Host-Microbe Relations:
A Phylogenomics-Driven Bioinformatic Approach
to the Characterization of Microbial DNA from Heterogeneous Sequence Data

Timothy Patrick Driscoll

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Genetics, Bioinformatics, and Computational Biology

Joseph J Gillespie
David R Bevan
Madhav V Marathe
T M Murali

May 1st, 2013
Blacksburg, Virginia

Keywords: phylogenomics, genome-mining, host-microbe interactions, genomics, bioinformatics, symbiosis, bacteria, lateral gene transfer

Host-Microbe Relations:
A Phylogenomics-Driven Bioinformatic Approach
to the Characterization of Microbial DNA from Heterogeneous Sequence Data

Timothy Patrick Driscoll

ABSTRACT

Plants and animals are characterized by intimate, enduring, often indispensable, and always complex associations with microbes. Therefore, it should come as no surprise that when the genome of a eukaryote is sequenced, a medley of bacterial sequences are produced as well. These sequences can be highly informative about the interactions between the eukaryote and its bacterial cohorts; unfortunately, they often comprise a vanishingly small constituent within a heterogeneous mixture of microbial and host sequences. Genomic analyses typically avoid the bacterial sequences in order to obtain a genome sequence for the host. Metagenomic analysis typically avoid the host sequences in order to analyze community composition and functional diversity of the bacterial component. This dissertation describes the development of a novel approach at the intersection of genomics and metagenomics, aimed at the extraction and characterization of bacterial sequences from heterogeneous sequence data using phylogenomic and bioinformatic tools.

To achieve this objective, three interoperable workflows were constructed as modular computational pipelines, with built-in checkpoints for periodic interpretation and refinement. The MetaMiner workflow uses 16S small subunit rDNA analysis to enable the systematic discovery and classification of bacteria associated with a host genome sequencing project. Using this information, the ReadMiner workflow comprehensively extracts, assembles, and characterizes sequences that belong to a target microbe. Finally, AssemblySifter examines the genes and scaffolds of the eukaryotic genome for sequences associated with the target microbe. The combined information from these three workflows is used to systemically characterize a bacterial target of interest, including robust estimation of its phylogeny, assessment of its signature profile, and determination of its relationship to the associated eukaryote.

This dissertation presents the development of the described methodology and its application to three eukaryotic genome projects. In the first study, the genomic sequences of a single, known endosymbiont was extracted from the genome sequencing data of its host. In the second study, a highly divergent endosymbiont was characterized from the assembled genome of its host. In the third study, genome sequences from a novel bacterium were extracted from both the raw sequencing data and assembled genome of a eukaryote that contained significant amounts of sequence from multiple competing bacteria. Taken together, these results demonstrate the usefulness of the described approach in singularly disparate situations, and strongly argue for a sophisticated, multifaceted, supervised approach to the characterization of host-associated microbes and their interactions.

Dedication

I dedicate this dissertation to my wife Charley, the most winsome and enchanting soul I've ever known, whose love, support, and unflagging encouragement never fail to humble and inspire me: Charley, may all of your days be sunny, may your ears be filled with love and laughter, and may your eyes some day alight upon a moose. To my daughter Samantha, who will always be the best of me, and who is, without exception or bias, the most beautiful girl in the world. To my as-yet-unborn daughter, who is the promise of tomorrow. To my mom and dad, a constant source of love and support, who never once in twenty-two years asked me when I was going to finish. To my brothers: Mark for lighting the way, Joe for exemplifying success through determination, and Jeff for showing me how to tread a little lighter. To my sisters: Lynn for being the big sister I always needed, and Sarah for jumping in with both feet. You are all an inspiration in your own way.

I also thank the many people who have guided me along the path to this day. My high school biology teacher, Mr. Lane (a.k.a. Mr. Greenjeans), who made biology interesting and fun and without whom I would have been an English major. John Boyer, my master's thesis advisor, for teaching me that biochemistry isn't so scary after all, and for instilling in me a sense of enthusiasm for scientific exploration that remains as strong today as ever. Eric Martz, for helping me to see that science and creativity are necessarily intertwined. And the countless teachers, professors, collaborators, and colleagues who have contributed their time, knowledge, and guidance with no expectation of reward. They are a constant source of inspiration.

Finally, I am wholeheartedly grateful for the many personal relationships that sustained me during my time in Blacksburg. The Reverend Doctor Bryan Lewis for his enduring friendship, the Beauty Loop (of course!), Stillers-Pats, and his utter Dudeness. Andrew Warren, Chris Lasher, Brian Gehrt, and Brian Yohn for their camaraderie and companionship through all manner of interesting adventures. To the Blue Ridge Mountains and the special places of solace therein, especially Pandapas Pond, Rock Castle Gorge, Douthat, Kelly's Knob, and Sinking Creek. Last but not least, I dedicate this dissertation to a famous man, a man whom I never met, but whose words and vision have been an inspiration nevertheless. "Somewhere, something incredible is waiting to be known." - Carl Sagan.

Acknowledgements

I acknowledge and appreciate the support of many people who helped enable the research presented in this dissertation, and to the GBCB program at Virginia Tech for the opportunity to carry out my graduate studies under their auspices.

I am indebted to all of my committee members for their participation, advice, and patience throughout the development of this dissertation.

Joe Gillespie, for his encouragement, mentorship, and collaborative spirit, for his infectious enthusiasm for biology in all its messy splendor, for always supporting my development as a scientist, and for his personal friendship.

T. M. Murali and Madhav Marathe for their patience, helpful comments, and willingness to see this process through to its conclusion.

David Bevan for his support and willingness to help carry this work across the final few months.

Additionally, I acknowledge Bruno Sobral and the members of the Molecular Genetics Lab and Cyberinfrastructure Division (past and present), for their support, technical expertise, and many thought-provoking conversations. The IT department at VBI, for their prompt assistance on all things related to the high-performance computing aspects of this project. Eric Nordberg for contributing his phylogenomics expertise and software. Nicolas Lartillot, author of PhyloBayes, for his assistance in running the MPI version of his software.

Finally, I am especially grateful to Dennie Munson, heart of the GBCB program, for her tireless dedication to helping her students meet the many administrative, academic, and personal challenges of being a graduate student. In recognition of her invaluable role in my graduate career, I will use a phrase that only Dennie could elicit from me: "Go Yankees!"

Attribution

Several colleagues aided in the writing and research behind one of the chapters presented as part of this dissertation. A brief description of their contributions is included here.

Chapter 4: Bacterial DNA sifted from the *Trichoplax adhaerens* (Animalia: Placozoa) genome project reveals a putative rickettsial endosymbiont.

Chapter 4 was published in the journal Genome Biology and Evolution with the following co-authors:

Joseph J. Gillespie Ph.D. (Virginia Bioinformatics Institute, Virginia Tech) was an equal contributor to this manuscript, wrote much of the paper, and provided expert analysis of the rickettsial biology.

Eric K. Nordberg (Virginia Bioinformatics Institute, Virginia Tech) was a co-author on this paper and provided the FastTree-based phylogenomic analyses.

Abdu F. Azad (Department of Microbiology and Immunology, University of Maryland School of Medicine) was a co-author on this paper and principal investigator for one of the grants supporting this research.

Bruno W. Sobral (Nestlé Institute of Health Sciences SA, Campus EPFL, Quartier de L'innovation, Batîment G, 1015 Lausanne, Switzerland) was a co-author on this paper and principal investigator for one of the grants supporting this research.

Table of Contents

List of Figures

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

List of Tables

CHAPTER 1

CHAPTER 2

CHAPTER 4

CHAPTER 5

List of Abbreviations

ANOVA: Analysis of variance
BLAST: Basic local alignment search tool
DNA: Deoxyribonucleic acid
Gb: Gigabase (one billion bases)
JGI: Joint Genome Institute
Kb: Kilobase (one thousand bases)
LGT: Lateral gene transfer
Mb: Megabase (one million bases)
MCL: Markov clustering algorithm
MGE: Mobile genetic element
NCBI: National Center for Biotechnology Information
OG: Ortholog group
PATRIC: Pathosystems Resource Integration Center
REIS: *Rickettsia* endosymbiont of *Ixodes scapularis*
RETA: *Rickettsia* endosymbiont of *Trichoplax adhaerens*
rDNA: Ribosomal DNA
RNA: Ribonucleic acid
SSU: Small subunit
T4P: Type IV pili
T4SS: Type IV secretion system
WGS: Whole-genome shotgun
XTAB: *Xenopus tropicalis* associated betaproteobacterium

Host-Microbe Relations:
A Phylogenomics-Driven Bioinformatic Approach
to the Characterization of Microbial DNA from Heterogeneous Sequence Data

CHAPTER 1. Introduction.

MOTIVATION

Plants and animals are characterized by intimate, enduring, and often indispensable associations with microbes. The human body, for example, contains more bacterial cells (10^{14}) than human (10^{13}) by an order of magnitude (**Wooley et al. 2010**), and those bacteria are largely ineradicable residents. The default condition of eukaryotic life includes microbes, and any study of the first must account for the second. Therefore, it should come as no surprise that when the genome of a eukaryote is sequenced, a medley of bacterial sequences are sequenced as well despite best efforts to avoid them. The relationship of the bacterial sequences to the eukaryote is often left undefined and, since the goal is typically the eukaryotic genome, ultimately discarded as contamination. This dissertation was motivated by the observation that bacterial sequences associated with the genomes of eukaryotes are prevalent, persistent, and can be highly informative about the interactions between a host and its bacterial cohorts (**table 1**). Recognizing that these sequences can be difficult to extract and characterize, the methodology developed in this research includes a versatile combination of phylogenomic and bioinformatic tools aimed at the extraction and characterization of bacterial sequences from heterogeneous sequence data, and applies them to three singularly different systems.

BACKGROUND

Prokaryotes comprise as many as 100 million distinct species and represent the largest proportion of individual organisms on Earth **(Sleator et al. 2008**). They have been found in all manner of environments from high in the atmosphere (**Maki et al. 2010**) to deep under the ocean (**Zinger et al. 2011**). Bacterial and archaeal communities are everywhere: the soil **(Vogel et al. 2009**), even when contaminated with toxins (**Hoffmann et al. 2003**); the outflow from terrestrial hot springs (**Deckert et al. 1998**); cooling towers and hot water systems (**Wéry et al. 2008**); glacial ice **(Simon et al. 2009**); the reproductive (**White et al. 2011**) and digestive **(Maynard et al. 2012**) tracts of animals. Bacteria live within the cells of other organisms **(Celli 2006**; **Chan et al. 2010**; **Chong & Celli 2010**; **Dandekar 2012**), and within organisms that themselves live within other organisms (**Gottlieb et al. 2012**; **McCutcheon & Dohlen 2011**). The genomes of these mostly uncultured prokaryotes encode a largely unexplored repository of metabolic capabilities and novel functions - information that can lead to new industrial applications for secondary metabolic pathways (**Anderson & Dawes 1990**; **Valdés et al. 2008**; **Singh 2009**), a better understanding of bacterial evolution and community ecology **(Biddle et al. 2008; DeLong et al. 2006**), and new insights into host-microbial interactions **(Augustin et al. 2012; Bright & Bulgheresi 2010**; **Dethlefsen et al. 2007**).

Direct molecular sequencing has so far provided close to ten thousand complete bacterial (**Gillespie et al. 2011**) and archaeal (**Pagani et al. 2012**) genomes, mainly from cultivable species, and nearly two hundred complete eukaryotic genomes. A rapidly growing variety of metagenomic techniques have revealed the existence of many thousands of additional microbial

species, and enabled powerful comparisons of the ecology and metabolic profiles of microbial communities without the need to cultivate pure cultures (**Sharpton et al. 2011**; **Mitra & Stark 2011**; **Liu et al. 2011**). Genomics and metagenomics are highly complementary fields of study. Genomics offers insights into genome structure and function across multiple species, the evolution of individual genes, and the development of new species. It is predicated on the availability of complete genome sequences, which provide a comprehensive view of each organism (functional repertoire, gene order and density, regulatory regions, etc.). Metagenomics is the study of DNA sequences taken directly from the environment. These sequences are generally short (20-700 nucleotides) and derived from a mixture of organisms; as a result they can rarely be assembled into contigs exceeding 5 Kb in length, and the reconstruction of whole genomes is generally not achievable (**Wooley et al. 2010**). Consequently, metagenomic studies are typically aimed at 1) community composition analysis, where one or more marker genes are used to identify and quantify the microbes present in a sample; and 2) functional metagenomics, where environmental DNA sequences are screened for particular functional activities (**Kunin et al. 2008**). Unlike genomics, metagenomics provides essential information about microbes in the context of their communities and habitats: their interactions with other organisms and their adaptation to environmental conditions.

The complete community of microbes that inhabit a particular environment is called a *microbiome*. Some of the best-studied microbiomes are those that exist in close association with plants (**Mendes et al. 2011**; **Krome et al. 2009**) and animals (**Gevers et al. 2012**; **White et al. 2011**; **Nelson et al. 2010**; **Ley et al. 2008**) and play vital roles in the development and health of their hosts (**Sekirov & Finlay 2006**; **Round & Mazmanian 2009**). In addition to harboring diverse microbiomes, many eukaryotes also entertain specialized associations with individual microbial pathogens or *endosymbionts*. These encounters are dynamic and remarkably diverse, potentially involving interactions between several microbes (**Venturi & Silva 2012**), multiple hosts (**Benson et al. 2004**), and environmental reservoirs or breeding grounds (**Molmeret et al. 2005**). Opportunistic infections occur when bacteria are able to exploit a temporarily vacant niche in a vulnerable eukaryote (**Oliver et al. 2005**). Facultative interactions cover a wide range of situations where a microbe can live freely but has developed specialized mechanisms to inhabit a frequently encountered host (**Chong & Celli 2010**). Finally, obligate symbioses involve bacteria that are unable to live independently from their host; they are often vertically inherited, live intracellularly or in specialized host tissues (*bacteriosomes*), and in certain cases induce co-dependence in their host as well (**Bright & Bulgheresi 2010**). Symbioses like these, especially intracellular symbioses, can have a dramatic impact on the genomes of both host (**Gladyshev et al. 2008**; **Acuña et al. 2012**; **Aikawa et al. 2009**; **Woolfit et al. 2009**) and microbe (**Nikoh et al. 2011**; **Heinz et al. 2012**; **Zientz et al. 2004**).

Through a combination of genomic and metagenomic studies, it has become clear that plants, animals, and fungi exist in close associations with complex and dynamic microbes and microbial communities. In general, those associations are riven in order to study either the host or the microbial component in relative isolation. There are sound practical reasons for this dissolution. Eukaryotic DNA is typically excluded from metagenomic analyses, either by careful selection of the sampling site or computational removal of sequences after sequencing, due to the enormous size, complexity, and low gene densities of eukaryotic genomes (**Kunin et al. 2008**). Microbial DNA in a eukaryote genome sequencing project is similarly either avoided or computationally removed (**Hellsten et al. 2010**; **Srivastava et al. 2008**); it can confound assembly and analysis

of the eukaryotic genome, especially given the relative dearth of reference genomes for cross-comparison and validation of the assembly. Nevertheless, heterogeneous sequence data containing a mixture of host and microbe sequences is prevalent. Indeed, a growing number of targeted studies have identified bacterial endosymbionts within the genome sequencing data of their hosts (**Salzberg et al. 2005**; **Gillespie et al. 2012**; **Chapman et al. 2010**), and a comprehensive analysis of these data is leading to important insights into host-microbe associations. The primary goal of this dissertation is to enable the rapid identification and efficient characterization of host-associated bacterial DNA from heterogeneous sequence data, using a combination of metagenomic and genomic tools.

The traditional eukaryotic genome sequencing project is aimed at obtaining the complete genome sequence for a single target organism, including the location and arrangement of its genes, regulatory sequences, etc., and typically includes six steps: 1) obtain as pure a DNA sample as possible; 2) randomly shear the DNA into fragments (*reads*) of 20-7000 nucleotides (depending on the sequencing technology being used) and sequence them; 3) computationally assemble overlapping reads into contiguous stretches of sequence (*contigs*); 4) remove contigs that match to known contaminant genomes (microbes, organelles, etc.); 5) arrange non-overlapping contigs into *scaffolds* using associated positional information; and 6) close the genome by manually re-sequencing the gaps between adjacent contigs. These last two steps are often highly laborious and time-consuming, especially in the absence of a closely-related reference genome; as a result they are often excluded, leading to published genome sequences in a variety of states of completion. In addition, the computational removal of non-eukaryote sequences (step 4) is often incomplete, leading to published genomes that contain microbial sequences (see **table 1**). Such microbial sequences can arise in several ways. The DNA sample preparation could have been contaminated, or cloning vector fragments used during sequencing could have been overlooked. Extracellular bacteria that occupy the same niche as the eukaryote could have been caught up in the sampling phase, or the eukaryote might harbor a closely-associated endosymbiont. Finally, they could represent lateral gene transfers (LGTs): stretches of DNA from a bacterial source that have inserted into the contiguous genome of the eukaryote (**Gladyshev et al. 2008**). While an argument could be made that LGTs are not strictly bacterial, it may be difficult for sequence similarity or genome assembly algorithms to make that distinction.

A central premise of this dissertation is that bacterial sequences in a eukaryotic genome sequencing represent a valuable potential resource of information about host-microbe interactions past and present, including rare or poorly characterized microbes which may be resistant to cultivation in the laboratory. The deep analysis of these data has remained a largely unmet challenge at the intersection of metagenomics and genomics, a challenge that this research aims to address.

ADVANTAGES

The approach presented here takes advantage of (though it is not restricted to) existing genome sequence projects; not only does this save the expense of having to generate new data, it can be applied immediately to a substantial number of available data sets. The workflows developed for this dissertation are also not limited to eukaryote genome sequencing projects; with some customization they can be applied to a range of heterogeneous sequence data.

This methodology currently incorporates parallel computing techniques for similarity matching and phylogenetic estimation, and is further scalable to even larger systems and concurrent analyses. Target microbes can be identified quickly by virtue of the integrated metagenomic techniques, and characterized in detail using phylogenomic inference and ortholog group construction. The speed of this methodology allows a broad evaluation of host-microbe associations, while the detailed characterization can help resolve the lifestyle of host-associated microbes, allowing for more focus in follow-up studies.

The multifaceted approach of this research allows it to be applied across a range of sub-optimal systems. With various combination of the programs developed in this dissertation, researchers can analyze genomes without available raw sequencing data, bacteria that are significantly divergent from known genomes, and different microbial targets within a single heterogeneous data set. In addition, the novel similarity measures introduced in this dissertation enable the rapid identification of candidate microbe-host LGTs from among tens of thousands of host coding sequences.

LIMITATIONS

Lateral gene transfer between prokaryotes is widespread and an important mechanism of genome diversification (**Kloesges et al. 2011**). It is anticipated that the methods presented here, many of which rely on sequence similarity, may perform poorly when extracting regions of a target genome dominated by bacteria-bacteria LGTs. This is not likely to interfere with phylogenomic inferences, which rely on conserved orthologs, but may complicate target genome assembly and functional characterization. Likewise, targets that are highly divergent from known bacteria will be more difficult to characterize in full, although the combination of workflows will at least provide some assessment.

The proportion of target bacterial cells in the initial sample extraction may be quite low, considering the sampling method is not likely to have enriched for bacterial cells and traditional sample preparation steps are geared toward minimizing bacterial sequences before sequencing. As a result, coverage of the target genome in the trace reads is likely to be poor, hampering reconstruction of the full genome. This situation is exacerbated in analyses of the assemble host genome, since the preparation of host reads for assembly generally includes computational removal of bacterial-like sequences. For many of the same reasons, evaluating existing eukaryotic genomes is likely to be limited to either closely-associated microbes (e.g., intracellular symbionts) or true contaminants.

PROJECT OVERVIEW

The following dissertation will demonstrate how bacterial sequences extracted from heterogeneous mixtures of eukaryote and prokaryote sequence data can be used to characterize novel host-associated bacteria and host-microbe interactions, and will present an integrated, multifarious methodology for the efficient application of this approach to various systems.

Chapter 2 presents the methodology as a collection of interoperable computational pipelines broadly grouped into three overarching workflows. MetaMiner uses metagenomics-inspired techniques to describe the community composition of a heterogeneous sequence data set and identify target microbes of interest. ReadMiner uses sequence similarity, genome assembly, and phylogenomic techniques to extract, assemble, and characterize target genomic sequences from the heterogeneous data set. AssemblySifter evaluates an assembled eukaryotic genome for mis-

incorporated microbial sequences and candidate bacteria-host LGTs. A specific investigation may require some or all of these workflows in different combinations, depending on the conditions of the study; consequently, these workflows are designed as flexible collections rather than strictly linear processes.

Chapters 3-5 present the application of the methodology described in Chapter 2 to three very different data sets, chosen to demonstrate the general applicability of the described approach under various real-world scenarios. Chapter 3 focuses on the characterization of a known obligate intracellular symbiont, *Rickettsia* endosymbiont of *Ixodes scapularis* (REIS), using the sequencing data of its host, *I. scapularis*, and provides some benchmarks on the effectiveness of MetaMiner and ReadMiner. Chapter 4 describes the identification and classification of a novel, highly divergent rickettsial endosymbiont associated with the primitive marine animal *Trichoplax adhaerens*, using only sequences sifted from the *T. adhaerens* assembly. In this case, previous independent studies provided the initial clues about the existence of an endosymbiont in this animal (**Schierwater 2005**). Chapter 5 describes the characterization of a novel betaproteobacterial microbe from the sequence data of the western clawed frog, *Xenopus* (Silurana) *tropicalis*; this study used all three workflows in the evaluation of a eukaryote with a wholly unknown bacterial component.

The methods employed in this research demonstrate the benefits of applying a combination of metagenomic and genomic analyses toward a deeper examination of genome sequencing data, especially vis-à-vis the characterization of rare or intractable bacterial symbionts. The following chapters and overall conclusions argue strongly that worthwhile and biologically relevant knowledge can be extracted from heterogeneous sequence data, leading to a deeper understanding of host-microbe associations.

LITERATURE CITED

Acuña R et al. 2012. Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. Proceedings of the National Academy of Sciences. 109:4197–4202. doi: 10.1073/pnas.1121190109.

Aikawa T et al. 2009. Longicorn beetle that vectors pinewood nematode carries many *Wolbachia* genes on an autosome. Proc Biol Sci. 276:3791–3798. doi: 10.1098/rspb.2009.1022.

Anderson AJ, Dawes EA. 1990. Occurrence, metabolism, metabolic role, and industrial uses of bacterial polyhydroxyalkanoates. Microbiol. Rev. 54:450–472.

Augustin R, Fraune S, Franzenburg S, Bosch TCG. 2012. Where simplicity meets complexity: hydra, a model for host-microbe interactions. Adv. Exp. Med. Biol. 710:71–81. doi: 10.1007/978-1-4419-5638-5_8.

Benson MJ, Gawronski JD, Eveleigh DE, Benson DR. 2004. Intracellular symbionts and other bacteria associated with deer ticks (*Ixodes scapularis*) from Nantucket and Wellfleet, Cape Cod, Massachusetts. Appl Environ Microbiol. 70:616–620. doi: 10.1128/AEM.70.1.616-620.2004.

Biddle JF, Fitz-Gibbon S, Schuster SC, Brenchley JE, House CH. 2008. Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment. Proceedings of the National Academy of Sciences. 105:10583–10588. doi: 10.1073/pnas.0709942105.

Bright M, Bulgheresi S. 2010. A complex journey: transmission of microbial symbionts. Nat Rev Micro. 8:218–230. doi: 10.1038/nrmicro2262.

Celli J. 2006. Surviving inside a macrophage: the many ways of *Brucella*. Res Microbiol. 157:93–98. doi: 10.1016/j.resmic.2005.10.002.

Chan YG-Y, Riley SP, Martinez JJ. 2010. Adherence to and invasion of host cells by Spotted Fever group *Rickettsia* species. Frontiers in Microbiology. 1:139. doi: 10.3389/fmicb.2010.00139.

Chapman JA et al. 2010. The dynamic genome of *Hydra*. Nature. 464:592–596. doi: 10.1038/nature08830.

Chong A, Celli J. 2010. The *Francisella* intracellular life cycle: toward molecular mechanisms of intracellular survival and proliferation. Frontiers in Microbiology. 1:138. doi: 10.3389/fmicb.2010.00138.

Dandekar T. 2012. *Salmonella enterica*: a surprisingly well-adapted intracellular lifestyle. 1–11. doi: 10.3389/fmicb.2012.00164/abstract.

Deckert G et al. 1998. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. Nature. 392:353–358. doi: 10.1038/32831.

DeLong EF et al. 2006. Community genomics among stratified microbial assemblages in the ocean's interior. Science. 311:496–503. doi: 10.1126/science.1120250.

Dethlefsen L, Mcfall-Ngai M, Relman DA. 2007. An ecological and evolutionary perspective on human-microbe mutualism and disease. Nature. 449:811–818. doi: 10.1038/nature06245.

Gevers D, Pop M, Schloss PD, Huttenhower C. 2012. Bioinformatics for the Human Microbiome Project Eisen, JA, editor. PLoS Comput Biol. 8:e1002779. doi: 10.1371/journal.pcbi.1002779.g002.

Gillespie JJ et al. 2012. A *Rickettsia* genome overrun by mobile genetic elements provides insight into the acquisition of genes characteristic of an obligate intracellular lifestyle. J Bacteriol. 194:376–394. doi: 10.1128/JB.06244-11.

Gillespie JJ et al. 2011. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. Infect Immun. 79:4286–4298. doi: 10.1128/IAI.00207-11.

Gladyshev EA, Meselson M, Arkhipova IR. 2008. Massive horizontal gene transfer in bdelloid rotifers. Science. 320:1210–1213. doi: 10.1126/science.1156407.

Gottlieb Y et al. 2012. A novel bacterial symbiont in the nematode *Spirocerca lupi*. BMC Microbiol. 12:1–1. doi: 10.1186/1471-2180-12-133.

Heinz E et al. 2012. The Genome of the Obligate Intracellular Parasite *Trachipleistophora*

hominis: New Insights into Microsporidian Genome Dynamics and Reductive Evolution Johnson, PJ, editor. PLoS Pathogens. 8:e1002979. doi: 10.1371/journal.ppat.1002979.g008.

Hellsten U et al. 2010. The genome of the Western clawed frog *Xenopus tropicalis*. Science. 328:633–636. doi: 10.1126/science.1183670.

Hoffmann D, Kleinsteuber S, Müller RH, Babel W. 2003. A transposon encoding the complete 2, 4-dichlorophenoxyacetic acid degradation pathway in the alkalitolerant strain *Delftia acidovorans* P4a. Microbiology (Reading, Engl). 149:2545–2556. doi: 10.1099/mic.0.26260-0.

Kloesges T, Popa O, Martin W, Dagan T. 2011. Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. Molecular Biology and Evolution. 28:1057–1074. doi: 10.1093/molbev/msq297.

Krome K, Rosenberg K, Bonkowski M, Scheu S. 2009. Grazing of protozoa on rhizosphere bacteria alters growth and reproduction of *Arabidopsis thaliana*. Soil Biology and Biochemistry. 41:1866–1873. doi: 10.1016/j.soilbio.2009.06.008.

Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P. 2008. A bioinformatician's guide to metagenomics. Microbiol Mol Biol Rev. 72:557.

Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI. 2008. Worlds within worlds: evolution of the vertebrate gut microbiota. Nat Rev Micro. 6:776–788. doi: 10.1038/nrmicro1978.

Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M. 2011. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. BMC Genomics. 12 Suppl 2:S4. doi: 10.1186/1471-2164-12-S2-S4.

Maki T et al. 2010. Phylogenetic analysis of atmospheric halotolerant bacterial communities at high altitude in an Asian dust (KOSA) arrival region, Suzu City. Sci. Total Environ. 408:4556–4562. doi: 10.1016/j.scitotenv.2010.04.002.

Maynard CL, Elson CO, Hatton RD, Weaver CT. 2012. Reciprocal interactions of the intestinal microbiota and immune system. Nature. 489:231–241. doi: 10.1038/nature11551.

McCutcheon JP, Dohlen von CD. 2011. An Interdependent Metabolic Patchwork in the Nested Symbiosis of Mealybugs. Curr Biol. 21:1366–1372. doi: 10.1016/j.cub.2011.06.051.

Mendes R et al. 2011. Deciphering the rhizosphere microbiome for disease-suppressive bacteria. Science. 332:1097–1100. doi: 10.1126/science.1203980.

Mitra S, Stark M. 2011. Analysis of 16S rRNA environmental sequences using MEGAN. BMC Genomics.

Molmeret M, Horn M, Wagner M, Santic M, Abu Kwaik Y. 2005. Amoebae as training grounds for intracellular bacterial pathogens. Appl Environ Microbiol. 71:20–28. doi:

10.1128/AEM.71.1.20-28.2005.

Nelson K et al. 2010. A catalog of reference genomes from the human microbiome. Science (New York, NY). 328:994–999.

Nikoh N, Hosokawa T, Oshima K, Hattori M, Fukatsu T. 2011. Reductive evolution of bacterial genome in insect gut environment. Genome Biol Evol. 3:702–714. doi: 10.1093/gbe/evr064.

Oliver JW, Stapenhorst D, Warraich I, Griswold JA. 2005. *Ochrobactrum anthropi* and *Delftia acidovorans* to bacteremia in a patient with a gunshot wound. Infectious Diseases in Clinical Practice. 13:78–81.

Pagani I et al. 2012. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucl. Acids Res. 40:D571–9. doi: 10.1093/nar/gkr1100.

Round JL, Mazmanian SK. 2009. The gut microbiota shapes intestinal immune responses during health and disease. Nat Rev Immunol. 9:313–323. doi: 10.1038/nri2515.

Salzberg SL et al. 2005. Serendipitous discovery of *Wolbachia* genomes in multiple *Drosophila* species. Genome Biol. 6:R23. doi: 10.1186/gb-2005-6-3-r23.

Salzberg SL, Puiu D, Sommer DD, Nene V, Lee NH. 2009. Genome sequence of the *Wolbachia* endosymbiont of *Culex quinquefasciatus* JHB. J Bacteriol. 191:1725. doi: 10.1128/JB.01731-08.

Schierwater B. 2005. My favorite animal, *Trichoplax adhaerens*. Bioessays. 27:1294–1302.

Sekirov I, Finlay BB. 2006. Human and microbe: united we stand. Nature Medicine, July 1 doi: 10.1038/nm0706-736.

Sharpton TJ et al. 2011. PhylOTU: A High-Throughput Procedure Quantifies Microbial Community Diversity and Resolves Novel Taxa from Metagenomic Data Béjà, O, editor. PLoS Comput Biol. 7:e1001061. doi: 10.1371/journal.pcbi.1001061.t001.

Simon C, Herath J, Rockstroh S, Daniel R. 2009. Rapid identification of genes encoding DNA polymerases by function-based screening of metagenomic libraries derived from glacial ice. Appl Environ Microbiol. 75:2964–2968. doi: 10.1128/AEM.02644-08.

Singh BK. 2009. Organophosphorus-degrading bacteria: ecology and industrial applications. Nat Rev Micro. 7:156–164. doi: 10.1038/nrmicro2050.

Sleator RD, Shortall C, Hill C. 2008. Metagenomics. Lett. Appl. Microbiol. 47:361–366. doi: 10.1111/j.1472-765X.2008.02444.x.

Srivastava M et al. 2008. The *Trichoplax* genome and the nature of placozoans. Nature. 454:955–960. doi: 10.1038/nature07191.

Valdés J et al. 2008. *Acidithiobacillus ferrooxidans* metabolism: from genome sequence to industrial applications. BMC Genomics. 9:597. doi: 10.1186/1471-2164-9-597.

Venturi V, Silva DPD. 2012. Incoming pathogens team up with harmless 'resident' bacteria. Trends Microbiol. doi: 10.1016/j.tim.2012.02.003.

Vogel T et al. 2009. TerraGenome: a consortium for the sequencing of a soil metagenome. Nat Rev Micro. 7:252.

Wéry N et al. 2008. Dynamics of *Legionella* spp. and bacterial populations during the proliferation of *L. pneumophila* in a cooling tower facility. Appl Environ Microbiol. 74:3030–3037. doi: 10.1128/AEM.02760-07.

White BA, Creedon DJ, Nelson KE, Wilson BA. 2011. The vaginal microbiome in health and disease. Trends Endocrinol. Metab. 22:389–393. doi: 10.1016/j.tem.2011.06.001.

Wooley JC, Godzik A, Friedberg I. 2010. A primer on metagenomics. PLoS Comput Biol. 6:e1000667.

Woolfit M, Iturbe-Ormaetxe I, McGraw EA, O'Neill SL. 2009. An ancient horizontal gene transfer between mosquito and the endosymbiotic bacterium *Wolbachia pipientis*. Molecular Biology and Evolution. 26:367–374. doi: 10.1093/molbev/msn253.

Zientz E, Dandekar T, Gross R. 2004. Metabolic interdependence of obligate intracellular bacteria and their insect hosts. Microbiol Mol Biol Rev. 68:745.

Zinger L et al. 2011. Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. PLoS ONE. 6:e24570. doi: 10.1371/journal.pone.0024570.

TABLES

Table 1. Examples of host-associated bacteria identified from within eukaryotic genome sequence data, including published reports and suspected from preliminary investigations (*Unk*). Micro-eukaryotes (protists, amoebae, etc.) are not included here. REIS: *Rickettsia* endosymbiont of *Ixodes scapularis*; RETA: Rickettsiales endosymbiont of *Trichoplax adhaerens*; XTAB: *Xenopus tropicalis* associated betaproteobacterium.

Group	Eukaryote	Associated Bacteria	Reference
Arthropod (Insect)	*Drosophila ananassae* (fruit fly)	*Wolbachia* endosymbiont of *D. ananassae*	(**Salzberg et al. 2005**)
Arthropod (Insect)	*Drosophila simulans*	*Wolbachia* endosymbiont of *D. simulans*	(**Salzberg et al. 2005**)
Arthropod (Insect)	*Drosophila willistoni*	*Wolbachia* endosymbiont of *D. willistoni*	(**Salzberg et al. 2005**)
Arthropod (Insect)	*Culex quinquefasciatus* (mosquito)	*Wolbachia* endosymbiont of *C. quinquefasciatus*	(**Salzberg et al. 2009**)
Animal (Cnidaria)	*Hydra magnipapillata*	*Curvibacter* putative endosymbiont of *H. magnipapillata*	(**Chapman et al. 2010**)
Arthropod (Insect)	*Ixodes scapularis* (deer tick)	REIS	(**Gillespie et al. 2012**)
Animal (Placazoa)	*Trichoplax adhaerens*	RETA	This study
Animal (Amphibian)	*Xenopus tropicalis* (Western clawed frog)	XTAB	This study
Animal (Porifera)	*Amphimedon queenslandica* (sponge)	*Unk*	
Arthropod (Insect)	*Nasonia longicornis* (parasitoid wasp)	*Unk*	
Animal (Cnidaria)	*Nematostella vectensis* (sea anemone)	*Unk*	
Arthropod (Insect)	*Solenopsis invictus* (fire ant)	*Unk*	
Animal (Nematode)	*Heterodera glycines* (soybean cyst nematode)	*Unk*	
Plant (Angiosperm)	*Populus* sp. (poplar)	*Unk*	
Arthropod (Crustacean)	*Daphnia* sp.	*Unk*	
Animal (Trematode)	*Clonorchis sinensis* (liver fluke)	*Unk*	

CHAPTER 2. A supervised, multifarious approach to the systemic characterization of bacterial DNA from heterogeneous sequence data.

ABSTRACT

The primary objective of this study was to design and implement a flexible bioinformatic methodology for characterizing bacterial DNA from heterogeneous sequence data (e.g., eukaryotic genome sequencing projects, metagenomic data sets, etc.). To achieve this objective, three interoperable workflows were constructed as modular computational pipelines, with built-in "checkpoints" to allow periodic *in situ* interpretation and manual refinement. The MetaMiner workflow determines the general taxonomic distribution and coverage of the commingled constituents, and guides the selection of taxa to be characterized. The ReadMiner and AssemblySifter workflows extract sequences for the identified target taxa from either the raw sequences (reads) or the assembled genome, respectively, of the genome sequencing project. The combined information from these three workflows is used to systemically characterize the bacterial target of interest, including robust estimation of its phylogeny, assessment of its signature profile, and determination of its relationship to the associated eukaryote. These workflows were applied to three eukaryotic systems with known or suspected bacterial components, the specific results of which are presented in the subsequent chapters. Taken together, these studies support the validity of this supervised multifarious approach for the systematic identification and characterization of eukaryote-associated bacteria from heterogeneous sequence data.

INTRODUCTION

Advances in high-throughput DNA sequencing technologies are continuing to enable the generation of genomic data at unprecedented levels. As of April 2013, The Genomes Online Database (GOLD) included sequences for 183 eukaryotic and 186 archaeal genomes (**Pagani et al. 2012**), while the number of sequenced bacterial genomes is rapidly approaching ten thousand (**Gillespie et al. 2011**) and the number of viral genomes is already well in excess of three-hundred thousand (**Pickett et al. 2012**). Complementing these genome sequencing projects is the field of metagenomics, where DNA is sequenced directly from environmental samples to gain insights into the composition of microbial communities. Through a combination of genomic and metagenomic studies, it is now clear that most eukaryotes are intimately associated with a complex and dynamic microbial community (its microbiome) that is vital to its development and maintenance of health **(Ley et al. 2008; Sekirov & Finlay 2006; Round & Mazmanian 2009; Mendes et al. 2011**). As a result, much attention has turned toward understanding the interplay between eukaryotic hosts and their associated microbes through the comprehensive analysis of sequence data.

Metagenomics derives from classical microbial genetics but does not require the cultivation of pure microbial cultures, thereby theoretically providing access to the genomes of uncultivable species. Metagenomic analyses typically center around three overlapping goals: 1) community composition analysis, where one or more marker genes are used to identify and quantify the microbes present in a sample; 2) functional metagenomics, where environmental DNA sequences are screened for particular functional activities; and 3) shotgun metagenomics, where entire microbial genomes are recovered from environmental DNA sequences (**Kunin et al. 2008**). These data invariably contain a heterogeneous mixture of sequences from different origins,

including prokaryotes but also viruses, bacteriophages, organelle genomes, and microbial eukaryotes. Microbial eukaryotes in particular are typically excluded from metagenomic analyses, either by careful selection of the sampling site or computational removal of DNA after sequencing, due to the enormous size, complexity, and low gene densities of their genomes. Similarly, microbiome samples contain mixtures of cells from both the microbial community plus its metazoan host, and sequences from the host are typically avoided.

A slightly different situation exists in eukaryote genome sequencing projects. When a eukaryote genome is sequenced, microbial DNA sequences (particularly bacterial in origin) are often generated as a byproduct and typically discarded without further analysis. Such bacterial sequences can arise in several ways. They can be accidental, the result of laboratory contamination of the DNA sample preparation or the failure to fully remove cloning vector or adapter sequences used during sequencing. They can be incidental, originating from extracellular bacteria that occupy the same niche as the eukaryote. They can be symbiotic in origin, arising from bacteria that are directly associated with the eukaryote either as extracellular or obligate intracellular symbionts. Finally, bacterial sequences found among eukaryotic genome sequences can represent bacteria-host LGTs (**Gladyshev et al. 2008**). These classifications are not always mutually independent; for example, LGTs in a eukaryotic genome may originate from a closely associated symbiont **(Hotopp et al. 2007; Nikoh et al. 2008; Ros & Hurst 2009**). In addition, it may be difficult to make a clear distinction between a facultative symbiont and a microbe that was merely coincident with the eukaryote at the time of sampling, especially in novel systems.

With the possible exception of accidental contamination, bacterial sequences in a eukaryotic genome sequencing project can provide clues about the biological relationships between eukaryotes and their associated bacterial species, particularly rare or poorly characterized species which may be resistant to cultivation in the laboratory. These sequences represent a valuable potential resource of information about intimate host-microbe interactions past and present, but the deep analysis of these data has remained an unmet challenge at the intersection of metagenomics and genomics. To address this need, the current study was undertaken to design and implement a flexible, extensible methodology for identifying, extracting, and characterizing bacterial sequences from heterogeneous sequence data.

METHODS

The workflows described in this chapter present a comprehensive description of the methodology developed in this dissertation. These workflows are flexible by design: the selection of specific combinations of relevant workflow elements allows the underlying strategy to be applied across a broad set of biological systems. Subsequent chapters demonstrate the application of the methodology to three different systems: 1) a host with a known, sequenced endosymbiont (Chapter 3); 2) a host with a novel, highly divergent symbiont (Chapter 4); and 3) a host with a suspected bacterium of unknown association (Chapter 5).

The term **target** (or **target microbe**) is used throughout this dissertation to refer to the individual bacterial constituent of a eukaryote sequencing project that is the subject of the current study. Similarly, the term **host** is used to refer specifically to the eukaryotic constituent of the eukaryote sequencing project. Host is also commonly used in biology to refer to an organism that harbors a parasite, symbiont, or pathogen; despite this connotation, no *a priori* assumptions are made here concerning the ecological relationship between the eukaryote and any of its associated microbial sequences (unless otherwise supported by data from other studies).

Data Generation

Download and decontamination of sequence data. Raw genome sequencing data (hereafter, **reads**) from a eukaryote sequencing project of interest were downloaded from the Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi) at the National Center for Biotechnology Information (NCBI), and partially compiled into fasta and qual files. When exact cloning vector sequences were available, reads were screened using `cross_match` as described in the `PHRAP` user manual (**Green 1996**). Bases identified as vector by `cross_match` were masked but not removed in this step. When exact cloning vector information was unavailable, reads were screened using `blastn` (**Camacho et al. 2009**) against the NCBI UniVec database of common cloning vector sequences (ftp://ftp.ncbi.nih.gov/pub/UniVec/). Contamination of read sequences by cloning vector is generally characterized by exact or near-exact sequence matches near the termini of reads; therefore, stringent `blastn` parameters (-q -5 -G 3 -E 3 -F "m D" -e 700 -Y $1.75x10^{12}$) and a scoring rubric similar to NCBI's VecScreen protocol (http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen_docs.html) were used to identify and mask likely vector sequences.

Reads that had more than 20% of their bases masked were subsequently discarded, and the remaining sequences trimmed from the masked bases to the closest terminus. Reads shorter than 100 bases after trimming were discarded outright. After vector removal, the remaining reads were fully compiled into a single fastq file and verified for correct syntax.

Sequencing read validation. After download, compilation, and decontamination, reads were assessed for initial quality using the `fastqc` program from the Babraham Bioinformatics group (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Low quality bases (Phred score < 25) were trimmed from the ends of reads using `fastx_trimmer`, part of the `fastx` package (**Hannon 2010**). Trimmed reads with a mean Phred score less than 25 or a total length less than 100 bases were discarded, and the remaining reads assessed again for quality using `fastqc`. Multiple rounds of trimming and/or decontamination were employed when necessary to ensure a final set of sequencing reads of acceptable quality.

Download of assembled genome data. Unless otherwise noted, annotated proteins and assembly scaffolds for all eukaryotic organisms were obtained from the NCBI Protein and Nucleotide databases, respectively. Small subunit (SSU) ribosomal DNA (rDNA) sequences for eukaryotes were obtained from the NCBI Nucleotide database. Unless otherwise noted, all genome sequences and annotated proteins for sequenced bacteria were obtained from the Pathosystems Resource Integration Center (PATRIC) (**Gillespie et al. 2011**).

MetaMiner: SSU rDNA Analyses

Isolation and classification of SSU rDNA sequences. In order to assess the distribution of bacterial sequences in a eukaryotic genome project, decontaminated and validated sequencing reads were mapped against a library of SSU rDNA sequences, including full-length bacterial 16S rDNA sequences from the Greengenes database (**DeSantis et al. 2006**) plus full-length 18S rDNA sequences from the eukaryote (host). If no 18S rDNA sequences were available for the host, sequences from the closest relative(s) were used instead. Sequencing reads were aligned to this SSU rDNA library using the Burrows-Wheeler Aligner (`BWA-SW`) (**Heng Li & Durbin 2009**) run under stringent match criteria (q=2, r=7, T=200). Reads that had at least one successful match in the SSU rDNA library were binned according to the taxonomic classification

of their matches (as derived from Greengenes), with the distribution subsequently visualized using `Krona` (**Ondov et al. 2011**). Reads that mapped to the host 18S rDNA were identified, quantified, and used to estimate the coverage of the host genome:

$$C = \frac{L_r}{L_g \cdot N} \quad (1)$$

Where:

C is the estimated coverage
L_r is the total length across all 18S rDNA reads
L_g is the length of the complete 18S rRNA gene
N is the copy number of the gene

The calculated coverage is expected to be roughly the same as (or slightly less than) the original coverage estimate for the sequencing project, and can serve as an initial rough verification on the mapping procedure.

Identification of minimally divergent sets of 16S rDNA reads. Taxa that may contain a potential target were identified through manual inspection of the SSU rDNA read distribution. Reads that mapped to bacterial 16S rDNA sequences from a target taxon of interest were extracted from the full set, combined with full-length 16S rDNA sequences from related bacterial species, and all-vs.-all pairwise gene divergences calculated. Gene divergence is defined here as the percent of non-identical positions between pairs of individual sequences. It is calculated by constructing an optimal global pairwise alignment (Needleman-Wunsch) between each pair of sequences using the `needle` program from `EMBOSS` (**Rice et al. 2000**) and computing the percentage of identical positions in each alignment:

$$D_g = \left(\frac{1-i}{L_a}\right) \cdot 100 \quad (2)$$

Where:

D_g is the gene divergence
i is the number of identical positions
L_a is the total length of aligned positions

Reads with a mean divergence less than 20% across all of the full-length 16S rDNA sequences from the target taxon were considered candidate reads to assemble the 16S rDNA sequence for the target.

Assembly of a target 16S rDNA sequence. Minimally divergent candidate 16S rDNA reads were combined with two additional data sets: 1) full-length 16S rDNA sequences from related bacterial species (PATRIC); and 2) at least one full-length 16S rDNA sequence from a distantly related (outgroup) bacterial species (PATRIC). All sequences were subsequently aligned using `MUSCLE` (**Edgar 2004**) with default parameters. Ambiguously aligned positions were identified and removed, using either `Gblocks` (with minimum length of a block -b4=5) or in cases where `Gblocks` led to weak support on key branches, by manual masking of the alignment.

Phylogenies were estimated under maximum likelihood using RAxML (**Stamatakis 2006**) with the GTR substitution model and estimation of GAMMA and the proportion of invariable sites. Branch support was measured with bootstrapping (1000 replications) (**Felsenstein 1985**).

Minimally divergent 16S rDNA reads that were also monophyletic in the resulting tree were determined to represent 16S rDNA sequences from a single target microbe. In order to stitch together a full-length 16S rDNA for this organism, MUSCLE was again used to align the monophyletic reads with the full-length 16S rDNA sequence from a close relative. The subsequent alignment was manually examined to identify and merge overlapping reads. Positions with less than 3X coverage across the merged reads were trimmed and a consensus sequence was subsequently determined.

Phylogenetic placement of the target 16S rDNA sequence. Results from two taxon sampling methods (neighborhood and cascading) and two tree-building algorithms (RAxML and PhyloBayes) were compared to estimate a robust phylogeny of the mined target 16S rDNA sequence. In the neighborhood sampling method, the target 16S rDNA sequence was used as the query in a single blastn search of the NCBI *refseq_genomic* database. Multiple sequences from the same taxon were culled to leave only the single best match for each taxon, and the best 200 of these **unique-taxon matches** were subsequently retained. In the cascading sampling method, the target 16S rDNA sequence was used as the query in a series of blastn searches against the NCBI *refseq_genomic* database. The exact taxonomic groups that comprise the cascade are determined by the particular study; in general, however, the immediate taxonomic parent of the target is first searched for the closest unique-taxon matches. Next, each sibling group to the target's parent is searched, followed by each sibling group to the parent of the parent, and so on until the level of phylum is reached. Finally, all eukaryotes (excluding the host) are searched for additional unique-taxon matches.

The resulting sequences from each search method were aligned using MUSCLE with default parameters. Ambiguously aligned positions were identified through manual inspection of the alignment and removed. Phylogenies were estimated under maximum likelihood using RAxML under the GTR substitution model, with estimation of GAMMA and the proportion of invariable sites. Branch support was measured with bootstrapping (1000 replications). Phylogenies were also estimated using PhyloBayes (**Lartillot et al. 2009**) under the CAT model of substitution, a nonparametric method for modeling site-specific features of sequence evolution **(Lartillot et al. 2007**). Two independent Markov chains were run in parallel using PhyloBayes-MPI under the CAT-GTR model, with the bipartition frequencies analyzed at various time points using the included bpcomp program. For tree-building, appropriate burn-in values were determined by plotting the log likelihoods for each chain over sampled generations (time). Analyses were considered complete when the maximum difference in bipartition frequencies between the two chains was less than 0.1. In general, a burn-in value of 1000 with sampling every two trees was sufficient to build a consensus tree.

Divergence from sequenced bacterial genomes. The extent of divergence of the target 16S rDNA from those of sequenced bacterial genomes is used to inform downstream read mining and assembly sifting workflows. Highly divergent 16S rDNA sequences suggest the need to use a broader set of genomes as bait in subsequent analyses. In order to calculate gene divergence, the target 16S rDNA sequence was combined with sequences from closely related bacteria that have available genome sequences, and pairwise divergence was calculated as described above.

ReadMiner: Analysis of Bacterial-Like Host Reads

Isolation and assembly of bacterial reads. Complete and whole genome shotgun (WGS) sequences for bacteria (excluding plasmids), with minimal divergence from the target as determined during SSU rDNA analysis, were downloaded from PATRIC and concatenated into a single multi-genome library. These genomes constitute the **best-matching clade** for the target genome. Similarly, genomes from taxa identified as potential contaminants during SSU rDNA analysis were also added to the library; these genomes constitute the **best-competing clade** for the target. Host reads were subsequently aligned to this genome library using `BWA-SW` under moderately stringent match criteria (q=2, r=7, T=100). Reads that had at least one successful match in the best-matching clade were retained, along with their mate-pairs where applicable. The extracted reads were de-duplicated and subsequently assembled *de novo* to the level of contig using `Mira` (job=denovo,accurate,sanger noclipping=all SANGER_SETTINGS) (**Chevreux 2005**). Gene models were predicted on the assembled contigs using the *ab initio* gene prediction program `fgenesb` (http://linux1.softberry.com/berry.phtml?topic=fgenesb&group=help&subgroup=gfindb) under default bacterial gene-finding parameters. In order to account for partial gene models and potential pseudogenes, the predicted protein sequences were queried against the *nr* database at NCBI using `blastp` at an e-value cutoff of 10^{-3}. Queries with no matches, or that aligned over less than 50% of the total subject length, were discarded; the remaining models constitute the initial set of mined bacterial genes.

Removal of mined contaminant genes. The presence of best-competing clade genomes during the initial read mining step is designed to titrate the number of contaminant reads mistakenly assigned to the target. To further reduce contamination, mined bacterial proteins were combined with protein sequences from the best-matching and best-competing clades and clustered into orthologous groups (OGs) using `FastOrtho`, a custom implementation of the OrthoMCL algorithm (**Li et al. 2003**). Mined bacterial proteins that cluster exclusively with proteins from the best-competing clade genomes were subsequently discarded, and the remaining genes were considered to be the final set of **mined target genes**.

Identification of a conserved signal among mined target genes. Mined target proteins were combined with all annotated proteins from selected genomes within the same taxonomic group (including the best-matching clade), and OGs were constructed using `FastOrtho` as described above. Mined target proteins with at least one ortholog in 80% or more of the genomes according to this analysis were subjected to *n*-taxon statement analysis as follows. Briefly, each protein was combined with its top `blastp` matches against at least three separate groups: 1) its parent taxonomic group; 2) all bacteria except its parent group; and 3) all eukaryotes except the host. The sequences were aligned using `MUSCLE` under default parameters, with phylogeny estimated using `RAxML` under the WAG substitution model (**Whelan & Goldman 2001**) with estimation of GAMMA. Branch support for each tree was measured with bootstrapping (100 replications). Mined target proteins that were most closely related to the top match within their immediate parent group were deemed **core genes**. Based on 1) similarity to best-matching clade sequences, 2) broad orthology across the parent group, and 3) monophyly with the best match in the parent group, it was concluded that the mined core genes were all vertically inherited from a common ancestral bacterium.

Genome-based phylogeny. Phylogeny estimated using 16S rDNA sequences is informative, but subject to various biases (**Wu & Eisen 2008**). To better understand the systematic position of the target microbe, a phylogeny inferred from multiple protein sequences was used. The mined core proteins were combined with at least three additional data sets: 1) all genomes from the best-matching clade; 2) a roughly equivalent number of genomes from the same parent group but outside of the best-matching clade; and 3) at least two outgroup taxa. For genome-based phylogeny estimation, an automated pipeline for protein family construction and tree building was used (**Gillespie et al. 2011**). Briefly, `BLAT` (refined `BLAST` algorithm) (**Kent 2002**) was used to identify similar protein sequences between all genomes, and `mcl` (**van Dongen 2000**) was used to cluster the results into OGs. The OGs were filtered to include only those with membership in at least 80% of the input genomes, and `MUSCLE` (default parameters) was used to align each OG, with regions of poor alignment masked using `Gblocks` (default parameters) (**Castresana 2000**). All alignments were subsequently concatenated into a single data set for phylogeny estimation using `FastTree` (**Price et al. 2010**) as described in Chapter 4.

Genome divergence. To assess the degree of divergence across the mined core sequences and their closest orthologs, the final alignment of the core proteins was processed to include only a single representative species from each genus in the best-matching clade (plus the target itself). All positions of the alignment that contained missing data were removed, and `DIVEIN` (**Deng et al. 2010**) was used to estimate percent protein divergence using both Blosum62 and WAG substitution matrices.

Identification of mined signature genes. Mined target genes with 1) orthologs in less then 20% of genomes outside of the best-matching clade, or 2) evidence of paraphyly or polyphyly from the *n*-taxon statement analysis, were analyzed manually for the presence of **signature genes**: functions specific to the ecological niche of the target or its relationship to the eukaryote from which it was mined. See Chapter 4 for an example of signature gene analysis.

AssemblySifter: Analysis of Assembled Host CDS

Identifying bacterial-like host proteins. A `BLAST`-based pipeline was used to identify candidate bacterial (especially target) proteins from within the host genome assembly, either as contaminants or actual components of host scaffolds. Each protein from the host was queried using `blastp` against three non-overlapping data sets: 1) all available proteins from bacteria that are closely related to the target, as determined by SSU rDNA analyses (**clade**); 2) all bacterial proteins excluding those in the clade data set (**bact**); and 3) all eukaryotic proteins excluding those from the host itself (**euk**). Multiple matches of a query to subjects from the same taxon were culled to leave only the single best-matching subject from that taxon (**unique-taxon matching**).

