High-Throughput Cultivation of Bacterioplankton from the Gulf of Mexico and Genomics of the First Cultured LD12 Representative

Michael Winslow Henson
Louisiana State University and Agricultural and Mechanical College

Follow this and additional works at: https://digitalcommons.lsu.edu/gradschool_dissertations

Part of the Biology Commons, and the Environmental Microbiology and Microbial Ecology Commons

Recommended Citation

This Dissertation is brought to you for free and open access by the Graduate School at LSU Digital Commons. It has been accepted for inclusion in LSU Doctoral Dissertations by an authorized graduate school editor of LSU Digital Commons. For more information, please contact gradetd@lsu.edu.
HIGH THROUGHPUT CULTIVATION OF BACTERIOPLANKTON FROM THE GULF OF MEXICO AND GENOMICS OF THE FIRST CULTURED LD12 REPRESENTATIVE

A Dissertation

Submitted to the Graduate Faculty of the Louisiana State University and Agricultural and Mechanical College in partial fulfillment of the requirements for the degree of Doctor of Philosophy in The Department of Biological Sciences

by
Michael Winslow Henson
B.A., Miami University, 2011
M.S., Central Michigan University, 2014
May 2019
This dissertation is dedicated to my brothers and sister-in-law, JP Henson, Matthew Henson, Bridgett Murphey Henson, to my parents, John and Michele Henson, and in the memories of my grandmother, Sarah Henson, (1924-2016) And grandfather, John Henson (1926-2007)
I am immensely indebted to Dr. J. Cameron Thrash, chair of my dissertation committee and mentor. Without him, my dissertation would have never been as successful as it has been. Dr. Thrash’s strive to put out strong, reproducible research has helped me develop into the researcher I am today. Further, his commitment to student development and health created a friendly environment that evoked the best in all the students in the lab. Dr. Thrash has made me a better thinker, writer, scientist, and more importantly a better person. During down times, when the world seemed dark and cold, he provided an open door and an ear for listening. I would also like to thank my committee members, Dr. Gary King, Dr. William Moe, and former member Dr. Brent Christner, for their guidance throughout my time at LSU. Their thoughtful feedback and questions pushed me to become a better researcher. Also, thank you Dr. D’Sa for providing an outside perspective and helping improve this dissertation.

I would also like to thank my funding support: the LSU Graduate School, LSU Department of Biological Sciences, the American Museum of Natural History, the Louisiana Environmental Education Comission, and The Louisiana State Board of Regents.

To my lab “sister” Celeste, thank you for always listening, giving me feedback, providing the never endings laughs, reminding me to leave the lab, and all the good times. Without those, I am not sure I would have stayed as happy and successful in my Ph.D. as I did. Thank you to Zach Rodriguez, my roommate, friend, and colleague. You have taught me many new lessons in life and reminded me that there is nothing better than a little bit of cooking, dancing, and egregious signing. You listened to my bickering even when it wasn’t necessary. Thank you for putting up with me for these last five years. Thanks also to the current Thrash Lab members Jordan Coelho and Alex Hyer for your support. A special thanks to all the Thrash lab undergraduates, past and present, for your help; especially, Anna Lucchesi, David Pitre, Emily Nall, and Jessica Weckhort. Without the UGs in the lab, none of my dissertation would have started or been able to proceed at the speed it did.

During my time at LSU, I have received enormous professional and personal support from many people both associated with LSU and beyond. Specifically, I would also like to thank all of my many friends and colleagues at LSU in the Department of Biological Sciences and at the Museum of Natural Sciences for providing both a strong social and academic support network.

To Dr. Jorge Santo Domingo, friend and mentor, I owe my career to you. Way back when you took a chance on me and gave me a position in your lab at the EPA. Without that opportunity, I would have never made it in this profession. The lessons and techniques learned in your lab have shaped the scientist I am today. And to Dr. Deric Learman, thank you for providing me with every opportunity I could have asked for. Because of you, I developed the knowledge and skills to navigate academia. Your continued support has been essential to my success.

Finally, I owe everything to my entire family (John, Michele, Matt, JP, and Bridget). To my Dad, thank you for being my rock. You taught me so much throughout these years. You had challenged me to do better when I didn’t think I could and supported me when I stumbled. Your love for this family created a bond amongst us all that has allowed us to thrive in the most challenging circumstances. You have shown me that no matter what challenges are put in front of you, being kind, loving, and caring is always first. To my mom, you are an inspiration every day. Since Dad’s stroke, you have shown me the importance of dedication, hard work, and love.
Without you, I am not sure I could say that I am a Ph.D. graduate. You have shown me the power of love and taught me to fight for what I want. You gave me the competitive edge to compete, no matter the opponent, but also taught me to be humble and gracious. To my brothers, thank you for the “gentle” reminders that no one is perfect, to stay humble, and to put my head down and go to work. Jp, without you, I would be in business. You reminded me to pursue my passions and not give up. Matt, you showed me my first peeks into Microbiology. I wouldn’t have been able to do it without you. You have led by example and are a scientist that I am humbled to be able to call my brother. And finally, to Bridgett, thank you for the daily reminders to stay grounded and for putting up with all that comes with being a Henson.
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS ............................................................................................................. iii

ABSTRACT ................................................................................................................................. vi

CHAPTER 1. INTRODUCTION .................................................................................................. 1

CHAPTER 2. ARTIFICIAL SEAWATER MEDIA FACILITATES CULTIVATING MEMBERS OF THE MICROBIAL MAJORITY FROM THE GULF OF MEXICO ............................................. 5

CHAPTER 3. CULTIVATION AND GENOMICS OF THE FIRST FRESHWATER SAR11 (LD12) ISOLATE ......................................................................................................................... 15

CHAPTER 4. COASTAL GULF OF MEXICO HIGH-THROUGHPUT CULTIVATION EXPANDS CULTIVAR DIVERSITY ........................................................................................................... 33

CHAPTER 5. CONCLUSIONS AND FUTURE DIRECTIONS ....................................................... 50

REFERENCES ............................................................................................................................ 53

APPENDIX A. SUPPLEMENTAL MATERIAL FOR CHAPTER 2 .................................................. 77

APPENDIX B. SUPPLEMENTAL MATERIAL FOR CHAPTER 3 .................................................. 81

APPENDIX C. SUPPLEMENTAL MATERIAL FOR CHAPTER 4 .................................................. 98

APPENDIX D. PERMISSION TO REPRODUCE CHAPTER 2 .................................................... 106

APPENDIX E. PERMISSION TO REPRODUCE CHAPTER 3 .................................................... 107

VITA ............................................................................................................................................ 108
ABSTRACT

Cultivation of microorganisms facilitates characterization of metabolism, interspecies dependencies, virus-host interactions, and other information necessary to resolve the functions and distribution of individual taxa. However, the metabolic and physiological capacities for the majority of microbes remains unresolved because of the lack of cultivated representatives for many groups limits our ability to test cultivation-independent observations. The Northern Gulf of Mexico offers a diversity of ecosystems under the continuous threat from natural and anthropogenic disturbances, yet little is known about its native bacterioplankton community. This dissertation sought to use high-throughput cultivation over three-years at six sites to isolate important coastal bacteria to uncover their role in biogeochemical cycling and ecosystem health. During the seventeen experiments, 7820 wells were inoculated, resulting in 328 repeatedly transferable cultivars. Isolates were placed into 49 monophyletic groups based on 16SR rRNA gene sequences, and represent multiple novel Species and Genera, including the first reported cultures of the SAR11 LD12 Alphaproteobacteria, OM241 Gammaproteobacteria, and acIV Actinobacteria clades. Cultivars also contribute to the expansion of cultured diversity of numerous cosmopolitan bacterioplankton such as SAR11 subclade IIIa and SAR116 Alphaproteobacteria, and BAL58, MWH-UniP1, and OM43 Betaproteobacteria. Physiological and genomic characterization of the first cultivated LD12 representative, Candidatus Fonsibacter ubiquis strain LSUCC0530, yielded novel insights into the potential metabolic capacity related to sulfur, ecotype differentiation based on temperature, as well as key gene losses associated with osmoregulation that provide a concise hypothesis for the evolution of salinity tolerance in SAR11. Comparison between the success of isolation and the relative abundance of the cultivar in the source water revealed that relative abundance was a good predictor of cultivation success for some frequently cultured clades, while it was unreliable for rarely cultivated clades such as SAR11 subclades IIIa, LD12 Alphaproteobacteria, acIV Actinobacteria, and HIMB59-type Alphaproteobacteria. We hypothesize that taxon-specific variations in dormancy and/or phenotypic variation rates among populations may affect the cultivation reliability of that clade rather than the isolation medium alone. The insights from this dissertation provide a new look at the complexity of cultivation even when providing an organism with its nutritional requirements while showcasing the importance of cultivation for answering ecological questions like the evolution of salinity tolerance.
CHAPTER 1.
INTRODUCTION

Understanding bacterioplankton and their genetic and metabolic potential is essential to predicting microbial community responses to natural and anthropogenic disturbances (Cho and Azam, 1988; Azam and Malfatti, 2007; King et al., 2015). Studies aimed at understanding the physiological and metabolic capacities of bacterioplankton require both in situ and in vitro observations to help validate the hypotheses generated by one another (Giovannoni and Stingl, 2007a). However, the vast majority of microorganisms remain uncultivated (Rappé and Giovannoni, 2003; Giovannoni and Stingl, 2007a; Epstein, 2013; Solden et al., 2016; Lloyd et al., 2018a), complicating cross-validation of in situ observations. This phenomenon was described in Staley and Konopka (1985) as the “great plate count anomaly,” where < 1% of taxa observed in the environment were present on agar plates (Staley and Konopka, 1985). Since 1985, we now know microbial diversity is much more vast than it was believed. In 1990, the sequencing of twenty 16S rRNA gene clones fundamentally changed our understanding of the bacterioplankton in our global ocean by identifying the previously unknown SAR11 clade (Giovannoni et al., 1990), that was later found to comprise up to 50% of surface bacterioplankton in ocean environments (Morris et al., 2002). Since then, the veil over microbial diversity has continued to be pulled away, and we are now confronted with staggering numbers of taxa known mostly only by gene sequence data. This study advances the task of the modern microbiologist to more clearly define the roles of previously uncultivated microorganisms.

The northern Gulf of Mexico (nGOM) provides an excellent study system for this effort because of its economic and ecological value (Rabalais et al., 1996; Adams et al., 2004), as well as its dynamic coastline that is subject to continuous natural and anthropogenic disturbances (Bianchi et al., 1999; Rabalais et al., 2002; King et al., 2012). While researchers have examined bacteria and archaea within the context of anthropogenic disturbances for their underlying metabolic capacities, community enrichment, and remediation potential (Bælum et al., 2012; Mason et al., 2012; Dubinsky et al., 2013; Gutierrez et al., 2013; Bristow et al., 2015; Gillies et al., 2015; Cameron Thrash et al., 2017), the natural bacterioplankton found throughout the coastal nGOM remain understudied (Olapade, 2010; King et al., 2012; Tolar et al., 2013). Prior to the Deepwater Horizon oil spill, some of the first samples analyzing natural microbial communities using molecular approaches in the nGOM and along the Louisiana coast found nGOM water dominated by Alphaproteobacteria and Bacteriodetes, specifically finding surface water dominated by SAR11 (Olapade, 2010; King et al., 2012). The surface microbial communities resembled those from other coastal and marine ecosystems (Cram et al., 2015; Chafee et al., 2018); however, that study was only a snapshot of the microbial communities during the 2010 season. It remains unknown how the natural bacterioplankton in the nGOM fluctuate over time, their roles in biogeochemical cycles, and their potential genetic and metabolic differences compared to other closely related organisms from other parts of the globe.

Motivated by the difficulty of bringing the majority of taxa into culture, numerous cultivation-independent techniques have been developed to explore the hidden diversity of ecosystems. These include metagenomics, metatranscriptomics, and single cell genomics (Handelsman, 2004; Rinke et al., 2013). Studies using these techniques have unearthed and described numerous abundant bacterioplankton clades and their potential functions found throughout aquatic ecosystems such as acl and OM1 Actinobacteria, SAR86 Gammaproteobacteria, and SAR406, a.k.a. Marinimicrobia (Rappé et al., 1999; Giovannoni and
These techniques have demonstrated the ebb and flow of these poorly understood taxa within temporal and spatial contexts, their corresponding transcripts, and the impacts of disturbances surrounding them (Azam and Malfatti, 2007; Pommier et al., 2007; Rusch et al., 2007; Gilbert et al., 2010; Rappé, 2013; Swan et al., 2013; Fuhrman et al., 2015). However, any hypotheses garnered continue to remain untested without experimental inquiry of cultivated isolates (Rappé and Giovannoni, 2003; Giovannoni and Stingl, 2007a). Therefore, the continued development of cultivation methods to isolate and characterize members from important uncultivated clades remains essential.

Towards this effort, many novel cultivation techniques have been developed and employed, such as diffusion chambers allowing the exchange of in situ nutrients (Kaeberlein et al., 2002; Bollmann et al., 2007; Steinert et al., 2014), sorting single cells based on characteristics like size and DNA content into natural media (Wang et al., 2009), encapsulation of cells in gel microdroplets for continuous metabolite and nutrient exchange (Zengler et al., 2002; Bruns et al., 2003), cell dilution and inoculation into natural seawater (Button et al., 1993; Connon and Giovannoni, 2002), and, simply, the use of low-nutrient agar plates (Hahnke et al., 2015). All of these different methodologies share the approach of using a low nutrient medium that mimics oligotrophic conditions.

The justification for lowering nutrient concentrations stems from the fact that most aquatic systems do not contain nutrients at the levels found in typical media used for cultivation (Poindexter, 1981; Giovannoni and Stingl, 2007a). As a result, many of the bacterioplankton isolated with these methods represent rare copiotrophic taxa that thrive in high-nutrient conditions (Poindexter, 1981; Overmann et al., 2017). Previously, many bacterioplankton in the ocean were theorized to be in states of starvation or inactivation, but Button et al., (1993) suggested these oligotrophic organisms may have unique genomic content and physiology characteristics that inhibited growth in high nutrient environments (Button et al., 1993; Schut et al., 1993). To avoid typical high nutrient media, cells were diluted to 1-10 cells in tubes containing unamended, autoclaved, 0.2 µm filtered natural seawater (Button et al., 1993). Button and colleagues predicted that by using natural oligotrophic conditions and diluting to near single cells, the fast-growing, less abundant copiotrophic organisms would not outcompete the obligate oligotrophic bacterioplankton. From these experiments, cultivability improved from the < 1% associated with the “great plate count anomaly” to 2-60% (Button et al., 1993). However, while novel oligotrophic bacterioplankton were isolated using this approach (Wang et al., 1996; Vancanneyt et al., 2001), this method was laborious and as such inhibited large-scale experiments aimed at cultivating numerous bacterioplankton.

In 2002, dilution-to-extinction cultivation was updated to a more efficient format by using 48 well plates coupled with corresponding 48 well cell arrays stained with DAPI and used to score growth with microscopy (Connon and Giovannoni, 2002; Rappé et al., 2002). The resulting high-throughput cultivation (HTC) approach led the isolation of the abundant and previously uncultivated marine clades SAR11 Alphaproteobacteria, OM43 Betaproteobacteria, and SAR92 and OM60 Gammaproteobacteria (Connon and Giovannoni, 2002). Subsequent improvements added flow cytometry, nutrient supplements, longer incubation times at lower temperatures, and Teflon plates (Stingl et al., 2007; Song et al., 2009). Together, these changes were essential to facilitate the screening of a high number of wells while avoiding culture materials (e.g. plastic) and cleaning detergents known to inhibit bacterial growth, and incorporating genomic and physiological characteristics to help improve cultivation of important
marine clades such as SAR11 (Stingl et al., 2007; Song et al., 2009). Other HTC efforts have continued to expand cultivar diversity of marine and freshwater bacterioplankton with the addition of cultivated representatives of the abundant acl Actinobacteria (Kim et al., 2018), the OPB35 soil group (Yang et al., 2016), and various oligotrophic marine Gammaproteobacteria (Cho and Giovannoni, 2004; Yang et al., 2016).

Many clades remain recalcitrant to cultivation efforts (e.g., SAR86, OM1), because of their unusual metabolisms, the difficulty in maintaining or providing appropriate growth conditions, slow growth rates, or metabolic interdependencies (Button et al., 1993; Connon and Giovannoni, 2002; Giovannoni and Stingl, 2007a; Epstein, 2013; Solden et al., 2016; Lloyd et al., 2018b). For example, the cultivation of HTCC1062 on artificial medium provided evidence of non-canonical metabolisms and specific nutrient requirements of the SAR11 clade that potentially led to the lack of cultivation and were difficult to predict using genomic-based evidence (Tripp et al., 2008; Carini et al., 2013; Paul Carini, Campbell, et al., 2014). The isolation and cultivation of Nitrosopumilus maritimus, the first mesophilic marine Group 1 Crenarcheota representative, was dependent on a targeted approach that prevented bacterial growth and selected for ammonia oxidation metabolism based on cultivation-independent observations and enrichment of Group 1 Crenarcheota in ammonia-oxidizing cultures (Könneke et al., 2005). In two different studies, researchers found that the addition of various growth factors (e.g., cAMP, LQPEV) improved cultivation efficiencies and induced growth of bacteria previously uncultivated by other experiments (Bruns et al., 2002; Nichols et al., 2008). Moreover, during the recent cultivation of the abundant freshwater acl Actinobacteria, catalase addition was necessary to maintain axenic growth despite the presence of the gene in the cultivar genomes (Kim et al., 2018). While not exhaustive, these exemplify the difficulty in cultivating novel organisms and the importance of using environmental conditions and genomic content as guides for developing innovative cultivation methods.

Experiments with environmental relevant isolates provide important context to the biological and biogeochemical anomalies observed in the environment. For example, scientists would not have been able to explain the “marine methane paradox” (Karl et al., 2008; Metcalf et al., 2012; Paul Carini, White, et al., 2014) or uncover some of the intricacies of phytoplankton-heterotrophic bacteria interactions (Aharonovich and Sher, 2016; Segev et al., 2016) without the cultivation of important phototrophic and heterotrophic marine clades. While some puzzles have been solved, many still need to be pieced together (Overmann et al., 2017). One question waiting for culture-based experimental testing is how physiological differences between ecotypes impact their distribution and abundance in dynamic coastal ecosystems (Fuhrman et al., 2015; Needham and Fuhrman, 2016; Chafee et al., 2018). While we can associate the fluctuations between ecotypes with environmental conditions or the presence and absence of certain taxa (Vergin et al., 2013; Needham and Fuhrman, 2016; Chafee et al., 2018), the underlying functional differences are harder to validate.

Cultivation-based experiments are also needed to explore how changes in salinity, a major driver of microbial community structure, has the potential to impact biogeochemical cycling by changing the functional capacity of the bacterioplankton present (Lozupone and Knight, 2007; Logares et al., 2009; Dupont et al., 2014). Cultivation-independent methods have shown that transitions between freshwater and marine environments cause switches in cellular energetic pathways and de novo synthesis versus uptake of various osmolytes and compounds (Dupont et al., 2014; Eiler et al., 2015).
Using environmentally relevant or closely related cultivars with disparate ecologies, scientists can conduct axenic and co-culturing experiments testing substrate utilization, gene expression, metabolite production, and competitive exclusion across various environmental conditions (e.g., salinity, temperature). These experiments could illuminate unknown or difficult to discern physiologies and metabolisms that fundamentally restrict a cultivar's distribution, as well as provide valuable insights into their evolutionary histories. Therefore, scientists must remain steadfast in cultivating both new organisms and different representative ecotypes within cultivated clades.

This dissertation sought to elucidate the genetic content, metabolic capacities, and physiologies of important bacterioplankton that inhabit the nGOM and differentiate them from closely related organisms from other systems. To do this, this study used a modified HTC methodology combined with cultivation-independent approaches, sampling from multiple sites along the Louisiana coastal nGOM. Because of the difficulty associated with using natural seawater and its limitations for use in the characterizations of organisms, defined, modular artificial media were developed and deployed in the HTC protocol. The efficacy of the cultivation approach was examined by comparing cultivars with cultivation-independent data from sampling sites. Then using in vitro experiments and comparative genomics, the most important novel cultivar obtained by these efforts, LSUCC0530, was studied to determine unique genomic and physiological traits associated with its distribution, as well as generate a hypothesis about the evolutionary transition between marine and freshwater habitats. The legacy of this project is over 300 isolates of abundant, co-occurring bacterioplankton from the nGOM, dozens of which have been genome-sequenced. Ongoing studies of several of these taxa are providing exciting information about their novel roles in coastal biogeochemical cycles.
CHAPTER 2.
ARTIFICIAL SEAWATER MEDIA FACILITATES CULTIVATING MEMBERS OF THE MICROBIAL MAJORITY FROM THE GULF OF MEXICO

2.1 INTRODUCTION

The study of microorganisms and their roles in remediation and biogeochemical cycling requires the observation of microbial communities and genetics in nature coupled with experimental testing of hypotheses both in situ and in laboratory settings. The latter is best accomplished by cultivation of microorganisms, yet the majority of microorganisms observed under a microscope are not readily cultivated (Staley and Konopka, 1985; Rappé, 2013; Fuhrman et al., 2015). High throughput, dilution-to-extinction culturing (HTC) with natural seawater has been responsible for the isolation of strains representing numerically abundant clades such as SAR11 (Rappé et al., 2002; Stingl et al., 2007), SUP05/Arctic96BD-19 (Marshall and Morris, 2013), OM43 (Giovannoni et al., 2008; Sosa et al., 2015), a small genome Roseobacter (Durham, Grote, Whittaker, Bender, Luo, Grim, Brown, Casey, Dron, Florez-Leiva, Krupke, et al., 2014), and others (Hahnke et al., 2015). HTC benefitted from the oligotrophy theory and research of Don Button and colleagues (Button et al., 1993; Schut et al., 1993), which demonstrated that because marine systems are frequently in a state of nutrient limitation, adaptation by the microbial denizens has led to unusual genomics and physiology which must be accounted for in cultivation strategies. These oligotrophic microorganisms are generally small, with streamlined genomes that eschew many complicated regulatory systems (Giovannoni et al., 2014), resulting in slow-growth and complicated nutrient requirements, including poor or no growth in high carbon concentrations (Button et al., 1993; Schut et al., 1993; Rappé et al., 2002; Giovannoni et al., 2008; Carini et al., 2012; Hahnke et al., 2015). Thus, successful HTC experiments have provided a non-competitive growth environment with naturally occurring compounds at in situ concentrations.

