-xylose
arabinose
ribose ABC
Na+
OAD
WXG 100
Glu-6-P
ATP
Leloir pathway
D-gal
AHG
Oaa
KDG
G3P
Pentose pathway
PEPCK
NADH
PEP
Vitamins
Bio
Fol
Pan
Rib
Thi
RfT
Riboflavin
ThiT
Thiamin
Amino acids
PpdK
Pyk
NADH
butanediol
BDH
acetoin
DR
acetolactic acid
ALS
Pyruvate
oligo-peptides
ABC
hydrophobic amino acids
CO2
MDH
formate
PFL
PFOR
Fd red
CS
Acetyl-CoA
glutamine
GOGAT
GS
glutamate
polar amino acids
AmT
NH+4
Urease
Mal
Oaa
Cit
Fum
TCA Cycle
Iso
Succ
Succ-CoA
2-oxo
H+
AdhE
Ethanol
Pta
AckA
Acetate
F1F0
ANR
NO3-
NO2-
FNT
Urea
TCT
C3-carboxylates
citrate
isocitrate

Table 3. List of key functional genes to indicate presence or absence of putative function or pathway.

Metabolic process	Key metabolic gene(s) or complexes	Gene product
Synthesis or transport of vitamins/cofactors		
A/retinol	*blh*	EC:1.13.11.63 Beta-carotene 15,15'-dioxygenase
B1/thiamin	*thiC, thiD, thiE, thiG, thiL*	EC:4.1.99.17 Phosphomethylpyrimidine synthase; EC:2.7.4.7 Hydroxymethylpyrimidine phosphate kinase; EC:2.5.1.3 thiamine phosphate synthase;EC:2.8.1.10 thiazole synthase; EC:2.7.4.16 thiamine monophosphate kinase
B2/riboflavin	*ribAB, ribE, ribH*	EC:4.1.99.12 3,4-dihydroxy 2-butanone 4-phosphate synthase; EC:2.5.1.78 6,7-dimethyl-8-ribityllumazine synthase; EC:2.5.1.9 riboflavin synthase
B3/nicotinamide	*nadA, nadB, nadC*	EC:2.5.1.72 quinolinate synthase;EC:1.4.3.16 L-aspartate oxidase;EC:2.4.2.19 Nicotinate-nucleotide pyrophosphorylase [carboxylating]
B5/pantothenate	*panC*	EC:6.3.2.1 pantothenate synthetase
B6/pyridoxal 5'-phosphate (PLP)	*pdxS/T*	EC:4.3.3.6 PLP synthase complex
B7/biotin	*bioA, bioB, bioD, bioF, bioH*	EC:2.6.162 Adenosylmethionine--8-amino-7-oxononanoate transaminase; EC:2.8.1.6 Biotin synthase; EC:6.3.3.3 Dethiobiotin synthase; EC:2.3.1.47 8-amino-7-oxononanoate synthase; EC:3.1.1.85 Pimeloyl-[acyl-carrier protein] methyl ester esterase
B9/folate	*folQ, folB,folK, folP, folC/folA,folC*	EC:1.5.1.3 Dihydrofolate reductase; EC:6.3.2.17 Tetrahydrofolate synthase; EC:6.3.2.12 Dihydrofolate synthase; EC:2.1.1.45 Thymidylate synthase; EC:1.5.1.5 Methylenetetrahydrofolate dehydrogenase;EC:3.5.4.9 Methenyltetrahydrofolate cyclohydrolase
B10/para-aminobenzoic acid (pABA)	*pabAB, pabC*	EC:2.6.1.85 4-amino-4-deoxychorismate synthase; EC:4.1.3.38 4-amino-4-deoxychorismate lyase
B12/cobalamin	*cobUT, cobC, cobSV*	EC:2.4.2.21 nicotinate-nucleotide-dimethylbenzimidazole phosphoribosyltransferase; EC:3.1.3.73 alpha-ribazole phosphatase; EC:2.7.8.26 cobalamin-5'-phosphate synthase
CoA	*coaX, coaBC, coaD, coaE*	EC:2.7.1.33 pantothenate kinase; EC:6.3.2.5/4.1.1.36 4'-phosphopantothenoylcysteine decarboxylase and phosphopantothenoylcysteine synthetase; EC:2.7.7.3 pantetheine-phosphate adenylyltransferase; EC:2.7.1.24 dephospho-CoA kinase

Table 3. Continued.

Metabolic process	Key metabolic gene(s) or complexes	Gene product
Synthesis of amino acids		
alanine	*iscSA/SB, alr*	EC:2.8.1.7 Cysteine desulfurase; EC:5.1.1.1 alanine racemase
arginine	*argJ, argB, argC, argD, argF, pyrAA, pyrAB, argG, argH*	EC:2.3.1.35/2.3.1.1 glutamate N-acetyltransferase / amino-acid N-acetyltransferase, EC:2.7.2.8 acetyleglutamate kinase, EC:1.2.1.38 N-acetyl-gamma-glutamyl-phosphate reductase, EC:2.6.1.11 Acetylornithine transaminase, EC:2.1.3.3 ornithine carbamoyltransferase, EC:6.3.5.5 Carbamoyl-phosphate synthase, EC:6.3.4.5 argininosuccinate synthase argininosuccinate lyase, EC:4.3.2.1
asparagine	*asnB*	EC:6.3.5.4 Asparagine synthase
aspartate (aspartic acid)	*aspB*	EC:2.6.1.1 Aspartate transaminase
cysteine	*cysE, cysK*	EC:2.3.1.30 Serine O-acetyltransferase; EC:2.5.1.47 Cysteine synthase
glutamate (glutamic acid)	*gltA, gltB or gdhA*	EC:1.4.1.13 Glutamate synthase (NADPH); EC:1.4.1.14 Glutamate synthase (NADH; EC:1.4.1.4 Glutamate dehydrogenase (NADP(+))
glutamine	*glnA*	EC:6.3.1.2 Glutamine synthetase
glycine	*ItaE*	EC:4.1.2.5 L-threonine aldolase
histidine	*hisG, hisIE, hisA, hisF/H, hisB, hisC, hisK, hisD*	EC:2.4.2.17 ATP phosphoribosyltransferase, EC:3.6.1.31/3.5.4.19 phosphoribosyl-ATP pyrophosphatase /phosphoribosyl-AMP cyclohydrolase, EC:5.3.1.16 1-(5-phosphoribosyl)-5- ((5-phosphoribosylamino)methylideneamino)imidazole-4-carboxamide isomerase, EC:4.1.3.-/2.4.2.- imidazole glycerol phosphate synthase/glutamine amidotransferase, EC:4.2.1.19 Imidazoleglycerol-phosphate dehydratase, EC:2.6.1.9 Histidinol-phosphate transaminase, EC:3.1.3.15 Histidinol-phosphatase, EC:1.1.1.23 Histidinol dehydrogenase
isoleucine	*ilvA, ilvBGIHN, ilvC, ilvD, ilvE*	EC:4.3.1.19 Threonine ammonia-lyase, EC:2.2.1.6 Acetolactate synthase ,EC:1.1.1.86 Ketol-acid reductoisomerase, EC:4.2.1.9 Dihydroxy-acid dehydratase, EC:2.6.1.42 Branched-chain-amino-acid transaminase

Table 3. Continued.

Metabolic process	Key metabolic gene(s) or complexes	Gene product
Synthesis of amino acids (cont.)		
leucine	*leuA, lueC, leuD, leuB, ilvE*	EC:2.3.3.13 2-isopropylmalate synthase, EC:4.2.1.33 3-isopropylmalate dehydratase, EC:1.1.1.85 3-isopropylmalate dehydrogenase, EC:2.6.1.42 Branched-chain-amino-acid transaminase
lysine	*dapA, dap, dapL, dapF, lysA*	EC:4.2.1.52 dihydrodipicolinate synthase, EC:1.31.26 dihydrodipicolinate reductase, EC:2.6.1.83 LL-diaminopimelate aminotransferase, EC:5.1.1.7 Diaminopimelate epimerase, EC:4.1.1.20 Diaminopimelate decarboxylase
methionine	*metA, metY, patB, metH*	EC:2.3.1.46 Homoserine O-succinyltransferase, EC:2.5.1.49 O-acetylhomoserine aminocarboxypropyltransferase, EC:4.4.1.8 Cystathionine beta-lyase, EC:2.1.1.13 Methionine synthase
phenylalanine*	*pheA, ilvE/aspB*	EC:5.4.99.5/4.2.1.51 chorismate mutase/prephenate dehyratase, EC:2.6.1.42/2.6.1.1 Branched-chain amino acid aminotransferase/Aspartate transaminase
proline	*proB, proA, proC*	EC:2.7.2.11 Glutamate 5-kinase, EC:1.2.1.41 Glutamate-5-semialdehyde dehydrogenase, EC:1.5.1.2 Pyrroline-5-carboxylate reductase
serine	*sdaA/B*	EC:4.3.1.17 L-serine ammonia-lyase
threonine	*lysC, asd, hom, thrB, thrC*	EC:2.7.2.4 Aspartate kinase, EC:1.2.1.11 Aspartate-semialdehyde dehydrogenase, EC:1.1.1.3 Homoserine dehydrogenase, EC:2.7.1.39 Homoserine kinase, EC:4.2.3.1 Threonine synthase
tryosine*	*pheA, trA, ilvE/aspB*	EC:5.4.99.5 chorismate mutase, EC:1.3.1.12 phrephanate dehydratase ,EC:2.6.1.42/2.6.1.1 Branched-chain amino acid aminotransferase/Aspartate transaminase
tryptophan*	*trpE, trpD, trpF, trpC, trpAB*	EC:4.1.3.27 anthranilate synthase, EC:2.4.2.18 anthranilate phosphoribosyltransferase, EC:5.3.1.24 phosphoribosylanthranilate isomerase, EC:4.1.1.48 indole-3-glycerol phosphate synthase, EC:4.2.1.20 tryptophan synthase alpha and beta chains

Table 3. Continued.

Metabolic process	Key metabolic gene(s) or complexes	Gene product
Synthesis of amino acids (cont.)		
valine	*ilvBGIHN, ilvC, ilvD, ilvE*	EC:2.2.1.6 Acetolactate synthase ,EC:1.1.1.86 Ketol-acid reductoisomerase, EC:4.2.1.9 Dihydroxy-acid dehydratase, EC:2.6.1.42 Branched-chain-amino-acid transaminase
*requires chorismate	*aroF, aroB, aroQ, aroE, aroK/L, aroA, aroC*	EC:2.5.1.54 3-deoxy-7-phosphoheptulonate synthase, EC:4.2.3.4 3-dehydroquinate synthase, EC:4.2.1.10 3-dehydroquinate dehydratase, EC:1.1.1.25 Shikimate dehydrogenase, EC:2.7.1.71 Shikimate kinase, EC:2.5.1.19 3-phosphoshikimate 1-carboxyvinyltransferase, EC:4.2.3.5 Chorismate synthase
Central metabolism pathways and conversions		
Leloir pathway	*galM, galK, galT, galE*	EC:5.1.3.3 Galactose mutarotase or aldose 1-epimerase, EC:2.7.1.6 galactokinase, EC:2.7.7.12 Gal-1-P uridylyltransferase, EC:5.1.3.2 UDP-galactose 4-epimerase
D-galactonate catabolism by DeLey-Doudoroff pathway	AHGDH, AHGI	EC:1.2.1.92 AHG dehydrogenase (AHG: 3,6-anhydro-L-galactose), EC: 5.5.1.25 AHG cycloisomerase (AHGA: 3,6 anhydrogalactose)
Entner-Doudoroff pathway (Galactonate interconversions)	*kdgK/dgoK, kdgA/dgoA/ eda*	EC:2.7.1.58 2-dehydro-3-deoxygalactonate (KDGal) kinase, EC:4.1.2.14 2-dehydro-3-deoxy-D-gluconate-6-phosphate (KDPGal) aldolase
D-mannitol degradation	PTS, *mtlD*, *pfk*	D-mannitol PTS (IIC component), EC:1.1.1.17 D-mannitol 1-phosphate 5-dehydrogenase, EC:2.7.1.11 6-phosphofructokinase
Pentose phosphate pathway	*rpe, rpi, tktA/B, talA/B*	EC:5.1.3.1 ribulose phosphate 3-epimerase, EC:5.3.1.6 ribulose 5-phosphate isomerase, EC:2.2.1.1 transketolase subunit A/B, EC:2.2.1.2 transaldolase subunit A/B

A pivot toward carboxylate transport dependency in "*Ca.* Industriepulo"

All A1 and A2 genomes code for a complete glycolysis pathway, but the rest of central carbon metabolism of these epulos differs in a few key ways (Fig. 7& 8). First, *Industriepulo* codes for two pathways linking glycolysis and the TCA cycle via the oxidative decarboxylation of pyruvate to acetyl-CoA and CO_2: the *por* pathway that uses pyruvate ferrodoxin/flavodoxin oxidoreductase and the NADH generating pathway that uses the pyruvate dehydrogenase complex (PDHC). To date, all epulos contain the *por* pathway. Uniquely, *Industriepulo* contains all three genes required for the PDHC: pyruvate decarboxylase (E1) for the decarboxylation of pyruvate, followed by the subsequent acylation of the lipoyl prosthetic groups by dihydrolipoamide acetyltransferase (E2) and dihydrolipoamide dehydrogenase (E3). During the last acylation step, NADH is generated. Similar to *Bacillus subtilis,* the genes appear to be clustered in an operon as putative *pdhA* (E1α), *pdhB* (E1β), *pdhC* (E2) and *pdhD* (E3) (61). Thus, the PDHC has the potential to provide more energy to *Industriepulo* compared to using the oxidoreductase pathway.

Second, *Industriepulo* is missing the citrate synthase gene (*gltA*) which converts oxaloacetate and acetyl-CoA to citrate in the TCA cycle. Known alternative enzymes, like *Re*-citrate synthase, were not found. Instead, the genomes contain putative citrate:Na^+ symporter(s) and multiple aconitases (3-4), suggesting that they are capable of converting citrate effectively to the essential carbon-intermediate α-ketoglutarate once citrate is imported. However, it is still unclear what role acetyl-CoA plays in the TCA cycle or what consequences this modification has on the fitness of *Industriepulo*.

Third, *Al*_SAG9 uniquely contains all putative genes for the reductive TCA cycle (rTCA). Most reactions in the oxidative TCA cycle can go in the reverse direction, and those that cannot require additional enzymes to complete the process.

These genes code for citrate lyase (*citDEF*) which converts citrate to oxaloacetate and acetyl-CoA, fumarate reductase (*frdA*) that converts fumarate to succinate, and 2-oxo-glutarate ferredoxin oxidoreductase (*korABCD*) that converts succinyl-CoA to α-ketoglutarate. The ability to fix CO_2 from the gut environment has been shown to be a successful strategy for gut microbes like *Bacteroides hydrogenotrophica* (62). Unusual for anerobic environments, the gut has a high concentration of CO_2 generated from microbial metabolic process (respiration and fermentation) as well as from the host. Although the TCA is incomplete, *Al*_SAG9 may be able to fix CO_2 with succinyl-CoA as an initiating substrate. Notably, strain *Al*_SAG7 and *E. viviparus* clade B also contain a putative citrate lyase that may support a gluconogenetic lifestyle suggesting a conserved strategy among epulos. The presence of fumarate reductase may not be associated with rTCA because it is in the middle of a pathway where genes for upstream (malate dehydrogenase and fumarase) and downstream (succinyl-CoA synthetase) reactions are missing. A putative dicarboxylate:Na^+ symporter (NaDC-1) was detected in the genome and is similar to transporters previously shown to translocate TCA C4-intermediates like succinate (63, 64). Instead, fumarate reductase may be used as a primitive form of electron transport similar to what was proposed for *B. hydrogenotrophica* (62). Fumarate reductase has a higher electron potential than pyruvate, making it a more energetically favorable terminal electron acceptor. There are two forms of fumarate reductase that use different electron donors, either quinol or NADH. All reductases found in both *Industriepulo* genomes have high sequence identity to the quinol form and are often annotated as "flavocytochrome *c*" on NCBI databases. However, these are more likely the NADH form because gene neighborhoods do not contain the additional subunits associated with quinol complexes.

Fourth, *Industriepulo* contains a PEP carboxykinase (PEPCK; EC:4.1.1.49)

which reversibly catalyzes oxaloacetate to phosphophenolpyruvate (PEP), requiring ATP. Theoretically, the reverse of this pathway, which generates oxaloacetate, may also generate ATP from the high energy containing PEP phosphate. PEP carboxykinase is found in all characterized epulo genomes (A1, A2, and B) suggesting a versatile central carbon metabolism that can repurpose biosynthetic intermediates and use supplementary mechanisms for generating energy. For *Industriepulo*, PEPCK is an alternative pathway to produce oxaloacetate, a necessary intermediate for amino acid synthesis.

Finally, citrate lyase reactions can also produce oxaloacetate and acetate from imported citrate. *Industriepulo* contains genes that encode for a citrate-specific transporter (CitM) and a citrate lyase complex (CitDEFCXG). Oxaloacetate can then be converted to pyruvate through a Na^+ translocating oxaloacetate decarboxylase (Oad). Altogether, these data suggest that epulos have a more complex carbon and energy program than previously appreciated.

Diverse energy and redox balance strategies in "*Ca.* Industriepulo"

With their predicted fermentative metabolism it had been suspected that epulos rely on substrate-level phosphorylation to generate ATP. Although we found multiple genes annotated as flavocytochromes (i.e. fumarate reductase/succinate dehydrogenase *frdA, sdhA),* we suspect that they are not cytochromes but flavoproteins because they lack the cytochrome *c* domain and are not associated with operons that contain genes that encode cytochromes. Many of these putative ferredoxins possess either a signal peptide or they are located in gene clusters coding secretory proteins. For example, both A1 genomes contain the *frdA* gene in a putative operon with a gene for a 4Fe-4S ferredoxin and a gene for the Tat proofreading chaperone TorD. Secretion of ferredoxin may help regulate the redox environment to promote enegetically favorable

conditions for desired metabolic reactions. The secretion of flavoproteins may also have a similar function, suggesting that the direct extracellular space of epulos may play a significant role in metabolic and energetic processes similar to the periplasm of Gram negative bacteria.

Additionally, we found genes that encode the Rnf complex (*rnfCDGEAB*), an alternative ion-motive electron transport chain. The Rnf complex is composed of membrane-bound iron-sulfur clusters and flavins that couple electron transfer from reduced ferredoxin to NAD^+ with the translocation of sodium ions. The Rnf complex has been previously described in *Acetobacterium woodii* and has been discovered in a number of clostridial genomes including *C. tetani*, *C. kluyveri*, *C. difficile*, *C. botulinum* and *Cellulosilyticum lentocellum* (65). *Industriepulo* contains genes for both the F-type and V-type ATPases, suggesting that they are capable of producing ATP via this alternative electron transport phosphorylation system. In *Fusobacterium nucleatum*, the F_1F_o-type ATPase is driven by Na^+ rather than H^+ (66). We hypothesize that *Industriepulo* relies on a gradient of sodium ions (sodium motive force, SMF) to generate energy using the Rnf complex and a Na^+-driven F-type ATPase. Using a SMF instead of a proton motive force (PMF) would be a useful strategy to reduce the stress of acidification generated through fermentation.

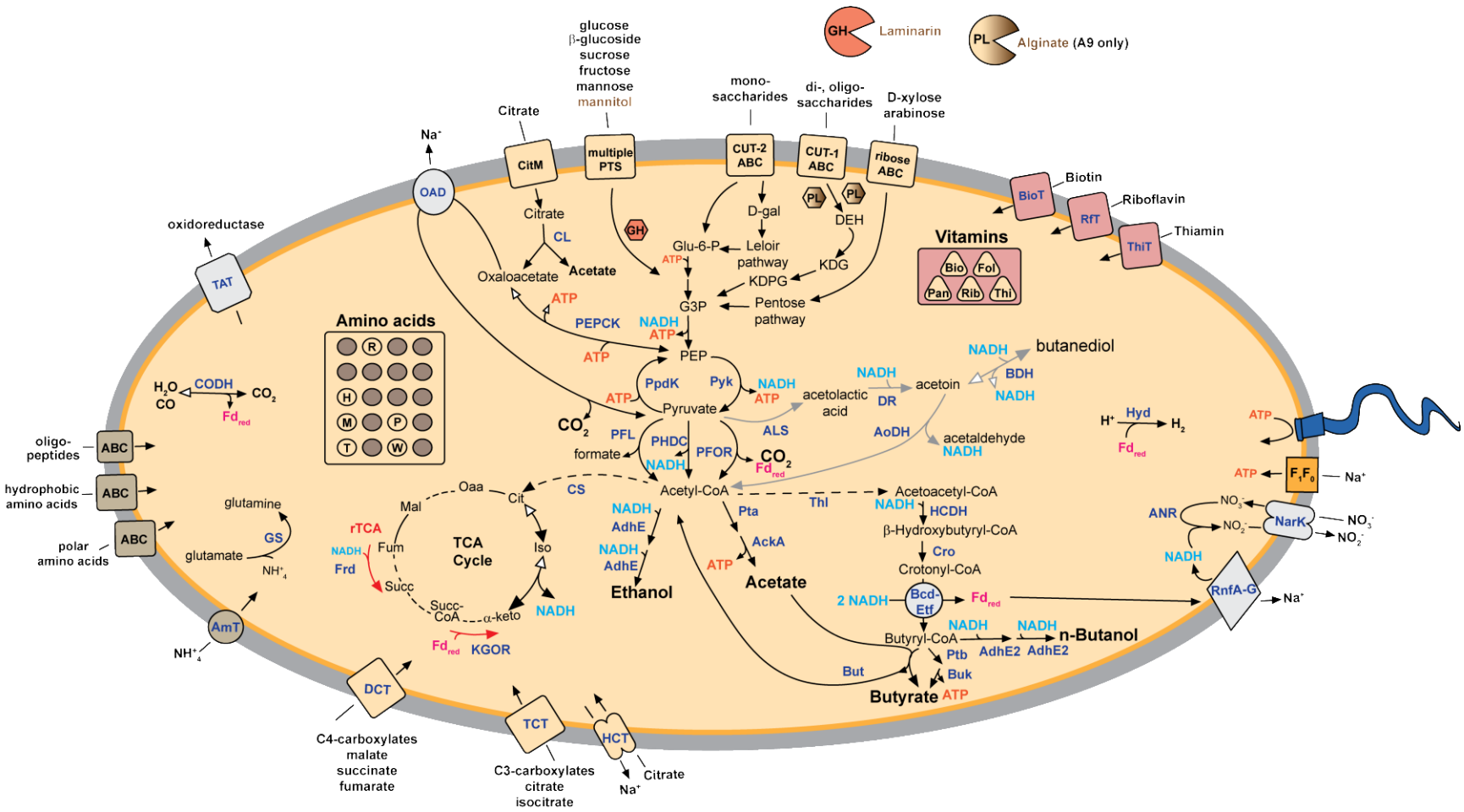
Furthermore, putative flavin-based electron-bifurcating (FBEB) enzyme complexes were coded for in the genomes of *Industriepulo*, providing a source of reduced ferredoxins. FBEB has recently been described as a third mode of energy conservation (67) along with substrate-level phosphorylation (SLP) and electron-transport-mediated phosphorylation (ETP). FBEB complexes are located in the cytoplasm and simulatenously mediate electron flow from individual electron donors (i.e. NADH) to multiple electron acceptors where one of these acceptors is a ferredoxin. FAD molelcules can be reduced, one electron at a time, forming stable

semiquinone flavin radicals (FADH) until they are completely reduced ($FADH_2$) by a second electron. Reduction of ferredoxin is an endergonic process (E_0' = -410 mV) and requires an electron donor with a more negative reduction potential than NADH (E_0' = -320 mV). However, by coupling this process with an exergonic reaction, ferredoxin can be reduced by NADH. For instance, *Industriepulo* contains in the same operon as FBEB genes (*etfBC*), the gene for butyryl-CoA dehydrogenase (*bcdA*) which coverts crotonyl-CoA to butyryl-CoA. Together, BcdA and EtfBC form a complex that has been demonstrated in various *Clostridium* spp. to catalyze the production of butyryl-CoA (E_0' = -10 mV) and reduction of ferredoxin (68, 69). In *Acidaminococcus fermentans*, the first flavoxin step has a reduction potential of -60 mV and the second step has E_0'= -430 mV. The more positive potential FAD reduces crotonyl-CoA and the more negative FAD reduces ferredoxin. As a result, the free energy retained in reduced ferredoxin is released and can be used to generate ATP via the Rnf complex/ATPase. Alternatively, the reduced ferredoxin can be used to catalyze the [Fe-Fe] hydrogenase (*hydA)* to convert protons released from acidogenic processes to hydrogen.

Genes for an NADH-dependent nitrate reductase and a nitrate:nitrite antiporter (*narK*) are present in the *Industriepulo* genomes. The nitrate reductase is predicted to be involved in a dissimilatory process (Nar) analogous to respiration but contains a diaphorase subunit rather than the membrane-bound cytochrome *b* subunit commonly found in assimilatory Nas enzymes. The gene cluster codes catalytic molybdoenzyme-containing subunit (*narG-like*), an FeS cluster (*narH*-like) and a NAD-dependent diaphorase subunit (*nasB*-like found in *B. subtilis*). The NarK transporter is not located in the *nar* cluster. We hypothesize that the NADH generated from the Rnf complex can be coupled to nitrate reduction as an alternative form of anaerobic respiration without the need for cytochromes or quinones (Fig. 8). Both *E.*

bomborense and *E. viviparus* contain NADH-dependent nitrate reductases although they do not contain genes for the Rnf complex. It has long been appreciated that respiration complexes have a modular composition. NADH-dependent reductases like Nar might resemble a primitive form of anaerobic respiration.