The top 25 (based on e-value) unique-taxon matches in each data set were pooled and ranked according to a comparative sequence similarity match score:

$$S_m(m,h) = b \cdot I \cdot Q \tag{3}$$

Where:

S_m is the comparative sequence similarity score
b is the raw bitscore of match *m* to host protein *h*

I is the percent identity
Q is the percent of *h* (the query) that aligned

By incorporating percent identity and alignment length in this manner, S_m is intended to de-emphasize highly significant matches to short stretches of query (*i.e.*, conserved domains) in favor of longer stretches of similarity. The top five (based on S_m) matches to each host protein from the pooled lists of subjects were ranked (1-5) and labeled with their source data set.

Assessing the bacterial nature of host proteins. In order to identify host proteins that may be either directly bacterial, or derived from a bacterial source, each host protein was assessed for the extent of its bacterial nature based on the makeup of its top-5 pooled S_m matches, using a comparative taxonomic nature score N_t. First, the S_m score for each match *m* to a ranked host protein *h* was transformed into a weighted score:

$$S_m^*(m,h) = S_m(m,h) \cdot \left[n(h) - R(m,h) + 1 \right] \quad (4)$$

Where

S_m^* is the weighted S_m score
n is the total number of pooled, ranked matches to *h*
R is the pooled rank order of *m* with respect to *h*

Next, a raw comparative nature score $N_t^{'}$ for *h* was calculated for each source data set *t*:

$$N_t^{'}(h,t) = \frac{\sum_t S_m^*(m,h)}{n(h)} \quad (5)$$

$N_t^{'}$ has meaning when comparing matches for each individual host protein to different data sets (*i.e.*, clade, bact, and euk), but not for comparisons between different host proteins. In order to facilitate between-protein comparisons, each $N_t^{'}$ value was normalized using the optimal (best possible) $N_t^{'}$ for host protein *h* assuming all of its pooled, ranked matches derive from a single data set:

$$N_t(h,t) = \frac{N_t^{'}}{N_t^{'}(h,opt)} \quad (6)$$

Final N_t values for each host protein to each data set *t* range from zero (no top-5 matches in data set *t*) to one (all top-5 matches in data set *t*).

Host proteins were binned according to N_t into four broad groups: **Candidate target proteins** included host proteins where N_{clade} was one and both N_{euk} and N_{bact} were zero; *i.e.*, all top-5 pooled matches to the host protein were from the clade data set. Proteins in this set were hypothesized to belong to the target itself. **Possible contaminants** included host proteins where N_{euk} was zero and both N_{clade} and N_{bact} were greater than zero. Proteins in this set were hypothesized to be bacterial, either from the target or from other bacteria. **Possible lateral gene transfer (LGT) proteins** included host proteins where N_{euk} greater than zero and at least one of

N_{clade} or N_{bact} were greater than zero. Proteins in this set were hypothesized to be bacterial in origin, and candidates for possible LGT events between host and target (or other bacteria). Finally, **fully eukaryotic proteins** included host proteins where N_{euk} was one and both N_{clade} and N_{bact} were zero.

Assessing the source of bacterial-like host genes. Several approaches were taken to determine the source of bacterial-like genes sifted from the host assembly. First, all but the fully eukaryotic genes from the N_t analysis were placed onto their genomic scaffolds, and mean N_t values were calculated from all genes on each scaffold. Short scaffolds that contained only candidate target genes were determined to belong to the target genome. Long scaffolds that had a significant eukaryotic nature were subjected to LGT analysis as described below. Scaffolds of an indeterminate nature were manually inspected to determine if they should be assigned to the target, analyzed for possible LGT, or discarded as non-target contamination.

Analysis of sifted target scaffolds. Bacterial gene models were called on sifted scaffolds determined to belong to the target genome, using `fgenesb` under default bacterial gene-finding parameters. These models constitute the initial set of **sifted target genes**. Each model was subjected to *n*-taxon statement analysis as described above, and genes monophyletic with the parent taxonomic group of the target microbe were determined to be **core genes**. Sifted core genes were used to estimate genome-based phylogeny of the target as described above for the mined core genes.

Evaluating microbial gene transfer to the host. Sifted scaffolds that were significantly eukaryotic in nature but contained bacterial-like genes were deemed possible loci of LGT events. To assemble evidence supporting or refuting this contention, bacterial-like genes on these scaffolds were first divided into single-exon and multi-exon groups. Each multi-exon gene (exons plus introns) was used as the query in a `blastx` search against the *nr* database. In addition, these entire gene models were analyzed with `fgenesb` using default bacterial gene-finding parameters, to determine discrepancies with the original eukaryotic gene predictions. Finally, individual protein phylogenies were estimated for both single-exon and multi-exon genes by aligning them to their original AssemblySifter blast matches using `MUSCLE` (default parameters), masking regions of poor alignment with `Gblocks` (default parameters), and building phylogenetic trees using RAxML with estimation of GAMMA and amino acid substitution models WAG and BLOSUM62. Branch support for each tree was measured with bootstrapping (100 replications). This LGT evaluation process allows for the distinction between true eukaryotic genes inadvertently categorized as bacterial (*e.g.*, nuclear genes encoding mitochondrial proteins), and bacterial LGTs undergoing transformation to eukaryotic-like gene structure (accumulation of introns, eukaryotic signal sequences, *etc*.). It can also reveal evidence against predicted introns by 1) identifying chimeric gene models comprised of multiple bacterial genes merged by the eukaryotic gene calling algorithms, 2) bacterial genes fragmented due to multiple start sites called by eukaryotic gene calling algorithms, and 3) gene models fused with additional, short ORFs that are likely to be artifacts.

RESULTS & DISCUSSION

All methods from this project are available as interoperable Perl and shell scripts, and will be deposited in the public software repository github (http://www.githib.com/). A complete list of

these scripts can be found in **appendix A1** and **appendix A2**. A versioned list of external executables used in these studies can be found in **appendix A3**. The use of these pipelines, either alone or in combination, is driven by the particular system under study; consequently, they are designed to be complementary but not interdependent. A general overview of the interoperability of the workflows is presented in **fig. 1**.

Data Retrieval and Preparation

Choice of long-read sequence data. Both MetaMiner and ReadMiner utilize genome sequence trace read data as input; the general pipeline for retrieving and processing sequencing reads is shown in **fig. 2**. The focus of the current data processing pipeline is long-read data (200+ bases per read) such as those produced by WGS or pyrosequencing projects. A primary advantage of long-read data is that they do not require assembly before analysis by MetaMiner or ReadMiner, effectively eliminating the possibility of artificial sequence mosaics due to misassembly. In addition, long reads are superior queries when searching for homologous sequences (**Wommack et al. 2008**). Finally, long-read data facilitate phylogenomic analyses using individual reads by increasing the number of useful positions. The disadvantages of long-read data are that they are generally more expensive and slower to produce, and provide more shallow genome coverage than the short-read data (25-100 bases per read) produced from more recent next-generation technologies.

Decontaminating input sequences. Raw genome data are generally stored in one of several formats at a sequence archive such as the NCBI Trace Archive. The archived reads from a typical eukaryote genome project routinely exceed 30 GB in total size and are largely unprocessed, containing variable amounts of unresolved sequence, low-quality regions, cloning vector or adapter sequences, and biological contaminants. These issues must be dealt with prior to any substantive analysis of the data. It is especially important to ensure the selective removal of as much cloning vector contamination as possible from WGS reads, since vector sequences are typically bacterial in nature and will complicate downstream analyses.

Once a full set of reads has been retrieved from the sequence archive, it is decontaminated by trimming read termini that match to known cloning vectors, and then removing any reads with excessive vector contamination. This process may remove up to 10% or more of the original sequence data, though much of this loss is attributable to read trimming rather than outright removal. After vector sequences have been eliminated, the reads are subjected to several rounds of quality control monitored using `fastqc` as described below.

Checkpoints and the quality control cycle. Since every input data set is different, checkpoints (workflow steps that allow manual inspection and control) are an important part of the data preparation pipeline. Each step in the quality control process utilizes a threshold (position, quality, or length) that is customizable depending upon manual inspection. In addition, each step or cycle can be repeated in order to achieve the desired overall data quality. For example, the quality per position across all reads typically follows a normal distribution in shotgun sequencing data, with poorer quality base calls appearing at the 5' and 3' ends; calculation and inspection of this distribution is the first step post-decontamination, and the results used to determine how far back to trim the reads. This step often serves the additional purpose of removing any remaining vestiges of cloning vector or adapter contamination.

Continual monitoring of the data set using `fastqc` during the preparation cycle determines when it is ready for analysis. The data preparation pipeline typically removes approximately 10-

20% of the input sequences, with the remaining reads trimmed to approximately 50-60% of their original length (mostly from the 3' end). These values can vary widely depending on the quality of the original data and the rigor of the preparation cycle.

MetaMiner: Identification of Bacterial Targets

Choice of the 16S SSU rRNA gene. MetaMiner is designed to enable the systematic discovery and classification of bacteria associated with a host genome sequencing project. The workflow extracts reads based on their similarity to known bacterial 16S rRNA genes, and facilitates the assembly of those reads into complete (or near-complete) 16S rDNA sequences through a combination of pairwise alignment and phylogenetic estimation (**fig. 3**). The bacterial 16S rRNA gene is a popular and reasonable choice for a single classifying genetic marker. It is relatively long (1500 bases), ubiquitous among bacteria, and has remained functionally stable over time. In addition, the existence of several comprehensive databases of bacterial 16S rDNA sequences (**DeSantis et al. 2006; Cole et al. 2009**) provides a rich set of data for comparative analyses. It has been shown previously that the 16S rRNA gene lacks phylogenetic resolution at the species level (**Janda & Abbott 2007**), and results from the three studies presented here reflect this inherent limitation (**fig. 4**): over 85% of all reads that mapped to bacterial 16S rDNA sequences could not be assigned to individual species, a direct result of the taxonomic assignment of the bait 16S rDNA sequences. Interestingly, the success rate for assigning 16S rDNA reads to genus or family varied widely between studies; this may be associated with the overall divergence of bacteria in the different read sets, although that hypothesis was not tested. All reads in all studies presented here could be assigned to the level of taxonomic order or better.

Determination of the best-matching and best-competing clades. In addition to providing insight into the community composition of the sequence data, MetaMiner also facilitates determination of the best-matching clade for the target microbe; genomes from this clade are used as the bait in subsequent ReadMiner and AssemblySifter analyses. In general, the most restrictive clade possible is selected, using the phylogenetic position of the target 16S rDNA sequence to dictate the initial locality. In the trivial case - when the target itself has a sequenced genome - the best-matching clade is simply the target genome (for example, see Chapter 3). In most cases, however, the scope of the clade is determined using a combination of factors, including 1) the quality and composition of the locality, 2) the available number of sequenced genomes, 3) the percent divergence of the target 16S rDNA sequence from its neighbors, 4) the reproducibility of the target's phylogenetic estimation across different combinations of taxon sampling and tree construction methods, and 5) the composition of off-target 16S rDNA-like reads, which may restrict the clade so as to avoid mining a mosaic of sequences.

MetaMiner is also used to determine the necessity for, and composition of, a best-competing clade. This clade (or clades) is comprised of genomes that may be present in the full set of reads and show sequence similarity to the best-matching clade (for example, see Chapter 5). The discussion of ReadMiner below provides more information on how the best-competing clade is used.

ReadMiner: Extracting Target Reads

Balancing best-matching and best-competing clades. ReadMiner (**fig. 5**) is designed to comprehensively extract sequencing reads that belong to the target microbe specified by MetaMiner. In the initial phase of the workflow, reads that are similar in sequence to genomes in the best-matching clade are extracted; these mined reads are considered candidate target

sequences. In some cases, off-target microbes in the input set may show significant sequence similarity to the target genome; if they are similar enough, sequences from these off-target genomes may be erroneously labeled as candidate target sequences. This problem is compounded by the observation that off-target competitors need not be phylogenetically related to the target; an example of this can be seen in Chapter 5. In order to account for contamination by off-target sequences, ReadMiner can include a best-competing clade in the extraction step. Reads that match preferentially to the best-competing clade are binned separately from those that prefer the best-matching clade. In addition, ReadMiner can use OG membership to identify and remove mined target genes that are only found in the best-competing clade.

Complications arising from bacteria-to-bacteria LGTs. The transfer of genetic material independent of vertical inheritance mechanisms, either as single genes or longer stretches of contiguous DNA, is widespread among prokaryotes and a major force in bacterial evolution (**Bapteste et al. 2004**). As a result, it is likely that any target genome identified by MetaMiner contains genetic material laterally transferred from other bacteria. If a transfer is relatively recent and originated from a distantly related species, the sequence may be distinct from anything in the best-matching clade of the target, and ReadMiner may be unable to identify the reads that comprise the transferred genetic material as belonging to the target. One possible approach to mine target LGT regions would be to extend the target contigs assembled by ReadMiner using an iterative algorithm, whereby reads from the input set that overlap at the ends of a target contig would be extracted and added to the contig.

Highly divergent target extraction. MetaMiner enables the discovery of targets that may be highly divergent from known genomes, by virtue of 1) the relatively high conservation of the 16S rRNA gene, and 2) the large number of bacterial 16S rDNA sequences (approximation of reality) compared to full bacterial genomes available (tip of the iceberg of microbial diversity). ReadMiner also relies on sequence similarity to identify genomic reads in the input set that belong to the target microbe; however, if the target shows a high degree of divergence from known genomes, ReadMiner may be unable to extract a significant number of target reads. An example of this can be seen in Chapter 4. One possible recourse is to relax the stringency for retaining matches, which may increase the number of mined sequences; another option is to define a broader best-matching clade. A disadvantage of both options is a possible increase in the level of contamination of the result set. The challenge of characterizing novel, highly divergent targets was part of the motivation for the development of AssemblySifter (see below).

AssemblySifter: Extracting Target Sequences from the Host Genome

Unlike ReadMiner, which uses the trace reads from a eukaryote sequencing project as input, AssemblySifter examines the genes and scaffolds that comprise the assembled host genome (**fig. 6**). Although the trace reads arguably represent a richer source of target sequence data since they have not been pre-filtered, there are several advantages to sifting the host assembly. First, the assembled genome may be the only source of data for a particular eukaryote. Second, AssemblySifter can assist in the identification of host scaffolds that may contain bacteria-to-host LGTs. Finally, AssemblySifter can help improve the quality of existing annotation databases by removing bacterial sequences from eukaryotic genomes, and also by re-annotating their component genetic and operonic features using bacterial models.

AssemblySifter uses two novel measures of similarity to identify candidate bacterial genes and bacteria-host LGTs: S_m is the bitscore of a match, attenuated by the percent identity and the

length of the match; N_t is the weighted relative proportion of the top S_m scores to each query database. Other measures have been used previously to identify possible LGTs; two popular methods are Alien Index and h_U (**Boschetti et al. 2012**). S_m and N_t have two advantages over previous measures that make them a viable alternative. First, they are based on multiple matches to a single query (the top five matches are used here, but theoretically any number can be used). This flexibility allows for the possibility that some matches may be incorrect, for example due to errors in annotation or failure to completely remove contaminating sequences from published genomes. Second, since S_m incorporates the length of the alignment between a query and its match, high-scoring domain-level matches are selected against.

Host scaffolds with a strong bacterial signal. Sifted genes are determined to be bacterial (*i.e.*, belonging to a distinct bacterial entity) if they exhibit a strong bacterial signal in their blast pattern and protein phylogeny, and are present on a scaffold that is itself comprised primarily or exclusively of bacterial genes. In these cases, the scaffold itself is also determined to be bacterial, and if the signal is confined to the best-matching clade for the target, the scaffold is considered to belong to the target specifically. It is assumed that such scaffolds have been included in the host genome assembly erroneously, although it is also conceivable that they could be both bacterial and contiguous with the host genome (**Nikoh et al. 2008**). In either case, predicting genes on bacterial scaffolds using a eukaryotic gene prediction algorithm may lead to various types of errors, including incorrect start site identification, erroneous intron splice site predictions, and missed genes due to differences in mean intergenic length between eukaryotes and prokaryotes. Consequently, AssemblySifter enables the re-annotation of bacterial scaffolds using a bacterial gene calling model; these results can be compared with the original predictions to further clarify the origin of these scaffolds.

Host scaffolds with a strong eukaryotic signal. Sifted genes are determined to belong to the host if they exhibit a strong eukaryote signal, and are present on primarily or exclusively eukaryotic scaffolds. Typically, these scaffolds are not analyzed further in the current workflow.

Host scaffolds with a hybrid bacterial/eukaryotic signal. A host scaffold may be a hybrid of bacterial and eukaryotic genes for a variety of reasons. Most significantly, a hybrid scaffold may arise as a result of the lateral transfer of genetic material into the host from a closely-associated microbe, like an endosymbiont. While traditionally thought to be rare, examples of such microbe-to-host LGTs are being discovered with increasing regularity (**Gladyshev et al. 2008; Chapman et al. 2010; Acuña et al. 2012**) and may play a significant role in the development of host-microbe symbioses (**Ros & Hurst 2009**).

Genes transferred into the host from a symbiont are anticipated to exhibit sequence similarity to other bacterial genes, but may contain distinctly eukaryotic features such as introns, enhancers, or eukaryotic promoters. There may also be changes to the translation start site, or transition to a more eukaryotic AT or codon bias as a result of host-specific evolutionary pressures. The presence and degree of these changes would depend to some extent on the relative age of the transfer event, with older transfers generally resembling the host genome more than newer transfers. To facilitate the evaluation of these attributes, AssemblySifter compiles a gene-by-gene analysis for each hybrid scaffold that includes the number of introns, intergenic distance fore and aft, comparison to a re-annotation of the sequence using a bacterial gene model, similarity to known bacterial and non-bacterial sequences, and phylogenetic estimations of both the protein and its re-annotated counterpart.

Alternatively, a hybrid scaffold may represent a region of the host genome that contains widely conserved genes, including ribosomal genes, polymerases, genes involved in core metabolic processes, etc. A hybrid scaffold may also be an assembly-induced artifact, a mosaic of host and bacterial reads that coincidentally overlap but are biologically incongruous. These can be identified in AssemblySifter as scaffolds that contain strongly eukaryotic genes interspersed with strongly bacterial genes that lack any eukaryotic features. In its current form, AssemblySifter ultimately relies on manual evaluation to distinguish these artifacts from possible lateral transfers, and definitive evidence will require follow-up experimental validation (*e.g.*, protein and mRNA expression) of the most promising candidates.

CONCLUSION

The methodology presented here employs sequence similarity and phylogeny to identify and characterize bacterial components within heterogeneous sequence data. It utilizes a multifarious approach that integrates the analysis of both pre- and post-assembly sequence data. Such an approach allows application of the underlying methodology in this dissertation to a broad variety of biological systems.

LITERATURE CITED

Acuña R et al. 2012. Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. Proceedings of the National Academy of Sciences. 109:4197–4202. doi: 10.1073/pnas.1121190109.

Bapteste E, Boucher Y, Leigh J, Doolittle WF. 2004. Phylogenetic reconstruction and lateral gene transfer. Trends Microbiol. 12:406–411. doi: 10.1016/j.tim.2004.07.002.

Boschetti C et al. 2012. Biochemical diversification through foreign gene expression in bdelloid rotifers. PLoS Genet. 8:e1003035. doi: 10.1371/journal.pgen.1003035.

Camacho C et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics. 10:421. doi: 10.1186/1471-2105-10-421.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.

Chapman JA et al. 2010. The dynamic genome of *Hydra*. Nature. 464:592–596. doi: 10.1038/nature08830.

Chevreux B. 2005. MIRA: An Automated Genome and EST Assembler. Ph.D. Thesis. 1–171.

Cole JR et al. 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucl. Acids Res. 37:D141–5. doi: 10.1093/nar/gkn879.

Deng W et al. 2010. DIVEIN: a web server to analyze phylogenies, sequence divergence, diversity, and informative sites. BioTechniques. 48:405–408. doi: 10.2144/000113370.

DeSantis TZ et al. 2006. Greengenes, a chimera-checked 16S rRNA gene database and

workbench compatible with ARB. Appl Environ Microbiol. 72:5069–5072.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res. 32:1792–1797. doi: 10.1093/nar/gkh340.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution. 783–791.

Gillespie JJ et al. 2011. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. Infect Immun. 79:4286–4298. doi: 10.1128/IAI.00207-11.

Gladyshev EA, Meselson M, Arkhipova IR. 2008. Massive horizontal gene transfer in bdelloid rotifers. Science. 320:1210–1213. doi: 10.1126/science.1156407.

Green P. 1996. Phrap documentation. http://www.phrap.org/phredphrap/phrap.html.

Hannon GJ. 2010. Fastx Toolkit. http://hannonlab.cshl.edu/fastx_toolkit/index.html.

Hotopp JCD et al. 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science. 317:1753–1756. doi: 10.1126/science.1142490.

Janda JM, Abbott SL. 2007. 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls. Journal of Clinical Microbiology. 45:2761–2764. doi: 10.1128/JCM.01228-07.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. Genome Research. 12:656–664. doi: 10.1101/gr.229202.

Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P. 2008. A bioinformatician's guide to metagenomics. Microbiol Mol Biol Rev. 72:557.

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol. 7 Suppl 1:S4. doi: 10.1186/1471-2148-7-S1-S4.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics. 25:2286–2288. doi: 10.1093/bioinformatics/btp368.

Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI. 2008. Worlds within worlds: evolution of the vertebrate gut microbiota. Nat Rev Micro. 6:776–788. doi: 10.1038/nrmicro1978.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25:1754–1760. doi: 10.1093/bioinformatics/btp324.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Research. 13:2178–2189. doi: 10.1101/gr.1224503.

Mendes R et al. 2011. Deciphering the rhizosphere microbiome for disease-suppressive bacteria. Science. 332:1097–1100. doi: 10.1126/science.1203980.

Nikoh N et al. 2008. *Wolbachia* genome integrated in an insect chromosome: evolution and fate of laterally transferred endosymbiont genes. Genome Research. 18:272–280. doi: 10.1101/gr.7144908.

Ondov BD, Bergman NH, Phillippy AM. 2011. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. 12:385. doi: 10.1186/1471-2105-12-385.

Pagani I et al. 2012. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucl. Acids Res. 40:D571–9. doi: 10.1093/nar/gkr1100.

Pickett BE et al. 2012. ViPR: an open bioinformatics database and analysis resource for virology research. Nucl. Acids Res. 40:D593–8. doi: 10.1093/nar/gkr859.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS ONE. 5:e9490. doi: 10.1371/journal.pone.0009490.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16:276–277.

Ros VI, Hurst GD. 2009. Lateral gene transfer between prokaryotes and multicellular eukaryotes: ongoing and significant? BMC Biol. 7:20. doi: 10.1186/1741-7007-7-20.

Round JL, Mazmanian SK. 2009. The gut microbiota shapes intestinal immune responses during health and disease. Nat Rev Immunol. 9:313–323. doi: 10.1038/nri2515.

Sekirov I, Finlay BB. 2006. Human and microbe: united we stand. Nature Medicine, July 1 doi: 10.1038/nm0706-736.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22:2688–2690. doi: 10.1093/bioinformatics/btl446.

van Dongen SM. 2000. Graph clustering by flow simulation.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 18:691–699.

Wommack KE, Bhavsar J, Ravel J. 2008. Metagenomics: read length matters. Appl Environ Microbiol. 74:1453–1463.

Wu M, Eisen JA. 2008. A simple, fast, and accurate method of phylogenomic inference. Genome Biol. 9:R151. doi: http://dx.doi.org/10.1186/gb-2008-9-10-r151.

FIGURES

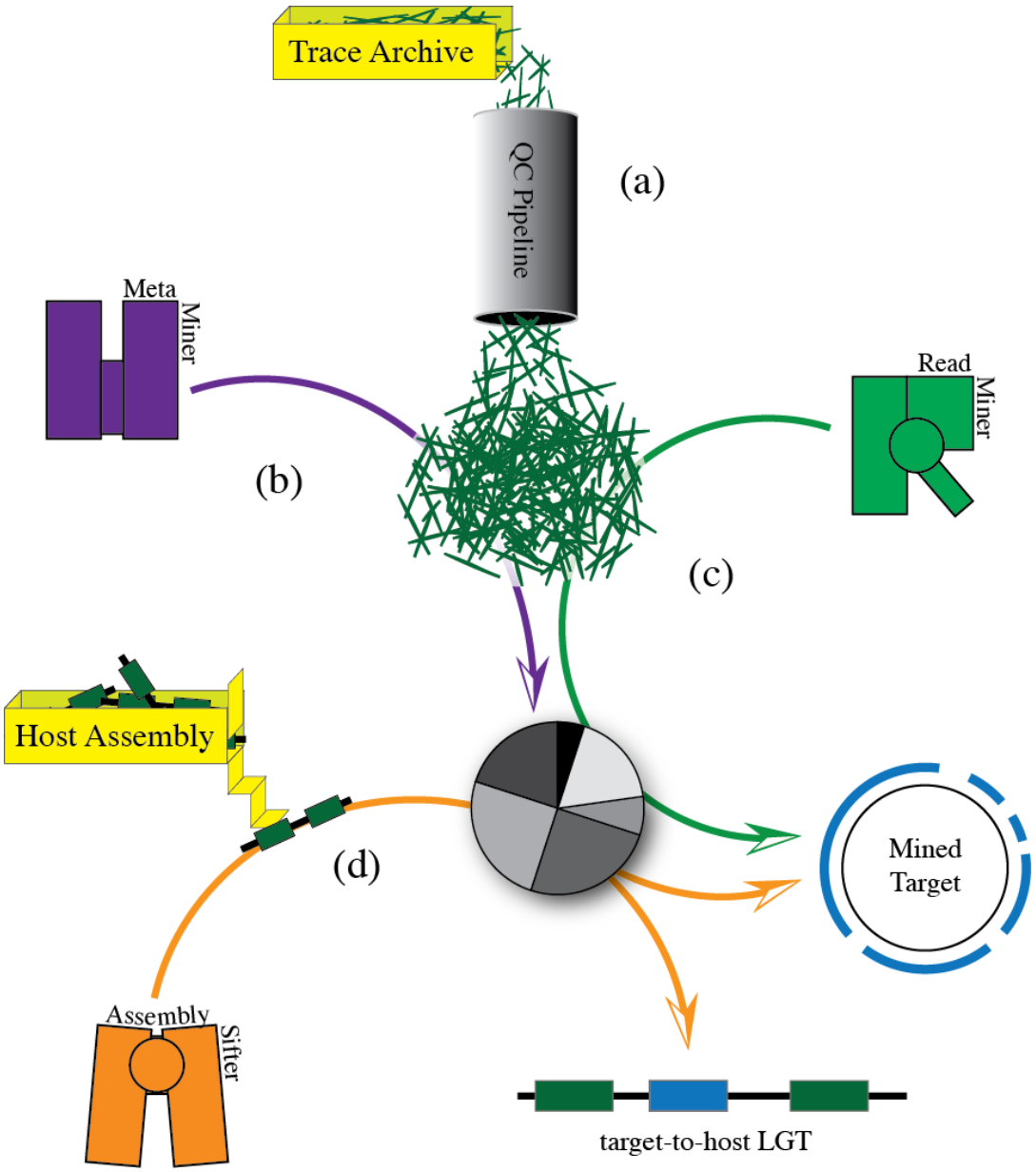


Figure 1. General schematic of the MetaMiner, ReadMiner, and AssemblySifter workflows. (**a**) QC Pipeline. Genome trace reads from a eukaryote of interest (host) are filtered for quality and length. (**b**) MetaMiner. Host reads are matched against a library of SSU sequences to identify bacterial components. (**c**) ReadMiner. Host reads are matched against a library of bacterial genomes to extract reads for a single bacterial target. (**d**) AssemblySifter. The assembled host genome is searched for bacterial-like genes and genomic fragments. See the text for full descriptions.

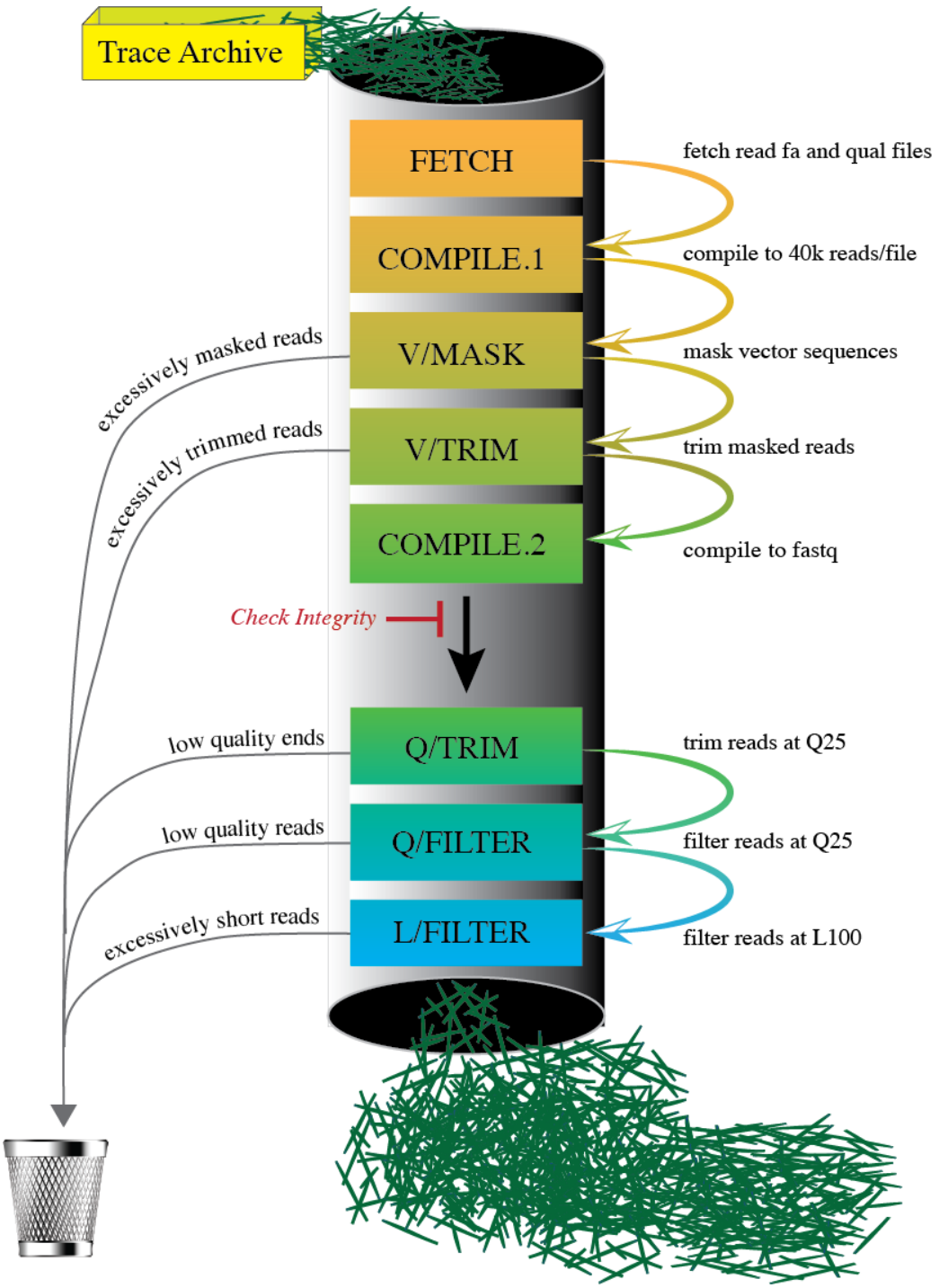


Figure 2. Schematic illustrating the major steps in the quality control pipeline. Genome trace reads from a host eukaryote are screened for cloning vector or adapter sequences, trimmed, and compiled. Remaining reads are trimmed to remove low quality termini, and reads with low overall quality or short length are removed. See the text for a full description.

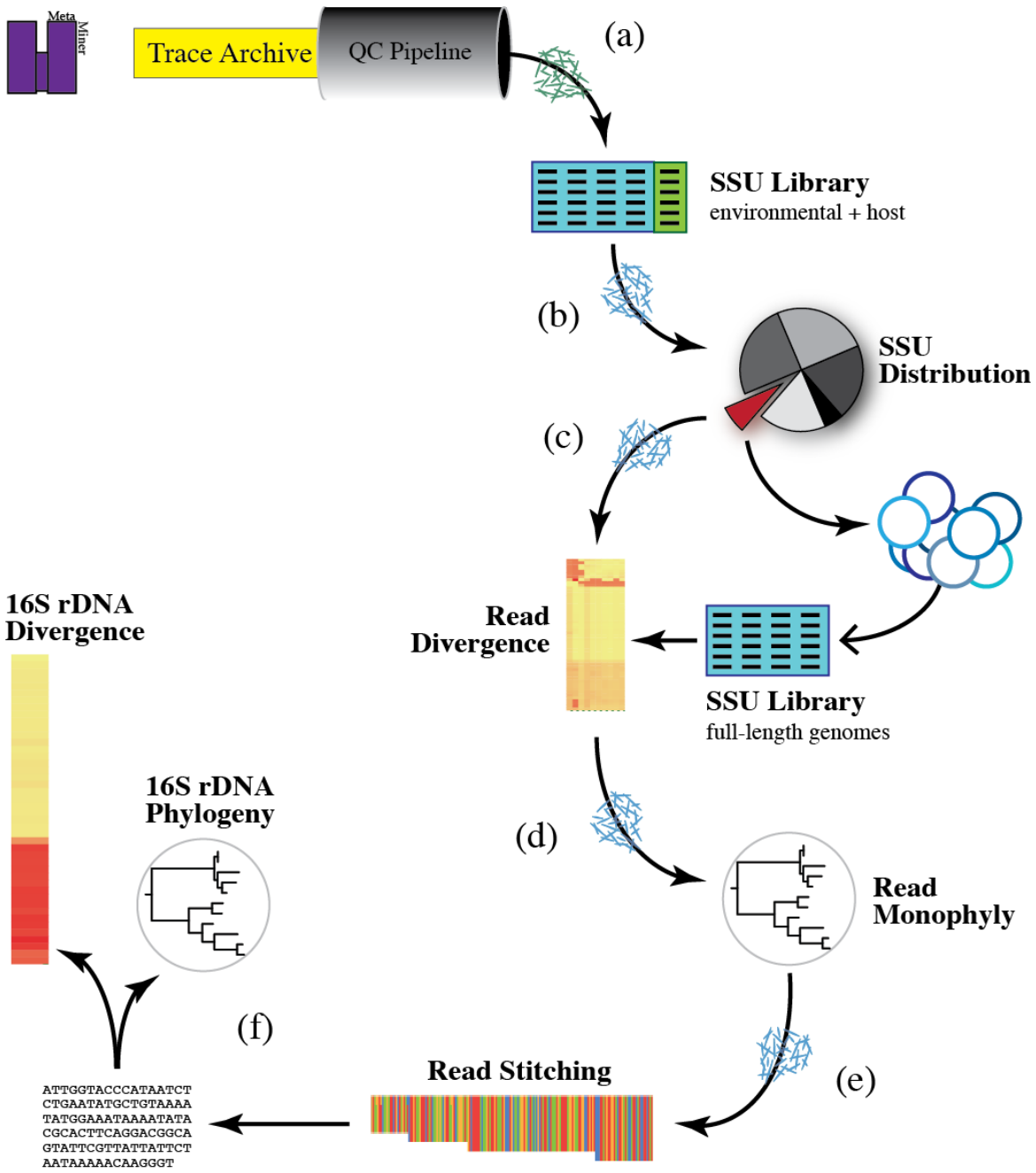


Figure 3. The MetaMiner workflow for identifying target microbes within host trace read data. (**a**) After quality control, trace reads are screened against a library of SSU sequences from bacteria plus the host. (**b**) Reads that match to the library are used to build a distribution and a target microbe is chosen. (**c**) Reads from the target are compared for minimal divergence from 16S rDNA sequences of related sequenced genomes and (**d**) tested for monophyly. (**e**) Minimally divergent, monophyletic reads are stitched together to construct a consensus target 16S rDNA. (**f**) Phylogeny estimation and divergence from 16S rDNA sequences from sequenced genomes are used to scope ReadMiner and AssemblySifter workflows.

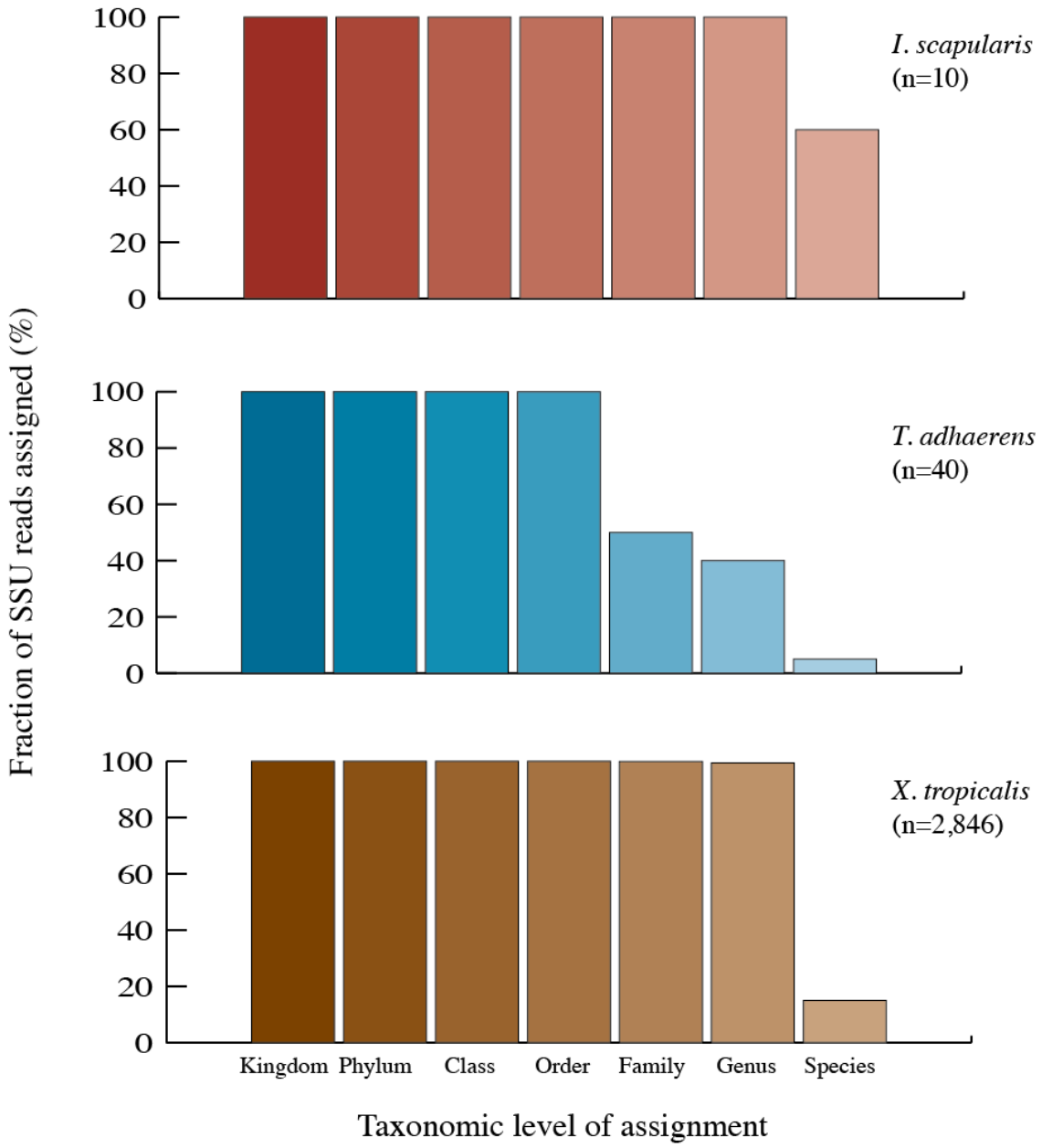


Figure 4. Assignment of taxonomy to bacterial 16S rDNA reads extracted from the genome sequence trace data sets of three hosts: *Ixodes scapularis* (top), *Trichoplax adhaerens* (middle), and *Xenopus tropicalis* (bottom). Assignments were made by transferring the most specific taxonomy for each match in the Greengenes database. See the text for a full description.

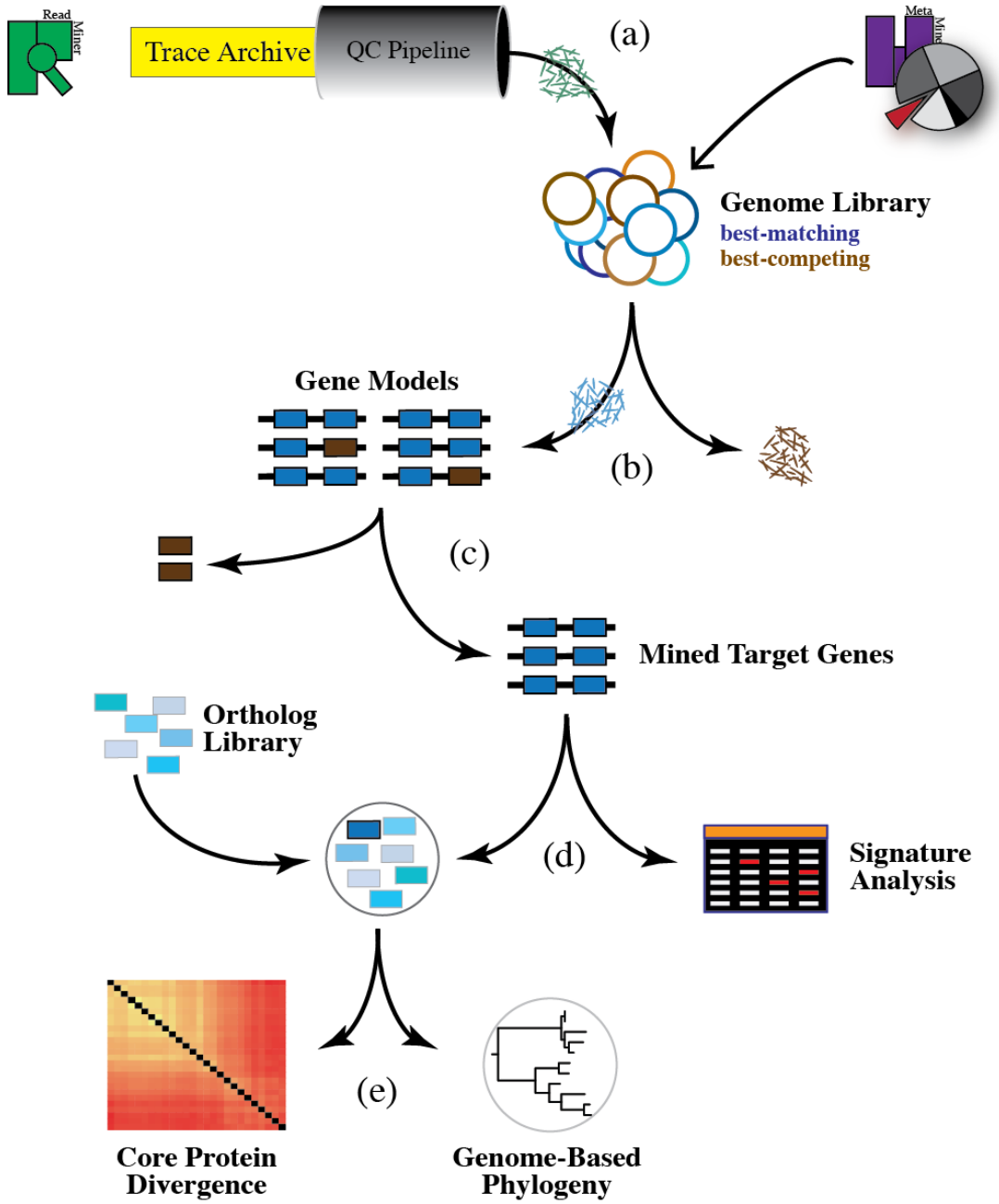


Figure 5. The ReadMiner workflow for extracting and characterizing target microbe sequences from within host trace read data. (**a**) After quality control, trace reads are screened against a library of genomes defined by MetaMiner. (**b**) Reads that match to genomes in the best-matching clade are assembled and gene models called. (**c**) Genes of appropriate length that are not exclusively orthologous to best-competing genes comprise the set of **mined target genes**. (**d**) Mined genes with widely distributed orthologs (**core genes**) are identified and compiled; other genes are subjected to specific signature and functional analysis. (**e**) Core genes are used to estimate phylogeny and divergence of the target.

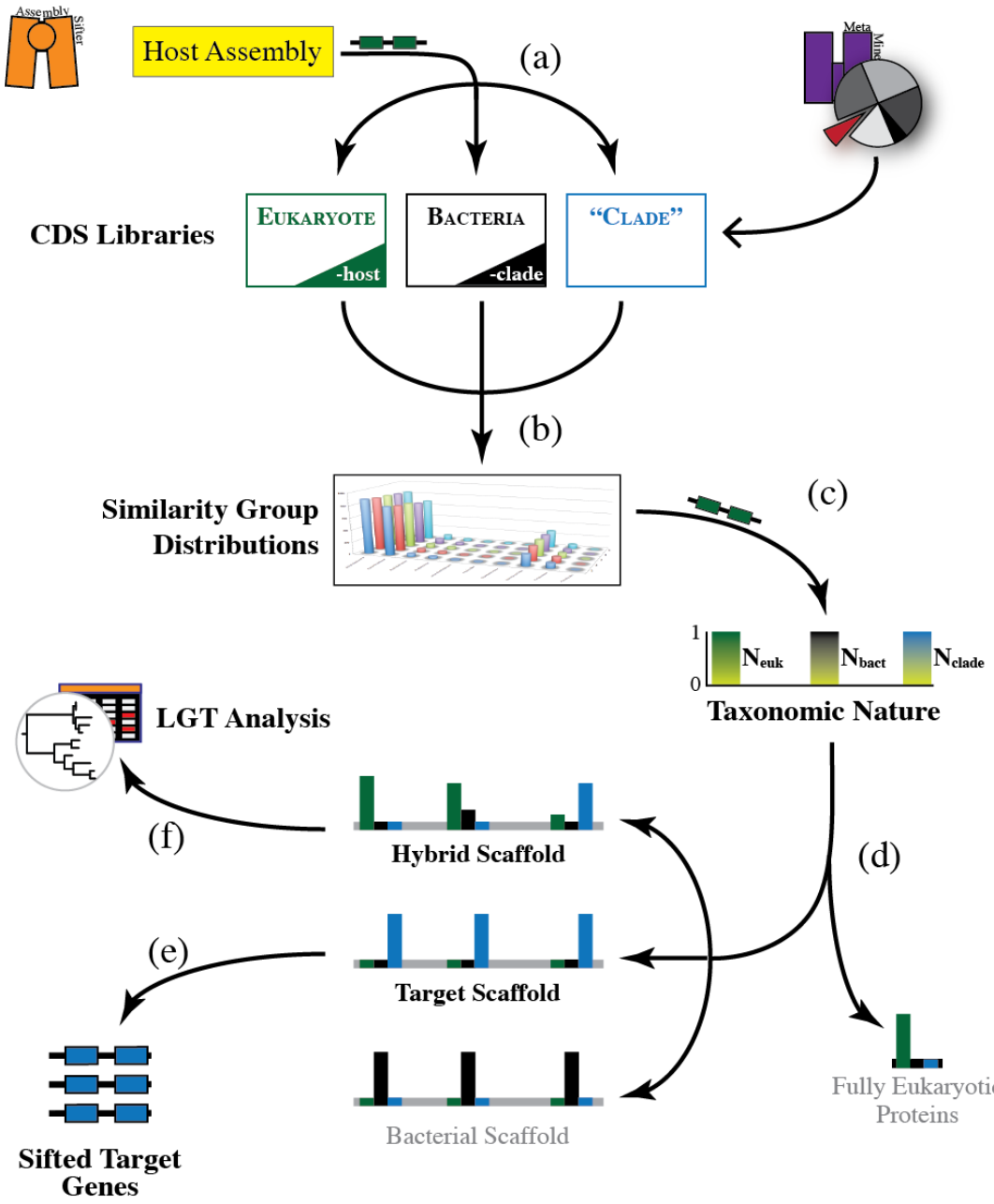


Figure 6. The AssemblySifter workflow for extracting and characterizing bacterial sequences from within host genome assembly data. (**a**) Each host protein is queried against three CDS libraries comprised of proteins from: 1) all eukaryotes except the host; 2) all bacteria except the "clade"; and 3) proteins from "clade" species as defined by MetaMiner. (**b**) Matches to each host protein are ranked by S_m score (see text). (**c**) Group distribution of the top five S_m scores are used to determine the taxonomic nature N_t of each protein. (**d**) Host proteins with any amount of bacterial nature are placed on their host scaffolds. (**e**) Short scaffolds comprised solely of fully "clade" proteins are analyzed as candidate sequences belonging to the target. (**f**) Long scaffolds comprised of mixtures of genes are evaluated as possible sites of bacteria-to-host LGTs.

CHAPTER 3. Bacterial DNA mined from the *Ixodes scapularis* (Wikel colony) genome project confirms the presence and genomic composition of its rickettsial endosymbiont.

ABSTRACT

The objective of this study was to demonstrate the effectiveness of MetaMiner and ReadMiner at identifying and extracting bacterial DNA from the sequencing reads of a eukaryotic genome project. To achieve this objective, a eukaryote was chosen (blacklegged tick, *Ixodes scapularis*) that is known to harbor a single endosymbiotic bacterium (*Rickettsia* endosymbiont of *Ixodes scapularis*, REIS) with a sequenced genome. Analysis of the *I. scapularis* sequencing reads using MetaMiner (excluding REIS) indicated the presence of 10 *Rickettsia*-like 16S rDNA reads, suggesting the presence of a *Rickettsia* genome at a sequencing depth of 4X. Assembly of the *Rickettsia*-like 16S rDNA reads produced a single contig, 1035 bases in length, with 100% identity to the 3' end of the known REIS 16S rDNA sequence. No other bacterial 16S rDNA sequences were discovered. Subsequent ReadMiner analysis of the *I. scapularis* sequence data with 22 (non-REIS) *Rickettsia* genomes yielded over 1.4 Mb in 655 contigs (N_{50}=2948). This represents approximately 65% of the sequenced REIS genome (including plasmids), or near complete coverage of the REIS genome outside of its diverse and numerous suites of mobile genetic elements. The results of this study demonstrate the usefulness of MetaMiner and ReadMiner in extracting and characterizing bacterial sequences from eukaryotic sequencing projects.

INTRODUCTION

The blacklegged tick (*Ixodes scapularis*) is an arthropod vector of several important pathogenic microbes, including *Anaplasma phagocytophilum* (human granulocytic anaplasmosis), *Bartonella* spp. (bartonellosis), *Babesia microti* (babesiosis), and *Borrelia burgdorferi* (Lyme disease) (**Healy et al. 1976; Eskow et al. 2001**), as well as numerous other symbiotic Rickettseaceae and Anaplasmataceae (**Benson et al. 2004**). A draft of the complete genome of *I. scapularis* using WGS sequencing has been released recently (GenBank accession NZ_ABJB000000000), deriving from heterogeneous tissue sampled from a colony of ticks inbred for twelve generations (Wikel colony). The tick genome assembly revealed the presence of a heretofore uncharacterized species of *Rickettsia*, the *Rickettsia* endosymbiont of *Ixodes scapularis* (REIS), that was extracted from the *I. scapularis* genome sequencing project data, assembled, and characterized in detail previously (**Gillespie et al. 2012**).

The current study was undertaken to evaluate the overall distribution of bacterial residents within the *I. scapularis* genome sequencing project, confirm the presence of REIS, and measure the general efficacy of the MetaMiner and ReadMiner workflows at extracting bacterial sequence from within a host.

METHODS

Data Preparation

A total of 10,707,500 WGS reads for the *Ixodes scapularis* (Wikel colony) genome project were downloaded from the Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi) at NCBI. Reads were decontaminated and validated as described in Chapter 2, using `blastn` and UniVec since exact cloning vector sequences were not available, resulting in 10,412,491 decontaminated

reads. Bases at the 5' and 3' ends of reads were trimmed so that the mean Phred (quality) score at all positions was 25 or higher. Finally, 1,359,513 reads with mean Phred scores below 25, and 385 additional reads shorter than 100 bases were removed, producing approximately $9x10^6$ reads that were subsequently analyzed for bacterial sequences.

SSU rDNA Analyses

MetaMiner was used to assess the distribution of bacterial sequences in the *I. scapularis* read data, identify a bacterial target, and assemble and analyze a target 16S rDNA sequence. A detailed description of MetaMiner can be found in Chapter 2. Briefly, prepared *I. scapularis* reads were mapped against 406,997 full-length bacterial 16S rDNA sequences from the 2012 release of the GreenGenes database (**DeSantis et al. 2006**), plus 18S SSU rDNA sequences from 76 *Ixodes* species obtained from NCBI (an *I. scapularis* 18S rDNA sequence was not available at the time of the study). Reads that had at least one match in this SSU rDNA library were binned according to the taxonomic classification of their matches. Reads that mapped to *Ixodes* spp. 18S rDNA sequences were used to calculate approximate coverage of the host genome and then removed.

Reads that mapped to *Rickettsia*-like 16S rDNA sequences were compared for gene divergence against full-length 16S rDNA sequences from 42 sequenced Rickettsiales genomes (excluding REIS) plus *Brucella suis* 1330, from the November 2012 release of PATRIC (**Gillespie et al. 2011**). For each read, an unpaired Student's t-test was used to evaluate the statistical significance of differences in mean divergence from the rickettsial and non-rickettsial genomes. Analysis of variance (ANOVA) was used to determine the statistical significance of differences in mean read divergence from individual genomes. The monophyly of the reads was evaluated with phylogeny estimation using `RAxML` (**Stamatakis 2006**) as described above. Minimally divergent, monophyletic reads were subsequently aligned and stitched together to construct a 16S rDNA sequence for the proposed rickettsial target. To assess the performance of MetaMiner, the target 16S rDNA sequence was aligned to the existing REIS sequence using the `needle` program from EMBOSS (**Rice et al. 2000**).