Until now, this has been accomplished predominantly via filtration and autoclaving of natural seawater using protocols (e.g. (Button et al., 1993; Schut et al., 1993; Connon and Giovannoni, 2002) that avoided excessive nutrients found in “traditional” artificial marine media (Giovannoni and Stingl, 2007b) (Table 2.3), and provided complex, but largely unknown, natural dissolved organic matter (DOM) components (Sosa et al., 2015). However, in spite of its success, there are many drawbacks to using a natural seawater medium. First, it requires regular access to large volumes (20 L or more) of seawater, which can be a major logistical hurdle for research labs not located near a coastal source and/or without vessel access. Second, natural seawater is inherently complex and undefined, thus limiting the amount of physiological characterization that can be accomplished post isolation (e.g., salinity growth optima or single-carbon substrate utilization). Third, seawater at a given location may experience significant intra- and interannual chemical fluxes, thereby creating “vintages” from specific sample collections that can prevent reproducible growth or repeated transfers when attempting to passage an organism from one vintage to another. We therefore developed a complex, yet defined, low

1 This chapter previously appeared as Henson MW, Pitre DM, Weckhorst JL, Lanclos VC, Webber AT, Thrash JC. 2016. Artificial seawater media facilitate cultivating members of the microbial majority from the Gulf of Mexico. mSphere 1(2):e00028-16. Reprinted with the permission of my co-authors and American Society for Microbiology.

2 This chapter previously appeared as Henson, Michael W., V. Celeste Lanclos, Brant C.
nutrient set of artificial seawater media (ASM) incorporating the theory of Button and colleagues and published measurements for marine systems including the Gulf of Mexico (GoM) that account for variable salinity in coastal ecosystems (Justič et al., 1995; Rabalais et al., 1996; Dagg et al., 2004; Carini et al., 2012; Thrash et al., 2015) (Tables S1).

2.2 RESULTS AND DISCUSSION

To create the JW1 medium (Tables 2.1 and 2.3), the basic salt mix was derived using values in the medium designed by Kester et al., 1967 (Kester et al., 1967). Concentrations for phosphate, iron, and trace metals were taken from a defined medium for SAR11 (Carini et al., 2012), with modified vitamin concentrations referencing data in Moore et al., 2007 and Sañudo-wilhelmy et al., 2012. DOM is one of the most important, yet least understood components of natural seawater, having varied carbon and nitrogen constituents that can be difficult to distinguish (Kujawinski, 2011). A complex but defined list of carbon and nitrogen compounds were used to generate our modular DOM mixes (Table S1), and the total carbon and nitrogen concentrations were based on previously reported GoM data (Justič et al., 1995; Rabalais et al., 1996; Dagg et al., 2004). Ultimately, our modular recipe allows for customization of the different carbon and nitrogen components, making it easily adaptable to a variety of environments.

Using the media in Table S1, we conducted seven HTC experiments with a combined total of 3360 distinct cultivation wells according to the protocol in (Justič et al., 1995; Rabalais et al., 1996; Dagg et al., 2004; Carini et al., 2012; Thrash et al., 2015), 2015, and included community characterization (via 16S rRNA gene tag sequencing) of the source water for each experiment. These experiments used water collected from six sites along the southern Louisiana coastline (Fig. 2.1) that represent varied estuarine-marine systems (Twilley et al., 1999). Our goals in sampling these different sites were to provide varied source inocula and to simultaneously collect coastal northern GOM microbial biogeography data. Initial cultivability (Button et al., 1993) ranged from 0-53.1% (Table 2.2), and 82 of the 231 positive cultures across all experiments were capable of repeated transfer after initial exchange from PTFE to polycarbonate flasks, and/or did not contain multiple organisms. The 82 isolates represent
### Table 2.1. JW1 medium constituents

<table>
<thead>
<tr>
<th>Basic components</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cl</td>
<td>0.54 M</td>
</tr>
<tr>
<td>Na</td>
<td>0.48 M</td>
</tr>
<tr>
<td>Mg</td>
<td>52 mM</td>
</tr>
<tr>
<td>S (SO$_4^{2-}$)</td>
<td>30 mM</td>
</tr>
<tr>
<td>Ca</td>
<td>10 mM</td>
</tr>
<tr>
<td>K</td>
<td>10 mM</td>
</tr>
<tr>
<td>Br</td>
<td>0.8 mM</td>
</tr>
<tr>
<td>B</td>
<td>0.42 mM</td>
</tr>
<tr>
<td>Sr</td>
<td>0.09 mM</td>
</tr>
<tr>
<td>F</td>
<td>0.055 mM</td>
</tr>
<tr>
<td>Fe</td>
<td>101 nM</td>
</tr>
<tr>
<td>P (PO$_4^{3-}$)</td>
<td>51 µM</td>
</tr>
<tr>
<td>HCO$_3^-$</td>
<td>10 mM</td>
</tr>
<tr>
<td>Total inorganic N*</td>
<td>45 µM</td>
</tr>
<tr>
<td>Total organic N**</td>
<td>17 µM</td>
</tr>
<tr>
<td>Total organic C***</td>
<td>71 µM</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trace metals</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mn</td>
<td>9 nM</td>
</tr>
<tr>
<td>Zn</td>
<td>1 nM</td>
</tr>
<tr>
<td>Co</td>
<td>0.5 nM</td>
</tr>
<tr>
<td>Mo</td>
<td>0.3 nM</td>
</tr>
<tr>
<td>Se</td>
<td>1 nM</td>
</tr>
<tr>
<td>Ni</td>
<td>1 nM</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vitamins</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1/Thiamine</td>
<td>500 nM</td>
</tr>
<tr>
<td>B2/Riboflavin</td>
<td>0.7 nM</td>
</tr>
<tr>
<td>B3/Niacin</td>
<td>800 nM</td>
</tr>
<tr>
<td>B5/Pantothenate</td>
<td>425 nM</td>
</tr>
<tr>
<td>B6/Pyridoxine</td>
<td>500 nM</td>
</tr>
<tr>
<td>B7/Biotin</td>
<td>4 nM</td>
</tr>
<tr>
<td>B9/Folic acid</td>
<td>4 nM</td>
</tr>
<tr>
<td>Vitamin B12</td>
<td>0.7 nM</td>
</tr>
<tr>
<td>Myo-inositol</td>
<td>500 nM</td>
</tr>
<tr>
<td>4-Aminobenzoic acid</td>
<td>60 nM</td>
</tr>
</tbody>
</table>

*as NO$_3^-$, NO$_2^-$, NH$_4^+$

**as urea, amino acids

*** as amino acids, carboxylic acids, sugars, fatty acids

Various marine clades in the *Gamma-, Beta-, Alphaproteobacteria,* and *Actinobacteria* (Table S2.1; Figs. S2.1-3) including five isolates from the Roseobacter clade, 23 isolates from the so-
called “Oligotrophic Marine Gammaproteobacteria” (Cho and Giovannoni, 2004), two isolates from OM43 clade, and most notably the first two reported isolates of SAR11 and SAR116 from the Gulf of Mexico (Table S1). Maximum observed cell yields for cultures grown in the JW1-4 media generally ranged between $10^5$ to $10^6$ cells mL$^{-1}$, as determined by flow cytometry, with growth rates spanning a wide range (Fig. S2.4).

Phylogenetic inspection shows that the ASM provided for similar cultivation success compared to natural seawater media, as evidenced by the numerous close relationships to cultivars obtained using natural seawater media (designated with HTCC, HIMB, and IMCC) (Figs. S2.1-3; Table S1). LSUCC0245 is only the second isolate from SAR11 subclade Va, and LSUCC0261 represents only the third isolate from SAR11 subclade IIIa (Grote et al., 2012; Vergin et al., 2013). Two clades were notable for their phylogenetic novelty in the Gammaproteobacteria—those containing LSUCC0096 and LSUCC0101. (Fig. S2.1). The clade containing LSUCC0096 was affiliated with a sole isolate, HIMB30, of the OM252 clade, and therefore LSUCC isolates represent a significant expansion of cultured representatives for this group. Similarly, LSUCC0101-type organisms matched best to an only recently recovered isolate, IMCC14953 (Yang et al., 2016), and this clade likely represents a novel gammaproteobacterial family based on depth of branching and best-blast identities of 91% to sister clade members (Fig. S2.1).

**Table 2.2. Initial cultivability statistics and salinity values for seven HTC experiments.**

<table>
<thead>
<tr>
<th>Site</th>
<th>Date</th>
<th>Inoculated Wells</th>
<th>Positive Wells</th>
<th>p</th>
<th>X</th>
<th>V*</th>
<th>% Cultivability</th>
<th>Medium</th>
<th>Medium Salinity</th>
<th>Source Salinity</th>
</tr>
</thead>
<tbody>
<tr>
<td>CJ</td>
<td>9/12/14</td>
<td>460</td>
<td>15</td>
<td>0.033</td>
<td>1.27</td>
<td>0.026</td>
<td>2.6</td>
<td>JWAMPFe</td>
<td>34.8</td>
<td>24.6</td>
</tr>
<tr>
<td>ARD</td>
<td>11/24/14</td>
<td>460</td>
<td>1</td>
<td>0.002</td>
<td>1.5</td>
<td>0.001</td>
<td>0.1</td>
<td>JW1</td>
<td>34.8</td>
<td>1.7</td>
</tr>
<tr>
<td>JLB</td>
<td>1/9/15</td>
<td>460</td>
<td>61</td>
<td>0.133</td>
<td>1.96</td>
<td>0.073</td>
<td>7.3</td>
<td>JW1</td>
<td>34.8</td>
<td>26</td>
</tr>
<tr>
<td>FWC</td>
<td>3/21/15</td>
<td>460</td>
<td>301</td>
<td>0.654</td>
<td>2</td>
<td>0.531</td>
<td>53.1</td>
<td>JW4</td>
<td>5.8</td>
<td>5.4</td>
</tr>
<tr>
<td>LB</td>
<td>6/9/15</td>
<td>460</td>
<td>0</td>
<td>0.089</td>
<td>1.56</td>
<td>0.06</td>
<td>6</td>
<td>JW4</td>
<td>5.8</td>
<td>1.4</td>
</tr>
<tr>
<td>Tbon</td>
<td>7/24/15</td>
<td>460</td>
<td>41</td>
<td>0.133</td>
<td>2</td>
<td>0.071</td>
<td>7.1</td>
<td>JW2</td>
<td>23.2</td>
<td>22.2</td>
</tr>
<tr>
<td>CJ2</td>
<td>10/1/15</td>
<td>460</td>
<td>61</td>
<td>0.133</td>
<td>2</td>
<td>0.071</td>
<td>7.1</td>
<td>JW2</td>
<td>23.2</td>
<td>22.2</td>
</tr>
</tbody>
</table>

* According to $V = -\ln(1-p)/X$, where $p = $ positive wells/inoculated wells, $X = $ # cells inoculated per well (Button et al., 1993). % Cultivability = $V \times 100$.  
** Sample was counted using microscopy and believed to have underestimated the total number of cells resulting in a larger % cultivability.

Comparison of isolate sequences with operational taxonomic units (OTUs) from whole community sequencing of the source water demonstrated that our method frequently captured some of the most abundant organisms in the system (Fig. 2.2 and 2.3; Table S1). For the most successful experiment, designated CJ2, 11 of the 30 isolates cultivated represented seven of the top 20 OTUs (Fig. 2.3). Specifically, LSUCC0245/ SAR11 subgroup Va was the 6th ranked OTU, while LSUCC0247/OM60/NOR5 clade was 7th, LSUCC0225 and 0226/SAR116 were 8th, and LSUCC0227 and 0268/OM43 were 10th (Fig. 2.3). Similar results were obtained in the other experiments, although success was varied (Fig. 2.2; Table S1). Often, if members of the dominant taxa were not recovered in a particular experiment, a representative of that OTU was recovered in a separate experiment. For example, the SAR11 subclade Va isolate LSUCC0245, represented by OTU10, was only recovered in the CJ2 experiment (Fig. 2.3), but it nevertheless
represented the number 3rd, 6th, 8th, and 8th ranks in CJ, JLB, LKB, and Tbon source waters, respectively (Fig. 2.2A,C,E,F). In an even more extreme example, cultivation was totally unsuccessful at LB and ARD, yet isolates from other cultivation attempts represented five and three of the top 25 OTUs, respectively, in the rank abundance curves from those experiments (Fig. 2.2B and E, respectively). While the SAR11 subclade Va organism and some other isolates were cultivated only once, several other closely related taxa were obtained more frequently. For example, 10 isolates of the OM252 clade, represented by OTU55, were cultured from 3 different experiments. This OTU showed varying relative abundance (ranks 20-540), but regular occurrence, being observed in four of the seven experiments within the top 35 ranks (Table S1). Continued cultivation efforts will provide opportunities for biogeographic analyses of individual isolates as well as efficacy and reproducibility of the cultivability of organisms from abundant clades.

Why some experiments were more successful than others can only be speculated upon at this time. There are numerous potential reasons why some organisms have not been isolated at all, or why some organisms can be cultivated in some circumstances, but not others, even though they are present in the starting inocula. It may have to do with a dependence on an unsupplied compound or combination of compounds, the specific concentration thresholds for compounds, or the requirement for a better surface than PTFE. There is also evidence that many taxa are dormant and can be cultivated after stochastic release from dormancy (Buerger et al., 2012b). Additionally, our threshold of 10^4 cells mL^−1 for calling a culture positive may prevent obtaining certain taxa that are always at low abundance, or those that have not achieved this threshold in the allotted time.

Beginning in March of 2015 we attempted to match salinity with measured values using the JW2-4 media by diluting the Basic salts and Mg/Ca components (excluding P) (Table 2.3). Our motivation for making these adaptations was the evidence for salinity as one of, if not the, most important determinants of microbial community composition in general (Lozupone and Knight, 2007; Logares et al., 2009) and in estuarine environments in particular (Crump et al., 2004; Fortunato et al., 2012; Dupont et al., 2014). The six sites ranged in salinity from 1.4 to 26 ppt (Table 2.2). Values at CJ, CJ2, and JLB were most similar to marine environments (> 22 ppt), while Tbon was brackish (14.6 ppt), and ARD, FWC, and LB nearly qualified as freshwater (< 6 ppt) (Table 2.2). We matched salinity as closely as possible for the FWC, LB, Tbon, and CJ2 experiments (Table 2.2), but the impact of these efforts was inconclusive. Although there may be a trend emerging across all our experiments whereby cultivability is correlated with the similarity between in situ and medium salinities, as of now there are too few data points to ascertain a significant pattern.

We attribute the success of the ASM to keeping constituent concentrations close to environmentally relevant values, and providing a complex suite of carbon and nitrogen sources, while simultaneously keeping total organic C low. However, there are still many taxa that remained uncultivated. Notably, the most dominant members of these communities were Actinobacteria, belonging either to the OM1 clade (i.e., “Marine Actinobacteria” (Rappé et al., 1999) in marine samples (OTU1- Fig. 2.2A,C,F and Fig. 2.3), or the hgcl clade in more freshwater samples (OTU2- Fig. 2.2B,D,E). Furthermore, it must be remembered that due to dark incubations, this method excluded cultivation of phototrophs, yet some like Synechococcus were among the most abundant organisms in some samples (e.g., OTU4- Fig. 2.2A). Light incubations could prove fruitful for isolation of phototrophs.
We did not attempt to precisely match our media concentrations of inorganic nutrients with those at sampled sites during these experiments, but this could also be a key place to improve cultivability. At 5, 2, and 38 µM the concentrations of ammonium, nitrite, and nitrate, respectively, in our ASM are near the measured values from the sites where we sampled for the cultivation experiments, however these values fluctuated (Table S1). Nitrate at site CJ was measured at 0.6 µM while ARD and Tbon were measured at 85 µM (Table S1). Our ASM phosphate concentration (51 µM) was based on previous success with the cultivation of SAR11 (Carini et al., 2012), however this was considerably higher than in situ levels (1-3 µM) (Table 2.2 and S1). Measurements of other constituent concentrations (e.g. DOM components) could help determine what drives the success in cultivating more of the microbial majority. Additionally, we expect physiological and genomic characterization of our isolates to shed additional light on why some taxa may have been preferentially isolated relative to others. Based on our observations, future experiments will include additional alterations to the medium to try and obtain additional key members of the microbial consortia.

Figure 2.2. Position of LSUCC isolates relative to matching OTUs within the top 60 ranks for the first six experiments. OTUs are ordered by decreasing relative abundance, according to the average of duplicate samples. Experiments are in order according to Table 2: A) CJ, B) ARD, C) JLB, D) FWC, E) LB, and F) Tbon. LSUCC isolates with best blast hits to the representative sequence for a given OTU are indicated in bold. Stars indicate OTUs with cultured representatives from other LSUCC experiments. Isolates with matching OTUs with ranks lower than 60 are detailed in Table S1.

Nevertheless, while there is always room for improvement, this is the first reported isolation of SAR11, SAR116, OM43, HIMB11-type members of the Roseobacter clade, and many other taxa in ASM, which validates the JW media design and provides a suite of isolates from the GoM that should prove valuable for linking genomics and physiology with biogeography. The modularity of our media also facilitates easy customization for targeted cultivation approaches informed by
genomics, metagenomics, or marine chemistry. Furthermore, the success of the ASM recipes in the HTC context fulfills the goal of providing a portable, reproducible cultivation strategy that can be utilized by researchers constrained in their capacity to obtain seawater, and thereby provides another valuable tool for cultivation of important marine clades.

All cultures are currently archived and available upon request as part of the Louisiana State University Culture Collection (LSUCC).

Figure 2.3. Position of LSUCC isolates relative to matching OTUs within the top 60 ranks for the most successful experiment, CJ2, which is enlarged to allow for OTU taxonomy labels. OTUs are ordered by decreasing relative abundance, according to the average of duplicate samples. LSUCC isolates with best blast hits to the representative sequence for a given OTU are indicated in bold. Stars indicate OTUs with cultured representatives from other LSUCC experiments. Isolates with matching OTUs with ranks lower than 60 are detailed in Table S1.

2.3 METHODS AND MATERIALS

2.3.1 SAMPLING

One liter of surface water was collected at each site (Table S1) in a sterile, acid washed polycarbonate bottle (Nalgene). Duplicate 120 mL volumes were filtered serially through 2.7 µm
Whatman GF/D and 0.22 µm Sterivex (Millipore) filters, which were immediately stored on ice and transferred to -20°C upon return to the lab until DNA extractions were performed. For biogeochemical measurements, duplicate 50 mL volumes of the 0.22 µm filtrate was collected and placed on ice for measurements of SiO$_4$, PO$_4$$_3$-, NH$_4$+, NO$_3$-, and NO$_2$- at the University of Washington Marine Chemistry Laboratory (http://www.ocean.washington.edu/story/Marine+Chemistry+Laboratory). 10 mL of whole water for cell counts was fixed with 10% formalin (final volume) and placed on ice immediately. Temperature, pH, DO, and salinity were measured using a handheld YSI 556 MPS.

**Table 2.3.** Comparisons of the composition of natural seawater, marine media 2216, and JW artificial seawater. Columns for seawater and Marine Broth 2216 are taken from Giovannoni & Stingl 2007.