Figure 8. Model of the carbohydrate and energy metabolism of "*Ca.* Industriepulo choatii" clade A1. Pathways were deduced from genome annotations and KEGG/Microcyc references. End products from fermentation are indicated in bold. Incomplete pathways are shown as dotted lines and reversible reactions contain white arrows. Buffering pathways that produce acetoin, butanediol and ethanol are indicated by grey arrows. Amino acid biosynthesis capability shown as filled in circles and incomplete pathways are indicated by amino acid single letter abbreviations. Synthesis of vitamins is indicated by shaded in triangle, whereas incomplete pathways are not shaded. Putative enzymes are abbreviated as follows: Pyk = pyruvate kinase; PFOR, pyruvate synthase; PDHC, pyruvate dehydrogenase complex; PFL, formate C-acetyltransferase; CS, citrate synthase; Frd, fumarate reductase; KGOR, 2-oxo-glutarate ferrodoxin oxidoreductase; PpdK, pyruvate, orthophosphate dinkinase; PEPCK, phosphoenolpyruvate carboxykinase (ATP); Pta = phosphate acetyltransferase; AckA, acetate kinase; AdhE, acetaldehyde dehydrogenase/alcohol dehydrogenase; AdhE2, putative butanol dehydrogenase; CL, ATP-requiring citrate lyase complex; ALS, acetolactate synthase; DR, (R/S)-acetoin dehydrogenase; BDH, (R,R/S,S)-butanediol dehydrogenase; AoDH, acetoin dehydrogenase complex; ThI, acetoacetyl-CoA thiloase; HCDH, 3-hydroxyacyl-CoA dehydrogenase; Cro, enoyl-CoA hydratase; Bcd-Etf, butyryl-CoA dehydrogenase electron-transferring flavoprotein complex; Ptb, phosphate butyryltransterase; Buk, butyrate kinase; But, butyryl-CoA:acetate CoA transferase; ANR, assimilatory nitrate reductase; CODH, anaerobic carbon-monoxide dehydrogenase; Hyd, ferredoxin hydrogenase; GS, glutamine synthetase; OAD, Na+ - translocating oxaloacetate decarboxylase; rTCA, reverse TCA cycle; GH, glycoside hydrolases; F_1F_0, predicated Na^+ -dependent ATPase; CitM, citrate-H^+ symporter, HCT, 2-hydroxycarboxylate transporter; TCT, tricarboxylate transporter (TTT) family; DCT, tripartite ATP-independent periplasmic transporter (TRAP-T) family; DAACS, Dicarboxylate/amino acid:cation symporter; AmT, ammonia transporter; TAT, twin-arginine translocation pathway; BioT, biotin transporter; RfT riboflavin transporter; ThiT, thiaminase transporter, NarK; nitrate/nitrite antiporter; RnfA-G, Rnf complex. Pathways are described in Table 3.


GH Laminarin
PL Alginate (A9 only)
glucose
β-glucoside
sucrose
fructose
mannose
mannitol
mono-saccharides
di-, oligo-saccharides
D-xylose
arabinose
Citrate
Na+
CitM
multiple PTS
CUT-2 ABC
CUT-1 ABC
ribose ABC
OAD
BioT
Biotin
RfT
Riboflavin
ThiT
Thiamin
Vitamins
Bio
Fol
Pan
Rib
Thi
oxidoreductase
TAT
Amino acids
R
H
M
P
T
W
Citrate
CL
Oxaloacetate
Acetate
GH
Glu-6-P
ATP
D-gal
Leloir pathway
KDPG
Pentose pathway
DEH
KDG
PL
G3P
ATP
PEPCK
NADH
ATP
PEP
ATP
PpdK
Pyk
NADH
ATP
Pyruvate
acetolactic acid
NADH
DR
acetoin
NADH
BDH
NADH
butanediol
AoDH
acetaldehyde
NADH
H2O
CO
CODH
CO2
Fdred
CO2
PFL
PHDC
PFOR
formate
NADH
CO2
Fdred
ALS
H+
Hyd
H2
Fdred
ATP
F1F0
Na+
oligo-peptides
hydrophobic amino acids
polar amino acids
ABC
ABC
ABC
glutamine
GS
glutamate
NH+4
AmT
NH+4
CS
Oaa
Cit
Mal
rTCA
NADH
Frd
Fum
TCA Cycle
Iso
Succ
Succ-CoA
α-keto
NADH
Fdred
KGOR
Acetyl-CoA
Thl
Acetoacetyl-CoA
NADH
HCDH
β-Hydroxybutyryl-CoA
Cro
Crotonyl-CoA
NADH
AdhE
NADH
AdhE
Ethanol
Pta
AckA
ATP
Acetate
2 NADH
Bcd-Etf
Fdred
NADH
NADH
Butyryl-CoA
AdhE2
AdhE2
n-Butanol
Ptb
Buk
But
Butyrate
ATP
ANR
NO3-
NO2-
NarK
NO3-
NO2-
NADH
RnfA-G
Na+
DCT
C4-carboxylates
malate
succinate
fumarate
TCT
C3-carboxylates
citrate
isocitrate
HCT
Na+
Citrate

Acidogenesis and Solventogenesis in epulos

Clostridia have complex and diverse cellular life cycles. Understanding the molecular mechanisms that regulate their transitions has been difficult to parse. Clostridia are obligate anaerobes that are challenging to culture, they have simple taxonomic characteristics and yet are phylogenetically diverse, and few can be genetically manipulated. In general, Clostridia are known for their ability to produce endospores goverened by the master regulator Spo0A. However, in tribute to this genetically and functionally diverse group, not all Clostridia that contain Spo0A are able to form endospores. In addition to sporulation, Spo0A activation regulates a life stage transition that results in a distinct cellular morhphology, the "clostridial-form"(70–72). In early exponential phase, most clostridia thrive as fermenters, a stage referred to as acidogenesis, in which the pH of their environment is reduced as the byproducts of their metabolism accumulate. In early stationary phase, their metabolism shifts to reassimilate the organic acids produced by fermentation (e.g. acetate and pyruvate)(73) and cease production of these acidic products. Cells transition to a solventogenic phase where proton-consuming reactions generate alcohols and neutral-carbon molecules like acetoin. The accumulation of acetoin and other neutral storage molecules within the cell correlates with the conversion to the clostridial-form (or cigar-shaped) cell shape. These energy-storing granules can sustain some cells during stationary phase while others transition to endosporulation. Remarkably, epulos appear to have maintained many of the genes that are characteristic of other Clostridia (Fig. 7 & 8).

As acidogens (acetate and butyrate)

Industriepulo is predicted to produce acetate and butyrate during acidogenesis. This genus contains two ways to produce acetate similar to *E. viviparus*. The first is through fermentation from pyruvate in the two step process (*pta-acka*) that generates

ATP. The second pathway generates acetate from imported citrate using the ATP-citrate lyase complex described above.

The butyrate synthesis pathway in *Industriepulo* resembles the acetyl-coenzyme A (CoA) pathway used by other Clostridia such as *C. acetobutylicum* (74, 75). Briefly, acetyl-CoA is converted to the intermediate crotonyl-CoA after a series of reactions. Crotonyl-CoA is then transformed to butyrl-CoA by a butyryl-CoA dehydrogenase electron-transferring flavoprotein complex (Bcd-Etfαβ). The specifics of this complex will be explained in a later section (energy). Finally, butyrl-CoA is converted to butyrate by either the more common butyryl-CoA:acetate CoA transferase (But) or butyrl-CoA is phosphorylated by phosphate butyryltransferase (Ptb) and then transformed to butyrate by butyrate kinase (Buk), forming ATP. *Industriepulo* contains all the genes except acetoacetyl-CoA thiolase (*thi*; EC:2.3.1.9) that initiates the pathway by converting acetyl-CoA to acetoacetyl-CoA and it is missing the *ptb* gene. *Industriepulo* may produce acetoacetyl-CoA through a novel pathway or import the substrate from the environment. The *ptb* gene is used for one of two alternative pathways for the final productin of butyrate, therefore its absence may suggest that is an ancestral vestage that is no longer required. Conversion of butyryl-CoA to butrate using the But pathway re-assimilates acetate to produce more acetyl-CoA. This pathway can be used to produce butyrate more efficiently, thus leading to a cyclic pathway that ultimately generates more ATP by coupling the ferredoxin reducing Bcd-Etf complex with the Rnf complex.

As solventogens (ethanol and butanol but not acetone)

Industriepulo have the potential to produce ethanol and butanol. They have the gene for a multifunctional alcohol dehydrogenase (*adhE*) that converts acetyl-CoA to acetaladehyde and then to ethanol. This gene or additional putative alcohol dehydrogenase genes (*bdhA* coding for AdhE2) may also be used to convert butyryl-

CoA to butanol. Although both genera are predicted to produce acetoin and butanediol, *Industriepulo* contains genes to produce all stereoisomers – (R,R), (S,S), and *meso*. The roles of these different iso-forms in the cell are not well understood, but all forms are of industrial importance for the aesthetic qualities in food products and as precursors for rubber and plastics. The formation of butanediol (via BDH) from acetoin is a reversible reaction. Acetoin can further be degraded to generate acetyl-CoA (via AcoAB) and ethanol. Overall, these pathways assimilate protons and oxidize NADH to buffer and maintain redox balance in the cell.

Although solventogenesis is correlated with changes to cell shape, it is still unclear how these solventogenic processes contribute to these changes. It is likely that they are part of a regulatory network rather than a direct cause for the morphological appearance. Futher studies will be needed to determine the contribution, if any, solventogenesis contributes to the clostridial-form in epulos.

Table 4. Transcript mapping results per sample. Reads were mapped to clade A2 representative genome *Al*_SAG11 and clade A1 genomes *Al*_SAG7 and *Al*_SAG9.

Sample	Total raw reads	*Al*_SAG11	*Al*_SAG7	*Al*_SAG9
G155	32,881,368	27,019,712	2,480,051	2,297,320
G156	25,312,036	15,185,777	4,548,596	4,014,542
G182	27,640,827	12,813,618	2,841,829	2,984,694
G186	26,838,298	17,575,951	5,133,326	4,419,491
G174	25,204,825	5,976,004	2,067,944	7,090,308
G187	30,722,343	12,522,364	8,410,845	1,938,238

Transcriptional profiles reveal time-dependent synchrony of symbionts

Cell development of epulos is often syncrhonized with time within surgeonfish hosts. To explore the gene pathways that are associated with predictable temporal oscillations, raw transcripts sampled from *A. lineatus* gut contents at duplicate time points spanning approximately nine hours of daylight (MM, AF, LAF; Table 1) were

mapped to a representative of clade A2 (*Al*_SAG11, which is nearly identical to *Al*_SAG6) and both representatives of clade A1 (*Al*_SAG7, *Al*_SAG9). All samples met the minimum mapping coverage threshold of ~2 million reads per genome (Table 4). Most samples had higher mapping coverage for clade A2 than clade A1, which correlates with relative abundances observed in clone library surveys. Differential expression analysis between the furthest time intervals MM and LAF yielded the highest number of genes (Fig. 9) (log2fold change > 1.5; $P_{adj.} < 0.05$) for all genomes. Duplicate samples per time interval were consistent across transcripts and showed a clear transcriptional shift between MM between LAF, and transitional levels during AF.

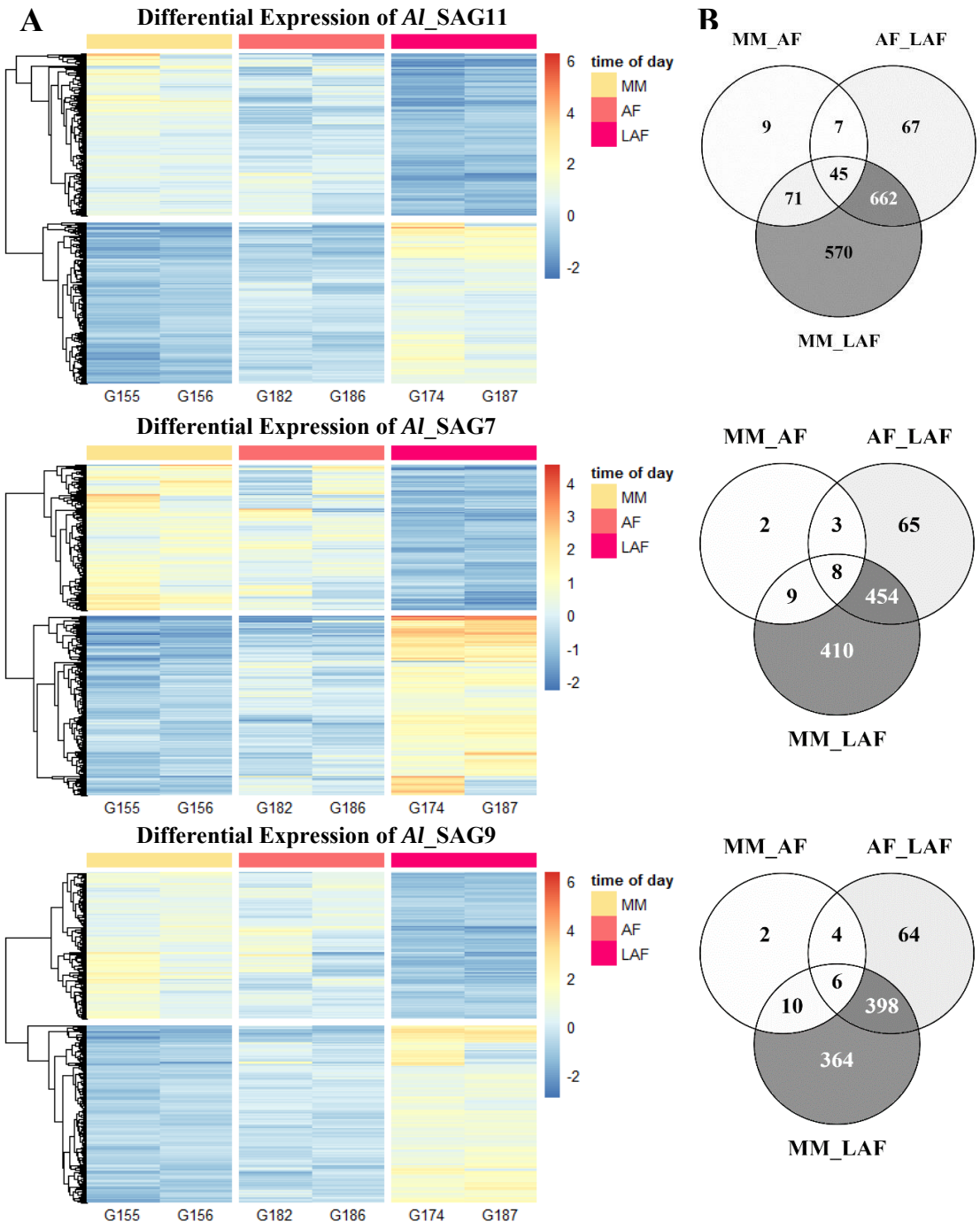


Figure 9. Differentially expressed genes of symbionts individually mapped to genomes *Al*_SAG11, *Al*_SAG7, and *Al*_SAG9. A) Heatmap of DESeq2 rlog transformed (log2fold change > 1.5; $P_{adj.} < 0.05$) values comparing symbionts collected during the middle morning (MM), afternoon (AF), and late afternoon (LAF). The number of significant DE genes for each genome were 1348, 881, and 778, respectively. B) Venn diagrams showing the number of DE genes per time of day comparison: using Venny v2.1.0.

Metabolic shift from active extracellular uptake and fermentation to solventogenesis

Many genes related to carbon metabolism were differentially expressed over time for all genomes (Fig. 10). To observe global trends, genes were categorized by general functions: organic carbon-specific membrane transport, complex polysaccharide degradation, and conversion/catabolism. Membrane transport-related genes (CUT1, CUT2, PTS, and ribose-ABC transporters) were highly expressed in the MM and were significantly reduced in LAF for all genomes. Clade A2 complex polysaccharide degradation genes (agarases, porphyranases, and lambda-carrageenases) were highly expressed at MM and reduced at LAF. Whereas intracellular glycoside hydrolases (GH31, GH43, and GH117) showed the reverse trend: high expression at LAF and reduced expression at MM. This inverse correlation indicates a shift from extracellular degradation to intracellular metabolism as the day progressed. Although *Al*_SAG9 contains polysaccharide lyases, none were significantly differentially expressed between these time intervals. With the exception of sample G182, the PLs were minimally expressed (<10 RPKM) suggesting that the presence of brown algae varies among samples and is likely a rare resource for epulos in *A. lineatus*.

The conversion/catabolic genes differentially expressed between MM and LAF shift from pyruvate catabolism to carbon storage/solventogenesis. In clade A2, gene expression for early transformations in the pentose phosphate pathways (e.g. *rpe*, *rpi*) are lower at LAF and genes for processes that occur late in the pathway are higher (e.g. *tkt*). Pyruvate is transitioning toward reassimilation observed as an increase of *ppdK* gene expressionat LAF. The shift towards carbon storage is more apparent in clade A1 wherein expression of putative granulose formation genes (*glgC, glgD*, and *glgP*) is higher in LAF. These genes are homologous to *C. acetobutylicum* granulose

genes that are expressed duringearly stationary phase (76).

Energy-related genes that were differentially expressed over time fell into the following categories: fermentation, redox balance, ATP synthetase subunits, carboxylate transport and H^+/Na^+ motive force formation (Fig. 11). Genes for ATP synthetases (V-type and F_1F_o-type), carboxylate transport and generation of H^+/Na^+ motive force (including oxaloacetate decarboxylase and the Rnf complex) were highly expressed at MM and reduced by LAF. Fermentation pathways that yield acetate (acetate kinase) and solvents (alcohol dehydrogenases) are uncoupled in epulos similar to other clostridia. Genes for acidogenesis are highly expressed during MM and expression patterns shift to solventogenesis at LAF. Redox balance genes that encode rubrerythrin, thioredoxin, and thioredoxin reductase are upregulated in LAF.

Epulos are likely responding to the feeding/fasting cycles of the host Transcriptional profiles of both carbon and energy metabolism correlate with the diurnal feeding habits of the host. *A. lineatus* commences feeding at dawn and continues to feed during the day (56). Peak expression of complex polysaccharide degradation genes and carbon-specific transport genes suggest epulos are responding to the influx of dietary inputs. Food accumulates in the digestive tract (stomach, intestine, and hindgut) of *A.lineatus* over the day, wherein the largest increase occurs in the intestines (56). The metabolic and shift towards solventogenesis and increase in redox balance gene expression indicate that metabolic byproducts accumulate in the intestine during the day.

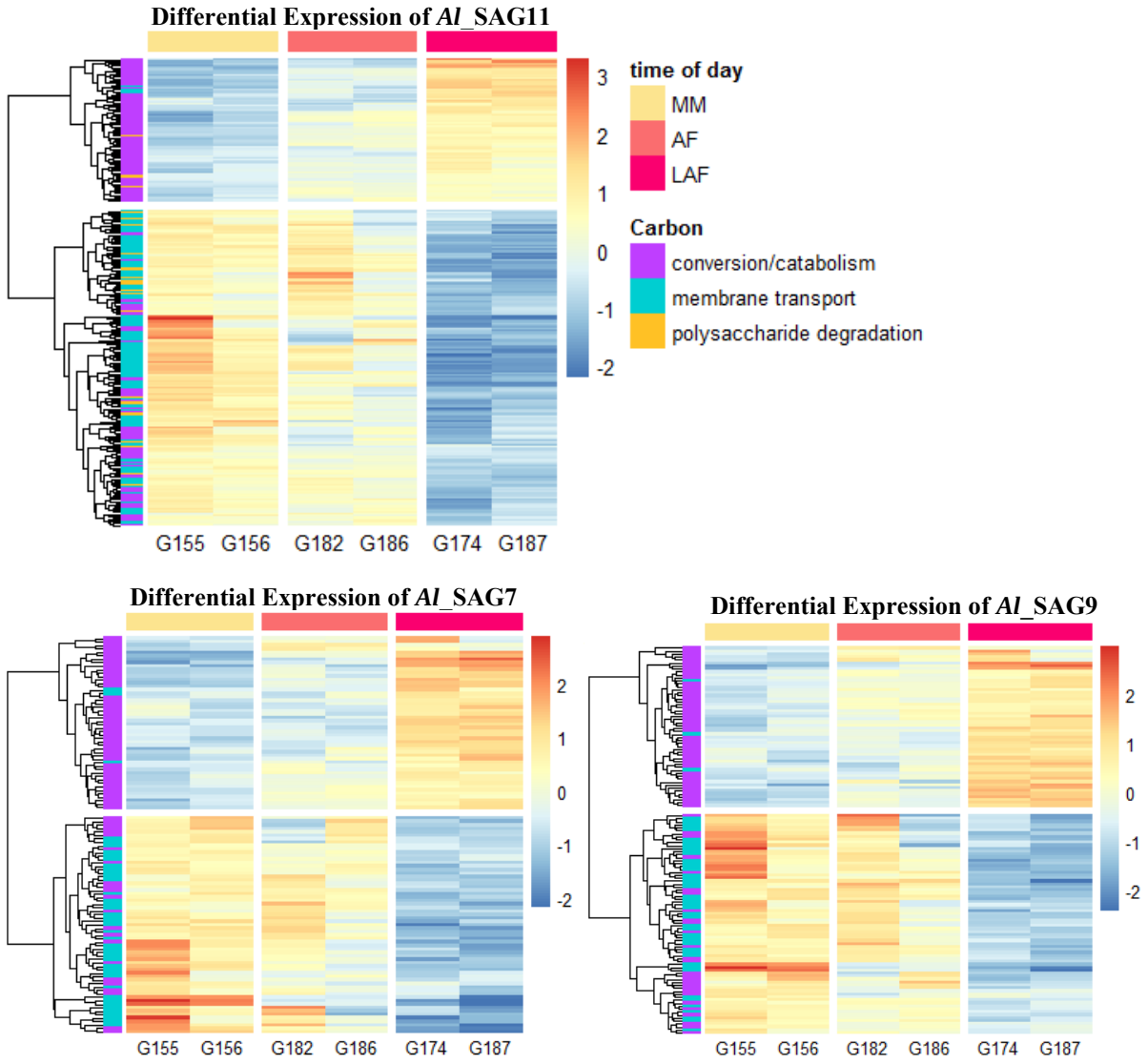


Figure 10. Heatmaps of carbon metabolism related genes (rows) that were differentially expressed between MM and LAF time intervals (columns) for each genome. Rows are annotated by general function: conversion/catabolism (e.g. TCA cycle, glycolysis, pentose phosphate pathway), membrane transport (e.g. carbon-substrate specific transporters), and polysaccharide degradation (e.g. glycoside hydrolases and polysaccharide lyases).

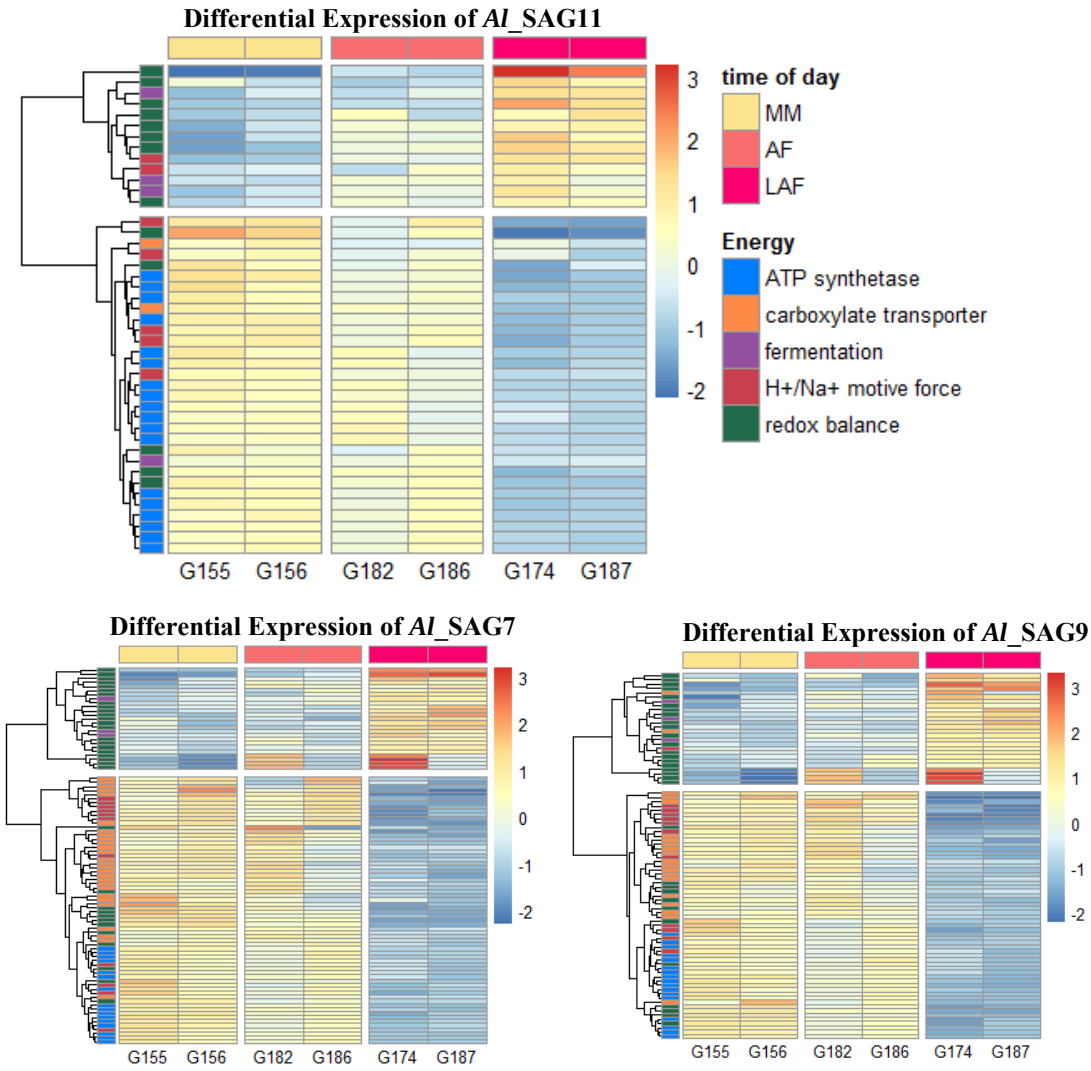


Figure 11. Heatmaps of energy related genes (rows) that were differentially expressed between MM and LAF time intervals (columns) for each genome. Rows are annotated by general function: ATP synthetases (V-type and F_1F_0-type), carboxylate transporters (citrate-H^+ symporter, TCT, DCT), fermentation (eg. Alcohol dehydrogenases, acetate kinase), H^+/Na^+ motive force (eg. oxaloacetate decarboxylase, Rnf complex), and redox balance (e.g. rubryethrin, thioredoxin, formate/nitrite transporter, pyruvate dehydrogenase complex).