Analyzing Rickettsial-like Host Reads

Two separate ReadMiner analyses were used to extract target sequences from the host reads. A detailed description of ReadMiner can be found in Chapter 2. In the first analysis, 83 complete and WGS Rickettsiales genomes from the November 2012 release of PATRIC (excluding REIS) were used as the best-matching clade. In the second analysis, 47 complete and WGS *Rickettsia* spp. genomes from PATRIC (excluding REIS) were used. For comparison, a third extraction was run using the REIS genome alone and a very high match stringency (m=500). No best-competing clade was used in any analysis. Sequences from each run were assembled separately using `Mira` (**Chevreux 2005**) as described, and mined target contigs were aligned to the REIS genome using `megablast` and an e-value threshold of 10^{-6}. Contigs that did not align to REIS were queried against the NCBI *refseq_genomic* database for the top 5 matches (e-value threshold 10^{-3}). For each analysis, an estimate of the precision was obtained from the total length of mined contigs that matched or didn't match to the REIS genome (e-value < 10^{-6}):

$$P = \frac{N_{hit}}{N_{hit} + N_{miss}} \quad (1)$$

Where:

P is the precision
N_{hit} is the total length of mined contigs that matched to REIS
N_{miss} is the total length of mined contigs that did not match to REIS

Similarly, an estimate of the recall (sensitivity) was calculated from the mined contigs that matched to REIS and the total length of the REIS genome:

$$R = \frac{N_{hit}}{N} \quad (2)$$

Where

R is the recall
N is the total length of the known REIS genome

RESULTS & DISCUSSION

From an initial set of 1.1×10^7 *I. scapularis* genome sequencing project reads, 85.5% (9.1×10^6) survived the decontamination and quality control pipeline. The final set had a mean read length of 644.5 bases and included 5.8×10^9 total bases. Full quality analysis results from `fastqc` can be found in **appendix B1** (before processing) and **appendix B2** (after processing). An atypical drop in the mean sequence quality per base around position 40 was noted, but ultimately deemed acceptable because of its short length and internal location.

Reconstruction of the REIS 16S rDNA

Minimally divergent reads. A slightly abridged version of the full MetaMiner workflow was used to assess the distribution of bacterial 16S rDNA sequences in the *I. scapularis* project reads. 8,248 reads matching *Ixodes* spp. 18S rDNA were identified, along with ten reads matching 16S rDNA from four different *Rickettsia* spp. (**table 1**). Pairwise gene divergence comparisons among these *Rickettsia*-like reads and 16S rDNA sequences from known rickettsial species are shown in **fig. 1** and **fig. 2**. REIS is shown for comparison, but was analyzed separately from all other genomes. All ten reads had significantly lower mean divergence ($p<0.001$) from *Rickettsia* spp. compared to non-*Rickettsia* spp. of Rickettsiales (**fig. 2a**). In general, the mean divergence from *Rickettsia* spp. genomes mirrored the divergence from REIS, whereas divergence from non-*Rickettsia* spp. genomes resembled *B. suis*. Analysis of variance ($p<0.001$) of mean read divergence to individual species (**fig. 2b**) revealed no significant difference in divergences among the *Rickettsia* spp., or among the non-*Rickettsia* spp. (including *B. suis*); however, there was a significant difference between *Rickettsia* spp. and non-*Rickettsia* spp. Removing *B. suis* (not in the Rickettsiales), or *Orientia tsutsugamuchi* (which showed intermediate divergence), or both from the non-*Rickettsia* spp. group did not affect the results. Taken together, these results suggest that the ten *Rickettsia*-like reads likely originate specifically from a *Rickettsia* sp. genome.

It is perhaps noteworthy that read divergence from REIS was non-zero for 70% of the *Rickettsia*-like reads, and generally indistinguishable from read divergence compared to other *Rickettsia* spp. genomes (see **fig. 1**). This seems at odds with the hypothesis that these reads arise from REIS, which implies they should not diverge at all from the REIS 16S rDNA sequence. Small divergences may be a reflection of underlying sequencing errors, including incorporation of incorrect nucleotides by the polymerase or mistakes in base-calling, especially near the read

termini. Such errors might be amplified in the divergence score if the region of overlap between a read and it's bait sequence is short. Such reads, despite their errors, would be expected to align well to a closely-related, full-length 16S rDNA sequence over most of their length. In contrast, more divergent reads may not be derived from the 16S rDNA at all, but incidentally pulled in because of short stretches of local similarity to a bait sequence. These reads would not be expected to align well to a full-length 16S rDNA. It is also possible that highly-divergent reads constitute part of a separate *Rickettsia*-like 16S rDNA sequence within the *I. scapularis* reads; however, the presence of two distinct, closely-related *Rickettsia* spp. occupying the same intracellular niche seems unlikely.

Monophyly of Rickettsia-like reads. Two of the *Rickettsia*-like 16S reads (ti=1133911933 and ti=1193421814) displayed consistently high inter-read divergence (**fig. 1**) and poor sequence alignment to *Rickettsia massiliae* (**fig. 3**), and were removed before phylogeny estimation. A third read (ti=1680655836) exhibited variable inter-read divergence, but alignment to *R. massiliae* indicated it aligned well to a region of the 16S rDNA that was not well-covered by the other reads; consequently, it was left in the analysis. The eight *Rickettsia*-like 16S reads with minimal divergence, plus 43 additional 16S rDNA sequences including sequenced Rickettsiales genomes and *B. suis* 1330, were used to construct a phylogenetic tree (**fig. 4a**). All reads were monophyletic in the tree, supporting a common origin for these sequences. In addition, the clade containing the reads was embedded among the Spotted Fever Group sequences, similar to the placement of REIS as previously described (**Gillespie et al. 2012**). Repeating the phylogenetic analysis including the REIS 16S rDNA sequence yielded a similar tree (**fig. 4b**), with REIS disrupting the monophyly of the reads as expected.

Construction and validation of the 16S rDNA sequence. Based on monophyly and low pairwise divergence from other *Rickettsia*, all eight *Rickettsia*-like 16S reads were aligned to *R. massiliae* and stitched together into a single 16S rDNA contig. The 1035-base consensus sequence was shown to align to the 3' end of the existing REIS sequence with 100% identity. As a result, it was concluded that this consensus sequence represents the REIS 16S rDNA. Since the phylogeny of REIS has been described previously (**Gillespie et al. 2012**), it was not repeated here.

Determination of the best-matching clade. In order to identify a suitable best-matching clade for ReadMiner, pairwise divergence was calculated between the mined REIS 16S rDNA sequence and the same 43 sequences (Rickettsiales plus *B. suis*) used to determine read monophyly (**fig. 5**). Mean divergence from 25 *Rickettsia* spp. sequences (excluding REIS) was 1.1%, significantly lower (Student's t-test; $p<0.001$) than the 16.8% mean divergence from 17 non-*Rickettsia* spp. sequences (excluding *B. suis*). In conclusion, MetaMiner identified the genus *Rickettsia* as the best-matching clade for the target, based on pairwise divergence and phylogeny of the mined 16S rDNA sequence. This is in agreement with the known systematic position of REIS within the Rickettsiaceae.

Analysis of Mined Rickettsia-like Reads

Assessing extraction efficiency. To assess the effect of the best-matching clade composition on the efficiency of ReadMiner, separate analyses were performed using two different clades: 1) 83 sequenced Rickettsiales genomes (except REIS); and 2) 43 sequenced *Rickettsia* spp. genomes (except REIS). A third run using the REIS genome alone was included for comparison. Each set of mined reads was assembled separately into contigs and compared against the REIS genome to

estimate the efficiency of the extraction (**fig. 6a**) and assembly (**fig. 6c**). Precision (the proportion of mined bases that truly belong to the target genome) increased slightly as the scope of the best-matching clade narrowed from all Rickettsiales (92.5%) to just *Rickettsia* spp. (99.9%). In contrast, recall (the proportion of actual target bases that were mined successfully) was consistent for both Rickettsiales and *Rickettsia* spp. (72%). When REIS alone was used as bait, there were no false positives and recall jumped to 86%. Ideally, the recall when using REIS alone as bait should be 100%. One explanation for the discrepancy is that the assembly process was not optimized, and some reads may not be included in the contigs. Another possibility is that some of the reads that comprise the true REIS genome were removed during the data preparation pipeline. This is to be expected, given the tremendous amount (~35% of the REIS genome) of mobile genetic elements identified within the REIS genome (**Gillespie et al. 2012**).

It is important to note that precision and recall used here assume the current REIS genome is complete, which may not be true; therefore, it is likely that the recall values in particular are slightly overestimated. To some extent, though, precision is arguably more important than recall in ReadMiner, since the primary goal of this workflow is not necessarily to assemble a complete genome, but to extract enough clean sequence to characterize the target using phylogeny and signature genomic elements.

GC content and novel REIS sequences. Percent GC content (%GC; **fig. 6b**) across all mined contigs that matched to REIS was consistent between runs (32.5%), and identical to the %GC for the REIS genome. This is substantially different from the *I. scapularis* %GC (45%), and highlights the difference between the two genomes. The three non-matching contigs found using *Rickettsia* spp. also had an REIS-like %GC (33.2%), in contrast to the non-matching contigs found using all Rickettsiales (49%). These three REIS-like, non-matching contigs (1818 bases) were queried against the NCBI *refseq_genomic* database using `megablast` to determine their possible origins; notably, all of the top hits to each contig were to the genomes of *Rickettsia* spp. (**table 2**). These results suggest that these three contigs may represent novel regions of the true REIS genome.

Characterizing missed sequences. The publication of the REIS genome provides a ready comparative tool for identifying and characterizing genomic sequences missed by ReadMiner. To this end, the mined contigs from the Rickettsiales, *Rickettsia* spp., and REIS extractions were plotted against the REIS genome (**fig. 7**). Two contiguous regions of REIS in particular were absent from the contigs mined using either Rickettsiales or *Rickettsia* spp. as bait. Missed region 1 (Mr1; NZ_CM000770:1.27-1.34 Mb) was 68.6 Kb long with a %GC of 31.8% and 65 annotated genes. It corresponds to a stretch of the REIS genome devoid of core rickettsial genes, but including a conjugative transfer (*tra*) operon, 28 hypothetical proteins, and a transposase. It is likely that Mr1 is part of a mobile genetic element (MGE) integrated into the REIS genome, and was missed by ReadMiner due to a lack of rickettsial signal in this region. Missed region 2 (Mr2; NZ_CM000770:1.57-1.61 Mb) was shorter, only 34.0 Kb long with a %GC of 32.0% and 28 annotated genes. Mr2 also likely represents part of an MGE, lacking any core rickettsial genes but containing 18 hypotheticals along with at least six *tra* genes. It is worth noting that Mr1 and Mr2 are not the only contiguous stretches of non-core genes in this REIS scaffold, for example a 0.5 Mb region near 0.25 Mb and an extended 1.3 Mb stretch near 0.66 Mb; however, none of these other regions (nor any others on the primary scaffold) were notably troublesome for

ReadMiner, suggesting they may be similar enough to rickettsial DNA to be amenable to extraction.

CONCLUSION

The results of this study demonstrate the effectiveness of MetaMiner at extracting enough information via 16S rDNA analysis to initially characterize bacterial residents of heterogeneous sequence read data. In addition, it was shown here that mining read data using ReadMiner and a well-supported set of sequenced genomes allows the assembly of a significant fraction of a target bacterial genome with high precision and moderately high recall. Regions of bacteria-to-bacteria LGT, while potentially an issue if they originated from unrelated bacteria, were not as problematic as expected given the prevalence of MGEs in the REIS genome.

LITERATURE CITED

Benson MJ, Gawronski JD, Eveleigh DE, Benson DR. 2004. Intracellular symbionts and other bacteria associated with deer ticks (*Ixodes scapularis*) from Nantucket and Wellfleet, Cape Cod, Massachusetts. Appl Environ Microbiol. 70:616–620. doi: 10.1128/AEM.70.1.616-620.2004.

Chevreux B. 2005. MIRA: An Automated Genome and EST Assembler. Ph.D. Thesis. 1–171.

DeSantis TZ et al. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 72:5069–5072.

Eskow E, Rao R-VS, Mordechai E. 2001. Concurrent infection of the central nervous system by *Borrelia burgdorferi* and *Bartonella henselae*: evidence for a novel tick-borne disease complex. Archives of neurology. 58:1357.

Gillespie JJ et al. 2012. A *Rickettsia* genome overrun by mobile genetic elements provides insight into the acquisition of genes characteristic of an obligate intracellular lifestyle. J Bacteriol. 194:376–394. doi: 10.1128/JB.06244-11.

Gillespie JJ et al. 2011. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. Infect Immun. 79:4286–4298. doi: 10.1128/IAI.00207-11.

Healy G, Speilman A, Gleason N. 1976. Human babesiosis: reservoir in infection on Nantucket Island. Science. 192:479–480. doi: 10.1126/science.769166.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16:276–277.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22:2688–2690. doi: 10.1093/bioinformatics/btl446.

FIGURES

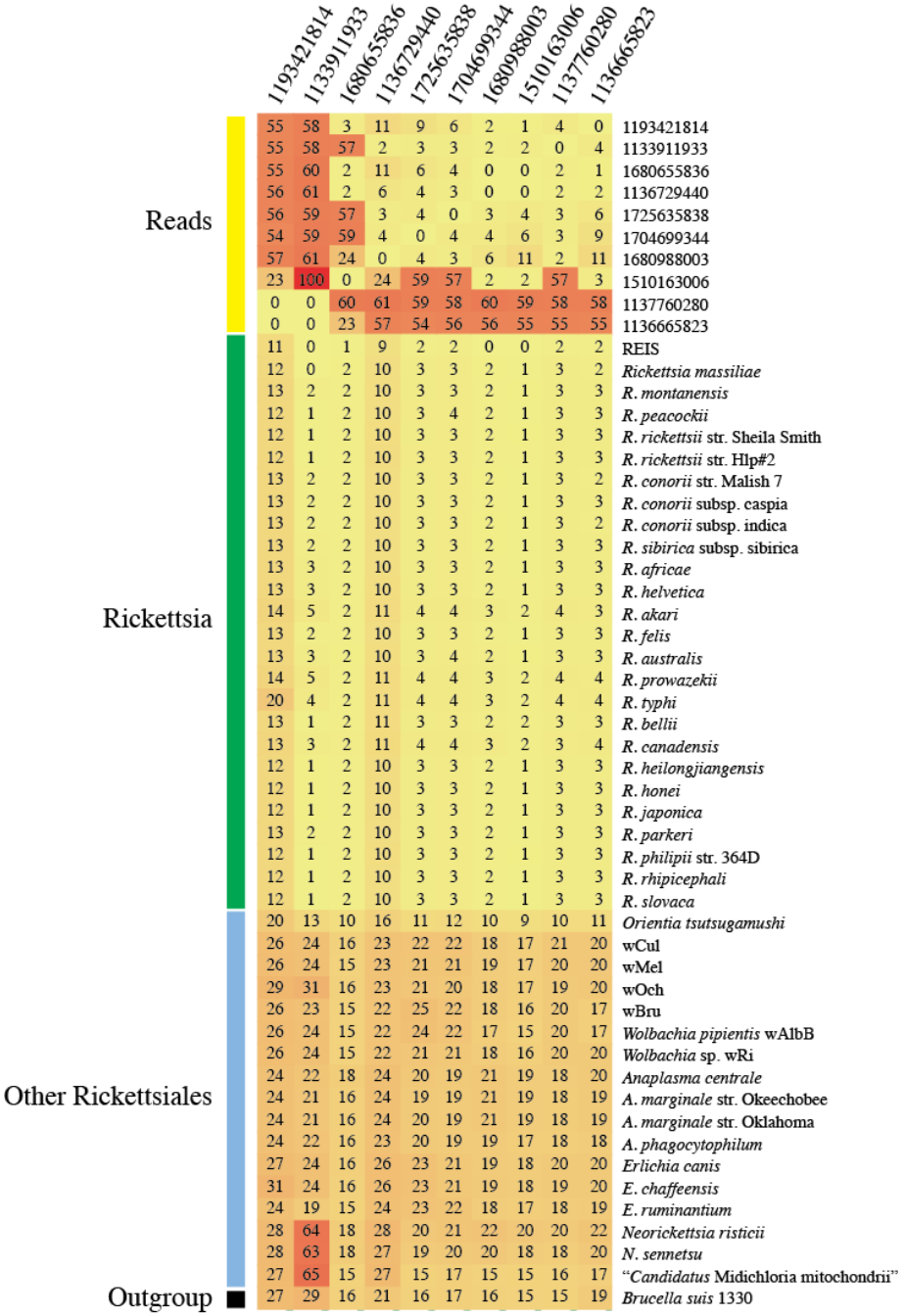


Figure 1. Pairwise divergence between 10 *Rickettsia*-like reads and full-length Rickettsiales 16S rDNA sequences. The green bar indicates *Rickettsia* spp., the blue bar indicates non-*Rickettsia* Rickettsiales, and the black bar indicates *Brucella suis* (outgroup). Read ids are displayed across the top and also in the first ten rows (yellow). Each cell contains the pairwise divergence between the corresponding sequences, colored from lowest (yellow) to highest (red) divergence.

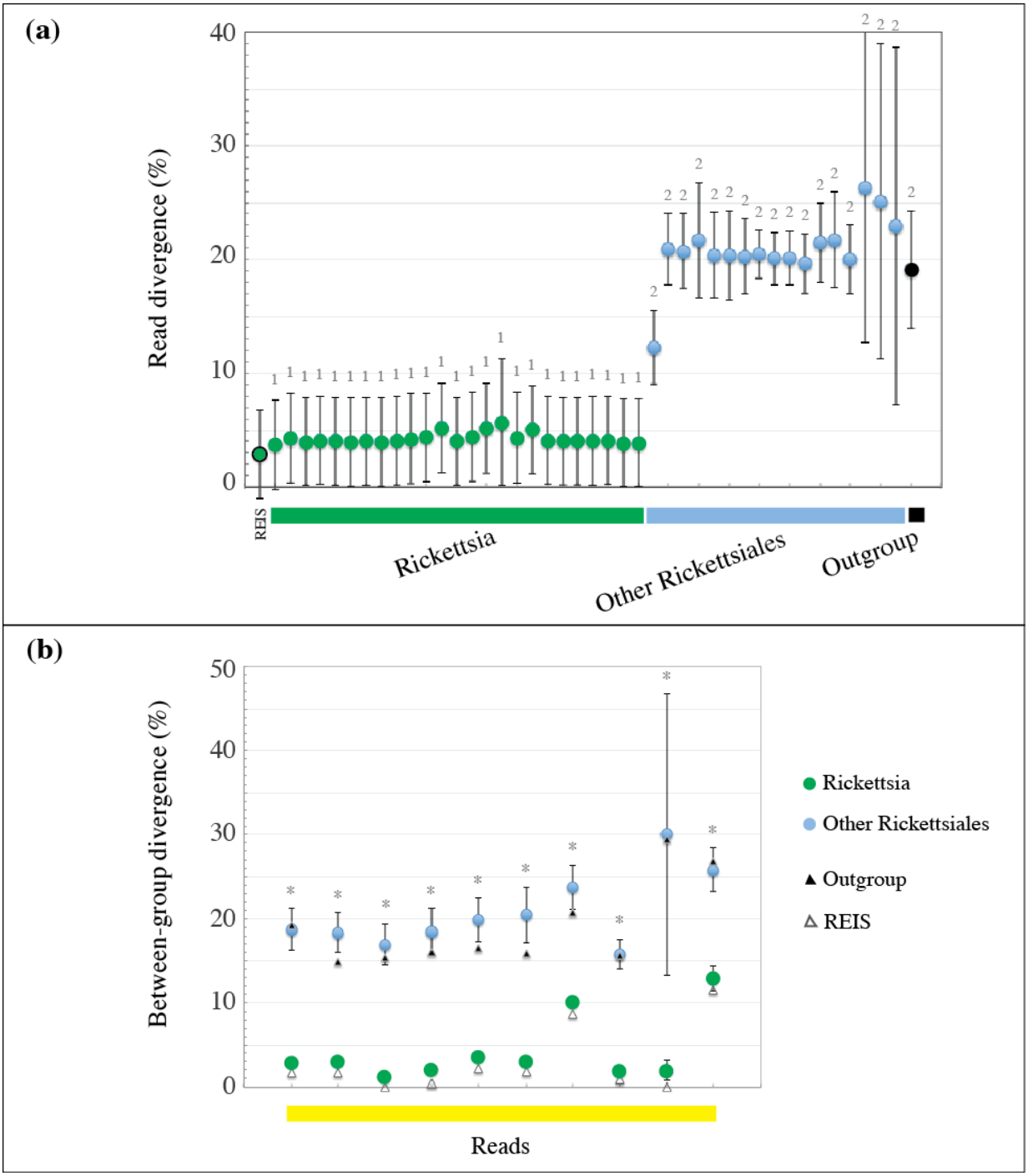


Figure 2. Mean read and between-group divergence for 10 mined *Rickettsia*-like 16S rDNA reads. (**a**) Mean read divergence from full-length 16S rDNA sequences of *Rickettsia* (green) and non-*Rickettsia* Rickettsiales (blue), plus *Brucella suis* (black). Numbers indicate results of an ANOVA at p<0.001. (**b**) Mean divergence of each mined read from *Rickettsia* (green) and non-*Rickettsia* Rickettsiales (blue) sequences. An asterisk indicates the between-group difference was significant (p<0.001). *B. suis* (black) and REIS (white) are shown for comparison.

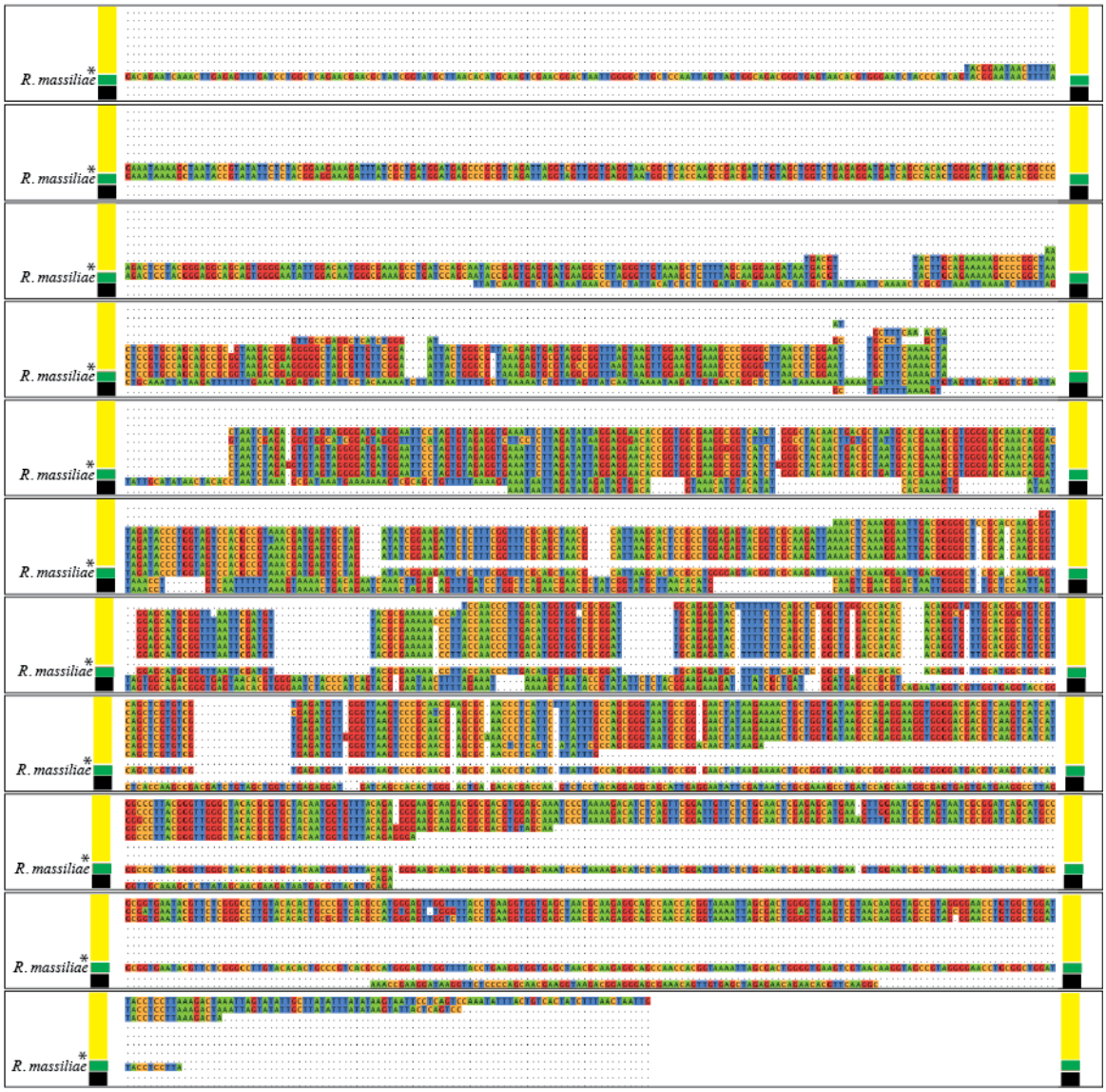


Figure 3. Alignment of ten *Rickettsia*-like 16S rDNA reads to the full-length *Rickettsia massiliae* 16S rDNA sequence (green). Reads above the *R. massiliae* sequence (yellow) were included in the subsequent estimation of monophyly. Reads below the sequence (black) aligned poorly and were discarded. Read ti-1680655836 (*) showed variable inter-read divergence but aligned well, so it was retained.

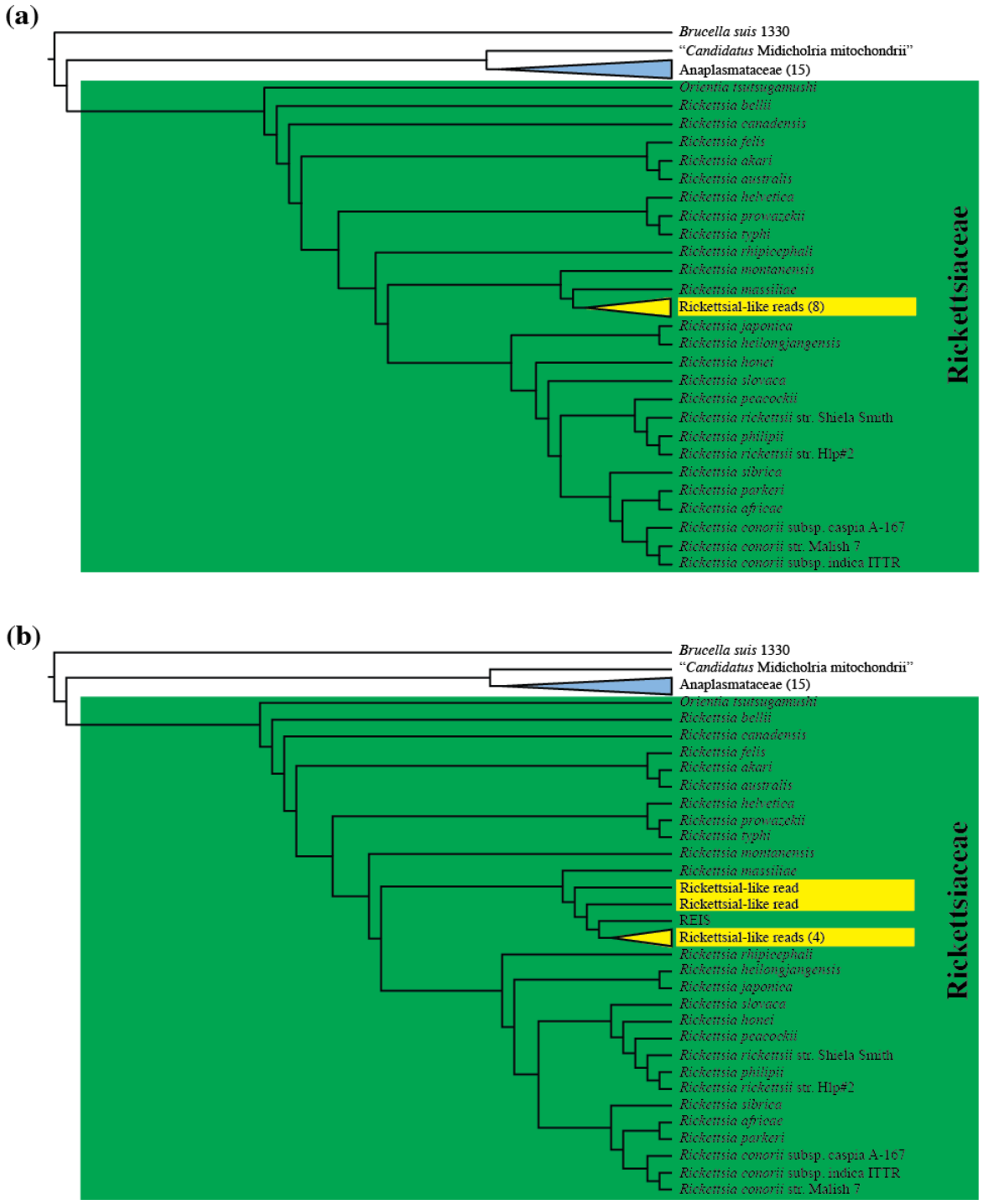


Figure 4. Phylogeny estimation of *Rickettsia*-like 16S rDNA reads from the *I. scapularis* genome trace data. Trees include all reads (yellow) plus 16S rDNA sequences from Rickettsiaceae (green), non-Rickettsiaceae Rickettsiales (blue), and *Brucella suis* 1030 (outgroup). (**a**) Phylogeny excluding REIS. (**b**) Phylogeny including REIS. See the text for more information.

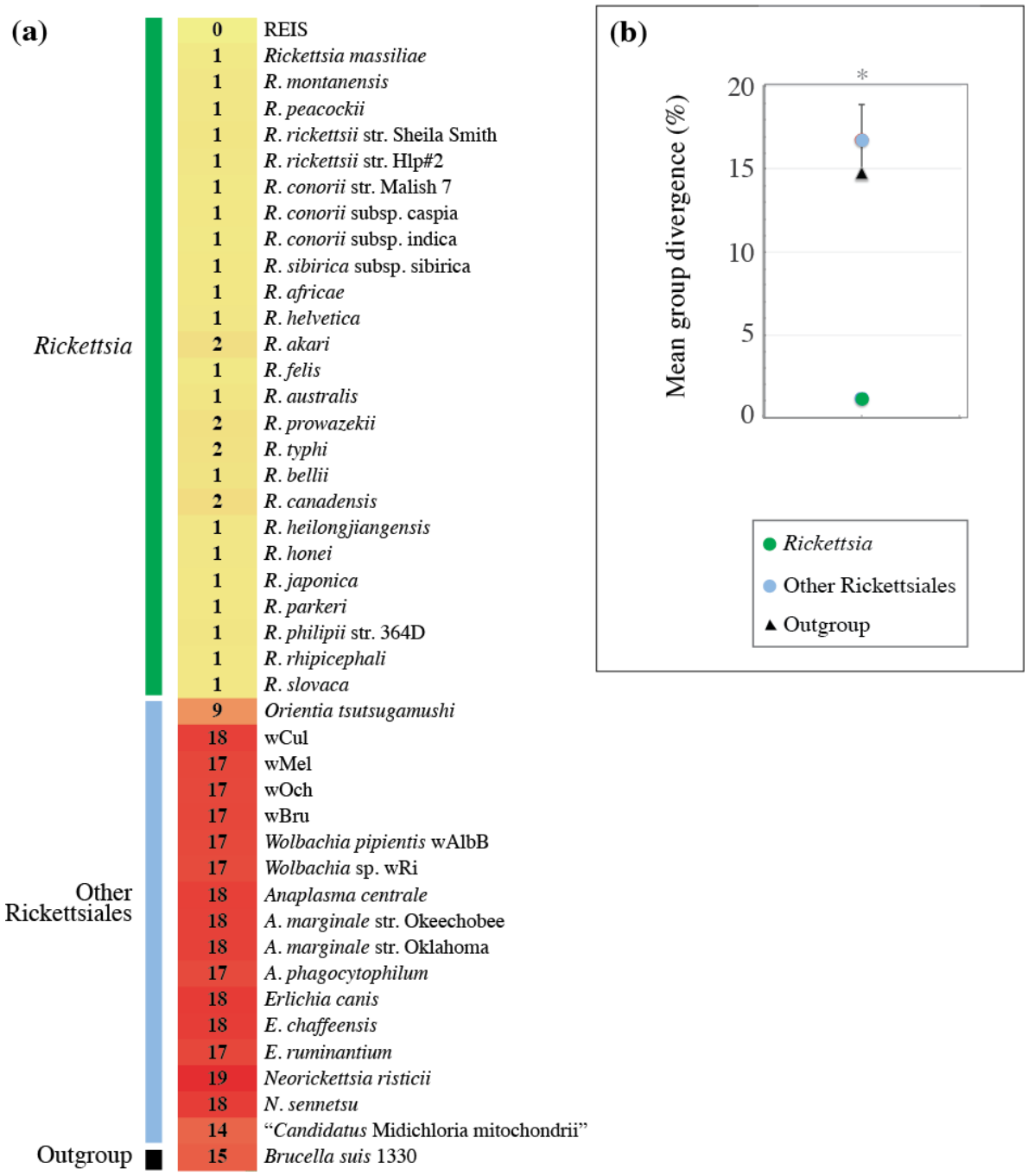


Figure 5. Pairwise divergence between the mined REIS 16S rDNA and full-length Rickettsiales sequences. (**a**) Heatmap coloring divergence from low (yellow) to high (red). The green bar indicates Rickettsia spp., the blue bar non-*Rickettsia* Rickettsiales, and the black bar *Brucella suis*. The full-length original REIS is at the top. (**b**) Mean divergence of the mined REIS from species of *Rickettsia* (green) and non-*Rickettsia* Rickettsiales (blue). The asterisk indicates the between-group difference is significant ($p<0.001$). *B. suis* is shown for comparison (black).

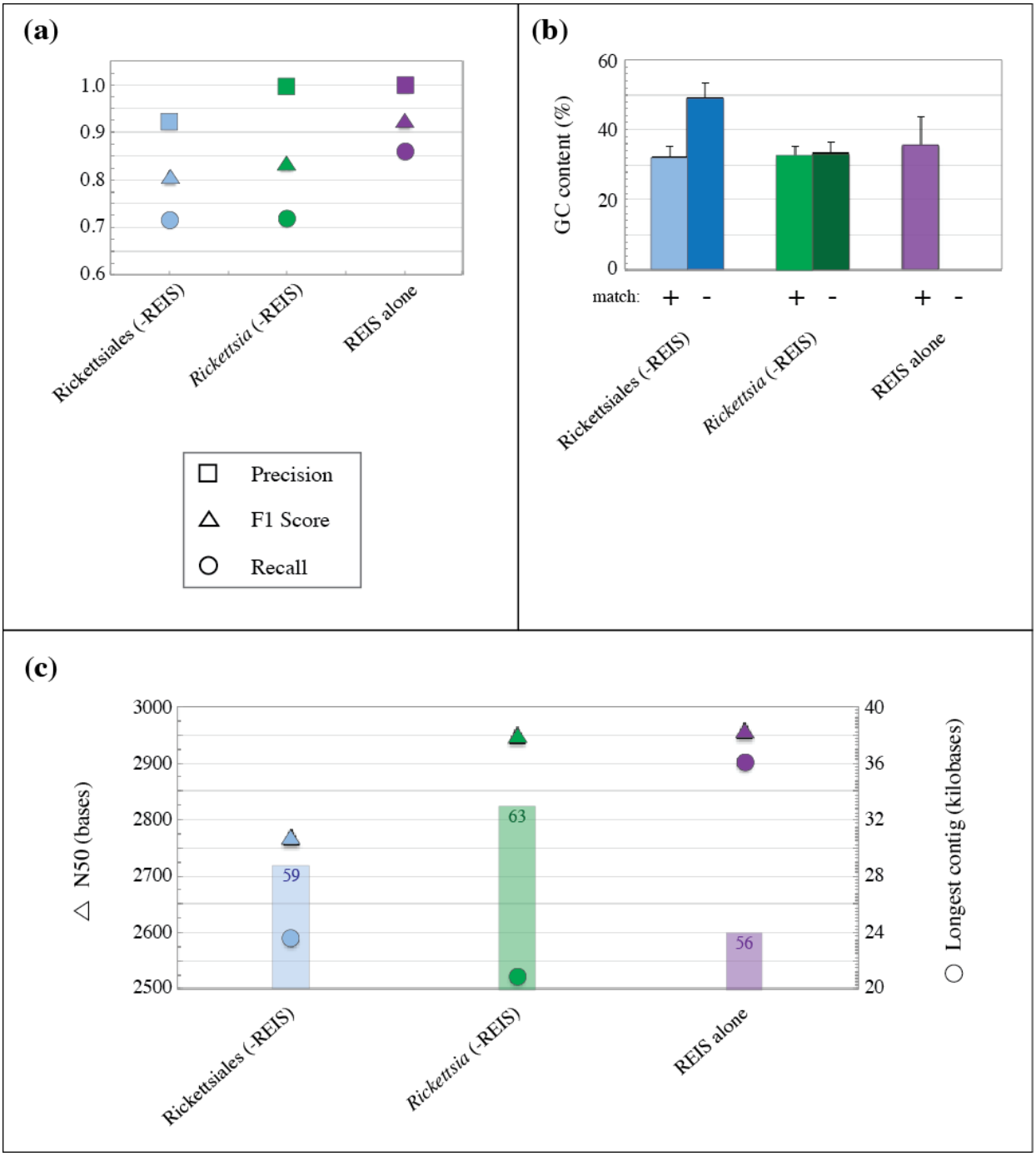


Figure 6. Extraction efficiency of ReadMiner. (**a**) Precision (squares), recall (circles), and F1 score (triangles) for contigs mined using all Rickettsiales (blue), *Rickettsia* spp. (green), or REIS alone (purple). (**b**) Mean %GC of mined contigs that matched (+) and did not match (-) the REIS genome. Colors as in (**a**). (**c**) Assembly N50 (triangles), longest contig (circles), and consensus quality (bars and numbers) for each ReadMiner analysis. Colors as in (**a**).

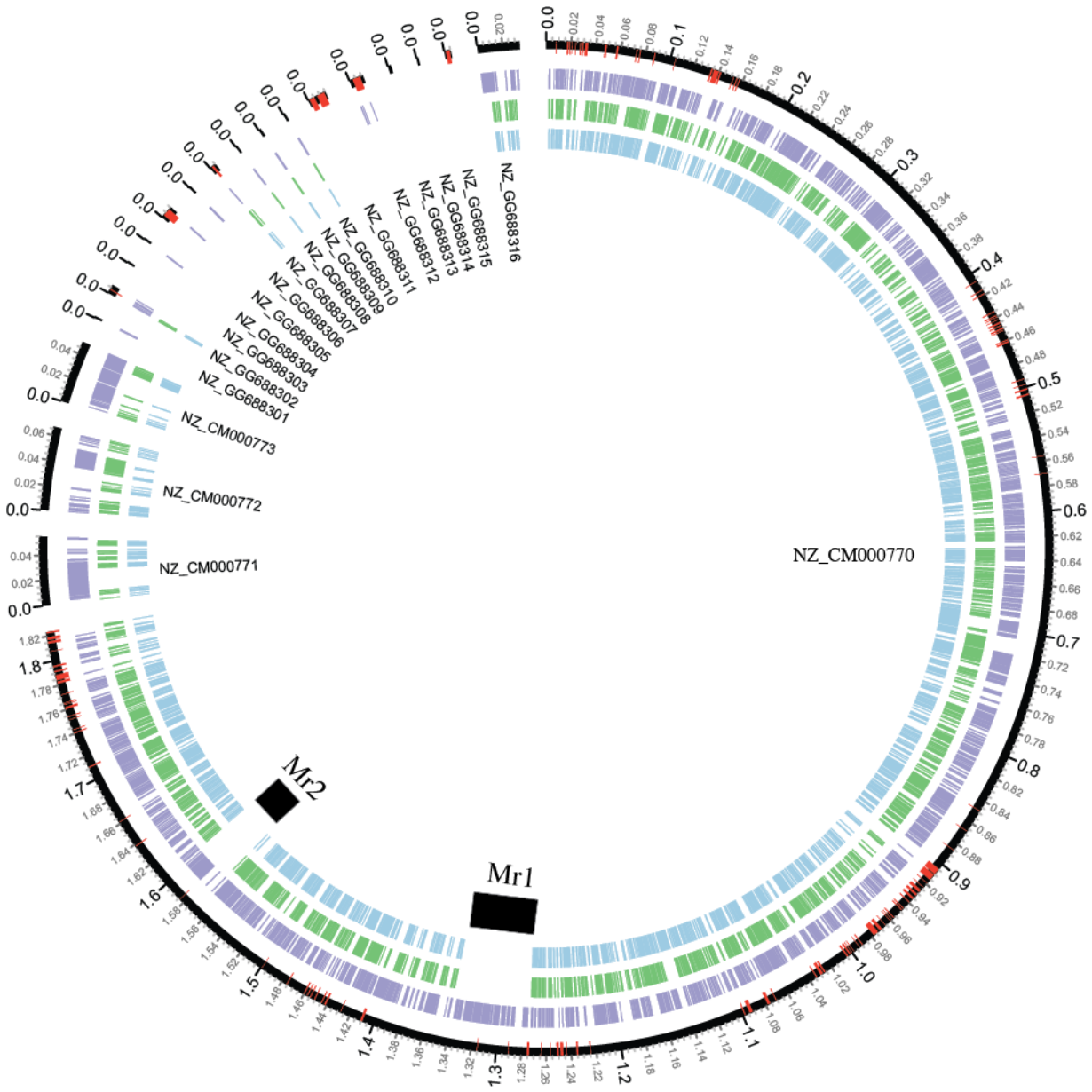


Figure 7. Circle plot of the REIS genome (black circle) in relation to ReadMiner contigs mined using all Rickettsiales (blue), just *Rickettsia* spp. (green), or REIS alone (purple) as bait. Gaps in the REIS assembly are shown in red. Missed region (Mr) 1 and Mr2 are shown as black bars in the interior of the plot. See the text for more information.

TABLES

Table 1. Distribution of all bacterial 16S rDNA reads extracted by MetaMiner from the *Ixodes scapularis* genome project trace data.

Taxonomy match	Greengenes ID	Reads
Rickettsia monacensis str. IrR/Munich	145296	5
Rickettsia sp. str. AT1	65333	3
Rickettsia conorii str. Malish 7	44963	1
new *Rickettsia* species, *Ixodes pacificus* ticks, USA: California Napa Valley clone 11122	340040	1

Table 2. Top-ranking BLAST matches to three contigs mined from the *Ixodes scapularis* genome trace reads using *Rickettsia* genomes that did not match to the REIS genome. %ID: percent identity of the alignment. %Q: percent of the query (contig) that aligned.

Contig (Q)	Genome of match (S)	E-value	%ID	%Q
is-rick_c545	*Rickettsia rhipicephali* str. 3-7-female6-CWPP	0	96.7	99.8
	Rickettsia massiliae str. AZT80	0	96.7	99.8
	Rickettsia japonica YH	0	96.7	99.8
	Rickettsia slovaca str. D-CWPP	0	96.5	99.8
is-rick_c478	*Rickettsia helvetica* C9P9	0	96.0	100
	Rickettsia felis URRWXCal2	0	95.1	100
	Rickettsia philipii str. 364D	0	95.1	99.8
	Rickettsia massiliae MTU5	0	95.1	99.8
is-rick_c574	"*Candidatus* Rickettsia amblyommii" str. GAT-30V	4e-103	93.2	46.3
	Rickettsia africae ESF-5	2e-101	92.8	46.3
	Rickettsia helvetica C9P9	2e-100	92.5	46.8
	Rickettsia felis URRWXCal2	8e-95	91.3	46.3
	"*Candidatus* Rickettsia amblyommii" str. GAT-30V	2e-36	78.0	46.5

CHAPTER 4. Bacterial DNA sifted from the *Trichoplax adhaerens* (Animalia: Placozoa) genome project reveals a putative rickettsial endosymbiont.

MANUSCRIPT

The manuscript presented in Chapter 4 was published under open access in the journal Genome Biology and Evolution on April 4, 2013. Supplementary tables S1-S5 and figures S6-S9 from the manuscript can be found in **appendix C**. Information relevant to this dissertation that is not in the manuscript, including trace read quality control results, can be found in **appendix D**.

The original manuscript is available online at: http://gbe.oxfordjournals.org/content/5/4/621

ABSTRACT

Eukaryotic genome sequencing projects often yield bacterial DNA sequences, data typically considered as microbial contamination. However, these sequences may also indicate either symbiont genes or lateral gene transfer (LGT) to host genomes. These bacterial sequences can provide clues about eukaryote-microbe interactions. Here, we used the genome of the primitive animal *Trichoplax adhaerens* (Metazoa: Placozoa), which is known to harbor an uncharacterized Gram-negative endosymbiont, to search for the presence of bacterial DNA sequences. Bioinformatic and phylogenomic analyses of extracted data from the genome assembly (181 bacterial CDS) and trace read archive (16S rDNA) revealed a dominant proteobacterial profile strongly skewed to Rickettsiales (*Alphaproteobacteria*) genomes. By way of phylogenetic analysis of 16S rDNA and 113 proteins conserved across proteobacterial genomes, as well as identification of 27 rickettsial signature genes, we propose a Rickettsiales endosymbiont of *Trichoplax adhaerens* (RETA). The majority (93%) of the identified bacterial CDS belong to small scaffolds containing prokaryotic-like genes; however, 12 CDS were identified on large scaffolds comprised of eukaryotic-like genes, suggesting that *T. adhaerens* might have recently acquired bacterial genes. These putative LGTs may coincide with the placozoan's aquatic niche and symbiosis with RETA. This work underscores the rich, and relatively untapped, resource of eukaryotic genome projects for harboring data pertinent to host-microbial interactions. The nature of unknown (or poorly characterized) bacterial species may only emerge via analysis of host genome sequencing projects, particularly if these species are resistant to cell culturing, as are many obligate intracellular microbes. Our work provides methodological insight for such an approach.

INTRODUCTION

Bacterial DNA sequences may be generated as a byproduct of eukaryotic genome sequencing. The source of this bacterial DNA can be 1) contamination (failure to separate incidental bacterial species from eukaryotic cell preparation, or even failure to completely eliminate sequencing adapters or cloning vector sequences), 2) environmental (extracellular bacteria sequenced as a consequence of occupying the same niche as the eukaryote), 3) symbiotic (extracellular or facultative/obligate intracellular bacterial species directly associated with the eukaryotic host), or 4) LGT [lateral gene transfer of bacterial sequences to the genome of the eukaryote]. Aside from contamination, the remaining sources of bacterial DNA sequences generated by eukaryotic genome projects provide clues about the biological relationships between eukaryotes and their

associated bacterial species. Thus, the creation of methods for the detection, extraction and characterization of microbial sequences generated by eukaryotic genome sequencing studies are of critical importance, particularly for gaining insight on poorly characterized bacterial species, some of which may be recalcitrant to cultivation. Furthermore, annotation of such bacterial reads or genomes and deposition into appropriate public databases is paramount for facilitating this approach.

Studies identifying endosymbiotic bacterial genomes within the data generated from eukaryotic sequencing projects are growing. Inspection by Salzberg et al. of disparate fruit fly (Arthropoda: Diptera: *Drosophila ananassae*, *D. simulans*, and *D. willistoni*) genome trace file archives resulted in the identification of three novel species of *Wolbachia* (*Alphaproteobacteria*: Rickettsiales: Anaplasmataceae) **(Salzberg et al. 2005b, a)**. The genome sequence of another *Wolbachia* strain was discovered within the whole-genome sequencing data for the mosquito *Culex quinquefasciatus* strain JHB **(Salzberg et al. 2009)**. Sequencing of the *Hydra magnipapillata* (Cnidaria: Hydrozoa) genome revealed the presence of an endosymbiont most closely related to species of *Curvibacter* (*Betaproteobacteria*: Burkholderiales: Comamonadaceae) **(Chapman et al. 2010)**. Most recently, the genome of a Rickettsiales endosymbiont of *Ixodes scapularis* (Rickettsiales: Rickettsiaceae: REIS) was assembled from mining the initial data generated from the deer tick sequencing effort **(Gillespie et al. 2012a)**. All of these studies have revealed genomic data essential for furthering the knowledge of bacterial endosymbioses within animal species. In the case of REIS, important characteristics of a non-pathogen came to light when compared to the genomes of closely related pathogenic spotted fever group rickettsiae **(Gillespie et al. 2012a)**.

Genomic analyses of several eukaryotes, such as the rotifers *Adineta vaga* and *A. ricciae* (Rotifera; Bdelloidea) **(Gladyshev et al. 2008, Boschetti et al. 2012)**, *H. magnipapillata* **(Chapman et al. 2010)**, the silkmoth *Bombyx mori* (Arthropoda: Lepidoptera) **(Li et al. 2011)**, and the spider mite *Tetranychus uriticae* (Arthropoda: Acari) **(Grbic et al. 2011)**, have revealed the presence of many genes originating from diverse bacterial species, illustrating the role of LGT in the diversification of eukaryotic genomes. For instance, a bacterial mannanase gene from *Bacillus* spp. (Firmicutes: Bacilliales) was recently reported in the genome of the coffee berry borer beetle, *Hypothenemus hampei* (Arthropoda: Coleoptera), and demonstrated to metabolize galactomannan, the major storage polysaccharide of coffee **(Acuna et al. 2012)**. Large portions of *Wolbachia* genomes have been identified in several arthropod host genomes, including the bean beetle *Callosobruchus chinensis* **(Kondo et al. 2002, Nikoh et al. 2008)**, the longicorn beetle *Monochamus alternatus* **(Aikawa et al. 2009)**, *D. ananassae*, **(Dunning Hotopp et al. 2007)**, the parasitoid wasp *Nasonia vitripennis* (Arthropoda: Hymenoptera) **(Werren et al. 2010)**, as well as several filarial nematode genomes **(Dunning Hotopp et al. 2007, McNulty et al. 2010)**, underscoring the prevalence of LGT between obligate intracellular bacterial species and their eukaryotic hosts. Intriguingly, several bacterial genes encoded in the genome of the pea aphid, *Acyrthosiphon pisum* (Arthropoda: Hemiptera), presumably foster its well-characterized mutualism with *Buchnera aphidicola* (*Gammaproteobacteria*: Enterobacteriales), potentially relegating the symbiont to aphid bacteriocytes **(Nikoh and Nakabachi 2009, Nikoh et al. 2010)**. These studies of bacterial gene incorporation into eukaryotic genomes illustrate the need to develop tools to distinguish congener bacterial genes serendipitously captured in eukaryotic sequencing projects from true LGT events.

In this study, we analyzed the genome project (reads and assembly) of the primitive metazoan *Trichoplax adhaerens* (Animalia: Placozoa) for the presence of bacterial DNA sequences. Published in 2008, the *T. adhaerens* genome revealed "cryptic complexity", as most genes encoding transcription factors and signaling pathways underpinning eumetazoan cellular differentiation and development are present in this simple animal **(Srivastava et al. 2008)**. *T. adhaerens* lacks nerves, sensory cells and muscle cells, with only four cell types previously described **(Grell 1971, Schierwater 2005)**. Morphologically, the animal resembles a flat disc of cells with two epithelial layers sandwiching a region of multinucleate fiber cells **(Grell and Ruthmann 1991, Guidi et al. 2011)**. *T. adhaerens* is known to harbor a Gram-negative endosymbiont within fiber cells **(Grell 1972, Grell and Benwitz 1974)**, with bacteria passed to developing oocytes via fiber cell extensions **(Eitel et al. 2011)**. Our motivation for analyzing the *T. adhaerens* genomic data for sequences belonging to this symbiont was generated by previous studies that included bacterial-like genes from *T. adhaerens* in phylogeny estimations **(Felsheim et al. 2009, Baldridge et al. 2010, Gillespie et al. 2010, Nikoh et al. 2010)**. As two of these genes are rickettsial signatures (*virD4*, plasmid-like *parA*), we considered it likely that the *T. adhaerens* fiber cell symbiont is a member of the obligate intracellular Rickettsiales.

We report an in-depth analysis of the *T. adhaerens* genome assembly and trace read archive, which divulged bacterial 16S rDNA sequences, 181 bacterial-like coding sequences [CDS] and many additional partial gene fragments of probable bacterial nature. Robust phylogenomic analyses grouped the *T. adhaerens* bacterium with the mitochondria invader "*Candidatus* Midichloria mitochondrii" (*Alphaproteobacteria*: Rickettsiales), albeit with only 53% conservation across the core proteins of these two species. Using this substantial molecular evidence, we name a Rickettsiales endosymbiont of *Trichoplax adhaerens* [RETA] and provide adjusted annotation and related genomic information for its genes deposited in the Pathosystems Resource Integration Center (PATRIC, http://www.patricbrc.org/). This work illustrates the rich resource of eukaryotic genome projects for data pertinent to diverse host-microbial interactions, and also demonstrates that highly divergent, poorly known microbial species can be characterized via in-depth mining and phylogenomic analyses of even minimal genetic information captured from these broad scale eukaryotic sequencing efforts.

METHODS

SSU rDNA Analyses

Read analysis. In order to assess the taxonomic distribution of bacterial species sequenced concomitantly with *Trichoplax adhaerens*, 1,230,612 WGS sequencing reads from the *T. adhaerens* genome project (Joint Genome Institute) were downloaded from the NCBI Trace Archive for analysis. Reads were cleaned of vector contamination using `cross_match` **(Ewing et al. 1998)** and screened for quality using the `fastqc` program from the Babraham Bioinformatics group (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). All reads with a Phred quality score greater than 20 were subsequently mapped against a library of small subunit [SSU] rDNA sequences, including full-length bacterial 16S rDNA sequences from the Greengenes database **(DeSantis et al. 2006)** and the full-length 18S rDNA sequence for *T. adhaerens* (NCBI acc. no. Z22783). *T. adhaerens* sequencing reads were aligned to this SSU library using the Burrows-Wheeler Aligner **(Li and Durbin 2010)** with the options `bwasw -t 4 -T 37`. Reads that had at least one successful match in the SSU library were binned according

to the taxonomic classification of their matches, and subsequently visualized using `Krona` v.2.0 **(Ondov et al. 2011)**.