<table>
<thead>
<tr>
<th>Element</th>
<th>Seawater (g/L)</th>
<th>Marine Broth 2216 (g/L)</th>
<th>JW1 (g/L)</th>
<th>JW2 (g/L)</th>
<th>JW3 (g/L)</th>
<th>JW4 (g/L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chlorine</td>
<td>19</td>
<td>19.8</td>
<td>19.2</td>
<td>12.8</td>
<td>6.4</td>
<td>3.2</td>
</tr>
<tr>
<td>Sodium</td>
<td>10.5</td>
<td>8.8</td>
<td>11</td>
<td>7.4</td>
<td>3.8</td>
<td>2</td>
</tr>
<tr>
<td>Magnesium</td>
<td>1.35</td>
<td>2.2</td>
<td>1.27</td>
<td>0.85</td>
<td>0.42</td>
<td>0.211</td>
</tr>
<tr>
<td>Sulphur</td>
<td>885</td>
<td>0.72</td>
<td>0.96</td>
<td>0.64</td>
<td>0.32</td>
<td>0.16</td>
</tr>
<tr>
<td>Calcium</td>
<td>0.4</td>
<td>0.65</td>
<td>0.41</td>
<td>0.28</td>
<td>0.14</td>
<td>0.07</td>
</tr>
<tr>
<td>Potassium</td>
<td>0.38</td>
<td>0.32</td>
<td>0.39</td>
<td>0.26</td>
<td>0.13</td>
<td>0.07</td>
</tr>
<tr>
<td>Bromine</td>
<td>0.65</td>
<td>0.054</td>
<td>0.063</td>
<td>0.042</td>
<td>0.022</td>
<td>0.01</td>
</tr>
<tr>
<td>Carbon</td>
<td>0.028</td>
<td>4.78</td>
<td>0.0008*</td>
<td>0.0008*</td>
<td>0.0008*</td>
<td>0.0008*</td>
</tr>
<tr>
<td>Boron</td>
<td>0.0046</td>
<td>0.0047</td>
<td>0.0046</td>
<td>0.003</td>
<td>0.0015</td>
<td>0.00075</td>
</tr>
<tr>
<td>Silicon</td>
<td>0.003</td>
<td>0.00085</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Fluorine</td>
<td>0.0013</td>
<td>0.00109</td>
<td>0.001</td>
<td>0.0009</td>
<td>0.00046</td>
<td>0.00023</td>
</tr>
<tr>
<td>Nitrogen</td>
<td>0.0005</td>
<td>0.72</td>
<td>0.00087**</td>
<td>0.00087**</td>
<td>0.00087**</td>
<td>0.00087**</td>
</tr>
<tr>
<td>Phosphorous</td>
<td>0.0007</td>
<td>0.045</td>
<td>0.0016</td>
<td>0.0016</td>
<td>0.0016</td>
<td>0.0008</td>
</tr>
<tr>
<td>Iron</td>
<td>0.00001</td>
<td>0.0226</td>
<td>0.000006</td>
<td>0.000006</td>
<td>0.000006</td>
<td>0.000006</td>
</tr>
</tbody>
</table>

*C as mixed amino acids, fatty acids, carboxylic acids, and sugars (Rabalais et al., 1996; Dagg et al., 2004)  
**N as NO3-, NO2-, NH4+, urea, and amino acids (Rabalais et al., 1996; Dagg et al., 2004). Differences in JW media for some element values compared with “Seawater” are related to the geochemical nature of the GoM compared to the Atlantic subtropical gyre.
2.3.2 HTC EXPERIMENTS

The remaining surface water was stored on ice and immediately returned to the lab for use in HTC experiments as described by (Thrash et al., 2015), 2015 (Thrash et al., 2015). Briefly, a subsample was filtered through a 2.7 \( \mu \text{m} \) GF/D filter, stained with 1x Sybr Green (Lonza), and enumerated using the Guava EasyCyte (Millipore) flow cytometer. For each experiment, a total of 480 dilution-to-extinction cultures were established using five 96 x 2.1 mL well PTFE plates (Radleys, Essex, UK) containing artificial media (Table S1). Each well was inoculated with an estimated 1-3 cells from the 2.7 \( \mu \text{m} \)-filtered water, plates were stored in the dark at in situ temperatures, and monitored for growth at 2-4 weeks, depending on the temperature, and again at 6 weeks. Growth was monitored using the Guava EasyCyte (Millipore) flow cytometer and isolates reaching at least 10\(^4\) cells mL\(^{-1}\) were transferred to 50 mL polycarbonate flasks, and cryopreserved in liquid N2 using 10% glycerol or 5% DMSO (final volume). Normally, all isolates reaching this minimum growth were transferred, with one exception, experiment FWC. In this experiment, cells were counted using microscopy (Christner et al., 2001), and we believe total cell concentration in our inoculum was underestimated. Resultant cultivability was over 53\%, and we therefore sub-selected only 60 wells for transfer. All transferred isolates, upon reaching at least 10\(^5\) cells mL\(^{-1}\), were collected on 25 mm, 0.22 \( \mu \text{m} \) polycarbonate filters (Millipore), and DNA extractions were performed using the MoBio PowerWater DNA kit following the manufacturer’s instructions. Additional cyrostocks were also made at this stage to increase the number preserved in the LSU culture collection (LSUCC). Isolate 16S rRNA genes were amplified using recombinant Taq (Invitrogen), 27F/1492R primers (S-D-Bact-0008-d-S-20/\( {\text{S}}^{-}\text{-Univ-1492-a-A-21} \)) (Klindworth et al., 2013) and the following PCR conditions: denaturation at 94°C for 30 s, annealing at 50.8°C for 30 s, elongation at 72°C for two min, and a final elongation step at 72°C for 10 min; repeated 35 times. Prior to Sanger sequencing, successful PCR products were cleaned using the QiaQuick PCR purification kit (Qiagen) to remove any inhibitors and PCR materials. Sanger sequencing of amplicons, using both the forward and reverse primer, was completed using the Finch software (Geospiza Finch Suite Distribution v 2.21.0) as provided by the Michigan State RTSF Genomic Core. Forward and reverse sequences were assembled, when sufficient overlap permitted, via the CAP3 webserver (http://doua.prabi.fr/software/cap3), after conversion of the reverse read to its reverse complement at http://www.bioinformatics.org/sms/rev_comp.html. Purity of isolates was based on evaluating the consistency of the quality scores across the length of the read. All chromatograms are available on the Thrash Lab website (http://thethrashlab.com/publications) with the reference to this manuscript linked to “Supplementary Information”.

2.3.3 COMMUNITY iTAG ANALYSIS

Community DNA was isolated from the sterivex filters by removing the filters from their housing under sterile conditions in a biosafety cabinet. Filters were then extracted with the PowerWater kit as well. DNA was sequenced at the 16S rRNA gene V4 region (515F 806R) with Illumina MiSeq 2x250bp paired-end sequencing at Argonne National Laboratories (Caporaso et al., 2012). Raw 16S rRNA gene amplicon data was analyzed with Mothur v.1.33.3 (Schloss et al., 2009) using the Silva v119 database (Pruesse et al., 2007). Briefly, 16S rRNA sequences were assembled into contigs and discarded if the contig had any ambiguous base pairs,
possessed repeats greater than 8 bp, or were greater than 253 bp in length. Contigs were aligned to the Silva rRNA v.119 database, checked for chimeras using UCHIME (Edgar et al., 2011), and classified using the Silva rRNA v.119 database. Contigs classifying to chloroplast, eukaryotes, mitochondria, or “unknown” affinities were removed from the data and the remaining contigs were clustered into distinctive operational taxonomic units (OTUs) using a 0.03 dissimilarity threshold (OTU<sub>0.03</sub>). Rank-abundance curves were arranged by plotting the mean relative abundance of two biological replicates collected and sequenced separately, with only the top 60 OTUs included for clarity. Complete OTU tables are available upon request. Plots were created in R (v. 3.2.1) (Ihaka and Gentleman, 1996) with the graphing program GGPLOT2 (v. 2.0.0) (Wickham, 2011). To determine which OTUs represented LSUCC isolates, sequences from the OTU representative fasta file, provided by mothur using get.oturep(), was used to create a database file, against which the LSUCC isolate 16S rRNA genes could be blasted. The following are the commands used (BLAST v 2.2.26):

```bash
formatdb -i database.faa -o T -p F
blastall -i queryfile.faa -d database.faa -p blastn -b 2 -v 2 -o outputfile.txt
```

All best hits ranked at 98% identity or above, with the exception of 3 sequences, which had 96-97% matches to very low ranking OTUs (LSUCC0175, 0115, and 0141). The blast results (BOR_bestblasthit), OTU rep file (BOR_16S.rep.fasta), OTU taxonomic assignments (BOR_16S.taxonomy) and OTU table (BOR_OTU.csv) are available at the Thrash Lab website (http://thethrashlab.com/publications) with the reference to this manuscript linked to “Supplementary Information”.

### 2.3.4 PHYLOGENETIC TREES

Full length or, if a contig was unavailable, forward 16S rRNA gene sequences from isolates were compiled with best hits from BLAST to the NR database and sequences of known representatives of the various marine clades to which isolates were matching (All fasta files are provided on the Thrash Lab website: http://thethrashlab.com/publications). Sequences were aligned with MUSCLE (Edgar, 2004), culled with Gblocks (Castresana, 2000), and phylogeny was inferred using FastTree2 (Price et al., 2010). These processes were linked with a custom shell script (FT_holder), also available at the Thrash Lab website (http://thethrashlab.com/publications) with the reference to this manuscript linked to “Supplementary Information”. Visualization was performed with Archaeopteryx (Han and Zmasek, 2009).

### 2.3.5 ACCESSION NUMBERS

Community 16S rRNA gene sequence fastq files are available at the NCBI Sequence Read Archive under the accession numbers: SRR3085688-3085701. Individual isolate sequences have been submitted to NCBI under the accession numbers: KU382357-KU382438.
CHAPTER 3.
CULTIVATION AND GENOMICS OF THE FIRST FRESHWATER SAR11 (LD12) ISOLATE

3.1 INTRODUCTION
Bacterioplankton in the SAR11 clade of *Alphaproteobacteria* are dominant heterotrophs in marine and freshwater systems. In the oceans, SAR11 can represent 25-50% of total planktonic cells (Morris et al., 2002; Schattenhofer et al., 2009). Numerous subclades with unique spatio-temporal distributions comprise SAR11 (Morris et al., 2002; Carlson et al., 2009; Giovannoni and Vergin, 2012; Thrash et al., 2014). At least nine subclades defined via 16S rRNA gene sequences occupy marine niches (Giovannoni and Vergin, 2012). However, in spite of its global distribution (Morris et al., 2002; Carlson et al., 2009; Giovannoni and Vergin, 2012; Thrash et al., 2014), massive predicted population size of $10^{28}$ cells (Morris et al., 2002), and an estimated divergence time from its last common ancestor of 1.1 billion years ago (Luo et al., 2013), the bulk of existing evidence suggests that SAR11 has only successfully colonized freshwater environments once in its natural history (Logares et al., 2010; Salcher et al., 2011; Eiler et al., 2014). Traditionally, all known freshwater SAR11 belong to subclade IIIb, a.k.a. LD12. However, a recent report challenges this assertion: a genome sister to subclade I was recovered in Lake Baikal metagenomic data (Cabello-Yeves et al., 2018). Regardless, the limited evolutionary diversification into less saline habitats has not prevented LD12 from achieving prominence in the ecosystems it inhabits. In many lotic and lentic environments, LD12 occupies similar relative abundances as its marine cousins (Newton et al., 2011; Salcher et al., 2011; Dupont et al., 2014; Henson et al., 2016; Garcia et al., 2018). Study of LD12 is important to understand SAR11 evolution, specifically, and how successful transitions between marine and freshwater environments occur in bacterioplankton (Logares et al., 2009), more generally.

Ecological, functional, and sequence-based inference from single amplified genomes (SAGs) and metagenomes support the hypothesis that LD12 bacterioplankton evolved from a genome-streamlined marine ancestor (Salcher et al., 2011; Luo et al., 2013; Zaremba-Niedzwiedzka et al., 2013; Eiler et al., 2015). As such, they share many of the same characteristics as marine SAR11, such as small cell volumes; adaptation to oligotrophic habitats; small, streamlined genomes; an obligate aerobic chemoorganoheterotrophic lifestyle with limited metabolic flexibility; preference for small molecular weight compounds like carboxylic and amino acids as carbon/energy sources; and auxotrophies for some amino acids and vitamins (Giovannoni and Stingl, 2005; Giovannoni, Tripp, et al., 2005; Dupont et al., 2014; Eiler et al., 2014, 2015). Previous research suggests that LD12 differ from their marine counterparts in specific elements of metabolic potential that indicate a greater emphasis on production, rather than uptake, of osmolytes, and important metabolic changes related to energy production (Dupont et al., 2014; Eiler et al., 2015). For example, metagenomic population data showed a correlation between decreased salinity and greater proportion of the Embden-Meyerhof-Parnass (EMP) vs. Entner-Doudoroff (ED) glycolysis pathways (Dupont et al., 2014). Comparative genomic analyses of SAGs from different SAR11 strains concurred: LD12 genomes contained the EMP pathway that is not found in most marine SAR11 (Grote et al., 2012; Eiler et al., 2015).

---

2 This chapter previously appeared as Henson, Michael W., V. Celeste Lanclos, Brant C. Faircloth, and J. Cameron Thrash. 2018. “Cultivation and Genomics of the First Freshwater SAR11 (LD12) Isolate.” The ISME Journal 12 (7): 1846–60. Reprinted with the permission of my co-authors.
SAG data also suggested that LD12 lacks the glyoxylate shunt and some single carbon (C1) metabolism (Eiler et al., 2015).

Despite what has been learned from cultivation-independent methods, the lack of cultured LD12 representatives has hampered a more detailed exploration of the group. Potential ecotypes within LD12 have been identified (Zaremba-Niedzwiedzka et al., 2013), and their population dynamics recently described with five-year time series data in freshwater lakes (Garcia et al., 2018). However, we cannot delineate what distinguishes ecotypes without better physiological and genomic data. Similarly, interpreting data on the ecological distribution of LD12 remains challenging without information on growth tolerances and optima for salinity and temperature. We also don’t understand whether a connection exists between more efficient energy production through EMP-based glycolysis and the freshwater lifestyle, or what other adaptations might explain LD12 evolution away from salt water.

The next steps in translating ‘omics-based predictions into measured data for integration with ecosystem models require living experimental subjects. For example, cultures of marine SAR11, such as HTCC1062 and HTCC7211, have facilitated testing of metabolism and growth (Giovannoni, Bibbs, et al., 2005; Tripp et al., 2008; Sun et al., 2011; Carini et al., 2012; P. Carini et al., 2014), structural organization (Zhao et al., 2017), and virus-host interactions (Zhao et al., 2013). We need cultivated representatives to provide this kind of understanding of other important bacterioplankton like LD12. In service of this goal, we pursued a systematic high throughput cultivation effort from coastal regions in the northern Gulf of Mexico. Here, we report the first isolation of a member of the LD12 clade, strain LSUCC0530, its complete genome sequence, preliminary physiological data, and an examination of the ecological distribution of this organism compared with other LD12 clade members. Our results provide evidence for temperature-dependent ecotype diversification within LD12 and a hypothesis about the evolutionary trajectory that led to LD12 colonization of fresh water.

3.2 RESULTS

3.2.1 ISOLATION AND GROWTH OF LSUCC0530

We isolated strain LSUCC0530 as part of a series of high throughput dilution-to-extinction cultivation experiments (HTC) conducted with inoculation water obtained across the southern Louisiana coast. The particular experiment that yielded LSUCC0530 utilized surface water from the coastal lagoon of Lake Borgne, 39 km southeast of New Orleans. At the time of collection, the surface water had a salinity of 2.39 and was 30.5°C. Subsequent community analysis demonstrated that LD12-type organisms represented 8.7% of the bacterioplankton population in the inoculum (see below). We diluted 2.7 μm-filtered Lake Borgne surface water for inoculation to an estimated 2 cells well−1 into our JW5 medium (salinity 1.45, Table S1), and incubated cultures for 30 days at 23°C (room temperature in the Thrash Lab). LSUCC0530 grew slowly, reaching 8.6 x 10⁶ cells mL⁻¹ after 24 days. It had a low fluorescence/low side scatter flow cytometric signature characteristic of other SAR11 cells (Fig. S3.1).

3.2.2 TAXONOMY AND MORPHOLOGY

The LSUCC0530 16S rRNA gene sequence had > 99% identity with eight SAG sequences from members of LD12 (Zaremba-Niedzwiedzka et al., 2013) and only 91% and 92% sequence
identity with sequences from the sister subclade IIIa representatives IMCC9063 and HIMB114, respectively. Of note for taxonomic incorporation of genomic data (Chun et al., 2018), the genome-generated 16S rRNA gene sequence matched the PCR-generated 16S rRNA gene sequence (GenBank accession number KY290650.1) at 1316 of 1320 sites. Two of these mismatches were single bp gaps the genome sequence, and three of the four mismatches occurred in the first 20 bp of the PCR-generated sequence. LSUCC0530 also had average amino acid identities (AAI) of 58.1% and 59.7% with HIMB114 and IMCC9063, respectively (Table S1). In contrast, AAI between LSUCC0530 and the LD12 SAGs ranged from 81.9% to 88.0%, corroborating that LSUCC0530 belongs to separate species from subclade IIIa (Konstantinidis and Tiedje, 2005). Phylogenetic inference using 16S rRNA genes placed LSUCC0530 within the Order Pelagibacteriales (a.k.a. SAR11; (Thrash et al., 2014), in monophyly with the original

---

**Figure 3.1. Phylogeny and morphology of LSUCC0530.** A) Phylogenetic position of LSUCC0530 within the SAR11 clade using a concatenation of 83 single copy marker genes and 23,356 amino acid positions. The tree was inferred using RAxML with 100 bootstrap runs. Values at the nodes indicate 100% bootstrap support unless otherwise noted. The tree was rooted on HIMB59. Subclades are indicated on the right, and LD12 microclusters are indicated A-C. Asterisk- we have included AAA028-C07 in microcluster A, even though it was excluded previously (Zaremba-Niedzwiedzka et al., 2013). B) Scanning electron micrograph of LSUCC0530 cells at 35,000x magnification. Scale bar represents 0.5 µm. C) Transmission electron micrograph of a LSUCC0530 cell thin section. Scale bar represents 0.5 µm.
LD12 clone library sequence (Fig. S3.2). Phylogenomic inference using 83 single copy marker genes from 49 SAR11 genomes also strongly supported placement of LSUCC0530 within LD12 as an early diverging member, with only the SAG AAA280-B11 from Lake Damariscotta diverging earlier (Fig. 3.1A). This tree also recovered the three microclusters within LD12 (A, B, C) observed previously (Zaremba-Niedzwiedzka et al., 2013) (Fig. 3.1A). Cells of strain LSUCC0530 were small, curved rods, approximately 1 µm x 0.1 µm (Fig. 3.1B, C), showing conserved morphology with the distantly related SAR11 strain HTCC1062 in subclade Ia (Rappé et al., 2002; Zhao et al., 2017).

3.2.3 GENOME CHARACTERISTICS

Genome assembly resulted in a single, circularized scaffold that we tested for completeness using multiple tools (Supplemental Information). The completed genome is 1,160,202 bp with a GC content of 29.02% and 1271 predicted genes (Fig. 3.2). There are 1231 putative protein coding genes and 40 RNA genes, including one each of the 5S, 16S, and 23S rRNA genes. We found no genomic evidence of lysogenic bacteriophage or a CRISPR-cas system. LSUCC0530 has the characteristics of genome streamlining previously reported for other SAR11 genomes, with an estimated coding percentage of 96.36% and the smallest complete SAR11 genome reported thus far (Giovannoni, Tripp, et al., 2005; Grote et al., 2012). Despite its smaller overall genome size, intergenic spacer regions are not significantly smaller than those of the sister subclade IIIa (Fig. S3.3). All LD12 genomes had much higher AAI and synteny with each other than with the subclade IIIa (Table S1), mirroring their phylogenomic relationships (Fig. 3.1A). However, within the LD12 clade, AAI and synteny percentages did not clearly delineate the same patterns we observed in the phylogenomic tree, which may have resulted from the LD12 SAGs being fragmented and incomplete.
Figure 3.2. Circular diagram of the LSUCC0530 genome. Rings indicate, from outside to the center: Position relative to the replication start site; Forward strand genes, colored by COG category; Reverse strand genes, colored by COG category; RNA genes (tRNAs green, rRNAs red, other RNAs black); GC content; GC skew.

The LSUCC0530 genome has the conserved architecture for hypervariable region 2 (HVR2) that has been observed in most other complete SAR11 genomes (Rodriguez-Valera et al., 2009; Grote et al., 2012) and other LD12 SAGs (Zaremba-Niedzwiedzka et al., 2013), between the 16S rRNA-tRNA-tRNA-23S rRNA operon and the 5S rRNA gene operon. This region is 54,755 bp and has a strong variation in GC content from neighboring parts of the genome (Fig. 3.2). Metagenomic recruitment also verified this region as hypervariable based on the lack of matching sequences in community data (Fig. S3.4). The predicted gene annotations in
HVR2 resemble those in the same location in other SAR11 genomes (Grote et al., 2012; Zaremba-Niedzwiedzka et al., 2013), namely genes likely functioning in membrane modification such as glycosyltransferases, epimerases, transaminases, and methyltransferases; as well as hypothetical proteins.

3.2.4 METABOLIC RECONSTRUCTION

![Metabolic reconstruction of LSUCC0530.](image)

**Figure 3.3** Metabolic reconstruction of LSUCC0530. Solid boxes indicate recovered genes, dashed boxes indicate missing genes. Color shading distinguishes major metabolic sub-systems, which are also labeled in bold text. ABC transporters are depicted with circles indicating subunits. Symporters and antiporters are depicted with ovals. Two component systems are depicted with large and small rectangles. The C4 TRAP transporter is depicted separately. Major subsystems are colored for easier identification. Multiple arrows indicate all genes present in a given pathway. Numbers identify genes according to the key in Table S1. Question marks indicate a single missing gene in an otherwise complete pathway (e.g., PRPP → His). Light blue fill indicates genes with no SAR11 orthologs outside of the LD12 subclade.

Like most other SAR11s (Grote et al., 2012; Thrash et al., 2014; Eiler et al., 2015), the LSUCC0530 genome encodes for an obligate aerobic lifestyle, with a complete oxidative phosphorylation pathway including a heme/copper-type cytochrome c oxidase, a proton-translocating NADH dehydrogenase, and a proton-translocating ATP synthase (Fig. 3.3). As previously detected in LD12 SAGs (Eiler et al., 2015), the LSUCC0530 genome encodes a
complete EMP glycolysis pathway, including phosphofructokinase and pyruvate dehydrogenase, and complete gluconeogenesis through nucleotide sugar production (Fig. 3). We recovered genes for the pentose phosphate pathway and a predicted fructokinase gene for conversion of D-fructose to β-D-fructose-6-P. As expected based on other SAR11 genomes, we did not find annotated genes for utilization of allose, fucose, fuculose, galactose, maltose, mannose, rhamnose, sucrose, starch, trehalose, or xylose. LSUCC0530 has a complete TCA cycle and glycolate oxidase, and like other SAR11s, the non-mevalonate pathway for isopentenyl diphosphate (Fig. 3).

We recovered the first isocitrate lyase in an LD12 genome, conferring a complete glyoxylate bypass in combination with malate synthase. The isocitrate lyase shares its ancestry with the majority of the SAR11 clade except subclade IIIa, where it appears to have been lost (Grote et al., 2012; Eiler et al., 2015) (Fig. S3.5). However, the LSUCC0530 malate synthase shared with other LD12 SAGs belongs to isoform A, whereas all other SAR11 clades contain isoform G (Anstrom et al., 2003) (Figs. 3.3; S3.6). Comparison of gene neighborhoods in a variety of SAR11 genomes revealed that both isoforms occur in the same location (Fig. S3.7).

We predict production of eighteen amino acids (Fig. 3.3) and partial synthesis pathways for three more (histidine, cysteine, and methionine), as well as the ability to interconvert, but not synthesize de novo, phenylalanine and tyrosine. The LSUCC0530 genome has lost some of the C1 and methylated compound pathways found in other SAR11s (Sun et al., 2011). For example, as in the LD12 SAGs (Eiler et al., 2015), LSUCC0530 does not have DMSP, methylamine, or glycine-betaine metabolism. However, it retains tetrahydrofolate metabolism, formate oxidation, and the glycine cleavage pathway (Fig. 3.3).

The genome has the same PII nitrogen response system and amtB ammonium transporter found in other SAR11s and is predicted to utilize ammonia as a nitrogen source in amino acid synthesis (Fig. 3.3). The genome also encodes a putative thiosulfate/3-mercaptopyruvate sulfurtransferase that may be involved in sulfur and thiocyanate reactions (Fig. 3.3). A possible product of this enzyme is sulfite, which could be further oxidized to sulfate in the periplasm via putative sorAB genes. This gene pair occurs infrequently in the SAR11 clade, but sulfite oxidation may serve, in some strains, as a means to generate additional proton motive force (Denger et al., 2008; Kappler, 2011). This is similar to the function conferred by proteorhodopsin (Béjà et al., 2001; Giovannoni, Bibbs, et al., 2005; Steindler et al., 2011), which the LSUCC0530 genome also contains. Also like other SAR11s, a complete assimilatory sulfate reduction pathway is missing, making LSUCC0530 dependent on reduced sulfur compounds for cellular sulfur requirements (Tripp et al., 2008). In our media, this was supplied in the form of sulfur-containing amino acids (Table S1). The genome encodes partial pathways for incorporation of sulfide into cysteine and methionine but is missing predicted serine O-acetyltransferase and homoserine O-succinyltransferase genes (Fig. 3.3).