Uncoupled onset of solventogenesis and sporulation in non-spore-forming epulos

Although neither clade A1 nor A2 have been observed to produce phase-bright spores, they have maintained some sporulation genes required for intracellular offspring formation, similar to clade B (77). The sporulation cascade in clostridia generally begins with the phosphorylation of Spo0A and activation of σ^H, followed by the activation of four sporulation-specific sigma factors σ^F, σ^E, σ^G, and σ^K. With the exception of σ^K, these sigma factors are activated in order (from σ^F, σ^E to σ^G) and are localized in either the forespore (σ^F and σ^G) or mother cell (σ^E and σ^K). Compared to *B. subtilis*, the role of sigma K varies among clostridia species by regulating genes during early stages of sporulation as well as late stages (78). Differential expression analyses from both epulo clades suggest that sporulation cascade follows the same order as described for clostridia (Fig. 12). Genes under the control of σ^H and σ^E are highly expressed earlier than genes in the σ^F, σ^G, and σ^K regulons. Differential expression of Spo0A was not apparent in this time interval and may have peak expression at an earlier time point or likely post-transcriptional or post-translational regulation of the transcription factor is more important.

In clostridia, solventogenesis and onset of sporulation occur simultaneously (79, 80). These genes are regulated by σ^H. The onset of sporulation correlates with the upregulation of genes under σ^H: *spoIIAA*, s*poIIAB* and *sigF*. The post-translational activation of the early forespore sigma factor also includes SpoIIE, which ultimately releases SigF from SpoIIAB when polar division is complete. Both clades contain the *spoIIAA-spoIIAB-sigF* operon and *spoIIE*, however only clade A1 contains the σ^H gene. Regulation of this operon in clade A2 is still unclear. In clade A1, *spoIIAA-spoIIAB-sigF* operon and *spoIIE* are highly expressed at MM and significantly reduced in LAF. In clade A2, only *spoIIAB* gene expressionis peaks at MM. Expression of these genes is markedly earlier than in clostridia, suggesting that epulos

have uncoupled solventogenesis and the sporulation-derived process of intracellular offspring formation. Early expression *spoIIE* was previously observed in type B using qPCR (81), further validating this phenomena in non-endospore-forming epulos.

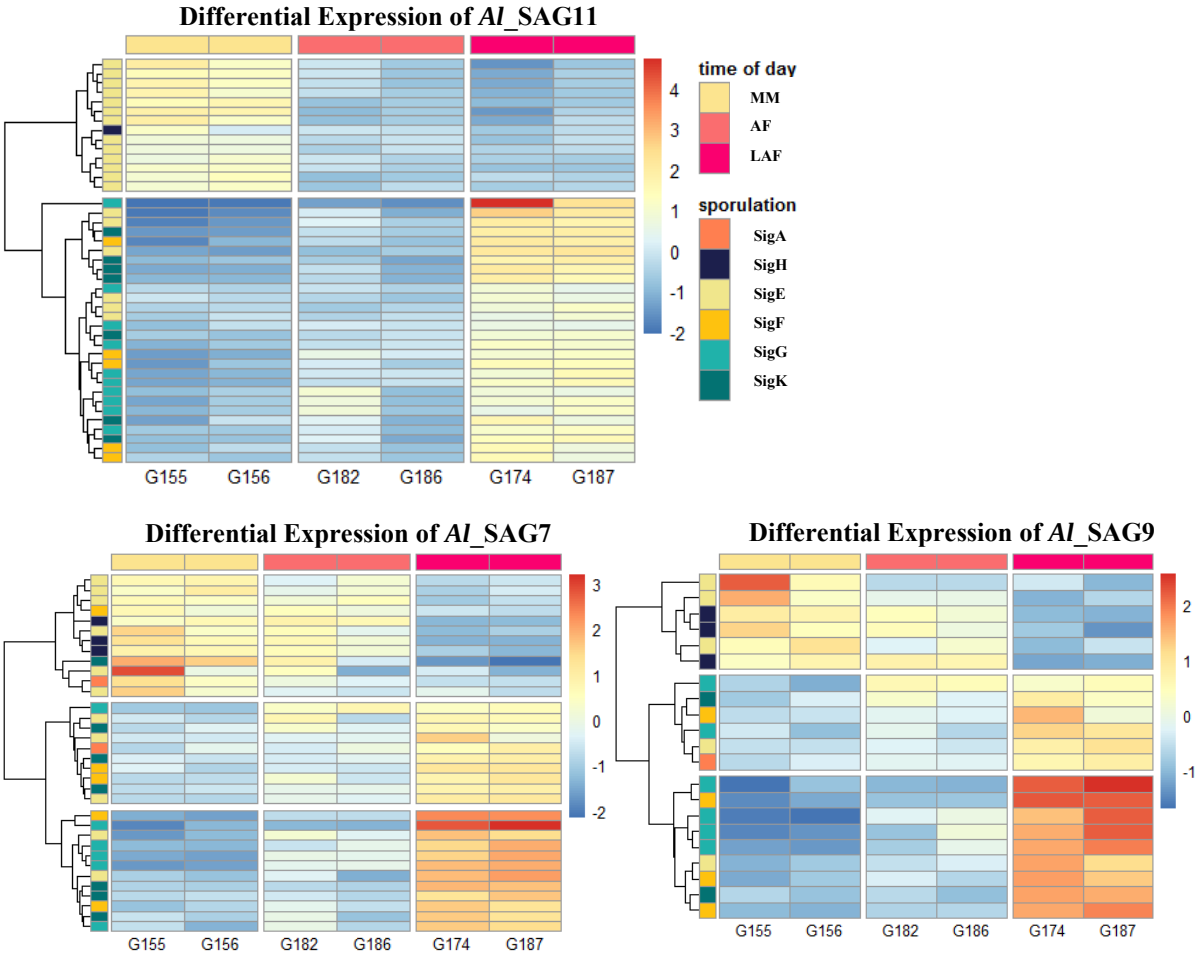


Figure 12. Heatmaps of sporulation gene homologs (rows) that were differentially expressed between MM and LAF time intervals (columns) for each genome. Rows are annotated by key sigma factors: SigA, SigH, forespore-specific SigF & Sig G and mother cell-specific SigE and SigK.

Secretome gene expression corresponds with time-dependent metabolic shifts

Both *E. bomborense* and *Industriepulo* contain genes for the complete Sec pathway, T2SS/pseudopilin (*gspE-pulF*), but have different transport systems for small folded proteins (WXG100 or Tat). *E. bomborense* encodes the WXG100 transport system that targets small proteins (~100 aa long) with a conserved WXG motif in the center of the protein (82). The genes that encode these small proteins are usually clustered near the genes for the WXG100 system. The function of these proteins (e.g. *yukE*) is unknown but speculated to be involved in DNA uptake or stress resistance (83, 84). Industriepulo encodes the Tat secretion system (*tatAC*). Tat-translocated proteins are often associated with maintaining redox balance in the periplasm of Gram-negative bacteria and in the cell wall or extracellular space of Gram-positive bacteria. Both WXG100 and Tat gene expression peaks at MM, which corresponds to high expression of genes for extracellular metabolism. Conversely, the Sec pathway gene expression peaks during LAF, correlating with solventogenesis. In toxin forming clostridia, secretion of toxins and other proteins use the Sec pathway during the solventogenic phase (80, 85).

Nitrogen assimilation and potential for amino acid cross-feeding between symbionts

The low protein composition of marine microalgae, including red algae, (58, 59) indicates that surgeonfish require additional sources of protein to supplement their nutritional demands. Fermentation of protein may not play a major role in nitrogen metabolism of *A. lineatus* due to the low levels of SCFA isovalerate present in the gut and blood (21). However, studies of other teleost fish confirm the presence of both ammonia and urea in the gut lumen (86), and likely provide a nitrogen source for surgeonfish. Genome predictions indicate that ammonia and urea assimilation vary

between epulo clades. All epulos contain ammonia transporters and can incorporate ammonia onto glutamate using glutamine synthase (*glnA*). All have amino-acid specific-ABC transporters while *Industriepulo* also contains a glutamate/Na^+ symporter. *E. bomborense* contains genes for urea transporters and ureases (*ureABC*) to generate ammonia whereas *Industriepulo* lacks this function (Fig. 7 & 8).

E. bomborense is predicted to be a prototroph while *Industriepulo* is auxotrophic for methionine, arginine, tyrosine, tryptophan, histidine, and proline. Most of the amino acid synthesis genes lacking in *Industriepulo* are for those amino acids that are more costly to produce (87); a characteristic that selects for cross-feeding among auxotrophs (88). Both clades contained ~70 genes related to organic nitrogen metabolism that were differentially expressed between EM and LAF (Fig. 13). Amino acid biosynthesis and amino acid transport genes were highly expressed in the morning and dropped by LAF. Although both clades contain polar (PAAT) and hydrophobic (HAAT) amino acid transporters, only *Industriepulo* showed peak gene expression for PAAT and HAAT during EM. These transporters may be indicative of cross-feeding of histidine, proline, and tyrosine between these two symbionts.

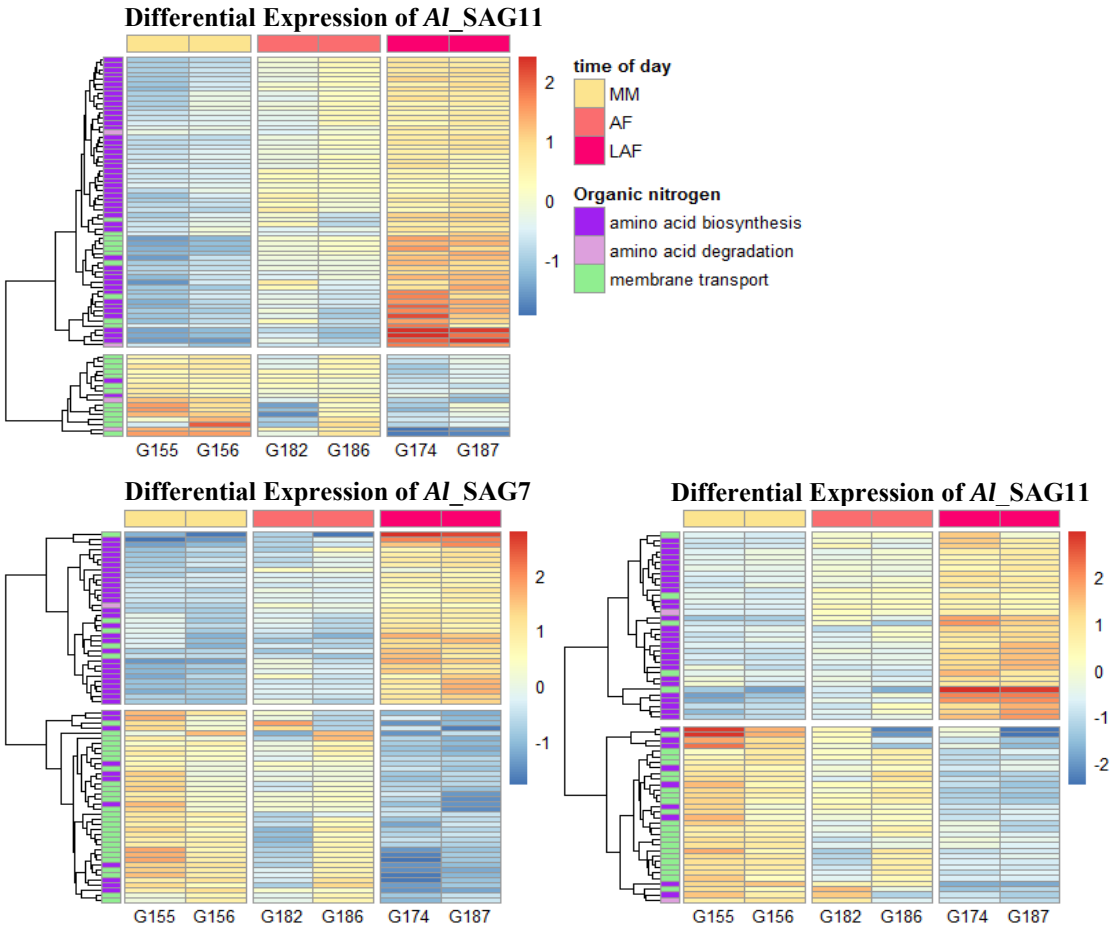


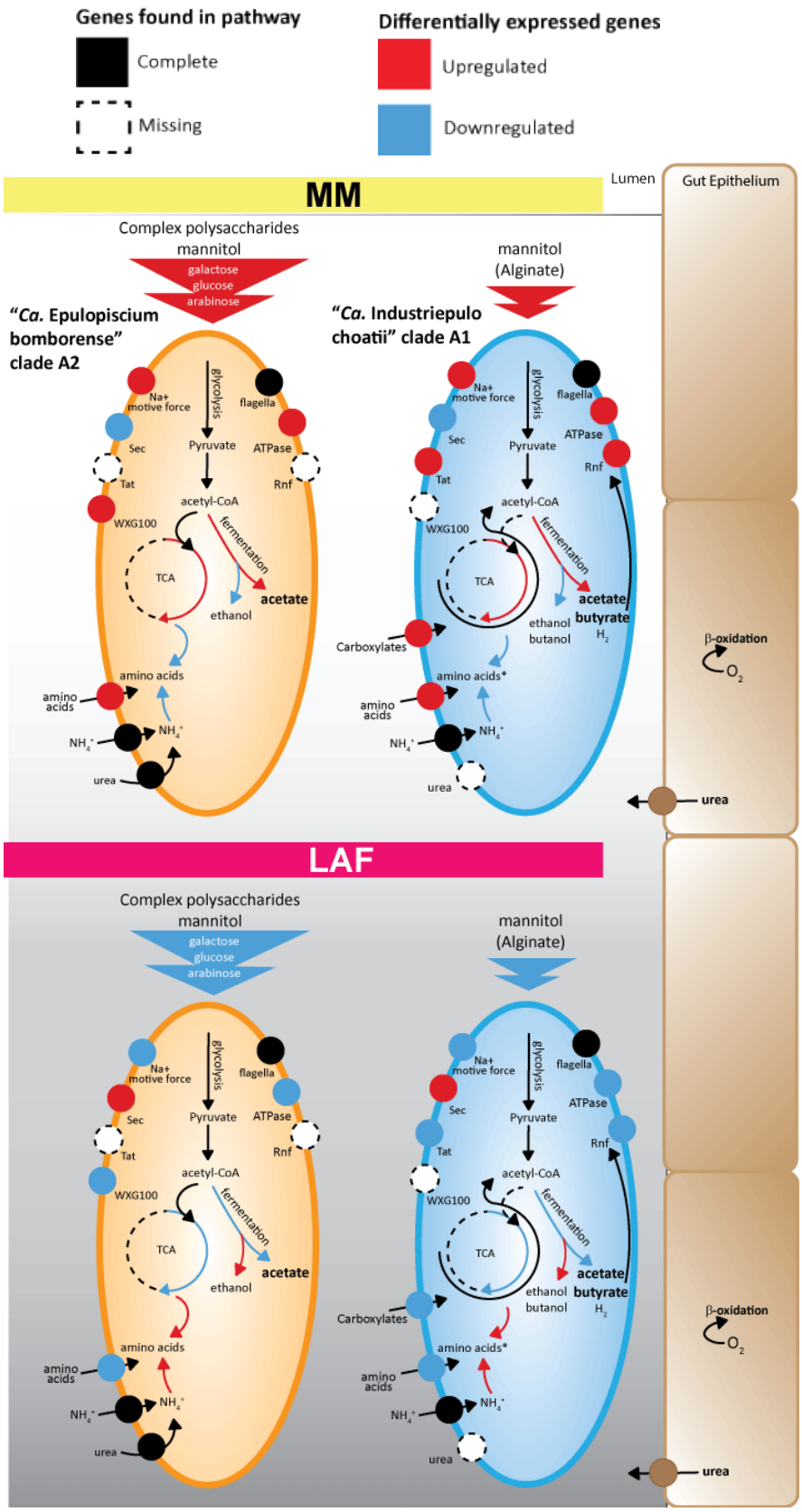
Figure 13. Heatmaps of nitrogen metabolism genes (rows) that were differentially expressed between MM and LAF time intervals (columns) for each genome. Rows are annotated by general function: amino acid biosynthesis, amino acid degradation (or conversion), and membrane transport (e.g. oligopeptide ATP transporters and amino-acid ABC transporters).

Model of epulo-host interactions

We propose that the metabolic capabilities of *E. bomborense* and *Industriepulo* are complementary and collectively contribute to supplying the host with essential SCFAs (Fig. 14). Transcriptional profiles of the symbionts are synchronized in the gut of *A. lineatus* and display time-dependent metabolic expression to allow for symbiont cooperation. In general, acidogenesis shifts to solventogenesis from MM to LAF by downregulating ATP generation via fermentation to acetate as well as generation of SMF to synthesize ATP via ATP synthetases. During MM, *E. bomborense* is hypothesized to degrade complex carbohydrates found in the host diet, mainly agarose, porphyran, carrageenan, and xyloglucans derived from red algae. This obligate fermenter contains multiple carbohydrate transporters and a complete central metabolism able to take up various oligosaccharides and convert them to essential intracellular housekeeping molecules as well as fermenting acetate for the host. *Acanthurus lineatus* is suspected to secrete ammonia and urea into the intestinal lumen where *E. bomborense* detoxifies the gut by importing ammonia and urea and synthesizing amino acids. *Industriepulo* is markedly deficient in the ability to breakdown complex polysaccharides and lacks the ability to channel acetyl-CoA into the TCA cycle. Instead, it relies on importing extracellular carboxylates and the reverse TCA to produce essential carbon intermediates. *Industriepulo* is hypothesized to bias its metabolic capacity to producing energy-rich molecules and maintain redox and pH balance. Carbon intermediates are diverted to the production of acetate and butyrate via energy conserving pathways that include an electron-bifurcation complex, an Rnf complex, and nitrate reduction. In mice, colonic cells preferentially take up butyrate (89). Butyrate is oxidized via an oxygen consuming pathway which results in the production of an anaerobic environment near the the gut wall. Thus, maintaining an environment conducive for anaerobic commensals but disadvantageous for

opportunistic facultative aerobic pathogens. *Industriepulo* lacks the genetic capacity to synthesize all of its amino acids and is proposed to rely on extracellularly-derived amino acids, likely from its neighbor, *E. bomborense.*

Figure 14. Proposed interactions between symbionts "*Ca.* Epulopiscium bomborense" clade A2 and "*Ca.* Industriepulo choatii" clade A1 and their acanthurid host. Top and bottom panels correspond to pathways that were differentially expressed between MM (0900-1200) and LAF (1500-1800), respectively. Presence (solid) and absence (dotted) of genes/pathways are indicated. Gene expression is shown as upregulated in red, downregulated in blue, or not differentially expressed in black.


Genes found in pathway
Complete
Missing
Differentially expressed genes
Upregulated
Downregulated
MM
Lumen
Gut Epithelium
Complex polysaccharides
mannitol
galactose
glucose
arabinose
mannitol
(Alginate)
"Ca. Epulopiscium bomborense" clade A2
"Ca. Industriepulo choatii" clade A1
Na+ motive force
glycolysis
flagella
Sec
Pyruvate
ATPase
Tat
Rnf
acetyl-CoA
WXG100
fermentation
TCA
acetate
ethanol
butanol
butyrate
H2
Carboxylates
amino acids
amino acids*
NH4+
urea
β-oxidation
O2
LAF

CONCLUSION

Epulos exemplify the diversity of gut commensal bacteria in the *Clostridiales* and the multiple ways that large heterotrophic bacteria maintain their extreme size. In this study, we generated high quality and near complete genomes to assess the physiology and ecological roles of these enigmatic anaerobes. By comparing the metabolic potential of two co-resident epulo species, *Epulopiscium bomborense* and *Industriepulo choatii*, we were able to identify key differences in their metabolic potential that inferred interspecies cross-feeding as well as bacterial-host interactions. We provided further evidence that natural populations of epulo symbionts are synchronized with time at the transcriptional level. Daily oscillations of these communities are likely dependent on host feeding and fasting cycles. Although epulos are currently uncultivable and studies specific to surgeonfish biology is limited, our work demonstrates the utility of single-cell genomic approaches on intestinal symbionts to better understand the ecology of non-model symbionts and their hosts.

REFERENCES

1. Youngblut ND, et al. (2019) Host diet and evolutionary history explain different aspects of gut microbiome diversity among vertebrate clades. *Nat Commun* 10(1):2200.

2. Arumugam M, et al. (2011) Enterotypes of the human gut microbiome. *Nature* 473(7346):174–80.

3. Gilbert SF (2014) A holobiont birth narrative: the epigenetic transmission of the human microbiome. *Front Genet* 5:282.

4. Colston TJ, Jackson CR (2016) Microbiome evolution along divergent branches of the vertebrate tree of life: what is known and unknown. *Mol Ecol* 25(16):3776–3800.

5. Angert ER, Losick RM (1998) Propagation by sporulation in the guinea pig symbiont *Metabacterium polyspora*. *Proc Natl Acad Sci U S A* 95(17):10218–23.

6. Ebert D (2013) The epidemiology and evolution of symbionts with mixed-mode transmission. *Annu Rev Ecol Evol Syst* 44(1):623–643.

7. Proctor L (2019) Priorities for the next 10 years of human microbiome research. *Nature* 569(7758):623–625.

8. Kamada N, et al. (2012) Regulated virulence controls the ability of a pathogen to compete with the gut microbiota. *Science* 336(6086):1325–9.

9. Byndloss MX, Litvak Y, Bäumler AJ (2019) Microbiota-nourishing immunity and its relevance for ulcerative colitis. *Inflamm Bowel Dis* 25(5):811–815.

10. Rakoff-Nahoum S, Foster KR, Comstock LE (2016) The evolution of cooperation within the gut microbiota. *Nature* 533(7602):255–259.

11. Clements KD, Sutton DC, Choat JH (1989) Occurrence and characteristics of unusual protistan symbionts from surgeonfishes (Acanthuridae) of the Great Barrier Reef, Australia. *Mar Biol* 102(3):403–412.

12. Miyake S, Ngugi DK, Stingl U (2015) Diet strongly influences the gut microbiota of surgeonfishes. *Mol Ecol* 24(3):656–72.

13. Miyake S, Ngugi DK, Stingl U (2016) Phylogenetic diversity, distribution, and cophylogeny of giant bacteria (*Epulopiscium*) with their surgeonfish hosts in the Red Sea. *Front Microbiol* 7:285.

14. Ngugi DK, et al. (2017) Genomic diversification of giant enteric symbionts reflects host dietary lifestyles. *Proc Natl Acad Sci U S A* 114(36):E7592–E7601.

15. Angert ER, Clements KD, Pace NR (1993) The largest bacterium. *Nature* 362(6417):239–41.

16. Flint JF, Drzymalski D, Montgomery WL, Southam G, Angert ER (2005) Nocturnal production of endospores in natural populations of *Epulopiscium*-like surgeonfish symbionts. *J Bacteriol* 187(21):7460–70.

17. Arroyo FA, Pawlowska TE, Choat JH, Clements KD, Angert ER (2019) Recombination contributes to population diversification in the polyploid intestinal symbiont *Epulopiscium* sp. type B. *ISME J* 13(4):1084–1097.

18. Angert ER, Clements KD (2004) Initiation of intracellular offspring in *Epulopiscium*. *Mol Microbiol* 51(3):827–35.

19. Hutchison E, et al. (2018) Developmental stage influences chromosome segregation patterns and arrangement in the extremely polyploid, giant bacterium *Epulopiscium* sp. type B. *Mol Microbiol* 107(1):68–80.

20. Angert ER (2006) The enigmatic cytoarchitecture of Epulopiscium spp. *Microbiology Monographs: Complex Intracelullar Structures in Prokaryotes*, ed Shively JM (Springer-Verlag, Berlin Heidelberg), pp 285–301. vol 2.

21. Clements KD, Choat JH (1995) Fermentation in tropical marine herbivorous fishes. *Physiol Zool* 68(3):355–378.

22. Turner S, Pryer KM, Miao VPW, Palmer JD (1999) Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. *J Eukaryot Microbiol* 46(4):327–338.

23. Kearse M, et al. (2012) Geneious Basic: an integrated and extendable desktop

software platform for the organization and analysis of sequence data. *Bioinformatics* 28(12):1647–9.

24. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–10.

25. Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29(8):1072–1075.

26. R Core Team (2014) R: A language and environment for statistical computing. Available at: http://www.r-project.org [Accessed November 2, 2015].

27. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 25(7):1043–55.

28. Kobayashi K, et al. (2003) Essential *Bacillus subtilis* genes. *Proc Natl Acad Sci U S A* 100(8):4678–83.

29. Commichau FM, Pietack N, Stülke J (2013) Essential genes in *Bacillus subtilis*: a re-evaluation after ten years. *Mol Biosyst* 9(6):1068.

30. Dupont CL, et al. (2012) Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J* 6(6):1186–1199.

31. Gil R, Silva FJ, Peretó J, Moya A (2004) Determination of the core of a minimal bacterial gene set. *Microbiol Mol Biol Rev* 68(3):518–37.

32. Huntemann M, et al. (2015) The standard operating procedure of the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4). *Stand Genomic Sci* 10(1):86.

33. Markowitz VM, et al. (2012) IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res* 40(D1):D115–D122.

34. Yin Y, et al. (2012) dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* 40(W1):W445–W451.

35. Saier MH, et al. (2016) The Transporter Classification Database (TCDB): recent advances. *Nucleic Acids Res* 44(D1):D372–D379.

36. Berman HM, et al. (2000) The protein data bank. *Nucleic Acids Res* 28(1):235–42.

37. Wang Y, Coleman-Derr D, Chen G, Gu YQ (2015) OrthoVenn: a web server for genome wide comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res* 43(W1):W78-84.