Phylogeny estimation. The rickettsial-like 16S rDNA sequences retrieved from the *T. adhaerens* read archive were used as subjects in `blastn` searches of the NCBI *nr* database for the closest bacterial sequences (of greater or equal length). Seven rickettsial-like sequences, from various environmental studies **(Revetta et al. 2010, Sunagawa et al. 2010, Revetta et al. 2011)**, were retrieved and combined with a dataset ($n = 47$) recently used to estimate Rickettsiales phylogeny **(Gillespie et al. 2012b)**. Additional mitochondrial SSU rDNA sequences and outgroup sequences (*Betaproteobacteria*, *Gammaproteobacteria*, *Alphaproteobacteria*: Rhodospirillales, Parvularculales, Rhizobiales) were included based on previous phylogenetic studies of *Alphaproteobacteria* **(Williams et al. 2007, Thrash et al. 2011, Viklund et al. 2012)**. Further rickettsial 16S rDNA sequences from recent studies **(Kawafune et al. 2012, Matsuura et al. 2012)** were added to entail robust sampling within the major Rickettsiales groups, bringing the dataset to 93 SSU rDNA sequences. Information pertaining to all analyzed SSU rDNA sequences, and the consensus rickettsial 16S rDNA sequence mined from the *T. adhaerens* read archive, are provided in supplementary table S1.

All sequences, plus a second set excluding the mitochondrial SSU rDNA sequences ($n = 83$), were aligned using `MUSCLE` v3.6 **(Edgar 2004a, b)** with default parameters. Ambiguously aligned positions, the majority being present within the variable regions of the SSU rRNA structure, were culled using `Gblocks` **(Castresana 2000, Talavera and Castresana 2007)**. Phylogenies were estimated under maximum likelihood using `RAxML` **(Stamatakis et al. 2008)**. The GTR substitution model was used with estimation of GAMMA and the proportion of invariable sites. Branch support was measured with bootstrapping (1000 replications).

CDS Analyses

Assembly analysis. A BLAST-based pipeline was used to identify candidate bacterial CDS within the *T. adhaerens* genome assembly. Each of the 11,540 predicted proteins of the Triad1 assembly was BLASTed (using `blastp`) against three databases: **1)** a scoping database consisting of all available Rickettsiales proteins (NCBI taxonomy id 766); **2)** all bacteria proteins (NCBI taxonomy id 2) excluding those in the Rickettsiales database; and **3)** all eukaryotic proteins (NCBI taxonomy id 2759) excluding *T. adhaerens*. The choice of Rickettsiales for the scoping database was informed by preliminary results from our SSU taxonomic distribution analysis and phylogeny estimation, as well as by previous studies that included *T. adhaerens* bacterial-like genes in phylogeny estimations **(Felsheim et al. 2009, Baldridge et al. 2010, Gillespie et al. 2010, Nikoh et al. 2010)**. For each *T. adhaerens* protein, the top 50 matches (based on e-value) in each database were pooled and ranked according to a comparative sequence similarity match score:

$$S_m(m,h) = b \cdot I \cdot Q \qquad (1)$$

Where:

S_m is the comparative sequence similarity score
b is the raw bitscore of match *m* to host protein *h*
I is the percent identity
Q is the percent of *h* (the query) that aligned

By incorporating %ID and match length, S_m is intended to de-emphasize highly significant matches to short stretches of query (*i.e.*, conserved domains) in favor of longer stretches of similarity.

The top five scoring matches from the pooled lists of subjects were retained and grouped according to hit number (1-5) and organism taxonomy. *T. adhaerens* proteins with no top-5 scoring matches to bacteria were excluded from further analyses (n = 9,843, or 85.3% of the total *T. adhaerens* proteins). The remaining proteins (n = 1697) were then subjected to cursory inspection of *T. adhaerens* annotation, as well as targeted `blastp` searches against various databases (*Alphaproteobacteria*, individual Rickettsiales genera, mitochondria, etc.) with manual inspection of functional annotations from top hits. A final dataset of probable bacterial CDS (n = 181) was constructed with annotations derived primarily from Uniprot **(2012)**, PATRIC **(Aziz et al. 2008, Gillespie et al. 2011)** and the NCBI Conserved domains database **(Marchler-Bauer et al. 2011)**. In some cases (e.g., rickettsial signature proteins) annotations from the literature were selected. We named a hypothetical organism, Rickettsiales endosymbiont of *Trichoplax adhaerens* [RETA], based on the hypothesis that these proteins define one single bacterial species. Each protein was assigned a unique identifier RETA0001-RETA0181. A complete list of the RETA proteins is available at PATRIC (http://enews.patricbrc.org/rickettsial-endosymbiont-of-trichoplax-adhaerens/) and provided in **supplementary table S2**.

Dataset classification. The 181 bacterial-like CDS extracted from the *T. adhaerens* assembly were divided into two groups based on manual inspection of `blastp` results. A Core Dataset of proteins with conserved domains (functions) that are generally vertically inherited, and hence not typical constituents of the bacterial mobilome, was constructed (n = 119). These bacterial-like proteins had one of three characteristics: **1)** top `blastp` hits to Rickettsiales with the next closest homologs in *Alphaproteobacteria*, or **2)** top `blastp` hits to *Alphaproteobacteria* with rickettsial homologs present or absent, or **3)** top `blastp` hits to other *Proteobacteria* but with highly similar rickettsial homologs. This relaxed criterion permitted the capture of putative rickettsial-like genes that may not be known from the available rickettsial (or even alphaproteobacterial) sequenced genomes. Further, it allowed for identifying CDS that may be difficult to detect due to extreme divergence of the symbiont genome. Finally, this approach also provided flexibility with interpretation of `blastp` results, which may be biased due to a number of characteristics in the query and/or subject sequences (e.g., truncated sequences, length heterogeneity across matches due to insertions and deletions, base compositional bias [BCB], etc.). Three instances of split ORFs were detected (*secA*, *mnmA* and *GlmS*), as well as three fused gene models (*tolC-sppA*, *rmuC-uvrD* and *kdsA-smpA*), bringing the number of Core Dataset genes to 116.

The remaining 62 bacterial-like CDS, or the Accessory Dataset, mostly encompassed proteins of the bacterial mobilome, especially those typically encoded by intracellular species. Aside from lacking a phylogenetic signal typical of conserved alphaproteobacterial proteins, CDS of the Accessory Dataset had one of the following characteristics: **1)** highly similar to Rickettsiales signature proteins, **2)** present in some (or all) Rickettsiales genomes yet divergent in sequence and phylogenetic signal, or **3)** unknown from Rickettsiales genomes. Proteins of the Accessory Dataset were analyzed separately (see below, Accessory Dataset analyses) since, while they could all depict proteins encoded by one putative rickettsial symbiont (RETA), it is also possible that some may be from additional microbes captured in the *T. adhaerens* genome sequencing (particularly those species for which 16S rDNA sequences were mined).

Genome comparison. Careful observation of the `blastp` profiles and preliminary phylogeny estimations revealed the mitochondria-associated rickettsial species "*Candidatus* Midichloria mitochondrii" (hereafter *M. mitochondrii*), **(Lo et al. 2004, Sacchi et al. 2004, Sassera et al. 2006)** as the closest relative (with available genome sequence data) to the majority of the mined bacterial-like CDS. Accordingly, an all-against-all `blastp` analysis was executed between *M. mitochondrii* (n = 1,211) and *T. adhaerens* (n = 11,540). The `blastp` results for 347 matches, including S_m scores and e-values, were mapped over a circular plot of the *M. mitochondrii* genome using Circos **(Krzywinski et al. 2009)**, with manual adjustment. Proteins of both the Core and Accessory Datasets with homologs in *M. mitochondrii* (n = 138) were highlighted, and regions of synteny between the *M. mitochondrii* genome and CDS from several *T. adhaerens* scaffolds were superimposed on the plot.

Core Dataset Analyses

Genome-based phylogeny. The RETA Core Dataset proteins were combined under the assumption that they were vertically inherited from an alphaproteobacterial ancestor. To better understand the systematic position of RETA, a total of 176 genomes were used for robust phylogeny estimation (supplementary table S3). Aside from the RETA Core Dataset proteins, the analysis included genomes from 80 Rickettsiales, 82 non-Rickettsiales *Alphaproteobacteria*, 12 mitochondria, and two outgroup taxa (*Betaproteobacteria* and *Gammaproteobacteria*). The *T. adhaerens* mitochondrial genome, which was generated separately from the whole genome sequencing project **(Dellaporta et al. 2006)**, was used. Taxon sampling was modeled after several previous studies on Rickettsiales phylogeny **(Sassera et al. 2011, Rodriguez-Ezpeleta and Embley 2012, Viklund et al. 2012)** for the purpose of presenting a comparable hypothesis. Trees from these previous studies are summarized in supplementary figure S1 to assist the interpretation of our current hypothesis.

For genome-based phylogeny estimation, an automated pipeline for protein family selection and tree building was implemented in Java. Bacterial protein sequences were downloaded from PATRIC **(Gillespie et al. 2011)**. The RETA and mitochondria proteins were extracted from NCBI, as were an additional 27 *Rickettsia* genomes not annotated at PATRIC at the time of analysis. BLAT (refined BLAST algorithm) **(Kent 2002)** searches were performed to identify similar protein sequences between all genomes, including the outgroup taxa. To predict initial homologous protein sets, mcl **(Van Dongen 2008)** was used to cluster BLAT results, with subsequent refinement of these sets using hidden Markov models as previously described **(Durbin et al. 1998)**. These protein families were then filtered to include only those with membership in >80% of the analyzed genomes (141 or more taxa included per protein family, excluding the mitochondrial genomes). Multiple sequence alignment of each protein family was performed using `MUSCLE` (default parameters) **(Edgar 2004a, b)**, and regions of poor alignment (length heterogeneous regions) were masked using `Gblocks` **(Castresana 2000, Talavera and Castresana 2007)**. All modified alignments were concatenated into a single dataset for phylogeny estimation.

Tree-building was initially performed using `FastTree` **(Price et al. 2010)**. Support for generated lineages was estimated using a modified bootstrapping procedure, with 100 pseudoreplications sampling only half of the aligned protein sets per replication (NOTE: standard bootstrapping tends to produce inflated support values for very large alignments). Local refinements to tree topology were attempted in instances where highly supported nodes have subnodes with low support. This refinement is executed by running the entire pipeline using only

those genomes represented by the node being refined (with additional sister taxa for rooting purposes). The refined subtree is then spliced back into the full tree.

Using `PhyloBayes` v3.3 **(Lartillot et al. 2009)**, we also analyzed the dataset with the CAT model of substitution, which is a nonparametric method for modeling site-specific features of sequence evolution **(Lartillot and Philippe 2004, 2006)**. Given the nature of the BCB of Rickettsiales and mitochondrial genomes, and the ability of the CAT model to accommodate saturation due to convergences and reversions **(Lartillot et al. 2007)**, this approach is of substantial importance for estimating Rickettsiales phylogeny **(Rodriguez-Ezpeleta and Embley 2012, Viklund et al. 2012)**. Two independent Markov chains were run in parallel using `PhyloBayesMPI` v.1.2e under the CAT-GTR model, with the bipartition frequencies analyzed at various time points using the `bpcomp` program. For tree-building, appropriate burn-in values were determined by plotting the log likelihoods for each chain over sampled generations (time). Analyses were considered complete when the maximum difference in bipartition frequencies between the two chains was less than 0.1. Ultimately, a burn-in value of 1000, with sampling every 2 trees, was used to build a consensus tree.

Finally, to further evaluate the rickettsial nature of the Core Dataset, all proteins were BLASTed against three databases: Rickettsiales, Bacteria (excluding Rickettsiales) and Eukaryota (excluding *T. adhaerens*). The proteins were then binned into three 'sub-datasets' (Ric, Bac or Euk) based on the highest S_m score against each database (supplementary figure S2). The resulting three sub-datasets (Ric-78, Bac-26 and Euk-9) were then run through the procedure described above for phylogeny estimation, resulting in one FastTree-based and one PhyloBayes-based tree for each sub-dataset (six total trees).

Genome divergence. To determine if the degree of divergence between the RETA Core Dataset proteins and their homologs in *M. mitochondrii* is typical for major rickettsial lineages, an approximation of genome divergence across the genera of Rickettsiales and the RETA Core Dataset was calculated. The final alignment of the Core Dataset was processed to include only one representative species from each Rickettsiales genus (*Odyssella*, *Midichloria*, *Neorickettsia*, *Wolbachia*, *Anaplasma*, *Ehrlichia*, *Orientia*, *Rickettsia*) plus the RETA Core Dataset proteins. All positions of the alignment containing missing data (?s) were removed, resulting in 8327 aa sites (8319 informative). The program `DIVEIN` **(Deng et al. 2010)** was used to estimate percent protein divergence using both the Blosum62 and WAG amino acid substitution models.

Accessory Dataset Analyses

The 62 *T. adhaerens* bacterial-like sequences lacking a typical alphaproteobacterial signal (Accessory Dataset) were separated from the Core Dataset proteins using `blastp` searches. The *nr* (All GenBank+RefSeq Nucleotides+EMBL+DDBJ+PDB) database was used, coupled with a search against the Conserved Domains Database **(Marchler-Bauer et al. 2011)**. Searches were performed across "all organisms", as well as "Rickettsiales" with composition-based statistics. No filter was used. Default matrix parameters (BLOSUM62) and gap costs (Existence: 11 Extension: 1) were implemented, with an inclusion threshold of 0.005. This process facilitated the division of the RETA Accessory Dataset into three groups: **1)** proteins with closest homologs to Rickettsiales (n = 27), **2)** proteins present in (some or all) Rickettsiales genomes but divergent from their rickettsial counterparts (n = 18), and **3)** proteins unknown from Rickettsiales (n = 17). The two groups containing rickettsial homologs were then used in subsequent `blastp` searches against the following five databases: **1)** "Rickettsiales", **2)** "*Alphaproteobacteria* (minus

Rickettsiales)", **3)** "*Proteobacteria* (minus *Alphaproteobacteria*)", **4)** "Bacteria (minus *Proteobacteria*)", and **5)** "minus Bacteria". The top 20-50 (query-dependent) subjects from each search resulting in significant (> 40 bits) alignments were retrieved, compiled and aligned using `MUSCLE` v3.6 (default parameters). Full alignments were used for subsequent analyses. In some instances, all sequences within alignments were screened for possible signal peptides using `SignalP` v.4.0 **(Petersen et al. 2011)**, `LipoP` v.1.0 **(Juncker et al. 2003)** and `Phobius` **(Kall et al. 2007)**. Potential transmembrane spanning regions were predicted using transmembrane hidden Markov model (`TMHMM`) v.2.0 **(Krogh et al. 2001)**.

Phylogenetic trees were estimated using `PAUP*` v4.0b10 (Altivec) **(Wilgenbusch and Swofford 2003)** under parsimony and implemented heuristic searches with 500 random sequence additions holding 50 trees per replicate. Single most parsimonious trees or consensus trees of equally parsimonious topologies were generated, with branch support assessed using bootstrapping (1000 pseudoreplications). Phylogenies were also estimated under maximum likelihood using `RAxML` v.7.2.8 **(Stamatakis et al. 2008)**. A gamma model of rate heterogeneity was used with estimation of the proportion of invariable sites. Branch support was assessed with 1000 bootstrap pseudoreplications. Finally, for the analyses of flagella (FlgG and FliG) and T4SS proteins (RvhD4 and RvhB6) alignments were combined and analyzed together using both `RAxML` and `PhyloBayes` (as described above).

Evaluating Bacterial Gene Transfer to the *T. adhaerens* Genome

Several approaches were made to determine if any of the 181 bacterial-like genes of the Core and Accessory Datasets have strong evidence for being a part of the *T. adhaerens* genome (as opposed to belonging to the genomes of RETA or other microbes). We first evaluated the scaffold properties that contain each bacterial-like gene, judging that bacteria-to-host LGTs could only be demonstrated on scaffolds greater than one gene and containing eukaryotic-like genes. The 181 RETA genes were divided into four categories: **1)** genes present on large (> 40 genes) scaffolds with predominately eukaryotic-like genes (n = 18); **2)** genes present on small (< 7 genes) "hybrid" scaffolds with both bacterial- and eukaryotic-like genes (n = 19), **3)** genes present on small (< 5 genes) scaffolds comprised entirely of bacterial-like genes (n = 59), and **4)** singleton-gene scaffolds (n = 85). Next, CDS within each category were split into single- and multi-exon genes. All multi-exon genes were then subjected to `blastx` searches using the entire gene models (exons + introns) as queries. These entire gene models were also analyzed with the bacterial gene prediction program `fgenesb` **(Tyson et al. 2004),** using the "generic BACTERIAL" model, to determine discrepancies with the original eukaryotic gene predictions within the *T. adhaerens* assembly. This intron evaluation allowed for the distinction between true eukaryotic genes inadvertently included within the RETA datasets (e.g., nuclear genes encoding mitochondrial proteins), and bacterial LGTs undergoing a transformation to eukaryotic-like gene structures (i.e., accruement of introns, gain of eukaryotic signal sequences, etc.). Importantly, the approach also revealed evidence against predicted introns due to **1)** chimeric gene models comprised of two or more genes (or gene fragments) that were "stitched" together by the eukaryotic gene calling algorithms, **2)** bacterial genes that were divided into fragments due to multiple start sites called by eukaryotic gene calling algorithms, and **3)** gene models that were fused with additional short (and likely spurious) ORFs. Finally, for the 18 RETA genes found on large scaffolds that are dominated by eukaryotic-like genes, individual protein phylogenies were estimated (see above, Accessory Dataset Analyses) to lend an additional level of support for discerning between true eukaryotic genes and LGTs to the *T. adhaerens* genome.

RESULTS

An overview of the methodology applied to the analysis of *T. adhaerens* genome project (sequence read archive and assembly) illustrates the various approaches implemented to identify bacterial DNA sequences (**fig. 1**). Totals for CDS and scaffolds are given for extracted data that suggest the presence of a rickettsial species, with more detailed information provided in the various sections below.

Bacterial DNA Mined from *Trichoplax*

SSU rDNA. Within the *T. adhaerens* genome project sequence read archive, a total of 289 SSU rDNA sequences were mined for analysis (**fig. 2a**). The majority (88.2%) of these sequences were identified as eukaryotic 18S rDNA genes belonging to the *T. adhaerens* genome. The remaining 34 SSU rDNA sequences were determined to have highest similarity with bacterial or plastid 16S rDNA-like sequences. Using the prokaryotic 16S rDNA sequences from the Greengenes database **(DeSantis et al. 2006)**, these sequences received the most accurate taxonomic assignment possible (**supplementary table S4**). Three major groups comprised 76.5% of the sequences: *Alphaproteobacteria* ($n = 9$), *Gammaproteobacteria* ($n = 4$) and eukaryotic chloroplasts ($n = 13$). The remaining eight sequences were grouped into a diverse array of taxa (*Betaproteobacteria*, *Deltaproteobacteria*, Spirochaetes, Firmicutes, and Plantomycetes).

Importantly, ten of the 16 taxonomic assignments were made for one individual 16S rDNA operational taxonomic unit, with the bacterial assignments for *Marivita* spp. (Rhodobacterales), *Limnobacter* spp. (Burkholderiales) and *Borrelia* spp. (Spirochaetes) possibly representing a single organism with multiple rDNA operons **(Kembel et al. 2012)**. The 13 cyanobacterial-like sequences had the best matches to chloroplasts of marine eukaryotes, such as haptophyte and cryptomonad algae, as well as heterokonts. These rDNA sequences may also be inflated due to a high copy number of plastid genomes. Finally, the two Rickettsiales sequences were determined to depict partial fragments of the same molecule, and thus were concatenated into one rDNA sequence and classified as RETA.

Bacterial CDS. Of the 11,540 predicted CDS within the *T. adhaerens* assembly, 14.7% ($n =$ 1,697) had at least one prokaryotic analog within the top five scoring hits in a reciprocal `blastp` analysis against the *nr* database (**fig. 2b**). Pooling the bacterial hits according to higher-level taxonomy illustrated a bias towards Rickettsiales, other *Alphaproteobacteria* and *Gammaproteobacteria*, with those three groups showing at least one representative taxon within the top five hits in 163, 175 and 211 `blastp` matches, respectively. The remaining higher-level taxa with more than 15 total hits (1-5) comprised a diverse group of prokaryotes. Importantly, aside from potential symbiont DNA and LGT products, many of these *T. adhaerens* proteins are nuclear genes encoding proteins that are trafficked to and imported by the mitochondria, with the diversity of bacterial groups in the `blastp` matches consistent with the genetic mosaicism of nuclear-encoded mitochondrial genes **(Thiergart et al. 2012)**. Along with Rickettsiales, other *Alphaproteobacteria*, and *Gammaproteobacteria*, four other higher-level taxonomic groups (*Betaproteobacteria*, *Deltaproteobacteria*, Spirochaetes, and Firmicutes) within this analysis had a corresponding 16S rDNA mined from the *T. adhaerens* sequence read archive (**fig. 2a**). Finally, within each taxonomic group containing scores to over 75 of the 1,697 *T. adhaerens* proteins, only the Rickettsiales showed a consistent representation across hits 1-5 ($R^2 = 0.006$), with other groups showing increasing or decreasing representation across hits 2-5 (avg. $R^2 =$

0.805 for *Gammaproteobacteria*, Other *Alphaproteobacteria*, Firmicutes, Acidobacteria, Cyanobacteria, Deinococcus-Thermus, Archaea, and Actinobacteria). Thus, in most cases where Rickettsiales was the top scoring hit to a *T. adhaerens* protein, hits 2-5 were also occupied by Rickettsiales, suggesting a strong rickettsial signal within these proteins.

Rickettsiales 16S rDNA Phylogeny

Despite mining a diverse set of 16S rDNA sequences from the *T. adhaerens* sequence read archive, we only estimated a phylogeny of the RETA 16S rDNA sequence with a diverse group of Rickettsiales for three primary reasons: 1) a match between the retrieved rickettsial 16S rDNA (**fig. 2a**) and CDS (**fig. 2b**), 2) the long-known presence of an intracellular bacterial symbiont associated with *T. adhaerens* **(Grell 1972, Grell and Benwitz 1974, Eitel et al. 2011)** and 3) evidence from other studies suggesting the presence of rickettsial CDS within the *T. adhaerens* assembly **(Felsheim et al. 2009, Baldridge et al. 2010, Gillespie et al. 2010, Nikoh et al. 2010)**. Phylogeny estimation of the SSU rDNA dataset grouped RETA in a clade of diverse rickettsial species that is sister to the traditional Anaplasmataceae *sensu stricto* (Anaplasmataceae *s. s.*) (**fig. 3**). Previous studies have also recovered this large clade within Rickettsiales **(Beninati et al. 2004, Davis et al. 2009, Vannini et al. 2010, Boscaro et al. 2012, Kawafune et al. 2012)**, which includes many species with diverse eukaryotic hosts, and we recently proposed the name "Midichloriaceae" as a sister family within the Anaplasmataceae *sensu lato* **(Gillespie et al. 2012b)**. Here we determined RETA to be part of a clade comprising poorly described bacteria from species of coral (*Gorgonia ventalina*, *Montastraea faveolata*) and sponge (*Cymbastela concentrica*) **(Revetta et al. 2010, Sunagawa et al. 2010, Revetta et al. 2011)**, consistent with the aquatic habitat of *T. adhaerens*. Importantly, this lineage is well diverged from the group comprising *M. mitochondrii* (89% identity between RETA and *M. mitochondrii* str. IricVA 16S rDNA sequences), which is comprised predominantly of bacteria identified in various arthropod species. The basal lineages of "Midichloriaceae" are comprised mostly of protist-associated rickettsial species and uncharacterized species collected via environmental sampling. Collectively, our analysis of the RETA SSU rDNA sequence retrieved from the *T. adhaerens* sequence reads is consistent with the presence of a rickettsial bacterial symbiont associated with Placozoa.

A Rickettsiales Genome Associated with *Trichoplax*

The 1697 "bacterial-like" proteins identified in our CDS mining of the *T. adhaerens* (**fig. 2b**) were further evaluated via manual inspection of annotation, as well as a series of `blastp` analyses against specific databases (bacterial groups, mitochondria), to yield a dataset of 181 probable bacterial CDS (**supplementary table S2**). The proteins were divided into RETA Core (n = 119) and Accessory (n = 62) Datasets and assigned unique identifiers (RETA0001-RETA0181). Given that the majority of `blastp` subjects from all-against-all blast analyses (between *T. adhaerens* and 'all bacteria') were from Rickettsiales genomes (**fig. 2b**), we compared the RETA proteins directly with the taxon containing the most top `blastp` subjects, *M. mitochondrii* (**fig. 4a**). An all-against-all blast analysis between *T. adhaerens* and *M. mitochondrii* yielded 347 hits above a set threshold (S_m >20), with 124 of these matches illustrating homologous proteins from the *M. mitochondrii* genome and the *T. adhaerens* assembly. Additional CDS (n = 14) were later identified as the best RETA-*M. mitochondrii* matches based on manual `blastp` analyses (**supplementary figure 3**)**.** Thus, a total of 138 RETA proteins were mapped to the *M. mitochondrii* genome, comprising 93.2% (n = 111) of the

Core and 41.5% (n = 27) of the Accessory RETA Datasets (**fig. 4b**). Despite being present on predominantly small scaffolds within the *T. adhaerens* (see below, Bacterial Genes in the *Trichoplax* Genome), seven regions of synteny were identified across RETA and *M. mitochondrii*, an understandable result given the lack of genome synteny across genera of Rickettsiales **(Gillespie et al. 2012b)**. Collectively, 76.2% of the RETA CDS were found to have highly similar homologs in the *M. mitochondrii* genome, including seven syntenic regions, suggesting that these bacterial CDS from the *T. adhaerens* genome project comprise a potential Rickettsiales bacterium.

Genome-Based Phylogenetic Position of RETA

An estimated phylogeny of the Core Dataset, which accommodated the strong inherent BCB in some of the data, unambiguously placed RETA with *M. mitochondrii* in a clade ("Midichloriaceae") within the Rickettsiales (**fig. 5**). The sampled mitochondrial genomes formed a lineage within the Rickettsiales, diverging after the Holosporaceae ("*Candidatus* Odyssella thessalonicensis", hereafter *O. thessalonicensis*) but prior to the derived rickettsial families (Anaplasmataceae, Rickettsiaceae). Thus the ancestral position of the Holosporaceae and branching point for the mitochondrial ancestor are consistent across trees estimated from SSU rDNA (**fig. 3**) and multiple proteins (**fig. 5**). The discrepancy in the placement of "Midichloriaceae" as ancestral to the Rickettsiaceae in the genome-based tree versus its sister relationship to the Anaplasmataceae *s. s.* in the SSU rDNA-based tree is explained by the lack of genomic data available for other members of the "Midichloriaceae".

While the tree generated without accommodating BCB also grouped RETA and *M. mitochondrii* together, the position of the SAR11 group of *Alphaproteobacteria* within the Rickettsiales was recovered (**supplementary figure 4**). Importantly, previous genome-based phylogenies estimated with methods that accommodate BCB unambiguously show that SAR11 are not closely related to mitochondria, yet are more derived in the *Alphaproteobacteria* **(Rodriguez-Ezpeleta and Embley 2012, Viklund et al. 2012)** (**supplementary figure 1**). Of note, our tree estimated without accommodating BCB placed *O. thessalonicensis* outside of the Rickettsiales as an early branching lineage of the remaining *Alphaproteobacteria*. While originally described as a member of the Holosporaceae **(Birtles et al. 2000)**, a study presenting genome-based phylogenies failed to group *O. thessalonicensis* within Rickettsiales **(Georgiades et al. 2011)**. This disparity with our hypothesis is discussed below in light of rickettsial signatures and the nature of Holosporaceae (see Discussion).

The robustness of the phylogenetic signal within the Core Dataset was determined by estimating trees from the three sub-datasets (Ric-78, Bac-26 and Euk-9), which were grouped based on top S_m score against three databases (Rickettsiales, Bacteria excluding Rickettsiales, and Eukaryota). Whether generated using `RAxML` (GTR model) or `PhyloBayes` (CAT model), estimated trees based on all three sub-datasets unambiguously grouped RETA within the Rickettsiales (**supplementary figure 5**). This suggests that, despite differences in top `blastp` hits using RETA proteins as queries (as revealed by S_m scores in **supplementary figure 2**), the proteins of the Core Dataset contain a phylogenetic signal that consistently places RETA within the Rickettsiales, distinct from the Rickettsiaceae and Anaplasmataceae.

RETA and Rickettsiales Genome Divergence

While some RETA CDS have a clear rickettsial signature (i.e., *rvhD4*, *parA* and *ampD* described previously **(Felsheim et al. 2009, Baldridge et al. 2010, Gillespie et al. 2010, Nikoh et al.**

2010), many of the mined CDS had limited %ID to any counterparts in the NR database, making distinction even between eukaryotic and prokaryotic proteins often difficult to assess by sequence comparison alone. Despite this, a predominant rickettsial signal emerged from the mined CDS (**fig. 5**), yet it is clear from the 16S rDNA tree that RETA is a member of a diverse rickettsial lineage with minimal genomic information available (*M. mitochondrii*). To determine if RETA is typical in its degree of divergence from other rickettsial lineages, we calculated genome divergence across the most conserved protein regions within the Core Dataset, sampling one representative member of each rickettsial genus, plus RETA (**table 1**). As expected, RETA was most similar to *M. mitochondrii*, with these genomes being 45% divergent from one another. The mean genome divergence across the major Rickettsiales genera was 51%, and ranged from 30% (*Anaplasma* vs. *Ehrlichia*) to 63% (RETA vs. *Neorickettsia*). Importantly, RETA had a mean divergence of 56% compared to other rickettsial genomes, just below *Neorickettsia* (57%) and not atypical from other rickettsial lineages. These results are consistent with our phylogeny estimation based on 16S rDNA (**fig. 3**) and the conserved proteins of the Core Dataset (**fig. 5**).

Variable Bacterial CDS Mined from *Trichoplax*

Of the total bacterial-like CDS identified from the *T. adhaerens* genome assembly, 32% ($n = 62$) were determined to lack a conserved alphaproteobacterial signal, meaning there is little support for the vertical inheritance of these genes from a common alphaproteobacterial ancestor. These 62 RETA Accessory Dataset proteins were further divided into three groups based on `blastp` analyses: **1)** proteins with closest homologs to Rickettsiales ($n = 27$), **2)** proteins present in (some or all) Rickettsiales genomes but divergent from their rickettsial counterparts ($n = 18$), and **3)** proteins unknown from Rickettsiales ($n = 17$). Plotting these CDS by their %GC revealed that the Rickettsiales-like proteins differed from the genes within the other two categories, possessing a mean %GC consistent with Rickettsiales genomes (**fig. 6**). Importantly, the mean %GC of these rickettsial-like CDS is the same as the mean %GC of the RETA Core Dataset CDS (29%), suggesting that at least these 27 CDS of the Accessory Dataset likely belong to RETA.

Phylogeny estimations substantiated the rickettsial nature of the entire 27 rickettsial-like CDS (**supplementary figure S6**). Most of these "rickettsial signature" CDS encode secretion system proteins (*rvh* T4SS, *apr*-like T1SS), variable transporters involved in metabolite scavenging from the host (*gltP*, *tlc1*, 2A0306, *citT*), and proteins involved in intracellular growth and survival (*spoT*, *sodA*, *proP*, patatins [cd07199], *ampG*, *mdlB*) (**supplementary table S5**). In addition, six other CDS in this set provided overwhelming evidence for the presence of a rickettsial symbiont associated with *T. adhaerens*. The *parA* gene, which is ubiquitous on rickettsial plasmids **(Baldridge et al. 2010, Gillespie et al. 2012a)**, hints at a possible plasmid associated with RETA. Furthermore, this CDS is fused with a short open reading frame encoding the antitoxin HicB, and *parA* and *hicB* are adjacent on the plasmids carried by "*Candidatus* Rickettsia amblyommii" strains, *R. massiliae* strains, and *R. rhipicephali*. PRK06567 encodes a putative bifunctional protein with both glutamate synthase subunit β (GltD) and 2-polyprenylphenol hydroxylase (UbiB) domains. Chimeric GltD-UbiB proteins are encoded in nearly every Rickettsiales genome (save *O. tsutsugamushi*), but sparsely encoded in other proteobacterial genomes and unknown from other bacterial genomes. The estimated phylogeny of p-stomatins, members of the SPFH (Stomatin-Prohibitin-Flotillin-HflC/K superfamily **(Hinderhofer et al. 2009)**, grouped the RETA protein with *M. mitochondrii* and Rickettsiaceae in a clade that is the likely origin of eukaryotic SPFH members **(Thiergart et al. 2012)**. Finally, three components of the bacterial flagella system (FlgG, FliG, FliH) grouped with homologs

from *M. mitochondrii* in the estimated tree, with this clade ancestral to all other alphapropteobacterial lineages containing flagella. Importantly, flagella were unknown from Rickettsiales until the recent genomes of *M. mitochondrii* and *O. thessalonicensis* revealed their presence **(Georgiades et al. 2011, Sassera et al. 2011)**, with a recent study demonstrating the expression of several *M. mitochondrii* flagellar components **(Mariconti et al. 2012)**. The putative presence of these three genes in RETA suggests that flagella may be common among rickettsial species outside of the traditional Rickettsiaceae and Anaplasmataceae.

The 18 CDS with divergent rickettsial homologs have a mean %GC higher than rickettsial genomes (39%) (**fig. 6**). Oddly, two (*trpS* and *rnhA*) were determined to be identical to sequences from the genome of the anaerobic thermohalophile *Halothermothrix orenii* H 168 (Firmicutes: Haloanaerobiales) **(Mavromatis et al. 2009)**. As both of these CDS are on the same *T. adhaerens* scaffold (NW_002061683), and the genes in the *H. orenii* genome are neighbors, these CDS are likely contamination. For the other 16 CDS, phylogeny estimations placed them in clades distinct from the Rickettsiales homologs (supplementary figure S7). While no dominant phylogenetic signal from any taxonomic group was observed, two genes (*cox3*, *rsmE*) and one gene fusion (*fkpA-trmH*) have clear chlamydial origins. Despite *fkpA* and *trmH* being adjacent in many chlamydial genomes, no chlamydial 16S rDNA was found in the *T. adhaerens* trace reads, casting doubt on the presence of a chlamydial species within the assembly. Encoding a chloroplast-targeted RpsP, *rpsP* was determined to be of algal (Haptophyceae) origin, consistent with the mining of cyanobacterial-like (chloroplast) 16S rDNA sequences (**fig. 2a**). Thus, it is likely that a small amount of algal contamination was present in the material used for sequencing, possibly originating from the food source of the cultured *T. adhaerens* laboratory colony. RETA proteins for the remaining 11 CDS (CcoP, PleD_2, MhpC, GNAT_1-2, PhrB_1-4, TraU, and PotA) grouped with various different bacterial species in the estimated trees, none of which had a clear association with the 16S rDNA sequences mined from the *T. adhaerens* trace reads.

Finally, the 17 CDS unknown from Rickettsiales had best scoring `blastp` hits to a wide range of bacterial species, most of which had no had clear associations with the extracted 16S rDNA sequences. However, three CDS (COG1506, hypothetical protein RETA0107 and PleD_1) had high similarity to sequences from aquatic Alteromonadales genomes, with the latter two proteins having 99% identity with sequences from the genome of *Alteromonas macleodii* ATCC 27126, a marine planktonic copiotroph **(Ivars-Martinez et al. 2008)**. This suggests likely contamination of the *T. adhaerens* assembly with at least one species of aquatic Alteromonadales, given the mining of three 16S rDNA sequences from this group (**fig. 2a**). Of note in this set, four sequences of a transposase (ISL3) of cyanobacterial origin, and a protein encoding SEL1 and other tetratricopeptide [TPR] motifs (COG0790), are strong candidates for being components of the RETA accessory genome. Several larger rickettsial genomes (e.g., REIS, *R. massiliae*, *R. felis* and *R. bellii*) have accessory genomes comprised of transposases of diverse origins, including Cyanobacteria **(Ogata et al. 2005, Ogata et al. 2006, Blanc et al. 2007, Gillespie et al. 2012a)**. Furthermore, bacterial proteins encoding eukaryotic domains (i.e., SEL1 and TPR motifs) are characteristic of a variety of intracellular species, some of which show evidence for extensive LGT with rickettsial genomes **(Schmitz-Esser et al. 2010, Penz et al. 2012)**. Importantly, aside from the probable contamination from algal, Alteromonadales and *H. orenii* genomes, most of the CDS either lacking or having divergent rickettsial homologs should be considered as possible components of the RETA accessory genome, given that diverse

elements are known to be encoded within rickettsial genomes (e.g., the aminoglycoside antibiotic biosynthesis cluster of REIS **(Gillespie et al. 2012a)**).

Bacterial Genes in the *Trichoplax* Genome

To determine if any of the 181 RETA CDS might instead represent lateral transfers into the *T. adhaerens* genome, we evaluated the assembly scaffolds that contained these CDS (**fig. 7a**). A total of 79.5% (n = 144) of the RETA CDS were present on either all-bacteria (n = 59) or singleton (n = 85) scaffolds. These CDS were removed from consideration, since LGT events could only be predicted on scaffolds with eukaryotic-like genes. Assessment of the non-RETA genes on small hybrid scaffolds did not reveal any strong candidates for complete eukaryotic-like genes, being comprised mostly of small *T. adhaerens* ORFans and partial sequences with little or no similarity to sequences in the *nr* database (data not shown). Thus, these CDS (n = 19) were likewise removed from consideration as LGT events. In further support of the above, none of the intron-containing RETA genes within these three small scaffold categories were determined to contain plausible introns using a bacterial gene prediction program (**supplementary table S2**). Of the four scaffold categories we created, only 10% (n = 18) of the total RETA CDS were found on large scaffolds dominated by eukaryotic-like genes, and were thereby amenable to LGT analysis.

The eight large eukaryotic-like scaffolds that include the 18 RETA CDS that we analyzed for LGT properties are major components of the *T. adhaerens* assembly, containing from 41-1541 genes (**fig. 7b**). Phylogeny estimations for all 18 RETA proteins (**supplementary figure S8**) clearly indicate that six of the CDS (*leuS*, *mutS*, *tilS*, *rpmA*, *prfA*, and *rplQ*) are eukaryotic genes that were erroneously obtained by our pipeline. Five of these genes (*leuS*, *mutS*, *tilS*, *rpmA*, *prfA*) grouped with eukaryotic genes encoding counterparts that are predicted to be imported by the mitochondria, with the latter four having rickettsial origins. This explains their initial characterization as RETA genes. Bacterial gene predictions across the scaffold regions encoding all of these genes, coupled with manual assessment, determined that the introns within all six genes are valid. In addition, homologs to all of these genes were identified in MitoCarta, a compendium of mammalian nuclear-encoded genes with strong support for mitochondrial localization **(Pagliarini et al. 2008)**. When we excluded these six genes from the RETA Core Dataset and re-estimated the genome phylogeny (as described above), the resultant tree (data not shown) had an identical topology to the one including these genes (**fig. 5**), suggesting the minimal eukaryotic signal within the RETA Core Dataset did not override the predominantly rickettsial signal.

Phylogeny estimations for the remaining 12 RETA CDS present on large eukaryotic-like scaffolds strongly implied that all of these genes are bacterial-like and present among eukaryotic genes within the *T. adhaerens* genome (**supplementary figure S8**). If they are indeed LGTs, ten of these genes are likely recent transfers, as they do not contain introns or other eukaryotic-like features (e.g., eukaryotic secretion signals). *dapF* was determined to encode a complete DapF-like protein when predicted as a bacterial gene, ruling out its two predicted introns. However, the single introns splitting the two coding regions within BPL_N_1 and BPL_N_2 are localized within conserved sites as compared with closely related BPL_N sequences (**supplementary figure S9**). The two BPL_N proteins are divergent from one another (45% amino acid identity), yet form a clade together with proteins encoded in *Rickettsia* and Chlamydiae genomes (**supplementary figure S8**). Furthermore, the sizes of both proteins (BPL_N_1, 267 aa; BPL_N_2, 264 aa) are consistent with bacterial BPL_N proteins rather than the larger eukaryotic

proteins that include a BirA-like domain in conjunction with the N-terminal region solely encoded by prokaryotic BPL_N proteins. Thus, the genes encoding BPL_N_1 and BPL_N_2 appear to be LGT products from bacteria that are undergoing transitions to eukaryotic-like gene structures.

Aside from the BPL_N proteins, genes encoding DapF and MurA also appear to be LGT products with rickettsial origins. Thus, all four of these *T. adhaerens* genes may be transfers from RETA and involved in maintenance of the symbiosis. The gene encoding rRNA small subunit methyltransferase E (RsmE) is clearly of chlamydial origin, and a strong candidate for LGT since the only eukaryotic-like *rsmE* genes encode products shipped to chloroplasts. The remaining candidate LGT products encode DNA photolyase repair enzymes (PhrB_1-4), GCN5-related N-Acetyltransferases (GNAT_1-2) and a bacterial lysophospholipase (MhpC). The phylogeny estimations for all of these proteins suggest a substantial degree of LGT underlying their distribution across bacteria, with several intracellular species grouping close to the RETA proteins. Collectively, our analysis of potential bacteria-to-*T. adhaerens* LGT events yielded 12 RETA genes that are clearly bacterial in origin and apparently encoded *within* the *T. adhaerens* genome.

DISCUSSION

For a variety of reasons (e.g., contamination, failure to purify the target organism from environmental microbiota, low-level capture of endosymbionts, LGT), eukaryotic genome sequencing projects often include DNA sequences from the non-target organism. Contigs and scaffolds are typically filtered from the assembly if there is evidence for organellar contamination, sequencing artifacts (e.g., bacterial and phage cloning vectors), and/or prokaryotic contamination. These sequences remain available within the trace read archives, and several studies have utilized these resources to assemble bacterial genomes and associate eukaryotes with their resident microbes **(Salzberg et al. 2005b, Salzberg et al. 2009)**. Still, other studies have capitalized on the concomitant sequencing of host and associated microbes and reported the presence of these microbes in conjunction with the eukaryotic sequencing project **(Chapman et al. 2010, Gillespie et al. 2012a)**. Whatever the route for identification of microbial DNA generated via eukaryotic genome sequencing projects, it is clear that it is no longer a rare event, and that methods are needed to facilitate the process of sifting out "who is who" amongst the generated sequence data. These methodological approaches will not only be applicable for identifying and analyzing microbial species from eukaryotic genome sequencing projects, but will also be practical for effectively processing data from metagenome and microbiotic studies **(Iverson et al. 2012)**.

Mining Bacterial Genes from the *Trichoplax* Genome

As an aquatic animal, *T. adhaerens* is known to feed on green algae (Chlorophyta), cryptomonad (Cryptophyta) species of the genera *Cryptomonas* and *Rhodomonas*, Cyanobacteria, and detritus from other organisms **(Schierwater and Kuhn 1998, Schierwater 2005)**. The "Grell" strain of *T. adhaerens*, which was fed a monoculture of the cryptophyte alga *Pyrenomonas helgolandii*, was the source for genome sequencing **(Srivastava et al. 2008)**. Despite the effort to purify tissues prior to genome sequencing, the results of our study show that a small, yet detectable portion of the generated sequence data was from a different source(s) than *T. adhaerens*. This is

not surprising, as the animals were not cultured axenically and may be associated with other free-living organisms in culture.

Our analysis of the trace read archive and genome assembly revealed considerable microbial diversity associated with the *T. adhaerens* sequencing project (**fig. 2**). According to the mined 16S rDNA sequences, many of the organisms with the highest sequence similarity are aquatic and probably inhabit similar niches as *T. adhaerens* (e.g., *Marivita* spp., *Alteromonas* spp, haptophyte and cryptomonad algae, heterokonts). Other 16S rDNA sequences extracted from the trace read archive (e.g., *Borrelia* spp., *Lawsonia intracellularis*, etc.) are harder to explain as environmental, and probably do not reflect any direct biological associations with *T. adhaerens*. Their presence suggests a low level of microbial contamination, as we did not find any substantial evidence for matching CDS in the assembly or trace read archive for these organisms. Notwithstanding, this minimal information may prove useful in future studies of placozoan biology, particular regarding ecological interactions, and it may also be important for testing the interpretations that we have presented here.

Importantly, the 181 CDS bacterial-like sequences we obtained from the *T. adhaerens* sequencing project were identified from the published assembly. According to the inclusion criteria from the original study **(Srivastava et al. 2008)**, scaffolds were removed from the assembly if they 1) were shorter than 1 Kb in total length, 2) were suspected to be organellar contaminants, and/or 3) possessed a distinct GC content or a prevalence of BLASTN alignments to prokaryotic genomes. In our analysis, 88.4% (n = 160) of bacterial-like CDS mined from the assembly were on small scaffolds (mean length of 2.7 Kb) just over the exclusion cutoff, with the remaining 21 CDS on much larger scaffolds. Only one CDS (a second copy of *rpsP*) was determined likely to originate from chloroplasts of haptophytic algae (Eukaryota; Haptophyceae). Finally, the mean %GC of all 181 CDS (31.7) was highly similar to that of the *T. adhaerens* genome (32.7), masking their detection as distinct non-*T. adhaerens* genes. Collectively, these characteristics of the identified bacterial-like CDS likely account for their inclusion in the *T. adhaerens* assembly.

Evidence for a Rickettsiales Endosymbiont of *Trichoplax adhaerens*

The rickettsial 16S rDNA sequence mined from the *T. adhaerens* sequence read archive was determined to be most similar to sequences isolated from the marine sponge *Cymbastela concentrica* (99% nt identity). Other highly similar 16S rDNA sequences have been reported from *Hydra oligactis*, diverse coral species and unknown hosts from environmental samplings **(Fraune and Bosch 2007, Revetta et al. 2010, Sunagawa et al. 2010, Revetta et al. 2011)**, and all of these sequences grouped in a clade outside of the well-studied rickettsial families Rickettsiaceae and Anaplasmataceae in our estimated phylogeny (**fig. 3**). A recent study on the microbiome of the euglenoid alga *Eutreptiella* sp. revealed the presence of a rickettsial symbiont **(Kuo and Lin 2012)** that is also a member of this clade (tree not shown; 97% nt identity to RETA 16S rDNA, NCBI acc. no. JQ337869). Thus, RETA belongs to a poorly known rickettsial lineage comprised of species associated with eukaryotic organisms from aquatic environments, consistent with the marine niche of *T. adhaerens*. Importantly, this lineage is divergent from other members of the "Midichloriaceae", including *M. mitochondrii*, and suggests that a substantial amount of diversity and host range underlay this large rickettsial assemblage **(Gillespie et al. 2012b)**.

The identification of a rickettsial 16S rDNA sequence was important for corroborating both the long-known presence of a Gram-negative intracellular symbiont in *T. adhaerens* **(Grell 1972,**

Grell and Benwitz 1974, Eitel et al. 2011) and the previous detection of rickettsial-like CDS associated with this genome **(Felsheim et al. 2009, Baldridge et al. 2010, Gillespie et al. 2010, Nikoh et al. 2010)**. It also steered the approach for analyzing the *T. adhaerens* assembly with the understanding that this rickettsial species was likely divergent from those with available genome sequences. The majority (81%, $n = 146$) of the 181 CDS mined from the *T. adhaerens* assembly indeed showed a consistent rickettsial signal. We observed that *M. mitochondrii* was the most similar rickettsial species to RETA with an available genome sequence, and subsequently mapped 138 of the total RETA CDS to the *M. mitochondrii* chromosome (**fig. 4**). In fact, the publication of the *M. mitochondrii* genome **(Sassera et al. 2011)** was an invaluable resource for effectively mining the RETA genes from the *T. adhaerens* assembly, as attempts prior to the availability of this genome yielded far fewer genes of interest. Principally, the strategy to mine rickettsial-like CDS from the *T. adhaerens* assembly was plagued by the divergent nature of the bacterial-like CDS, and in many instances (e.g., universal proteins, nuclear encoded organellar proteins) it was difficult to discern between top `blastp` subjects in the eukaryotic and prokaryotic databases. Thus, a thorough manual component of our methodology was invoked, coupled with rigorous phylogeny estimation of all the CDS. Calculating an approximation of the genome divergence across RETA and the major Rickettsiales lineages illustrated that the observed divergence of RETA genes, in relation to other lineages, is consistent for rickettsial genomes (**table 1**). Collectively, the divergent nature of the mined RETA genes is in agreement with the SSU rDNA-based phylogeny that suggests the RETA-containing group of aquatic Rickettsiales is a well-diverged clade within the "Midichloriaceae".

A majority (66%, $n = 119$) of the RETA CDS, named the Core Dataset, contained characteristics of vertically inherited alphaproteobacterial genes; as such, their predicted proteins were used to estimate a robust genome-based phylogeny across *Alphaproteobacteria* (**fig. 5**). This phylogeny estimation grouped RETA with *M. mitochondrii*, to the exclusion of the derived rickettsial families Anaplasmataceae and Rickettsiaceae. Without genome sequences for any other members of the "Midichloriaceae", coupled with the limited data identified for RETA, it is difficult to make conclusions regarding the factors that distinguish this group from the other derived Rickettsiales lineages. However, our genome-based phylogeny estimation did agree with the overall higher level divergences within Rickettsiales based on the SSU rDNA tree (**fig. 3**), especially regarding the ancestral position of Holosporaceae and the branching point of the mitochondrial ancestor prior to the diversification of the Anaplasmataceae, Rickettsiaceae and "Midichloriaceae". The uniqueness of Rickettsiales within the *Alphaproteobacteria*, particularly regarding the relatedness of Holosporaceae, will become clearer with the generation of more genome sequences from this poorly understood taxon. Notwithstanding, the clade containing RETA and its aquatic relatives will provide useful genomic information regarding the diversification of the three derived rickettsial families, especially regarding the evolution of vertebrate pathogenicity from a seemingly vast array of obligate intracellular symbionts of virtually every major eukaryotic lineage **(Gillespie et al. 2012b)**.

The remaining 62 identified CDS, named the Accessory Dataset (**fig. 6**), provided additional evidence for a rickettsial symbiont associated with *T. adhaerens*. Nearly half (43.5%, $n = 27$) of these genes are signatures of all or some rickettsial genera (**supplementary table S5**), and phylogeny estimation unambiguously supports the rickettsial nature of these RETA CDS (**supplementary figure S6**). In conjunction with the genes of the Core Dataset, a profile emerged implicating the probable metabolic dependency of RETA on eukaryotic cells; e.g., the

presence of genes for the uptake of host ATP (*tlc1*), carbohydrates (*citT*) and amino acids (2A0306, *proP*, *gltP*). Other rickettsial signature genes encode proteins involved in the establishment of osmoregulation (*proP*), antioxidant defense (*sodA*), regulation of the stringent response (*spoT*), and peptidoglycan recycling/β-lactamase induction (*ampG*). Aside from the six rickettsial hallmark genes described above that encode ParA-HicB, PRK06567, p-stomatin, and several flagellar proteins (FlgG, FliG, FliH), two additional rickettsial signatures are noteworthy. First, a gene encoding a patatin phospholipase (cd07199) was identified that is highly similar to the rickettsial Pat1 proteins recently demonstrated to function in the invasion of host cells **(Rahman et al. 2013)**. Second, a gene was identified that encodes a MdlB-like transporter with unknown specificity that is present within the genomes of many intracellular bacterial species **(Gillespie et al. 2012a)**. We have previously identified many of these rickettsial signatures (e.g., Tlc, ProP, GltP, SpoT, MdlB) as components of integrative conjugative elements **(Gillespie et al. 2012a)**. Their presence in RETA, and phylogeny estimations provided in this study (**supplementary figure S6**), suggests that these genes are widely dispersed across Rickettsiales and are likely critical factors that orchestrate the obligate intracellular life cycle of these bacteria. Aside from broadening our perspective on rickettsial biology and genomics, particularly the range and nature of the rickettsial mobilome, these signatures will be useful for identifying Rickettsiales in future metagenomic, microbiome and environmental studies.

In relation to the phylogeny of Rickettsiales, two notable observations were made from the analysis of the Accessory Dataset. First, several RETA genes were identified that encode components of the *rvh* T4SS **(Gillespie et al. 2009, Gillespie et al. 2010)**; however, we did not find any of the *rvh* genes in the genome of *O. thessalonicensis*, suggesting that either the rickettsial lineages branching off after Holosporaceae acquired a P-T4SS, or the Holosporaceae have secondarily lost the P-T4SS. Second, regarding the flagella system genes identified for RETA, only *M. mitochondrii* and *O. thessalonicensis* are known to contain flagellar genes among Rickettsiales. This suggests that the two well-studied groups Anaplasmataceae and Rickettsiaceae have secondarily lost the requirement for flagella, moving effectively within and across eukaryotic cells without them. More genome sequences from the Holosporaceae and "Midichloriaceae" are needed to test these evolutionary scenarios for the gain and loss of P-T4SSs and flagella across Rickettsiales.

Finally, 56.5% ($n = 35$) of the Accessory Dataset is comprised of CDS that are not rickettsial in origin, with 17 of these genes not having any significant homologs in Rickettsiales genomes. Some of these CDS, which are highly similar to sequences from algal, Alteromonadales and *H. orenii* genomes, hint at a low level of contamination within the *T. adhaerens* assembly. However, the majority of these mined CDS may depict genes of the RETA accessory genome, especially considering that none of them have corresponding 16S rDNA sequences extracted from the trace read archive. Additionally, these CDS do not encode proteins that expand the typical metabolic capacity of Rickettsiales genomes. The complete sequencing of the RETA genome will be essential for determining if these sequences are indeed encoded within its genome; regardless, even excluding these tenuous CDS of the Accessory Dataset, the rickettsial-like CDS of the Accessory Dataset together with all CDS of the Core Dataset provide substantial evidence for RETA as the intracellular denizen of *T. adhaerens* fiber cells.

Bacterial Genes Encoded in the *T. adhaerens* Genome

Our analysis of the 181 mined bacterial-like CDS revealed that 12 of these genes are present on large scaffolds primarily composed of eukaryotic-like genes (**fig. 7**). Phylogeny estimation of

these LGT products suggests that four of these genes encode proteins (BPL_N_1, BPL_N_2, DapF, and MurA) with strong homology to rickettsial counterparts, possibly reflecting RETA transfers to the *T. adhaerens* genome (**supplementary figure S8**). While the two BPL_N encoding genes have acquired introns (**supplementary figure S9**), they are likely not complementing any host deficiency related to biotin ligation that is associated with eukaryotic BirA proteins that carry the BPL_N domain. Three lines of evidence support this hypothesis. First, the genes only encode the N-terminal domain of BirA proteins, which is not the region responsible for ligating biotin to biotin-dependent enzymes. Instead, this domain has homology to type 1 glutamine amidotransferases (cd03144), which function in the transfer of the ammonia groups of Gln residues to other substrates. Second, despite the presence of introns, both BPL_N genes encode predicted proteins similar in size (~250 aa) to bacterial-like BPL_N proteins, none of which contain a BirA domain. Finally, the *T. adhaerens* genome encodes a separate BirA gene (XP_002116544) located on a scaffold (NW_002060958) distinct from those encoding the BPL_N genes, with the encoded protein 44% identical to the human holocarboxylase synthetase. This suggests that *T. adhaerens* is capable of ligating biotin to biotin-dependent enzymes without the need for BPL_N proteins.