In addition to the PII nitrogen sensor, LSUCC0530 also has putative two component systems handling phosphate limitation, pH, osmotic upshift via potassium, and redox state (Fig. 3.3). The LSUCC genome encodes ABC transporters for phosphate (ptsABCS), lipoprotein, and branched-chain and L amino acid uptake. We identified tripartite ATP-independent periplasmic (TRAP) transporter genes for C4-dicarboxylate and mannitol/chloroaromatic compounds. However, the gene cluster appears to be missing the substrate binding subunit, and we did not find predicted genes for mannitol utilization.

In general, LSUCC0530 and the LD12 SAGs share many of the same pathways for osmolyte synthesis as the subclade IIIa and other SAR11 organisms, such as N-acetyl-ornithine,
glutamate, glutamine, and proline (Table 3.1). In some cases, there appear to be analogous substitutions in osmolyte transport systems. For example, LD12 genomes have potassium antiporters, but not the trkAH genes for potassium transport in other SAR11s. Similarly, a sodium/proton antiporter orthologus cluster comprised genes exclusively from LD12 genomes, but a separate sodium/proton antiporter occurred in all other SAR11 genomes (Table 3.1). Perhaps the most important difference in genomic content concerning osmolytes occurs in two key transporter losses: LSUCC0530 and the other LD12s have no genes for the glycine betaine/proline or ectoine/hydroxyectoine ABC transporters that are present in HIMB114 and IMCC9063 and other subclades of SAR11 (Table 3.1).

3.2.5 PHYSIOLOGY

![Figure 3.4](image)

**Figure 3.4.** Growth rates for LSUCC0530 according to 
A) salinity and 
B) temperature. The boxes indicate the interquartile range (IQR) of the data, with vertical lines indicating the upper and lower extremes according to 1.5 x IQR. Horizontal lines within each box indicate the median. The underlying data points of individual biological replicates, calculated from corresponding growth curves in Figure S8, are plotted on top of each box.

Under isolation conditions (23°C, JW5 medium), strain LSUCC0530 grew to a density of $2 \times 10^7$ cells mL$^{-1}$ with an average growth rate of 0.52 day$^{-1}$. Our physiological experiments to determine the salinity tolerance and growth optima revealed that LSUCC0530 could grow at salinities between 0.36 and 4.7 (Fig. 3.4A). The optimal salinity for LSUCC0530 was at or below the isolation medium of 1.45. Growth rates decreased between salinities of 2.9 and 4.7 (Fig. 3.4A), and cells died (determined by rapid loss of flow cytometric signal) in salinities above 8 (Fig. S3.8). We observed growth after a very extended lag phase in a subset of cultures at salinity 5.8, but could not always reproduce this growth in repeated experiments (Fig. S3.8). Among the
temperatures tested, LSUCC0530 grew optimally at 23°C, and we observed growth up to 30°C, but not at 12°C or below, nor at 35°C or above (Fig. 3.4B). However, we did not observe a clear loss of signal at 4, 12, 35, or even 40°C (Fig. S3.8), even after over 33 days, suggesting that LSUCC0530 can endure extended periods at these temperatures in a non-growth state. Notably during growth experiments, one of the five replicates during any given experiment frequently grew at a reduced rate than the others (Fig. S3.8).

3.2.6 ECOLOGY

Figure 3.5. Results of competitive metagenomic recruitment to all genomes used in the phylogenomic tree. A) 95% identity RPKM values for LD12 genomes only, by site, with data aggregated for all genomes in microclusters A-C, according to the key. Colors indicate broad environmental categories. B-D) Density plots of metagenomic sequence recruitment according to percent identity of the best hit. Data was aggregated according to the key, with LD12 separated as in A, HIMB59 separately depicted as “V”, and all other SAR11 genomes together as “Other.” B) Two sites where LSUCC0530 dominated recruitment with high percent identity. C) Two sites where the LSUCC0530 only recruited sequences at ~90% identity and no other group had high recruitment at higher identity. D) Two sites where other LD12 microclusters showed similar patterns as LSUCC0530 in B and C.

We quantified the abundance of LD12 taxa in the coastal Louisiana ecosystems using 16S rRNA gene sequences clustered into operational taxonomic units (OTUs). These sequences were obtained as part of the combined sampling and HTC campaign that included the successful isolation of LSUCC0530. Based on BLASTn sequence similarity, the representative sequence for OTU7 and LSUCC0530 shared 100% identity over the 250bp v4 region of the 16S rRNA gene, confirming the Mothur classification of OTU7 as belonging to the SAR11 LD12 clade. Over the three-year sampling period, OTU7 was the fifth most abundant OTU based on total read abundance and sixth ranked OTU based on median read abundance. At the time of sampling for the experiment that resulted in the isolation of LSUCC0530, OTU7 had a relative abundance of 8.7% in Lake Borgne. In some sites, such as the Atchafalaya River Delta (salinity 0.18), LD12
represented as much as 16% of total bacterioplankton (Fig. S3.9). Generally, the LD12 OTU7 relative abundance correlated negatively with salinity. At seven sites with salinities below 6, OTU7 had a relative abundance of 4.0% or greater with a median relative abundance of 8.8% (Fig. S3.9). At the remaining ten sites above a salinity of 6 (median 19.44), OTU7 had relative abundances around 2% or lower (Fig. S3.9).

Table 3.1. Compatible solute production and transport in SAR11

<table>
<thead>
<tr>
<th>Osmolyte system</th>
<th>LSUC0530</th>
<th>LD12 SAGs</th>
<th>HIMB114</th>
<th>IMCC9063</th>
<th>Other SAR11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine synthesis</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>D-myo-inositol phosphate synthesis</td>
<td>+/-</td>
<td>+/-</td>
<td>+/-</td>
<td>+/-</td>
<td>+/-</td>
</tr>
<tr>
<td>Ectoine/hydroxyectoine transport</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Glutamate synthesis</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Glutamate transport</td>
<td>-</td>
<td>+*</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Glutamine synthesis</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Glycine synthesis</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Glycine betaine synthesis</td>
<td>-</td>
<td>+/-*</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Glycine betaine/proline ABC transporter</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>N-α-Acetyl-ornithine synthesis</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Potassium transport (antiporters K03549, K03455)</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Potassium transport (trkAH + K16052)</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Proline synthesis</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Proline transport (symporter)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Sodium transport (antiporter K03313)</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Sodium transport (other antiporter)</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sorbitol synthesis</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Taurine transport</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>TMAO synthesis</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

+ genes present; - genes absent; +/- partial pathway; *B11 LD12 SAG only. Table shows only systems with predicted genes in at least one SAR11 genome. Table only shows systems for which genes could be identified in the SAR11 genomes examined in this study. For a complete list of all systems searched, see Table S1.

The phylogenetic separation of LSUC0530 from B11 and the IIIB A-C microclusters raised the question of whether LSUC0530 represents a novel ecotype with unique spatiotemporal distribution compared to other LD12 taxa. Therefore, we explored the distribution of LD12 genomes in 90 fresh and brackish water environments through competitive recruitment of metagenomic sequences to all the SAR11 genomes in the phylogenomic tree and the outgroup sequence HIMB59. We quantified relative abundance via recruited reads (per kpb of metagenomic sequence per Mbp of genome sequence- RPKM), and aggregated LD12 data based on the A, B, and C microclusters (Fig. 3.1A), the B11 SAG, and LSUC0530. Figure 5A displays RPKM values for reads with ≥ 95% identity, roughly corresponding to a species delineation via average nucleotide identity (ANI) (Konstantinidis and Tiedje, 2005), but we also provide additional comparisons at ≥ 85, 90, 92, 98, and 100% identity (Fig. S3.10).

All LD12 subgroups had their highest RPKM values in very low salinity sites, although we could detect LD12 across various brackish salinities (Fig. 3.5A). In general, IIIB.A and
LSUCC0530 averaged the highest relative abundances across all sites with salinities between 0 and 5.75. All subgroups showed site-specific variation in abundances (Fig. 3.5A). Microcluster IIIb.A clearly dominated recruitment at Lake Michigan, whereas LSUCC0530 represented the most abundant genome at Feitsui Reservoir and Lake Gatun (Fig. 3.5A).

We repeated the previous examination of recruitment density to the different microclusters across a range of percent identities, now including LSUCC0530 (Garcia et al., 2018). We corroborate observed patterns indicating additional LD12 diversity beyond that represented in the current genomes, despite the inclusion of our new representative. For example, whereas the LSUCC0530 genome recruited the majority of high percent identity reads in the Feitsui Reservoir and Lake Gatun samples (Fig. 3.5B), in some locations the majority of recruitment to this genome centered around lower percent identity hits (Fig. 3.5C). We observed the same kind of alternative patterns for the other microclusters as well (Fig. 3.5D). As noted by our colleagues, since reads were competitively recruited against all close relative genomes, these cases where a microcluster primarily recruited reads ~90-92% identity likely mean that we lack a representative genome for that fraction of the metagenomic data (Garcia et al., 2018).

3.3 DISCUSSION

Our successful cultivation of an LD12 strain for the first time has demonstrated that this organism shares many physiological traits with cultured marine SAR11 representatives, and also has some important distinctions. LSUCC0530 doubles roughly once every two days, similar to taxa in subclade Ia (Rappe et al., 2002) and Ib (Jimenez-Infante et al., 2017). It has an aerobic, chemoorganoheterotrophic lifestyle, thriving in oligotrophic media with simple carbon compounds and likely depends on reduced organosulfur compounds and ammonium as its sulfur and nitrogen sources, respectively. Cell size and shape showed considerable morphological conservation (small, curved rods roughly 1 x 0.1 µm) across large evolutionary distances (LD12 vs. subclade Ia ~89% 16S rRNA gene identity). Although a systematic survey of salinity tolerance has not been conducted for other cultured SAR11 representatives, we have shown that LD12 cannot survive in media with salinities over eight, whereas all other cultured SAR11 grow in seawater salinities (Rappe et al., 2002; Stingl et al., 2007; Song et al., 2009; Carini et al., 2013; Jimenez-Infante et al., 2017). LSUCC0530 thrives in relatively high temperatures (23-30˚C), and we could not confirm growth at low temperatures in which other SAR11 have been isolated (Song et al., 2009).

LSUCC0530 shares many genomic characteristics with marine SAR11, such as a small, streamlined genome with few paralogs, short intergenic spaces, and a conserved hypervariable region, confirming predictions from LD12 SAGs and metagenomes (Newton et al., 2011; Zaremba-Niedzwiedzka et al., 2013; Eiler et al., 2015). The LSUCC0530 genome also encodes much of the same metabolic potential as what has been reported through LD12 SAGs and metagenomic data. For example, we confirm the EMP glycolysis pathway, TCA cycle, non-mevalonate terpenoid biosynthesis, numerous amino acid synthesis pathways, TRAP transporters for C4 dicarboxylate compounds, and ABC transporters for various amino acids. We can confirm the lack of a phosphoenolpyruvate carboxylase (ppc) and missing pathways for DMSP, glycine-betaine, and methyamine metabolism.

We have also recovered important genetic information previously unknown for LD12. At 1.16 Mbp, LSUCC0530 has a smaller genome, with fewer genes, than any other SAR11. LSUCC0530 has a complete glyoxylate bypass, which differs from previous inferences using
SAGs (Eiler et al., 2015). The unusual LD12 malate synthase isoform A gene occupies the same genomic location in LSUCC0530 as the isoform G copy in most subclade Ia, Ic, and IIIa genomes, with the exception of HIMB114 (Fig. S3.7). Therefore, the most parsimonious explanation for the presence of isoform A in LSUCC0530 is a homologous recombination event that replaced the isoform G homolog. The LSUCC0530 genome also contains putative pathways for oxidation of sulfide, thiosulfate, and sulfite, which have not yet been explored in any cultured SAR11.

The relative abundance data from coastal Louisiana and the isolation of an LD12 strain from a coastal environment supports previous work indicating that LD12 occupies a much wider range of aquatic habitats than inland freshwater systems (Salcher et al., 2011; Piwosz et al., 2013; Dupont et al., 2014; Herlemann et al., 2014). Questions have been raised as to whether these organisms occur in brackish waters as active and growing community members, or whether there are present simply as a result of freshwater input. LD12 cells have been shown to uptake thymidine in brackish water (Piwosz et al., 2013). Our observation that LSUCC0530 can grow in slightly brackish water provides further evidence that such habitats may support LD12 cells as active community members. However, the relative competitiveness of LD12 within brackish microbial assemblages requires further research.

Metagenomic recruitment provides evidence that LD12 has diversified into multiple ecotypes with unique distributions (Fig. 3.5A) (Zaremba-Niedzwiedzka et al., 2013; Garcia et al., 2018). Furthermore, our data suggests that LSUCC0530 likely represents a novel LD12 ecotype because it preferentially recruits metagenomic sequences at high percent identity in different locations than the other microclusters of LD12 (Fig. 3.5). What distinguishes the distribution of these ecotypes remains unclear. The observed patterns may represent temporal variation captured in these sampling snapshots (Garcia et al., 2018). Another possibility is that small variations in salinity optima for different ecotypes may drive diversification. However, given the paucity of low-brackish sites in available datasets, we cannot establish whether salinity plays a role in ecotype diversification. Furthermore, without complete genomes from other LD12 ecotypes, we cannot rule out whether variant ionic strength management strategies, or some other metabolic characteristic, help define evolutionary differences associated with ecotypes. For example, the ability to utilize different sulfur sources may distinguish LSUCC0530 from other LD12. The sorAB sulfite dehydrogenase and the thiosulfate/3-mercaptopropionate sulfurtransferase genes have uneven distribution among SAR11 genomes, including within the LD12 subclade.

An alternative hypothesis is that temperature has driven diversification between ecotypes. A striking feature of LSUCC0530 is its growth preference for relatively high temperatures and inability to grow at low temperatures. While robust growth at 30°C may seem intuitive given that this represents typical Lake Borgne surface water temperature, the inability to grow at 12°C or below would likely place LSUCC0530-type LD12 cells at a fitness disadvantage in places like Lake Michigan, where annual water temperatures rarely get above 22°C and average temperatures remain below 15°C for the majority of the year (https://coastwatch.glerl.noaa.gov/). The metagenomic recruitment data support the temperature hypothesis: the samples in which LSUCC0530 dominated recruitment (Lake Gatun, Feitsui Reservoir) were all collected from locations having average temperatures above 19°C. Conversely, other microclusters dominated recruitment of samples from colder environments (e.g. IIIb.A in Lake Michigan) (Fig. 3.5). More sequence data may strengthen the relationship between temperature and ecotypes, but since
temperature optima are extremely hard to infer from genomes alone, additional cultivars will be needed to test the temperature hypothesis for ecotype diversification.

A critical question remains for LD12. What has made these cells uniquely adapted to low salinity waters? Genomic analyses here and elsewhere (Dupont et al., 2014; Eiler et al., 2015) have helped define the differences in LD12 metabolic potential compared to other SAR11, but very few adaptations appear related to salinity. It has been noted that the EMP pathway provides more ATP and NADH than the ED pathway, thus conferring greater energy conservation through glycolysis for LD12 cells than other SAR11s (Dupont et al., 2014; Eiler et al., 2015). If or how this energy conservation relates to salinity adaptation has not been pursued. The presence of unique LD12 cation-proton antiporters (Fig. 3) could provide a connection. LD12 cells may benefit from additional NADH for the production of a proton motive force that drives the potassium-proton antiporters, thus improving active transport to lower internal ionic strength. However, while this adaptation may improve the ability of LD12 cells to live in fresh water, it doesn’t explain why they have lost the ability to survive in salt water where their marine SAR11 cousins thrive.

The most likely explanation comes in the distribution of genes for osmolyte uptake. Critically, LSUCC0530, and LD12 in general, have lost the proVWX glycine-betaine/proline ABC transporter shared by all other marine SAR11 (Table 3.1). Although a relatively minimal number of genes, this uptake capability likely represents a significant loss of function for LD12 salinity tolerance. In Escherichia coli, the proVWX operon has a strong promoter that upregulates these genes two orders of magnitude in response to increased salinity (Dattananda and Gowrishankar, 1989). Indeed, rapid proline uptake is a common response to osmotic upshift in many bacteria (Empadinhas and da Costa, 2008). Furthermore, LSUCC0530 and other LD12 genomes have lost the ABC transporters for ectoine/hydroxyectoine present in the other SAR11 genomes. Given the slow growth of LD12, rapid production of new intracellular osmolytes in response to increased salinity seems unlikely. Thus, the loss of these transport systems for quickly equilibrating intracellular osmolarity probably prevents successful dispersal of this organism into higher salinity waters.

This finding allows us to speculate further about the evolutionary trajectory that led to successful colonization of freshwater environments by LD12. All evidence suggests that the last common SAR11 ancestor was a marine organism with a streamlined genome (Logares et al., 2010; Luo et al., 2013; Zaremba-Niedzwiedzka et al., 2013). During further diversification into subclade III, additional genes and pathways were lost, for example, ppc and genes for DMSP and some C1 compound metabolisms (Eiler et al., 2015; Sun et al., 2016). Therefore, the common ancestor of subclade IIIa and LD12 would have had to adapt to freshwater environments starting with an already reduced genome and limited metabolic flexibility. A simple way to inhabit lower salinity environments would have been to decrease expression of, or altogether lose, genes for uptake and/or production of osmolytes. If the adaptation were reversible, we would expect to see examples of LD12 organisms growing in marine systems. However, existing evidence suggests that while LD12 organisms can grow (established herein) and be active (Piwosz et al., 2013) in low brackish conditions, they are not successful in higher salinity brackish or marine water (Logares et al., 2010; Salcher et al., 2011; Dupont et al., 2014; Herlemann et al., 2014). Thus, we propose that LD12 adaptation to fresh water involved irreversible loss of function, which decreased osmotic response capability in LD12 cells, thereby preventing recolonization of higher salinity environments. The missing ABC transporters for proline/glycine betaine and ectoine/hydroxyectoine in LD12 are likely candidates for this loss of function. Experimental tests
on the differences in production and uptake of osmolytes among subclade IIIa and LD12 representatives will provide critical insight into how these different evolutionary paths have resulted in unique salinity ranges.

For the first cultured representative of the LD12 clade, we propose the provisional taxonomic assignment for strain LSUCC0530 as ‘Candidatus Fonsibacter ubiquis’,

**Description of Fonsibacter gen. nov.**

*Fonsibacter* (Fons L. noun; fresh water, spring water, -bacter, Gr. adj.; rod, bacterium. Fonsibacter referring to the preferred low salinity habitat of this organism).

Aerobic, chemoorganoheterotrophic, and oligotrophic. Cells are small, curved rods roughly 1 x 0.1µm. Non-motile. Based on phylogenomics and 16S rRNA gene phylogenetics, subclade IIIb/LD12 occurs on a separate branch within the *Pelagibacteraceae* (SAR11), sister to subclade IIIa containing strains HIMB114 and IMCC9063. Due to the depth of branching between these clades using concatenated protein-coding genes, and the 92% and 91% 16S rRNA gene sequence identity with HIMB114 and IMCC9063, respectively, we propose LSUCC0530 and subclade IIIb/LD12 as a novel genus in the *Pelagibacteraceae*.

**Description of Fonsibacter ubiquis sp. nov.**

*Fonsibacter ubiquis* (L. adv. ubiquitous).

In addition to the properties given in the genus description, the species is described as follows. Growth occurs at temperatures between 23˚C and 30˚C, but not at 11˚C or below, nor at 35˚C or above. Optimal salinity is 1.45 and below, and growth occurs between 0.36 and 4.68. At optimal temperature and salinity, cells divide at an average rate of 0.52 day⁻¹. Genome size is 1.16 Mbp, with 1271 predicted genes, and a GC content of 29.02% (calculated). LSUCC0530 had an ANI of 75.3% and 75.9% and an AAI of 58.1% and 59.7% with HIMB114 and IMCC9063, respectively. The genome is available on GenBank under accession number CP024034.

The type strain, LSUCC0530T, was isolated from brackish water from the coastal lagoon Lake Borgne off the coast of southeastern Louisiana, USA.

### 3.4 MATERIALS AND METHODS

#### 3.4.1 ISOLATION AND IDENTIFICATION OF LSUCC0530

Isolation, propagation, and identification of strain LSUCC0530 (subclade IIIb) was completed as previously reported (Henson *et al.*, 2016). Briefly, water was collected from the surface of the coastal lagoon Lake Borgne (Shell Beach, Louisiana) (29.872035 N, -89.672819 W) on 1 July, 2016. We recorded salinity on site using a YSI 556 MPS (YSI, Solon, OH, USA) (Table S1). Whole water was filtered through a 2.7 µm filter (Whatman), stained with 1X SYBR Green (Lonza, Basel, Switzerland), and enumerated using the Guava EasyCyte 5HT flow cytometer (Millipore, Massachusetts, USA). After serial dilution, water was inoculated into five 96 x 2.1 mL well PTFE plates (Radleys, Essex, UK) containing 1.7 mL of JW5 medium at an estimated 2
cells well\(^{-1}\). Cultures were incubated at room temperature in the dark for three weeks and evaluated for growth using flow cytometry. Positive wells (> 10\(^4\) cells mL\(^{-1}\)) were transferred in duplicate to capped, 125 mL polycarbonate flasks (Corning, New York, USA) containing 50 mL of JW5 medium. Upon reaching a density of > 5.0 x 10\(^5\) cells mL\(^{-1}\), cells were filtered onto 25 mm 0.22 \(\mu\)m polycarbonate filters (Millipore, Massachusetts, USA), and DNA extractions were performed using the MoBio PowerWater DNA kit (QIAGEN, Massachusetts, USA) following the manufacturer’s instructions. DNA was amplified as previously reported (Henson et al., 2016) and sequenced at Michigan State University RTSF Genomics Core. Sanger sequences were evaluated using the freely available software 4Peaks (v. 1.7.1)(http://nucleobits.com/4peaks/). Forward and reverse sequences were assembled using the CAP3 webserver (http://doua.prabi.fr/software/cap3), after reverse reads were converted to their reverse complement at http://www.bioinformatics.org/sms/rev_comp.html. The PCR-generated 16S rRNA gene sequence for strain LSUCC0530 is accessible on NCBI GenBank under the accession number KY290650.