38. DeLong EF, Wickham GS, Pace NR (1989) Phylogenetic stains: ribosomal RNA-based probes for the identification of single cells. *Science* 243(4896):1360–3.

39. Angert ER (1995) Molecular phylogenetic analyses of *Epulopiscium*-like symbionts. Dissertation (Indiana University).

40. Fox J, Weisberg S (2011) *An {R} Companion to Applied Regression* (Sage, Thousand Oaks CA). 2nd Ed. Available at: https://socialsciences.mcmaster.ca/jfox/Books/Companion/ [Accessed November 26, 2018].

41. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–7.

42. Guindon S, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59(3):307–21.

43. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 14(6):587–589.

44. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS (2018) UFBoot2: Improving the ultrafast bootstrap approximation. *Mol Biol Evol* 35(2):518–522.

45. Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA Sequences. *Am Math Soc Lect Math Life Sci* 17:57–86.

46. Ronquist F, et al. (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* 61(3):539–42.

47. Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK (2016)

Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal Methods* 8(1):12–24.

48. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–20.

49. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Available at: http://arxiv.org/abs/1303.3997.

50. Montgomery WL, Pollak PE (1988) *Epulopiscium fishelsoni* N. G., N. Sp., a protist of uncertain taxonomic affinities from the gut of an herbivorous reef fish. *J Protozool* 35(4):565–569.

51. Caccamo PD, Brun Y V. (2018) The molecular basis of noncanonical bacterial morphology. *Trends Microbiol* 26(3):191–208.

52. van Teeseling MCF, de Pedro MA, Cava F (2017) Determinants of bacterial morphology: From fundamentals to possibilities for antimicrobial targeting. *Front Microbiol* 8:1264.

53. Cesar S, Huang KC (2017) Thinking big: the tunability of bacterial cell size. *FEMS Microbiol Rev* 41(5):672–678.

54. Flood BE, et al. (2016) Single-cell (meta-)genomics of a dimorphic *Candidatus* Thiomargarita nelsonii reveals genomic plasticity. *Front Microbiol* 7:603.

55. Bowers RM, et al. (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* 35(8):725–731.

56. Choat JH, Robbins WD, Clements KD (2004) The trophic status of herbivorous fishes on coral reefs: II. Food processing modes and trophodynamics. *Mar Biol* 145(3):445–454.

57. Choat JH, Clements KD, Robbins W. (2002) The trophic status of herbivorous fishes on coral reefs I. Dietary analysis. *Mar Biol* 140(3):613–623.

58. Linn Montgomery W, Gerking SD Marine macroalgae as foods for fishes: an evaluation of potential food quality. *Env Biol Fish* 5(2):143–153.

59. Renaud SM, Luong-Van JT (2006) Seasonal variation in the chemical composition of tropical australian marine macroalgae. *J Appl Phycol* 18(3–5):381–387.

60. Hutcheson SW, Zhang H, Suvorov M (2011) Carbohydrase systems of Saccharophagus degradans degrading marine complex polysaccharides. *Mar Drugs* 9(4):645–65.

61. Gao H, Jiang X, Pogliano K, Aronson AI (2002) The E1beta and E2 subunits of the *Bacillus subtilis* pyruvate dehydrogenase complex are involved in regulation of sporulation. *J Bacteriol* 184(10):2780–8.

62. Fischbach MA, Sonnenburg JL (2011) Eating for two: How metabolism establishes interspecies interactions in the gut. *Cell Host Microbe* 10(4):336–347.

63. Pajor AM (1999) Sodium-coupled transporters for Krebs cycle intermediates. *Annu Rev Physiol* 61(1):663–682.

64. Markovich D, Murer H (2004) The SLC13 gene family of sodium sulphate/carboxylate cotransporters. *Pflugers Arch* 447(5):594–602.

65. Köpke M, et al. (2010) *Clostridium ljungdahlii* represents a microbial production platform based on syngas. *Proc Natl Acad Sci U S A* 107(29):13087–92.

66. Schulz S, et al. (2013) A new type of Na+-driven ATP synthase membrane rotor with a two-carboxylate ion-coupling motif. *PLoS Biol* 11(6):e1001596.

67. Buckel W, Thauer RK (2013) Energy conservation via electron bifurcating ferredoxin reduction and proton/Na+ translocating ferredoxin oxidation. *Biochim Biophys Acta - Bioenerg* 1827(2):94–113.

68. Herrmann G, Jayamani E, Mai G, Buckel W (2008) Energy conservation via electron-transferring flavoprotein in anaerobic bacteria. *J Bacteriol* 190(3):784–91.

69. Li F, et al. (2008) Coupled ferredoxin and crotonyl coenzyme A (CoA) reduction with NADH catalyzed by the butyryl-CoA dehydrogenase/Etf complex from *Clostridium kluyveri*. *J Bacteriol* 190(3):843–50.

70. Harris LM, Welker NE, Papoutsakis ET (2002) Northern, morphological, and fermentation analysis of *spo0A* inactivation and overexpression in *Clostridium acetobutylicum* ATCC 824. *J Bacteriol* 184(13):3586–97.

71. Ravagnani A, et al. (2000) Spo0A directly controls the switch from acid to solvent production in solvent-forming clostridia. *Mol Microbiol* 37(5):1172–85.

72. Wilkinson SR, Young DI, Morris G, Young M (1995) Molecular genetics and the initiation of solventogenesis in *Clostridium beijerinckii* (formerly *Clostridium acetobutylicum*) NCIMB 8052. *FEMS Microbiol Rev* 17(3):275–285.

73. Paredes CJ, Alsaker K V., Papoutsakis ET (2005) A comparative genomic view of clostridial sporulation and physiology. *Nat Rev Microbiol* 3(12):969–978.

74. Vital M, Howe AC, Tiedje JM (2014) Revealing the bacterial butyrate synthesis pathways by analyzing (meta)genomic data. *mBio* 5(2):e00889.

75. Walter KA, Nair R V., Cary JW, Bennett GN, Papoutsakis ET (1993) Sequence and arrangement of two genes of the butyrate-synthesis pathway of *Clostridium acetobutylicum* ATCC 824. *Gene* 134(1):107–111.

76. Alsaker K V, Papoutsakis ET (2005) Transcriptional program of early sporulation and stationary-phase events in *Clostridium acetobutylicum*. *J Bacteriol* 187(20):7103–18.

77. Miller DA, Suen G, Clements KD, Angert ER (2012) The genomic basis for the evolution of a novel form of cellular reproduction in the bacterium *Epulopiscium*. *BMC Genomics* 13:265.

78. Al-Hinai MA, Jones SW, Papoutsakis ET (2015) The *Clostridium* sporulation programs: diversity and preservation of endospore differentiation. *Microbiol Mol Biol Rev* 79(1):19–37.

79. Jones SW, et al. (2008) The transcriptional program underlying the physiology of clostridial sporulation. *Genome Biol* 9(7):R114.

80. Dürre P (2014) Physiology and sporulation in *Clostridium*. *Microbiol Spectr* 2(4):TBS-0010-2012.

81. Miller DA, Choat JH, Clements KD, Angert ER (2011) The *spoIIE* homolog of

Epulopiscium sp. type B is expressed early in intracellular offspring development. *J Bacteriol* 193(10):2642–6.

82. Pallen MJ (2002) The ESAT-6/WXG100 superfamily – and a new Gram-positive secretion system? *Trends Microbiol* 10(5):209–212.

83. Coros A, Callahan B, Battaglioli E, Derbyshire KM (2008) The specialized secretory apparatus ESX-1 is essential for DNA transfer in *Mycobacterium smegmatis*. *Mol Microbiol* 69(4):794–808.

84. Rösch TC, Golman W, Hucklesby L, Gonzalez-Pastor JE, Graumann PL (2014) The presence of conjugative plasmid pLS20 affects global transcription of its *Bacillus subtilis* host and confers beneficial stress resistance to cells. *Appl Environ Microbiol* 80(4):1349–1358.

85. Mukherjee K, Karlsson S, Burman LG, Kerlund TAH (2002) Proteins released during high toxin production in *Clostridium difficile*. *Microbiology* 148(Pt 7):2245–53.

86. Bucking C, LeMoine CMR, Craig PM, Walsh PJ (2013) Nitrogen metabolism of the intestine during digestion in a teleost fish, the plainfin midshipman (*Porichthys notatus*). *J Exp Biol* 216(15):2821–2832.

87. Akashi H, Gojobori T (2002) Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc Natl Acad Sci U S A* 99(6):3695–700.

88. Mee MT, Collins JJ, Church GM, Wang HH (2014) Syntrophic exchange in synthetic microbial communities. *Proc Natl Acad Sci U S A* 111(20):E2149.

89. Byndloss MX, et al. (2017) Microbiota-activated PPAR-γ signaling inhibits dysbiotic Enterobacteriaceae expansion. *Science* 357(6351):570–575.

CHAPTER 3

CHARACTERIZING THE DAILY TRANSCRIPTIONAL OSCILLATIONS OF A GIANT ENDOSPORE-FORMING GUT SYMBIONT FROM *NASO UNICORNIS*

ABSTRACT

Most animals exhibit circadian rhythms that likely impact the intimate relationships they have with their symbiotic partners. The dominant members of the herbivorous surgeonfish gut microbiome, *Epulopiscium* spp. and their relatives, display synchronized intracellular offspring development which may reflect this interplay. In this study, the daily transcriptional oscillations of the endospore-forming clade C1 from wild-caught surgeonfish (*Naso unicornis*) in the Great Barrier Reef were explored. A population-amplified genome (*Nu*_PAG1) was assembled from spores to aid in functional predictions. The C1 clade symbionts are specialized brown algae and urea degraders, which support their contribution of acetate and amino acids to their host. A shift from active metabolism to spore-formation in C1 was coordinated with the feeding/fasting cycle of the host. Gene expression of pathways for ATP synthesis and nutrient assimilation peaked in the early morning and dropped in the afternoon. Endospore germination and initiation of sporulation in populations occurred nearly simultaneously in the early morning and development progressed to the formation of mature endospores late in the evening. Oxygen and hydrogen concentration gradients from the gut wall to the lumen were inversely correlated; oxygen concentrations decreased and hydrogen increased with depth. Changes in gas concentration gradients over the course of the day corresponded with the shift from active metabolism to sporulation of the symbiont. These data suggest that the metabolic and reproductive cycles of symbiont clade C1 are synchronized with the

host gut environment. Maintaining rhythm with the host likely explains the ecological success of these widespread intestinal symbionts in herbivorous surgeonfish.

INTRODUCTION

Many organisms (1), have mechanisms to anticipate and prepare for the day/night cycle by evolving circadian clocks: endogenous feedback loops that follow a 24 hour periodicity. A common external cue (zeitgeber) that entrains these rhythms is light. Cyanobacteria in culture (2) and in the open ocean (3) demonstrate diel transcriptional oscillations to synchronize their photosynthetic metabolism with daylight hours. Some non-heterocystous cyanobacteria synchronize nitrogen fixation with the night so that this oxygen-sensitive reaction is temporally separated from oxygenic photosynthesis (4, 5). Diel periodicity is also observed among heterotrophic bacterioplankton taxa within the community of cyanobacteria in the open ocean (3, 6). Although the external cues for these heterotrophs are not known, the coordination of metabolic and biogeochemical processing between heterotrophic bacteria and their photoautotrophic neighbors demonstrate how circadian rhythmicity of one organism can directly affect the life cycles of ecosystems.

Day/night cycles are known to affect community composition in mammalian gut microbiomes (7–9). Host-microbial communities are dependent on the host for nutrients and as such, these microbes experience environmental changes due to host feeding/fasting cycles, activities and daily variations in host immunity. Murine studies show a peak in *Firmicutes* numbers during feeding while *Bacteroidetes* dominate during fasting periods. Members of the *Firmicutes* (Lachnospiraceae and Clostridia) thrive when dietary glycans are available. Their ability to break down complex polysaccharides extracellularly at the onset of dietary input establishes members of the *Firmicutes* as primary consumers that produce fermentation byproducts such as

acetate and other short chain fatty acids that can be used by the host. As the dietary glycans diminish, *Firmicutes* numbers diminish. Meanwhile, fasting allows *Bacteroidete*s to flourish due to their ability to utilize host-derived glycans, like mucin. Gene expression for mucus degradation and motility of mucosa-associated community members are upregulated during fasting (10). Predictable temporal changes in the gut habitat have likely selected for microbes that can anticipate environmental changes; successful commensals would be prepared to use the increased input of nutrients in the morning and transition quickly as the fasting phase approaches and the environment becomes perhaps less hospitable (e.g. decreased pH, accumulation of toxic metabolites).

Many members of *Firmicutes* are obligate anaerobes that are able to produce endospores. This allows them a unique option for adapting to host fasting while improving the likelihood of transmission to another host (11). Segmented filamentous bacteria (SFB) are a group of vertebrate-associated commensals that bind specifically to the host epithelium without invading the tissue (12). SFBs undergo a complex life cycle where holdfast cells bind to the ileal epithelium, develop into a long, segmented filament, and later differentiate to produce either endospores or intracellular holdfast cells that are eventually released into the lumen. The filament remains attached to the ileum for 2-3 days depending on the frequency of epithelium turnover. Under favorable conditions, the holdfast form is produced to recolonize tissue in the intestinal tract. Hostile conditions (e.g. unfavorable oxygen concentrations) promote spore formation. Spores are shed in the feces of the host animal and may be transmitted to other hosts. Although SFBs appear to adapt to the changing gut environment, they do not follow any perceptible daily cycle.

Epulopiscium sp. and their relatives, known as epulos, are good models for investigating circadian oscillations of host-associated gut microbes. Epulos are a

morphological diverse group of bacteria from the Lachnospiraceae XIVb lineage that inhabit the intestinal tract of herbivorous surgeonfish (13–17). The reproductive strategies among epulos varies. The largest cell types, "*Ca.* Epulopiscium*"* spp. and *"Ca.* Industriepulo*",* have the ability to produce multiple intracellular offspring while lacking the ability to undergo binary fission (13, 14, 18). Other epulos, such as clade C, likewise do not use binary fission but have maintained the ability to produce multiple, mature endospores. Epulos exhibit a predictable, diurnal reproductive cycle evident in natural populations (16, 18, 19). Generally, intracellular offspring formation and growth occurs over the course of the day and offspring are released from the mother cell the following morning. In spore-forming epulos, forespores develop and grow during the day then transition to phase-bright endospores by night (16). Some spores germinate inside the host intestinal tract while others are shed in the early morning, at the onset of host feeding, and thus may be transmitted to other hosts. The triggers and timing of sporulation are unknown, but may be influenced by changes in the gut environment.

In this study, I characterized the daily oscillations of endospore-forming clade C epulos. To accomplish this, I assembled a population-amplified genome to predict the functional capacity of endospore-forming epulos in *Naso unicornis*. I also characterized the daily transcriptional and developmental oscillations of gut symbionts in wild-caught surgeonfish from the Great Barrier Reef. I identified the pathways that were coordinated with time of day (including sporulation, carbon and nitrogen metabolism) and examined the reproducibility of synchronization at the transcriptional level among naturally occurring epulo populations. I monitored various physiochemical parameters *in situ* along a transverse axis of a region of the surgeonfish intestine where clade C cells predominate with the goal of identifying daily fluctuations of the gut environment.

METHODS

Sample summary

Spores used for DNA extraction and genome assembly were collected from a single *Naso unicornis* individual that was kindly provided by Dr. Edward DeMartini, NOAA National Marine Fisheries Service, Honolulu, HI. All other samples were from *N. unicornis* or *N. lituratus* caught by spear near Lizard Island, Great Barrier Reef, Australia. Fish were kept on ice until processed. With the exception of microsensor monitoring, the gut was removed, untangled, and sectioned as shown in Fig. 1. The intestinal contents containing large numbers of epulos were fixed in 80% ethanol for DNA and microscopy, spores were preserved in 20% glycerol, and all were stored at −20°C. Samples used for transcriptome analyses were not fixed but processed immediately to extract RNA. For microsensor monitoring, the intestinal tract was removed from the fish and kept coiled (Fig. 1A). Sampling information and demographic data for each fish host was recorded along with descriptions of corresponding epulo symbiont.

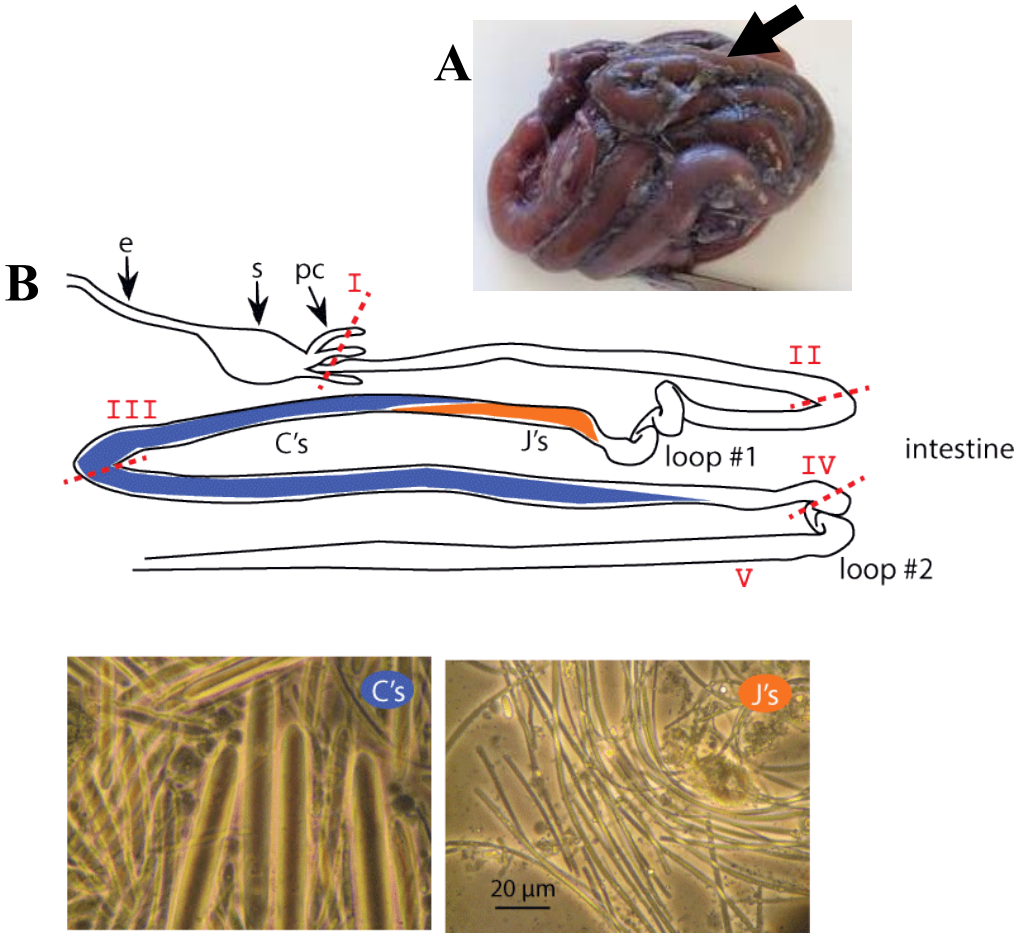


Figure 1. Gut anatomy of *Naso* species and distribution of epulo morphotypes. (A) Coiled gut used for microsensor analyses with an arrow indicating location of section IV monitoring. (B) A diagram of the uncoiled gastrointestinal tract of *Naso* species (*N. unicornis* and *N. lituratus*) was divided into 5 sections as indicated for sample collections. Section I contains the esophagus (e), stomach (s), and pyloric cecae (pc). Sections II-V comprise the intestine and characteristic hairpin loops found in these surgeonfish. The sections of the intestine are subdivided relative to the hairpin loops as indicated. Distribution of the two epulo morphotypes C and J are shown in sections III and IV. Micrographs below the drawing show contents from the C and J regions of the intestine.

Laser capture of endospores and genome amplification

Spores of *Epulopiscium* C1 were selected by laser capture micro-dissection catapult (LCM) (Zeiss PALM Microbeam, Imaging at The Biotechnology Resource Center Imaging Facility, Cornell University) as follows. An aliquot from the glycerol

stock was washed 4X in PBS and 2X in 70% ethanol, resuspended in 70% ethanol and transferred to a UV-treated membrane slide (PEN 1.0, Zeiss). The slide was air-dried completely in a biosafety cabinet prior to laser-capture. Five sets of 20 spores were captured onto AdhesiveCap centrifuge tubes (Zeiss) and processed separately. The spore coat was removed by soaking spores in decoating solution (50 mM Tris, pH 8.0; 1% SDS; 8 M urea; 50 mM DTT; and 10 mM EDTA) at 37°C for 90 minutes. Spores were then rinsed 5X in wash buffer (10 mM Tris, pH 7.2; 10 mM EDTA; 150 mM NaCl) and then rinsed in nuclease-free sterile water. Finally, decoated spores were washed with Agencourt AMPure magnetic beads and subjected to whole-genome amplification as described in Chapter 2, methods.

Amplified genomic DNA (gDNA) was visualized on a 1% agarose gel and samples with high molecular weight DNA were purified using QIAmp DNA mini kit (Qiagen). Purified gDNA was screened via PCR (HotStarTaq Master Mix, Qiagen) using Epulo-specific primer sets 515F/1423R and screened for purity using bacterial universal primers set 8F/1492R. After purification with QIAquick PCR purification system kit (Qiagen), amplicons were sequenced (in both directions) on an ABI 3730 automated sequencer (Biotechnology Resource Center, Cornell University). Sequences were trimmed and analyzed using Geneious version 6.0.6 (http://www.geneious.com, Kearse *et al.* 2012) and compared to nucleotide sequences in Genbank using BLAST (21). Samples that shared >98% 16S rRNA gene sequence identity with *Epulopiscium* clade C1 were used for sequencing. The amplified gDNA of these candidate samples were de-branched using Nuclease S1 (Promega) and purified by phenol/chloroform prior to library preparation.

Whole genome sequencing and assembly

The amplified gDNA from 100 pooled laser-captured C1 spores (5 sets of 20) were sequenced at the Biotechnology Resource Center (Cornell University). Paired-

end reads (2x300 bp) were sequenced using the Illumina MiSeq platform.

Raw reads were processed by the BBtools software package (https://sourceforge.net/projects/bbmap/). Paired-end, interleaved reads were assembled using the metaSpades mode of SPAdes 3.7. Trimmed reads were mapped back to their concurrent assembly to determine read coverage using bbmap and samtools. Contamination control, genome completeness and genome finishing were performed as described in Chapter 2, methods.

Gene annotation and discovery

The draft genome was submitted to the Integrated Microbial Genomes (IMG) database for annotation using the MGAP v4 pipeline (22). Gene content and gene neighborhood comparisons were initially analyzed on the IMG comparative analysis system (23). Further gene discovery for polysaccharide metabolic enzymes, secretory proteins, and substrate specificity of transport proteins was performed as previously described (Chapter 2, methods). Genes that were previously identified in (24) to be involved in the development of endospores for *Epulopiscium* sp. and related bacteria were compared to the draft genome *Nu*_PAG1.

Clone library surveys and phylogenetic analysis

To further examine the phylogenetic diversity of morphotype C cells, DNA was extracted from ethanol-fixed intestinal contents of *N. unicornis.* Cell lysis, DNA extraction, and clone libraries were performed as described in Chapter 2 (Methods: 16S rRNA gene clone library surveys).

16S rRNA gene maximum likelihood (ML) trees were constructed in PhyML version 3.0 (25) using the generalized Time Reversible nucleotide substitution model (26) plus invariant sites (I) and Γ rate heterogeneity. Bootstrap values were generated from 1,000 replicates. To further characterize the phylogenetic relationships of these clades, the average nucleotide identity (ANI) was calculated using the Kostas lab tool

(http://enve-omics.ce.gatech.edu/ani/) with default parameters.

RNA extraction and sequencing

Fish caught between time intervals 0600-0900 hours (EM), 0900-1200 (MM), 1200-1500 (AF), and 1500-1800 (LAF) were used for transcriptomic analyses of symbionts over time. Total RNA was extracted from the gut fluid of section III and IV (Fig. 1B) immediately upon arrival to the research station. Cells were lysed by bead beating in RLT buffer (Qiagen) containing beta-mercaptoethanol and Ambion RNase inhibitor. RNA was further purified using the Qiagen RNAeasy kit following manufacturer's instructions, stored in liquid nitrogen for international shipment, and stored at -80°C. DNA was removed using the TURBO DNA-free kit (Invitrogen). The quantity and quality of RNA was determined by Qubit spectrofluorometry and fragment analysis (ABI 3730xl, Cornell University Genomics facility), respectively. Ribosomal RNA from the host and gut community was removed using Illumina Ribozero Gold rRNA removal kit (epidemiology). Libraries were prepared as directional (stranded) and ~75 nt single reads were sequenced using Illumina NextSeq500 by the RNA sequencing core (Cornell University).

Reads were trimmed using Trimmomatic v0.36 (27) and residual rRNA sequences were removed by mapping known rRNA sequences with bwa mem (28). Processed reads were mapped to the draft genome *Nu*_PAG1 using bwa mem. Raw count summaries and RPKM values were calculated and organized using samtools and the mRNAtool.pl script (https://github.com/kentnf/KTools/blob/master/mRNAtool.pl). Differential expression across time intervals (EM, MM, AF, and LAF) was analyzed using R packages *DeSeq2* 1.22 and *vsn* 3.50 and visualized as heatmaps using *pheatmap*. After exploratory analysis, samples within time interval EM were re-grouped into EM1 and EM2. These additional subgroups allowed for finer temporal resolution.

Additional variables, beyond time of day, were explored that might have influenced transcriptional oscillations. These variables included year sampled, geographic location of sampled fish, section of gut sampled, and developmental stage of symbiont. Developmental stage was quantified as the proportion of the average length of forespores compared to the length of their mother cell. As the cell cycles progresses, the proportion of offspring occupancy increases until spores mature. Variables were annotated alongside heatmaps and separately analyzed as PCAs in R.