The presence in the *T. adhaerens* genome of two rickettsial genes encoding cell envelope synthesis enzymes (DapF, and MurA) is reminiscent of aphid (LdcA, RlpA, AmiD) and rotifer (Ddl) genomes, wherein such bacterial-like genes are functional **(Gladyshev et al. 2008, Nikoh and Nakabachi 2009, Nikoh et al. 2010)**. In the case of the aphid, *ldcA*, *rlpA* and *amiD* have acquired introns and (in some instances) eukaryotic signal sequences and are highly expressed in bacteriocytes that house its obligate symbiont *B. aphidicola*. This remarkable phenomenon likely restricts symbiont growth to these specific tissues **(Nikoh et al. 2010)**. It is tempting to speculate that putative transfers of RETA *dapF* and *murA* genes to the *T. adhaerens* genome may operate in a similar fashion. Recent images of the *T. adhaerens* symbiont in oocytes (transferred from fiber cell extensions) **(Eitel et al. 2011)** show a Gram-negative cell envelope structure, suggesting the presence of peptidoglycan. Thus, tissue specific expression of *murA* and *dapF* by *T. adhaerens* could limit the growth of RETA to the fiber cells, which are primarily where the symbiotic bacteria are observed **(Grell 1972, Grell and Benwitz 1974)**.

Four of the eight remaining putative LGT products encode copies of a bacterial-like photolyase (PhrB_1-4) that may play a functional role in *T. adhaerens*, or the maintenance of RETA, or both. PhrB is a photoreactivation enzyme, functioning in the repair of pyrimidine dimers **(Schul et al. 2002)**. Genes encoding these enzymes are frequently detected in marine metagenomic studies and are likely important factors of aquatic microbes inhabiting surface waters **(DeLong et al. 2006, Frias-Lopez et al. 2008, Singh et al. 2009)**. Photolyase genes are known to be differentially gained and lost throughout bacterial evolution **(Lucas-Lledo and Lynch 2009)**, and our phylogeny estimation suggests a substantial degree of LGT shaping the distribution of these genes in bacteria (**supplementary figure S8**). Furthermore, it was recently demonstrated that species of the SAR11 clade of *Alphaproteobacteria* encode some PhrB and PhrB-like genes that are most similar to those encoded within cynaobacterial genomes **(Viklund et al. 2012)**. Thus, transfer to the *T. adhaerens* genome of these genes from a non-rickettsial source could function to repair DNA damage to its resident symbiont, particularly if there is limited symbiont gene flow due to its vertical transmission. Alternatively, the PhrB proteins may function to benefit *T. adhaerens* in its aquatic niche if it is vulnerable to substantial UV light. It has been proposed that loss of photolyases in eukaryotic species would enhance deleterious

mutation rates **(Lucas-Lledo and Lynch 2009)**, and given that we did not detect any PhrB or related cryptochrome genes within the *T. adhaerens* genome (data not shown), the four bacterial-like *phrB* genes may be under strong selective pressure to repair damage to *T. adhaerens* genes.

The remaining four putative LGT products (GNAT_1, GNAT_2, MhpC, RsmE) are not of rickettsial origin, and encode enzymes known solely from prokaryotes or plastids; consequently, it is difficult to envision these bacterial-like genes complementing functions lost in the *T. adhaerens* genome. They may, however, function in the maintenance of RETA, despite their various evolutionary sources, possibly representing parts of the RETA accessory genome that were laterally acquired and subsequently transferred to the *T. adhaerens* genome. There is precedence for such an event; for instance, most of the identified gene transfers to the aphid genome that support its gammaproteobacterial symbiont (*B. aphidicola*) are of rickettsial and not gammaproteobacterial origin **(Nikoh et al. 2010)**.

Whatever the source of the 12 identified bacterial-like genes encoded in the *T. adhaerens* genome, their significance in placozoan biology and possible role they may play in fostering its symbiosis with RETA is an exciting area of future research. Importantly, if LGTs to the *T. adhaerens* genome support the maintenance of a rickettsial symbiont, it may be that growth of RETA in other eukaryotic cells or cell lines may not be possible without these *T. adhaerens* bacterial genes.

CONCLUSION

From the genomic sequence data generated for a placozoan (*Trichoplax adhaerens*), we identified and analyzed bacterial DNA sequences and compiled evidence for a rickettsial endosymbiont of *T. adhaerens*. This genomic profile of RETA potentially confirms the long suspected presence of a bacterial symbiont associated primarily with *T. adhaerens* fiber cells. Based on available genome sequences for Rickettsiales, all of which are either obligate intracellular symbionts or pathogens of various metazoan species, the data we present here likely depicts approximately 20% of the entire RETA genome. Despite the lack of a complete genome, our phylogeny estimations and other analyses place RETA solidly within the "Midichloriaceae" clade. This rickettsial lineage is poorly understood, but based on the closest relative with an available genome sequence (*M. mitochondrii*), is quite different than the >80 genome sequences currently available for the well-studied species of the Anaplasmataceae and Rickettsiaceae. Thus, a better understanding of the “Midichloriaceae” clade of Rickettsiales is needed, particularly for 1) highlighting features involved in vertebrate pathogenicity in species of Anaplasmataceae and Rickettsiaceae, 2) determining the diversifying factors that define obligate intracellular life cycles of a wide range of eukaryotic hosts, and 3) deciphering the processes that shaped the transition of a rickettsial symbiont to an organelle (mitochondria) of eukaryotic cells.

Placozoans are an early-branching metazoan lineage **(Miller and Ball 2008, Schierwater et al. 2009, Schierwater and Kamm 2010)**, and publication of the *T. adhaerens* genome sequence has proven invaluable to the fields of evolutionary genetics, developmental biology and animal phylogenetics, among others. However, given previous morphological evidence, as well as our data presented for RETA, the relevance of a symbiotic genome must be included in discussions and analyses of placozoan biology. Of primary importance is understanding whether or not RETA is primitive in gene repertoire compared to other rickettsial species that invade more complicated metazoan species with much more diverse cellular networks. Also critical is determining the function of bacterial-like genes encoded in the *T. adhaerens* genome, and

whether these functions pertain to the placozoan itself or symbiosis with RETA (or both). Given the complex genetics underlying the aphid-*Buchnera* mutualism (discussed above) and other metazoan-bacterial symbioses **(Woyke et al. 2006, Wu et al. 2006, McCutcheon and Moran 2007, McCutcheon et al. 2009, McCutcheon and Moran 2010, McCutcheon and von Dohlen 2011, Engel et al. 2012)**, as well as the recent demonstration of a tripartite metabolic co-dependency across genomes of an obligate biotrophic fungus (*Gigaspora margarita*), its plant hosts, and its betaproteobacterial endosymbiont (*Candidatus* Glomeribacter gigasporarum) **(Ghignone et al. 2012)**, it is clear that symbioses pose challenges for high-throughput genomics. Complete metabolic and cellular pathways can only be garnered by obtaining the genes from all participant genomes that underlay the complete "organism", particularly if some (or all) of the players are inextricably tied together. Thus, isolating RETA from *T. adhaerens* and determining the nature of this symbiosis is critical for understanding this aspect of placozoan biology.

ACKNOWLEDGMENTS

We thank members of the Cyberinfrastructure Division (Virginia Bioinformatics Institute at Virginia Tech) and the Azad laboratory (U. Maryland, School of Medicine) for invaluable feedback and discussion throughout the duration of this project. We are grateful to Shrinivasrao Mane (Dow AgroSciences) for bioinformatics assistance in the preliminary phase of this work. We thank Bernd Schierwater and Tina Herzog (University of Veterinary Medicine Hannover, Foundation) for providing literature on *Trichoplax adhaerens*, and Leo Buss (Yale University) for fielding inquiries about placozoan biology.

This work was supported by the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH), Department of Health and Human Services [HHSN272200900040C to B.W.S., R01AI017828 and R01AI59118 to A.F.A.]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIAID or the NIH.

LITERATURE CITED

2012. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res **40**:D71-75.

Acuna, R., B. E. Padilla, C. P. Florez-Ramos, J. D. Rubio, J. C. Herrera, P. Benavides, S. J. Lee, T. H. Yeats, A. N. Egan, J. J. Doyle, and J. K. Rose. 2012. Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. Proc Natl Acad Sci U S A **109**:4197-4202.

Aikawa, T., H. Anbutsu, N. Nikoh, T. Kikuchi, F. Shibata, and T. Fukatsu. 2009. Longicorn beetle that vectors pinewood nematode carries many *Wolbachia* genes on an autosome. Proc Biol Sci **276**:3791-3798.

Aziz, R. K., D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass, M. Kubal, F. Meyer, G. J. Olsen, R. Olson, A. L. Osterman, R. A. Overbeek, L. K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G. D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke, and O. Zagnitko. 2008. The RAST Server: rapid annotations using subsystems technology. BMC Genomics **9**:75.

Baldridge, G. D., N. Y. Burkhardt, M. B. Labruna, R. C. Pacheco, C. D. Paddock, P. C. Williamson, P. M. Billingsley, R. F. Felsheim, T. J. Kurtti, and U. G. Munderloh. 2010. Wide dispersal and possible multiple origins of low-copy-number plasmids in rickettsia species associated with blood-feeding arthropods. Appl Environ Microbiol **76**:1718-1731.

Beninati, T., N. Lo, L. Sacchi, C. Genchi, H. Noda, and C. Bandi. 2004. A novel alpha-Proteobacterium resides in the mitochondria of ovarian cells of the tick Ixodes ricinus. Appl Environ Microbiol **70**:2596-2602.

Birtles, R. J., T. J. Rowbotham, R. Michel, D. G. Pitcher, B. Lascola, S. Alexiou-Daniel, and D. Raoult. 2000. '*Candidatus* Odyssella thessalonicensis' gen. nov., sp. nov., an obligate intracellular parasite of *Acanthamoeba* species. Int J Syst Evol Microbiol **50 Pt 1**:63-72.

Blanc, G., H. Ogata, C. Robert, S. Audic, J. M. Claverie, and D. Raoult. 2007. Lateral gene transfer between obligate intracellular bacteria: evidence from the *Rickettsia massiliae* genome. Genome Res **17**:1657-1664.

Boscaro, V., S. I. Fokin, M. Schrallhammer, M. Schweikert, and G. Petroni. 2012. Revised Systematics of *Holospora*-Like Bacteria and Characterization of "*Candidatus* Gortzia infectiva", a Novel Macronuclear Symbiont of *Paramecium jenningsi*. Microb Ecol.

Boschetti, C., A. Carr, A. Crisp, I. Eyres, Y. Wang-Koh, E. Lubzens, T. G. Barraclough, G. Micklem, and A. Tunnacliffe. 2012. Biochemical Diversification through Foreign Gene Expression in Bdelloid Rotifers. PLoS Genet **8**:e1003035.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol **17**:540-552.

Chapman, J. A., E. F. Kirkness, O. Simakov, S. E. Hampson, T. Mitros, T. Weinmaier, T. Rattei, P. G. Balasubramanian, J. Borman, D. Busam, K. Disbennett, C. Pfannkoch, N. Sumin, G. G. Sutton, L. D. Viswanathan, B. Walenz, D. M. Goodstein, U. Hellsten, T. Kawashima, S. E. Prochnik, N. H. Putnam, S. Shu, B. Blumberg, C. E. Dana, L. Gee, D. F. Kibler, L. Law, D. Lindgens, D. E. Martinez, J. Peng, P. A. Wigge, B. Bertulat, C. Guder, Y. Nakamura, S. Ozbek, H. Watanabe, K. Khalturin, G. Hemmrich, A. Franke, R. Augustin, S. Fraune, E. Hayakawa, S. Hayakawa, M. Hirose, J. S. Hwang, K. Ikeo, C. Nishimiya-Fujisawa, A. Ogura, T. Takahashi, P. R. Steinmetz, X. Zhang, R. Aufschnaiter, M. K. Eder, A. K. Gorny, W. Salvenmoser, A. M. Heimberg, B. M. Wheeler, K. J. Peterson, A. Bottger, P. Tischler, A. Wolf, T. Gojobori, K. A. Remington, R. L. Strausberg, J. C. Venter, U. Technau, B.

Hobmayer, T. C. Bosch, T. W. Holstein, T. Fujisawa, H. R. Bode, C. N. David, D. S. Rokhsar, and R. E. Steele. 2010. The dynamic genome of *Hydra*. Nature **464**:592-596.
Davis, A. K., J. L. DeVore, J. R. Milanovich, K. Cecala, J. C. Maerz, and M. J. Yabsley. 2009. New findings from an old pathogen: intraerythrocytic bacteria (family Anaplasmatacea) in red-backed salamanders *Plethodon cinereus*. Ecohealth **6**:219-228.
Dellaporta, S. L., A. Xu, S. Sagasser, W. Jakob, M. A. Moreno, L. W. Buss, and B. Schierwater. 2006. Mitochondrial genome of *Trichoplax adhaerens* supports placozoa as the basal lower metazoan phylum. Proc Natl Acad Sci U S A **103**:8751-8756.
DeLong, E. F., C. M. Preston, T. Mincer, V. Rich, S. J. Hallam, N. U. Frigaard, A. Martinez, M. B. Sullivan, R. Edwards, B. R. Brito, S. W. Chisholm, and D. M. Karl. 2006. Community genomics among stratified microbial assemblages in the ocean's interior. Science **311**:496-503.
Deng, W., B. S. Maust, D. C. Nickle, G. H. Learn, Y. Liu, L. Heath, S. L. Kosakovsky Pond, and J. I. Mullins. 2010. DIVEIN: a web server to analyze phylogenies, sequence divergence, diversity, and informative sites. Biotechniques **48**:405-408.
DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol **72**:5069-5072.
Dunning Hotopp, J. C., M. E. Clark, D. C. Oliveira, J. M. Foster, P. Fischer, M. C. Munoz Torres, J. D. Giebel, N. Kumar, N. Ishmael, S. Wang, J. Ingram, R. V. Nene, J. Shepard, J. Tomkins, S. Richards, D. J. Spiro, E. Ghedin, B. E. Slatko, H. Tettelin, and J. H. Werren. 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science **317**:1753-1756.
Durbin, R., S. Eddy, A. Krogh, and G. Mitchison. 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press.
Edgar, R. C. 2004a. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5**:113.
Edgar, R. C. 2004b. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32**:1792-1797.
Eitel, M., L. Guidi, H. Hadrys, M. Balsamo, and B. Schierwater. 2011. New insights into placozoan sexual reproduction and development. PLoS One **6**:e19639.
Engel, P., V. G. Martinson, and N. A. Moran. 2012. Functional diversity within the simple gut microbiota of the honey bee. Proc Natl Acad Sci U S A **109**:11002-11007.
Ewing, B., L. Hillier, M. C. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res **8**:175-185.
Felsheim, R. F., T. J. Kurtti, and U. G. Munderloh. 2009. Genome sequence of the endosymbiont *Rickettsia peacockii* and comparison with virulent *Rickettsia rickettsii*: identification of virulence factors. PLoS One **4**:e8361.
Fraune, S. and T. C. Bosch. 2007. Long-term maintenance of species-specific bacterial microbiota in the basal metazoan *Hydra*. Proc Natl Acad Sci U S A **104**:13146-13151.
Frias-Lopez, J., Y. Shi, G. W. Tyson, M. L. Coleman, S. C. Schuster, S. W. Chisholm, and E. F. Delong. 2008. Microbial community gene expression in ocean surface waters. Proc Natl Acad Sci U S A **105**:3805-3810.

Georgiades, K., M. A. Madoui, P. Le, C. Robert, and D. Raoult. 2011. Phylogenomic analysis of *Odyssella thessalonicensis* fortifies the common origin of Rickettsiales, *Pelagibacter ubique* and *Reclimonas americana* mitochondrion. PLoS One **6**:e24857.

Ghignone, S., A. Salvioli, I. Anca, E. Lumini, G. Ortu, L. Petiti, S. Cruveiller, V. Bianciotto, P. Piffanelli, L. Lanfranco, and P. Bonfante. 2012. The genome of the obligate endobacterium of an AM fungus reveals an interphylum network of nutritional interactions. ISME J **6**:136-145.

Gillespie, J. J., N. C. Ammerman, S. M. Dreher-Lesnick, M. S. Rahman, M. J. Worley, J. C. Setubal, B. S. Sobral, and A. F. Azad. 2009. An anomalous type IV secretion system in *Rickettsia* is evolutionarily conserved. PLoS One **4**:e4833.

Gillespie, J. J., M. S. Beier, M. S. Rahman, N. C. Ammerman, J. M. Shallom, A. Purkayastha, B. S. Sobral, and A. F. Azad. 2007. Plasmids and rickettsial evolution: insight from *Rickettsia felis*. PLoS One **2**:e266.

Gillespie, J. J., K. A. Brayton, K. P. Williams, M. A. Diaz, W. C. Brown, A. F. Azad, and B. W. Sobral. 2010. Phylogenomics reveals a diverse Rickettsiales type IV secretion system. Infect Immun **78**:1809-1823.

Gillespie, J. J., V. Joardar, K. P. Williams, T. Driscoll, J. B. Hostetler, E. Nordberg, M. Shukla, B. Walenz, C. A. Hill, V. M. Nene, A. F. Azad, B. W. Sobral, and E. Caler. 2012a. A *Rickettsia* genome overrun by mobile genetic elements provides insight into the acquisition of genes characteristic of an obligate intracellular lifestyle. J Bacteriol **194**:376-394.

Gillespie, J. J., E. K. Nordberg, A. F. Azad, and B. W. Sobral. 2012b. Phylogeny and Comparative Genomics: The Shifting Landscape in the Genomics Era. Pages 84-141 *in* A. F. Azad and G. H. Palmer, editors. Intracellular Pathogens II: Rickettsiales. American Society of Microbiology, Boston.

Gillespie, J. J., A. R. Wattam, S. A. Cammer, J. L. Gabbard, M. P. Shukla, O. Dalay, T. Driscoll, D. Hix, S. P. Mane, C. Mao, E. K. Nordberg, M. Scott, J. R. Schulman, E. E. Snyder, D. E. Sullivan, C. Wang, A. Warren, K. P. Williams, T. Xue, H. S. Yoo, C. Zhang, Y. Zhang, R. Will, R. W. Kenyon, and B. W. Sobral. 2011. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. Infect Immun **79**:4286-4298.

Gillespie, J. J., K. Williams, M. Shukla, E. E. Snyder, E. K. Nordberg, S. M. Ceraul, C. Dharmanolla, D. Rainey, J. Soneja, J. M. Shallom, N. D. Vishnubhat, R. Wattam, A. Purkayastha, M. Czar, O. Crasta, J. C. Setubal, A. F. Azad, and B. S. Sorbal. 2008. *Rickettsia* Phylogenomics:Unwinding the Intricacies of Obligate Intracellular Life. PLoS One **3**:e2018.

Gladyshev, E. A., M. Meselson, and I. R. Arkhipova. 2008. Massive horizontal gene transfer in bdelloid rotifers. Science **320**:1210-1213.

Grbic, M., T. Van Leeuwen, R. M. Clark, S. Rombauts, P. Rouze, V. Grbic, E. J. Osborne, W. Dermauw, P. C. Ngoc, F. Ortego, P. Hernandez-Crespo, I. Diaz, M. Martinez, M. Navajas, E. Sucena, S. Magalhaes, L. Nagy, R. M. Pace, S. Djuranovic, G. Smagghe, M. Iga, O. Christiaens, J. A. Veenstra, J. Ewer, R. M. Villalobos, J. L. Hutter, S. D. Hudson, M. Velez, S. V. Yi, J. Zeng, A. Pires-daSilva, F. Roch, M. Cazaux, M. Navarro, V. Zhurov, G. Acevedo, A. Bjelica, J. A. Fawcett, E. Bonnet, C. Martens, G. Baele, L. Wissler, A. Sanchez-Rodriguez, L. Tirry, C. Blais, K. Demeestere, S. R. Henz, T. R. Gregory, J. Mathieu, L. Verdon, L. Farinelli, J. Schmutz, E. Lindquist, R. Feyereisen, and Y. Van de

Peer. 2011. The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. Nature **479**:487-492.
Grell, K. G. 1971. *Trichoplax adhaerens*, F.E. Schulze und die Entstehung der Metazoen. Naturwiss. Rundsch. **24**:160–161.
Grell, K. G. 1972. Eibildung und Furchung von *Trichoplax adhaerens* F.E. Schulze (Placozoa). Z. Morph Tiere **73**:297-314.
Grell, K. G. and G. Benwitz. 1974. Spezifische Verbindungsstruktnren der Faserzellen yon *Trichoplax adhaerens* F.E. Schulze. Z. Naturforsch. **29e**:790.
Grell, K. G. and A. Ruthmann. 1991. Placozoa. Pages 13-28 *in* F. W. Harrison and J. A. Westfall, editors. Microscopic Anatomy of Invertebrates. Wiley-Liss, New York.
Guidi, L., M. Eitel, E. Cesarini, B. Schierwater, and M. Balsamo. 2011. Ultrastructural analyses support different morphological lineages in the phylum Placozoa Grell, 1971. J Morphol **272**:371-378.
Hinderhofer, M., C. A. Walker, A. Friemel, C. A. Stuermer, H. M. Moller, and A. Reuter. 2009. Evolution of prokaryotic SPFH proteins. BMC Evol Biol **9**:10.
Ivars-Martinez, E., A. B. Martin-Cuadrado, G. D'Auria, A. Mira, S. Ferriera, J. Johnson, R. Friedman, and F. Rodriguez-Valera. 2008. Comparative genomics of two ecotypes of the marine planktonic copiotroph *Alteromonas macleodii* suggests alternative lifestyles associated with different kinds of particulate organic matter. ISME J **2**:1194-1212.
Iverson, V., R. M. Morris, C. D. Frazar, C. T. Berthiaume, R. L. Morales, and E. V. Armbrust. 2012. Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. Science **335**:587-590.
Juncker, A. S., H. Willenbrock, G. Von Heijne, S. Brunak, H. Nielsen, and A. Krogh. 2003. Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci **12**:1652-1662.
Kall, L., A. Krogh, and E. L. Sonnhammer. 2007. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. Nucleic Acids Res **35**:W429-432.
Kawafune, K., Y. Hongoh, T. Hamaji, and H. Nozaki. 2012. Molecular identification of rickettsial endosymbionts in the non-phagotrophic volvocalean green algae. PLoS One **7**:e31749.
Kembel, S. W., M. Wiu, J. A. Eisen, and J. L. Green. 2012. Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance. PLoS Comp Biol **8**:e1002743.
Kent, W. J. 2002. BLAT--the BLAST-like alignment tool. Genome Res **12**:656-664.
Kondo, N., N. Nikoh, N. Ijichi, M. Shimada, and T. Fukatsu. 2002. Genome fragment of *Wolbachia* endosymbiont transferred to X chromosome of host insect. Proc Natl Acad Sci U S A **99**:14280-14285.
Krogh, A., B. Larsson, G. von Heijne, and E. L. Sonnhammer. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol **305**:567-580.
Krzywinski, M., J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. 2009. Circos: an information aesthetic for comparative genomics. Genome Res **19**:1639-1645.

Kuo, R. C. and S. Lin. 2012. Ectobiotic and Endobiotic Bacteria Associated with *Eutreptiella* sp. Isolated from Long Island Sound. Protist.
Lartillot, N., H. Brinkmann, and H. Philippe. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol **7 Suppl 1**:S4.
Lartillot, N., T. Lepage, and S. Blanquart. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics **25**:2286-2288.
Lartillot, N. and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol **21**:1095-1109.
Lartillot, N. and H. Philippe. 2006. Computing Bayes factors using thermodynamic integration. Syst Biol **55**:195-207.
Li, H. and R. Durbin. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics **26**:589-595.
Li, Z. W., Y. H. Shen, Z. H. Xiang, and Z. Zhang. 2011. Pathogen-origin horizontally transferred genes contribute to the evolution of Lepidopteran insects. BMC Evol Biol **11**:356.
Lo, N., T. Beninati, L. Sacchi, C. Genchi, and C. Bandi. 2004. Emerging rickettsioses. Parassitologia **46**:123-126.
Lucas-Lledo, J. I. and M. Lynch. 2009. Evolution of mutation rates: phylogenomic analysis of the photolyase/cryptochrome family. Mol Biol Evol **26**:1143-1153.
Marchler-Bauer, A., S. Lu, J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, F. Lu, G. H. Marchler, M. Mullokandov, M. V. Omelchenko, C. L. Robertson, J. S. Song, N. Thanki, R. A. Yamashita, D. Zhang, N. Zhang, C. Zheng, and S. H. Bryant. 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res **39**:D225-229.
Mariconti, M., S. Epis, L. Sacchi, M. Biggiogera, D. Sassera, M. Genchi, E. Alberti, M. Montagna, C. Bandi, and C. Bazzocchi. 2012. On the presence of flagella in the Rickettsiales: the case of Midichloria mitochondrii. Microbiology.
Matsuura, Y., Y. Kikuchi, X. Y. Meng, R. Koga, and T. Fukatsu. 2012. Novel clade of alphaproteobacterial endosymbionts associated with stinkbugs and other arthropods. Appl Environ Microbiol **78**:4149-4156.
Mavromatis, K., N. Ivanova, I. Anderson, A. Lykidis, S. D. Hooper, H. Sun, V. Kunin, A. Lapidus, P. Hugenholtz, B. Patel, and N. C. Kyrpides. 2009. Genome analysis of the anaerobic thermohalophilic bacterium *Halothermothrix orenii*. PLoS One **4**:e4192.
McCutcheon, J. P., B. R. McDonald, and N. A. Moran. 2009. Convergent evolution of metabolic roles in bacterial co-symbionts of insects. Proc Natl Acad Sci U S A **106**:15394-15399.
McCutcheon, J. P. and N. A. Moran. 2007. Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. Proc Natl Acad Sci U S A **104**:19392-19397.
McCutcheon, J. P. and N. A. Moran. 2010. Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. Genome Biol Evol **2**:708-718.
McCutcheon, J. P. and C. D. von Dohlen. 2011. An interdependent metabolic patchwork in the nested symbiosis of mealybugs. Curr Biol **21**:1366-1372.
McNulty, S. N., J. M. Foster, M. Mitreva, J. C. Dunning Hotopp, J. Martin, K. Fischer, B. Wu, P. J. Davis, S. Kumar, N. W. Brattig, B. E. Slatko, G. J. Weil, and P. U. Fischer. 2010.

Endosymbiont DNA in endobacteria-free filarial nematodes indicates ancient horizontal genetic transfer. PLoS One **5**:e11029.
Miller, D. J. and E. E. Ball. 2008. Animal evolution: Trichoplax, trees, and taxonomic turmoil. Curr Biol **18**:R1003-1005.
Nikoh, N., J. P. McCutcheon, T. Kudo, S. Y. Miyagishima, N. A. Moran, and A. Nakabachi. 2010. Bacterial genes in the aphid genome: absence of functional gene transfer from *Buchnera* to its host. PLoS Genet **6**:e1000827.
Nikoh, N. and A. Nakabachi. 2009. Aphids acquired symbiotic genes via lateral gene transfer. BMC Biol **7**:12.
Nikoh, N., K. Tanaka, F. Shibata, N. Kondo, M. Hizume, M. Shimada, and T. Fukatsu. 2008. *Wolbachia* genome integrated in an insect chromosome: evolution and fate of laterally transferred endosymbiont genes. Genome Res **18**:272-280.
Ogata, H., B. La Scola, S. Audic, P. Renesto, G. Blanc, C. Robert, P. E. Fournier, J. M. Claverie, and D. Raoult. 2006. Genome sequence of *Rickettsia bellii* illuminates the role of amoebae in gene exchanges between intracellular pathogens. PLoS Genet **2**:e76.
Ogata, H., P. Renesto, S. Audic, C. Robert, G. Blanc, P. E. Fournier, H. Parinello, J. M. Claverie, and D. Raoult. 2005. The genome sequence of *Rickettsia felis* identifies the first putative conjugative plasmid in an obligate intracellular parasite. PLoS Biol **3**:e248.
Ondov, B. D., N. H. Bergman, and A. M. Phillippy. 2011. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics **12**:385.
Pagliarini, D. J., S. E. Calvo, B. Chang, S. A. Sheth, S. B. Vafai, S. E. Ong, G. A. Walford, C. Sugiana, A. Boneh, W. K. Chen, D. E. Hill, M. Vidal, J. G. Evans, D. R. Thorburn, S. A. Carr, and V. K. Mootha. 2008. A mitochondrial protein compendium elucidates complex I disease biology. Cell **134**:112-123.
Penz, T., S. Schmitz-Esser, S. E. Kelly, B. N. Cass, A. Muller, T. Woyke, S. A. Malfatti, M. S. Hunter, and M. Horn. 2012. Comparative Genomics Suggests an Independent Origin of Cytoplasmic Incompatibility in *Cardinium hertigii*. PLoS Genet **8**:e1003012.
Petersen, T. N., S. Brunak, G. von Heijne, and H. Nielsen. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods **8**:785-786.
Price, M. N., P. S. Dehal, and A. P. Arkin. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One **5**:e9490.
Rahman, M. S., J. J. Gillespie, S. J. Kaur, K. T. Sears, S. M. Ceraul, M. S. Beier-Sexton, and A. F. Azad. 2013. *Rickettsia typhi* possesses phospholipase A2 enzymes that are involved in infection of host cells. PLoS Pathogens.
Revetta, R. P., R. S. Matlib, and J. W. Santo Domingo. 2011. 16S rRNA gene sequence analysis of drinking water using RNA and DNA extracts as targets for clone library development. Curr Microbiol **63**:50-59.
Revetta, R. P., A. Pemberton, R. Lamendella, B. Iker, and J. W. Santo Domingo. 2010. Identification of bacterial populations in drinking water using 16S rRNA-based sequence analyses. Water Res **44**:1353-1360.
Rodriguez-Ezpeleta, N. and T. M. Embley. 2012. The SAR11 Group of Alpha-Proteobacteria Is Not Related to the Origin of Mitochondria. PLoS One **7**:e30520.
Sacchi, L., E. Bigliardi, S. Corona, T. Beninati, N. Lo, and A. Franceschi. 2004. A symbiont of the tick Ixodes ricinus invades and consumes mitochondria in a mode similar to that of the parasitic bacterium Bdellovibrio bacteriovorus. Tissue Cell **36**:43-53.

Salzberg, S. L., J. C. Dunning Hotopp, A. L. Delcher, M. Pop, D. R. Smith, M. B. Eisen, and W. C. Nelson. 2005a. Correction: Serendipitous discovery of *Wolbachia* genomes in multiple *Drosophila species*. Genome Biol **6**:402.

Salzberg, S. L., J. C. Dunning Hotopp, A. L. Delcher, M. Pop, D. R. Smith, M. B. Eisen, and W. C. Nelson. 2005b. Serendipitous discovery of *Wolbachia* genomes in multiple *Drosophila species*. Genome Biol **6**:R23.

Salzberg, S. L., D. Puiu, D. D. Sommer, V. Nene, and N. H. Lee. 2009. Genome sequence of the *Wolbachia* endosymbiont of *Culex quinquefasciatus* JHB. J Bacteriol **191**:1725.

Sassera, D., T. Beninati, C. Bandi, E. A. Bouman, L. Sacchi, M. Fabbi, and N. Lo. 2006. 'Candidatus Midichloria mitochondrii', an endosymbiont of the tick Ixodes ricinus with a unique intramitochondrial lifestyle. Int J Syst Evol Microbiol **56**:2535-2540.

Sassera, D., N. Lo, S. Epis, G. D'Auria, M. Montagna, F. Comandatore, D. Horner, J. Pereto, A. M. Luciano, F. Franciosi, E. Ferri, E. Crotti, C. Bazzocchi, D. Daffonchio, L. Sacchi, A. Moya, A. Latorre, and C. Bandi. 2011. Phylogenomic evidence for the presence of a flagellum and cbb(3) oxidase in the free-living mitochondrial ancestor. Mol Biol Evol **28**:3285-3296.

Schierwater, B. 2005. My favorite animal, *Trichoplax adhaerens*. Bioessays **27**:1294-1302.

Schierwater, B., M. Eitel, W. Jakob, H. J. Osigus, H. Hadrys, S. L. Dellaporta, S. O. Kolokotronis, and R. Desalle. 2009. Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. PLoS Biol **7**:e20.

Schierwater, B. and K. Kamm. 2010. The early evolution of Hox genes: a battle of belief? Adv Exp Med Biol **689**:81-90.

Schierwater, B. and K. Kuhn. 1998. Homology of Hox genes and the zootype concept in early metazoan evolution. Mol Phylogenet Evol **9**:375-381.

Schmitz-Esser, S., P. Tischler, R. Arnold, J. Montanaro, M. Wagner, T. Rattei, and M. Horn. 2010. The genome of the amoeba symbiont "*Candidatus* Amoebophilus asiaticus" reveals common mechanisms for host cell interaction among amoeba-associated bacteria. J Bacteriol **192**:1045-1057.

Schul, W., J. Jans, Y. M. Rijksen, K. H. Klemann, A. P. Eker, J. de Wit, O. Nikaido, S. Nakajima, A. Yasui, J. H. Hoeijmakers, and G. T. van der Horst. 2002. Enhanced repair of cyclobutane pyrimidine dimers and improved UV resistance in photolyase transgenic mice. EMBO J **21**:4719-4729.

Singh, A. H., T. Doerks, I. Letunic, J. Raes, and P. Bork. 2009. Discovering functional novelty in metagenomes: examples from light-mediated processes. J Bacteriol **191**:32-41.

Srivastava, M., E. Begovic, J. Chapman, N. H. Putnam, U. Hellsten, T. Kawashima, A. Kuo, T. Mitros, A. Salamov, M. L. Carpenter, A. Y. Signorovitch, M. A. Moreno, K. Kamm, J. Grimwood, J. Schmutz, H. Shapiro, I. V. Grigoriev, L. W. Buss, B. Schierwater, S. L. Dellaporta, and D. S. Rokhsar. 2008. The *Trichoplax* genome and the nature of placozoans. Nature **454**:955-960.

Stamatakis, A., P. Hoover, and J. Rougemont. 2008. A rapid bootstrap algorithm for the RAxML Web servers. Syst Biol **57**:758-771.

Sunagawa, S., C. M. Woodley, and M. Medina. 2010. Threatened corals provide underexplored microbial habitats. PLoS One **5**:e9554.

Talavera, G. and J. Castresana. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol **56**:564-577.

Thiergart, T., G. Landan, M. Schenk, T. Dagan, and W. F. Martin. 2012. An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. Genome Biol Evol **4**:466-485.

Thrash, J. C., A. Boyd, M. J. Huggett, J. Grote, P. Carini, R. J. Yoder, B. Robbertse, J. W. Spatafora, M. S. Rappe, and S. J. Giovannoni. 2011. Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. Sci Rep **1**:13.

Tyson, G. W., J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature **428**:37-43.

Van Dongen, S. 2008. Graph Clustering Via a Discrete Uncoupling Process. SIAM. J. Matrix Anal. & Appl. **30**:121-141.

Vannini, C., F. Ferrantini, K. H. Schleifer, W. Ludwig, F. Verni, and G. Petroni. 2010. "Candidatus anadelfobacter veles" and "Candidatus cyrtobacter comes," two new rickettsiales species hosted by the protist ciliate Euplotes harpa (Ciliophora, Spirotrichea). Appl Environ Microbiol **76**:4047-4054.

Viklund, J., T. J. Ettema, and S. G. Andersson. 2012. Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. Mol Biol Evol **29**:599-615.

Werren, J. H. and S. Richards and C. A. Desjardins and O. Niehuis and J. Gadau and J. K. Colbourne and L. W. Beukeboom and C. Desplan and C. G. Elsik and C. J. Grimmelikhuijzen and P. Kitts and J. A. Lynch and T. Murphy and D. C. Oliveira and C. D. Smith and L. van de Zande and K. C. Worley and E. M. Zdobnov and M. Aerts and S. Albert and V. H. Anaya and J. M. Anzola and A. R. Barchuk and S. K. Behura and A. N. Bera and M. R. Berenbaum and R. C. Bertossa and M. M. Bitondi and S. R. Bordenstein and P. Bork and E. Bornberg-Bauer and M. Brunain and G. Cazzamali and L. Chaboub and J. Chacko and D. Chavez and C. P. Childers and J. H. Choi and M. E. Clark and C. Claudianos and R. A. Clinton and A. G. Cree and A. S. Cristino and P. M. Dang and A. C. Darby and D. C. de Graaf and B. Devreese and H. H. Dinh and R. Edwards and N. Elango and E. Elhaik and O. Ermolaeva and J. D. Evans and S. Foret and G. R. Fowler and D. Gerlach and J. D. Gibson and D. G. Gilbert and D. Graur and S. Grunder and D. E. Hagen and Y. Han and F. Hauser and D. Hultmark and H. C. t. Hunter and G. D. Hurst and S. N. Jhangian and H. Jiang and R. M. Johnson and A. K. Jones and T. Junier and T. Kadowaki and A. Kamping and Y. Kapustin and B. Kechavarzi and J. Kim and B. Kiryutin and T. Koevoets and C. L. Kovar and E. V. Kriventseva and R. Kucharski and H. Lee and S. L. Lee and K. Lees and L. R. Lewis and D. W. Loehlin and J. M. Logsdon, Jr. and J. A. Lopez and R. J. Lozado and D. Maglott and R. Maleszka and A. Mayampurath and D. J. Mazur and M. A. McClure and A. D. Moore and M. B. Morgan and J. Muller and M. C. Munoz-Torres and D. M. Muzny and L. V. Nazareth and S. Neupert and N. B. Nguyen and F. M. Nunes and J. G. Oakeshott and G. O. Okwuonu and B. A. Pannebakker and V. R. Pejaver and Z. Peng and S. C. Pratt and R. Predel and L. L. Pu and H. Ranson and R. Raychoudhury and A. Rechtsteiner and J. T. Reese and J. G. Reid and M. Riddle and H. M. Robertson and J. Romero-Severson and M. Rosenberg and T. B. Sackton and D. B. Sattelle and H. Schluns and T. Schmitt and M. Schneider and A. Schuler and A. M. Schurko and D. M. Shuker and Z. L. Simoes and S. Sinha and Z. Smith and V. Solovyev and A. Souvorov and A. Springauf and E. Stafflinger and D. E. Stage and M. Stanke and Y. Tanaka and A. Telschow and C. Trent and S. Vattathil

and E. C. Verhulst and L. Viljakainen and K. W. Wanner and R. M. Waterhouse and J. B. Whitfield and T. E. Wilkes and M. Williamson and J. H. Willis and F. Wolschin and S. Wyder and T. Yamada and S. V. Yi and C. N. Zecher and L. Zhang and R. A. Gibbs. 2010. Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. Science **327**:343-348.

Wilgenbusch, J. C. and D. Swofford. 2003. Inferring evolutionary trees with PAUP*. Curr Protoc Bioinformatics **Chapter 6**:Unit 6 4.

Williams, K. P., B. W. Sobral, and A. W. Dickerman. 2007. A robust species tree for the alphaproteobacteria. J Bacteriol **189**:4578-4586.

Woyke, T., H. Teeling, N. N. Ivanova, M. Huntemann, M. Richter, F. O. Gloeckner, D. Boffelli, I. J. Anderson, K. W. Barry, H. J. Shapiro, E. Szeto, N. C. Kyrpides, M. Mussmann, R. Amann, C. Bergin, C. Ruehland, E. M. Rubin, and N. Dubilier. 2006. Symbiosis insights through metagenomic analysis of a microbial consortium. Nature **443**:950-955.

Wu, D., S. C. Daugherty, S. E. Van Aken, G. H. Pai, K. L. Watkins, H. Khouri, L. J. Tallon, J. M. Zaborsky, H. E. Dunbar, P. L. Tran, N. A. Moran, and J. A. Eisen. 2006. Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. PLoS Biol **4**:e188.

FIGURES

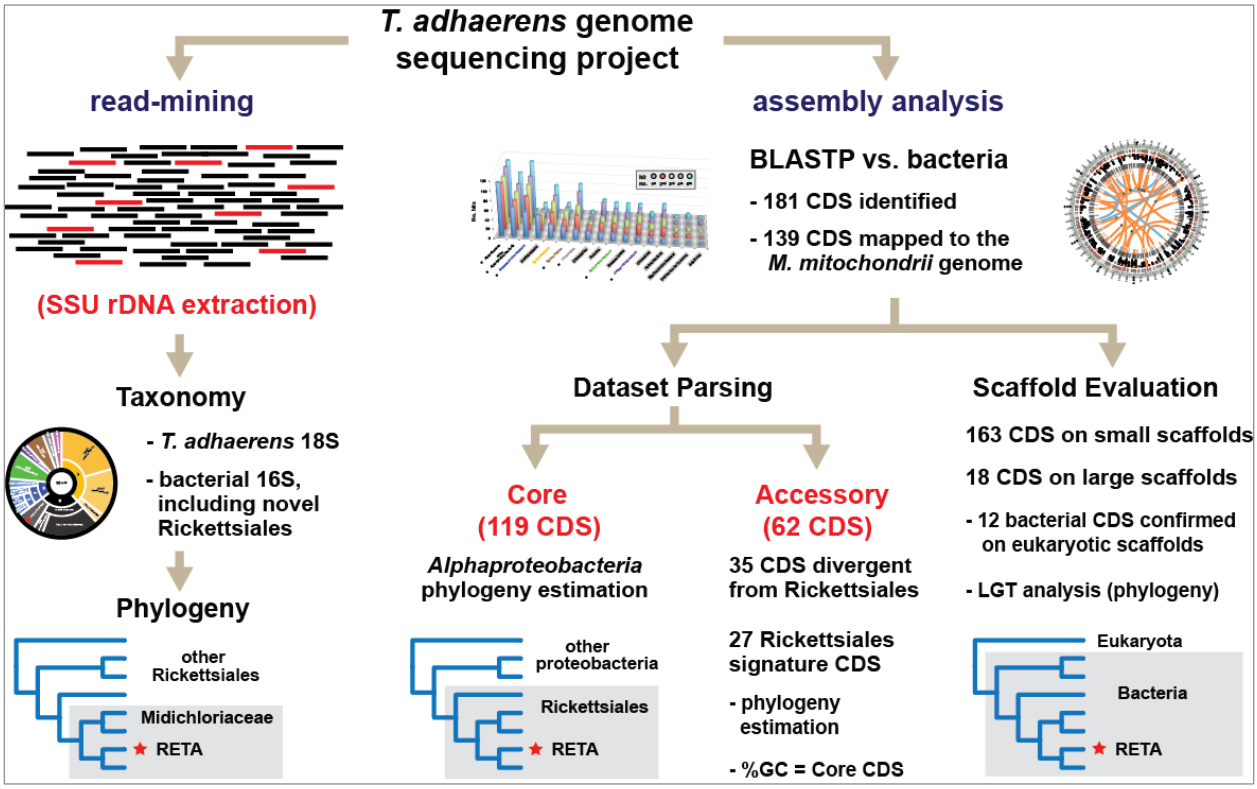


Figure 1. Overview of the methodology used to identify bacterial DNA sequences within the *Trichoplax adhaerens* genome project. Bacterial SSU rDNA sequences were mined from the trace read archive (left), with rickettsial sequences further analyzed via phylogeny estimation. Bacterial CDS identified in the assembly (right) were determined to be primarily rickettsial-like based on phylogeny estimation and identification of rickettsial signature genes. The distribution of bacterial CDS on small (bacterial-like) and large (eukaryotic-like) scaffolds is shown, with CDS on the latter further evaluated for LGT via phylogeny estimation.

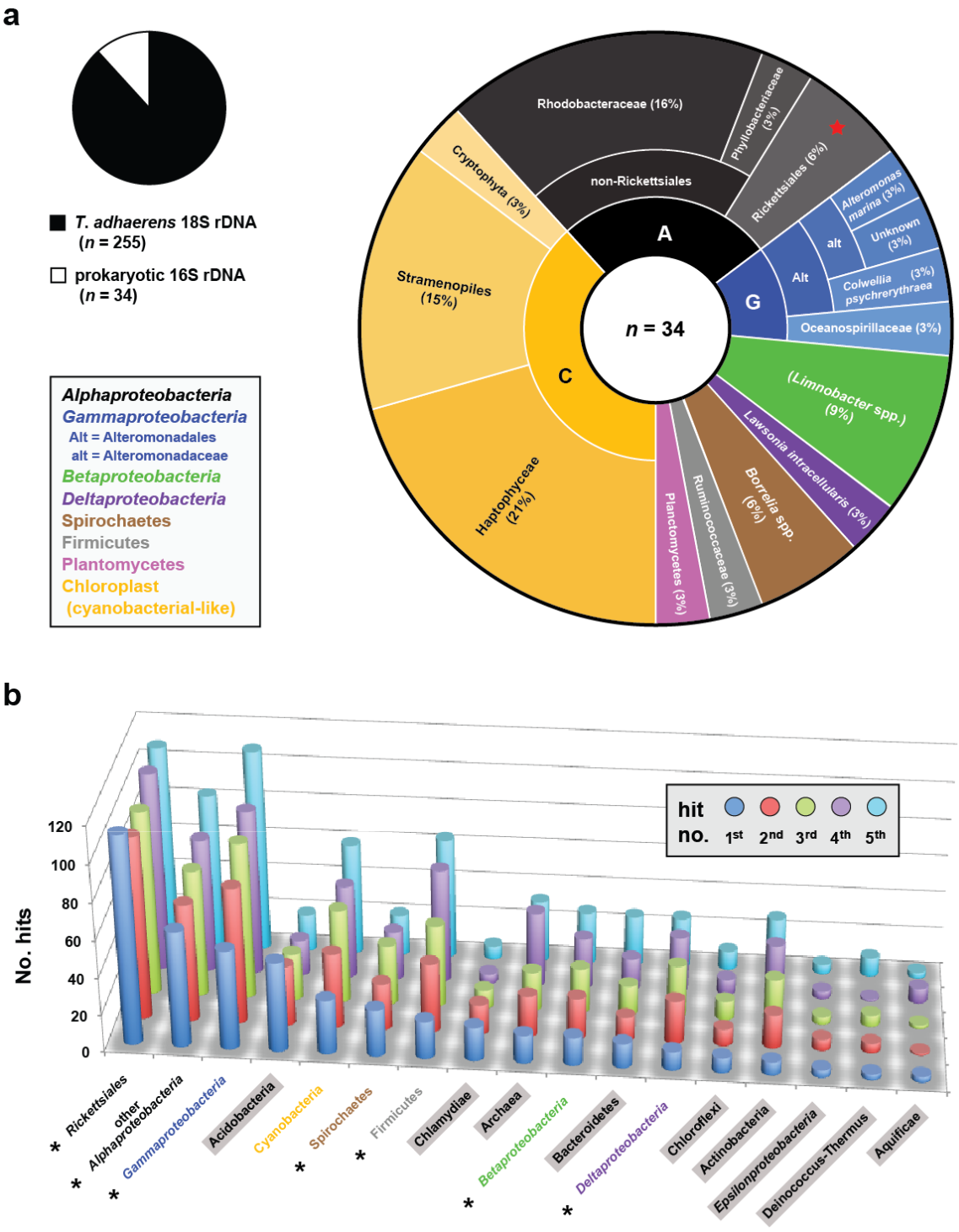


Figure 2. Identification of bacterial DNA sequences within the *Trichoplax adhaerens* genome trace read archive and assembly. (**a**) Illustration of 289 SSU rDNA sequences identified in the *T. adhaerens* trace read archive (http://genome.jgi-psf.org/Triad1/Triad1.download.ftp.html). The pie chart at top left illustrates the 34 prokaryotic 16S rDNA sequences detected among 255 *T. adhaerens* 18S rDNA sequences. Larger graph at right illustrates the taxonomic distribution of

the 34 prokaryotic 16S rDNA sequences (see text for details on taxonomic assignment). Sequences are grouped into nested sectors according to hierarchical taxonomy, progressing from the interior to exterior of the plot. Color scheme is explained in box at bottom left. Cyanobacterial hits correspond to chloroplast rDNA sequences of cyanobacterial origin. Plot made with Krona v.2.0 **(Ondov et al. 2011)** with manual adjustment. (**b**) Illustration of *T. adhaerens* proteins that have strong similarity to their prokaryotic counterparts. All proteins encoded within the *T. adhaerens* assembly (NCBI, ASM15027v1, n = 11,540) were BLASTed against prokaryotic and eukaryotic proteins within the *nr* database (NCBI), with subjects ranked according to S_m score (see text for details). Graph depicts the taxonomic distribution of the top five scores per *T. adhaerens* protein that included a prokaryotic protein (n = 1,697). The taxa are arrayed along the x-axis in decreasing order according to the number of top hits (blue). Prokaryotic groups with less than 15 total hits per group (sum 1-5) are not shown. Asterisks depict taxonomic groups that also have a 16S rDNA sequence illustrated in panel **a**.

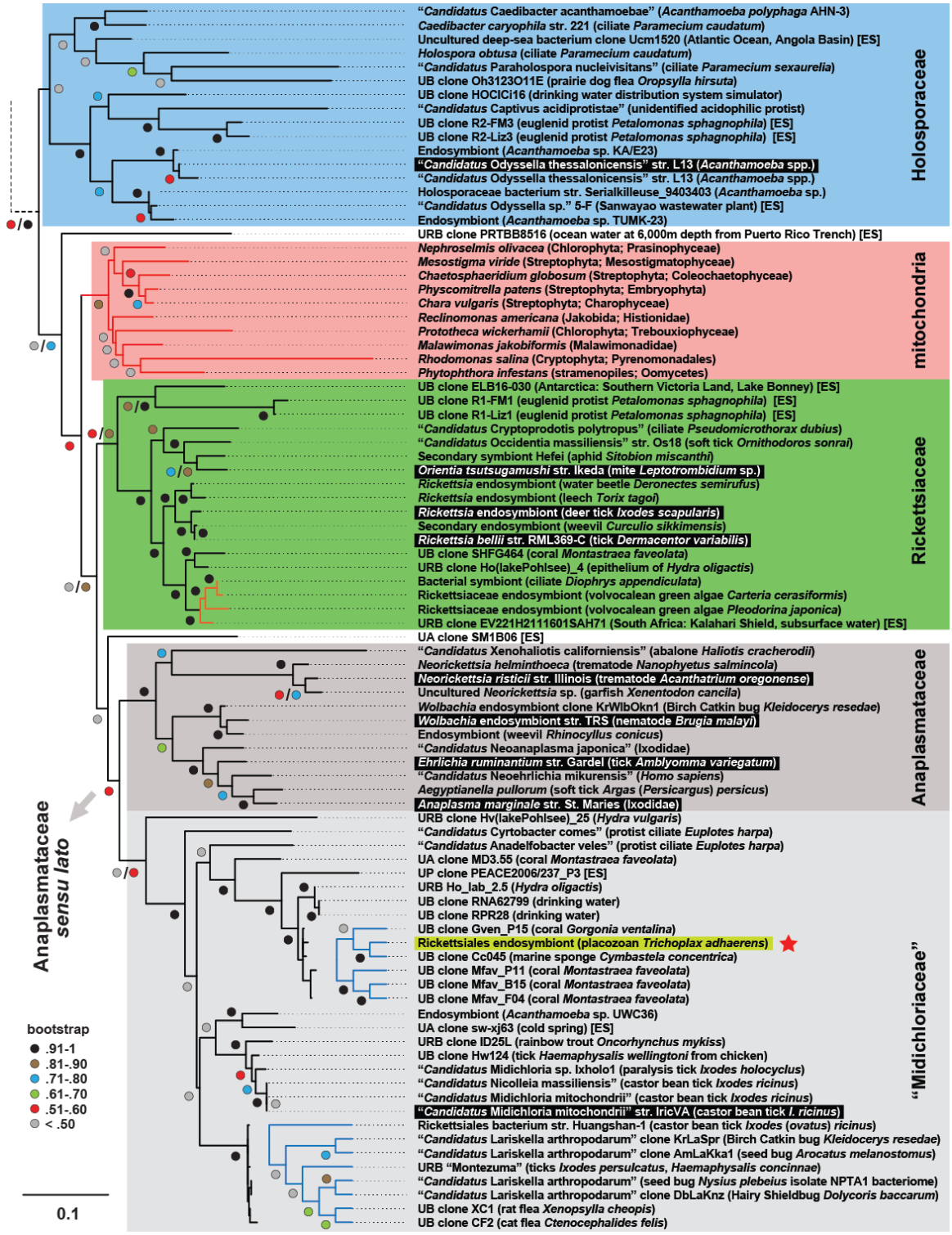


Figure 3. Phylogeny of SSU rDNA sequences estimated for 78 Rickettsiales taxa, ten mitochondria, and five outgroup taxa. See text for alignment and tree-building methods. Tree is final optimization likelihood: (-22042.321923) using GTR substitution model with GAMMA and proportion of invariant sites estimated. Branch support is from 1000 bootstrap

pseudoreplications. For nodes represented by two bootstrap values, the left is from the analysis that included ten mitochondrial sequences, with the right from the analysis without the mitochondrial sequences. All nodes with single bootstrap values had similar support in both analyses. Red (mitochondria) and orange (within Rickettsiaceae) branches are reduced 75% and increased 50%, respectively. Blue cladograms depict minimally resolved lineages within the "Midichloriaceae". The divergence point of five outgroup taxa from *Betaproteobacteria* ($n = 1$), *Gammaproteobacteria* ($n = 1$), and other *Alphaproteobacteria* ($n = 3$) is shown with a dashed branch. For each taxon, associated hosts are within parentheses, with ES depicting an environmental sample. Other abbreviations: UB, uncultured bacterium; UP, uncultured proteobacterium; UA, uncultured alphaproteobacterium; URB, uncultured Rickettsiales bacterium. Taxa within black boxes have available genome sequence data. The 16S rDNA sequence mined from the *T. adhaerens* trace archive is boxed green and noted with a red star. Accession numbers for all sequences are provided in **supplementary table S1**.