### 3.4.2 COMMUNITY iTAG ANALYSIS

Community DNA from 17 sites of various salinities was filtered, extracted, and analyzed following our previously reported protocol (Henson et al., 2016). The sites sampled were Lake Borgne (LKB; Shell Beach, LA), Bay Pomme D’or (JLB; Buras, LA), Terrebonne Bay (TBon; Cocodrie, LA), Atchafalaya River Delta (ARD; Franklin, LA), Freshwater City (FWC; Kaplan, LA), Calcasieu Jetties (CJ; Cameron, LA). Each site was visited three times with the exception of TBon, which was visited only twice. Briefly, extracted DNA was sequenced at the 16S rRNA gene V4 region (515F, 806R) (Apprill et al., 2015; Walters et al., 2016), with Illumina MiSeq 2x250bp paired-end sequencing at Argonne National Laboratories. 16S rRNA gene amplicon data was analyzed and classified (OTU\(_{0.03}\)) with Mothur v.1.33.3 (Schloss et al., 2009) using the Silva v119 database (Pruesse et al., 2007). To determine the best matching OTU for LSUCC0530, sequences from the OTU representative fasta files, provided by mothur using `get.oturep()`, were used to create a blast database (`formatdb`) against which the LSUCC isolate 16S rRNA genes could be searched via blastn (BLAST v 2.2.26). All best hits ranked at 100% identity. OTU abundance analysis was conducted within the R statistical environment v3.2.3, within the package PhyloSeq (v1.14.0) (McMurdie and Holmes, 2013). Sequences were rarefied using the command `rarefy_even_depth()` and then trimmed for OTUs with at least 2 reads in at least 90% of the samples. Abundances of an OTU were averaged across biological duplicates and a rank abundance matrix was calculated using the command `transform_sample_counts()` with the argument function of function(x) \(x / \text{sum(x)}\). Completed rarefied rank abundance OTU tables were plotted using the graphing program GGPlot2 (v. 2.0.0) (Wickham, 2011). All iTag sequences are available at the Short Read Archive with accession number SRR6235382-SRR6235415.

### 3.4.3 GROWTH EXPERIMENTS

Salinity tolerance was tested by growing LSUCC0530 in different media distinguished by proportional changes in the concentration of major ions. All other media elements (carbon, iron, phosphate, nitrogen, vitamins, and trace metals) were kept consistent. Media included JW1, JW2, JW3, and JW4, previously published (Henson et al., 2016), and JW3.5, JW4.5, JW5, and JW6.
Recipes for all media used in this study are provided in Table S1. Salinity tolerance experiments were conducted in quintuplicate at room temperature. Temperature range and sole carbon substrate use were tested using the isolation medium, JW5. Temperature range experiments were also conducted in quintuplicate. For all experiments, growth was measured using flow cytometry as described above.

3.4.4 MICROSCOPY

Strain LSUCC0530 cells were grown to near max density (1x10⁷ cells mL⁻¹) in JW5 medium. Cells for transmission electron microscopy were prepared as previously described (Rappé et al., 2002) with one minor change: centrifuged cells were concentrated in JW5 medium with no added fixative. Cells were visualized under a JEOL JSM-2011 High Resolution transmission electron microscope at the LSU Socolofsky Microscopy Center, Baton Rouge, LA. For scanning electron microscopy, cultured cells were fixed for four hours in 4% Glutaraldehyde buffered with 0.2M Cacodylate buffer and filtered onto a 0.2 μm nylon filter (Pall, Michigan, USA). Samples were dehydrated in an ethanol series (30%, 50%, 70%, 96 and 100%) for 15 minutes each, followed by critical point drying. Cells were visualized under a JEOL JSM-6610 scanning electron microscope at the LSU Socolofsky Microscopy Center, Baton Rouge, LA.

3.4.5 GENOME SEQUENCING, ASSEMBLY, AND ANNOTATION

Cells were grown in 1 L JW5 medium (Table S1) and filtered onto 0.2 μm nylon filters (Pall, Michigan, USA). DNA was obtained via phenol-chloroform extraction and sequenced using both Illumina HiSeq and MiSeq approaches. HiSeq. 350 ng of DNA was sheared to ~500 bp with a Qsonica Q800R (QSonica) using 16 cycles with a amplitude of 25% and pulse rate of 20 seconds on, 20 seconds off. DNA was visualized on a 1.5% agarose gel stained with Gel Red (Biotum) to determine average sheared length and quantified using Qubit high sensitivity dsDNA assay kit (Invitrogen). Library prep was conducted following a modified Kapa Hyper Prep Kit protocol (Kapa Biosystems, Inc.). Briefly, we ligated end-repaired and adenylated DNA to universal Y-yoke oligonucleotide adapters and custom Illumina iTru dual-indexed primers (Glenn et al., 2016). Prior to index amplification, post-ligation LSUCC0530 DNA was cleaned using 1.2x SPRI (Rohland and Reich, 2012). Following a brief, 8 cycle PCR amplification of the library to add indexes, the genomic library was quantified using a Qubit broad range dsDNA assay kit (Invitrogen). Illumina HiSeq 3000 sequencing was performed at Oklahoma Medical Research Facility using SBS kit chemistry to generate 150 bp paired-end reads, insert size 400 bp. MiSeq. DNA was sent to the Argonne National Laboratory Environmental Sample Preparation and Sequencing Facility which performed library prep and sequencing, generating 250 bp paired-end reads, insert size 550 bp.

Due to the large number of reads from the HiSeq sequencing, we created a random subset using 14% of the total reads by using the program seqtk (v. 1.0-r75-dirty) and the flag –s100. We trimmed reads for adapter contamination and quality using Trimmomatic (Bolger et al., 2014). Initial assembly with the subset of reads was done with SPAdes (Bankevich et al., 2012). This generated a large single scaffold with overlapping ends. We performed a quality assessment of the single scaffold via iterations of Reapr (Hunt et al., 2013)-advised scaffold breaks and SSPACE (Boetzer et al., 2011) extensions using both the full set of HiSeq reads and the MiSeq reads. Final assessment of the scaffold with overlaps removed from the ends was performed with
Pilon (Walker et al., 2014) using all HiSeq reads mapped to the scaffold with BWA (Li and Durbin, 2009). No issues were reported. An analysis of the GC skew was performed on the final scaffold using the gc_skew script provided via Brown et al., 2015 (Brown et al., 2015). Contamination and completion were also evaluated with CheckM (Parks et al., 2015). Detailed commands and outputs for the assembly and quality assessment are provided in Supplementary Information. The final circular chromosome scaffold was annotated by IMG (Markowitz et al., 2011) and is publicly available with IMG Taxon ID number 2728369501. It is also publicly available in GenBank with accession number CP024034. Genome sequencing reads are available in SRA with accession numbers SRR6257553 (MiSeq) and SRR6257608 (HiSeq).

### 3.4.6 COMPARATIVE GENOMICS

Orthologous clusters were determined using the Get Homologues pipeline (Contreras-Moreira and Vinuesa, 2013) using the LSUCC0530 genome and 48 publicly available SAR11 genomes from IMG, including the 10 LD12 SAGs analyzed previously (Zaremba-Niedzwiedzka et al., 2013; Eiler et al., 2015) (Table S1). Clusters were determined from amino acid sequences using the OrthoMCL option in Get Homologues. All clusters are provided in cluster_list.txt as part of Supplemental Information. Average amino acid identity (AAI) and synteny were calculated as previously reported (Dick et al., 2009). Average nucleotide identity (ANI) was calculated with the Konstantinidis Lab website ANI calculator (http://enve-omics.ce.gatech.edu/ani/index) as reported (Goris et al., 2007). Compatible solutes identified in Table S1 were identified in SAR11 genomes using KO numbers, annotations, and cross-referenced with orthologous clusters.

### 3.4.7 BACTERIOPHAGE SEARCHES

We looked for signatures of bacteriophage in the LSUCC0530 genome using two methods, VirSorter (Roux et al., 2015) and PHASTER (Arndt et al., 2016) with default settings. CRISPR regions are identified as part of the standard IMG annotation process, and none were found in LSUCC0530.

### 3.4.8 PHYLOGENOMICS

We selected single-copy genes present in at least 39 of the 49 genomes as determined via Get Homologues (above). Amino acid sequences from these 83 markers were separately aligned with MUSCLE (Edgar, 2004), culled with Gblocks (Castresana, 2000), and concatenated into a superalignment totaling 23,356 positions. Inference was performed with RAxML (Stamatakis et al., 2008) using the PROTCATLG model and 100 bootstrapping runs. All scripts (with settings) used in the alignment, culling, concatenation, and inference processes are available in Supplemental Information. Node labels were renamed with Newick Utilities (Junier and Zdobnov, 2010) nw_rename, and the tree was visualized with Archaeopteryx (Han and Zmasek, 2009).

### 3.4.9 SINGLE GENE PHYLOGENY

The 16S rRNA gene sequence from the LSUCC0530 genome was aligned with 284 alphaproteobacterial sequences and seven outgroup sequences. We completed phylogenetic
inference for aceA and malA in an analogous manner. For each, the LSUCC0530 gene was searched against the NCBI nr database using blastp (v. 2.2.28). The top 100 sequence hits were combined with all homologous SAR11 sequences identified with Get Homologues for the aceA phylogeny, and both the malA and malG sequences in all SAR11s for the malA phylogeny. Sequences from redundant taxa were removed. For all trees, alignment with MUSCLE (Edgar, 2004), culling with Gblocks (Castresana, 2000), and tree inference with FastTree2 (Price et al., 2010) was completed with the FT_pipe script, provided in Supplemental Information, along with the initial fasta file containing all sequences and the newick tree output file. Node labels were changed with Newick Utilities (Junier and Zdobnov, 2010) nw_rename and visualization was performed with Archaeopteryx (Han and Zmasek, 2009).

3.4.10 METAGENOMIC RECRUITMENT

Scaffolds for the 49 SAR11 genomes, including strain LSUCC0530, used in the phylogenomic analyses were used for competitive recruitment of metagenomic sequences from 89 different samples. Prior to recruitment, multi-scaffold assemblies were condensed into single scaffolds, and all single scaffolds were used to generate a BLAST database with the command makeblastdb (BLAST v. 2.2.28). Metagenomic datasets from 90 different environmental samples of various salinities from the Baltic Sea (Dupont et al., 2014), Amazon River, freshwater lakes (Martinez-Garcia et al., 2012; Eiler et al., 2014), Lagoon Albufera (Ghai et al., 2012), Columbia River (Smith et al., 2013; Fortunato and Crump, 2015), Lake Lanier (Oh et al., 2011), Lake Michigan (Denef et al., 2016), Feitsui Reservoir (Tseng et al., 2013), and GOS (Rusch et al., 2007) were downloaded from the Short Read Archive (SRA) and European Nucleotide Archive (ENA) using ftp and SRA toolkit (v 2.8.2-1), respectively. Reads from pyrosequencing were dereplicated with the program CD-HIT (v 4.6) (Fu et al., 2012) using the cd-hit-454 command and flag -c 0.95. Illumina sequence data was converted from fastq to fasta files with the FastX toolkit (v. 0.0.13.2) (http://hannonlab.cshl.edu/fastx_toolkit) using the command fastq_to_fasta with the flag –Q33. Metagenomic sequences were separated into 1 million sequence files and recruited against the SAR11 database with blastn (v. 2.2.28) with –max_target_seqs 1. Best hits for each metagenomic sequence were kept if the alignment length was > 90% of the median length of all metagenomic sequences for a given sample. Metagenomic recruitment was expressed in read density per genome using ggplot2 geom_density() and reads per kilobase of genome per million mapped reads (RPKM) as previously reported (Cameron Thrash et al., 2017).

3.4.11 SCRIPT AVAILABILITY

All scripts used in this work can be found on the Thrash Lab website with the manuscript link at http://thethrashlab.com/publications.
CHAPTER 4.
COASTAL GULF OF MEXICO HIGH THROUGHPUT CULTIVATION EXPANDS CULTIVAR DIVERSITY

4.1 INTRODUCTION

Understanding the physiological and genetic characteristics of bacterioplankton is essential to building a predictive capacity for how alterations to community membership will impact biogeochemical cycling and ecosystem health. While cultivation-independent approaches have provided important genomic-based inferences about these differences (Handelsman, 2004; Rappé, 2013; Rinke et al., 2013; Swan et al., 2013), numerous factors still complicate the interpretation of data generated from these methods, including many genes of unknown function, cryptic metabolisms, and likely multifaceted interactions (Rappé, 2013; Overmann et al., 2017; Lloyd et al., 2018b). Cultivation-based experiments of environmentally relevant organisms provide a valuable method for experimentally testing hypotheses generated from in situ observations. However, the vast majority of organisms remain uncultivated - an observation coined the “great plate count anomaly” (Staley and Konopka, 1985).

To address the paucity of cultivated representatives from important bacterioplankton clades, scientists have employed various cultivation techniques (Kaeberlein et al., 2002; Zengler et al., 2002; Nichols et al., 2010; Steinert et al., 2014; Tandogan et al., 2014), with dilution-to-extinction of natural communities combined with high throughput culturing (HTC) achieving considerable success (Schut et al., 1993; Connon and Giovannoni, 2002; Stingl et al., 2007; Hahnke et al., 2015; Sosa et al., 2015; Yang et al., 2016; Henson, Lanclos, et al., 2018). Pioneered by Don Button and colleagues for cultivation of oligotrophic bacteria (Button et al., 1993), this method dilutes microbial communities to near single cells using sterile natural seawater as the medium (Button et al., 1993; Schut et al., 1993). These conditions avoid the challenge of slow-growing, obligate oligotrophic bacterioplankton from being outcompeted by fast-growing, less abundant copiotrophic organisms, and provide these taxa with the same chemical surroundings in which they are normally found (Button et al., 1993). Improvements to this method in multiple labs has increased the number of inoculated wells and decreased the time needed to detect growth (Connon and Giovannoni, 2002; Stingl et al., 2007; Song et al., 2009).

In a previous study, we showed that this approach could also be successfully utilized with artificial seawater media (Henson et al., 2016). Thus far, HTC methods using both artificial and natural seawater have been responsible for the first ever isolation of many numerically abundant marine and freshwater groups, such as marine SAR11 Alphaproteobacteria (Rappé et al., 2002; Stingl et al., 2007; Song et al., 2009; Henson et al., 2016; Henson, Lanclos, et al., 2018), the freshwater SAR11 LD12 clade (Henson, Lanclos, et al., 2018), SUP05/Arctic96BD-19 Gammaproteobacteria (Marshall and Morris, 2013; Shah et al., 2017), OM43 Betaproteobacteria (Connon and Giovannoni, 2002; Huggett et al., 2012; Sosa et al., 2015; Yang et al., 2016), HIMB11-type Roseobacter (Durham, Grote, Whittaker, Bender, Luo, Grim, Brown, Casey, Dron, Florez-Leiva, and Others, 2014; Henson et al., 2016), numerous so-called oligotrophic marine Gammaproteobacteria (Cho and Giovannoni, 2004), and acl Actinobacteria (Kim et al., 2018).

The isolation of these organisms and others derived from different cultivation efforts has provided important genomic, metabolic, and physiological data that compliment cultivation-independent derived hypotheses. Experiments using Ca. Nitrosopumilus maritimus strain SCM-1...
and Candidatus Pelagibacter sp. strain HTCC7211 to demonstrate the production of methylphosphonic acid and its respiration to methane, respectively, were critical for explaining methane saturation in the surface ocean (Karl et al., 2008; Metcalf et al., 2012; Paul Carini, White, et al., 2014). Further, the role of proteorhodopsin could only be theorized until experiments using SAR11 strain HTCC1062, Flavobacterium, and other cultures revealed that it helps with maintaining cellular functions during states of carbon starvation (Gómez-Consarnau et al., 2007; Steindler et al., 2011). The use of cultivation strategies in concert with cultivation-independent approaches and environmental measurements will continue to be crucial to validating unknown functions, metabolisms, and cross-feeding interactions that drive biogeochemical cycling (Giovannoni and Stingl, 2007b).

Northern Gulf of Mexico (nGOM) ecosystems harbor economically important (Rabalais et al., 1996; Adams et al., 2004) and ecologically diverse environments (Bianchi et al., 1999) that are under the influence of freshwater, estuarine, and marine systems and host unique coastal microbial assemblages (Rappe et al., 1997; Yeo et al., 2013; Chafee et al., 2018). While studies in the nGOM have been valuable for detailing the microbial communities and their metabolic capacities associated with the “dead zone” (Gillies et al., 2015; Cameron Thrash et al., 2017) and oil spills (Mason et al., 2012; King et al., 2015), the microbial communities found throughout the coastal area remain understudied (Olapade, 2010; King et al., 2012). In one study of bacterioplankton communities prior to the 2010 Deepwater Horizon Oil spill, the nGOM surface waters were found to be dominated by Alphaproteobacteria and Bacteriodetes, a result supported by a clonal study of Florida coastal GOM sites from 2008-2009 (Olapade, 2010), with a majority (> 60%) of Alphaproteobacteria belonging to the SAR11 clade (King et al., 2012). Analysis of the communities from the 17 sites found that while the composition of the bacteria present varied significantly, the overall structure did not. The authors concluded that environmental filtering largely drove community composition, while other processes such as competitive exclusion also played a role (King et al., 2012). However, (King et al., 2012), 2012 was a snapshot of the nGOM microbial community during the 2010 season, and only two of the sites were near the coastal nGOM. Further, Olapade, 2010 only sampled three sites along a well-established beach and the majority of clone libraries belonged to the “unknown” category, hampering the conclusions made in the study. While these studies provided evidence of community structure and composition, the potential unique physiologies and metabolisms of native coastal bacterioplankton still needs to be further explored.

We combined HTC cultivation and amplicon-sequencing community assessment over a three-year campaign at six sites along the coastal nGOM. Previously, we established a robust protocol for HTC using artificial seawater in these systems (Henson et al., 2016). Here we expand our report beyond our original seven HTC experiments to evaluate the results from a total of seventeen. Many of the Louisiana State University Culture Collection (LSUCC) cultivars belonged to putatively novel Species and Genera, expanding the diversity of cultured representives of clades from the nGOM and beyond. Salinity was a major driver in the distribution of most LSUCC cultivars, but some clades (e.g., OM43 and SAR11 subclade IIIa) were found to be continually abundant despite the dynamic environmental conditions. While relative abundance was a good predictor of cultivation success for some well-represented clades in the culture collection, it was unrelated to cultivation success and the number of isolates for others. This study provides valuable insights into the distribution of many abundant aquatic clades in the nGOM, expands the cultivar diversity, and explores the mechanisms underlying cultivation success.
4.2 METHODS AND MATERIALS

4.2.1 SAMPLING

Surface water samples were collected at six sites yearly for three years, except for Terrebonne Bay, which was collected twice. The sites sampled were Lake Borgne (LKB; Shell Beach, LA), Bay Pomme d’Or (JLB; Buras, LA), Terrebonne Bay (Tbon; Cocodrie, LA), Atchafalaya River Delta (ARD; Franklin, LA), Freshwater City (FWC; Kaplan, LA), and Calcasieu Jetties (CJ; Cameron, LA) (Table S1). Water collection for biogeochemical and biological analysis followed the protocol in (Henson et al., 2016). Briefly, we collected surface water in a sterile, acid-washed polycarbonate container. Duplicate 120 ml water samples were filtered serially through 2.7 μm Whatman GF/D (GE Healthcare, Little Chalfont, UK) and 0.22 μm Sterivex (Millipore, Massachusetts, USA) filters and placed on ice until transferred to -20°C in the laboratory (maximum 3 hours on ice). The University of Washington Marine Chemistry Laboratory analyzed duplicate subsamples of 50 ml 0.22 μm-filtered water collected in sterile 50 ml falcon tubes (VWR, Pennsylvania, USA) for concentrations of SiO$_4^{4-}$, PO$_4^{3-}$, NH$_4^+$, NO$_3^-$, and NO$_2^-$. Samples for cell counts were filtered through a 2.7-μm GF/D filter, fixed with 10% formaldehyde, and stored on ice until enumeration (maximum 3 hours).

4.2.2 HIGH THROUGHPUT CULTURING AND ISOLATE IDENTIFICATION

Isolation, propagation, and identification of isolates were completed as reported in Henson et al., 2016, 2018 (Henson et al., 2016; Henson, Lanclos, et al., 2018). A subsample of 2.7 μm filtered surface water was stained with 1X SYBR Green (Lonza, Basal, Switzerland), and enumerated using a flow cytometer as described in Thrash et al., 2015 (Thrash et al., 2015). After serial dilution to a predicted 1-3 cells/μL, 1-2 μl water was inoculated into five, 96-well PTFE plates (Radleys, Essex, UK) containing 1.7 ml artificial seawater medium (Table S1) to achieve an estimated 1-3 cells well$^{-1}$. Salinity of the medium was chosen based on in-situ salinity measurements after experiment JLB (January 2015) (Table S1). Cultures were incubated at in-situ temperatures (Table S1) in the dark for three to six weeks and evaluated for positive growth ($> 10^4$ cells ml$^{-1}$) by flow cytometry. 200 μl from positive wells was transferred to duplicate 125 ml polycarbonate flasks (Corning, New York, USA) containing 50 ml of medium. At FWC, FWC2, JLB2c, and JLB3, not all positive wells were transferred because of the large number of positive wells. At each site, 45/301, 60/403, 60/103, and 60/146 of the positive wells were transferred, respectively. Cultures reaching maximum density, or $> 1 \times 10^5$ cells ml$^{-1}$, had 35 mL of the 50 mL volume filtered for identification via 16S rRNA gene PCR onto 25-mm 0.22-μm polycarbonate filters (Millipore, Massachusetts, USA). DNA extractions were performed using the MoBio PowerWater DNA kit (QIAGEN, Massachusetts, USA) following the manufacturer’s instructions and eluted in sterile water. The 16S rRNA gene was amplified as previously reported in (Henson et al., 2016) and sequenced at Michigan State University RTSF Genomics Core. Evaluation of Sanger sequences quality was performed with the freely available software 4Peaks (v. 1.7.1) (http://nucleobytes.com/4peaks/) and forward and reverse complement sequences (http://www.bioinformatics.org/sms/rev_comp.html) were assembled where overlap was sufficient using the CAP3 web server (http://doua.prabi.fr/software/cap3).
4.2.3 COMMUNITY iTAG SEQUENCING

Community DNA from the 34 samples were filtered, extracted, and analyzed using our previously reported protocols and settings (Henson et al., 2016; Henson, Hanssen, et al., 2018). Extracted DNA was sequenced at the 16S rRNA gene V4 region using the primers 515F, 806R (Apprill et al., 2015; Walters et al., 2016), with Illumina MiSeq 2x250bp paired-end sequencing at Argonne National Laboratories. Sequencing of the 16S rRNA gene amplicons resulted in 13,485,352 raw sequences. Using Mothur v1.33.3 (Schloss et al., 2009) and the Silva v119 database (Pruesse et al., 2007; Quast et al., 2013), 16S rRNA gene amplicon data were clustered into distinctive OTUs with a 0.03 dissimilarity threshold (OTU$_{0.03}$) and classified. After these steps, 55,256 distinct OTUs$_{0.03}$ remained.