Environmental monitoring of *N. unicornis* gut section IV

Duplicate/triplicate fish were caught at four time intervals to track environmental fluctuations in the gut of host *N. unicornis* over the course of 12 hours (Table 1). Concentrations of oxygen, hydrogen, and redox potentials in the gut were monitored *in vivo* using microsensors (Unisense, Aarhus, Denmark) with a 25 μm tip diameter. At least three technical replicates were taken for each fish sample. Transverse profiling of the intact gut at section IV (Fig. 1A) were conducted immediately upon arrival to the research station. At the end of monitoring each fish, a subsample was observed microscopically to verify presence of symbionts. The hydrogen microsensor was calibrated using hydrogen gas-saturated water concentrations of 0, 10, 50, and 100 μM. Oxygen microsensors were calibrated by the difference between air-saturation and anoxic conditions (0.1 M ascorbate in 0.1 M NaOH). A hose with an air stone adapter was submerged in water and attached to an air pump which was used to generate air-saturated water. The redox microsensor was calibrated with saturated solutions of quinhydrone in pH standards of pH 4.0 and 7.0. During redox measurements, the intact gut was placed in PBS and the electric potential was measured against the Ag-AgCl reference electrode submerged in PBS. Multiple readings for all conditions were collected at 25 μM-depth intervals from the surface of the gut to a maximum depth of 4 mm.

Duplicate hindgut contents (~1 g) were collected and weighed in 20-ml, N_2-flushed vials and sealed. The headspace (1 mL) from these vials were immediately transferred to a secondary 20-mL, N_2-flushed and sealed vial. Although the gut-containing vials were refrigerated, the secondary vials provided a more conservative measure of the concentration of methane at the time of sampling. Methane was measured by gas chromatography (Perking-Elmer 3920B) and calibrated at 1000 ppm.

Table 1. Sample descriptions for microsensor and methane analyses from *N. unicornis* gut section IV.

Time of day*	Sample ID	Year sampled	Time caught	Gut depth (cm)	Sample location near Lizard Isand, AUS	Sex	Sampled for methane analysis
EM	G74	12/7/2016	655	0.7	Vicki's Reef	n.d.	
	G83	12/8/2016	655	1.2	Corner Reef	female	
MM	G41	12/4/2016	1000	1	South Island	female	
	G63	12/6/2016	1000	1	North Direction	male	x
AF	G107	12/10/2016	1245	0.8	Washing Machine	immature	
	G87	12/8/2016	1300	0.8	Mac's Reef	male	x
	G47	12/4/2016	1330	1	North Reef	immature	
LAF	G59	12/5/2016	1620	1.5	Granite Bluffs	female	
	G71	12/6/2016	1545	0.7	North Direction	n.d.	x
	G100	12/9/2016	1650	0.8	Washing Machine	male	

*Time of day intervals correspond with early-morning (EM: 0600-0900), middle-morning (MM: 0900-1200), afternoon (AF: 1200-1500), and late-afternoon (LAF: 1500-1800).

RESULTS

Genome statistics and phylogenetic placement

To improve the resolution of clade C, 16S rRNA clone surveys from *N. unicornis* collected in the Great Barrier Reef (GBR) were sequenced (Fig. 2). Currently available sequences contained members of clade C from *N. lituratus*

collected near Hawaii, and *N. unicornis* and *N. elegans* sampled in the Red Sea. These new sequences provided additional support for the previously described clades C1 and C2 (16) and support for a new clade, C3.

A high-quality draft population-amplified genome (PAG) was assembled from a pooled subset of 100 LCM spores herein referred to as *Nu*_PAG1 (Table 2). The spores were collected from gut contents of a single *N. unicornis* individual, caught near Oahu, HI, USA. Phylogenetic reconstructions and sequence similarity of 16S rRNA genes (>99%) place this population among endospore-forming epulo clade C1 from the Pacific (GBR and Hawaii), regardless of host species (*N. unicornis* and *N. lituratus*). A previously described metagenome assembled draft genome for clade C (PG-Nuni2H_MBin01) from the Red Sea (29) shares 88-89% average nucleotide identity (ANI) and 97% 16S rRNA gene similarity with *Nu*_PAG1, indicating that *Nu*_PAG1 is a novel species and likely a novel genus. Therefore, the name "*Candidatus* Epulonibulbus gigasporus" clade C1 is proposed for *Nu*_PAG1 to describe its distinct morphology as bulbus-shaped cells that produces giant endospores (Latin: Epulonibulbus, a bulb-shaped guest at a banquet; gigasporus, having giant spores). Members of clade C1 are found in *N. unicornis* and *N. lituratus* from Hawaii and GBR as well as *N. unicornis* and *N. elegans* from the Red Sea (Fig. 2).

Using isolated spores for genome assembly proved successful in this study and produced high quality DNA. The spore decoating procedure helped remove vegetative bacterial cells, host cells, and other potential sources of DNA that were present in the intestinal sample. By limiting the number of spores to 100 and isolating spores based on distinct morphological attributes (e.g. length, width) via LCM, population heterogeneity was limited and this aided the assembly of a draft PAG to 19 contigs.

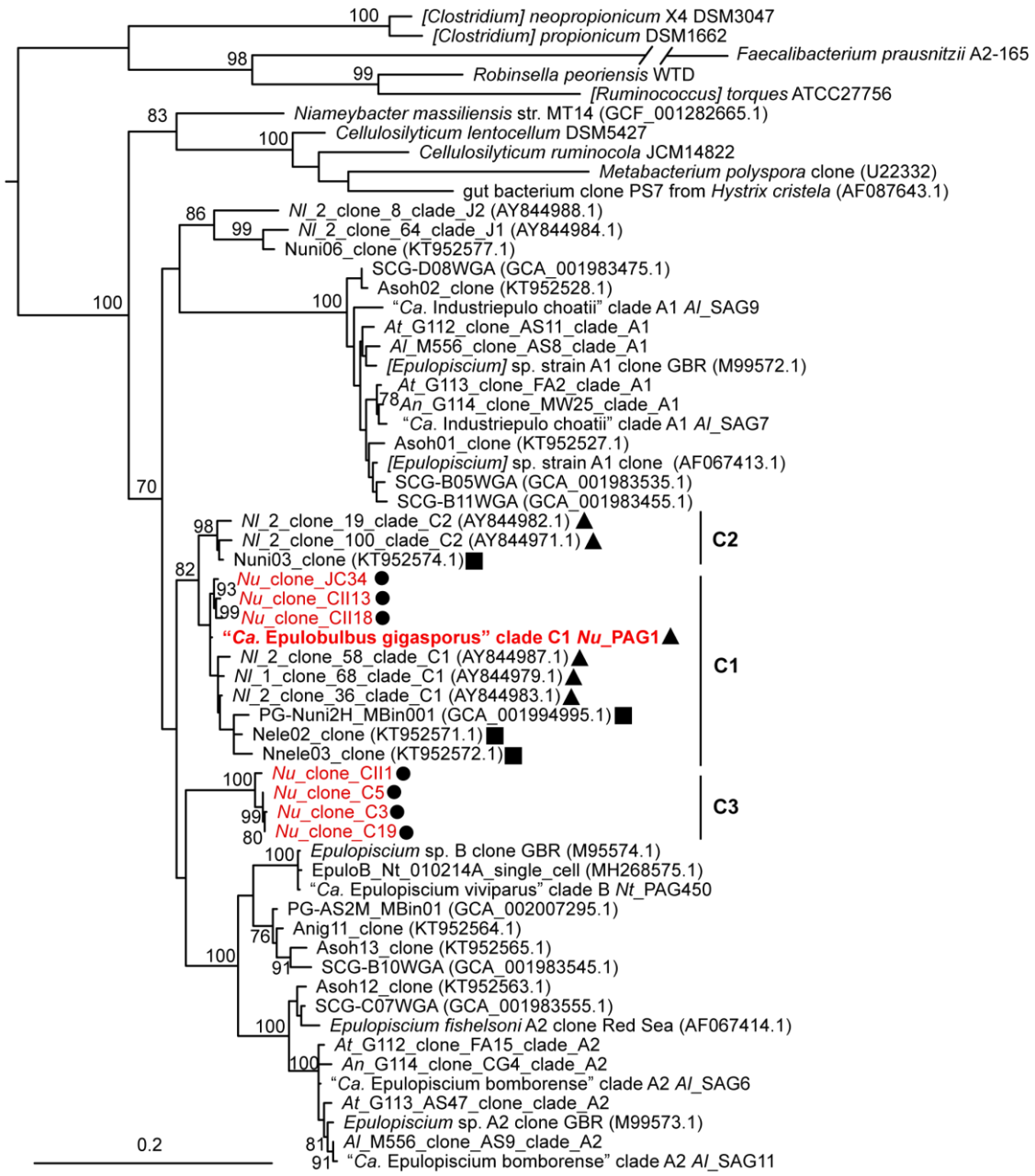


Figure 2. Phylogenetic analysis of *Epulopiscium*-like C morphotypes. A maximum likelihood tree was constructed from 16S rRNA gene sequence comparisons of previously described *Epulopiscium* clones and draft genomes. Bootstrap values greater than 70% are provided at nodes (1000 replicates). Tree was rooted using members of the Lachnospiraceae XIVa and *F. prausnitzii*. Three morphotypes (C1, C2, and C3) identified using clade specific oligonucleotide probes are shown. Sequences derived in this study are in red font and the genome in bold font. Host fish are abbreviated as *N. unicornis* (*Nu* or Nuni), *N. lituratus* (*Nl*), and *N. elegans* (Nele). Geographic location for clade C sequences are shown as GBR (circle), Hawaii (triangle), and Red Sea (square). Scale bar represents 0.2 nucleotide substitutions.

Table 2. Genome statistics for epulo C1 PAG derived from endospores.

	"*Ca.* Epulonibulbus gigasporus" clade C1
Genome ID	*Nu*_PAG1
Assembly quality & analysis type[1]	High quality draft PAG
Number of spores in population	100
Average genome coverage	1,236x
Genome size (bp)	3,054,003
Contigs	19
%GC	34.17
Contig N50/L50	371,083/3
Max contig length (bp)	829,938
Protein coding genes (percent from total genes)	2,544 (96.07%)
CRISPR/Cas genes	7
Giant genes (code for proteins > 2000 aa)	12
rRNA genes present	16S, 23S, 5S
tRNA genes (incl. all 20 aa)	60
%Completeness (CheckM[2]/Reference genome[3])	96/93
%Contamination/ Heterogeneity	0/0
Fish Host	*Naso unicornis*
Location	Oahu, HI, USA

[1] As defined by (30).
[2] CheckM Class: Clostridia.
[3] Compared to the closest relative with a complete genome, *Cellulosilyticum lentocellum*.

Transcriptional oscillations correlated with time and cell development

To explore the daily transcriptional oscillations of epulos, *N. unicornis* individuals were collected over a 12-hr period (i.e. EM, MM, AF, and LAF) in December 2016 and Dec 2017. In total, 27 samples from *N. unicornis* met our minimum mapping threshold of ~1.5 M reads (Table 3). Three additional samples were collected from host *N. lituratus* within the same time interval MM, for comparisons.

Differential expression (DE) analysis of the four time intervals were initially explored using the entire dataset of 30 samples (Fig. 3). The largest time range comparison, EM to LAF, had the greatest number of genes significantly differentially expressed over all other pairwise comparisons (1,288 genes; log2fold change > 1.5; $P_{adj.} < 0.05$). Using this broad dataset, additional metadata variables were explored via principal component analysis (PCA) to determine if other conditions corresponded with the transcriptional profiles. Two different hosts (*N. unicornis* and *N. lituratus*) collected at the same time interval clustered together (Fig. 3B), suggesting that symbiont daily transcriptional oscillations occur regardless of host species. Not surprising, cell developmental stage was significantly correlated with time ($r = 0.94$, $P = 1.65 \times 10^{-8}$) and daily transcriptional profiles (Fig. 3A & 3C). Transcripts originating from different sections of the gut largely clustered together indicating that many genes are synchronized across this area of the intestine (Fig. 3E). Sampling year and location do not appear to influence daily transcriptional oscillations (Fig. 3D & 3F).

Transcription profiles from the EM time interval formed two distinct clusters (Fig. 3B). This observation was not explained by the other sampling variables examined (year, sample location, etc.). The heatmap shows a distinct shift with the corresponding samples (G75C, G75J, G76C, and G86C). Therefore, the EM samples were further divided into two subgroups EM1 and EM2. To reduce bias by

oversampling a single host, samples that were collected from both gut sections were limited to one representative, with preference for gut section IV. To further simplify the interpretations, *N. lituratus* samples were omitted from further analysis. Samples that were transcriptional outliers of their respective time interval were removed. Altogether, this new refined sample set consisting of 19 samples (Table 3, black font) was used for the high resolution DE analysis for the rest of the study.

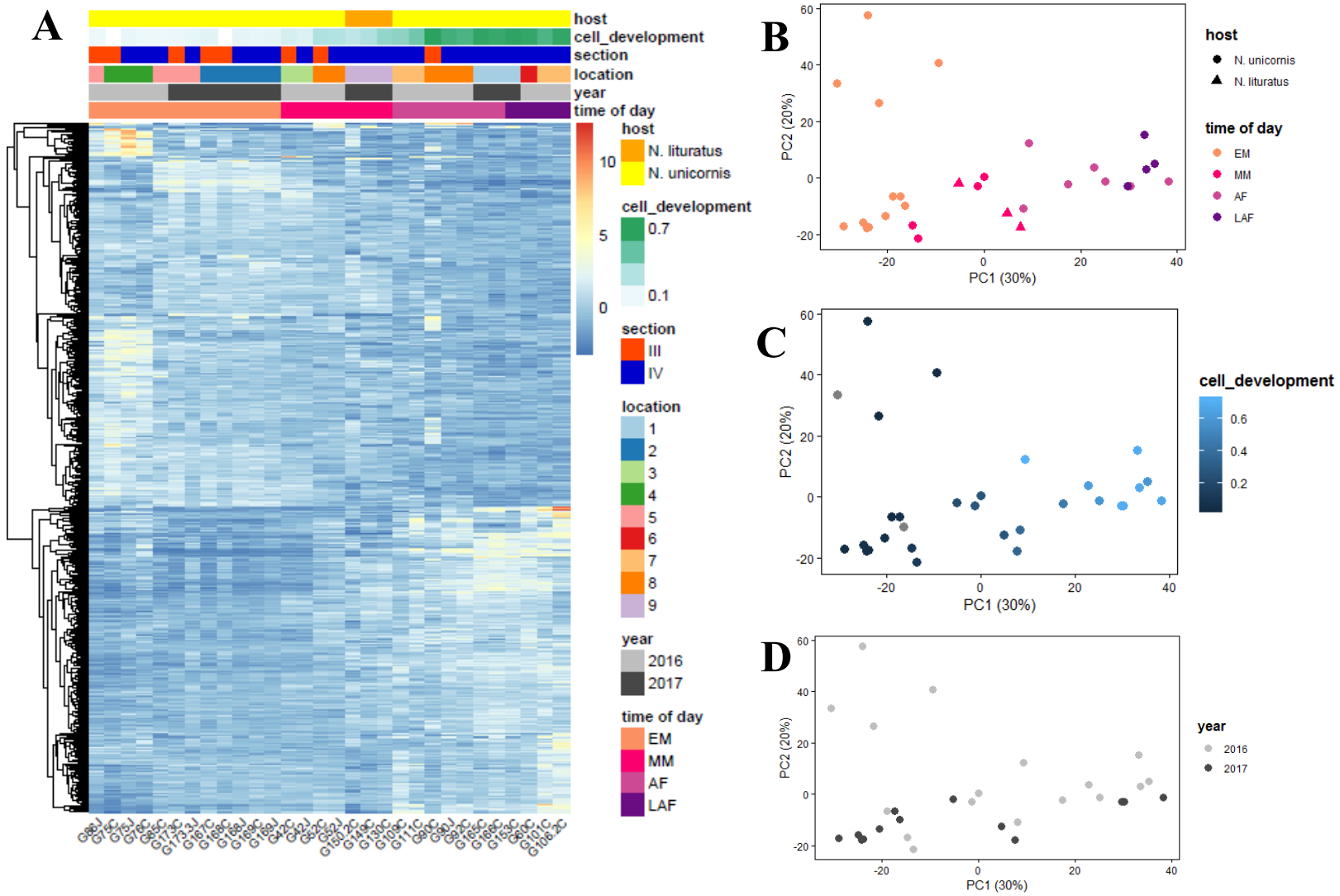
To improve the resolution of the DE analysis, significantly differentially expressed genes between the broad time scale EM1_LAF (Fig. 4A) and daily progressing pairwise comparisons from EM1 to LAF (Fig. 4B) were combined. In total 1,917 differentially expressed genes were organized into functional categories using COG, followed by manual annotations. Major categories that are described in this study include carbon metabolism and transport (158 DE genes), energy (71 DE genes), nitrogen metabolism and transport (193 DE genes), sporulation (64 DE genes), and motility (103 DE genes).

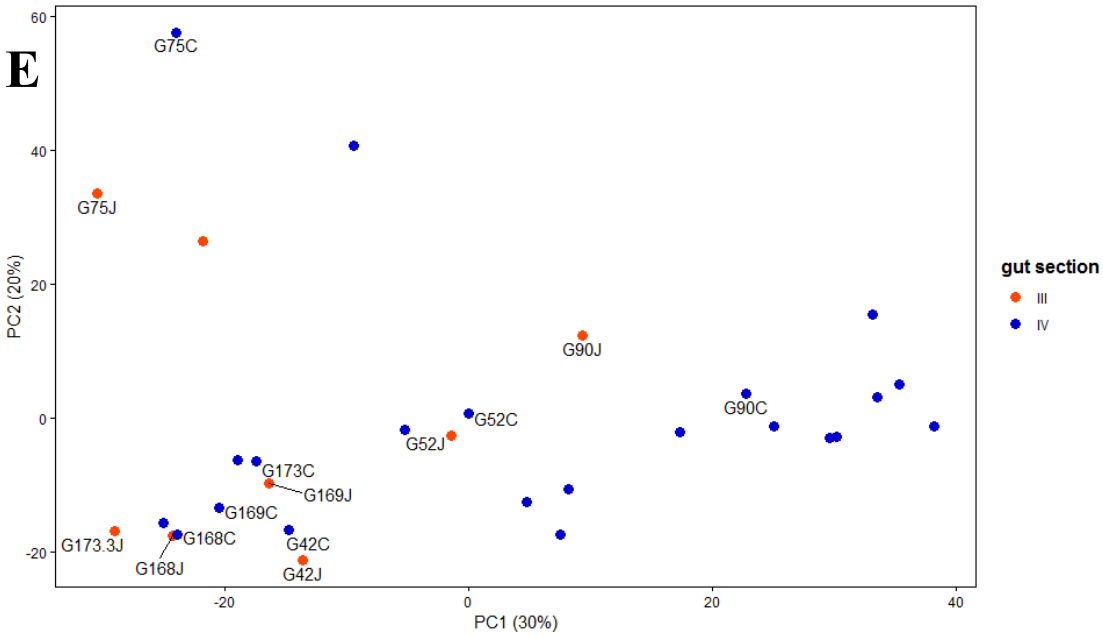
Table 3. Sample description and RNA mapping statistics to *Nu*_PAG1. Samples shown in red font were dropped from the fine scale DE analysis because they did not meet minimum read threshold, were redundant samples from the same individual host, were from host *N. lituratus*, or were outliers.

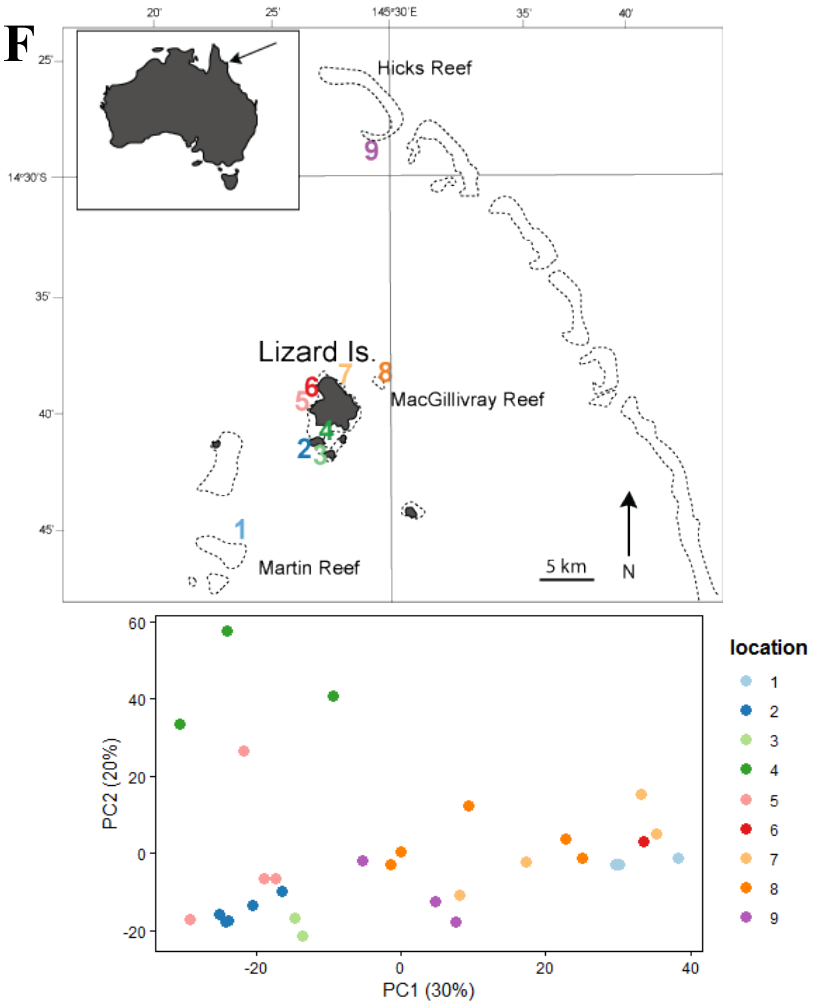
Time of day[1]	Sample ID	Time caught	Host	Raw reads	Mapped reads
EM1	G86J[2]	645	*N. unicornis*	41,692,945	5,011,676
EM1	G75C[3]	650	*N. unicornis*	16,797,413	6,168,219
EM1	G75J[2]	650	*N. unicornis*	30,383,023	1,763,900
EM1	G76C[3]	705	*N. unicornis*	18,482,329	1,806,202
EM1	G85C[2]	710	*N. unicornis*	16,180,659	6,367,772
EM2	G173.3J[2]	830	*N. unicornis*	21,271,404	2,663,753
EM2	G173C[3]	830	*N. unicornis*	14,096,714	11,672,825
EM2	G167C[3]	850	*N. unicornis*	22,654,631	12,272,317
EM2	G168C[3]	850	*N. unicornis*	14,948,765	1,624,328
EM2	G168J[2]	850	*N. unicornis*	8,909,439	4,074,568
EM2	G169C[3]	850	*N. unicornis*	9,631,131	4,085,596
EM2	G169J[2]	850	*N. unicornis*	23,525,247	6,687,223
MM	G42C[3]	1000	*N. unicornis*	21,328,501	18,585,690
MM	G42J[2]	1000	*N. unicornis*	18,577,512	12,467,207
MM	G52C[3]	1045	*N. unicornis*	24,425,721	10,519,515
MM	G52J[2]	1045	*N. unicornis*	31,092,226	4,137,206
MM	G150.2C[3]	1120	*N. lituratus*	19,325,496	6,332,854
MM	G149C[3]	1145	*N. lituratus*	23,892,397	5,455,566
MM	G130C[3]	1150	*N. lituratus*	6,433,720	4,342,969
AF	G109C[3]	1210	*N. unicornis*	13,814,325	2,074,770
AF	G111C[3]	1230	*N. unicornis*	3,095,036	2,298,465
AF	G90C[3]	1300	*N. unicornis*	22,841,636	1,492,846
AF	G90J[2]	1300	*N. unicornis*	30,640,716	20,137,404
AF	G92C[3]	1300	*N. unicornis*	17,193,505	11,330,710
AF	G165C[3]	1450	*N. unicornis*	17,519,899	11,806,266
AF	G166C[3]	1450	*N. unicornis*	16,802,272	10,582,225
LAF	G153C[3]	1530	*N. unicornis*	10,493,241	7,748,192
LAF	G60C[3]	1530	*N. unicornis*	18,357,073	14,511,649
LAF	G101C[3]	1630	*N. unicornis*	20,948,493	7,056,018
LAF	G106.2C[3]	1640	*N. unicornis*	17,982,002	9,946,389

[1]Time of day intervals correspond with early- morning (EM: 0600-0900; further subdivided as EM1:0600-0800, EM2:0800-0900), middle-morning (MM: 0900-1200), afternoon (AF: 1200-1500), and late-afternoon (LAF: 1500-1800).
[2]Gut section III.
[3]Gut section IV.

Figure 3. Differential expression (DE) and exploratory analyses of the impact of sampling variables. (A) Heatmap of DE values from type C epulos comparing the two extreme time intervals EM and LAF. Values are rlog transformed for 1,228 genes significantly differentially expressed (log2fold change > 1.5; $P_{adj.} < 0.05$). Samples are arranged in order, from left to right, by time sampled. Metadata information for each sample is mapped along the top. (B-F) PCA analysis of each sample according to metadata condition or variable: (B) host (*N. unicornis* or *N. lituratus*) and time interval (EM, MM, AF, LAF); (C) developmental stage of symbionts calculated as proportion of daughter cell length to mother cell length. Light gray dots indicate samples did not contain cell development observations; (D) year sampled; (E) sample IDs are indicated for fish that yielded type C RNAseq data from both sections III (or J) and IV (or C); (F) fish collection locations numbered 1-9, corresponding to the map of Lizard Island and outer reef. Samples that cluster together represent fish replicates taken at the same time on the same day.