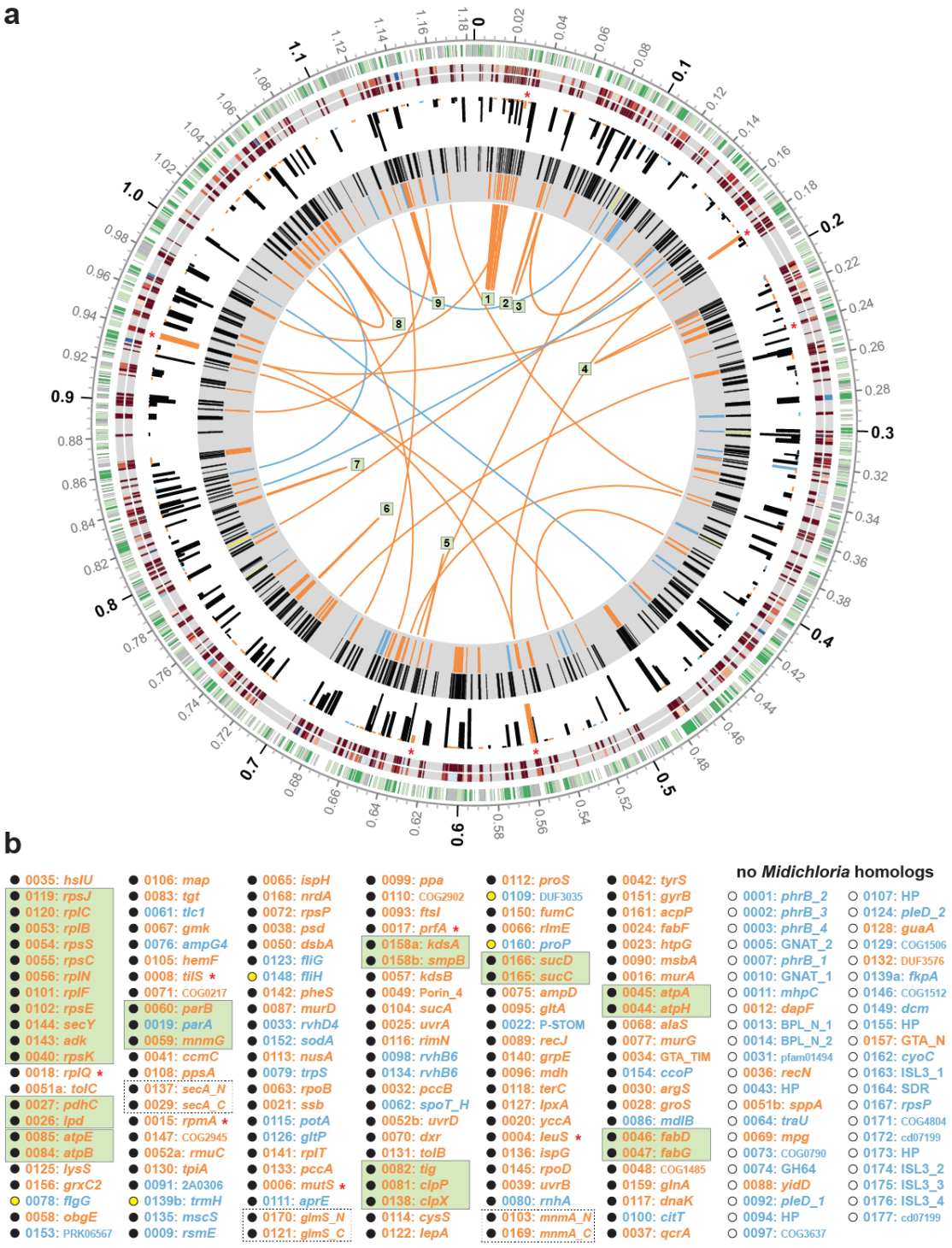


Figure 4. Bacterial CDS identified within the *Trichoplax adhaerens* genome assembly. (**a**) Results of an all-against-all BLASTP analysis between the genomes of *T. adhaerens* Grell-BS-1999 (n = 11,540) and "*Candidatus* Midichloria mitochondrii" str. IricVA (n = 1,211), hereafter *M. mitochondrii*. Outer black circle is a scale with coordinates (in Mb) for the *M. mitochondrii*

genome, with the putative origin of replication positioned at 12 o'clock as previously determined **(Sassera et al. 2011)**. Four rings inside the scale as follows: **(1)** 1,211 CDS of the *M. mitochondrii* genome, with operons and transcriptional units (predicted using fgenesb **(Tyson et al. 2004)**) colored green and gray, respectively; **(2)** heat maps for S_m scores > 20 (outer) and corresponding E-values (inner) for 347 *T. adhaerens*-*M. mitochondrii* protein matches, with S_m scores from 20 (dark blue) to 576 (burgundy) and E-values from 1 (dark blue) to 1.00E-500 (burgundy); **(3)** histograms depicting the number of contigs on each scaffold that contain the identified *T. adhaerens* gene: eukaryotic-like CDS (black), Rickettsiales endosymbiont of *T. adhaerens* [RETA] CDS of Core Dataset (orange), RETA CDS of Accessory Dataset (blue); **(4)** All 347 *T. adhaerens* CDS (outer, black) and 138 RETA CDS (inner, orange, blue) (see supplementary figure 3 for linear histogram and further information). NOTE: five yellow CDS (outer) were below the S_m 20 cutoff but were determined to be RETA CDS via manual inspection. RETA CDS present on the same *T. adhaerens* scaffold are linked in the interior of the plot, with boxes (1-7) depicting syntenic regions across *M. mitochondrii* and RETA. Plot made using Circos **(Krzywinski et al. 2009)** with manual adjustment. (**b**) List of 181 RETA CDS identified within the *T. adhaerens* assembly. RETA identifier (0001-0181) followed by gene symbol or predicted product description (complete annotations in **supplementary table S2**). Core Dataset CDS (orange) comprise 119 ORFs corresponding to 116 genes, with three split genes (dashed boxes). Accessory Dataset CDS (blue) comprise 62 genes. Black circles depict RETA CDS with homologs present in the *M. mitochondrii* genome (n = 138), and are listed according to their clockwise arrangement in ring 4 of the plot in panel **a**. Yellow circles depict the five genes added manually (S_m < 20) to ring 4. Green boxes enclose the seven syntenic regions illustrated in the interior on the plot. Open circles depict the 42 RETA CDS that do not have significant homologs in the *M. mitochondrii* genome. Red asterisks denote six CDS that were subsequently determined to be likely nuclear encoded mitochondrial genes (see text). Red asterisks also mark the location of these CDS in panel a between rings 2 and 3.

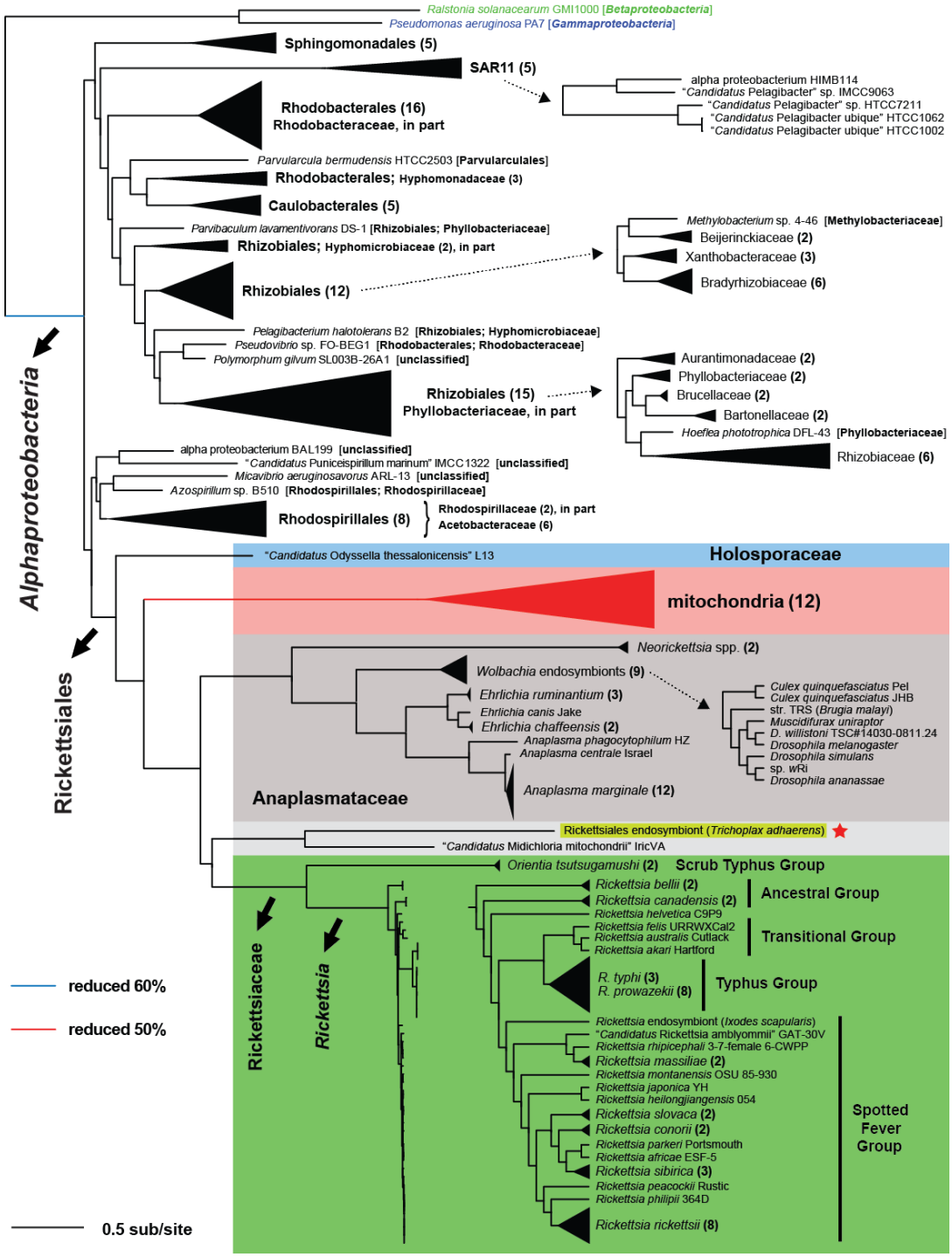


Figure 5. Genome-based phylogeny estimated for Rickettsiales endosymbiont of *T. adhaerens* [RETA], 162 alphaproteobacterial taxa, twelve mitochondria, and two outgroup taxa. RETA core proteins (n = 113) were included in the phylogenetic pipeline that entails ortholog group (OG) generation, OG alignment (and masking of less conserved positions), and concatenation of aligned OGs (see text). Tree was estimated using the CAT-GTR model of substitution as implemented in PhyloBayes v3.3 **(Lartillot and Philippe 2004, 2006)**. Tree is a consensus of

1522 trees (post burn-in) pooled from two independent Markov chains run in parallel. Branch support was measured via posterior probabilities, which reflect frequencies of clades among the pooled trees. RETA is boxed green and noted with a red star. Classification scheme for *Rickettsia* spp. follows previous studies **(Gillespie et al. 2007, Gillespie et al. 2008)**. Taxon names, PATRIC genome IDs (bacteria) and NCBI accession numbers (mitochondria) for the 176 genomes are provided in **supplementary table S3**.

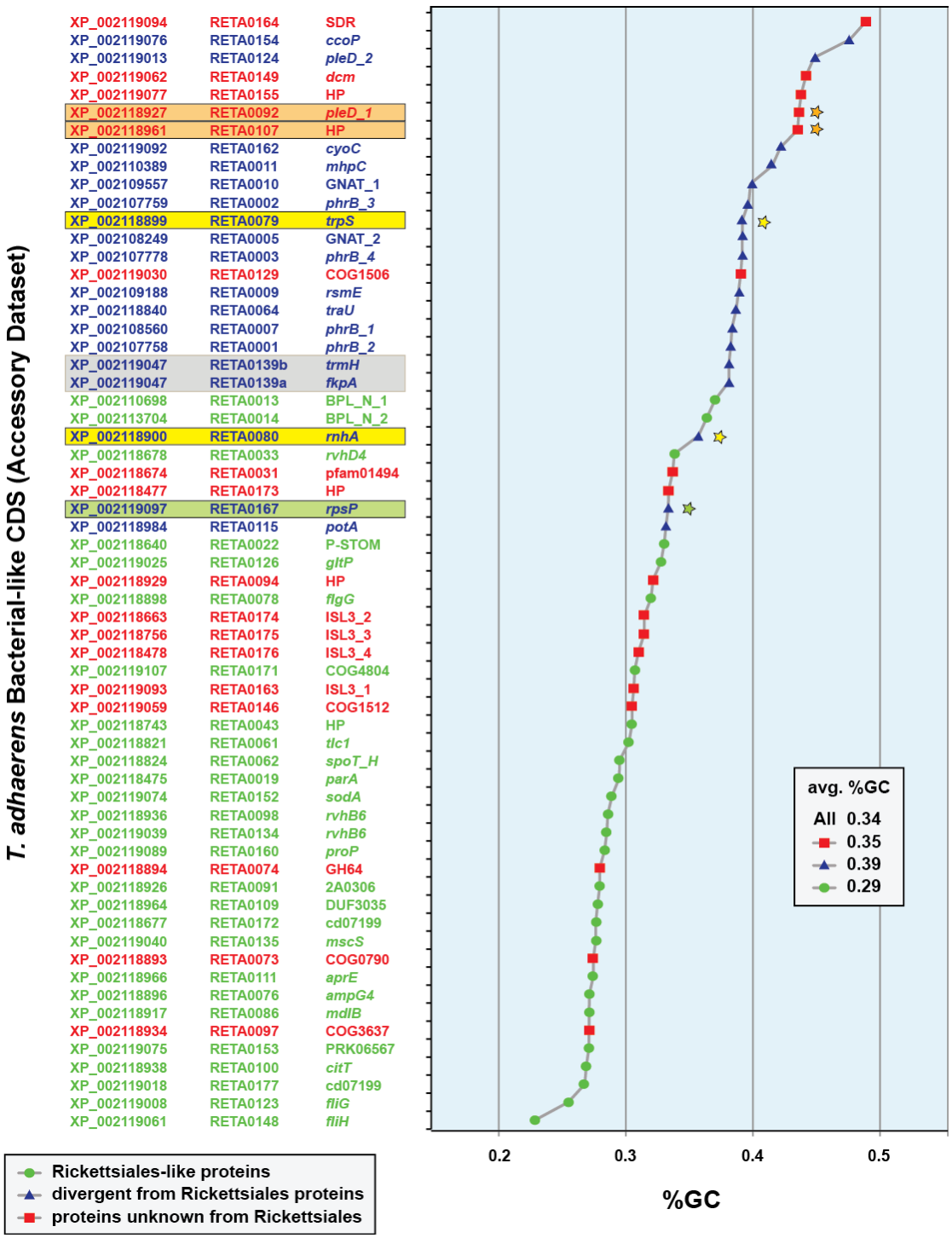
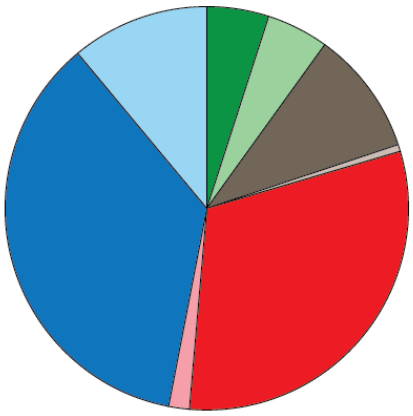


Figure 6. Bacterial **CDS** (Accessory Dataset) identified within the *Trichoplax adhaerens* genome assembly. These 62 CDS were determined to lack the profile of typical Rickettsiales genes inherited vertically from an alphaproteobacterial ancestor (see text). Rickettsiales endosymbiont of *T. adhaerens* [RETA] CDS are plotted by %GC (*x*-axis). *T. adhaerens* protein accession numbers (NCBI), RETA IDs and gene/protein names are listed on the *y*-axis, with

color scheme as follows: green, highly similar to Rickettsiales signature proteins (trees shown in **supplementary figure 6**); blue, present in some (or all) Rickettsiales genomes yet divergent in sequence and phylogenetic signal (trees shown in **supplementary figure 7**); red, unknown from Rickettsiales genomes. Inset shows the avg. %GC for all 62 CDS, as well as for the three groups. Stars depict the following: yellow, identical to sequences from the genome of *Halothermothrix orenii* H 168 (Firmicutes: Haloanaerobiales); orange, 99% aa identity with sequences from the genome of *Alteromonas macleodii* ATCC 27126 (*Gammaproteobacteria*: Alteromonadales); green, most similar to chloroplast sequences of haptophytic algae (Eukaryota; Haptophyceae). Colored boxes on the *y*-axis correspond with stars on the plot, with the gray box illustrating a fused gene model (*trmH-fkpA*).

a



Assembly features	1 exon	introns
Eukaryotic scaffolds (*n* = 18)	9	9
Small hybrid scaffolds (*n* = 19)	18	1
All-bacteria scaffolds (*n* = 59)	56	3
Singleton scaffolds (*n* = 85)	65	20

b

Scaffold ID/ No. genes	CDS (RETA ID)	Protein	Sub-dataset	No. exons	Coverage (*T. adhaerens*)	Discrepancy	Coverage (RETA)	Phylogeny
NW_002060943 1541	0001	PhrB_2	AD-B	1	1.00	-	0.93	A; Nitrospirae/Legionellales
	0002	PhrB_3	AD-B	1	1.00	-	0.95	A; Nitrospirae/Legionellales
	0003	PhrB_4	AD-B	1	0.80	N	0.96	A; Nitrospirae/Legionellales
	0004	LeuS	CD-B	12	0.06	N	0.56	B; mitochondria
	0005	GNAT_2	AD-B	1	1.00	-	0.98	C; Spirochaetes/Odyssella
	0006	MutS	CD-B	3	0.83	C	0.88	D; mitochondria
	0007	PhrB_1	AD-B	1	1.00	-	0.95	A; Nitrospirae/Legionellales
	0008	TilS	CD-R	7	0.15	C	0.44	E; mitochondria
NW_002060944 1091	0009	RsmE	AD-B	1	1.00	-	0.97	F; Chlamydiae
	0010	GNAT_1	AD-B	1	1.00	-	0.94	C; Spirochaetes/*Odyssella*
NW_002060945 1071	0011	MhpC	AD-B	1	1.00	-	0.98	G; Cyanobacteria
	0012	DapF	CD-R	3	0.66	C	0.99	H; Rickettsiales
	0013	BPL_N_1	AD-R	2	0.39	C	0.90	I; *Rickettsia* spp./Chlamydiae
NW_002060948 745	0014	BPL_N_2	AD-R	2	0.07	C	0.85	I; *Rickettsia* spp./Chlamydiae
NW_002060955 173	0015	RpmA	CD-B	3	0.00	-	0.00	J; Rickettsiales/mitochondria
NW_002060959 152	0017	PrfA	CD-E	7	0.00	-	0.00	K; mitochondria
NW_002060958 145	0016	MurA	CD-R	1	0.88	N	0.98	L; Rickettsiales
NW_002060975 41	0018	RplQ	CD-B	3	0.07	N	0.52	M; Eukaryota/Chloroflexi

Figure 7. Evidence for bacterial-like genes encoded in the *Trichoplax adhaerens* genome. (**a**) Division of the 181 Rickettsiales endosymbiont of *T. adhaerens* [RETA] CDS into four categories based on the composition of their scaffolds: eukaryotic scaffolds, CDS present on

large (> 40 genes) scaffolds with predominately eukaryotic-like genes (*n* = 18); small hybrid scaffolds, CDS present on small (< 7 genes) scaffolds with both bacterial- and eukaryotic-like genes (*n* = 19); all-bacteria scaffolds, CDS present on small (< 5 genes) scaffolds comprised entirely of bacterial-like genes (*n* = 59); and singleton-gene scaffolds (*n* = 85). Each category is further divided into single exon genes and genes possessing one or more introns (as predicted within the original *T. adhaerens* assembly). (**b**) Eight *T. adhaerens* scaffolds contain 18 RETA CDS. Scaffold IDs and number of encoded genes are from the *T. adhaerens* assembly (see text). RETA IDs and protein names are further described in **supplementary table S2**. For Core Dataset CDS: CD-R, CD-B and CD-E correspond to the sub-datasets Ric-78, Bac-26 and Euk-9, respectively (**supplementary figure S2**). For Accessory Dataset CDS: AD-R, highly similar to Rickettsiales signature proteins; AD-B, present in some (or all) Rickettsiales genomes yet divergent in sequence and phylogenetic signal (see **fig. 6**). The number of exons for each CDS is shown. The results of gene predictions by fgenesb **(Tyson et al. 2004)** (headings for three columns colored green) are described as follows: 'Coverage (*T. adhaerens*)', percentage of bps in the eukaryotic gene prediction matching those in the fgenesb prediction; 'Discrepancy', differences at either the N- or C-terminus across eukaryotic and fgenesb predictions; ''Coverage (RETA)', percentage of bps in the fgenesb prediction matching those in the eukaryotic gene prediction. The most related sequences as determined by phylogeny estimation are listed, with full trees provided in (**supplementary figure S8**). Potential bacteria-to-*T. adhaerens* LGT products are highlighted yellow.

TABLES

Table 1. Comparison of sequence divergence across RETA and Rickettsiales genera [1, 2].

	RETA	Midi	Odys	Neor	Wolb	Ehrl	Anap	Orie	Rick
RETA		***0.47***	0.56	0.66	0.61	0.60	0.64	0.60	0.54
Midi	***0.45***		0.47	0.60	0.53	0.53	0.55	0.53	0.46
Odys	0.54	0.45		0.58	0.52	0.53	0.54	0.51	0.44
Neor	0.63	0.58	0.56		0.55	0.55	0.56	0.63	0.59
Wolb	0.58	0.51	0.50	0.53		0.36	0.40	0.55	0.52
Ehrl	0.58	0.51	0.51	0.53	0.36		0.30	0.54	0.51
Anap	0.61	0.53	0.52	0.54	0.39	0.30		0.58	0.54
Orie	0.57	0.51	0.50	0.60	0.53	0.52	0.56		0.39
Rick	0.52	0.45	0.43	0.56	0.50	0.49	0.52	0.39	

[1] Calculated % divergence (8327 aa sites of core dataset) with DIVEIN (Deng et al. 2010), using the WAG (top) and Blosum62 (bottom) amino acid substitution models. Color scheme as follows: light yellow, < 35 (% divergence); yellow, 36-40; tan, 41-45; light orange, 46-50; orange, 51-55; dark orange, 56-60; red, > 60. Values for the closest taxon to RETA (*Midichloria*) are in bold italics.

[2] Taxon abbreviations as follows: RETA, Rickettsiales endosymbiont of *Trichoplax adhaerens*; Midi, "*Candidatus* Midichloria mitochondrii" str. IricVA; Odys, "*Candidatus* Odyssella thessalonicensis" str. L13; Neor, *Neorickettsia risticii* str. Illinois; Wolb, *Wolbachia* endosymbiont str, TRS of *Brugia malayi*; Ehrl, *Ehrlichia ruminantium* str. Gardel; Anap, *Anaplasma phagocytophilum* str. HZ; Orie, *Orientia tsutsugamushi* str. Ikeda; Rick, *Rickettsia bellii* str. RML369-C.

CHAPTER 5. Bacterial sequences mined from the *Xenopus* (Silurana) *tropicalis* genome project and assembly provide evidence for an associated betaproteobacterial bacterium.

ABSTRACT

The objective of this study was to identify and characterize a putative *Xenopus tropicalis* associated bacterium from the genome sequencing data of *Xenopus* (Silurana) *tropicalis* (Western clawed frog). A survey of bacterial 16S SSU rDNA sequences within the unassembled *X. tropicalis* reads revealed a substantial number of 16S rDNA sequences primarily encompassing three bacterial taxa: *Betaproteobacteria*, *Gammaproteobacteria*, and Firmicutes. Seventeen betaproteobacterial 16S rDNA reads were assembled into a single contig, 1660 bases in length, with no match to any known 16S rDNA sequence in the NCBI database. Phylogeny estimation placed this 16S rDNA sequence among the Comamonadaceae family of *Betaproteobacteria*. Analysis of the *X. tropicalis* reads with ReadMiner using 34 sequenced Comamonadaceae genomes yielded over 1.3 Mb of sequence on 1067 contigs. Phylogeny estimation and genome divergence using 323 mined orthologs conserved across 80-100% of *Betaproteobacteria* genomes confirmed the placement of this bacterium within the Comamonadaceae. Corresponding analysis of the assembled *X. tropicalis* genome using AssemblySifter identified at least 291 scaffolds (891 genes) as likely genomic elements from bacteria, including 129 scaffolds (331 genes) as possible components of a betaproteobacterial genome. The results of this study confirm the presence of a betaproteobacterium associated with the *X. tropicalis* genome sequencing project, and demonstrate the effectiveness of a combined application of the MetaMiner, ReadMiner, and AssemblySifter workflows in extracting and characterizing bacterial sequences for a single target among many non-target species.

INTRODUCTION

The African clawed frogs (genus *Xenopus*) are among the most well-studied amphibians, and model organisms in the study of vertebrate developmental biology. They are easily maintained in the laboratory, develop externally, and have large oocytes tractable for the manipulation of embryogenesis. A draft of the complete genome sequence of the Western clawed frog *Xenopus* (Silurana) *tropicalis* was published in 2010 (**Hellsten et al. 2010**), using random WGS sequencing data at 7.6-fold redundancy and covering approximately 1.51 Gb of the estimated 1.7 Gb genome. Nuclear DNA for sequencing was derived from erythrocytes, and was assembled prior to the removal of vector, non-cellular debris, and a known prokaryotic contaminant. Twenty-six scaffolds of varying sizes were classified as prokaryotic and removed from the *X. tropicalis* assembly, including a few with similarity to sequences from *Betaproteobacteria* (mostly *Burkholderia* and *Ralstonia* species) (ftp://ftp.xenbase.org/pub/Genomics/JGI/Xentr4.0/v4.0_contamination/contamination.txt). *Betaproteobacteria* are a highly diverse class of bacteria commonly found in soil environments. Many species are isolated from wastewater or contaminated soil and exhibit an array of biodegradative abilities (**Bruland et al. 2009; Khan et al. 2002**); still more are known to be pathogens (**Bahar et al. 2011; Rosenstein et al. 2001; Saldias & Valvano 2009; Farshad et al. 2012**) or symbionts (**Ghignone et al. 2011; Pinel et al. 2008; Lund et al. 2009; McCutcheon & Dohlen 2011; Du et al. 1994**) of both plants and animals.

Based on multiple observations of *X. tropicalis* proteins with extremely high similarity to betaproteobacterial sequences, the diversity of known betaproteobacterial symbionts in nature,

and the characteristics of the *X. tropicalis* assembly, the existence of a *X. tropicalis* associated betaproteobacterial bacterium (XTAB) was hypothesized. The current study was undertaken to verify the existence of XTAB, extract its genome from the *X. tropicalis* genome sequencing project reads, and sift assembled *X. tropicalis* scaffolds for likely bacterial genomic elements, including possible bacteria-to-host LGTs. The results presented here demonstrate the ability of the described read mining approach to extract a single novel bacterial target from among many bacterial constituents within a eukaryotic genome sequencing project. Moreover, this study also illustrates the effectiveness of the described computational approach to sifting host genome scaffolds for sequences that may represent bacterial LGTs into the host genome.

METHODS

Data Preparation

A total of 22,132,045 WGS reads for the *Xenopus* (Silurana) *tropicalis* genome project were downloaded from the Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi) at NCBI. Reads were decontaminated and validated as described in Chapter 2 using cloning vector sequences obtained from the Joint Genome Institute (JGI), resulting in 21,517,851 decontaminated reads. Bases at the 5' and 3' ends of reads were trimmed so that the mean Phred (quality) score at all positions was 25 or higher. Finally, 3,510,992 reads with mean Phred scores below 25 or lengths shorter than 100 bases were removed, producing approximately 1.8×10^7 reads that were subsequently analyzed for bacterial sequences.

SSU rDNA Analyses

Read analysis. MetaMiner was used to assess the distribution of bacterial sequences in the *X. tropicalis* read data, identify a bacterial target, and assemble and characterize a target 16S rDNA sequence. A detailed description of MetaMiner can be found in Chapter 2. Briefly, prepared *X. tropicalis* reads were mapped against 406,997 full-length bacterial 16S rDNA sequences from the 2012 release of the Greengenes database (**DeSantis et al. 2006**), plus the 18S SSU rDNA sequence (NCBI GI 65094) from *Xenopus laevis* obtained from NCBI (*X. laevis* was the closest relative to *X. tropicalis* with an available 18S rDNA sequence at the time of the study). Reads that had at least one match in this SSU rDNA library were binned according to the taxonomic classification of their matches.

Reads that mapped to betaproteobacterial-like (20) and gammaproteobacterial-like (206) 16S rDNA sequences were compared for pairwise gene divergence against full-length 16S rDNA sequences from the sequenced genomes of 33 betaproteobacterial species (Burkholderiales plus *Neisseria gonorrhoeae* NCCP11945), 7 *Pseudomonas* spp. (*Gammaproteobacteria*), and 5 *Paenibacillus* spp. (Firmicutes), all from the November 2012 release of PATRIC **(Gillespie et al. 2011)** (**table 1**). An unpaired Student's t-test was used to evaluate the statistical significance of differences in mean divergence. In addition, ANOVA was used to determine the statistical significance of differences in mean read divergence to three different groups of genomes (*Betaproteobacteria*, *Pseudomonas* spp., and *Paenibacillus* spp.). The monophyly of the betaproteobacterial-like reads was evaluated with phylogeny estimation, in the presence and absence of the full-length Pseudomonad 16S rDNA sequences, using `RAxML` (**Stamatakis 2006**) as described above. Minimally divergent, monophyletic betaproteobacterial-like reads were subsequently aligned and stitched together to construct a single 16S rDNA sequence. A phylogeny was estimated for the mined 16S rDNA using both neighborhood and cascading

taxon-sampling methods as described above. Nine taxonomic groups were used in the cascading method, based on the predicted position of the target in the Comamonadaceae family (**table 2**). Additionally, pairwise gene divergence of the mined 16S rDNA was calculated using the same 45 full-length 16S rDNA sequences used in the read analysis (see **table 1**).

Whole-Genome Read Analysis

ReadMiner was used to extract Comamonadaceae-like sequences from the *X. tropicalis* read data. A detailed description of ReadMiner can be found in Chapter 2. Briefly, 34 complete and WGS Comamonadaceae genomes, and 181 complete and WGS Pseudomonadaceae genomes were used as the best-matching and best-competing clades, respectively (**table 3**). All genome sequences were obtained from the November 2012 release of PATRIC, with plasmid genomes removed prior to analysis. Reads that matched to Comamonadaceae, but not Pseudomonadaceae genomes were extracted, de-duplicated, assembled using `Mira` **(Chevreux 2005**), with gene models constructed using `fgenesb` (http://linux1.softberry.com/berry.phtml?topic=fgenesb&group=help&subgroup=gfindb) as described above. Predicted protein sequences were queried using `blastp` for the top ten unique-taxon matches (e-value<10^{-3}) against NCBI's *nr* database (excluding the *X. tropicalis* group); queries with at least one alignment over 50% or more of a subject comprised the initial set of mined sequences, and were annotated using the GOanna online resource (**McCarthy et al. 2006**). To further account for possible contamination, mined sequences were combined with 429,155 annotated proteins from 82 sequenced Comamonadaceae and Pseudomonadaceae genomes and OGs were constructed as described above. Notably, no mined sequences were found that clustered exclusively with Pseudomonadaceae proteins.

Identification of core target genes. Mined sequences were combined with 217,625 annotated proteins from 54 sequenced betaproteobacterial genomes from PATRIC (**table 4**) and OGs were constructed using `FastOrtho` as described above. Mined proteins with at least one ortholog in 80% of the genomes were subjected to *n*-taxon statement analysis using three taxonomic sub-groups of NCBI's *nr* database: all Comamonadaceae (NCBI taxid=80864), all bacteria (NCBI taxid=2) except Comamonadaceae, and all eukaryotes (NCBI taxid=2759) except the *X. tropicalis* group (NCBI taxid=8363). Orthologs that were most closely related to other Comamonadaceae proteins were hypothesized to comprise a core set of target genes from a single *Xenopus tropicalis* associated betaproteobacterium (XTAB).

Published eukaryotic genome assemblies have been shown to contain bacterial scaffolds (see Chapter 4), which have the potential to mislead *n*-taxon statement analyses by falsely identifying mined XTAB orthologs as eukaryotic in origin. A reciprocal blast approach was taken to identify mined orthologs that had been incorrectly excluded from the core set of XTAB genes during *n*-taxon analysis. Mined orthologs that were more closely related to apparent eukaryote proteins were queried using `blastp` for the top 25 matches (by e-value) against NCBI's *nr* database. Any match to a eukaryote was queried against *nr* in turn, and the distribution of its top hits examined for a preponderance of hits to bacterial proteins, and/or hits to proteins annotated with unknown, hypothetical, or obviously bacterial functions. This evidence was subsequently used to overturn clearly artifactual *n*-taxon statements.

Genome-based phylogeny. Core XTAB proteins were combined under the assumption that they were vertically inherited from a betaproteobacterial ancestor, and total of 56 genomes were used for robust phylogeny estimation, including the 54 betaproteobacterial genomes used to build

OGs (see **table 4**) plus two outgroup species: *Pseudomonas aeruginosa* PA7 (*Gammaproteobacteria*) and *Pseudovibrio* sp. FO-BEG1 (*Alphaproteobacteria*). Genome-based phylogenies were estimated using both `FastTree` (**Price et al. 2010**) and `PhyloBayes` (**Lartillot et al. 2009**) as described above. In addition to phylogeny, an estimate of genome divergence was calculated using the same set of core XTAB proteins. The final alignment of proteins was processed to include only a single representative species from each genus in the Comamonadaceae (plus XTAB) and all positions with missing data were removed. Percent protein divergence was estimated with the online program `DIVEIN` (**Deng et al. 2010**) using both the WAG and BLOSUM62 amino acid substitution models.

Analysis of mined flagella genes. The initial set of mined sequences was examined for proteins that matched to flagella proteins from other bacteria. Eleven putative XTAB flagella proteins were extracted (**table 5**) and queried using `blastp` for the top 500 (unique-taxon) matches to bacteria (taxid=2) in NCBI's *nr* database. The results were pruned to contain only hits to taxa common across all 11 queries, and separated based on query id. Each subset was aligned independently using `MUSCLE` (default parameters) (**Edgar 2004**), and regions of poor alignment were masked using `Gblocks` (default parameters) (**Castresana 2000**). All alignments were subsequently concatenated into a single data set for phylogeny estimation using `RAxML` with estimation of GAMMA, under four different amino acid substitution models: WAG, BLOSUM62, CPREV (plastid), and RTREV (retrovirus). Branch support for each tree was measured with bootstrapping (100 replications).

Evidence for symbiosis. Several approaches were taken to determine whether the mined sequence data suggested a symbiotic lifestyle for XTAB. First, AT bias in the mined contigs was compared to AT bias in the genomes of fifteen species of *Betaproteobacteria*: five free-living species, five facultative host-associated species, and five obligate intracellular symbionts (**table 6**). Second, codon usage bias was compared between all shared orthologs from these thirteen genomes plus XTAB, using the effective number of codons prime (`ENCprime`) software (**Novembre 2002**), which accounts for heterogeneity in nucleotide composition. Third, annotated XTAB proteins, especially those from OGs enriched for symbiont species, were examined for functions relevant to a symbiotic lifestyle, including nucleotide importers, amino acid transporters, type IV secretion system (T4SS) genes, and other niche-specific functions. Finally, the *X. tropicalis* genome was analyzed for the presence of integrated betaproteobacterial-like genes (see below).

Evaluating Bacterial Gene Transfer to the *X. tropicalis* Genome

AssemblySifter was used to ascertain whether any XTAB sequences included in the *X. tropicalis* assembly had strong evidence for being integrated in the *X. tropicalis* genome. A detailed description of AssemblySifter can be found in Chapter 2. Briefly, each of the 22,832 predicted proteins from the *X. tropicalis* assembly v4.1 was queried using `blastp` against three taxonomic subsets of the *nr* database at NCBI: 1) all *Betaproteobacteria* (NCBI taxid=28216); 2) all bacteria (NCBI taxid=2) except *Betaproteobacteria*; and 3) all eukaryotes (NCBI taxid=2759) except the *X. tropicalis* group (NCBI taxid=8363). The matches were pooled, ranked by S_m, and used to calculate comparative taxonomic nature scores (N_t). Bacterial-like *X. tropicalis* proteins were subsequently placed on the genomic scaffolds that contain their coding regions, and the scaffolds sifted for 1) short, purely bacterial scaffolds that likely represent bacterial genomic sequences, and 2) longer scaffolds mainly comprised of purely eukaryotic genes but containing

some bacterial-like genes. This second group was evaluated for potential bacteria-to-host LGTs by a combination of Nt score comparisons, exon count, scaffold length, and functional analysis of the top-5 AssemblySifter matches.

RESULTS & DISCUSSION

From an initial set of $2.2x10^7$ *X. tropicalis* genome sequencing project reads, 81.4% ($1.8x10^7$) survived the decontamination and quality control pipeline. The final set had a mean read length of 643.6 bases and included $1.2x10^{10}$ total bases. Full quality analysis results from `fastqc` can be found in **appendix D1** (before processing) and **appendix D2** (after processing).

SSU rDNA Analysis

A total of 8,190 SSU rDNA sequences were mined from the *X. tropicalis* genome project. The majority of reads (65.2%) were identified as eukaryotic, matching to the full-length 18S SSU rDNA sequence from *X. laevis*. The remaining 2849 reads were assigned to taxonomic groups according to the Greengenes annotations of their matches (**fig. 1**). All but one read matched to bacterial 16S rDNA sequences; the majority (90.6%) were assigned to *Paenibacillus* spp. (Firmicutes), with an additional 206 reads (7.2%) assigned to *Pseudomonas* spp. (*Gammaproteobacteria*) and 20 reads to various genera in the order Burkholderiales (*Betaproteobacteria*).

Pairwise divergence of bacterial 16S rDNA reads. The presence of a large number of *Pseudomonas*-like 16S rDNA reads in the *X. tropicalis* data complicated the analysis of 16S rDNA phylogeny, and the subsequent mining of betaproteobacterial reads as well, because of its similarity to betaproteobacterial *Acidovorax* spp. (Comamonadaceae) (**Willems et al. 1990**). Consequently, pairwise gene divergence was calculated using both the betaproteobacterial and gammaproteobacterial reads, and compared to full-length 16S rDNA sequences from both betaproteobacterial and pseudomonad genomes (**fig. 2**). Mean divergence of betaproteobacterial-like reads from each full-length betaproteobacterial sequence was low and significantly different ($p<0.001$, or $p<0.005$ for *Neisseria gonorrhoeae*) compared to the divergence of gammaproteobacterial-like reads from the same sequence. Similarly, mean divergence of gammaproteobacterial-like reads from each full-length 16S rDNA sequence from *Pseudomonas* spp. was low and significantly different ($p<0.001$) compared to mean divergence of betaproteobacterial-like reads from the same sequence. In contrast, mean divergence to full-length 16S rDNA sequences from *Paenibacillus* spp. was high for both betaproteobacterial-like and gammaproteobacterial-like reads, and not significantly different from each other ($p<0.01$).

Similar results were obtained by examining each read separately (**fig. 3**). The mean divergence of each betaproteobacterial-like read from the full-length betaproteobacterial sequences was lower and significantly different ($p<0.001$) when compared to its mean divergence from the full-length sequences from *Pseudomonas* spp. Conversely, the mean divergence of each gammaproteobacterial-like 16S rDNA read from pseudomonads was lower and significantly different ($p<0.001$) compared to its mean divergence from *Betaproteobacteria* (data not shown).

Finally, ANOVA of mean divergence of the betaproteobacterial-like 16S rDNA reads from 33 *Betaproteobacteria* sequences was significantly different ($p<0.001$) than divergence from either the 7 sequences from *Pseudomonas* spp. or the 5 sequences from *Paenibacillus* spp., further supporting the classification of these reads as betaproteobacterial. Moreover, ANOVA

also revealed a significant difference ($p<0.001$) in mean read divergence from 21 Comamonadaceae sequences compared to divergence from 12 *Betaproteobacteria* outside of the Comamonadaceae (see **fig. 2**). Taken together, these results indicate that 1) the betaproteobacterial-like 16S rDNA reads contain no contamination from either *Pseudomonas* spp. or *Paenibacillus* spp., and 2) the betaproteobacterial-like reads likely belong in the Comamonadaceae family specifically.

Phylogeny of betaproteobacterial-like 16S rDNA reads. All twenty betaproteobacterial-like 16S rDNA reads, along with the 33 full-length 16S rDNA sequences from *Betaproteobacteria* used to analyze pairwise divergence (see **table 4**), were used to construct a phylogenetic tree (**fig. 4**). A separate phylogeny (**fig. 5**) was also estimated using the same reads and full-length *Betaproteobacteria* sequences plus seven full-length rDNA sequences from *Pseudomonas* spp. (see **table 4**). Three reads were isolated away from the Comamonadaceae in both trees (two reads near the *Burkholderia* clade, one among the *Polaromonas*/*Variovorax* clade), and were subsequently discarded.

The remaining reads were embedded in the Comamonadaceae, monophyletic in the absence of sequences from *Pseudomonas* spp. (**fig. 4**) but split into two distinct groups of ten reads and seven reads in the presence of sequences from *Pseudomonas* spp. (**fig. 5**). The Pseudomonad sequences also introduced substantial reshuffling of *Acidovorax* spp. in general, consistent with these two genera being somewhat confounding (**Willems et al. 1990; STANIER et al. 1966**). Manual inspection of the original sequence alignment showed the two groups of ten and seven Comamonadaceae-like reads clustered at the 5' and 3' ends of the alignment, respectively, with adequate overlap; consequently, all seventeen reads were assembled into a single consensus sequence of 1660 bases (minimum of 3X coverage). Based on the results of pairwise divergence, multiple sequence alignment, and phylogeny estimation of the read data, it was concluded that this consensus sequence represents the 16S rDNA sequence from a single betaproteobacterial target, XTAB.

Phylogeny of the assembled target SSU rDNA. Under the neighborhood taxon sampling method, sequences from 91 taxa (including the target) aligned across 998 conserved positions were used to construct a phylogenetic tree (**fig. 6**). In this hypothesis, XTAB was placed with the earthworm symbiont *Verminephrobacter eiseniae,* in a branch of the Comamonadaceae that also included *Alicycliphilus denitrificans*, *Comamonas* spp. and *Delftia* spp., as well as various (polyphyletic) *Acidovorax* spp. The four families of Bukholderiales represented in this tree (Comamonadaceae, Burkholderiaceae, Alcaligenaceae, and Burkholderiales genera *incertae sedis*) showed little or no polyphyly, and branch support was generally high. A second phylogeny (**fig. 7**) was estimated by aligning sequences from 162 taxa across 1275 positions obtained by the cascading taxon sampling method. This tree was far less resolved, with generally poor branch support, very little cohesion even at the level of genus, and sequences from the three *Proteobacteria* taxa intermingled. XTAB appears on its own branch in a clade with several *Comamonas* spp. and *Acidovorax* spp. *V. eiseniae* was not present in this tree so its relationship to XTAB could not be evaluated. In addition, this tree contains a considerable number of uncultured and poorly classified sequences which are undoubtedly contributing to the noise. The cascading method yielded similar results using NCBI's *nr* database rather than *refseq_genomic*, suggesting this taxon sampling method may require additional refinement.

In conclusion, the 16S rDNA phylogeny supports an initial classification that places XTAB within the Comamonadaceae family, and more tentatively within a clade that contains species of *Comamonas*, *Delftia*, *Alicycliphilus*, *Verminephrobacter*, and *Acidovorax*.

Whole-Genome Read Analysis

From within the *X. tropicalis* genome sequence project, a total of 221,809 reads matched to 35 Comamonadaceae (best-matching) plus 135 Pseudomonad (best-competing) genomes (**fig. 8**). 6405 reads that preferentially binned with Comamonadaceae genomes were extracted, de-duplicated, and assembled into 1067 contigs (N_{50}=1333) for a consensus length of 1.3 Mb and an average %GC of 65.6%. This assembly constitutes the initial mined XTAB genome. The size of the XTAB assembly is smaller than sequenced Comamonadaceae genomes (3-7 Mb), including the vertically-inherited extracellular symbiont *V. eiseniae* (5.6 Mb). It is possible that XTAB is undergoing some degree of genome reduction, leading to a complete genome size similar to obligate intracellular *Betaproteobacteria* (e.g., the 1.6 Mb genome of *Polynucleobacter necessarius* subsp. necessarius, or the 1.73 Mb genome of "*Candidatus* Glomeribacter gigasporum"); however, it is more likely that the XTAB genome is small due to incomplete extraction or suboptimal assembly.

The mined XTAB contigs contained 1170 predicted gene models with at least 50% alignment to a known protein. An initial set of OGs constructed with proteins from XTAB, Comamonadaceae, and *Pseudomonas* spp. revealed that none of these XTAB proteins were strict orthologs of the *Pseudomonas* spp. Therefore, these 1170 sequences were used in all subsequent analyses as the set of mined XTAB genes (proteins).

Core XTAB data and whole-genome phylogeny. Several steps were taken to identify XTAB genes that are widely conserved among the *Betaproteobacteria*. First, 18,194 OGs were constructed with proteins from XTAB plus 54 select betaproteobacterial genomes (chosen to provide robust sampling across known *Betaprotebacteria* genera). A total of 388 XTAB proteins were found in OGs composed of at least one protein from 80% or more of all taxa. For each of these XTAB proteins, an *n*-taxon statement was prepared to determine its closest phylogenetic neighbor (see **fig. 9**). Out of 388 XTAB orthologs, 323 (83%) were most closely related to the Comamonadaceae branch of their *n*-taxon tree. These orthologs comprised the core data set for XTAB.

An estimated phylogeny of the 323 core XTAB orthologs with FastTree (**fig. 10**) placed XTAB on its own branch, ancestral to the *Comamonas*/*Delftia* clade within the Comamonadaceae. Bootstrap support for the specific XTAB branch is lower (0.66) than the surrounding branches (0.97-1.0), suggesting the exact placement of XTAB may improve with refinement. The consensus from a second phylogeny estimated with PhyloBayes (**fig. 11**) was similar to the FastTree phylogeny, and displayed an identical arrangement of the XG including XTAB, although branch support for family-level divisions in the Burkholderiales was consistently low (0.56-0.68). The primary difference between the two trees is the placement of "*Candidatus* Tremblaya princeps", which is coincident with "*Candidatus* Zinderia insecticola" (Oxalobacteraceae) in **fig. 10** but not in **fig. 11**; this is most likely a reflection of PhyloBayes and its ability to account for long branch attraction. In conclusion, while both of these results require further refinement, they are generally consistent with the phylogeny estimated using 16S rDNA (neighborhood sampling) shown in **fig. 6**.

Many of the bacteria along the branch of Comamonadaceae that includes XTAB mediate processes in environmental toxicology, including the biodegradation of aromatic hydrocarbons and heavy metals. *Delftia acidovorans* was originally isolated from acetamide-enriched soil (**Wen et al. 1999; Hoffmann et al. 2003**) and has the ability to precipitate gold particles, presumably as a defense mechanism (**Johnston et al. 2013**). *Alicycliphilus denitrificans* was originally isolated from wastewater and has been shown to degrade aromatic hydrocarbons (**Mechichi et al. 2003**). *C. testosteroni* is an environmental microbe found in diverse environments, including marshes, marine habitats, and contaminated sludge; at least one strain (CNB-2) has been shown to utilize aromatic compounds as its sole source of carbon and nitrogen (**Ma et al. 2009**). *Acidovorax* sp. KKS102 is known to degrade polychlorinated biphenyl compounds (**Ohtsubo et al. 2012**). Recognizing the unusual xenobiotic character of many of the species in this clade, it is proposed that this branch, including *Alicycliphilus*, *Comamonas*, *Verminephrobacter*, XTAB, and several *Acidovorax* spp., be tentatively called the Xenobiotic Group (XG) of the Comamonadaceae.

To determine whether XTAB is typical in its divergence from other *Betaproteobacteria*, and other XG species in particular, genome divergence was calculated using the most conserved 2004 positions within the core data and sampling one representative member of each betaproteobacterial genus (five species from the ill-defined *Acidovorax* genus). XTAB was most similar to *Acidovorax ebreus* TPSY (5% divergence), but no more than 8% divergent from any other XG genome (**fig. 12**). Mean divergence of XTAB with all other XG was 6.2%, and typical of this group as a whole (5.8%). Mean divergence of XTAB with all *Betaproteobacteria* was higher (20.4%) but still not unusual; mean divergence across all *Betaprotebacteria* was 18.9% and ranged from 1% to 32%. These results are consistent with the estimated phylogeny based on 16S rDNA (**fig. 6**) and conserved core proteins (**figs. 10** and **11**).

Exploring evidence for a symbiotic lifestyle. In addition to bioremediation, several bacteria in the Xenobiotic Group are known to have pathogenic or symbiotic characteristics. *C. testosteroni* (**Farshad et al. 2012**) and *D. acidovorans* (**Oliver et al. 2005**) are occasional clinical isolates and opportunistic pathogens of humans, typically associated with bacteremia. *Acidovorax citrulli* (**Bahar et al. 2011**) and *Acidovorax avenae* (**Xie et al. 2011**) are agriculturally important facultative phytopathogens of Cucurbitaceae (melons, gourds, and squashes). *V. eiseniae* is a vertically transmitted, extracellular symbiont of earthworms that resides in the nephridia and may be involved in nitrogen recycling (**Pinel et al. 2008; Lund et al. 2009**). Consequently, the possibility that XTAB is a symbiont of *X. tropicalis* was explored.

An estimated phylogeny of eleven core structural flagellar proteins (**fig. 13**) placed XTAB with *D. acidovorans* in a clade dominated by Burkholderiales species. Several known betaproteobacterial symbionts also appear in this clade, including *Acidovorax citrulli*, several *Verminephrobacter* spp., and *Curvibacter* putative symbiont of *Hydra magnipapillata*. Intriguingly, *A. citrulli* requires both its polar flagellum and Type IV pili (T4P) for full virulence on melon plants (**Bahar et al. 2011**), and *V. eiseniae* requires the same two structures to successfully colonize earthworm nephridia (**Dulla et al. 2011**). Preliminary analysis has so far revealed at least sixteen candidate flagellar and four candidate T4P genes in XTAB, suggesting that it may at a minimum possess these structures; however, no firm conclusions can yet be drawn concerning the significance of these genes in XTAB with respect to pathogenesis.

Several additional proteins suggest that XTAB may be capable of symbiotic or pathogenic associations. The protein XTAB000910 is over 90% identical to the virulence-associated protein

VapD from *V. eiseniae*. VapD-like proteins have been found in many bacteria (**Kwon et al. 2012**); although their exact function remains unclear, *vap* genes in general are known as regulators of other virulence genes (**Katz et al. 1992**). In addition to VapD, XTAB contains several T4SS-like proteins. T4SSs originally evolved from bacterial conjugation systems and play a role in effector delivery or genetic exchange during pathogenesis (**Zhang et al. 2009**). While these results are tantalizing, more work is needed to confirm the expression and activity of VapD, T4SS, or other pathogenesis-related proteins.

Many bacterial symbionts, especially bacteriosome-associated symbionts of insects, are characterized by a nucleotide composition that favors adenosine and thymine (AT bias) (**Moran et al. 2008**). This is reflected in low genomic %GC values; some symbiont genomes in the Rickettsiales, for example, exhibit %GC values around 25% or lower (see Chapter 3). Mean %GC of the mined XTAB contigs was high (68%) and showed no AT bias (**fig. 14**). However, %GC in the genomes of other *Betaproteobacteria* were similarly high, ranging from 45% (free-living *Polynucleobacter necessarius* subsp. *asymbioticus*) to 69% (host-associated *Acidovorax citrulli*). Furthermore, no systematic AT bias was apparent in either the free-living (58.4%), host-associated (63.1%), or obligate endosymbiotic (56.0%) species. In addition, no significant differences ($p<0.1$) were found between the mean %GC values for these three groups. A slight AT bias was detected in *P. necessarius*, but %GC was nearly identical for both the free-living (44.8%) and endosymbiotic (45.6%) subspecies. These results suggest that AT bias is not a general feature of endosymbiotic genomes in the *Betaproteobacteria*.

Negative (purifying) selection ensures that deleterious mutations are removed from the genome, thereby contributing to the long-term stability of biological structures. Bacterial endosymbionts often exhibit relaxed purifying selection (**Kjeldsen et al. 2012**), as ecological pressure from the host induces the symbiont to adapt to an unstable environment. Codon usage bias is one way to identify the extent of purifying selection: less bias indicates a relaxation of purifying selection, as slightly deleterious mutations are allowed to accumulate leading to suboptimal codon usage (**Hershberg & Petrov 2008**). No significant difference was found in mean codon usage bias between XTAB, the free-living *Acidovorax* sp. JS42, and the facultative phytopathogenic *Acidovorax citrulli* (data not shown).

Bacterial Genes in the *X. tropicalis* Genome

To investigate the possibility that the *X. tropicalis* genome harbors LGTs of betaproteobacterial origin, the *X. tropicalis* assembly was evaluated for the presence of betaproteobacterial-like sequences. Of the 22,832 predicted *X. tropicalis* proteins, 94.5% (21,662) were selected for further analysis based on the presence of at least one known match from outside of the *X. tropicalis* group. Of these known proteins, 5.0% (1083) matched best to bacterial proteins, and 4.1% (893) to proteins from *Betaproteobacteria* in particular (**fig. 15a**). Of those betaproteobacterial best matches, 89% (794) occurred in the Comamonadaceae specifically (**fig. 15b**), strongly suggesting the presence of a sizable bacterial component skewed heavily toward the Comamonadaceae within the *X. tropicalis* assembly. This was not due to over-representation of strains from *Betaproteobacteria* or Comamonadaceae, since these groups accounted for only 20.8% of the total bacterial strains (**fig. 15c**). This is consistent with the results from mining the reads directly, and strongly suggests that XTAB sequences are present in the *X. tropicalis* assembly.