4.2.4 AMPLICON SINGLE NUCLEOTIDE VARIANTS

A second approach using minimum entropy decomposition (MED) was used to discern amplicon single nucleotide variants (ASV) across the three-year community dataset (Eren et al., 2014). MED uses Shannon entropy to identify nucleotide variation within amplicon datasets that are useful to partition reads into finer scale clusters. Reads were first analyzed using Mothur as described above up to the screen.seqs() command. The cleaned reads fasta file was converted to MED-compatible headers with the ‘mothur2oligo’ tool renamer.pl using the fasta output from screen.seqs() and the Mothur group file (https://github.com/michberr/MicrobeMiseq/tree/master/mothur2oligo). In total, 4,217,322 reads were analyzed using MED (v. 2.1) with the flags -M 60, -V 3, and -d 1. MED resulted in 2,813 refined ASVs. ASVs were classified in Mothur using classify.seqs(), the Silva v119 database, and a cutoff bootstrap value of 80%. After classification, reads identified as “chloroplast”, “mitochondria”, or “unknown” were removed from the dataset.

4.2.5 COMMUNITY DATA ANALYSIS

OTU (OTU$_{0.03}$) and ASV abundances were analyzed within the R statistical environment v.3.2.1 (R Development Core Team 2015) following previously published protocols (Henson et al., 2016; Henson, Hanssen, et al., 2018; Henson, Lanclos, et al., 2018). Using the package PhyloSeq (McMurdie and Holmes, 2013), OTUs and ASVs were rarefied using the command rarefy_even_depth() and OTUs/ASVs not appearing at least two times in $\geq 11\%$ of the samples were removed. After filtering, the datasets contained 1,073 and 1,323 unique 16S rRNA gene clusters for OTUs and ASVs, respectively. For site-specific community comparisons, beta-diversity between sites was examined using Bray-Curtis distances via ordination with non-metric multidimensional scaling (NMDS). The influence of Hellinger transformed environmental parameters on beta-diversity was calculated in R with the envfit function. OTU and ASV reads were averaged across biological duplicates, and relative abundance matrices were calculated using the command transform_sample_counts() with the argument function of function(x) x / sum(x). Completed rarefied rank abundances were plotted using the graphing program GGPLOT2 (v. 2.0.0) (Wickham, 2011). To determine the best matching OTU or ASV for a given LSUCC isolate, the OTU representative fasta file, provided by Mothur using get.oturep(), and the ASV fasta file were used to create a BLAST database (makeblastdb) against which the LSUCC isolate 16S rRNA genes could be searched via blastn (BLAST v 2.2.26). For a LSUCC
isolate 16S rRNA gene to be associated with an ASV sequence, hits had to share 99-100% sequence identity.

4.2.6 SPARCC

We used SPARCC (Friedman and Alm, 2012) to examine covarying taxa following a similar protocol as in Chafee et al., 2017. To implement SparCC, the command SparCC.py was used to iteratively (-i 20) calculate the log-transformed relative abundance and compute the Pearson correlations with a cutoff correlation value of 0.25. One hundred bootstrap replicates were run using MakeBootstraps.py and then SparCC.py was re-run with 20 iterations to calculate one-sided p-values with PseudoPvals.py. The correlation output was then converted to a network file with the tool cor_to_network.py (https://github.com/hallamlab/utilities/tree/master/SparCC). To focus on strong, significant interactions, correlations greater than or equal to 0.80 were used for all further analysis.

4.2.7 ACTUAL VERSUS EXPECTED NUMBER OF CULTIVARS

We estimated the number of cultivars we should expect, E, for a particular taxon i from any given cultivation experiment from sample j based on the following calculation:

\[ E = C_{ij} \times d \times w \]

where \( C_{ij} \) is the estimated number of cells ml\(^{-1}\) for a taxon i based on the product of its relative abundance \( R_i \) and the total cells ml\(^{-1}\) of the inoculum \( C_j \): \( C_{ij} = R_i \times C_j \); the dilution factor \( d = (\text{cells well}^{-1}/1.7 \text{ ml})/\text{cells ml}^{-1} \); and \( w \) is the number of inoculated wells. Total volume in the wells was 1.7 mL. The expected and actual number of cultivars was plotted using the graphing program GGPLOT2 (v. 2.0.0) (Wickham, 2011).

4.2.8 16S rRNA GENE PHYLOGENY

The 16S rRNA gene phylogeny was completed as reported in Henson et al., 2016. Briefly, 16S rRNA genes of LSUCC cultivars were searched against the nt database (accessed August 2018) and a selection of best hits was generated from each blast search, combined with sequences from known aquatic clades, aligned with MUSCLE (Edgar, 2004), culled with Gblocks (Castresana, 2000), and trees were inferred with FastTree2 (Price et al., 2010) via the FT_pipe script. All trees were visualized with Archaeopteryx (Han and Zmasek, 2009). Fasta files and alignments are provided in the supplemental information.

4.2.9 ASSESSMENT OF CULTIVAR NOVELTY

The NCBI nt database was downloaded on August 8th, 2018 and used to assess the percent identity of LSUCC isolate 16S rRNA genes to known cultivars in NCBI. LSUCC isolates were placed into 49 taxonomic groups based on sharing ≥ 94% identity and/or occurring in monophyletic groups within our 16S rRNA gene trees (Figs. S4.1-5). The first cultivated isolate was used as the representative sequence for blastn searches. All representative sequences were searched against the nt database using blastn (BLAST+v. 2.7.1) with the flags -perc_identity 84, -eval 1E-6, -task blastn, -outfmt "6 qseqid sseqid pident length slen mismatch evalue bitscore sscinames sblastnames stitle ", and -negative_gilist to remove uncultured and
environmental sequences. The negative GI list was obtained by searching "environmental
samples"[organism] OR metagenomes[orgn]" in the NCBI Nucleotide database (accessed
September 12th, 2018) and hits were downloaded in GI list format. The resultant hits from the nt
database search were curated to remove sequences classified as single cell genomes, clone
libraries, duplicates, and previously deposited LSUCC isolates.

We observed that many HTCC, IMCC, and HIMB isolates that fell within our clades
(Figs. S4.1-5) were missing from nt hits, so isolate accession numbers from (Page et al., 2004;
Stingl et al., 2007, 2008; Song et al., 2009; Marshall and Morris, 2013; Hahnke et al., 2015;
Sosa et al., 2015; Yang et al., 2016; Kim et al., 2018), 2016, (Rappé et al., 2002; Stingl et al.,
2007; Song et al., 2009; Carini et al., 2013; Jimenez-Infante et al., 2017), Connon and
Giovannoni 2002, and (Page et al., 2004; Stingl et al., 2007, 2008; Song et al., 2009; Marshall
and Morris, 2013; Hahnke et al., 2015; Sosa et al., 2015; Yang et al., 2016; Kim et al., 2018).
2004 were extracted from the nt database via blastdbcmd and made into a separate HTC database
using makeblastdb. Accession numbers matching hits from the nt search were removed. Finally,
the 328 LSUCC isolate sequences were compared to each other to quantify all pairwise
identities. The HTC and LSUCC searches were done using the command blastn (BLAST+v.
2.7.1) with the flags -perc_identity 84, -task blastn. Hits for all searches with \( \leq 85\% \) of the length
of at least the subject or query sequence were removed. The resultant three lists of hits were
combined and any HTC database hits that fell below the lowest percent identity hit to the nt
database were removed. Strains from the HTC culture collections were labeled with the
corresponding collection name, while all other hits were labeled as “General”. Plots were created
using the graphing program GGPlOT2 (v. 2.0.0) (Wickham, 2011).

4.2.10 DATA ACCESSIBILITY

All iTag sequences are available at the Short Read Archive with accession numbers
SRR6235382-SRR6235415. PCR-generated 16S rRNA gene sequences are accessible on NCBI
GenBank under the accession numbers XXXXX-XXXXX.

4.3 RESULTS

We conducted a total of 17 HTC experiments combined with microbial community
characterization of source waters over three years and across six coastal Louisiana sites. We
inoculated a combined 7,820 distinct cultivation wells with an estimated 1-3 cells well\(^{-1}\). A total
of 1448 wells were considered positive (> \(10^4\) cells ml\(^{-1}\)), and 738 of these were transferred to
larger 125 mL polycarbonate flasks (Table 4.1). Of the 738 transferred wells, 328 were capable
of being repeatedly transferred. Site percent cultivability, using the calculations from Button et
al., 1993 (Button et al., 1993), ranged from 0% to 104% with a median and mean percent
cultivability of 6.7 and 14.7%, respectively (Table 4.1) (Button et al., 1993). FWC was the only
site counted using microscopy and represents an extreme outlier in the calculated percent
cultivability. We believe the > 100% calculation was most likely caused by the underestimation
of the number of cells inoculated into each well.
Figure 4.1. Percent identity of LSUC cultivars compared to others that are publicly available in NCBI (“general”, gray dots) or from the culture collections IMCC (gold dots), HTCC (blue dots), and HIMB (green dots). Purple dots represent LSUC isolates belonging to that group based on ≥ 94% identity and phylogenetic placement. Above the graph is the percent identity to the closest non-LSUC cultivar to the representative LSUC isolate. Groups labeled in red represent clades putatively novel genera, whereas orange indicates putatively novel species.

The 328 isolates belonged to three Phyla: Proteobacteria (n=319), Actinobacteria (n=8), and Bacteroidetes (n=1) (Figs. S4.1-S4.5). We placed these cultivars into 49 groups based on their occurrence in monophyletic groups (Figs. S4.1-S4.5) and having ≥ 94% 16S rRNA gene sequence identity, and provided a general nomenclature ID based on other cultivated representatives (Table S1). Cultures from six groups represent putatively novel genera, with < 94.5% 16S rRNA gene identity to a previously cultured representative. These include the acIY Actinobacteria subclades A and B, and one other unnamed Actinobacterial group; an undescribed Acetobacteraceae (Alphaproteobacteria) clade; the freshwater SAR11 LD12 (Candidatus Fonsibacter ubiquis, (Henson, Lanclos, et al., 2018), and the MWH-UniPo-type
Betaproteobacteria (Fig. 4.1, Table S1). Cultures in seven groups represent novel species (between 94.6 and 96.9% 16S rRNA gene sequence identity), including three groups in the Commamonadaceae and Burkholderiales (Betaproteobacteria), as well as the Porticocccaceae (Gammaproteobacteria) (Fig. 4.1). LSUCC isolates belonging to the groups BAL58 Betaproteobacteria (Fig. S4.3), OM252 Gammaproteobacteria, HIMB59 Alphaproteobacteria, and the LSUCC0101-type Gammaproteobacteria had close matches to other cultivated taxa at the species level, however, none of those taxa have been formally described (Fig. 4.1; Fig. S4.1-5; Table S1). In total, the OM252, BAL58, and MWH-UniPo-type clades contained 127 of our 328 cultivars (Table S1). Thus, with the isolation of numerous putative novel Species and Genera, as well as other important aquatic clades, we have dramatically increased the cultivated diversity of nGOM organisms and within poorly described groups.

To assess the relative efficacy of our cultivation approach for isolating taxa with different natural abundances, we matched LSUCC isolate 16S rRNA gene sequences with those amplified from whole bacterioplankton communities in all our samples. We analyzed bacterioplankton communities using both operational taxonomic units (OTUs), and amplicon single nucleotide variants (ASVs). OTUs are groups of clustered sequences based on 3% dissimilarity, while ASVs are OTUs that were further differentiated by the location of the nucleotide variation found in the amplified 16S rRNA gene region. We used ASVs because of their ability to delineate within-OTU diversity and to help explore potential changes in cultivation efficiency that may have been associated with relative abundance. Since many abundant bacterioplankton clades contain only single copies of the 16S rRNA gene, ASVs can provide beneficial resolution to further differentiate these groups (Eren et al., 2014; Needham and Fuhrman, 2013).
Figure 4.2. Rank abundances of the 50 most abundant taxa at salinities less than 7 (top) and greater than 12 (bottom) for OTUs (A, B) and ASVs (C, D) across all sites. The boxes indicate the interquartile range (IQR) of the data, with vertical lines indicating the upper and lower extremes according to 1.5 x IQR. Horizontal lines within each box indicate the median. The data points comprising the distribution are plotted on top of the boxplots. The shade of the dot represents the salinity, while the color of the box is the Phylum of the OTU. OTUs and ASVs with cultured representatives from the LSU culture collection are labeled with LSUCC.

LSUCC cultivars and their associated OTUs (Fig. 4.2A,B and S4.6A) and ASVs (Fig. 4.2C,D and S4.6B) represented many members of the microbial majority during the three-year cultivation effort in the coastal nGOM. Overall, we found 14 and 10 LSUCC associated OTUs and ASVs, respectively, within the 50 most abundant taxa (Fig. S4.6; Table S1). These include the acIIV subclade A, SAR11 subclades IIIa and LD12, HIMB11-types, and the OM43 clade. Because salinity was the primary driver separating communities based on beta diversity (ANOSIM, OTU: R2=0.89, P = 0.001, ASV: R2= 0.842, P = 0.001), we explored rank abundances of our LSUCC cultivars in salinities below 7 and above 12. At sites with salinities < 7, we found 12 and 8 OTUs and ASVs, respectively, associated with LSUCC cultivars in the top
50, including the 2nd most abundant taxa in both analyses, SAR11 LD12 (Fig. 4.2A and B, Table S1). At salinities > 12, there were 13 and 11 LSUCC associated OTUs and ASVs, respectively, present in the 50 most abundant taxa (Fig. 4.2C and D, Table S1). These taxa included groups from SAR11 subclade IIIa.1, OM43 clade, OM182 clade, and LSUCC0101-type (Fig. 4.2C and D). Some groups such as SAR11 subclade IIIa.1 and OM43 clade were present in the top 15 most abundant taxa regardless of the salinity (Fig. 4.2, Table S1). These results provide further evidence of the successful cultivation of taxa from various environmental conditions, specifically salinity, that represent taxa within the microbial majority at both broad scale (OTUs) and fine-scale resolutions (ASV).

**Figure 4.3.** Relative abundance of the ASVs represented by ten cultivars according to site salinity. The color of the line represents the different LSUCC associated ASVs. Non-linear regressions have been added for reference. Triangles represent site samples from which isolates were cultivated.

We asked if cultivation success, defined as the isolation of an organism using our artificial media, was related to relative abundance. For cultivars that were rarely isolated; such as aclV, SAR11 subclade III, and HIMB59-type clades; the isolation of a cultivar appeared unrelated to relative abundance. Of the most abundant isolate associated ASVs for those clades, we found that none were cultivated at sites where the LSUCC-ASV rank abundance was within the top two most abundant (Fig. S4.8, Table S1). For instance, SAR11 subclade IIIa.1 was cultivated at experiments when the LSUCC-ASV was the fourth, eighth, and seventeenth most abundant (Fig. 4.3). Similarly, we isolated the LSUCC0101-type cultivars at sites where the associated ASV had mid-rank relative abundances (CJ2 and TBON3). We did not cultivate any LSUCC0101-type isolates during the spike of relative abundance observed at JLB3 (Fig. 4.3, Table S1). Comparatively, the OM252 clade, HIMB11-type, BAL58, and MWH-UniPo-type clades were often isolated in experiments when their ASV was at peak relative abundances, suggesting relative abundance was a good predictor of cultivation success for these clades (Fig. 4.3).
4.3). However, we did observe variation between relative abundance and cultivation for different ASVs within those well represented clades. For example, the LSUCC cultivars from the MWH-UniPo-type associated with type 6 (ASV7146) were cultivated at sites with the highest relative abundance for the LSUCC-ASV, but we failed to isolate type 1 (ASV7025) from experiments with similar relative abundances or at the same site during the 1st and 3rd year (Fig. 4.3, Table S1).

Since cultivation success for clades now well represented in our LSUCC appeared related to relative abundance, we further explored if relative abundance could successfully predict the number of organisms isolated from that clade (Fig. 4.4). Relative abundance poorly predicted the number of cultivars from an experiment for isolates from the clades SAR11 LD12, SAR11 subclade III, and acIV (Fig. 4.4, Table S1). For well-represented clades such as OM252,

![Figure 4.4](image)

**Figure 4.4.** The number of LSUCC isolates obtained in a given experiment versus an estimated frequency of cultivation given their relative abundance in the inoculum (A). Bacterial clades of interest are colored, while all others are gray. The blue line represents the expected one to one ratio of expected versus actual if all conditions for cultivation were met. Inset: An enhanced view to show isolates inside the lower range.

OM252, MWH-UniPo-type, and BAL58, we found that at sites when the associated ASV had high relative abundance, it often predicted the number of cultivars isolated during that experiment (Fig 4.4 inset, Table S1). However, as the relative abundance decreased for the associated ASV,
the ability to predict the number of isolates became more inconsistent (Table S1). For example, we found that for experiments from the top three sites based on relative abundance of MWH-UniPo-type ASV type 6 the predicted number of isolates was near the actual number of cultivars (Table S1). However, at the next four sites, the relative abundance of the ASV only predicted the number of isolates from two experiments (Table S1). These predictions based on ASV relative abundance were only made for ASVs that were successfully cultivated on our artificial medium. The variation we observed between the relative abundance and its ability to predict the number of isolates showcases the complexity associated with cultivation even when a medium is known to provide the nutrient requirements that support the isolation and growth of a taxon.

<table>
<thead>
<tr>
<th>Site</th>
<th>Date</th>
<th>Inoculated Wells</th>
<th>Positive Wells</th>
<th>p</th>
<th>X</th>
<th>V</th>
<th>% Culturability*</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>CJ</td>
<td>September 2014</td>
<td>460</td>
<td>15</td>
<td>0.033</td>
<td>1.27</td>
<td>0.026</td>
<td>2.6</td>
<td>(Henson et al., 2016)</td>
</tr>
<tr>
<td>ARD</td>
<td>November 2014</td>
<td>460</td>
<td>1</td>
<td>0.002</td>
<td>1.5</td>
<td>0.001</td>
<td>0.1</td>
<td>(Henson et al., 2016)</td>
</tr>
<tr>
<td>JLB</td>
<td>January 2015</td>
<td>460</td>
<td>61</td>
<td>0.133</td>
<td>1.96</td>
<td>0.073</td>
<td>7.3</td>
<td>(Henson et al., 2016)</td>
</tr>
<tr>
<td>FWC</td>
<td>March 2015</td>
<td>460</td>
<td>301</td>
<td>0.654</td>
<td>2</td>
<td>0.531</td>
<td>53.1</td>
<td>(Henson et al., 2016)</td>
</tr>
<tr>
<td>LKB</td>
<td>May 2015</td>
<td>460</td>
<td>0</td>
<td>0</td>
<td>1.8</td>
<td>0</td>
<td>0</td>
<td>(Henson et al., 2016)</td>
</tr>
<tr>
<td>TBON</td>
<td>August 2015</td>
<td>460</td>
<td>41</td>
<td>0.089</td>
<td>1.56</td>
<td>0.06</td>
<td>6</td>
<td>(Henson et al., 2016)</td>
</tr>
<tr>
<td>CJ2</td>
<td>September 2015</td>
<td>460</td>
<td>61</td>
<td>0.133</td>
<td>2</td>
<td>0.071</td>
<td>7.1</td>
<td>(Henson et al., 2016)</td>
</tr>
<tr>
<td>FWC2</td>
<td>April 2016</td>
<td>460</td>
<td>403</td>
<td>0.876</td>
<td>2</td>
<td>1.044</td>
<td>104.4</td>
<td>This study</td>
</tr>
<tr>
<td>ARD2c</td>
<td>June 2016</td>
<td>460</td>
<td>7</td>
<td>0.015</td>
<td>2</td>
<td>0.008</td>
<td>0.8</td>
<td>This study</td>
</tr>
<tr>
<td>JLB2c</td>
<td>May 2016</td>
<td>460</td>
<td>103</td>
<td>0.224</td>
<td>2</td>
<td>0.127</td>
<td>12.7</td>
<td>This study</td>
</tr>
<tr>
<td>LKB2</td>
<td>July 2016</td>
<td>460</td>
<td>39</td>
<td>0.085</td>
<td>2</td>
<td>0.044</td>
<td>4.4</td>
<td>This study</td>
</tr>
<tr>
<td>Tbon3</td>
<td>July 2016</td>
<td>460</td>
<td>78</td>
<td>0.170</td>
<td>2</td>
<td>0.093</td>
<td>9.3</td>
<td>This study</td>
</tr>
<tr>
<td>CJ3</td>
<td>September 2016</td>
<td>460</td>
<td>69</td>
<td>0.15</td>
<td>2</td>
<td>0.081</td>
<td>8.1</td>
<td>This study</td>
</tr>
<tr>
<td>FW3</td>
<td>November 2016</td>
<td>460</td>
<td>27</td>
<td>0.059</td>
<td>2</td>
<td>0.03</td>
<td>3</td>
<td>This study</td>
</tr>
<tr>
<td>ARD3</td>
<td>December 2016</td>
<td>460</td>
<td>58</td>
<td>0.126</td>
<td>2</td>
<td>0.067</td>
<td>6.7</td>
<td>This study</td>
</tr>
<tr>
<td>JLB3</td>
<td>January 2016</td>
<td>460</td>
<td>146</td>
<td>0.317</td>
<td>2</td>
<td>0.191</td>
<td>19.1</td>
<td>This study</td>
</tr>
<tr>
<td>LKB3</td>
<td>February 2016</td>
<td>460</td>
<td>38</td>
<td>0.083</td>
<td>2</td>
<td>0.043</td>
<td>4.3</td>
<td>This study</td>
</tr>
</tbody>
</table>

*According to V = -ln(1-p)/X, where p = positive wells/inoculated wells, X = # cells inoculated per well (Button et al., 1993). % Cultivability = V x 100.
Our data also illuminates the relationship of many poorly understood taxa with coastal marine chemistry, specifically with regards to salinity. We found that the HIMB59-type clade OTUs associated with LSUCC isolates had >1% relative abundance at nearly every site (Fig. S4.7), while ASV-based relative abundances showed HIMB59 was most abundant at salinities > 15 (Fig. S4.8). For the two LSUCC associated OTUs belonging to the BAL58 clade, the relative abundance was highest at salinities between 1-7 (Fig. S4.7) with ASV-based relative abundances following a similar trend except for one type (type 5), which was most abundant at salinities < 2 (Fig. 4.3). The different relationships with OTU- and ASV-based relative abundances and salinity demonstrates the increased resolution provided by ASV analysis for a given taxon. Spikes in relative abundance were found for many LSUCC associated ASVs and OTUs, such as OM252, OM241, and unc. Porticococcaceae3, at the site JLB3 (Fig. S4.7 and S4.8). While the increase in relative abundance was not significantly correlated to any measured variables (Wilcoxon Rank Sum Test, P > 0.05), spikes in relative abundance could be connected to rises in the concentration of nutrients not measured in this study or the bloom or presence of another organism.