A
host
cell_development
section
location
year
time of day
host
N. lituratus
N. unicornis
cell_development
0.7
0.1
section
III
IV
location
1
2
3
4
5
6
7
8
9
year
2016
2017
time of day
EM
MM
AF
LAF
10
5
0
B
PC2 (20%)
PC1 (30%)
host
N. unicornis
N. lituratus
time of day
EM
MM
AF
LAF
C
PC2 (20%)
PC1 (30%)
cell_development
0.6
0.4
0.2
D
PC2 (20%)
PC1 (30%)
year
2016
2017


E
G75C
G75J
G90J
G90C
G52C
G52J
G173C
G169J
G169C
G173.3J
G168C
G168J
G42C
G42J
PC2 (20%)
PC1 (30%)
gut section
III
IV


F
Hicks Reef
Lizard Is.
MacGillivray Reef
Martin Reef
5 km
N
PC2 (20%)
PC1 (30%)
location
1
2
3
4
5
6
7
8
9

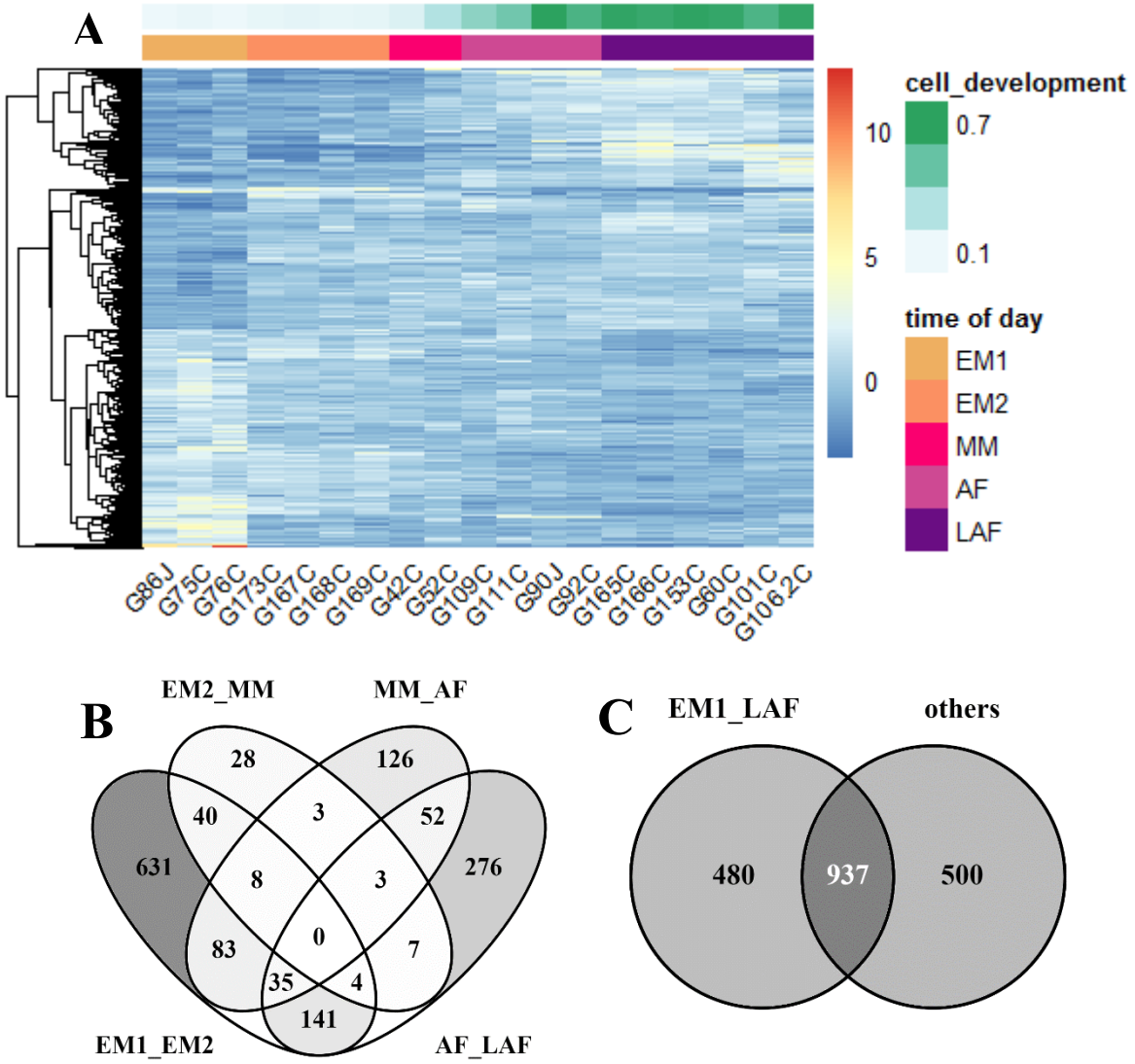


Figure 4. Differentially expressed genes for new time intervals EM1, EM2, MM, AF, and LAF. (A) Heatmap of expression values, rlog transformed for 1,417 genes significantly differentially expressed (log2fold change > 1.5; $P_{adj.} < 0.05$) between EM1 and LAF. Metadata variables that best correlated with symbiont transcription profiles are mapped on the top. (B) Venn diagram indicating the number of significantly differentially expressed genes between each progressive time interval. (C) Venn diagram indicating the number of significantly differentially expressed genes between EM1_LAF and all other progressive time intervals from B.

Specialized brown algae degrader

Naso unicornis and *N. lituratus* share similar feeding preferences for brown thallate algae (31, 32). Epulos are hypothesized to breakdown and ferment the complex polysaccharides derived from brown algae (*Phaeophyceae*) into short chain fatty acids (SCFAs) for the host. The major SCFA found in the intestine of both *Naso* species is acetate, attributing to 85-90% of the total SCFAs (33). Brown algae is composed mainly of alginate, laminarin, and mannitol. Generally, microbial degradation of brown algae begins with extracellular breakdown of the structural components (alginate and laminarin) into glucose, uronic acid, and mannitol (34, 35). The resulting mannitol (which is also the major photosynthate product of brown algae) and monosaccharides are transported into a cell via specific transporters. Glucose can be used via glycolysis while additional enzymes are required to assimilate mannitol (D-mannitol 1-phosphate 5-dehydrogenase, MPDH; 6-phosphofructokinase, PFK; and D-mannitol PTS).

*Nu*_PAG1 encodes for all the genes required for brown algae degradation (Table 4). Seven putative alginate lyase genes (PL7), two cytoplasmic exo-alginate lyases (PL15) and one oligoalginate lysase (PL17) were detected for the extracellular degradation of alginate. One putative endo-1,3-beta-D-glucosidase (GH16) was detected for the hydrolysis of laminarin. All three genes necessary for mannitol assimilation were present in the genome. *Nu*_PAG1 also encodes genes for the transport and use of simpler sugars including glucose and fructose but not galactose. No other genes for the breakdown of complex polysaccharides (e.g. agarase) were identified, suggesting that *Epulonibulbus* is a specialized brown algae degrader.

All alginate degradation genes were differentially transcribed over the course of the day (Fig. 5). The majority reached peak expression during EM1 and were significantly reduced by LAF. Two lyases remained highly expressed in the LAF

(Ga0192388_10754, Ga0192388_10291), suggesting that the substrate was still present and being used as a carbon source.

Table 4. Brown algae degradation genes detected by CAZyDB and IMG annotation.

Gene ID	Catalytic domain (qty)	Enzyme[1]	Reference match[1]	Other Domains (qty)	Signal Peptide[3]	Amino acid length
Alginate						
Ga0192388_10754	PL7	Alginate lyase	PDB:3NFV	-	-	383
Ga0192388_102178	PL7	Alginate lyase	PDB:3NFV	-	-	383
Ga0192388_119116	PL7	Alginate lyase	PDB:2CWS	SLH (2)	yes	1519
Ga0192388_119118	PL7	Alginate lyase	PDB:2CWS	GH120, SLH	yes	3134
Ga0192388_119161	PL7	Alginate lyase	PDB:2CWS	-	yes	320
Ga0192388_119201	PL7 (2)	Alginate lyase	PDB:2CWS	SLH (2)	-	2665
Ga0192388_119233	PL7 (2)	Alginate lyase	PDB:2CWS	-	-	795
Ga0192388_10291	PL15	Exo-alginate lyase	PDB:3A0O	-	-	704
Ga0192388_119636	PL15	Exo-alginate lyase	PDB:3A0O	-	-	698
Ga0192388_119635	PL17	Oligoalginate lyase	PDB:4NEI	-	-	627
Laminarin						
Ga0192388_119581	GH16	Endo-1,3-β-glucanase	PDB:4CRQ	SLH(3), CBM4 (6), CBM54, CBM56(2)	yes	2394
Mannitol						
Ga0192388_10836	-	D-mannitol PTS	-	-	-	621
Ga0192388_10837	-	D-mannitol 1-phosphate 5-dehydrogenase	-	-	-	379
Ga0192388_119488	-	6-phospho-fructokinase	-	-	-	325

[1]Enzyme identity was confirmed by sequence homology in the PDB database.
[2]Secretion signals were detected using SignalPeptide software.

Acidogenesis, solventogenesis and energy potential

Similar to previously described epulos (Sannino, unpublished & Chapter 2), *Epulonibulbus* has an incomplete TCA cycle, the capacity to ferment acetate and ethanol, and use of a sodium motive force (SMF) to generate ATP (Fig. 8). *Nu*_PAG1 encodes for sodium-translocating oxaloacetate decarboxylase (OAD) to covert oxaloacetate into pyruvate while transporting sodium out of the cell and a sodium-proton antiporter (NHA). Two types of ATP synthetases are encoded in the genome: V-type and F1-type. Additional pathways are present to help sustain the SMF via OAD. Multiple transporters of carboxylate intermediates (DCT, TCT, and DAACS) are encoded for in the genome that can supplement the TCA cycle. The citrate lyase pathway (CL) and specialized citrate-proton symporter (CitM) are predicted to convert imported citrate to oxaloacetate independent of the TCA cycle.

Daily transcriptional cycles of *Epulonibulbus* (Fig. 6) generally follow the metabolic trend previously observed for *Epulopiscium bomborense* and *Industriepulo chotii*, described in Chapter 2: uncoupling of the two fermentative stages, acidogensis (active metabolism) and solventogenesis (fasting). However, variation among some pathways were observed. The majority of carbon and carboxylate transporters as well as OAD were highly expressed in the morning but transcription levels decreased over the course of the day. The expression of genes for F-type ATP synthetase subunits peaked in the morning and appeared to be drop starting around AF. Conversely, expression of V-type ATP synthetases were lowest in the morning and appeared highly expressed at AF. In other epulos, both ATP synthetases were highly expressed in the morning and reduced at LAF. The difference may reveal a metabolically important distinction between spore-forming and non endospore-forming epulos. Further studies are needed to understand the roles of each ATP synthetase among epulos. Similar to other epulos, redox balance genes (rubrethyrin, thioredoxin, and

thioredoxin reductase) reach peak expression in the AF and LAF, suggesting a change in redox potential and/or oxygen stress.

Genes involved in acidogeneis and solventogenesis differ between non-endospore-forming epulos and endospore-forming *Epulonibulbus*. For non-endospore-forming epulos acidogenesis, marked by peak gene expression of acetate kinase (*ackA*), occured in the morning, while the marker for solventogenesis, alcohol dehydrogenase (*adhE*), peaked later in the day. However, the three genes that encode for AdhE in *Epulonibulbus* have variable transcriptional profiles. One gene (Ga0192388_117322) follows the solventogenic switch as described above, while Ga0192388_119516 peaked at EM1 and Ga0192388_10624 fluctuated during the day. In non-endospore-forming epulos, peak gene expression for solventogensis occurred during LAF and corresponded with peak expression of putative granulose-forming genes *(glgC*, *glgD*, and *glgP*) (Fig. 5). However, most granulose formation genes in *Epulonibulbus* were more highly expressed earlier in the day. Putative genes for carbon storage (*eutN*, *eutM*) were identified in the *Epulonibulbu*s genome. These genes are predicted to encode for a carboxysome shell and allow for ethanolamine utilization. Homologs are also found in other clade C epulo genomes from the NCBI database (Nuni2H_MBin001 and Nuni2H_MBin003). The carbon storage genes were differentially expressed at EM1_EM2 and EM2_AF. Furthermore, *Epulonibublus* lacks the acetoin buffering pathways that direct the conversion of pyruvate to acetoin/butanediol in *Epulopiscium bomborense* and *Industriepulo*. Peak expression of carbon storage and granulose synthesis genes early in the day by *Epulonibulbus* indicates a likely strategy to store carbon in the developing forespore. The presence of an acetoin buffering pathway in non-endopsore-forming epulos may have allowed for the evolution of intracellular offspring by providing protection against an acidifying mother cell, a hostile condition that spore-formers eventually escape.

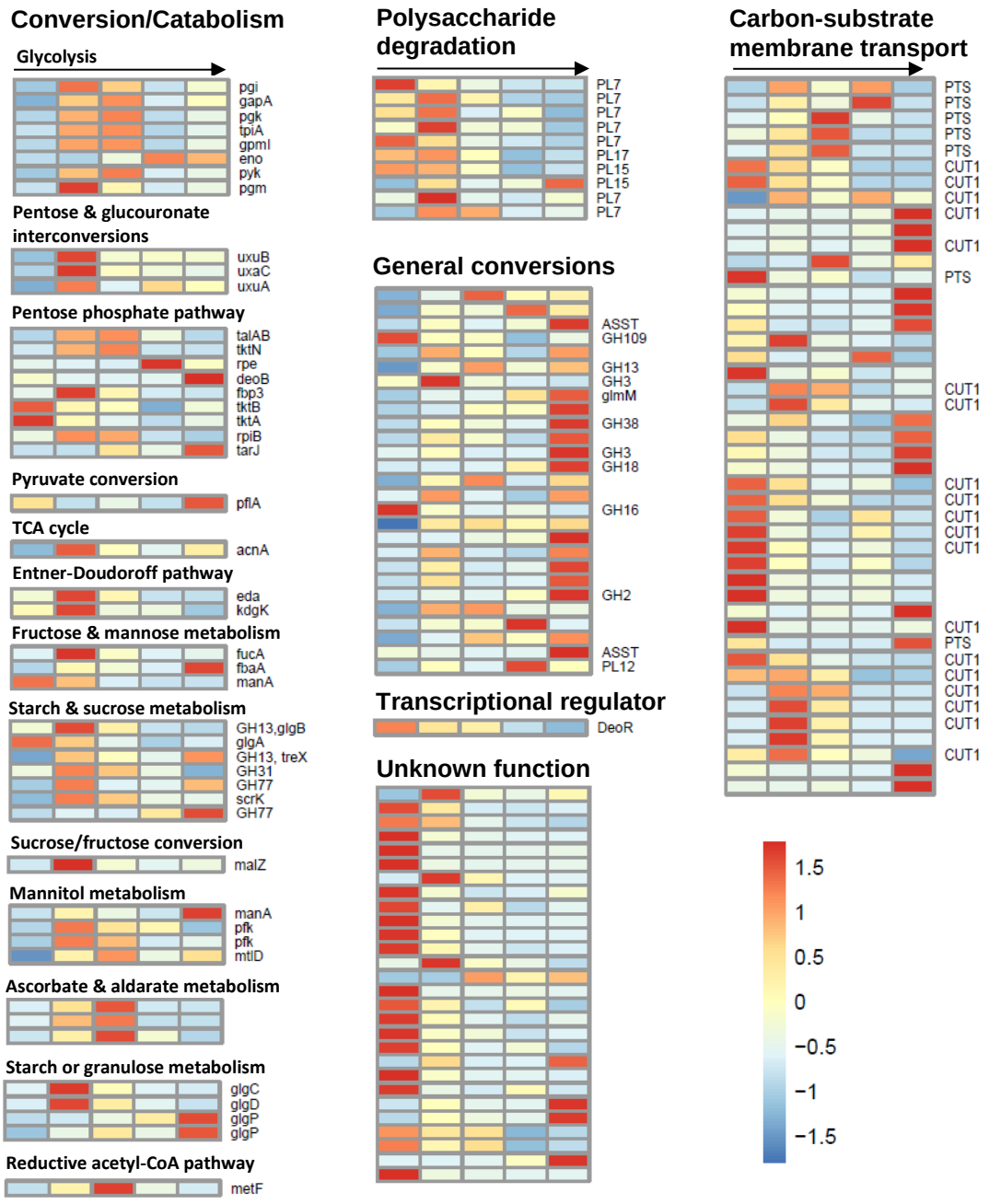


Figure 5. Transcriptional oscillations of carbon metabolism. Heatmaps showing significantly differentially expressed genes. Values are average RPKM for each consecutive time interval (EM1, EM2, MM, AF, LAF) and scaled by row. Genes are grouped in general categories: conversion/catabolism (e.g. TCA cycle, glycolysis, Pentose phosphate pathway), general conversions, membrane transport (e.g. carbon-substrate specific transporters), transcriptional regulators, and polysaccharide degradation (e.g. glycoside hydrolases and polysaccharide lyases).

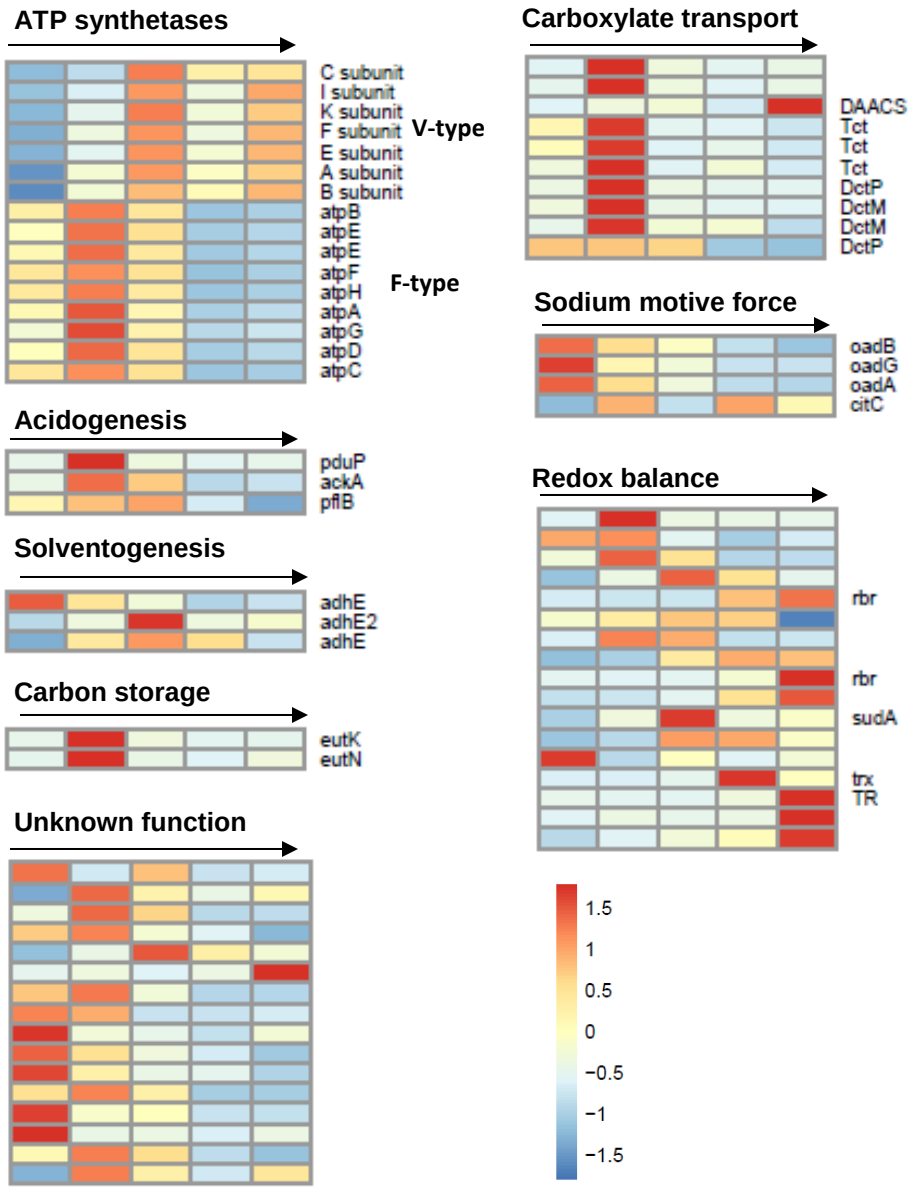


Figure 6. Transcriptional oscillation of energy metabolism. Heatmaps showing significantly differentially expressed genes. Values are average RPKM for each consecutive time interval (EM1, EM2, MM, AF, LAF) and scaled by row. Rows are annotated by general function: ATP synthetases (V-type and F_1F_0-type), carboxylate transporters (citrate-H^+ symporter, TCT, DCT), acidogenesis (e.g. acetate kinase), solventogenesis (e.g. Alcohol dehydrogenases), carbon storage, sodium motive force (eg. oxaloacetate decarboxylase), redox balance (including *rbr* – rubryethrin, *trx* – thioredoxin, and TR – thioredoxin reductase), and gene of unknown function.

Contribution of amino acids via urea degradation

Epulonibulbus is predicted to be an amino acid prototroph (Fig. 8) and have the capacity to use ammonia and urea as initial substrates. Although there are no studies available that characterize nitrogen metabolism and waste regulation in surgeonfish, studies of other teleost fish confirm the presence of both ammonia and urea in the gut lumen (36). Genome annotations predict two pathways for *Nu*_PAG1 to assimilate nitrogen. All epulos studied thus far contain ammonia transporters and can incorporate ammonia onto glutamate using glutamine synthase (GS). In *Epulonibulbus*, glutamate can be generated by converting glutamine to glutamate using glutamate synthase (GOGAT). Transport of glutamate varies among species wherein all have amino-acid specific-ABC transporters and a glutamate/Na^+ symporter. *Nu*_PAG1 codes for urea transporters and urease (*ureABC*) to convert urea to ammonia. Compared to previously described epulos, type C uniquely codes for an additional pathway composed of the enzymes urea carboxylase (UC: urea + ATP + HCO_3^- → urea-1-carboxylate) and allophanate hydrolase (AH: urea-1-carboxylate → $NH_3 + CO_2$). These genes occur in a gene cluster with additional nitrogen regulatory genes, NitT/TauT family transport genes, and hypothetical genes. Temporal expression patterns of these genes suggest that they might be in the same operon. All genes for urea degradation were highly expressed at EM2 and quickly dropped at AF (Fig. 7).

During the early morning stage, most genes related to ammonia transport, amino acid/oligopeptide transport and amino acid biosynthesis were most highly transcribed. This corresponds with an active carbon metabolism in the morning versus the evening. Synchronization of carbon and nitrogen activities during the initial feeding cycle of the host may provide epulos an advantage to use resources quickly

before conditions change drastically within a few hours.

A gene cluster containing *nifH* is present in the genome of type C and one of the type A1 *Industriepulo* genomes (*Al*_SAG7). The gene cluster encodes for the NifH nitrogenase, followed by an iron complex transport permease, an iron complex transport system ATP-binding protein, two *nifE* genes that encode a molybdenum-cofactor synthesis protein and an ABC-type Fe^{3+}-hydroxamate transport sytem substrate-binding protein. This *nifH* gene cluster does not appear to be the canonical *nifHDK* cluster that is used for nitrogen fixation in other bacteria. Phylogenetically the epulo *nifH* gene falls into the uncharacterized *nifH* cluster IV group. If this *nifH* gene is related to nitrogen fixation, expression of this gene would be inhibited by the presence of organic nitrogen sources or ammonia. Expression of the *Epulonibulbus nifH* peaked at LAF, which corresponded to times of reduced gene expression of organic nitrogen and ammonia transport. Although the timing of peak expression of *nifH* in Epulonibulbus may indicate a relation to nitrogen fixation, sequence similarity and lack of accessory *nif* genes (*nifD*, *nifK*) continues to suggest that the functional role of cluster IV *nifH* is unclear, and likely related to other metabolic processes.