Initial analysis of top blast matches also indicated that 41% (78) of the bacterial best matches from outside of the *Betaproteobacteria* were hits to *Pseudomonas* spp. proteins. This number is

smaller than expected, since the large number of mined *Pseudomonas*-like 16S rDNA reads (206) implied a significant presence of this gammaproteobacterium; therefore it seems likely that many of the Pseudomonad reads were successfully removed before assembly. Also surprisingly, only one *X. tropicalis* protein had a top blast match to a protein from a *Paenibacillus* species, a result that is inconsistent with the enormous number of *Paenibacillus*-like 16S rDNA reads in the original data (2579). Intriguingly, only 2,718 reads total were extracted by ReadMiner using six *Paenibacillus* spp. genomes from PATRIC, a number largely unaffected by the presence of Comamonadaceae or Pseudomonad genomes. The reason for this discrepancy between the results from MetaMiner, ReadMiner, and AssemblySifter for *Paenibacillus* spp. is not known.

Taxonomic nature of X. tropicalis proteins. To address more comprehensively whether any of the *X. tropicalis* proteins might be bacterial, either from a bacterial genomic fragment in the assembly or as an LGT in the *X. tropicalis* genome, each protein was evaluating according to pairwise N_t calculations (**fig. 16**). Each protein's N_t score for a single group (eukaryote, bacteria not *Betaproteobacteria*, and *Betaproteobacteria*) is a weighted, normalized score of its top-5 blast matches to that group. They range from a maximum of one (all top-5 matches in that group) to a minimum of zero (no top-5 matches in that group). Pairwise N_t scores for each protein are calculated by simple subtraction of its N_t scores for two groups. An A-B pairwise N_t value of one for a protein means that protein has all top-5 matches in group A and none in group B; a value of -1 means all top-5 matches for that protein are in group B and none in group A.

Pairwise N_t analysis of *X. tropicalis* proteins revealed three distinct categories (**fig. 16**). 3.3% (713 proteins) scored exactly -1 for both the eukaryote-*Betaproteobacteria* and bacteria-*Betaproteobacteria* comparisons, indicating all of their top-5 matches were betaproteobacterial. An additional 2.3% (503 proteins) were intermediate, with a mixture of N_t scores; the remaining 94.4% (20,446 proteins) scored one for eukaryote-*Betaproteobacteria* and zero for bacteria-*Betaproteobacteria*, indicating all of their top-5 matches were eukaryotic. This third and largest group of proteins was determined to be comprised of true *X. tropicalis* proteins and was not analyzed further in this study.

Taxonomic nature of sifted scaffolds. The 1216 proteins that exhibited at least some bacterial nature according to pairwise N_t analysis were encoded by genes found across 446 genomic scaffolds in the *X. tropicalis* assembly (24 genes could not be associated with a *X. tropicalis* scaffold id in NCBI, and were set aside). A total of 348 sifted scaffolds (78%) had bacterial-to-total gene ratios of one (**fig. 17**), indicating they contained no purely eukaryotic genes. A subset (291) of these scaffolds also had combined mean bacterial nature scores of one (mean betaproteobacterial N_t plus mean non-betaproteobacterial N_t), indicating that none of their bacterial-like genes had any eukaryotic nature. The complete lack of eukaryotic genes or bacterial-like genes with eukaryotic nature, plus the relatively sparseness of these 291 scaffolds (none contained more than 18 total genes), suggested they were bacterial genomic sequences.

The remaining 155 sifted scaffolds were binned into two groups. Group I scaffolds (98) had bacterial-to-total gene ratios less than one, indicating that they contained a combination of bacterial-like and purely eukaryotic genes (**fig. 18**), and some contained over a hundred total genes. Due to their makeup (few bacterial genes embedded among many eukaryotic genes) and relatively large size, Group I scaffolds were deemed most likely locations for bacterial LGT events. A second group of 57 scaffolds had bacterial-to-total gene ratios of one (no purely eukaryotic genes) and mean bacterial nature scores less than one, indicating that all of their

component genes evinced a mixture of bacterial and eukaryotic nature. These Group II scaffolds (data not shown) were fairly sparse (none contained more than eleven total genes) and were considered more likely to be bacterial in origin.

A total of 103 *X. tropicalis* proteins were mapped to Group I scaffolds and evaluated for the possible existence of bacterial LGTs into the *X. tropicalis* genome. In general, Group I scaffolds were dominated by retrotransposable elements, widely conserved homologues, and proteins with repeat element domains with similarity to both bacteria and eukaryotes. There were seven betaproteobacterial-like proteins, two of which had introns. One of these (XP_002942295.1) had eleven introns and was 6% betaprotebacterial as computed by N_t analysis, and its non-hypothetical matches indicated significant similarity to polyhydroxybutyrate (PHB) depolymerase, including one from *Ramlibacter tataouinensis* (45% identity, >75% alignment over both query and subject). PHB is synthesized and stored intracellularly by a variety of free-living and endosymbiotic bacteria as a carbon storage form during periods of unbalanced growth (**Jendrossek 2002; Lodwig et al. 2005**). It is stored in inclusion bodies and can accumulate to more than 90% of dry cell mass (**Jendrossek 1998**). Extracellular PHB depolymerases are produced by various microbes (including bacteria, fungi, and mycetes) and act to degrade PHB polymers in soil, seawater, freshwater, and activated sludge (**Kasuya et al. 1997**). Extracellular PHB depolymerase activity in bacteria is negatively regulated by soluble carbon sources that the bacteria can utilize as preferential energy sources (**Jendrossek 1998**). PHB depolymerase activity is unknown in amphibians; however, *Betaproteobacteria*, and members of the Comamonadaceae in particular, have been shown to be very effective at degrading extracellular PHB under denitrifying conditions (**Khan et al. 2002**). Therefore, it is possible that this *X. tropicalis* protein represents an LGT into the *X. tropicalis* genome from an associated microbe such as XTAB. However, more work is needed to confirm its phylogenetic origin and expression as a functional protein.

CONCLUSION

The results of this study reveal the presence of a novel betaproteobacterial (Comamonadaceae) organism (XTAB) from within the genome sequencing reads of *Xenopus* (Silurana) *tropicalis*. In addition, a significant bacterial presence heavily skewed toward the Comamonadaceae was also detected and isolated from the assembled *X. tropicalis* genome using AssemblySifter. Combined analyses of mined XTAB genes plus regions of sifted *X. tropicalis* scaffolds that were deemed most likely sites of bacterial LGTs revealed suggestions of, but no strong evidence to support, a lasting symbiotic relationship between XTAB and *X. tropicalis*.

These results demonstrate the combined effectiveness of the ReadMiner workflow at extracting and assembling sufficient genomic sequence for a single target to enable robust whole-genome phylogeny and functional analysis, even in the presence of large amounts of non-target bacterial sequences. They also demonstrate the ability of the AssemblySifter workflow to computationally sift the assembled host genome for bacterial genomic elements, and identify potential regions of bacterial LGT into the host genome.

LITERATURE CITED

Bahar O, Levi N, Burdman S. 2011. The cucurbit pathogenic bacterium *Acidovorax citrulli* requires a polar flagellum for full virulence before and after host-tissue penetration. Mol. Plant Microbe Interact. 24:1040–1050. doi: 10.1094/MPMI-02-11-0041.

Bruland N, Bathe S, Willems A, Steinbüchel A. 2009. *Pseudorhodoferax soli* gen. nov., sp. nov. and *Pseudorhodoferax caeni* sp. nov., two members of the class Betaproteobacteria belonging to the family Comamonadaceae. INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY. 59:2702–2707. doi: 10.1099/ijs.0.006791-0.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.

Chevreux B. 2005. MIRA: An Automated Genome and EST Assembler. Ph.D. Thesis. 1–171.

Deng W et al. 2010. DIVEIN: a web server to analyze phylogenies, sequence divergence, diversity, and informative sites. BioTechniques. 48:405–408. doi: 10.2144/000113370.

DeSantis TZ et al. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 72:5069–5072.

Du Y, Maslov DA, Chang KP. 1994. Monophyletic origin of beta-division proteobacterial endosymbionts and their coevolution with insect trypanosomatid protozoa *Blastocrithidia culicis* and *Crithidia* spp. Proc Natl Acad Sci USA. 91:8437–8441.

Dulla GFJ, Go RA, Stahl DA, Davidson SK. 2011. *Verminephrobacter eiseniae* type IV pili and flagella are required to colonize earthworm nephridia. ISME J. 6:1166–1175. doi: 10.1038/ismej.2011.183.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res. 32:1792–1797. doi: 10.1093/nar/gkh340.

Farshad S, Norouzi F, Aminshahidi M, Heidari B, Alborzi A. 2012. Two cases of bacteremia due to an unusual pathogen, *Comamonas testosteroni* in Iran and a review literature. J Infect Dev Ctries. 6:521–525.

Ghignone S et al. 2011. The genome of the obligate endobacterium of an AM fungus reveals an interphylum network of nutritional interactions. ISME J. 6:136–145. doi: 10.1038/ismej.2011.110.

Gillespie JJ et al. 2011. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. Infect Immun. 79:4286–4298. doi: 10.1128/IAI.00207-11.

Hellsten U et al. 2010. The genome of the Western clawed frog *Xenopus tropicalis*. Science. 328:633–636. doi: 10.1126/science.1183670.

Hershberg R, Petrov DA. 2008. Selection on codon bias. Annu Rev Genet. 42:287–299. doi: 10.1146/annurev.genet.42.110807.091442.

Hoffmann D, Kleinsteuber S, Müller RH, Babel W. 2003. A transposon encoding the complete 2, 4-dichlorophenoxyacetic acid degradation pathway in the alkalitolerant strain *Delftia acidovorans* P4a. Microbiology (Reading, Engl). 149:2545–2556. doi: 10.1099/mic.0.26260-0.

Jendrossek D. 2002. Extracellular Polyhydroxyalkanoate (PHA) Depolymerases: The Key Enzymes of PHA Degradation. Biopolymers Online.

Jendrossek D. 1998. Microbial degradation of polyesters: a review on extracellular poly (hydroxyalkanoic acid) depolymerases. Polymer degradation and stability. 59:317–325. doi: 10.1016/S0141-3910(97)00190-0.

Johnston CW et al. 2013. Gold biomineralization by a metallophore from a gold-associated microbe. Nat Chem Biol. 9:241–243. doi: 10.1038/nchembio.1179.

Kasuya K et al. 1997. Biochemical and molecular characterization of the polyhydroxybutyrate depolymerase of *Comamonas acidovorans* YM1609, isolated from freshwater. Appl Environ Microbiol. 63:4844–4852.

Katz ME, Strugnell RA, Rood JI. 1992. Molecular characterization of a genomic region associated with virulence in *Dichelobacter nodosus*. Infect Immun. 60:4586–4592.

Khan ST, Yamamoto M, Hiraishi A. 2002. Members of the family Comamonadaceae as primary poly(3-hydroxybutyrate-co-3-hydroxyvalerate)-degrading denitrifiers in activated sludge as revealed by a polyphasic approach. Appl Environ Microbiol. 68:3206–3214.

Kjeldsen KU et al. 2012. Purifying selection and molecular adaptation in the genome of *Verminephrobacter*, the heritable symbiotic bacteria of earthworms. Genome Biol Evol. 4:307–315. doi: 10.1093/gbe/evs014.

Kwon A-R et al. 2012. Structural and biochemical characterization of HP0315 from *Helicobacter pylori* as a VapD protein with an endoribonuclease activity. Nucl. Acids Res. 40:4216–4228. doi: 10.1093/nar/gkr1305.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics. 25:2286–2288. doi: 10.1093/bioinformatics/btp368.

Lodwig EM et al. 2005. Role of polyhydroxybutyrate and glycogen as carbon storage compounds in pea and bean bacteroids. Mol. Plant Microbe Interact. 18:67–74. doi: 10.1094/MPMI-18-0067.

Lund MB et al. 2009. Diversity and host specificity of the *Verminephrobacter*-earthworm symbiosis. Environmental Microbiology. doi: 10.1111/j.1462-2920.2009.02084.x.

Ma Y-F et al. 2009. The complete genome of *Comamonas testosteroni* reveals its genetic adaptations to changing environments. Appl Environ Microbiol. 75:6812–6819. doi: 10.1128/AEM.00933-09.

McCarthy FM et al. 2006. AgBase: a functional genomics resource for agriculture. BMC Genomics. 7:229. doi: 10.1186/1471-2164-7-229.

McCutcheon JP, Dohlen von CD. 2011. An Interdependent Metabolic Patchwork in the Nested Symbiosis of Mealybugs. Curr Biol. 21:1366–1372. doi: 10.1016/j.cub.2011.06.051.

Mechichi T, Stackebrandt E, Fuchs G. 2003. *Alicycliphilus denitrificans* gen. nov., sp. nov., a cyclohexanol-degrading, nitrate-reducing beta-proteobacterium. INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY. 53:147–152.

Moran NA, McCutcheon JP, Nakabachi A. 2008. Genomics and evolution of heritable bacterial symbionts. Annu Rev Genet. 42:165–190. doi: 10.1146/annurev.genet.41.110306.130119.

Novembre JA. 2002. Accounting for background nucleotide composition when measuring codon usage bias. Mol Biol Evol. 19:1390–1394.

Ohtsubo Y, Maruyama F, Mitsui H, Nagata Y, Tsuda M. 2012. Complete genome sequence of *Acidovorax* sp. strain KKS102, a polychlorinated-biphenyl degrader. J Bacteriol. 194:6970–6971. doi: 10.1128/JB.01848-12.

Oliver JW, Stapenhorst D, Warraich I, Griswold JA. 2005. *Ochrobactrum anthropi* and *Delftia acidovorans* to bacteremia in a patient with a gunshot wound. Infectious Diseases in Clinical Practice. 13:78–81.

Pinel N, Davidson SK, Stahl DA. 2008. *Verminephrobacter eiseniae* gen. nov., sp. nov., a nephridial symbiont of the earthworm Eisenia foetida (Savigny). INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY. 58:2147–2157. doi: 10.1099/ijs.0.65174-0.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS ONE. 5:e9490. doi: 10.1371/journal.pone.0009490.

Rosenstein NE, Perkins BA, Stephens DS, Popovic T, Hughes JM. 2001. Meningococcal disease. N Engl J Med. 344:1378–1388. doi: 10.1056/NEJM200105033441807.

Saldias MS, Valvano MA. 2009. Interactions of *Burkholderia cenocepacia* and other *Burkholderia cepacia* complex bacteria with epithelial and phagocytic cells. Microbiology. 155:2809–2817. doi: 10.1099/mic.0.031344-0.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22:2688–2690. doi: 10.1093/bioinformatics/btl446.

STANIER RY, PALLERONI NJ, DOUDOROFF M. 1966. The aerobic pseudomonads: a taxonomic study. J. Gen. Microbiol. 43:159–271.

Wen A, Fegan M, Hayward C, Chakraborty S, Sly LI. 1999. Phylogenetic relationships among members of the Comamonadaceae, and description of *Delftia acidovorans* (den Dooren de Jong 1926 and Tamaoka et al. 1987) gen. nov., comb. nov. Int. J. Syst. Bacteriol. 49 Pt 2:567–576.

Willems A et al. 1990. Acidovorax, a new genus for *Pseudomonas facilis*, *Pseudomonas delafieldii*, E. Falsen (EF) group 13, EF group 16, and several clinical isolates, with the species *Acidovorax facilis* comb. nov., *Acidovorax delafieldii* comb. nov., and *Acidovorax temperans* sp. nov. Int. J. Syst. Bacteriol. 40:384–398.

Xie G-L et al. 2011. Genome sequence of the rice-pathogenic bacterium *Acidovorax avenae* subsp. avenae RS-1. J Bacteriol. 193:5013–5014. doi: 10.1128/JB.05594-11.

Zhang R, LiPuma JJ, Gonzalez CF. 2009. Two type IV secretion systems with different functions in Burkholderia cenocepacia K56-2. Microbiology (Reading, Engl). 155:4005–4013. doi: 10.1099/mic.0.033043-0.

FIGURES

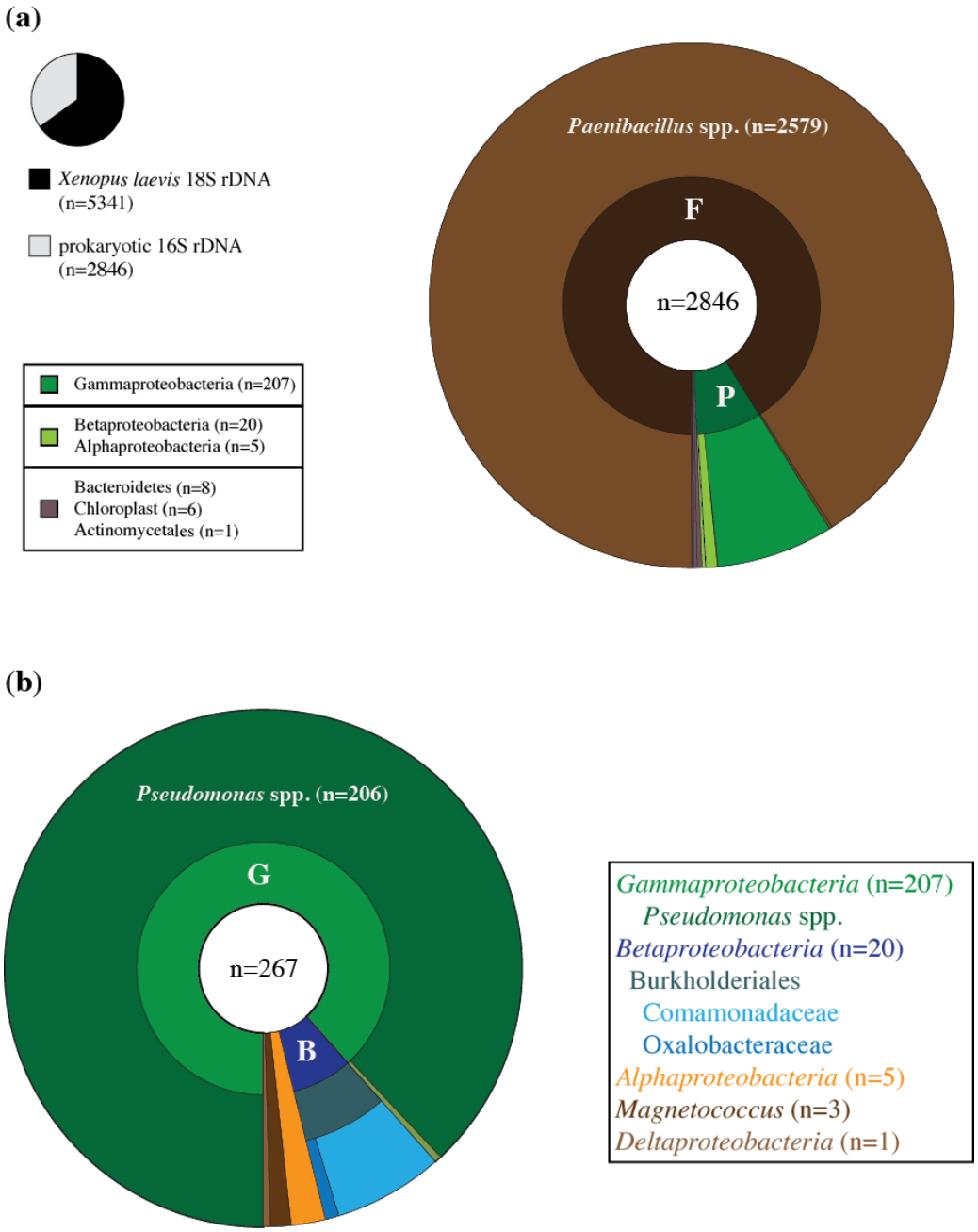


Figure 1. Distribution of bacterial 16S rDNA reads within the *Xenopus tropicalis* genome trace read archive. (**a**) Identification of 2846 prokaryotic 16S rDNA sequences among 5341 *Xenopus* spp. 18S rDNA sequences (top left), and distribution of the prokaryotic sequences between Firmicutes (F), *Proteobacteria* (P), and other higher-order taxa (top right). (**b**) Distribution of 267 proteobacterial 16S rDNA sequences between *Gammaproteobacteria* (G), *Betaproteobacteria* (B), and other taxa.

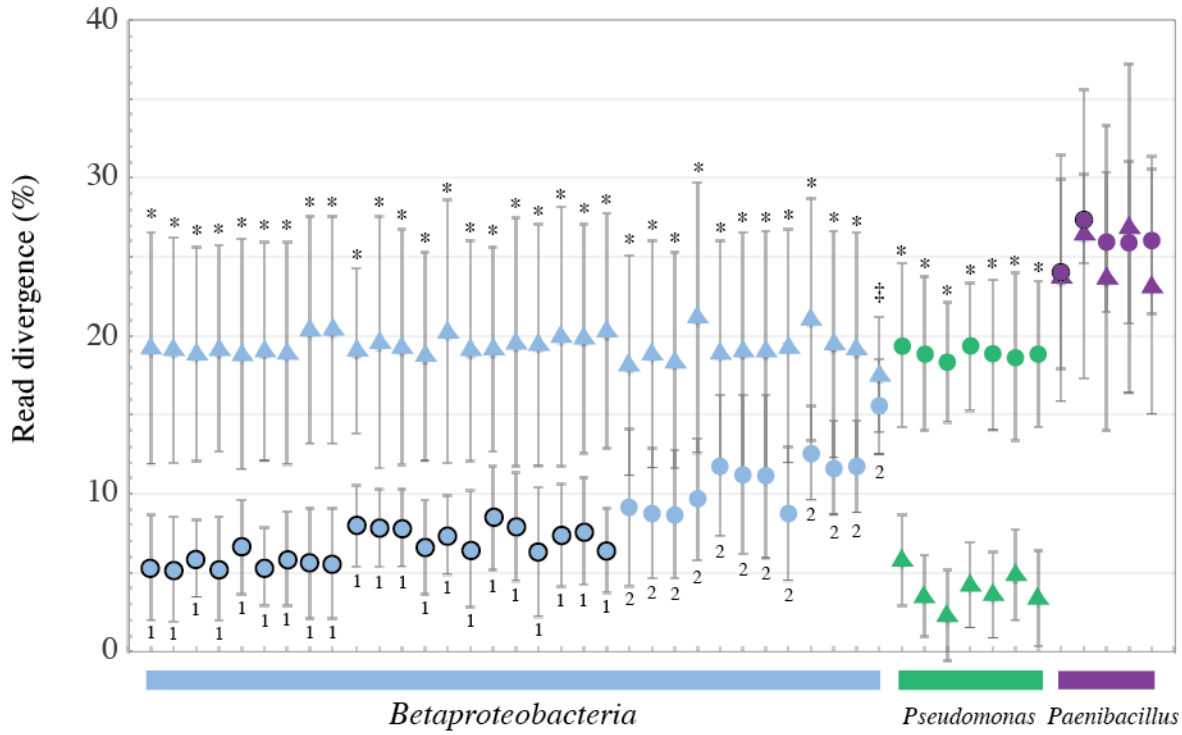


Figure 2. Mean read divergence for 20 mined betaproteobacterial-like (circles) and 206 mined gammaproteobacterial-like (triangles) 16S rDNA reads. Divergence was calculated from full-length 16S rDNA sequences of 33 *Betaproteobacteria* (blue), 7 *Psuedomonas* spp. (green), and 5 *Paenibacillus* spp. (purple). An asterisk ($p<0.001$) or a double-hash ($p<0.005$) denotes a significant difference between the corresponding betaproteobacterial-like and gammaproteobacterial-like means for that species. Numbers indicate the results of an ANOVA ($p<0.001$) comparing the betaproteobacterial-like read divergence from two sub-groups of *Betaproteobacteria*: 1) 21 Comamonadaceae (blue circles with heavy black borders); and 2) 12 non-Comamonadaceae *Betaproteobacteria* (blue circles with no borders).

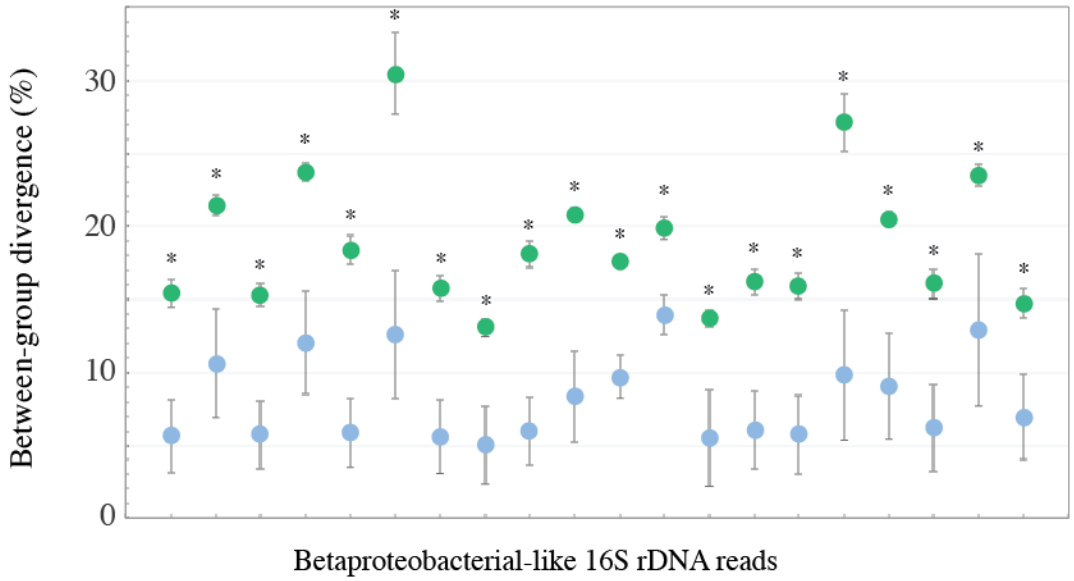


Figure 3. Mean between-group divergence for each of 20 mined betaproteobacterial-like 16S rDNA reads, from 33 full-length *Betaproteobacteria* (blue) and 7 full-length *Pseudomonas* spp. (green) sequences. An asterisk denotes a significant between-group difference ($p<0.001$). Error bars show plus or minus one standard deviation.

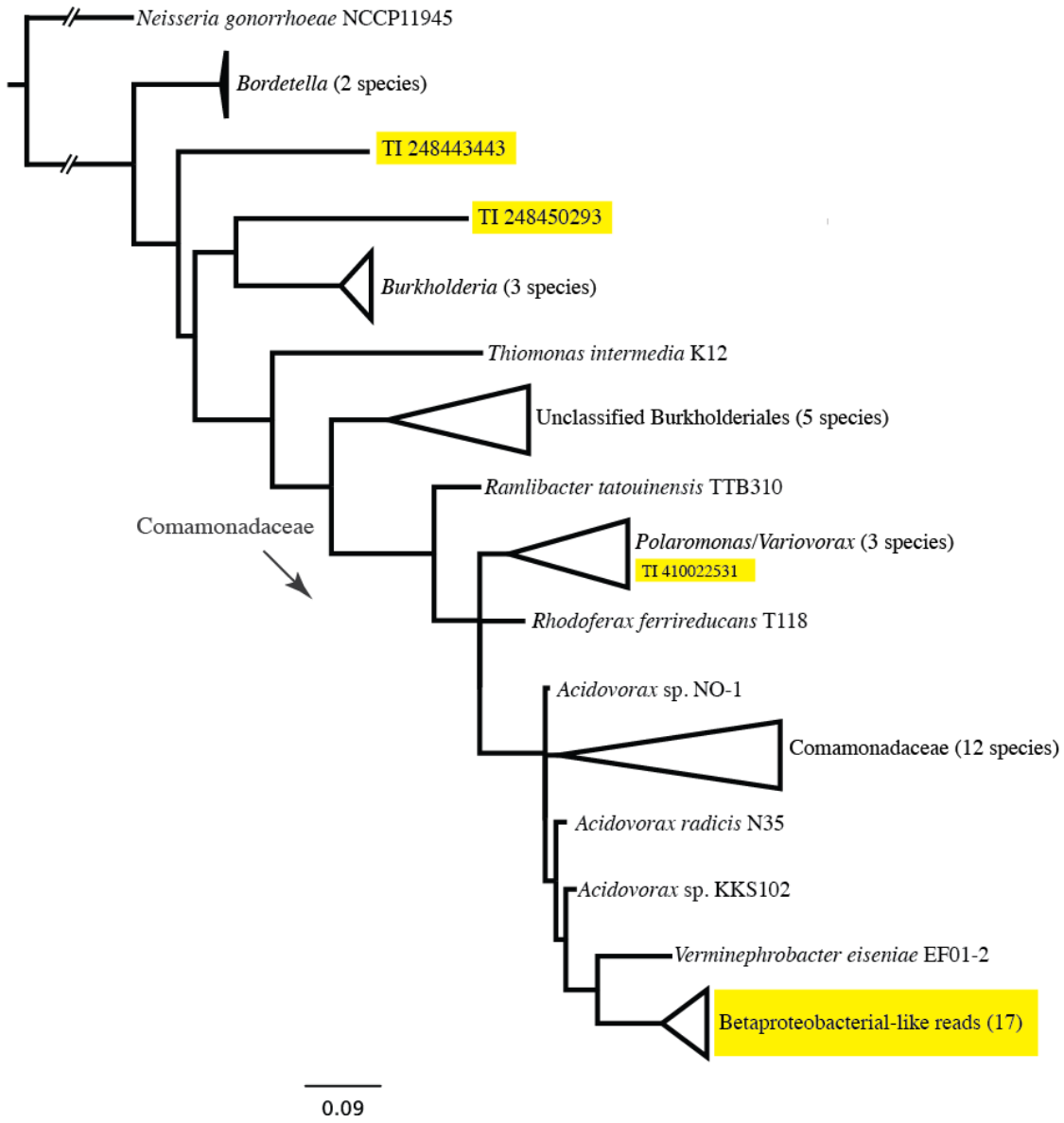


Figure 4. Phylogeny of SSU rDNA sequences estimated for 20 betaproteobacterial-like 16S rDNA reads (yellow) plus full-length 16S rDNA sequences from 33 *Betaproteobacteria* taxa. See text for alignment and tree-building methods.

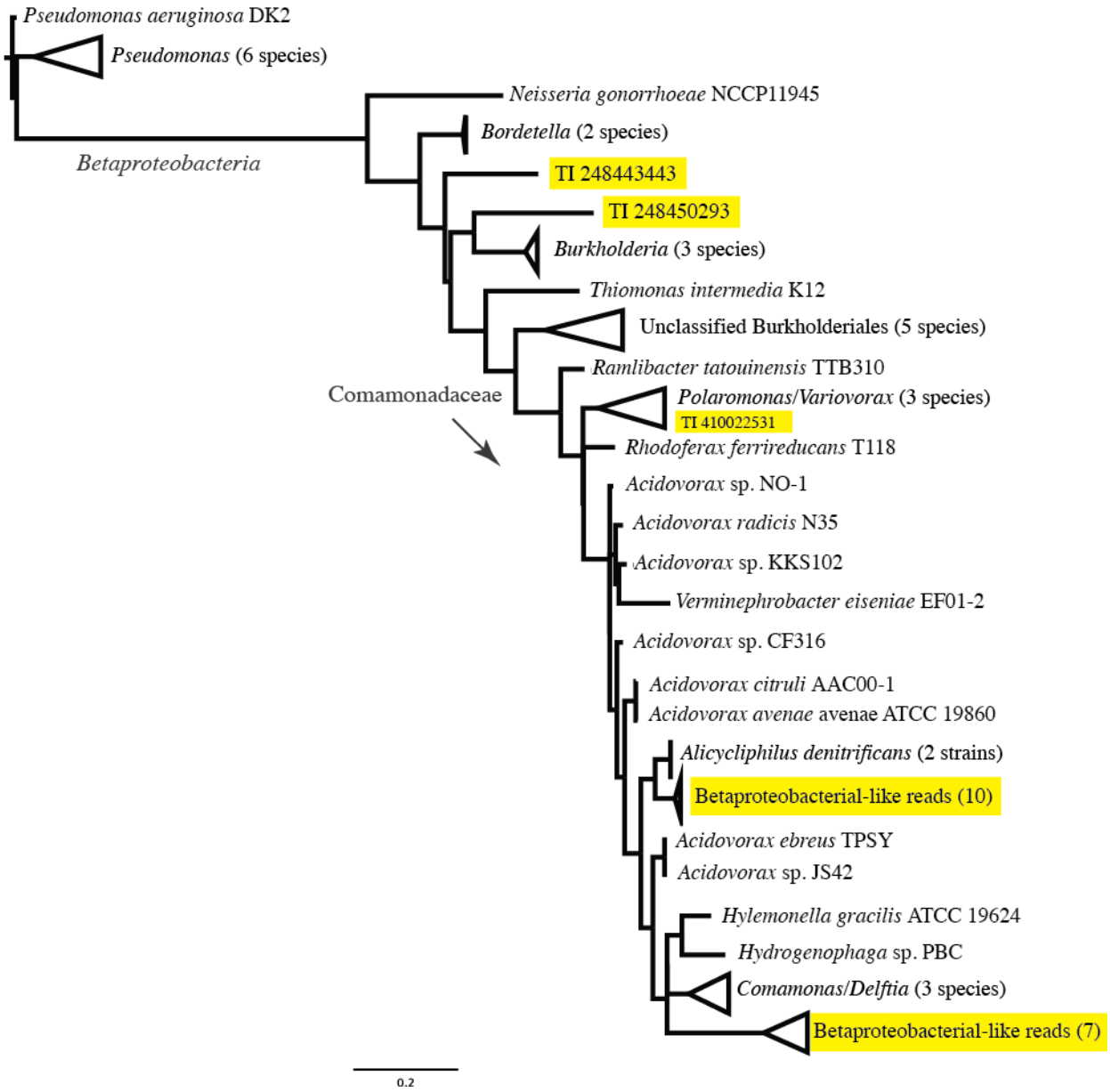


Figure 5. Phylogeny of SSU rDNA sequences estimated for 20 betaproteobacterial-like 16S rDNA reads (yellow), plus full-length 16S rDNA sequences from 33 *Betaproteobacteria* taxa and 7 *Pseudomonas* species. See text for alignment and tree-building methods.

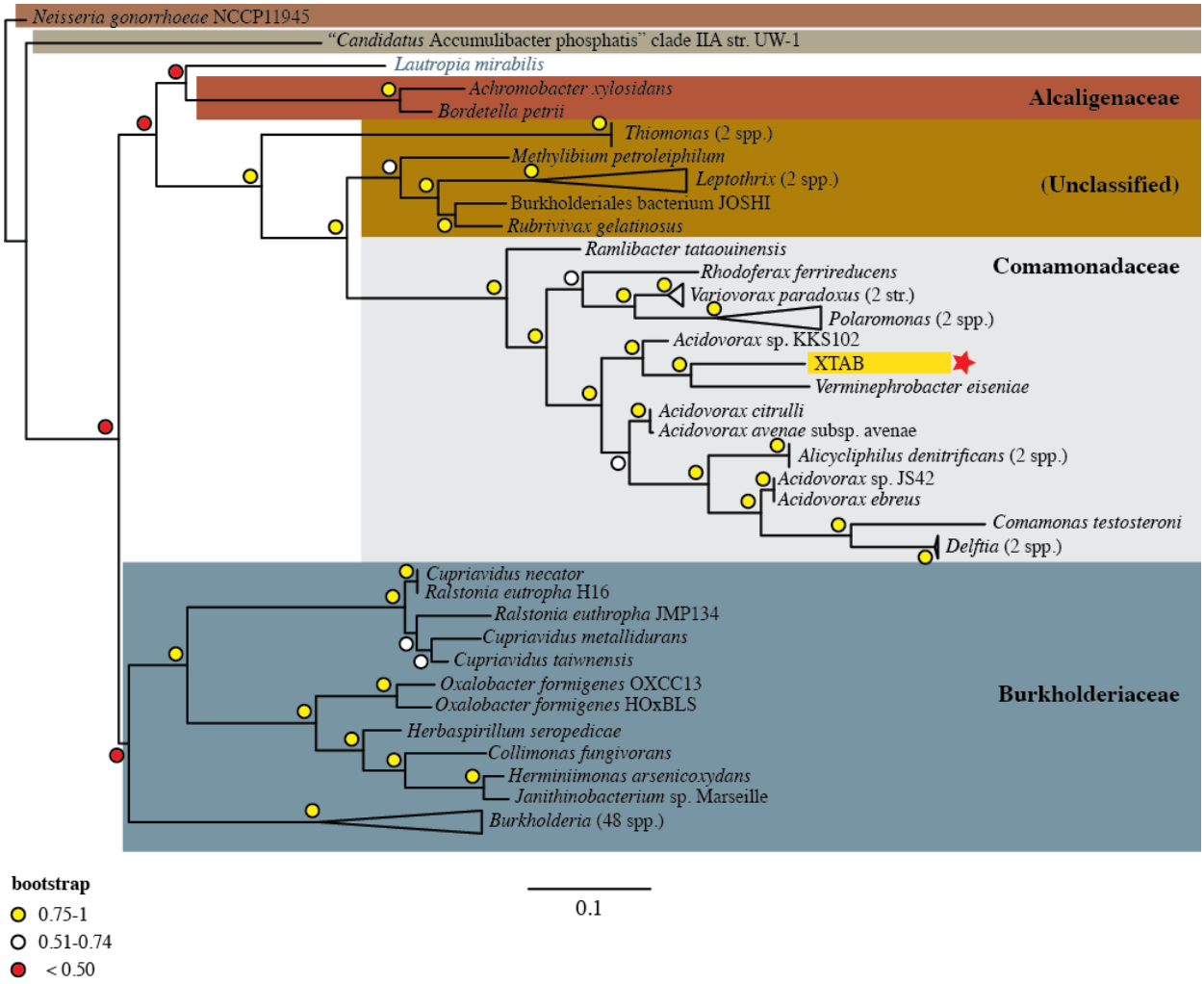


Figure 6. Phylogeny of SSU rDNA sequences estimated for 90 Burkholderiales taxa and 2 outgroup taxa. See text for alignment and tree-building details. Taxa were chosen by the neighborhood sampling method using the XTAB 16S rDNA sequence as the query.

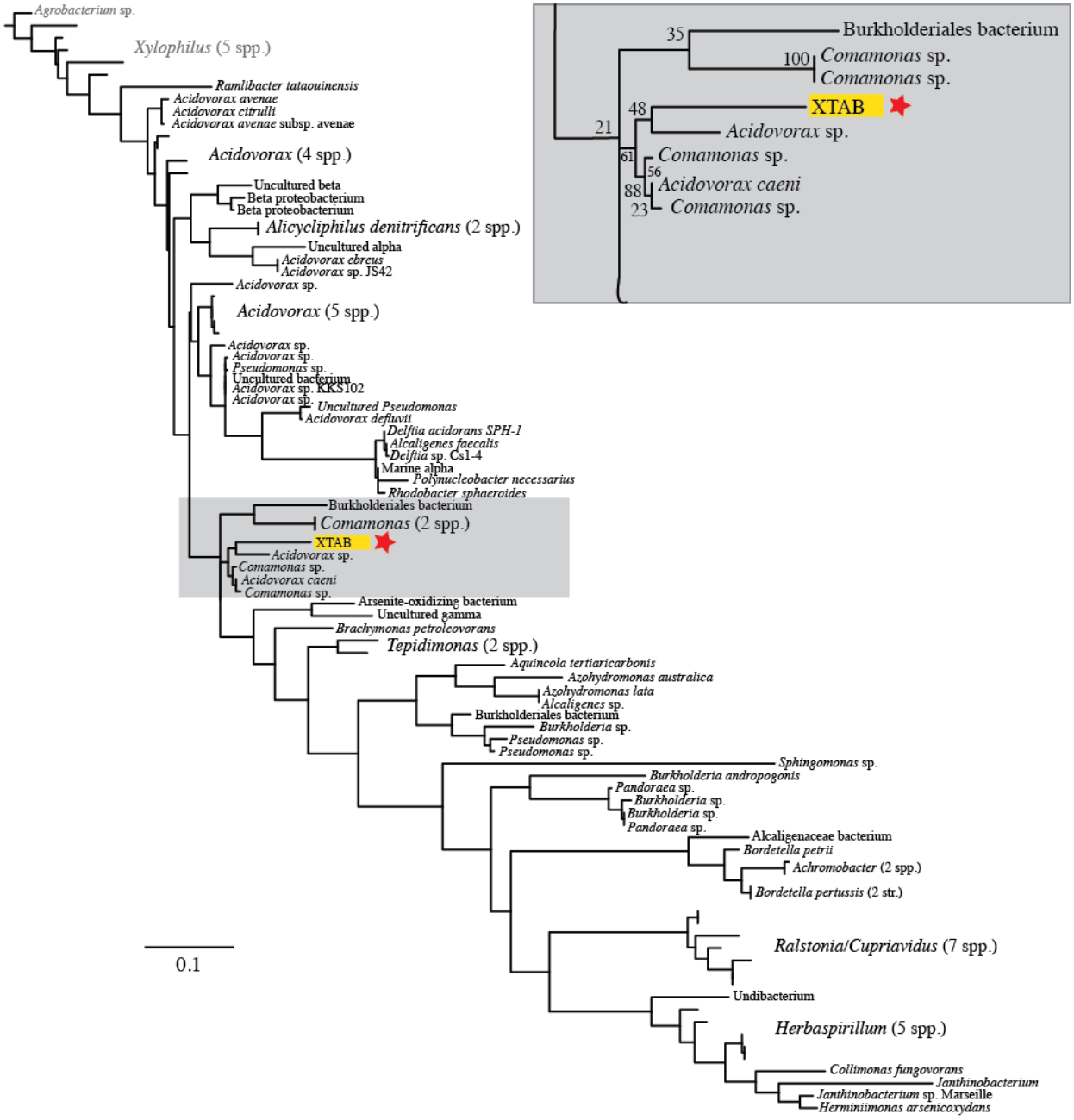


Figure 7. Phylogeny of SSU rDNA sequences estimated for 75 Burkholderiales, 5 *Betaproteobacteria* (not Burkholderiales), 5 *Alphaproteobacteria*, and 5 *Gammaproteobacteria* taxa. See text for alignment and tree-building details. Taxa were chosen by the cascading sampling method using the XTAB 16S rDNA sequence as the query. The inset shows detail for the dark gray boxed area of the main tree.

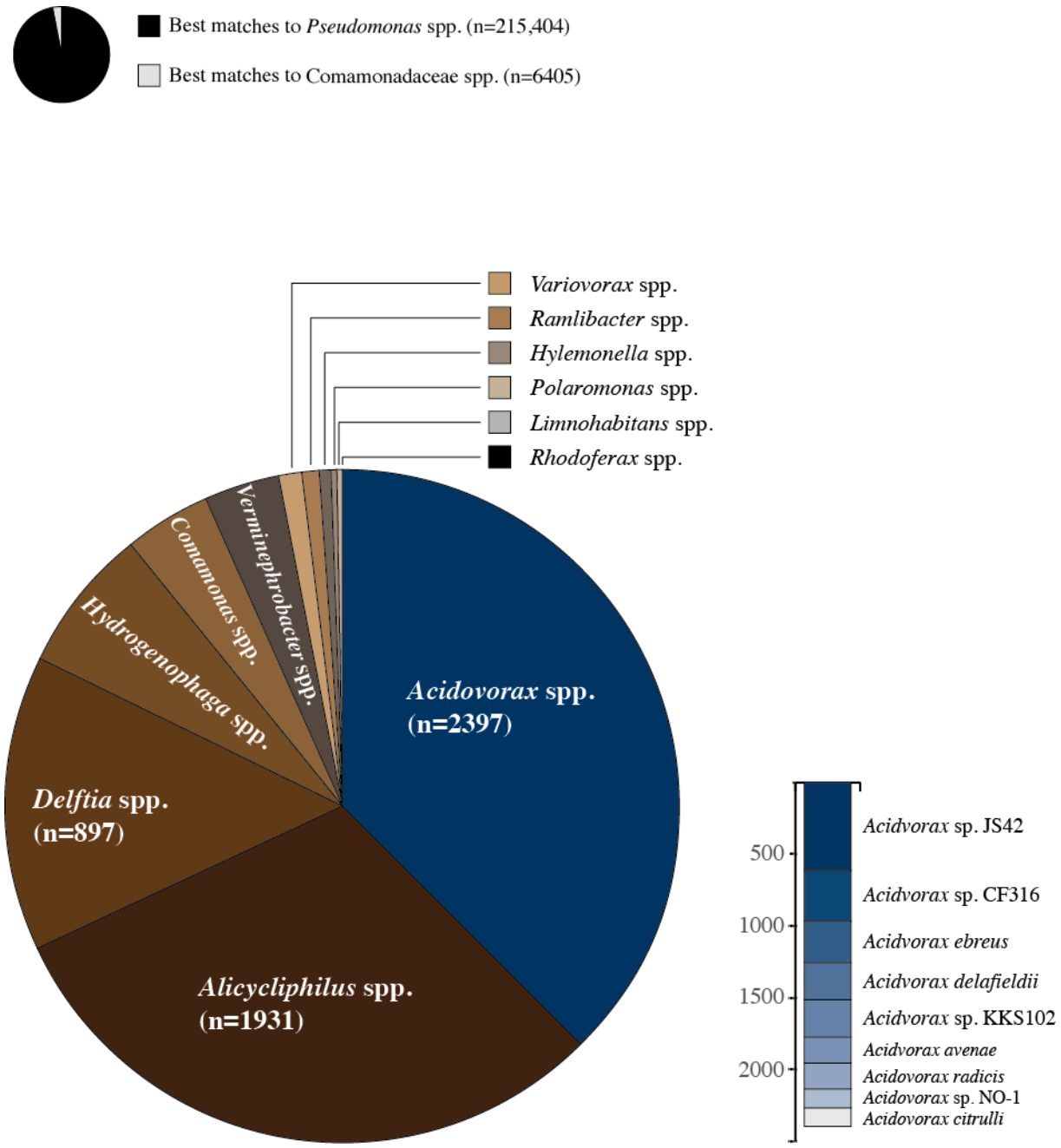


Figure 8. Identification of bacterial sequences within the *Xenopus tropicalis* genome trace read archive. Sequences were extracted with ReadMiner using best-matching (Comamonadaceae) and best-competing (Psuedomonad) clades; see text for details. **Top**: Distribution of 221,809 bacterial sequences among *Pseudomonas* spp. genomes and genomes from the Comamonadaceae. **Bottom**: The large pie chart illustrates the distribution of sequences among genera of Comamonadaceae; the stacked bar displays the distribution of 2397 sequences among the various species of *Acidovorax*.

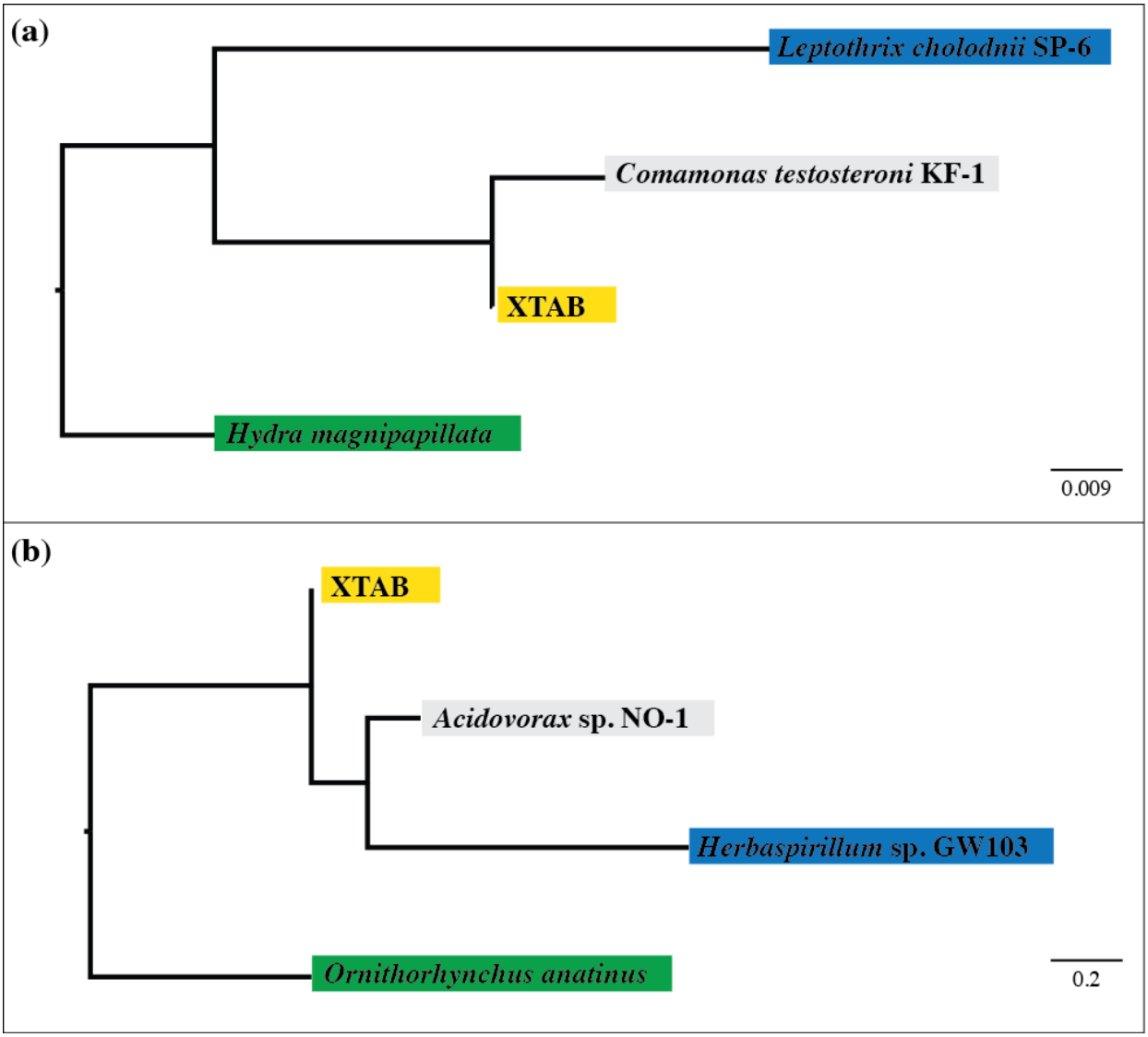


Figure 9. *N*-taxon statements for two mined XTAB core proteins. Each XTAB protein (yellow) was aligned with its best matches from non-*Xenopus* eukaryotes (green), non-Comamonadaceae bacteria (blue), and Comamonadaceae (gray) from the NCBI *nr* database. Trees were estimated as described in the text and rooted on the eukaryotic taxon. (**a**) An *n*-taxon statement evaluated as true; XTAB and its Comamonadaceae match are sibling branches in the tree. (**b**) An *n*-taxon statement evaluated as false; XTAB and its Comamonadaceae match are not siblings.

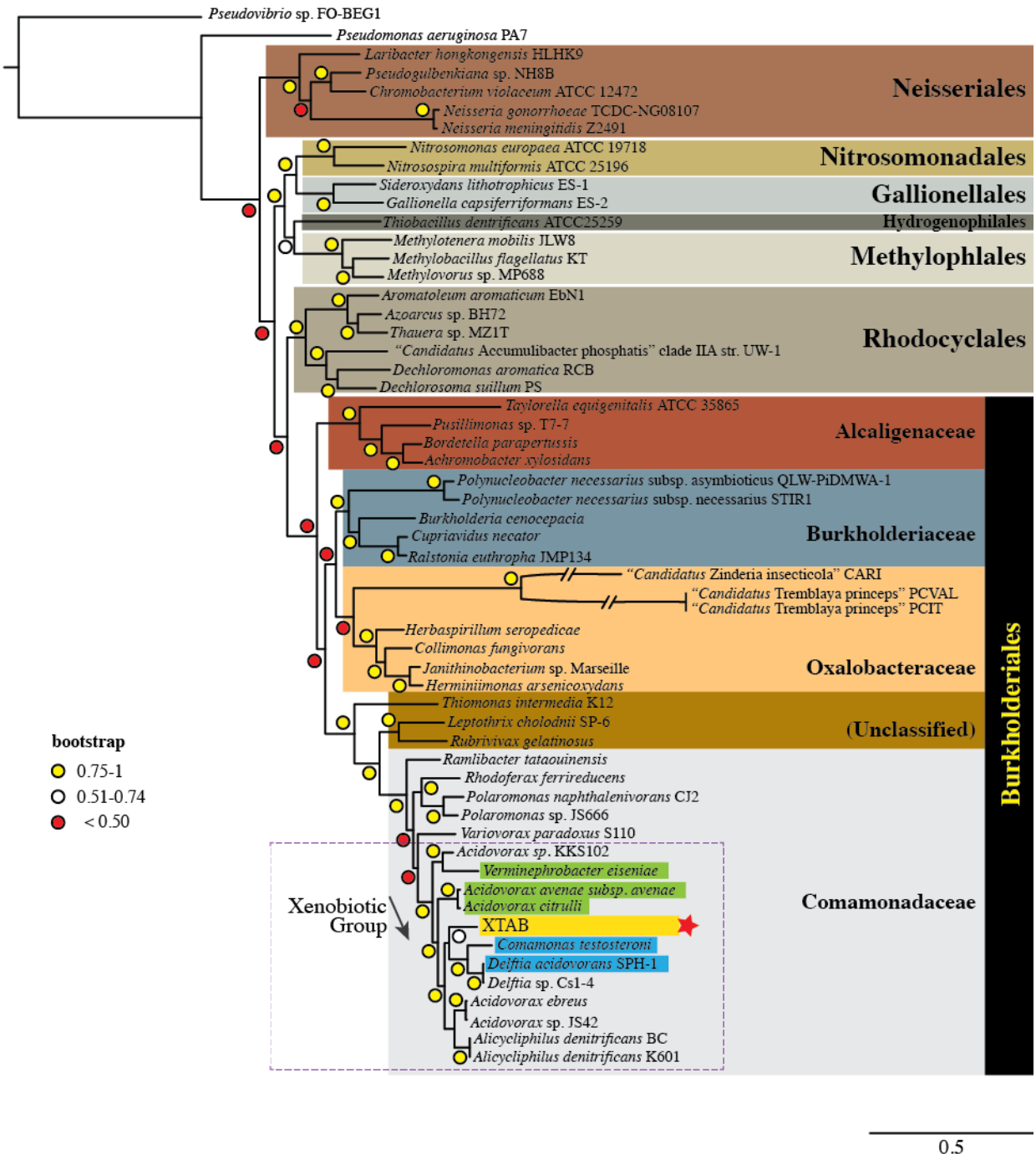


Figure 10. Genome-based phylogeny estimated for XTAB, 54 *Betaproteobacteria* taxa, and 2 outgroup taxa using `FastTree`. See text for alignment and tree-building details. The Xenobiotic Group (XG) of Comamonadaceae, which includes XTAB, is indicated by the dashed outline. Opportunistic human pathogens in the XG are highlighted in blue; extracellular symbionts in the XG are highlighted in green.