Many recent studies have shown the importance of interactions between co-occurring organisms (Biller et al., 2016; Aharonovich and Sher 2016; Amin et al., 2012; Amin et al., 2015; Segev et al., 2016); however, little is known about the potential for these interactions to occur between heterotrophs. Using a co-occurrence network inferred from ASV relative abundances, we found many strong correlations between heterotrophic bacteria, specifically between LSUCC isolates (Fig. 4.5). Twenty-five LSUCC isolate associated ASVs were found in the network correlation analysis (Fig. 4.5). Within the top 50 most connected ASVs, we found seven LSUCC associated ASVs of which two (ASV5512/OM252 clade, ASV7518/unc. Gammaproteobacteria) belonged to ASVs putatively novel at the Genus level. Of the numerous other correlations, twenty-two were between LSUCC associated ASVs. Specifically, we found that ASV5073 (HIMB11-type clade) and ASV5512 (OM252 clade) had the most co-correlations with 7 and 6, respectively (Fig. 4.5). Further, these two LSUCC-ASVs had the 4th (ASV5512) and 11th (ASV7518) most correlations to their node, respectively (Table S1). Our observations of the many LSUCC associated ASV co-occurrences suggests our isolates represent important organisms within the microbial communities in the nGOM. While the co-occurrences could be an artifact of relative abundances, they may indicate potential cross feeding or other interaction between two heterotrophs, a prediction testable with co-cultivation of the LSUCC isolates.
**4.4 DISCUSSION**

With 7820 wells inoculated over three years, this work represents the largest continual HTC effort to date (Connon and Giovannoni, 2002; Cho and Giovannoni, 2004; Page *et al.*, 2004; Stingl *et al.*, 2007, 2008; Song *et al.*, 2009; Yang *et al.*, 2016). The range of cultivability from these experiments was similar to or improved from the average and median values reported in HTC experiments using natural freshwater and seawater (Connon and Giovannoni, 2002; Page *et al.*, 2004; Brandon, 2006; Stingl *et al.*, 2007; Yang *et al.*, 2016), while vastly improving upon the < 1% associated with the “great plate count anomaly” and the more recently calculated value of 0.5% reported in Lloyd *et al.*, 2018 (Staley and Konopka, 1985; Lloyd *et al.*, 2018b). These improvements support the observation that the suites of JW and MWH artificial media are effective for cultivating important aquatic clades while providing a streamlined approach to characterizing those organisms. By combining amplicon sequencing with cultivation experiments, these experiments have provided the first isolates for groups such as SAR11 LD12 (as reported in Henson *et al.*, 2018), acIV subclades A and B, the OM241 clade, and MWH UniPo-type organisms. Moreover, while relative abundance was certainly related to isolation success for some taxa that are now well-represented in the culture collection, for many other groups it poorly predicted cultivation success. This discrepancy highlights the complexity of cultivation efforts even when the appropriate nutrient requirements are met.

As culture collections of environmentally relevant microbial clades continue to grow, examining interactions between microbes using co-cultivation will be essential to elucidating the interactions occurring between organisms (Bertrand *et al.*, 2015). Many previous co-culturing experiments have focused on phototrophic and heterotrophic interactions (e.g., (Amin *et al.*, 2012, 2015; Aharonovich and Sher, 2016; Biller *et al.*, 2016; Segev *et al.*, 2016). While...
understanding these interactions are important, potentially equally as important are interactions between heterotrophic bacteria (Solden et al., 2018, Fig. 4.5). Within the co-occurrence network, we found a highly significant connection between ASVs associated with LSUCC0261 (ASV7471, SAR11 subclade IIIa.1 group) and LSUCC0227 (ASV7241, OM43 group) (Fig. 4.5). LSUCC0261 and LSUCC0227 were both isolated using the same medium (JW2) and during the same cultivation experiment (CJ2, September 2015; Table S1), providing a streamlined opportunity to test the potential interactions occurring between two organisms known to occur at the same site together.

Despite being found in coastal studies and their presence in other HTC experiments together (Morris et al., 2006; Hugerth et al., 2015; Henson et al., 2016; Yang et al., 2016), specific interactions between these two clades are currently unknown. One potential explanation for the observed co-occurrence of these two heterotrophic isolates could be that it is an artifact of their relative abundances. If both organisms co-occur with a eukaryotic phototroph, not sequenced as part of this study, then the association between the two cultivars associated ASVs may be related to their presence with that organism rather than each other. An alternative hypothesis is that the two heterotrophs co-correlated because of a cross-feeding or synergistic interaction. Both clades are known to have undergone genome streamlining, an evolutionary process associated with auxotrophies and the dependency on cross-feeding interactions (Giovannoni, Tripp, et al., 2005; Lynch, 2006; Giovannoni et al., 2014). Therefore, these isolates make for an interesting case study to examine the potential interactions between heterotrophic bacteria. Future studies could use sequenced genomes to help predict auxotrophies generated during genome streamlining and use the different metabolic capabilities to direct experiments testing putative cross-feeding interactions.

The ability to predict the successful cultivation of LSUCC associated ASVs or the number of isolated organisms based on relative abundance was variable depending on the clade being analyzed. One explanation for this discrepancy is the viable but nonculturable state (VBNC), where bacteria decrease their metabolic activity and are no longer actively dividing (Oliver, 2005; Epstein, 2009, 2013; Lloyd et al., 2018b). VBNC state is caused by natural stressors, changes in ideal conditions, or perhaps placement in a cultured condition (Zengler et al., 2002; Oliver, 2005). These viable but non-active cells may prevent otherwise cultivable cells from growing in favorable conditions in wells and explain why organisms abundant in the environment (e.g., SAR11) are not consistently cultivated (Oliver, 2005; Epstein, 2013; Hug, 2018). One potential release from the VBNC state is the activation of scout cells (Buerger et al., 2012a). The central premise in the scout hypothesis is that dormant cells stochastically release active, scout cells, into the environment (Epstein, 2009; Buerger et al., 2012a). Under favorable conditions, the cell will trigger, “awaken,” other cells, while under adverse conditions the scout cell will die, and the cell population will remain dormant (Epstein, 2009; Buerger et al., 2012a). Therefore, it is plausible that the percent of VBNC cells in a population or the time it takes for an organism to release into a scout cell varies between clades. These fluctuations may help explain why some clades were rarely cultivated, while others were more successful cultivated and connected to relative abundance, despite providing the nutrient required for all cultivated organisms. Previously, cultivation of the SAR11 clade was found to improve with duration of experiment with 60-84% of the identified wells from week 20 and 24 belonging to SAR11, while none were found at four weeks (Song et al., 2009). With this considered, future experiments should increase the number of wells and the duration of the experiments to help improve cultivation success.
Another explanation may be phenotypic heterogeneity (Ackermann, 2015; Kell et al., 2015), where the metabolic or genomic differences in subpopulations, not detectable by standard practices, may prevent or allow the growth of a previously cultivable or uncultured taxon, respectively. For example, in the recent cultivation of two acl clade species, the acl cultivars had a phenotypic variation with different preferred reduced sulfur sources, a trait difficult to predict using genomic analyses (Kim et al., 2018). Variation of the phenotype in a population may explain why at one site the number of actual cultivars nearly matched the expected and were capable of growing in the medium provided, while at other sites more variation or no cultivation was observed.

However, the differences observed between actual versus expected may not be purely biological. It is important to note that at some sites (FWC, FWC2, JLB2c, and JLB3) we did not transfer all of the positive wells. For these sites, the number of cultured may not represent the true “actual” number, but instead be more of a conservative estimate of the ratio. While at most sites, all positive wells were transferred, some wells were never successfully transferred to polycarbonate flasks, DNA extracted, or PCR amplified. These issues may have prevented the successful identification of cultures present in the wells and negatively impact the expected versus actual ratio. Further, starting at TBON3, the medium was switched between JW-series to MWH-series. The suite of MWH artificial media were updated to include many vital osmolytes, reduced sulfur, and other compounds (Vila-Costa et al., 2006; Mou et al., 2007; Carini et al., 2012; Paul Carini, Campbell, et al., 2014; Stein, 2015; Repeta et al., 2016; Curson et al., 2017; Widner and Mulholland, 2017) (Table S1). These differences could have improved or decreased the chances of successful cultivation of an organism and, as a result, impacted the successful cultivation of that clade. Therefore, these ratios should represent a conservative view of the cultivation success of these experiments.

Despite the success of HTC experiments using both natural seawater (Page et al., 2004; Stingl et al., 2007, 2008; Song et al., 2009; Marshall and Morris, 2013; Hahnke et al., 2015; Sosa et al., 2015; Yang et al., 2016; Kim et al., 2018) and artificial aquatic media (Henson et al., 2016), the many unknown factors impacting cultivation may explain why some of the most abundant members of the coastal GOM communities continue to remain uncultivated (Button et al., 1993; Connon and Giovannoni, 2002; Epstein, 2013; Hug, 2018; Lloyd et al., 2018b). While nutrient additions (e.g., DMSP, glycine betaine) may be necessary for the growth and cultivation of organisms (Stingl et al., 2007; Carini et al., 2012; Tripp, 2013), nutrients may not be the sole factor impacting their cultivation (Epstein, 2009). For example, the recent cultivation of the acl clade by Kim et al., 2018 was facilitated by the addition of catalase to the medium. Based on genomic analyses, both genomes of the cultivars contained katG, yet in axenic conditions, the catalase was unable to lower H2O2 concentrations to levels facilitating growth (Kim et al., 2018). The cultivation of Psychrobacter sp. MSC33 was induced by a dose-dependent concentration of a short-chain signaling peptide, LQPEV. Without the peptide, MSC33 was uncultivated when using unamended artificial medium, suggesting the importance of considering microbial interactions in the cultivation of some organisms (Kaeberlein et al., 2002; Nichols et al., 2008). These examples show the complexity of microbial growth in vitro and highlight the importance of incorporating environmental conditions, genomic content, and in situ interactions when developing anexic cultivation conditions.
4.4.1 CONCLUSION

The cultivation of microorganisms remains an important step in understanding the ecological and physiological hypotheses derived from cultivation-independent techniques (Giovannoni and Stingl, 2007b). These experiments resulted in the cultivation of previously uncultivated lineages such as SAR11 LD12, acIv subclade A and B, OM241, and MWH UniPo-type and greatly increases the cultivated diversity of nGOM bacterioplankton and many well known and poorly understood clades. The connection between relative abundance and cultivation success was highly variable between clades, suggesting that the cultivation of some oligotrophic bacterioplankton may be reliant on mimicking both environmentally relevant conditions and other non-nutrient related dependencies such as isolation of active cells, growth time, or in situ syntrophic interactions. To overcome this, future HTC experiments must better incorporate cell activity, metabolic interdependencies, and an increased number of wells and duration of experiments to help improve cultivation of currently uncultured organisms.
CHAPTER 5.
CONCLUSIONS AND FUTURE DIRECTIONS

The broad goals of this dissertation were to isolate important bacterioplankton to help elucidate genomic, metabolic, and physiological differences that impact their abundance and distribution across different ecosystems and provide insights into the factors controlling cultivation of these organisms. To support this, we used a two-pronged approach of HTC and amplicon sequencing to provide context to the abundance and distribution of those organisms in the nGOM. We demonstrated that artificial media could be used to cultivate and characterize bacterioplankton apart of the microbial majority. Experiments using the first SAR11 LD12 representative Candidatus Fonsibacter ubiquis strain LSUCC0530 provided evidence for two hypotheses on the diversification of the subclade based on temperature and salinity. Further, we hypothesized that the transition of SAR11 LD12 into freshwater environments was related to the loss and gain of genes involved in the production and import of osmolytes. Finally, we highlighted the complexity involved in the cultivation of organisms despite providing the appropriate nutrients for cultivation by demonstrating the disconnect between relative abundance and the expected number of positive wells of an isolate. This dissertation underscores the importance of using cultivars to verify cultivation-independent derived hypotheses and observations, the ability for artificial media-based HTC experiments to isolate novel and abundant aquatic clades, and demonstrates the intricacies impacting cultivation once an organism is provided with the nutritional requirements necessary for growth.

Over the 17 experiments, many important and novel cultivars were isolated, such as LSUCC0530 from the globally abundant SAR11 LD12 clade. The continued effort to understand the physiology and metabolic capacities of these organisms will be essential for verifying their unique capacities to inhabit various environments in the nGOM but also help put them in context with other taxa. For Candidatus Fonsibacter ubiquis strain LSUCC0530, we hypothesized that the loss of genes involved in osmolyte uptake and changes in energy production prevent the SAR11 LD12 clade from inhabiting marine environments. With the isolation of LSUCC0530 and LSUCC0261 from the marine SAR11 subclade IIIa, we have the opportunity to test the underlying genetic differences in osmolyte production and uptake that allowed LD12 to transition to freshwater but also elucidate differences in the metabolic capacity that enable both subclades to cohabitate an environment. Experiments should incorporate both transcriptomic- and metabolomics-based analyses in various salinities to examine the differences in cellular responses and metabolite production of energetic pathways and important osmolytes, respectively, to changes in salinities. Validated with in situ based samples, these experiments can unveil how salinity transitions occur within a well-known clade, and the potential impact global changes in salinity related to climate change affect microbial community composition and their functional capacity for biogeochemical cycling.

While metagenomic recruitment of reads from various freshwater studies to the LSUCC0530 genome was high at some sites, sequences from inland lakes were poorly recruited. Specifically, recruitment was highest at sites associated with high water temperatures (e.g., Lake Gatun and Feitsui Reservoir). Therefore, an alternative hypothesis to salinity driven diversification of the LD12 clade is one based on temperature. Temperature optima experiments with LSUCC0530 supported the metagenomic-based observation and found a limited temperature range between 24°C and 30°C. However, because of the difficulty in inferring physiological traits such as temperature optima from genomic data, the isolation of more LD12 representatives from inland
bodies is needed. These cultivars would facilitate experiments on the different temperature ranges within LD12 ecotypes and allow the testing of other potential modes diversification such as metabolism (e.g., sulfur).

Numerous other cultivars in the collection provide exciting opportunities to study novel microorganisms. A current ongoing project will explore the first two cultivars from the acIV Actinobacteria clade. The acIV clade, like SAR11 LD12, is a globally abundant freshwater group that has been mainly reported in inland water studies (Debroas et al., 2009; Newton et al., 2011; Ghai et al., 2012; Garcia et al., 2018). While a lot of research and isolation efforts have focused on its sister clade acI (Ghai et al., 2012; Ghylin et al., 2014; Kang et al., 2017; Garcia et al., 2018; Neuenschwander et al., 2018), little to no information on the metabolic capacity or genomic content of acIV organisms is available. To date, both of our isolates have been sequenced using Oxford Nanopore technology, resulting in one circularized genome (LSUCC0897) and one single contig genome (LSUCC0889). Because of the error rate associated with MinION sequencing, the genomes will need to be corrected using Illumina short reads. Future work should explore if these coastal isolates represent novel ecotypes within the acIV clade and how they are physiologically and genomically different from organisms belonging to the acIV clade found in inland lakes and rivers. If the two isolates have similar salinity ranges to SAR11 LD12, as suggested by relative abundances, the examination of the potential genomic and metabolic similarity and differences between the two clades could provide a more general hypothesis on how coastal freshwater clades have adapted to fluctuating salinity.

Many other important contributions associated with LSUCC cultivars are still waiting to be further tested. For instance, LSUCC0245 represents only the second isolate within the HIMB59 clade that was previously thought to be a part of the SAR11 clade, but instead likely represents a novel streamlined Alphaproteobacterial lineage (Vergin et al., 2013; Viklund et al., 2013; Thrash et al., 2014; Martijn et al., 2018). Analysis of the genome of LSUCC0245 has discovered two proteorhodopsin genes, one of which contains a unique domain within the gene that may confer a functionality previously not known. Another discovery was found in the SAR subclade IIIa.1 (LSUCC0261) genome: the complete set of urease genes. If LSUCC0261 were capable of using urea as an organic nitrogen source, this would represent a new capability within the SAR11 clade that would impact models of nitrogen cycling in coastal oceans.

An important part of the future success of HTC experiments is the continued improvement of the methodology. As part of the current protocol, all positive wells are transferred manually to 125 mL polycarbonate flasks containing 50 mL of medium. This process can be laborious and without people assisting, limits the number of potential wells that can be transferred. Identification of the numerous transfers often led to experiments lasting three months before the identity of each organism was known. Modifications to this protocol should consider the use of ultracentrifugation and colony PCR to facilitate the identification of positive wells. To do this, after cyrostocks are made, a subset of the remaining volume could be placed into ultracentrifuge tubes and spun to pellet cells. An ultracentrifuge would be required because many organisms do not pellet using a standard centrifuge. Then, a colony PCR on the pelleted cells using the standard 27F, 1492R primers would be performed. These improvements would avoid using DNA extraction and allow the targeted approach of only transferring wells of organisms of interest. A different identification method that could be used is matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF) (Biswas and Rolain, 2013; Singhal et al., 2015). MALDI-TOF would provide a rapid, cost-effective method for identifying various isolates, while also avoiding the need for DNA extractions and PCR (Biswas and Rolain, 2013;
Efforts must continue to incorporate metabolic information garnered from cultivation-independent studies to isolate yet-to-be cultivated organisms. Single-cell genomic- and metagenomic-based inferences can elucidate potential metabolisms and auxotrophies, while techniques such as Nanoscale secondary ion mass spectrometry (NanoSIMS) and bioorthogonal noncanonical amino acid tagging (BONCAT) can track substrate incorporation and utilization (Behrens et al., 2008; Musat et al., 2012; Hatzenpichler and Connon, 2016). Together, this data can help facilitate the creation of media explicitly designed for an uncultivated organism of interest by allowing researchers to uncover metabolites necessary for growth. For instance, a high molecular weight organic matter enriched media was used with HTC to cultivate the organisms that helped elucidate the metabolic pathways and genes involved in the degradation of semi-labile dissolved organic matter, a process poorly understood (Sosa et al., 2015). Moreover, cell sorting using taxon-specific markers or targeting metabolically active cells offers another promising area to improve cultivation success. If the hypothesis that many cells are in viable but nonculturable states and waiting for their stochastic release as scout cells is correct (Oliver, 2005; Buerger et al., 2012a), then increasing the number of wells inoculated and the duration of experiments could increase the chance that a well contains an active cell. One potential method may employ two currently available technologies: fluorescence-activated cell sorting (FACS) and fluorescence in situ hybridization (FISH) (Yilmaz et al., 2010). By utilizing an in-solution fixation-free FISH method (Yilmaz et al., 2010), probes targeting active cells or specific taxa (e.g., OM1, SAR86) could be used to dramatically increase the number of wells and the chance of isolation by avoiding non-active cells or organisms not of interest.

This dissertation has the potential to serve as a blueprint for future projects using isolates to understand the underlying biology of important clades. Further, the protocol developed within can be used in an educational format to teach students the importance of cultivation in microbiology by providing hands-on experience with active research and allowing individuals to isolate, identify, and characterize bacteria (Bakshi et al., 2017). The numerous new cultivars provided by this work will be the legacy of this dissertation for years to come and will help researchers answer important fundamental questions in microbiology.
REFERENCES


Figure S2.1. 16S rRNA gene phylogeny of LSUCC isolates in the *Gammaproteobacteria*. Scale bar represents 0.01 changes per position. Values at internal nodes indicate Shimodaira-Hasegawa “like” test values.
Figure S2.1. 16S rRNA gene phylogeny of LSUCC isolates in the *Alphaproteobacteria*. Scale bar represents 0.01 changes per position. Values at internal nodes indicate Shimodaira-Hasegawa “like” test values.
Figure S2.3. 16S rRNA gene phylogeny of LSUCC isolates in the Betaproteobacteria. Scale bar represents 0.01 changes per position. Values at internal nodes indicate Shimodaira-Hasegawa “like” test values.
Figure S2.4. Cell counts of the 82 cultures, in duplicate, during their initial transfers from plates to flasks. Concentration is indicated on the y-axes, time on the x-axes, in days. Note time scales for each culture are independent. LSUCC numbers are indicated in the grey boxes above each plot. Dots indicate sample points.
**Figure S3.1.** Flow cytometry plot of strain LSUCC0530 during the initial experiment by which it was isolated. Side scatter is plotted vs. green fluorescence.
Figure S3.2. Phylogenetic inference of the Alphaproteobacteria, including LSUCC0530 and other reference SAR11 sequences, using 16S rRNA genes. Values at nodes indicate Shimodaira-Hasegawa like values from FastTree2.
**Figure S4.3.** Box plots of intragenic spacer distributions for subclade IIIa and IIIb (LD12) with Wilcoxon rank-sum results indicated.