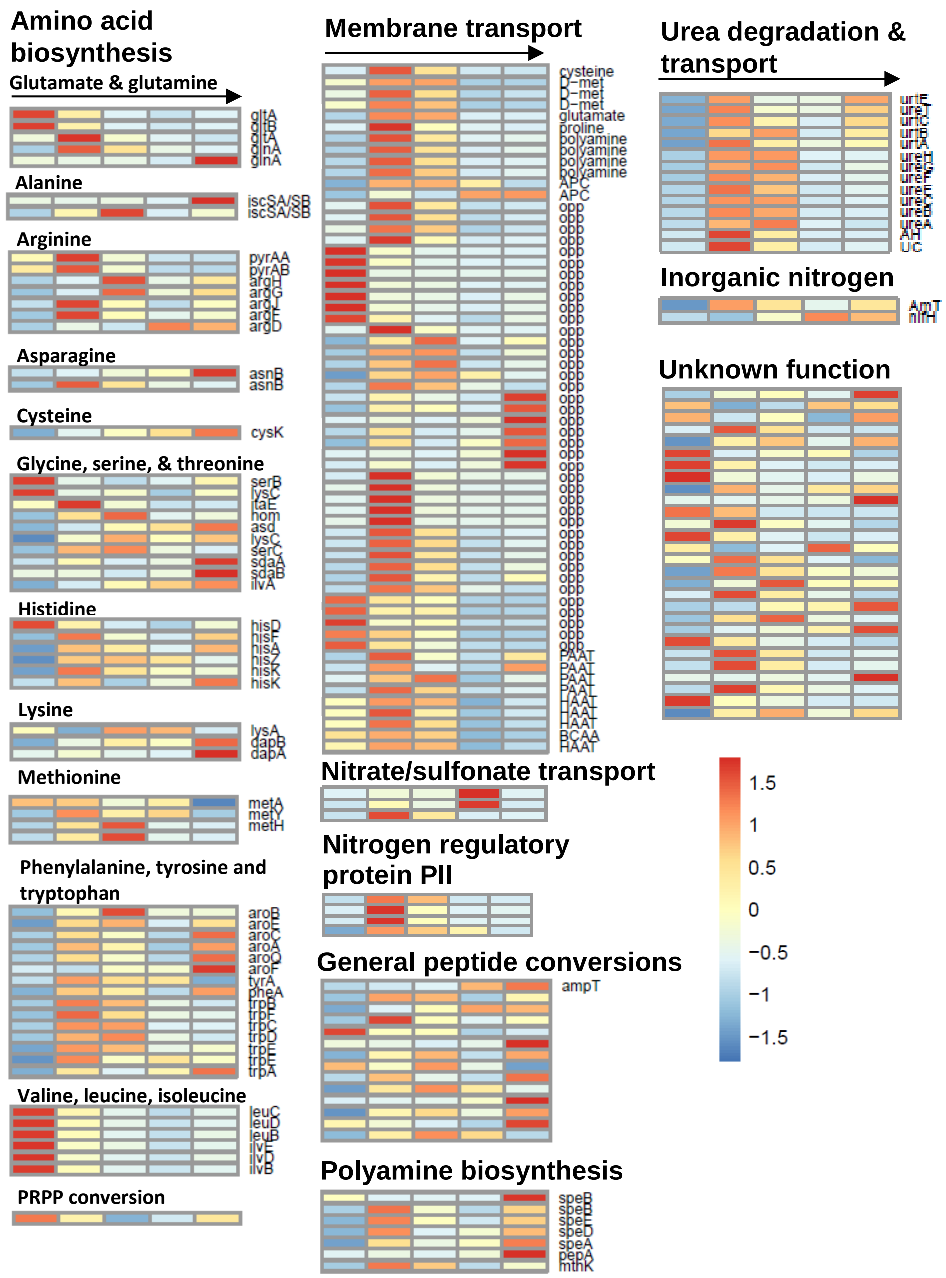
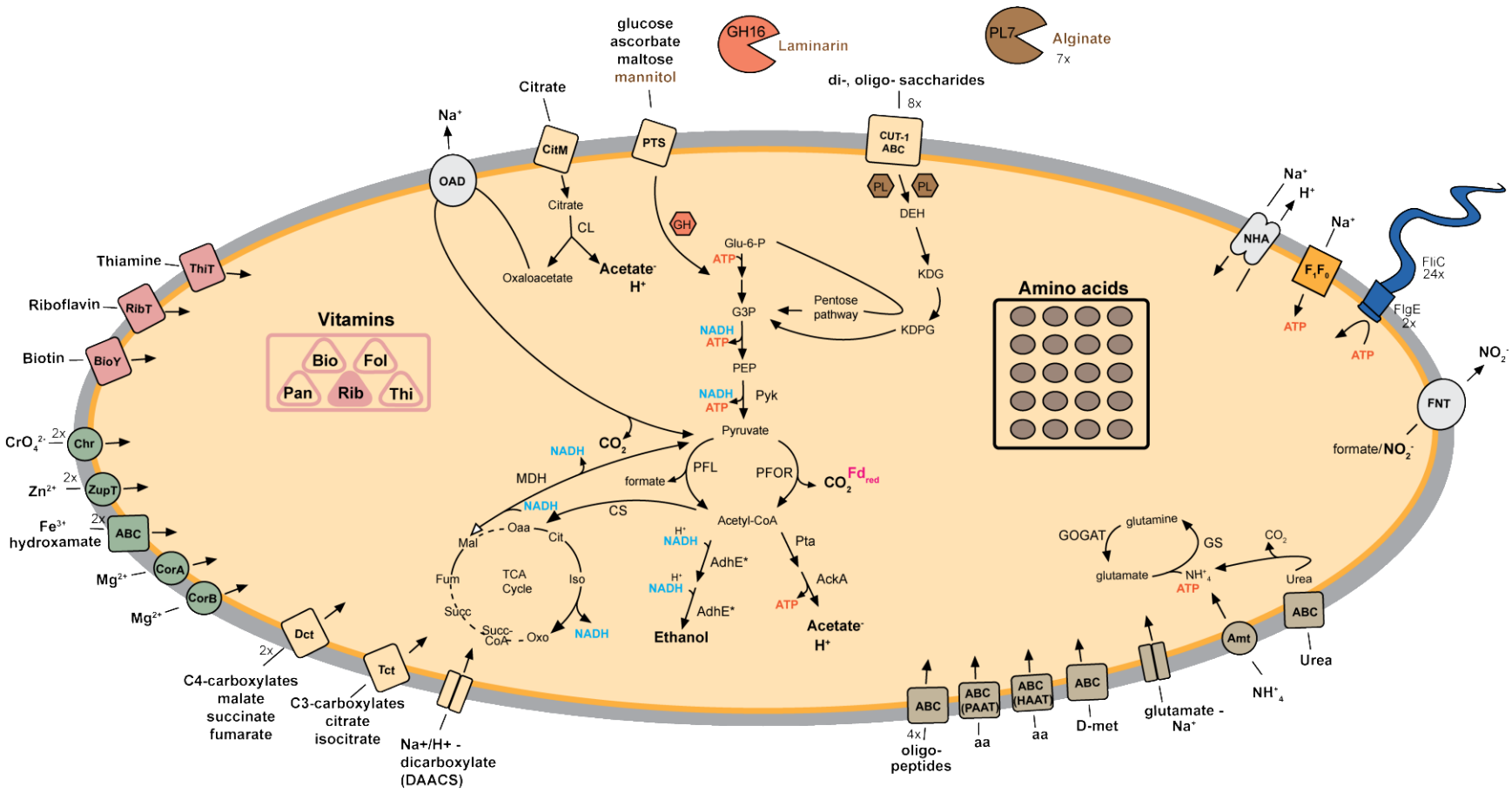


Figure 7. Transcriptional oscillations of nitrogen metabolism. Heatmaps showing significantly differentially expressed genes. Values are average RPKM for each consecutive time interval (EM1, EM2, MM, AF, LAF) and scaled by row. Genes are grouped in general categories: amino acid biosynthesis, membrane transport (e.g. oligopeptide ATP transporters and amino-acid ABC transporters), urea degradation and transport, nitrate/sulfonate transport, nitrogen regulatory protein (PII), inorganic nitrogen (ammonia transport & putative *nifH*), polyamine biosynthesis (amino acid degradation), and unknown functions.

Figure 8. Model of the carbohydrate and energy metabolism of "*Ca.* Epulonibulbus gigasporus" clade C1. Pathways were deduced from genome annotations and KEGG/Microcyc references. End products from fermentation are indicated in bold. Incomplete pathways are shown as dotted lines and reversible reactions contain white arrows. Amino acid biosynthesis capability shown as filled in circles and incomplete pathways are indicated by amino acid single letter codes. The potential to synthesize a particular vitamin is indicated by a shaded triangle, whereas incomplete pathways are not shaded. Putative enzymes are abbreviated as follows: Pyk, pyruvate kinase; PFOR, pyruvate synthase; PFL, formate C-acetyltransferase; CS, citrate synthase; MDH, malate dehydrogenase; PpdK, pyruvate orthophosphate dinkinase; PEPCK, phosphoenolpyruvate carboxykinase (ATP); Pta, phosphate acetyltransferase; AckA, acetate kinase; AdhE*, acetaldehyde dehydrogenase/alcohol dehydrogenas; GS, glutamine synthetase; GOGAT, glutamate synthase; OAD, Na^+ -translocating oxaloacetate decarboxylase; CitM, citrate-H^+ symporter; CL, ATP-requiring citrate lyase complex; GH, glycoside hydrolases; PL, polysaccharide lyases; F_1F_0, predicated Na^+ -dependent ATPase; DAACS, Na^+/H^+-dicarboxylate_symporter; TCT, tricarboxylate transporter (TTT) family; DCT, tripartite ATP-independent periplasmic C4-dicarboxylate transporter (TRAP-T) family; AmT, Ammonia transporter; PAAT, Polar amino acid transporter; HAAT, Hydrophobic amino acid transporter; FNT, formate/nitrate transporter; NHA, Na^+-H^+ antiporter; RifT, riboflavin-H^+ symporter; ThiT, thiamine transporter' BioY, Biotin transporter. Pathways are described in Chapter 2, Table 3.


glucose
ascorbate
maltose
mannitol
GH16
Laminarin
PL7
Alginate
7x
Citrate
di-, oligo- saccharides
8x
Na+
CitM
PTS
CUT-1
ABC
OAD
PL
PL
DEH
GH
Glu-6-P
ATP
KDG
Citrate
CL
Oxaloacetate
Acetate-
H+
Pentose
pathway
G3P
KDPG
NADH
ATP
PEP
NADH
ATP
Pyk
Pyruvate
NADH
CO2
MDH
formate
PFL
PFOR
CO2
Fdred
NADH
CS
Acetyl-CoA
Mal
Oaa
Cit
H+
NADH
Pta
Fum
TCA
Cycle
Iso
AdhE*
AckA
Succ
Succ-
CoA
Oxo
NADH
H+
NADH
AdhE*
ATP
Ethanol
Acetate-
H+
Vitamins
Bio
Fol
Pan
Rib
Thi
Thiamine
ThiT
Riboflavin
RibT
Biotin
BioY
CrO4 2-
2x
Chr
Zn2+
2x
ZupT
Fe3+
hydroxamate
2x
ABC
Mg2+
CorA
Mg2+
CorB
2x
Dct
C4-carboxylates
malate
succinate
fumarate
Tct
C3-carboxylates
citrate
isocitrate
Na+/H+ -
dicarboxylate
(DAACS)
Amino acids
Na+
H+
NHA
Na+
F1F0
ATP
ATP
FliC
24x
FlgE
2x
NO2-
FNT
formate/NO2-
GOGAT
glutamine
GS
CO2
glutamate
NH+4
ATP
Urea
Amt
ABC
Urea
NH+4
ABC
4x
oligo-
peptides
ABC
(PAAT)
aa
ABC
(HAAT)
aa
ABC
D-met
glutamate -
Na+

Making do with less: endosporulation in spore-forming epulos is tightly regulated and follows a daily cycle

The sporulation genes found in the *Epulopiscium* sp. type B (*E. viviparus*) draft genome were previously compared to the closest spore-forming relative with a complete genome *Cellulsilyticum lentocellum* that was available at that time (24). After comparing a core sporulation gene list of 146 genes between *B. subtilis*, *Ce. lentocellum*, and *Epulopiscium* sp. type B, 87 homologs were found in *Ce. lentocellum* and 57 homologs were found in type B. All the homologs found in epulo type B were present in *Ce. lentocellum*, having retained many of the genes for early stages of sporulation, but epulo B lacked genes for producing a mature spore. Comparison of this same gene list to spore-forming *Epulonibulbus* resulted in 73 homologs common to *Ce. lentocellum*. The sporulation genes present in type B were all found in *Epulonibulbus*, and the additional genes that were found in *Epulonibulbus* and not in type B were mainly from late stages of sporulation. This suggests that *Epulonibulbus* still manages to produce mature endospores despite the reduced core sporulation gene content compared to *Ce. lentocellum*.

Nearly all epulo genomes examined to date are missing σ^H , which regulates the transtion of *B. subtilis* cells from late expoential phase to sporulation. Only clade A1 genomes code for a homolog. There is a limited number of σ^H-dependent genes conserved between *B. subtilis* and *C. difficle* (37) and genes that were dual regulated by σ^A do not show a dependence on σ^H (38). These data suggest that σ^H may be conditionally essential for spore formation. The transition state may be more important to regulate solventogenesis in A1, which exhibits a fuller complement of genes involved in butanoate, propanoate metabolism than other epulos.

The fine scale transcriptional profile in this study (Fig. 9) captured transitions in the sporulation cascade that were not apparent in a previous epulo study (Chapter

2). The sporulation regulatory and signaling cascade in the *Epulonibulbus* roughly follows the scheme in other Clostridia (Fig. 10). The gene for the master regulator Spo0A, is transcribed around the onset of sporulation. In its phosphorylated form, Spo0A initiates the sequential activation of sporulation-specific sigma factors in the forespore (σ^F and σ^G) and the mother cell (σ^E and σ^K). As a cell progress through sporulation, it undergoes many of the morphological changes as have been described in *B. subtilis* and Clostridia (39–41). In *B. subtilis*, stage 0 occurs during normal vegetative growth while cells are undergoing binary fission and cells show no signs of forespore development. Stage I is often indicated by the expression of the Spo0A and the tethering of replication origins to the poles and formation of an axial filament. Polar division is evident during Stage II. The peptidoglycan between the mother cell and forespore is enzymatically degraded and the mother-cell membrane begins to wrap around the forespore. Stage III indicates complete engulfment of the forespore by the mother cell. Stage IV is cortex formation between the mother cell and forespore membranes. The cortex is a modified peptidoglycan (PG). Spore coat formation occurs during Stage V. A major component of the spore, dipicolinate acid (DPA), is synthesized and transported into the forepore. The developing spore matures during Stage VI. Stage VII encompasses mother-cell lysis and release of the mature spores.

Stage I-II: Spo0A expression and polar septa formation

Since *Epulonibulbus* does not undergo binary fission, and newly germinated cells already have forespores associated with the poles, stage 0 is omitted from their developmental cycle. In Clostridia, genes under the control of σ^H (*sigF-spoIIAA-spoIIAB* operon) are expressed in Stage I, indicating the start of both sporulation and solventogenesis. However, *Epulonibulbus* lacks the *sigH* gene and its consensus recognition sequence near the promoter region of *sigF-spoIIAA-spoIIAB* operon was not found. Yet, these hallmark genes (*sigF-spoIIAA-spoIIAB*) were upregulated during

the earliest time interval (EM1), along with other early sporulation genes (*spo0A, spoIIIE, jag, spoIIIJ, spoIIE*) that are under the control of the major housekeeping sigma factor, σ^A. This suggests that the *sigF-spoIIAA-spoIIAB* operon is regulated by σ^A in *Epulonibulbus*. The expression of *spoIIE* in epulos occurs earlier in development than in *Clostridium* and *Bacillus* spp. (Chapter 2, (42)). During this early time interval, phase-bright mature endospores are present in the *Epulonibulbus* population and some cells have condensed DNA or have divided at the poles. Bypassing σ^H-stationary phase regulation may have been selected to support rapid onset of sporulation in epulos so that the process can be completed in a time frame that complements host activity and environmental shifts.

Stage III: Engulfment

During EM2, genes under the control of σ^E were significantly differentially expressed. These genes include the *spoIIIAA-AH* which are expressed in the mother cell and involved in forespore engulfment. *Epulonibulbus* cells at this time interval were observed as having either condensed DNA at the polar ends of the mother cell or cells that have completed forespore engulfment. Genes related to spore structure peaked, including a factor involved in spore cortex formation (*ylbJ*), a sporulation-specific autolysin (*spoIIP*) and spore maturation/dehydration protein (*spmB*). The *sigE* gene was significantly differentially expressed, but RPKM and clear expression of the σ^E regulon indicate that this gene was expressed during EM2. In *B. subtilis*, an inactive form of this sigma factor (pro-σ^E) is first formed and requires cleavage by the combined interaction of the membrane-bound SpoIIGA protease (under σ^A control) and forespore localized SpoIIR (σ^F). Neither gene was significantly expressed, but RPKM values in a few samples, spanning the EM2-AF time interval, suggest that the two proteins may be present and play a role in pro-σ^E cleavage.

Stage IV: Cortex formation

A few genes from the forespore σ^F and σ^G regulons and mother cell σ^E-dependent genes were differentially expressed during AF. Genes that are under control of σ^F include *sigG* and the transcriptional regulator of σ^G-dependent genes (*spoVT*) as well as genes associated with later spore development such as a D-alanine-D-alanine carboxypeptidase that manages the degree of crosslinking of the spore PG (*dacF*), the regulatory protease for processing pro-σ^K (*spoIVB*), and *ytfJ* (hypothetical protein). Genes that are expressed under the control of σ^G include the DPA uptake genes (*spoVAA-AE*), an N-acetylmuramic acid deacetylase gene involved in generating muramic δ-lactam in the spore cortex PG (*pdaA*), spore germination receptor subunit genes (*gerKA-B*), and small acid-soluble spore proteins (*sspB* and *sspC/F*). Genes under σ^E control include additional cortex formation genes *dacB* and *cwlD*. At this stage of development, forespores have reached ~1/2 the length of the mother cell but were not phase bright.

Stage V: Spore coat formation

Nearing the end of daylight hours (LAF), *sigK*, σ^K-dependent genes, and σ^E-dependent *spoIIID* were significantly expressed. In *B. subtilis* and *C. difficile*, SpoIIID inhibits members of the σ^E regulon in the mother cell and activates *sigK* gene expression (43, 44). This positive interaction is also likely in *Epulonibulbus* since σ^K-dependent genes peak during this stage. Genes significantly expressed under σ^K control included a gene that encodes for DPA synthase subunit A (*spoVFA*), and two genes that encode a sporulation-specific ABC transporter (*ytlC* and *ytlD*). During this time interval, the majority of cells contain forespores that were nearly full grown within the mother cell and one sample (G106.2C) contained phase-bright endospores.

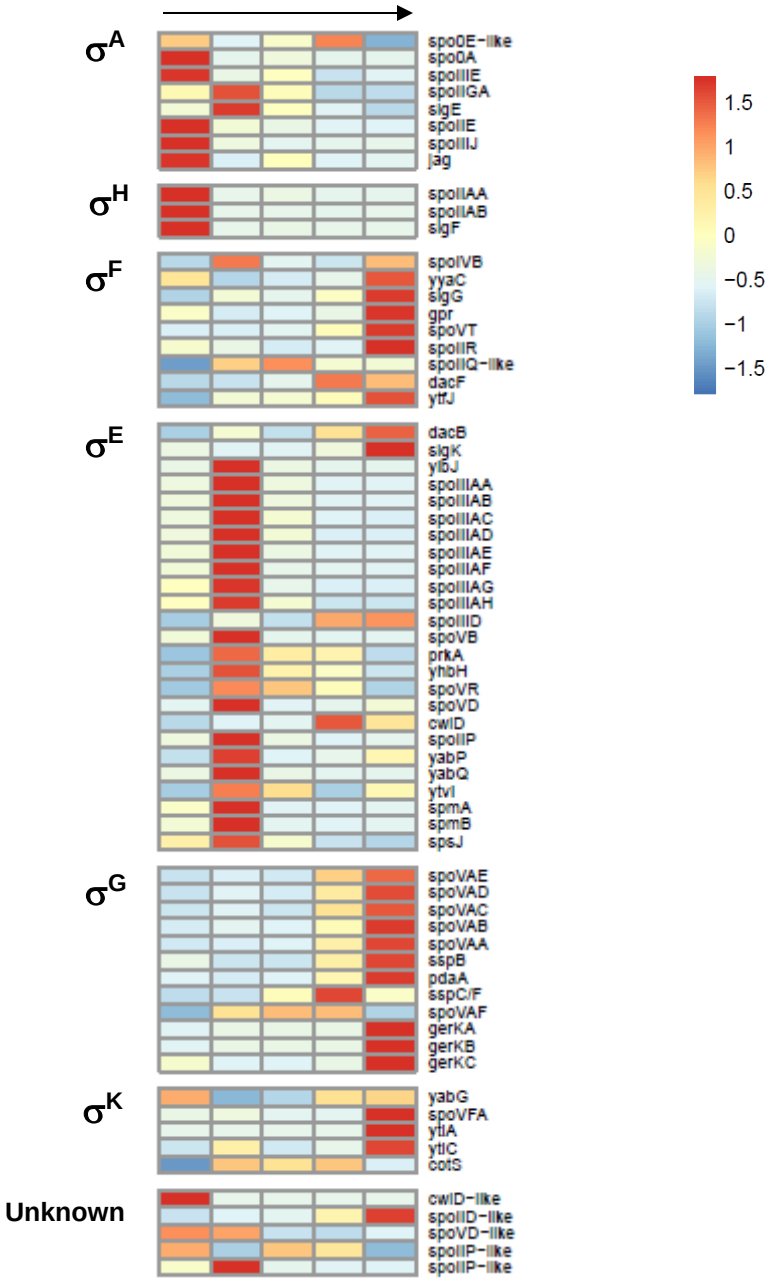


Figure 9. Transcriptional oscillations of sporulation genes. Heatmaps showing significantly differentially expressed (DE) genes. Genes that were not significantly DE are indicated by an asterisk. Values are average RPKM for each consecutive time interval (EM1, EM2, MM, AF, LAF) and scaled by row. Genes are grouped by key sigma factors: SigA, SigH, forespore-specific SigF & Sig G and mother cell-specific SigE and SigK. Unknown are duplicate genes with an unknown sigma factor dependency.

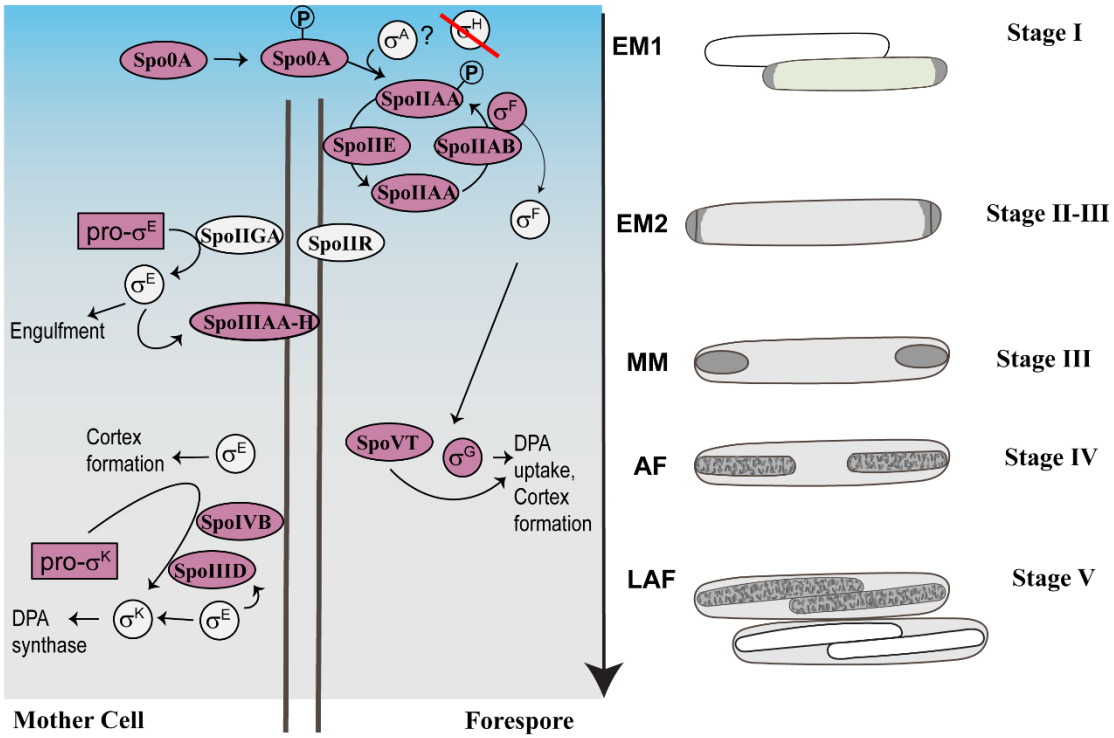


Figure 10. Summary of the sporulation cascade and cell development cycle of "*Ca.* Epulonibulbus gigaspora" clade C1. The daily progression of spore-formation is indicated by the large black arrow. Key sigma factors and sporulation-specific proteins localized in the mother cell or forespore are shown. Significantly differentially expressed transcripts that encode these proteins are highlighted in purple while proteins that were not significant are shown in white. Timing of the non-significant transcripts hypothesized based on peak average RPKM values. Genes that are encoded by differentially expressed transcripts are shown along with their cognate sigma factor. Cell development associated with these time intervals are illustrated along with predicted sporulation-related developmental stage (I-V). Vegetative cells are in grey and phase bright spores are in white. Condensed DNA of developing forespore is in dark grey. Polar septum are shown as vertical lines in stage II-III.

Multiple flagellin genes for an intestinal symbiont

Epulonibulbus contains 24 unique *fliC* genes which are mostly clustered in two separate loci. Each cluster contains additional flagellin-associated genes including the flagellin-specific chaperonin *fliS* gene. Expression of these genes varied over the course of the day, however the majority were expressed during EM1 along with flagellar motor genes *motAB,* hook-associated genes (*flgE*, *flgL*, *flgK*) (Fig. 11). Motility is often inversely correlated with later stages of sporulation in other spore forming lineages. Maintaining motility during the active phase of carbon and nitrogen metabolism would allow for cells to gain access to the nutrient source. Flagellin genes expressed later in the day may indicate movement away from inhospitable environmental conditions.

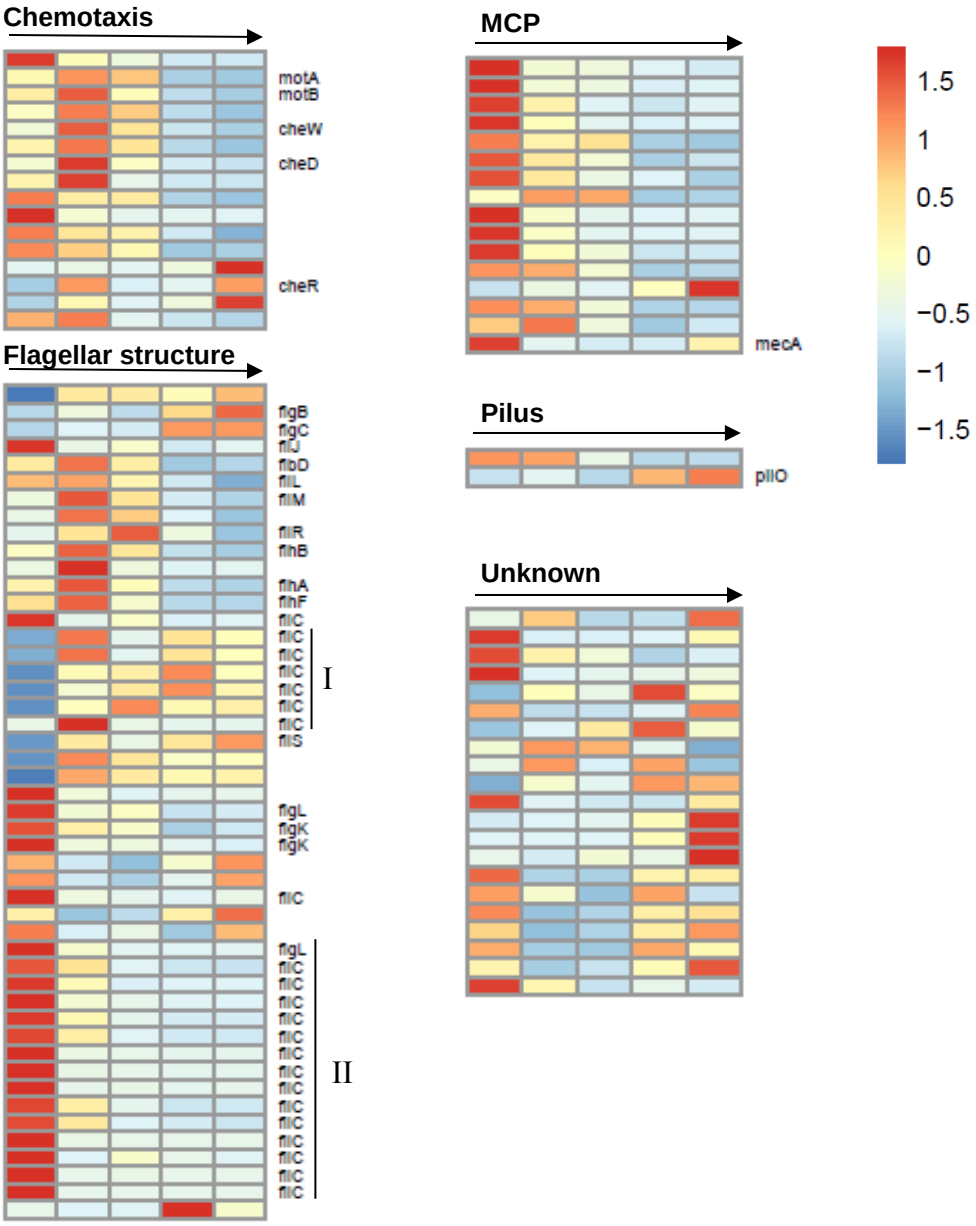


Figure 11. Transcriptional oscillation of motility-related genes. Heatmaps showing significantly differentially expressed genes. Values are average RPKM for each consecutive time interval (EM1, EM2, MM, AF, LAF) and scaled by row. Genes are grouped by function: flagellar structural genes (e.g. *fliW*, *fliC*), chemotaxis genes (e.g. *motA*, *cheA/D*), MCPs (Methyl-accepting chemotaxis protein), typeIV pilin genes (e.g. *pilB*), and genes of unknown function. The two largest flagellin-containing gene clusters are indicated by a vertical line.

Hydrogen and oxygen *in situ* transverse profiles shift with time

The most well studied models for microbial biogeography of the vertebrate gut are mice and humans (10, 45, 46). A physiochemical gradient occurs along both the longitudinal axis of the intestinal tract and the transverse axis in the small and large intestine. Along the longitudinal axis (proximal to distal), pH increases from acidic to near neutral while oxygen decreases. The cross-sectional oxygen gradient also decreases dramatically from the epithelium towards the lumen. Microbial populations show spatial heterogeneity along these gradients. Microbes that are able to maintain a close association with the host intestinal epithelium are heavily influenced by host defenses such as mucus and products of the immune system. Daily oscillations in the mammalian gut have been recently explored but have focused on community compositional changes to mucus-associated microbiota. However, the daily oscillations of important physiochemical gradients, such as oxygen, have not been investigated. The physiochemical gradients along the cross-sectional axis of *N. unicornis* were monitored over 12-hour period during the day (EM, MM, AF, and LAF) (Table 1). Section IV of the gut, which harbors the largest concentration of type C epulos, was analyzed to explore possible correlations of time-dependent metabolic activity and gut environment oscillations.

Oxygen decreases with depth away from the gut wall, with oxygen starting at 240 μm/L near the epithelial surface it became depleted at depths of 1000-3500 μm (Fig. 12A). At time intervals MM and AF, the oxygen concentration was the lowest and reached a shallower depth of depletion compared to EM and LAF (linear mixed model: $P < 0.05$). On the other hand, hydrogen concentrations increased with depth, starting at ~3 μm/L near the surface and increasing to 10-54 μm/L at depths of 2500-4000 μm (Fig 12B). Periodic differences were also significant between MM/AF and EM/LAF. MM/AF consistently maintained the lowest concentration of hydrogen

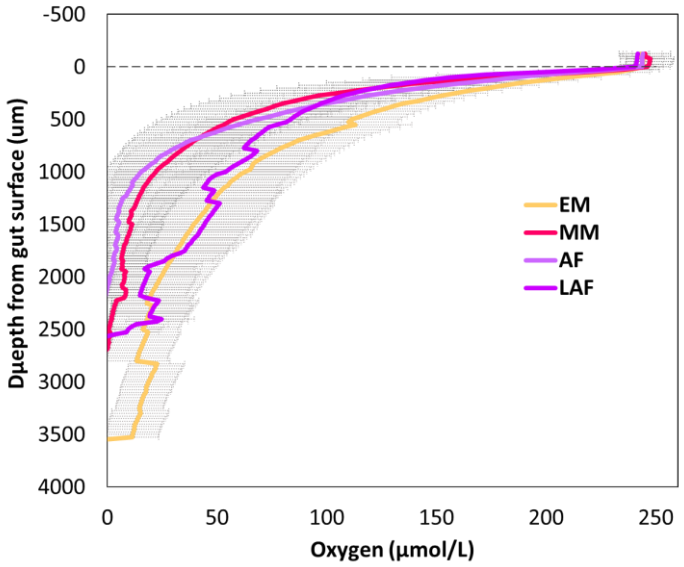
compared to EM/LAF.

The inverse relationship between oxygen and hydrogen concentrations across time suggests that host-microbe interactions are linked. Oxygen is likely diffusing from the host epithelium and circulatory system (45). Oxygen entering the lumen is consumed by facultative anaerobes, removed by non-respiratory activities by anaerobes (e.g. endogenous reductants or H_2-dependent acetogenesis) (47–49), or consumed by respiring gut epithelial cells. The high amount of oxygen observed at EM suggests that oxygen consumption was reduced during the night, thus allowing oxygen to accumulate. During MM, oxygen consuming reactions peaked, correlating with the lowest levels of oxygen observed. As the day continued, oxygen slowly accumulated as metabolic activities decline. Resistance to oxygen radicals via increased expression of *sod2* or redox enzymes (e.g. rubrerythrin) as the day proceeds might allow epulos to colonize more diverse gut locations, including sites close to the intestinal wall, despite the presence of oxygen.

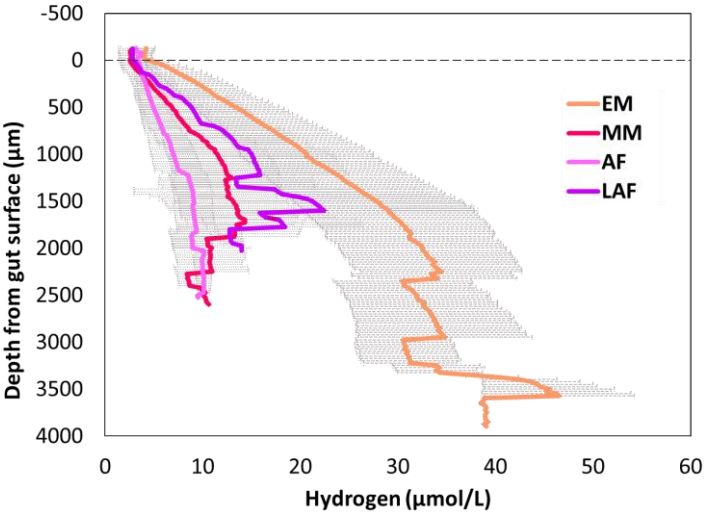
Hydrogen generation and consumption are likely driven by other community members of the gut. *Epulonibulbus* does not possess the mixed fermentation pathways that generate hydrogen. Future community surveys would help identify key players of this process. Methane was not detected, suggesting that methanogens are not present to impact hydrogen concentrations. Redox measurements were also collected, however shifts in redox potential were not observed at measured depths (Fig. 12C).

Figure 12. Time series gut profile of (A) oxygen concentration, (B) hydrogen concentration, and (C) redox potential. Time of day are 0600-0900 (EM), 0900-1200 (MM), 1200-1500 (AF), 1500-1800 (LAF). Values are means and error bars represent standard deviation (morning n = 2; afternoon n = 3). To determine if these environmental parameters showed evidence of periodicity, we constructed a linear mixed model of concentration/mV at five depths (500-2500μm) as a function of time of day. Significant periodicity ($P < 0.05$) was observed for oxygen and hydrogen concentrations when grouped as early morning-late (EM & LAF) vs middle of the day (MM & AF).

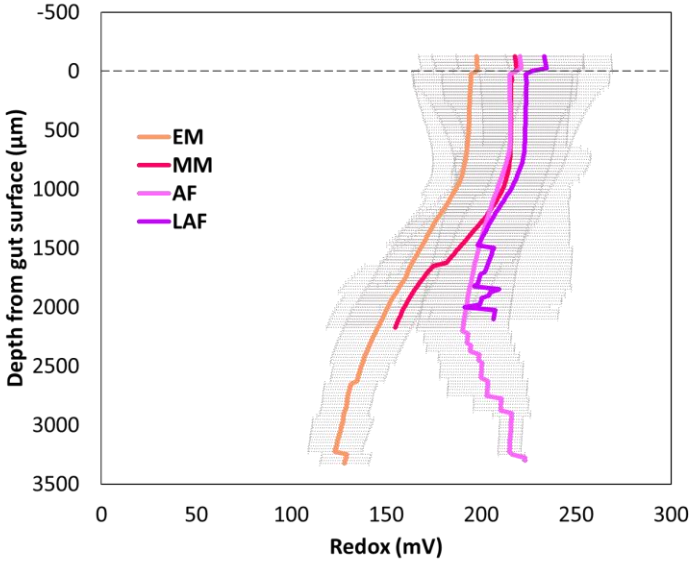
A


-500
0
500
1000
1500
2000
2500
3000
3500
4000
Dμepth from gut surface (um)
0
50
100
150
200
250
Oxygen (μmol/L)
EM
MM
AF
LAF

B


-500
0
500
1000
1500
2000
2500
3000
3500
4000
Depth from gut surface (μm)
0
10
20
30
40
50
60
Hydrogen (μmol/L)
EM
MM
AF
LAF

C


-500
0
500
1000
1500
2000
2500
3000
3500
Depth from gut surface (μm)
0
50
100
150
200
250
300
Redox (mV)
EM
MM
AF
LAF

DISCUSSION

Natural populations of epulos exhibit a daily synchronization of cell development over a broad range of surgeonfish species. A diversity of metabolic and reproductive strategies have allowed epulos to adapt to their competitive and changing environment. Surgeonfish follow a diurnal activity and feeding cycle, which begins at dawn and ends at sunset. Gut environment fluctuations in herbivorous surgeonfish over the course of the day are evident. In the morning, proximal intestinal regions (i.e. sections II – IV) are mostly empty, thin, and full of fluid. As the day progresses, algal fodder accumulates (32). As the fish sleeps, the proximal gut empties and a large amount of partially digested food remains in the distal gut. At the onset of feeding, material that was retained in the intestinal tract is expelled (50). Symbiont transcriptional profiles follow the feeding and fasting cycles of the host. Active metabolism is indicated by the peak expression of genes that encode for brown algae degradation (polysaccharide lyases, mannitol assimilation, and laminarinase), acidogenesis, carbon and nitrogen transport, amino acid biosynthesis, urea degradation, F-type ATP synthetase, SMF generation via OAD, and motility. Whereas preparation for fasting is indicated by the high expression of genes for solventogenesis, maintaining redox balance (rubrerythrin, thioredoxin, and thioredoxin reductase) and the reduced expression of the active phase genes. The transcriptional shift from genes for the active phase to fasting occurs quickly and early. The most significant differential expression of genes were between the EM and every other time interval examined. This time period spans only a couple of hours of daylight and yet had the most dynamic changes in transcription. *In situ* monitoring of hydrogen and oxygen concentrations in the gut also showed a clear shift from EM compared to other time points. Altogether, these data suggest that epulos must have had to adapt to this

rapid change by anticipating the diurnal activity of the host. Further studies are needed to determine if the metabolic synchrony of the epulo symbionts are a result of entrainment or as a response to the host's diurnal feeding behavior. Criteria for a true circadian oscillator requires persistent oscillation under constant conditions, entrainment or resetting of oscillations to light/dark cycle by environmental cues, and temperature compensation (5, 51, 52).

This is the first study to characterize the sporulation cascade of a spore-forming epulo, "*Ca*. Epulonibulbus gigaspora" clade C1. The high resolution time intervals informed our understanding of DE for 88% of the sporulation-specific genes identified in *Epulonibulbus*. The sporulation cascade was tightly correlated with microscopic observations of cell development. Stages I-V were identified over the course of the 12-hour time span. The sporulation cascade followed a progression similar to other Clostridia, albeit within a shorter time scale. The absence of *sigH* and earlier expression of *spoIIE* hastened the process compared to other Clostridia. The reproductive cycle of *Epulonibulbus* has adapted to the demands of keeping rhythm with the host. Sporulation is energetically costly. By corresponding metabolic activity and carbon storage with early stages of sporulation, epulos maintain the metabolic requirements to complete the sporulation cycle. Production of spores provides epulos with the option to remain in the gut of their host during the fasting period or be transmitted to another host.

The ecological success of this commensal is observed by its worldwide distribution in herbivorous surgeonfish. Among four *Naso* species, *Epulonibulbus*, in particular, is globally distributed and found across the Pacific Ocean and the Red Sea. The quick and well managed sporulation cascade may not only enhance *Epulonibulbus* population resilience, compared to non-spore-forming epulos, but it may also have a positive impact on the distribution and success of its surgeonfish hosts.

REFERENCES

1. Bell-Pedersen D, et al. (2005) Circadian rhythms from multiple oscillators: lessons from diverse organisms. *Nat Rev Genet* 6(7):544–556.

2. Ito H, et al. (2009) Cyanobacterial daily life with Kai-based circadian and diurnal genome-wide transcriptional control in *Synechococcus elongatus*. *Proc Natl Acad Sci U S A* 106(33):14168–73.

3. Ottesen EA, et al. (2014) Ocean microbes. Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial assemblages. *Science* 345(6193):207–12.

4. Stal LJ, Krumbein WE (1985) Nitrogenase activity in the non-heterocystous cyanobacterium *Oscillatoria* sp. grown under alternating light-dark cycles. *Arch Microbiol* 143:67–71.

5. Grobbelaar N, Huang TC, Lin HY, Chow TJ (1986) Dinitrogen-fixing endogenous rhythm in *Synechococcus* RF-1. *FEMS Microbiol Lett* 37(2):173–177.

6. Ottesen EA, et al. (2013) Pattern and synchrony of gene expression among sympatric marine microbial populations. *Proc Natl Acad Sci U S A* 110(6):E488-97.

7. Thaiss CA, et al. (2014) Transkingdom control of microbiota diurnal oscillations promotes metabolic homeostasis. *Cell* 159(3):514–529.

8. Zarrinpar A, Chaix A, Yooseph S, Panda S (2014) Diet and feeding pattern affect the diurnal dynamics of the gut microbiome. *Cell Metab* 20(6):1006–1017.

9. Liang X, FitzGerald GA (2017) Timing the microbes: The circadian rhythm of the gut microbiome. *J Biol Rhythms* 32(6):505–515.

10. Thaiss CA, et al. (2016) Microbiota diurnal rhythmicity programs host transcriptome oscillations. *Cell* 167(6):1495-1510.e12.

11. Moeller AH, Suzuki TA, Phifer-Rixey M, Nachman MW (2018) Transmission modes of the mammalian gut microbiota. *Science* 362(6413):453–457.

12. Chase DG, Erlandsen SL (1976) Evidence for a complex life cycle and endospore formation in the attached, filamentous, segmented bacterium from murine ileum. *J Bacteriol* 127(1):572–83.

13. Fishelson L, Montgomery WL, Myrberg Jr. A (1985) A unique symbiosis in the gut of tropical herbivorous surgeonfish (Acanthuridae: Teleostei) from the Red Sea. *Science* 229(4708):49–51.

14. Clements KD, Sutton DC, Choat JH (1989) Occurrence and characteristics of unusual protistan symbionts from surgeonfishes (Acanthuridae) of the Great Barrier Reef, Australia. *Mar Biol* 102(3):403–412.

15. Angert ER (2005) Alternatives to binary fission in bacteria. *Nat Rev Microbiol* 3(3):214–224.

16. Flint JF, Drzymalski D, Montgomery WL, Southam G, Angert ER (2005) Nocturnal production of endospores in natural populations of *Epulopiscium*-like surgeonfish symbionts. *J Bacteriol* 187(21):7460–70.

17. Angert ER, Clements KD, Pace NR (1993) The largest bacterium. *Nature* 362(6417):239–41.

18. Montgomery WL, Pollak PE (1988) *Epulopiscium fishelsoni* N. G., N. Sp., a protist of uncertain taxonomic affinities from the gut of an herbivorous reef fish. *J Protozool* 35(4):565–569.

19. Angert ER, Clements KD (2004) Initiation of intracellular offspring in *Epulopiscium*. *Mol Microbiol* 51(3):827–35.

20. Kearse M, et al. (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28(12):1647–9.

21. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–10.

22. Huntemann M, et al. (2015) The standard operating procedure of the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4). *Stand Genomic Sci* 10(1):86.

23. Markowitz VM, et al. (2012) IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res* 40(D1):D115–D122.

24. Miller DA, Suen G, Clements KD, Angert ER (2012) The genomic basis for the evolution of a novel form of cellular reproduction in the bacterium *Epulopiscium*. *BMC Genomics* 13:265.

25. Guindon S, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59(3):307–21.

26. Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA Sequences. *Am Math Soc Lect Math Life Sci* 17:57–86.

27. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–20.

28. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Available at: http://arxiv.org/abs/1303.3997.

29. Ngugi DK, et al. (2017) Genomic diversification of giant enteric symbionts reflects host dietary lifestyles. *Proc Natl Acad Sci U S A* 114(36):E7592–E7601.

30. Bowers RM, et al. (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* 35(8):725–731.

31. Choat JH, Clements KD, Robbins W. (2002) The trophic status of herbivorous fishes on coral reefs I. Dietary analysis. *Mar Biol* 140(3):613–623.

32. Choat JH, Robbins WD, Clements KD (2004) The trophic status of herbivorous fishes on coral reefs: II. Food processing modes and trophodynamics. *Mar Biol* 145(3):445–454.

33. Clements KD, Choat JH (1995) Fermentation in tropical marine herbivorous fishes. *Physiol Zool* 68(3):355–378.

34. Ji S-Q, Wang B, Lu M, Li F-L (2016) *Defluviitalea phaphyphila* sp. nov., a novel thermophilic bacterium that degrades brown algae. *Appl Environ Microbiol* 82(3):868–877.

35. Thomas F, et al. (2012) Characterization of the first alginolytic operons in a marine bacterium: from their emergence in marine Flavobacteriia to their independent transfers to marine Proteobacteria and human gut Bacteroides. *Environ Microbiol* 14(9):2379–2394.

36. Bucking C, LeMoine CMR, Craig PM, Walsh PJ (2013) Nitrogen metabolism of the intestine during digestion in a teleost fish, the plainfin midshipman (*Porichthys notatus*). *J Exp Biol* 216(15):2821–2832.

37. Britton RA, et al. (2002) Genome-wide analysis of the stationary-phase sigma factor (sigma-H) regulon of *Bacillus subtilis*. *J Bacteriol* 184(17):4881–90.

38. Saujet L, Monot M, Dupuy B, Soutourina O, Martin-Verstraete I (2011) The key sigma factor of transition phase, SigH, controls sporulation, metabolism, and virulence factor expression in *Clostridium difficile*. *J Bacteriol* 193(13):3186–3196.

39. Higgins D, Dworkin J (2012) Recent progress in *Bacillus subtilis* sporulation. *FEMS Microbiol Rev* 36(1):131–148.

40. Hilbert DW, Piggot PJ (2004) Compartmentalization of gene expression during *Bacillus subtilis* spore formation. *Microbiol Mol Biol Rev* 68(2):234–62.

41. Stragier P, Losick R (1996) Molecular genetics of sporulation in*Bacillus subtilis*. *Annu Rev Genet* 30(1):297–341.

42. Miller DA, Choat JH, Clements KD, Angert ER (2011) The *spoIIE* homolog of *Epulopiscium* sp. type B is expressed early in intracellular offspring development. *J Bacteriol* 193(10):2642–6.

43. Saujet L, et al. (2013) Genome-wide analysis of cell type-specific gene

transcription during spore formation in *Clostridium difficile*. *PLoS Genet* 9(10):e1003756.

44. Eichenberger P, et al. (2004) The program of gene transcription for a single differentiating cell type during sporulation in *Bacillus subtilis*. *PLoS Biol* 2(10):e328.

45. Donaldson GP, Lee SM, Mazmanian SK (2015) Gut biogeography of the bacterial microbiota. *Nat Rev Microbiol* 14(1):20–32.

46. Vaishnava S, et al. (2011) The antibacterial lectin RegIIIgamma promotes the spatial segregation of microbiota and host in the intestine. *Science* 334(6053):255–8.

47. Tegtmeier D, Thompson CL, Schauer C, Brune A (2016) Oxygen affects gut bacterial colonization and metabolic activities in a gnotobiotic cockroach model. *Appl Environ Microbiol* 82(4):1080–1089.

48. Boga HI, Brune A (2003) Hydrogen-dependent oxygen reduction by homoacetogenic bacteria isolated from termite guts. *Appl Environ Microbiol* 69(2):779–86.

49. Riebe O, Fischer R-J, Wampler DA, Kurtz DM, Bahl H (2009) Pathway for H_2O_2 and O_2 detoxification in *Clostridium acetobutylicum* . *Microbiology* 155(1):16–24.

50. Clements KD (1991) Gut microorganisms of surgeonfishes (Family Acanthuridae). Dissertation (James Cook University of North Queensland).

51. Huang TC, Tu J, Chow TJ, Chen TH (1990) Circadian rhythm of the prokaryote *Synechococcus* sp. RF-1. *Plant Physiol* 92(2):531–3.

52. Schweiger H-G, Hartwig R, Schweiger M (1986) Cellular aspects of circadian rhythms. *J Cell Sci Suppl* (4):181–200.

APPENDIX A

PHYSIOCHEMICAL ANALYSIS OF THE HINDGUT CHAMBER FROM *POMACANTHUS SEXSTRIATUS*

METHODS

P. sexstriatus was caught by spear fishing near Lizard Island, Great Barrier Reef, Australia (Table 1). Fish were kept on ice until processing within hours of capture. The intestinal tract was removed from the fish and kept intact during microsensor monitoring and sampling of the hindgut (Fig. 1). *In vivo* hindgut concentrations of hydrogen and sulfide were monitored using microsensors (Unisense, Aarhus, Denmaark) with a 25 µm tip diameter. The hydrogen microsensor was calibrated using hydrogen gas-saturated water concentrations of 0, 10, 50, and 100 µM. The sulfide microsensor was calibrated using a 0.1 M citric acid buffered and anaerobic solution (N2-flushed) containing 0, 100, 300, and 500 µM Na2S. Duplicate readings were collected at 25 µM-depth intervals from the surface of the hindgut to a maximum depth of 1 mm.

Methane was measured by gas chromatography (Perking-Elmer 3920B) and calibrated at 1000 ppm. Duplicate hindgut contents (~1 g) were collected and weighed in 20-ml, N2-flushed vials and sealed. The headspace (1mL) from these vials were immediately transferred to a secondary 20-mL, N2-flushed and sealed vial. Although the gut-containing vials were refrigerated, the secondary vials serve as a more conservative measure of the concentration of methane at the time of sampling. The concentration of methane from the secondary vials is reported in table 2.

Due to the fragility of the microsensors, duplicate hindgut contents (~1g) were also collected for the purpose of measuring sulfide concentrations. These contents were aliquoted to 20-ml vials containing 1M HCl, flushed with N2 gas, sealed and

mixed thoroughly. The headspace (1 mL) was immediately transferred to a sealed, N2-flushed, 20-ml vial containing 1% zinc acetate (1 mL). Sulfide concentrations were determined using the colorimetric assay described by Cline (1969) and measurements were performed on a UV/vis spectrophotometer (Beckman Coulter DU730).

For pH, hindgut samples were collected in cryovial tubes and immediately frozen in liquid nitrogen. Samples will be measured using a pH microsensor (50 µM tip diameter) calibrated with commercial pH standard solutions of pH 4.0, 7.0, and 10.0. This analysis is still pending.

Table 1. Sampling summary and description of *P. sexstriatus.*

Fish#	Time caught	Date	Hindgut depth (cm)	Sample location	Sex
G37	1000	12/3/2016	2.5	Day Reef (N)	male
G68	1000	12/6/2016	2.0	North Direction	male
G69	1000	12/6/2016	2.0	North Direction	immature
G95	1000	12/9/2016	2.0	Channel entrance	female
G96	950	12/9/2016	2.2	Bird Islet	female

Figure 1. The hindgut of *P. sexstriatus*. The red arrow indicates the location for microsensor monitoring and sampling. The posterior end of the hindgut is clamped by a hemostat.

RESULTS

The hindgut chamber contains detectable concentrations of methane, hydrogen and sulfide (Table 2). Concentrations of methane, hydrogen, and sulfide vary widely between fish but variation within fish is minimal. There does not appear to be any correlation between the concentrations of these gases.

Table 2. Physiochemical conditions of the hindgut chamber given as a weighted subset or *in vivo* reading from microsensors.*

Fish#	nmol/g fresh wt.		*In vivo* (µmol/L) @ depth 0.8 mm	
	Methane	Sulfide	Hydrogen	Sulfide
G37	260 ± 75	-	32.5	90.4 ± 42.0
G68	1,487 ± 614	956 ± 525	28.4 ± 1.2	-
G69	990 ± 239	3934 ± 456	35.8 ± 3.5	-
G95	245 ± 9	118 ± 45	47.9 ± 4.6	-
G96	577 ± 29	116 ± 17	17.8 ± 0.2	-
Average	712 ± 559	1281 ± 1608	32.5 ± 10.7	

* Values are mean ± standard deviation for n = 2.