**Figure S3.4.** Metagenomic recruitment to the LSUCC0530 genome scaffold using sequences from Feitsui Reservoir (A) and Lake Gatun (B). Recruitment is plotted according to percent identity of read hit to the genome, which is depicted linearly. HVR2 is visible as the region of very poor recruitment on the left side of each plot, highlighted with the red bars.
Figure S3.5. Phylogenetic tree of *aceA* (isocitrate dehydrogenase) sequences. Scale bar indicates 0.1 changes per position. Values at nodes indicate Shimodaira-Hasegawa like values from FastTree2.
Figure S3.6. Phylogenetic tree of malate synthase sequences. Scale bar indicates 0.1 changes per position. Values at nodes indicate Shimodaira-Hasegawa like values from FastTree2.
Figure S3.7. Gene neighborhood of malate synthase (red, center) in the LSUCC0530 genome and those of a subset of other SAR11 genomes. Conserved region is boxed.
**Figure S3.8.** Results of growth experiments testing salinity range (A) and temperature range (B). Results are plotted as cell concentration by hour, and are centered at zero as the beginning of logarithmic growth to allow for ease of comparison across multiple independent experiments. Replicates are plotted as separate colors. Replicates showing aberrant growth (e.g., black line in A, 2.9) were not plotted in Figure 4.
Figure S3.9. Relative abundance of the LD12 OTU in coastal samples from the northern Gulf of Mexico. Bar graph indicates OTU7 relative abundance compared with all other SAR11 OTUs, with sites ordered according to increasing salinity. Inset depicts the LD12 OTU7 only data with a linear regression, equation and R$^2$ value included, and 95% confidence intervals shaded.
Figure S3.10. Relative abundance of LD12 genome microclusters at 85, 90, 92, 98, and 100% percent identity, for comparison with Figure 5 (95% identity). RPKM values are listed by site, with data aggregated for all genomes in microclusters A-C, according to the key. Colors indicate broad environmental categories.
Supplemental Text for Chapter 3

*Genome assembly and quality assessment*

I. SPAdes assembly
Step 1. Subset the reads.
$ seqtk sample -s100 530_R1.fastq 1000000 > 530_subset1.r1.fastq
$ seqtk sample -s100 530_R2.fastq 1000000 > 530_subset1.r2.fastq

Adapter file for Trimmomatic (Bolger *et al.*, 2014):
> i7
GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGGTTCGAATCTCGTTATGATGCGGCCGTCTTCTTG
> i5
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTCTCCTAGGATATGTAGATCTCGGTGTGTTCGTATCATT

Step 2. Trim pretty aggressively to remove adapters and bases < q20 at ends and over sliding window

$ java -jar ~/bin/trimmomatic-0.30.jar PE -threads 12 -phred33 -trimlog trimmed/530_trimlog.log untrimmed/530_subset1.r1.fastq untrimmed/530_subset1.r2.fastq trimmed/530_subset1.r1.fastq trimmed/530_subset1.r1.unpaired.fastq trimmed/530_subset1.r2.fastq trimmed/530_subset1.r2.unpaired.fastq ILLUMINA_CLIP:/scratch/jcthrash-test/530_adapters.fa:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:13:20 MINLEN:40

Step 3. Assemble w/ SPAdes (Bankevich *et al.*, 2012):

$ python ~/src/SPAdes-3.10.1-Linux/bin/spades.py --threads 12 --memory 46 --cov-cutoff auto --pe1-1 trimmed/530_subset1.r1.fastq --pe1-2 trimmed/530_subset1.r2.fastq --s1 trimmed/530_subset1.r1.unpaired.fastq --s2 trimmed/530_subset1.r2.unpaired.fastq
This resulted in a single contig with overlapping ends, plus three very small sequences. The first was 83 bp and contained only repeats of “CCCTAA,” the second was 78 bp of “C,” the third, 78 bp of “A.”

II. Quality assessment
The single scaffold from the original SPAdes assembly was evaluated with Reapr (Hunt et al., 2013), manually broken based on the results (two FCD errors), and the gaps were extended using SSPACE (Boetzer et al., 2011) and all the Illumina HiSeq reads (Step 1). Overlaps were determined with megablast and the scaffolds were manually rejoined with overlaps removed. Another run of Reapr (Step 2) identified only one FCD error. The break and SSPACE extension process was repeated, and overlaps determined again with megablast. To evaluate the quality of the overlaps at the outer ends of the scaffold, we artificially broke the scaffold in two, reorganized the ends so they were now internal, and removed overlaps. Reapr did not identify any additional errors at this step, but verified the previous single FCD error (Step 3). The break-extension process was repeated one more time (Step 4) and then again using the MiSeq reads, with a final re-evaluation with reapr identified only the single FCD error in the same location (Step 5). To assess the validity of the error, we repeated the quality assessment of the final scaffold using a separate algorithm, Pilon (Walker et al., 2014) after mapping all the HiSeq reads to the scaffold using BWA (Li and Durbin, 2009) (Step 6). This resulted in no errors.

Corroborating information was obtained using CheckM (Parks et al., 2015) (Step 7) and by evaluating the GC skew (Brown et al., 2015) (Step 8). Detailed commands and outputs are as follows, according to the steps identified in this synopsis.

Step 1.
Reapr
$ reapr facheck 530_spades_sc.fasta
no error
$ reapr perfectmap 530_spades_sc.fasta 530_subset1.r1.fastq 530_subset1.r2.fastq 300 perfect
$ reapr smaltmap 530_spades_sc.fasta 530_subset1.r1.fastq 530_subset1.r2.fastq 530_bwa.bam
$ reapr pipeline 530_spades_sc.fasta 530_bwa.bam 530_reapr_output perfect

This gave similar output, with only two FCD errors and two small sections removed (and replaced with Ns). The scaffold was manually broken in these locations, with new fasta headers added, resulting in three scaffolds → 04.break.broken_assembly_broken.fa.

SSPACE
$ perl SSPACE_Standard_v3.0.m.pl -l library.txt -s 04.break.broken_assembly_broken.fa -x 1 -v 1 -T 16 -b 530_sspacev2.1
The library.txt file specified all the original HiSeq reads. All three scaffolds were extended. Overlaps were identified with megablast by comparing each pair of scaffolds.

\$ cat 530_sspacev2.1.final.scaffolds.fasta | fastaToTab | grep scaffold1 | tabToFasta > scaffold1.fasta
\$ cat 530_sspacev2.1.final.scaffolds.fasta | fastaToTab | grep scaffold2 | tabToFasta > scaffold2.fasta
\$ cat 530_sspacev2.1.final.scaffolds.fasta | fastaToTab | grep scaffold3 | tabToFasta > scaffold3.fasta

\$ makeblastdb -dbtype nucl -in scaffold1.fasta -out scaffold1db -parse_seqids -hash_index
\$ makeblastdb -dbtype nucl -in scaffold2.fasta -out scaffold2db -parse_seqids -hash_index

\$ blastn -query scaffold2.fasta -db scaffold1db -out scaffold2v1blast
\$ blastn -query scaffold3.fasta -db scaffold1db -out scaffold3v1blast
\$ blastn -query scaffold3.fasta -db scaffold2db -out scaffold3v2blast

The order starting with the beginning of the longest scaffold is 1—>3—>2, with a loop back to 1. I removed overlaps and named the file 530_sspace_scaffold1.fasta.original.

To smooth out missing lines:
\$ cat 530_sspace_scaffold1.fasta.original | fastaToTab | tabToFasta > 530_sspace_scaffold1.fasta

Step 2.
\$ reapr facheck 530_sspace_scaffold1.fasta
no error
\$ reapr perfectmap 530_sspace_scaffold1.fasta 530_R1.fastq 530_R2.fastq 300 perfect
\$ reapr smaltmap -n 16 530_sspace_scaffold1.fasta 530_R1.fastq 530_R2.fastq
530_sspace_scaffold1.bam
\$ reapr pipeline 530_sspace_scaffold1.fasta 530_sspace_scaffold1.bam
530_sspace_scaffold1_output perfect

Just one RCD error.

Step 3.
Check to see if that is the same and only error if the 2—>1 gap is closed. I’ve redone the scaffold assembly such that the scaffolds are 3—>2 —>1: 530_sspace_scaffold1_ro.fasta.original

\$ cat 530_sspace_scaffold1_ro.fasta.original | fastaToTab | tabToFasta >
530_sspace_scaffold1_ro.fasta
The 04.break.broken_assembly_bin.fa files for both 530_sspace_scaffold1 and 530_sspace_scaffold1_ro are 100% identical across their alignments (NCBI megablast), but slightly different sizes. Importantly this also means that the scaffold end join is not an issue.

Step 4.
Now try to eliminate the FCD error with a final break and extension of 530_sspace_scaffold1.fasta.

Take the 04.break.broken_assembly.fa file from reapr_sspacev2.1a, manually remove the Ns, break the scaffold and rename the headers —> 04.break.broken_assembly_broken.fa.original

$ cat 04.break.broken_assembly_broken.fa.original | fastaToTab | tabToFasta > 530_sspace_scaffold1_broken.fasta

run SSPACE with the same library.txt file as above.

$ mv sspace.o630499 library.txt q_sspace 530_sspace_scaffold1_broken.fasta 530_sspacev2.2/
$ cd 530_sspacev2.2/

Although the scaffolds were not joined, the larger one was extended.

$ cat 530_sspacev2.2.final.scaffolds.fasta | fastaToTab | grep scaffold1 | tabToFasta > scaffold1.fasta
$ cat 530_sspacev2.2.final.scaffolds.fasta | fastaToTab | grep scaffold2 | tabToFasta > scaffold2.fasta
$ makeblastdb -dbtype nucl -in scaffold1.fasta -out scaffold1db -parse_seqids -hash_index
$ blastn -query scaffold2.fasta -db scaffold1db -out scaffold2v1blast

megablast showed that the beginning of scaffold two has a 616 bp overlap with the end of scaffold 1. Also, the end of scaffold 2 has a 462 bp overlap with the beginning of scaffold 1, as it should.
Join the two and run reaper another time —> 530_sspace_scaffold2.fasta.original
$ cat 530_sspace_scaffold2.fasta.original | fastaToTab | tabToFasta > 530_sspace_scaffold2.fasta
$ cp 530_sspace_scaffold2.fasta ../
$ cd ../
$ reapr facheck 530_sspace_scaffold2.fasta
  no error
$ reapr perfectmap 530_sspace_scaffold2.fasta 530_R1.fastq 530_R2.fastq 300 perfect
$ reapr smaltmap -n 16 530_sspace_scaffold2.fasta 530_R1.fastq 530_R2.fastq
  530_sspace_scaffold2.bam
$ reapr pipeline 530_sspace_scaffold2.fasta 530_sspace_scaffold2.bam
  530_sspace_scaffold2_output perfect

This is still showing an FCD error in the same place. Attempt to close with MiSeq reads.

Step 5.
Manually break 04.break.broken_assembly.fa —> 04.break.broken_assembly_broken.fa.original.

$ cat 04.break.broken_assembly_broken.fa.original | fastaToTab | tabToFasta > 04.break.broken_assembly_broken.fa

Run SSPACE as above but with MiSeq reads.

This only extended one scaffold. Run megablast to identify overlaps:
$ cat 530_sspacev3.final.scaffolds.fasta | fastaToTab | grep scaffold1 | tabToFasta > scaffold1.fasta
$ cat 530_sspacev3.final.scaffolds.fasta | fastaToTab | grep scaffold2 | tabToFasta > scaffold2.fasta
$ makeblastdb -dbtype nucl -in scaffold1.fasta -out scaffold1db -parse_seqids -hash_index
$ blastn -query scaffold2.fasta -db scaffold1db -out scaffold2v1blast

There was the typical overlap from the scaffold ends, 462 bp in this case. There was only a 236 bp overlap between the end of scaffold 1 and the beginning of scaffold 2. I joined these manually for another reapr check —> 530_sspacev3_scaffold2.fasta.original
$ cat 530_sspacev3_scaffold2.fasta.original | fastaToTab | tabToFasta > 530_sspacev3_scaffold2.fasta

Run reapr with the miseq sequences and the hiseq sequences and compare FCD reports.
$ reapr facheck 530_sspacev3_scaffold2.fasta
no error
$ reapr perfectmap 530_sspacev3_scaffold2.fasta
/project/jcthrash/genome_fastqs/LSUCC0530_S1_L001_R1_001.fastq
/project/jcthrash/genome_fastqs/LSUCC0530_S1_L001_R2_001.fastq 440 perfect
$ reapr smaltmap -n 16 530_sspacev3_scaffold2.fasta
/project/jcthrash/genome_fastqs/LSUCC0530_S1_L001_R1_001.fastq
/project/jcthrash/genome_fastqs/LSUCC0530_S1_L001_R2_001.fastq
530_sspacev3_scaffold2.bam
$ reapr pipeline 530_sspacev3_scaffold2.fasta 530_sspacev3_scaffold2.bam
530_sspacev3_scaffold2_output perfect

This actually yielded no FCD errors, but many errors regarding low fragment coverage within the contig. Try running the hiseq reads across this new assembly.

$ reapr perfectmap 530_sspacev3_scaffold2.fasta /project/jcthrash/genome_fastqs/530_R1.fastq
/project/jcthrash/genome_fastqs/530_R2.fastq 300 perfect
$ reapr smaltmap -n 16 530_sspacev3_scaffold2.fasta
/project/jcthrash/genome_fastqs/530_R1.fastq /project/jcthrash/genome_fastqs/530_R2.fastq
530_sspacev3_scaffold2_hs.bam
$ reapr pipeline 530_sspacev3_scaffold2.fasta 530_sspacev3_scaffold2_hs.bam
530_sspacev3_scaffold2_hs_output perfect

This yielded the same erroneous region it did for all the checks (using all the reads). Time to cross-evaluate with a different algorithm.

Step 6.
Manually remove the scaffold overlaps at the ends of 530_sspacev3_scaffold2.fasta and change the header —> LSUCC0530_final_assembly.fasta

bwa
$ cd hiseq_bwa/
$ bwa index LSUCC0530_final_assembly.fasta
$ bwa aln -n 0 -t 16 LSUCC0530_final_assembly.fasta
/project/jcthrash/genome_fastqs/530_R1.fastq > 530_R1.sai
$ bwa aln -n 0 -t 16 LSUCC0530_final_assembly.fasta
/project/jcthrash/genome_fastqs/530_R2.fastq > 530_R2.sai
$ bwa sampe LSUCC0530_final_assembly.fasta 530_R1.sai 530_R2.sai /
/project/jcthrash/genome_fastqs/530_R1.fastq /project/jcthrash/genome_fastqs/530_R2.fastq >
530_R1R2.sam
$ samtools faidx LSUCC0530_final_assembly.fasta
$ samtools import LSUCC0530_final_assembly.fasta.fai 530_R1R2.sam 530_R1R2.bam
$ samtools sort 530_R1R2.bam 530_R1R2.sorted
$ samtools index 530_R1R2.sorted.bam

On my local machine

pi

$ java -Xmx16G -jar ../../../Applications/pilon-1.22.jar --genome
LSUCC0530_final_assembly.fasta --frags 530_R1R2.sorted.bam

Pilon version 1.22 Wed Mar 15 16:38:30 2017 -0400
 Genomic: LSUCC0530_final_assembly.fasta
 Fixing snps, indels, gaps, local
 Input genome size: 1160202
 Scanning BAMs
 530_R1R2.sorted.bam: 14264790 reads, 0 filtered, 9778615 mapped, 9537494 proper, 30284
 stray, FR 100% 311+/−109, max 639
 Processing LSUCC0530_final_assembly:1-1160202
 frags 530_R1R2.sorted.bam: coverage 1018
 Total Reads: 9989452, Coverage: 1018, minDepth: 102
 Confirmed 1160121 of 1160202 bases (99.99%)
 Corrected 0 snps; 0 ambiguous bases; corrected 0 small insertions totaling 0 bases, 0 small
 deletions totaling 0 bases
 # Attempting to fix local continuity breaks
 LSUCC0530_final_assembly:1-1160202 log:
 Finished processing LSUCC0530_final_assembly:1-1160202
 Writing updated LSUCC0530_final_assembly_pilon to pilon.fasta
 Mean frags coverage: 1018
 Mean total coverage: 1018

Step 7.
$ checkm lineage_wf -x .fasta -t 16 -f 530_sspacev3_scaffold2 530_sspacev3_in/
530_sspacev3_out/

CheckM results still predict the scaffold to be 100% complete with 0% contamination, and it
clades in the same space with the subclade III taxa. Output:
<table>
<thead>
<tr>
<th>Bin Id</th>
<th>Marker lineage</th>
<th># genomes</th>
<th># markers</th>
<th># marker sets</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5+</th>
<th>Completeness</th>
<th>Contamination</th>
<th>Strain heterogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td>530_sspacev3_scaffold2</td>
<td>k_Bacteria (UID2495)</td>
<td>2993</td>
<td>139</td>
<td>83</td>
<td>0</td>
<td>139</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Step 8.

$ gc_skew -f LSUCC0530_final_assembly.fasta

This final assembly was submitted to IMG (Markowitz et al., 2014) for genome annotation. It is publically available with IMG Taxon ID 2728369501, and at GenBank under accession number CP024034.
Figure S4.1. 16S rRNA gene phylogeny of LSUCC cultivars in the Phylum Actinobacteria. Scale bar represents 0.01 changes per position. Values at internal nodes indicate Shimodaira-Hasegawa “like” test values.
Figure S4.2. 16S rRNA gene phylogeny of LSUCC cultivars in the Class Alphaproteobacteria. Scale bar represents 0.01 changes per position. Values at internal nodes indicate Shimodaira-Hasegawa “like” test values.
Figure S4.3. 16S rRNA gene phylogeny of LSUCC cultivars in the Phylum Bacteroidetes. Scale bar represents 0.01 changes per position. Values at internal nodes indicate Shimodaira-Hasegawa “like” test values.
Figure S3.4. 16S rRNA gene phylogeny of LSUCC cultivars in the Class Betaproteobacteria. Scale bar represents 0.01 changes per position. Values at internal nodes indicate Shimodaira-Hasegawa “like” test values.
Figure S4.5. 16S rRNA gene phylogeny of LSUCC cultivars in the Class *Gammaproteobacteria*. Scale bar represents 0.01 changes per position. Values at internal nodes indicate Shimodaira-Hasegawa “like” test values.
Figure S4.6. Rank abundance plot of the top 50 most abundant (A) OTUs and (B) ASVs across all sites. The boxes indicate the interquartile range (IQR) of the data, with vertical lines indicating the upper and lower extremes according to 1.5 x IQR. Horizontal lines within each box indicate the median. The underlying data points are each individual OTU’s sample relative abundances. The shade of the dot represents the salinity, while the color of the box is the Phylum of the OTU. OTUs with cultured representatives from the LSU culture collection are labeled with one LSUCC cultivar.
Figure S4.7. Relative abundance of cultivated OTUs according to site salinity. The color of the line represents the different ASVs classified within the LSUCC group. Non-linear regressions have been added for reference.

Figure S4.8. Relative abundance of cultivated ASVs according to site salinity. The color of the line represents the different ASVs classified within the LSUCC group. Non-linear regressions have been added for reference. Triangles represent sites where isolates were cultivated.
Figure S4.9. Relative abundance of ASVs classified to SAR11 clade according to site Salinity. ASVs with an associated LSUCC group have been labeled. Non-linear regressions have been added for reference.
APPENDIX D
PERMISSION TO REPRODUCE CHAPTER 2

Publisher policy:

ASM also grants the author the right to republish discrete portions of his/her article in any other publication (including print, CD-ROM, and other electronic formats) of which he or she is author or editor, provided that proper credit is given to the original ASM publication. An ASM author also retains the right to reuse the full article in his/her dissertation or thesis. “Proper credit” means either the copyright lines shown on the top of the first page of the PDF version, or “Copyright © American Society for Microbiology, [insert journal name, volume number, year, page numbers and DOI]” of the HTML version. For technical questions about using Rightslink, please contact Customer Support via phone at (877) 622-5543 (toll free) or (978) 777-9929, or e-mail Rightslink customer care at customercare@copyright.com.
APPENDIX E
PERMISSION TO REPRODUCE CHAPTER 3

Publisher policy:

The author of articles published by SpringerNature do not usually need to seek permission for re-use of their material as long as the journal is credited with initial publication. Ownership of copyright in original research articles remains with the Author, and provided that, when reproducing the contribution or extracts from it or from the Supplementary Information, the Author acknowledges first and reference publication in the Journal, the Author retains the following non-exclusive rights:
To reproduce the contribution in whole or in part in any printed volume (book or thesis) of which they are the author(s).
The author and any academic institution where they work at the time may reproduce the contribution for the purpose of course teaching.
To reuse figures or tables created by the Author and contained in the Contribution in oral presentations and other works created by them.
To post a copy of the contribution as accepted for publication after peer review (in locked Word processing file, of a PDF version thereof) on the Author's own website, or the Author's institutional repository, or the Author's funding body's archive, six months after publication of the printed or online edition of the Journal, provided that they also link to the contribution on the publisher's website.
The above use of the term 'Contribution' refers to the author's own version, not the final version as published in the Journal.
VITA

Michael Winslow Henson was born in Toledo, Ohio on October 10th, 1988. There, Michael attended St. Joseph’s (K), Highland Elementary (1-5), McCord Middle School (6-8), and St. John’s Jesuit (9-12). In high school, Michael worked at Highland Meadows Golf Course as a lifeguard and graduated from St. John’s in May of 2007.

Michael attended the Miami University in Oxford, Ohio from 2007-2011, where he majored in Microbiology. While at Miami University, Michael worked in the lab of Marcia Lee and at the local restaurant Skipper’s Pub. In the summer of 2010, Michael worked as a student intern at the Environmental Protection Agency in Cincinnati, Ohio with Dr. Jorge Santo Domingo. After graduation in May 2011 with a B.A. in Microbiology, he moved to Cincinnati, Ohio to work full time as a student contractor in Dr. Jorge Santo Domingo lab. While at the EPA, Michael worked on the project investigating the meta-’omics of hospital and drinking water microbiomes.

In 2012, Michael began a master’s program in Biological Sciences at Central Michigan University in the lab of Dr. Deric Learman, where he studied the strain-level genomic and physiological variation in four Microbacterium spp. chromate reducers. While at CMU, he was invited to take part in an international collaboration between Jiangxi Normal University and CMU as a student representative. The project focused on the impacts of anthropogenic inputs on microbial communities in the sediment of Poyang Lake, the largest freshwater lake in Nanchang, China compared to sediment from EPA Areas of Concern in the Great Lakes region. Michael graduate in the summer of 2014 with a Master’s of Science.

In 2014, Michael began his PhD at Louisiana State University in the Department of Biological Sciences in the lab of Dr. J. Cameron Thrash. He will defend his dissertation on November 19, 2018, and will be joining Dr. Thrash as a PostDoc at the University of Southern California.