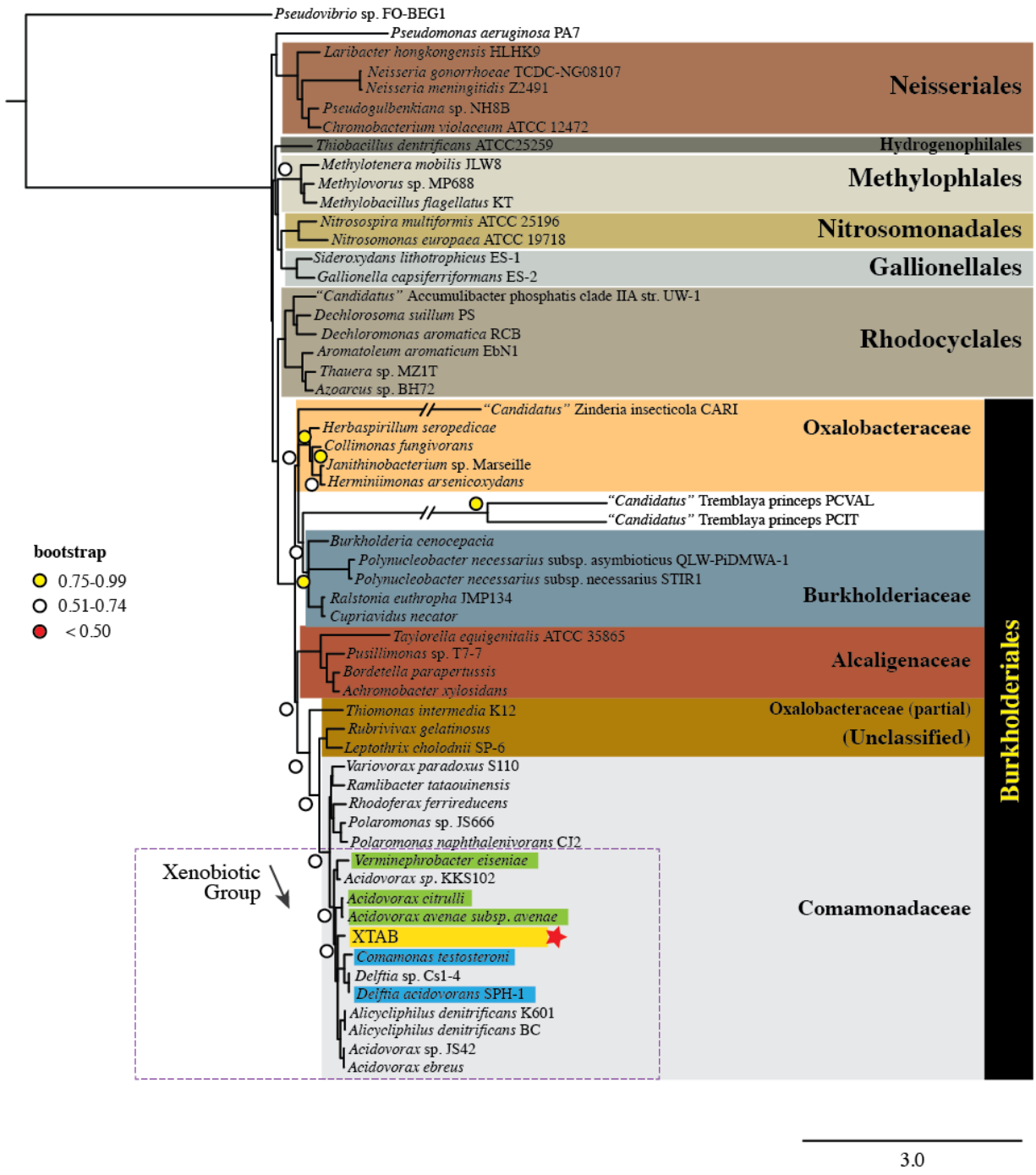


Figure 11. Genome-based phylogeny estimated for XTAB, 54 *Betaproteobacteria* taxa, and 2 outgroup taxa using `PhyloBayes`. See text for alignment and tree-building details. The Xenobiotic Group (XG) of Comamonadaceae, which includes XTAB, is indicated by the dashed outline. Opportunistic human pathogens in the XG are highlighted in blue; extracellular symbionts in the XG are highlighted in green. Bootstrap values of exactly 1.0 have been omitted for clarity.

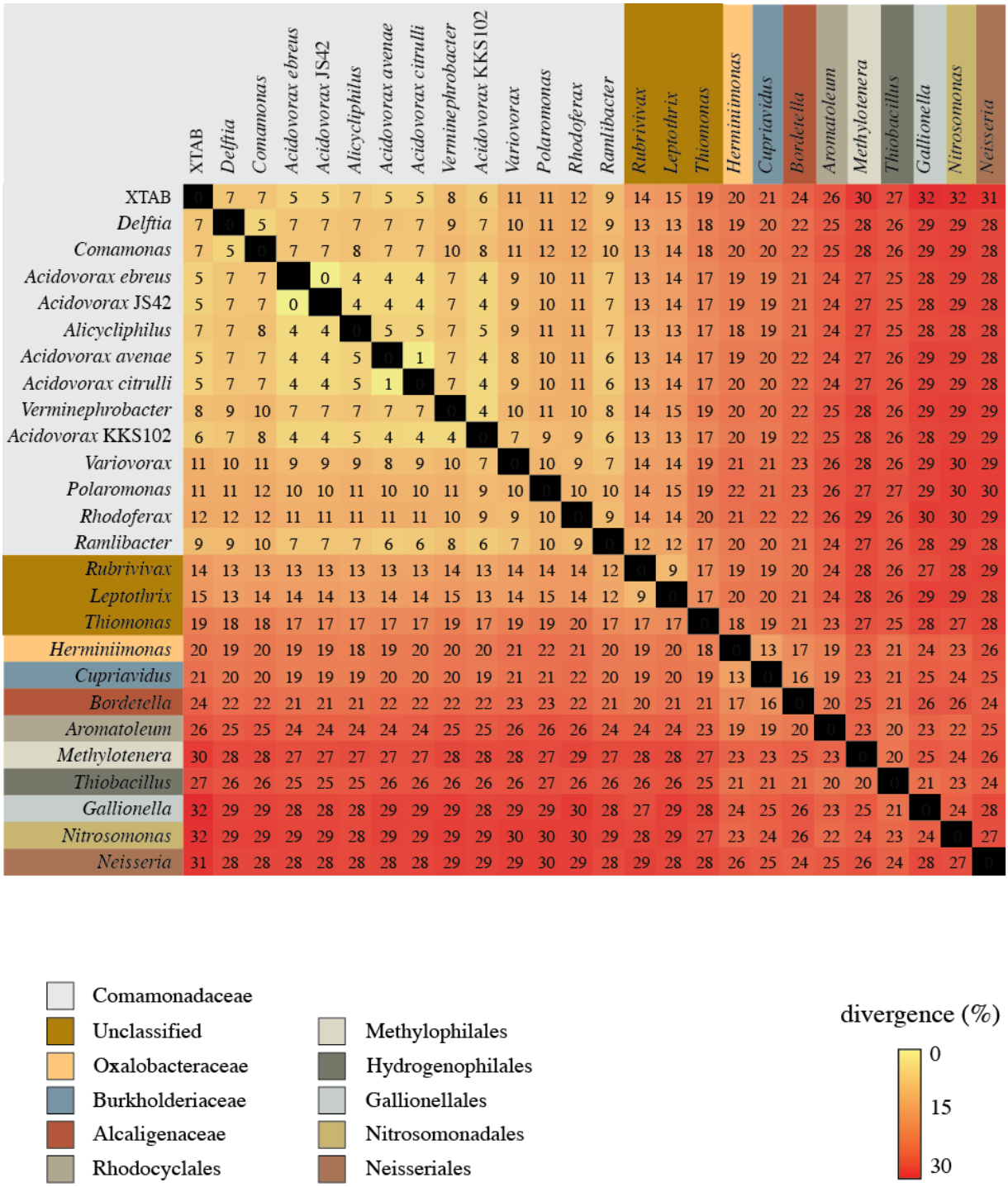


Figure 12. Comparison of sequence divergence across XTAB and *Betaproteobacteria* genera. A single species from each Comamonadaceae and Unclassified genus is shown here, plus a single representative genus from other Burkholderiales families and *Betaproteobacteria* orders. Divergence was calculated on 2004 amino acid sites of the core data set with DIVEIN, using the WAG substitution model; see text for details.

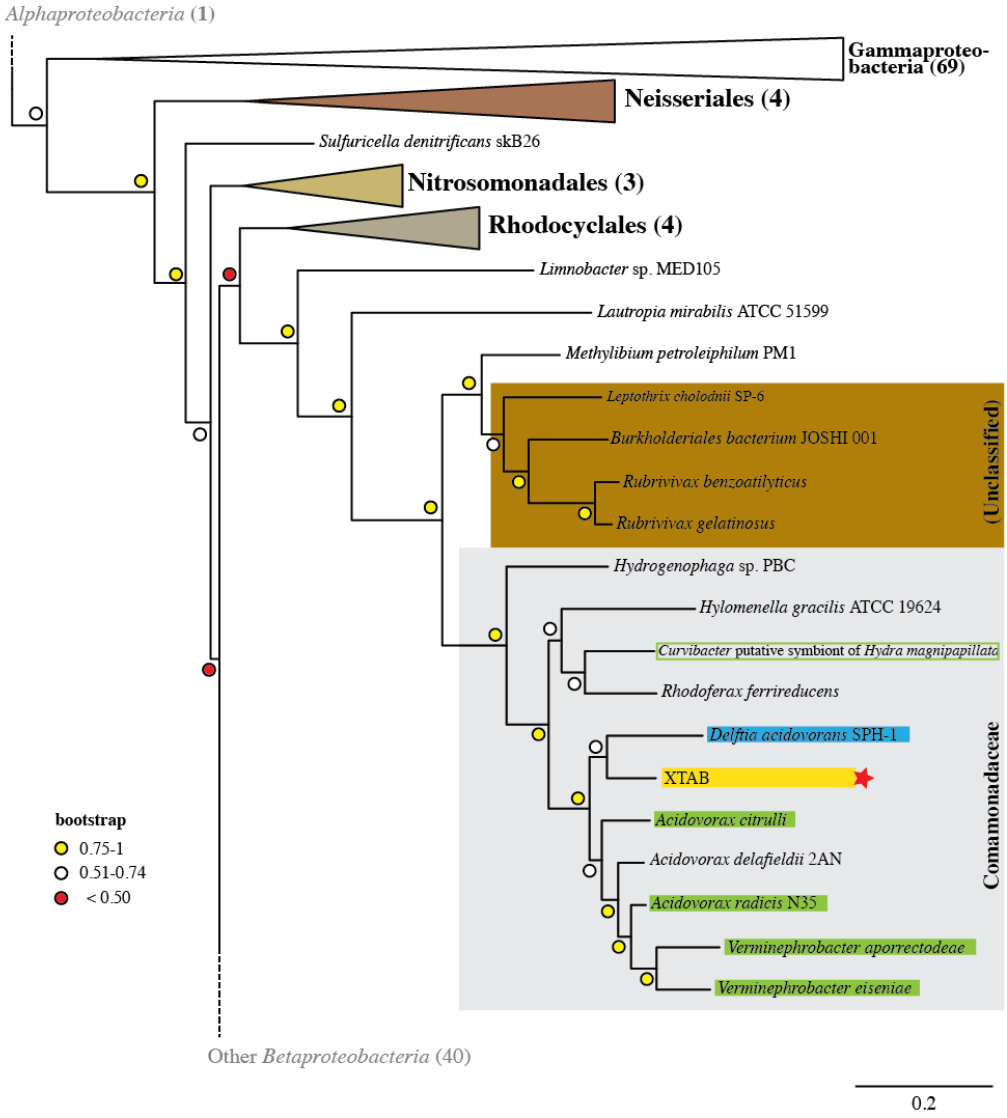


Figure 13. Phylogeny estimated (WAG model) for 11 flagellar proteins from NCBI's *nr* database shared among XTAB and 148 bacterial taxa. See text for alignment and tree-building details. Opportunistic human pathogens in the Comamonadaceae are highlighted in blue; extracellular symbionts are highlighted in green. Putative symbionts are indicated by a green box outline. Only the Comamonadaceae and neighboring branches are shown here, for clarity. The branch off the top of the visible tree includes an outgroup *Alphaproteobacteria* species; the branch off the bottom of the visible tree includes 40 additional *Betaproteobacteria* species.

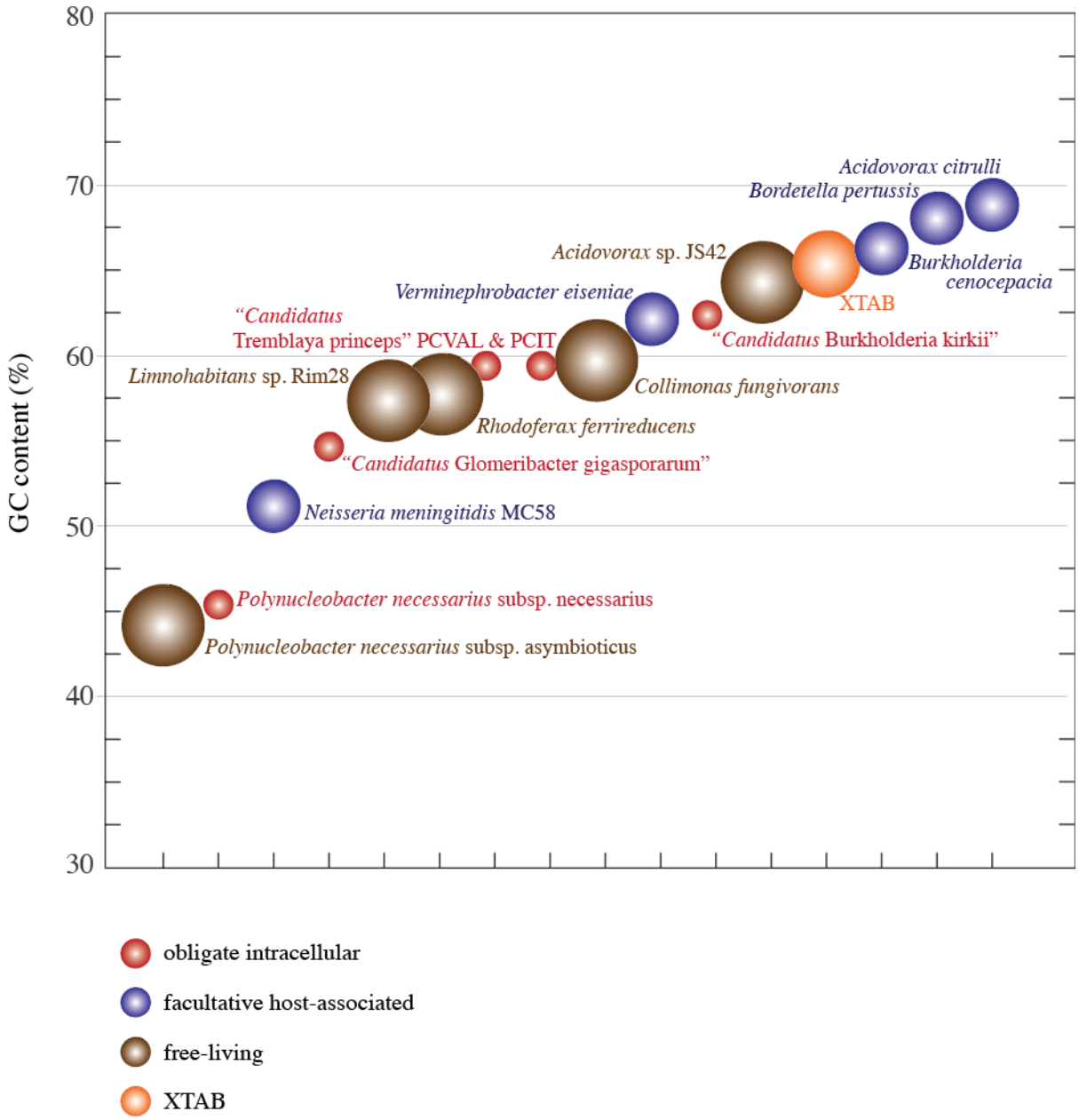


Figure 14. Base composition bias in XTAB compared to *Betaproteobacteria* from different environments. Percent GC (%GC) is plotted along the y axis for the genomes of five obligate intracellular (small, red), five host-associated (medium, blue), and five free-living (large, brown) *Betaproteobacteria*. Mean contig %GC values are used for WGS genomes that contain multiple contigs.

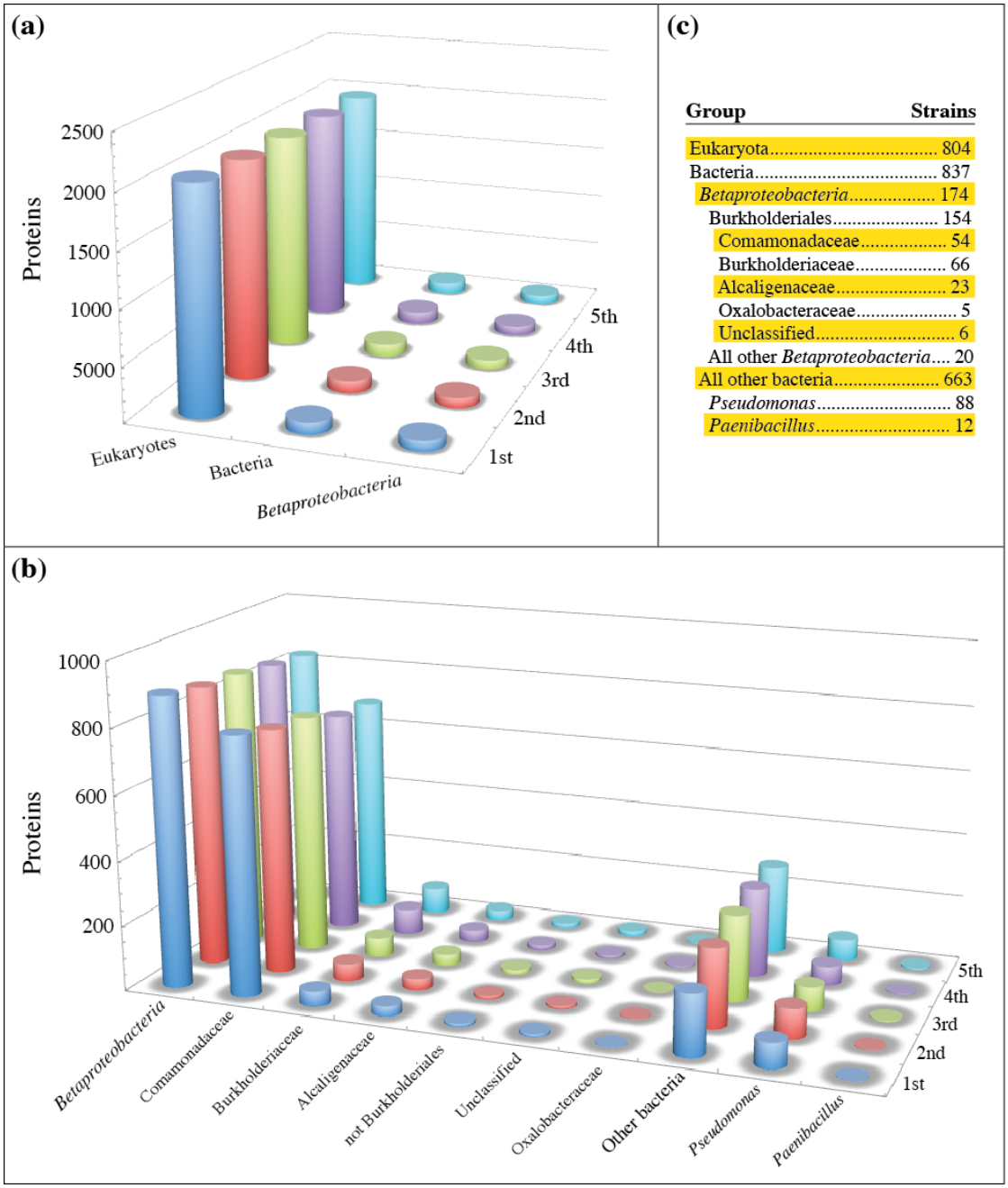


Figure 15. Classification of bacterial-like proteins from within the *Xenopus tropicalis* genome assembly. 21,662 *X. tropicalis* proteins with at least one match in the NCBI *nr* database (excluding the *X. tropicalis* group) were ranked by S_m score. (**a**) Taxonomic distribution of the top-5 scores for each *X. tropicalis* protein by major group. "Eukaryotes" totals are exclusive of *X. tropicalis*; "Bacteria" totals are exclusive of *Betaproteobacteria*. (**b**) Taxonomic distribution of the top-5 scores by major bacterial group. "*Betaproteobacteria*" totals include the next six groups along the x axis, and "Other bacteria" totals include *Pseudomonas* and *Paenibacillus* counts. (**c**) Distribution of major strain types among top-5 matches of all *X. tropicalis* proteins.

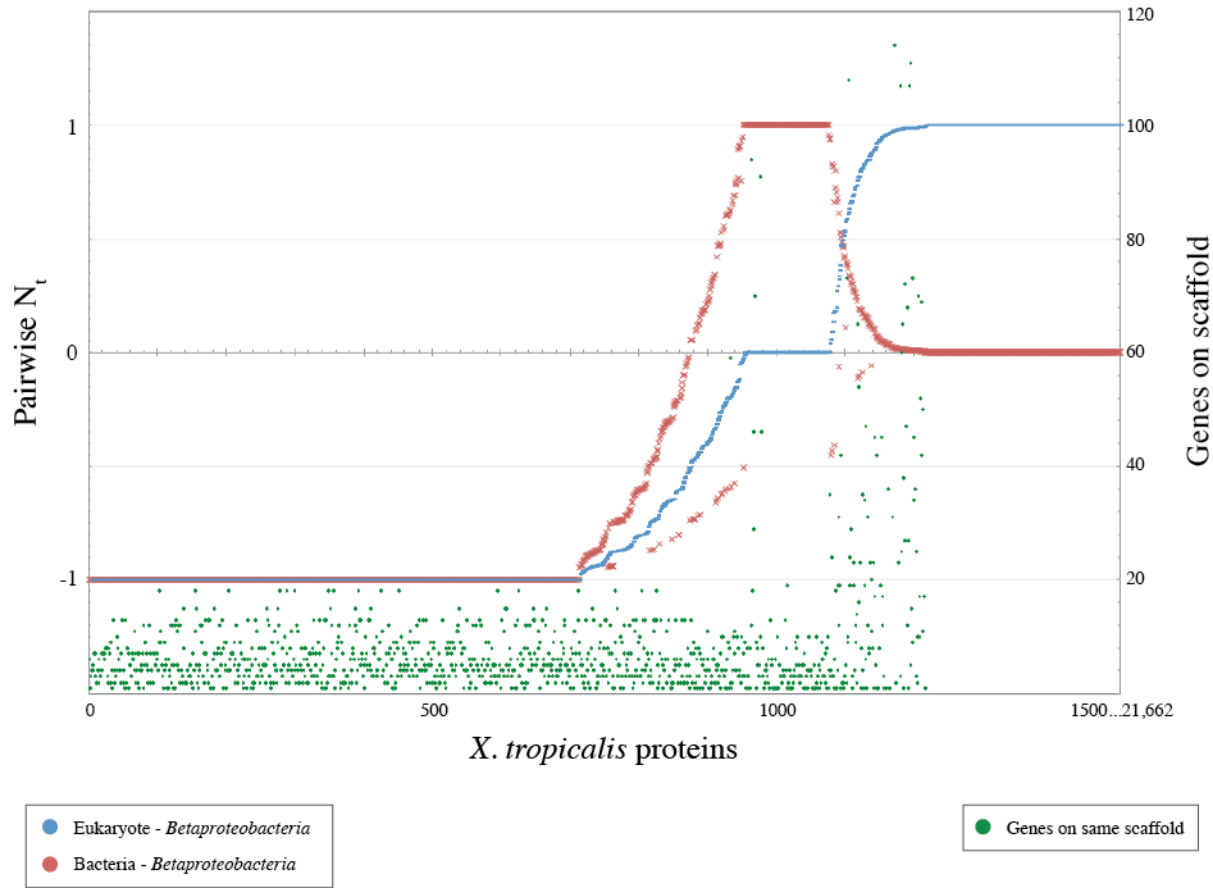


Figure 16. Identification of *Xenopus tropicalis* proteins with similarity to bacterial or betaproteo-bacterial proteins. Pairwise N_t scores for 21,662 *X. tropicalis* proteins with available S_m values are arranged along the x axis from smallest (-1) to largest (1) pairwise eukaryote-*Betaproteobacteria* N_t scores (blue circles). The first 1216 proteins had eukaryote-*Betaproteobacteria* scores less than 1 and non-zero bacteria-*Betaproteobacteria* scores (red circles); these proteins were chosen to pursue detailed scaffold analysis. The number of genes on the assembly scaffolds for these 1216 proteins (green circles) are plotted on the secondary y axis. the remaining 20,446 proteins had eukaryote-*Betaproteobacteria* scores of exactly 1 and bacteria-*Betaproteobacteria* scores of exactly 0; these were classified as fully eukaryotic. For clarity, only the first 284 fully eukaryotic proteins are shown in this figure.

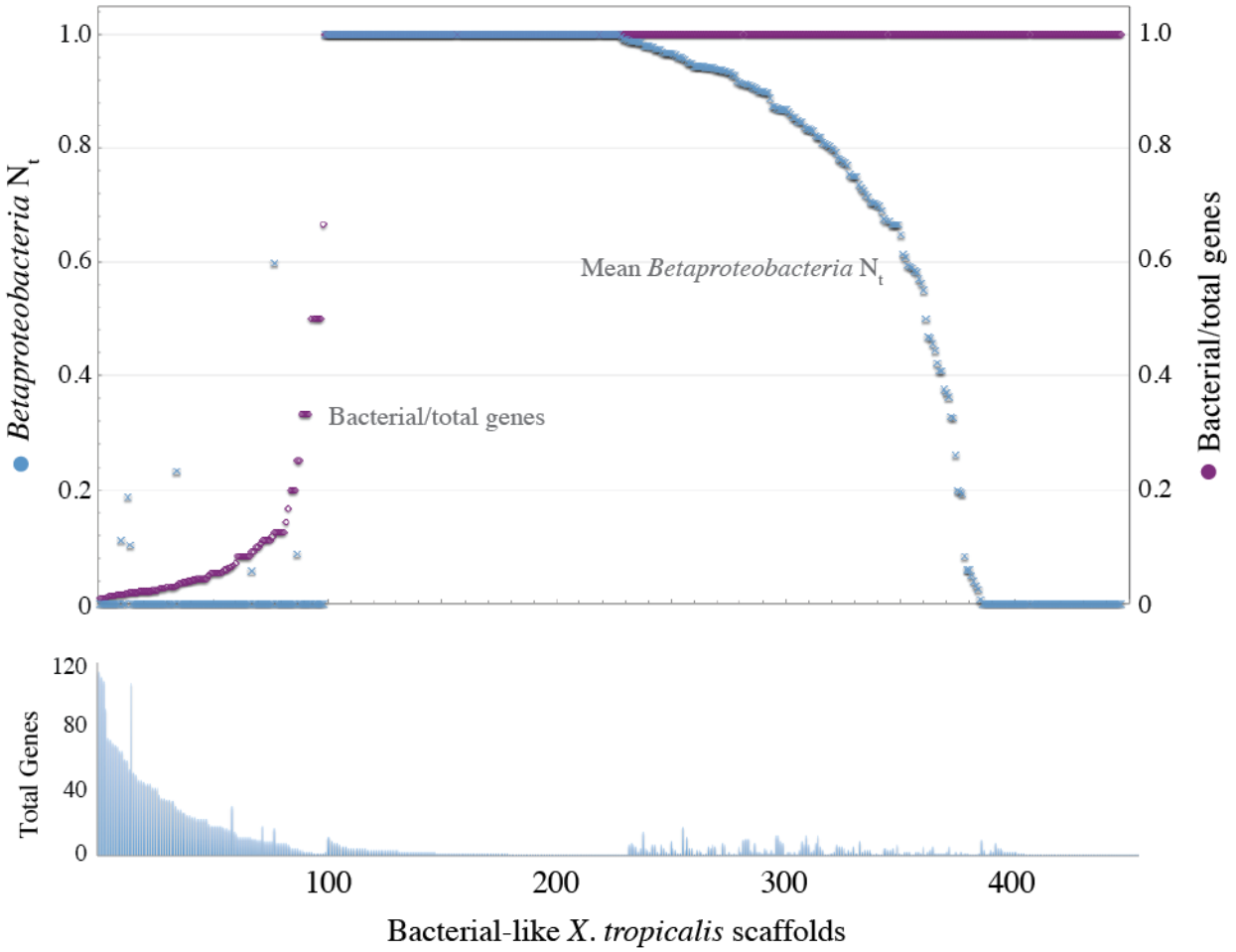


Figure 17. Analysis of genomic scaffolds from the *Xenopus tropicalis* assembly that contain bacterial-like genes. **Top**: 1216 *X. tropicalis* proteins with some bacterial-like characteristics were mapped to their genomic scaffolds. For each scaffold, the mean N_t score for all betaproteobacterial-like genes (blue crosses) was compared to the ratio of bacterial-like genes to total genes (purple circles) on that scaffold. Scaffolds are arranged along the x axis from smallest (0) to largest (1) ratio of bacterial to total genes. **Bottom**: Bar graph showing the total gene counts for each scaffold.

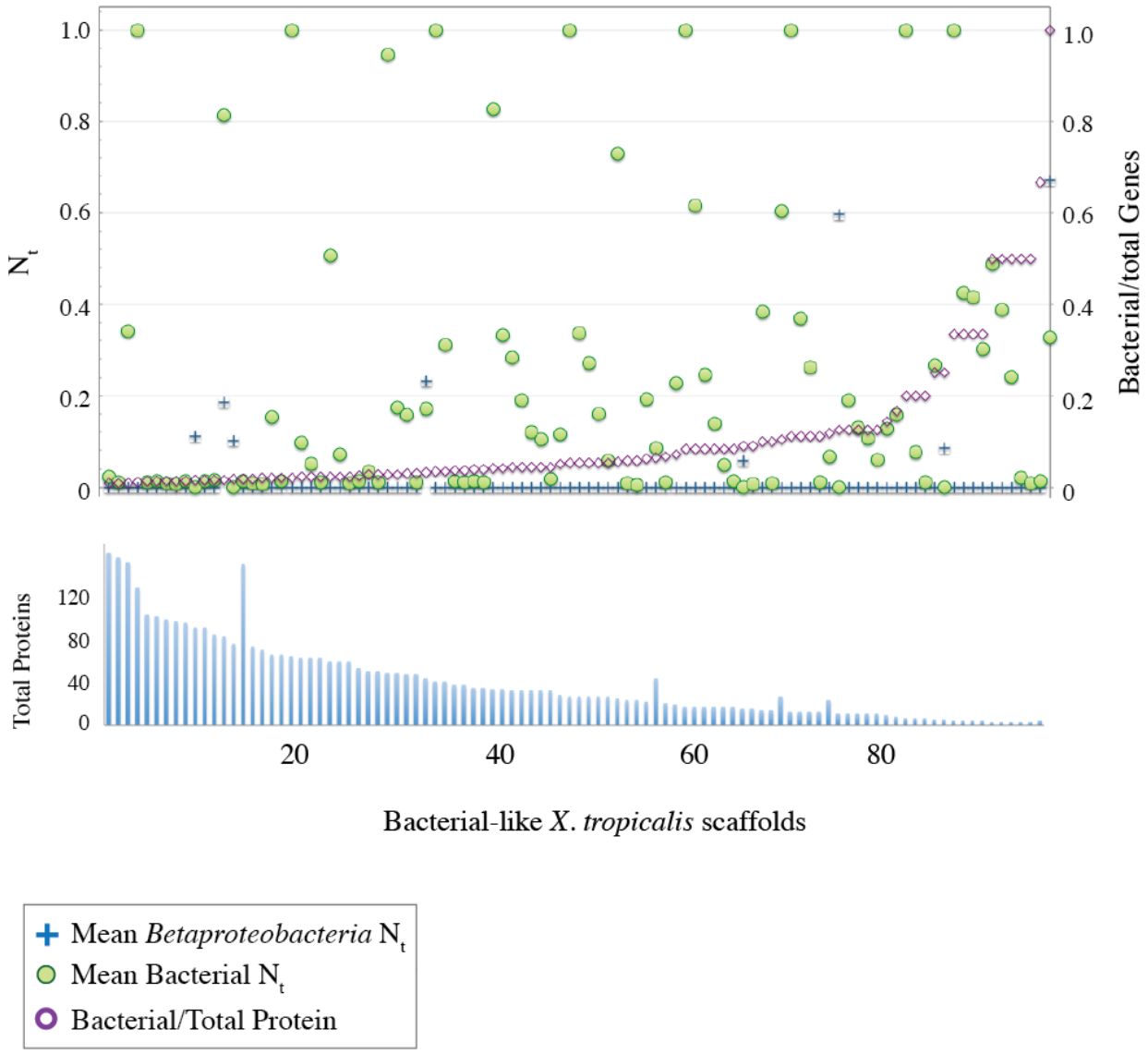


Figure 18. Identification of bacterial-like genes on primarily eukaryotic genomic scaffolds of the *Xenopus tropicalis* assembly. **Top**: 98 *X. tropicalis* scaffolds with a mixture of bacterial-like and eukaryotic genes are arranged along the x axis from smallest to largest ratio of bacterial to total genes (purple circles). Mean N_t scores for non-betaproteobacterial bacteria (green circles) and *Betaproteobacteria* (blue crosses) are plotted on the primary y axis. The bacteria to total gene ratio is plotted on the secondary y axis. Mean N_t scores are calculated using only the bacterial-like genes on a scaffold, as described in the text. **Bottom**: Bar graph showing the total gene counts for each scaffold.

TABLES

Table 1. Full-length 16S rDNA sequences used for pairwise gene divergence comparisons to betaproteobacterial and gammaproteobacterial 16S rDNA reads mined from the *Xenopus tropicalis* genome trace data. All sequences were obtained from genomes available in PATRIC.

Betaproteobacteria (33)

Comamonadaceae (21)

Acidovorax avenae subsp. avenae ATCC 19860
Acidovorax citrulli AAC00-1
Acidovorax ebreus TPSY
Acidovorax radicis N35
Acidovorax sp. CF316
Acidovorax sp. NO-1
Acidovorax sp. JS42
Acidovorax sp. KKS102
Alicycliphilus denitrificans BC
Alicycliphilus denitrificans K601
Comamonas testosteroni CNB-2
Delftia acidovorans SPH-1
Delftia sp. Cs1-4
Hydrogenophaga sp. PBC
Hylemonella gracilis ATCC 19624
Polaromonas naphthalenivorans CJ2
Polaromonas sp. JS666
Ramlibacter tataouinensis TTB310
Rhodoferax ferrireducens T118
Variovorax paradoxus EPS
Verminephrobacter eiseniae EF01-2

Unclassified (6)

Burkholderiales bacterium JOSHI_001
Leptothrix cholodnii SP-6
Leptothrix ochracea L12
Rubrivivax benzoatilyticus JA2
Rubrivivax gelatinosus CBS
Thiomonas intermedia K12

Burkholderiaceae (3)

Burkholderia cenocepacia J2315
Burkholderia mallei PRL-20
Burkholderia pseudomallei Pasteur 52237

Alcaligenaceae (2)

Bordetella pertussis Tohama I
Bordetella parapertussis 12822

Neisseriaceae (1)

Neisseria gonorrhoeae NCCP11945

Gammaproteobacteria (7)

Pseudomonas aeruginosa DK2
Pseudomonas entomophila L48
Pseudomonas fluorescens F113
Pseudomonas mendocina NK-01
Pseudomonas putida KT2440
Pseudomonas stutzeri DSM 4166
Pseudomonas syringae pv. syringae B728a

Firmicutes (5)

Paenibacillus mucilaginosus K02
Paenibacillus polymyxa
Paenibacillus sp. JDR-2
Paenibacillus sp. Y412MC10
Paenibacillus terrae HPL-003

Table 2. Search terms used in the cascading taxon sampling method for estimating phylogeny of the target 16S rDNA sequence mined from the *Xenopus tropicalis* genome trace data. Searches were run as Entrez queries against NCBI's *nt* database, excluding environmental samples and metagenomes.

Entrez Query String	Matches Kept
Comamonadaceae OR *Leptothrix* OR *Rubrivivax* OR *Thiomonas*	30
Oxalobacteraceae	10
Alcaligenaceae	10
Burkholderiaceae NOT *Burkholderia*	10
Burkholderia	5
Burkholderiales NOT (Comamonadaceae OR *Leptothrix* OR *Rubrivivax* OR *Thiomonas* OR Oxalobacteraceae OR Alcaligenaceae OR Burkholderiaceae)	5
Betaproteobacteria NOT Burkholderiales	5
Alphaproteobacteria	5
Gammaproteobacteria	5

Table 3. Genomes comprising the best-matching (Comamonadaceae) and best-competing (Pseudomonadaceae) clades for mining target reads from the Xenopus tropicalis genome trace data with ReadMiner. All genomes were obtained from PATRIC and plasmid sequences were removed before use.

Comamonadaceae (34)

Acidovorax avenae subsp. avenae ATCC 19860
Acidovorax avenae subsp. avenae RS-1
Acidovorax citrulli AAC00-1
Acidovorax delafieldii 2AN
Acidovorax ebreus TPSY
Acidovorax radicis N35
Acidovorax radicis N35v
Acidovorax sp. CF316
Acidovorax sp. NO-1
Acidovorax sp. JS42
Acidovorax sp. KKS102
Alicycliphilus denitrificans BC
Alicycliphilus denitrificans K601
Alicycliphilus sp. CRZ1
Comamonas testosteroni ATCC 11996
Comamonas testosteroni CNB-2
Comamonas testosteroni KF-1
Comamonas testosteroni S44
Delftia acidovorans SPH-1
Delftia sp. Cs1-4
Hydrogenophaga sp. PBC
Hylemonella gracilis ATCC 19624
Limnohabitans sp. Rim28
Limnohabitans sp. Rim47
Polaromonas naphthalenivorans CJ2
Polaromonas sp. CF318
Polaromonas sp. JS666
Ramlibacter tataouinensis TTB310
Rhodoferax ferrireducens T118
Variovorax paradoxus EPS
Variovorax paradoxus S110
Variovorax sp. CF313
Verminephrobacter aporrectodeae subsp. tuberculatae At4
Verminephrobacter eiseniae EF01-2

Pseudomonadaceae (181)

Pseudomonas aeruginosa (36 strains)
P. agarici NCPPB 2289
P. avellanae BPIC 631
P. brassicacearum subsp. brassicacearum NFM421
P. chlororaphis O6
P. chlororaphis subsp. aureofaciens 30-84
P. chlororaphis subsp. chlororaphis GP72
P. entomophila L48
P. extremaustralis 14-3 substr. 14-3b
P. fluorescens (19 strains)
P. fragi (2 strains)
P. fulva 12-X
P. fuscovaginae (2 strains)
P. gingeri NCPPB 3146
P. luteola XLDN4-9
P. mandelii JR-1
P. mendocina (3 strains)
P. monteilii QM
P. pseudoalcaligenes (2 strains)
P. psychrophila HA-4
P. psychrotolerans L19
P. putida (14 strains)
P. savastanoi pv. savastanoi NCPPB 3335
Pseudomonas sp. (32 strains)
P. stutzeri (10 strains)
P. synxantha BG33R
P. syringae Cit 7
P. syringae pv. aceris str. M302273
P. syringae pv. actinidiae (8 strains)
P. syringae pv. aesculi (3 strains)
P. syringae pv. aptata str. DSM 50252
P. syringae pv. avellanae (2 strains)
P. syringae pv. glycinea (3 strains)
P. syringae pv. japonica str. M301072
P. syringae pv. lachrymans (2 strains)
P. syringae pv. maculicola str. ES4326
P. syringae pv. mori str. 301020
P. syringae pv. morsprunorum str. M302280
P. syringae pv. oryzae str. 1_6
P. syringae pv. panici str. LMG 2367
P. syringae pv. phaseolicola (2 strains)
P. syringae pv. pisi str. 1704B
P. syringae pv. syringae (3 strains)
P. syringae pv. tabaci (3 strains)
P. syringae pv. theae NCPPB 2598
P. syringae pv. tomato (5 strains)
P. tolaasii (2 strains)

Table 4. 54 genomes used for ortholog group (OG) construction in conjunction with mined sequences from the *Xenopus tropicalis* genome trace data. All protein sequences were obtained from PATRIC.

***Betaproteobacteria* (54)**	
Achromobacter xylosoxidans A8	*Janthinobacterium* sp. Marseille
Acidovorax avenae subsp. avenae ATCC 19860	*Laribacter hongkongensis* HLHK9
Acidovorax citrulli AAC00-1	*Leptothrix cholodnii* SP-6
Acidovorax ebreus TPSY	*Methylobacillus flagellatus* KT
Acidovorax sp. JS42	*Methylotenera mobilis* JLW8
Acidovorax sp. KKS102	*Methylovorus* sp. MP688
Alicycliphilus denitrificans BC	*Neisseria gonorrhoeae* TCDC-NG08107
Alicycliphilus denitrificans K601	*Neisseria meningitidis* Z2491
Aromatoleum aromaticum EbN1	*Nitrosomonas europaea* ATCC 19718
Azoarcus sp. BH72	*Nitrosospira multiformis* ATCC 25196
Bordetella parapertussis 12822	*Polaromonas naphthalenivorans* CJ2
Burkholderia cenocepacia J2315	*Polaromonas* sp. JS666
"*Candidatus* Accumulibacter phosphatis" clade IIA str. UW-1	*Polynucleobacter necessarius* subsp. asymbioticus QLW-P1DMWA-1
"*Candidatus* Tremblaya princeps" PCIT	*Polynucleobacter necessarius* subsp. necessarius STIR1
"*Candidatus* Tremblaya princeps" PCVAL	*Pseudogulbenkiania* sp. NH8B
"*Candidatus* Zinderia insecticola" CARI	*Pusillimonas* sp. T7-7
Chromobacterium violaceum ATCC 12472	*Ralstonia eutropha* JMP134
Collimonas fungivorans Ter331	*Ramlibacter tataouinensis* TTB310
Comamonas testosteroni CNB-2	*Rhodoferax ferrireducens* T118
Cupriavidus necator N-1	*Rubrivivax gelatinosus* IL144
Dechloromonas aromatica RCB	*Sideroxydans lithotrophicus* ES-1
Dechlorosoma suillum PS	*Taylorella equigenitalis* ATCC 35865
Delftia acidovorans SPH-1	*Thauera* sp. MZ1T
Delftia sp. Cs1-4	*Thiobacillus denitrificans* ATCC 25259
Gallionella capsiferriformans ES-2	*Thiomonas intermedia* K12
Herbaspirillum seropedicae SmR1	*Variovorax paradoxus* S110
Herminiimonas arsenicoxydans	*Verminephrobacter eiseniae* EF01-2

Table 5. Eleven putative flagella proteins extracted from the initial set of mined XTAB sequences. Each XTAB protein is shown along with its best match in NCBI's *nr* database. XTAB annotations were transferred from each best match. %ID: percent identity of the match; %Q: percent of the query (XTAB) that aligned; %S: percent of the subject that aligned.

XTAB ID	Annotation		Best Match			
			Ref	%ID	%Q	%S
XTAB000079	FliG	flagellar C ring	*Acidovorax* sp. CF316			
			ZP_10389688.1	83.3	100	87.3
XTAB001397	FlhB	flagellar export apparatus	*Acidovorax* sp. CF316			
			ZP_10389096.1	71.1	97.9	61.7
XTAB000869	FlgC	flagellar proximal rod	*Delftia acidovorans* SPH-1			
			YP_001561663.1	96.2	100	98.5
XTAB001396	FlhA	flagellar export apparatus	*Acidovorax ebreus* TPSY			
			YP_002554526.1	90.6	99.6	39.9
XTAB000748	FliM	flagellar C ring	*Comamonas testosteroni* CNB-2			
			YP_003276507.1	87.1	100	35.0
XTAB000717	MotA	flagellar motor	*Alicycliphilus denitrificans* BC			
			YP_004128461.1	85.4	100	96.1
XTAB000343	FliC	flagellar filament	*Acidovorax delafieldii* 2AN			
			ZP_04761711.1	72.6	98.6	43.5
XTAB000716	MotB	flagellar motor	*Alicycliphilus denitrificans* BC			
			YP_004128462.1	91.6	100	23.0
XTAB001200	FlgL	flagellar hook-filament	*Acidovorax* sp. JS42			
			YP_988021.1	88.1	100	60.4
XTAB000517	FlgF	flagellar proximal rod	*Alicycliphilus denitrificans* BC			
			YP_004128477.1	83.3	100	100
XTAB000516	FlgE	flagellar hook	*Alicycliphilus denitrificans* BC			
			YP_004128476.1	88.6	100	20.6

Table 6. *Betaproteobacteria* genomes used in the evaluation of XTAB as a possible symbiont. For species that have a known association with a host organism, both the (primary) host and PubMed ID of a reference describing the association are provided.

Lifestyle	Genome Name	Host	PubMed
obligate intracellular 	"*Candidatus* Burkholderia kirkii" UZHbot1	*Psychotria kirkii*	12508863
	"*Candidatus* Glomeribacter gigasporarum" BEG34	*Gigaspora margarita*	21866182
	"*Candidatus* Tremblaya princeps" PCIT	*Planococcus citri*	21914892
	"*Candidatus* Tremblaya princeps" PCVAL	*Planococcus citri*	21914892
	Polynucleobacter necessarius subsp. necessarius STIR1	*Euplotes aediculatus*	13009932
facultative, host-associated 	*Acidovorax citrulli* AAC00-1	*Cucurbitaceae*	21554180
	Bordetella pertussis 12822	*Homo sapiens*	12010271
	Burkholderia cenocepacia J2315	*Homo sapiens*	19542002
	Neisseria meningitidis MC58	*Homo sapiens*	6126800
	Verminephrobacter eiseniae EF01-2	*Eisenia foetida*	18768621
free-living 	*Acidovorax* sp. JS42		
	Collimonas fungivorans Ter331		
	Limnohabitans sp. Rim28		
	Polynucleobacter necessarius subsp. asymbioticus QLW-P1DMWA-1		
	Rhodoferax ferrireducens T118		

Conclusions

CHAPTER 2

Chapter 2 presents the methodology used throughout this dissertation, employing sequence similarity and phylogeny to identify and characterize bacterial components within heterogeneous sequence data. This approach utilizes a multifarious approach that integrates the analysis of both pre- and post-assembly sequence data. Such an approach allows application of the underlying methodology in this dissertation to a broad variety of biological systems. In addition, the novel similarity metrics (S_m and N_t) developed here may prove generally useful in identifying microbial LGTs within metazoan genomes.

CHAPTER 3

Chapter 3 demonstrates the use of the genome sequence trace read analysis workflows (MetaMiner and ReadMiner) to identify and extract the genome for REIS from the sequencing data of its host, *Ixodes scapularis*. The results of this study show the effectiveness of MetaMiner at extracting enough information via 16S rDNA analysis to initially characterize bacterial residents of heterogeneous sequence read data. In addition, it was shown that mining read data using ReadMiner and a well-supported set of sequenced genomes allows the assembly of a significant fraction of a target bacterial genome with high precision and moderately high recall. Regions of bacteria-to-bacteria LGT, while potentially an issue if they originated from unrelated bacteria, were not as problematic as expected given the prevalence of MGEs in the REIS genome.

CHAPTER 4

Chapter 4 describes the identification and characterization of bacterial sequences from a rickettsial endosymbiont, from the genomic sequence data generated for a placozoan (*Trichoplax adhaerens*). This genomic profile of RETA potentially confirms the long suspected presence of a bacterial symbiont associated primarily with *T. adhaerens* fiber cells. Based on available genome sequences for Rickettsiales, all of which are either obligate intracellular symbionts or pathogens of various metazoan species, the data presented here likely depicts approximately 20% of the entire RETA genome. Despite the lack of a complete genome, phylogeny estimations and other analyses place RETA solidly within the "Midichloriaceae" clade. This rickettsial lineage is poorly understood, but based on the closest relative with an available genome sequence (*M. mitochondrii*), is quite different than the >80 genome sequences currently available for the well-studied species of the Anaplasmataceae and Rickettsiaceae. Thus, a better understanding of the “Midichloriaceae” clade of Rickettsiales is needed, particularly for 1) highlighting features involved in vertebrate pathogenicity in species of Anaplasmataceae and Rickettsiaceae, 2) determining the diversifying factors that define obligate intracellular life cycles of a wide range of eukaryotic hosts, and 3) deciphering the processes that shaped the transition of a rickettsial symbiont to an organelle (mitochondria) of eukaryotic cells.

CHAPTER 5

Chapter 5 combines both read mining workflows (MetaMiner and ReadMiner) to reveal the presence of multiple bacterial entities associated with the genome sequencing trace data of

Xenopus (Silurana) *tropicalis*, and to characterize a novel *Xenopus*-associated betaproteobacterium (XTAB). In addition, a significant bacterial presence heavily skewed toward the Comamonadaceae was also detected and isolated from the assembled *X. tropicalis* genome using AssemblySifter. Combined analyses of mined XTAB genes plus regions of sifted *X. tropicalis* scaffolds that were deemed most likely sites of bacterial LGTs revealed suggestions of, though no strong evidence to support, a lasting symbiotic relationship between XTAB and *X. tropicalis*. The results presented in this chapter demonstrate the combined effectiveness of the ReadMiner workflow at extracting and assembling sufficient genomic sequence for a single target to enable robust whole-genome phylogeny and functional analysis, even in the presence of large amounts of non-target bacterial sequences. They also demonstrate the ability of the AssemblySifter workflow to computationally sift the assembled host genome for bacterial genomic elements, and identify potential regions of microbial LGT into the host genome.

FUTURE WORK

A number of possible microbial symbionts remain to be investigated, and they will be analyzed using combinations of MetaMiner, ReadMiner, and AssemblySifter similar to the work demonstrated here. In addition, preliminary results from collaborators indicate the likely presence of rickettsial endosymbionts in other species of *Trichoplax*, and these will be evaluated and compared to RETA. Also possible is the targeted re-sampling of *T. adhaerens* to isolate and sequence RETA directly.

The decision to use Mira as the genome assembly software was made early in the development of ReadMiner; while it performs adequately, several options are available and should be explored. For example, Velvet is a popular assembly package for bacterial genomes and may provide higher quality assemblies of mined bacterial genomes.

Given the intriguing xenobiotic functionality of the bacteria related to XTAB, it may be illuminating to analyze secondary metabolic pathways represented among the mined XTAB genes. This can be done using the GO annotations already assigned, or by plotting XTAB genes onto pathway modules from other primary sources such as KEGG. In addition, the current work identified a manageable set of candidate LGTs in the *X. tropicalis* genome, which can be evaluated further using phylogeny, and possibly targeted for expression studies. In addition, the strongly betaproteobacterial sequences sifted from the X. tropicalis genome will be compared to the mined XTAB sequences to evaluate differences between the two sets.

The similarity measures for LGT finding that were developed as part of AssemblySifter extend what is available in the current literature, and may improve our ability to generally identify microbe-host LGTs. Another future goal of this work will be to apply these measures to eukaryotes with known LGTs (e.g., bdelloid rotifers) and compare with published methods to assess their performance.

APPENDICES

Appendix A. Custom software and external executables used in this project.

Table A1. Primary custom software created for this project. The code (including documentation) will be made available on the public software repository github (http://www.github.com/).

MetaMiner.sh	AssemblySifter_prep.sh
MetaMiner_search.pl	AssemblySifter.sh
MetaMiner_extract_reads_by_id.sh	AssemblySifter_search_remote.pl
MetaMiner_extract_reads_by_tax.sh	AssemblySifter_search_THREADS.pl
MetaMiner_msa_reads.sh	AssemblySifter_compile.sh
MetaMiner_raxml_reads.sh	AssemblySifter_process.sh
	AssemblySifter_tweak_eukann.pl
ReadMiner.sh	AssemblySifter_analyze_introns.pl
ReadMiner_call_genes.pl	AssemblySifter_hits2circos.pl
ReadMiner_goodhits.pl	AssemblySifter_gene2scaf.pl
ReadMiner_goodreads.pl	AssemblySifter_scaffold2gene.pl
	AssemblySifter_scaffold_summarize.pl

Table A2. Ancillary utilities created for this project. The code (including documentation) will be made available on the public software repository github (http://www.github.com/).

bin_by_col.pl	bin_generic.pl	bin_genes_by_scaffold.pl
blast_ann.pl	blast_cat.pl	blast_cfg.pl
blast_extract_best.pl	blast_extract_byID.pl	blast_filter.pl
blast_ggann.pl	blast_intersect.pl	blast_rm.pl
blast_score_Nt.pl	blast_sort_2D.pl	blast_sort.pl
blast_split.pl	blast_tab.pl	blast2circos.pl
blast2fasta.pl	calc_readlen.pl	calc_ttest_prob.pl
calc_dvg.pl	calc_gc.pl	calc_len.pl
calc_skew.pl	cat4db.sh	contig_blastn.sh
count_genes_on_scaffold.pl	count_introns.pl	extract_nohits.pl
extract_tax.pl	fa_1defline.pl	fa_addgroup.pl
fa_compare.pl	fa_count.pl	fa_extract_by_tax.pl
fa_extract_features.pl	fa_extract_id.pl	fa_extract_org.pl
fa_extract_seq.pl	fa_extract_subseq.pl	fa_filter_len.pl
fa_filter.pl	fa_findgaps.pl	fa_getcommon.pl
fa_grep.pl	fa_hdr2ref.pl	fa_headmap.pl
fa_headswap.pl	fa_lengths.pl	fa_randx.pl
fa_remap.pl	fa_rm_dupheads.pl	fa_rm_plasmid.pl
fa_slice.pl	fa_sort.pl	fa_splice.pl
fa_split.pl	fa_transform.pl	fa_verify.pl
fa2bc.pl	fa2fq.pl	fa2karyotype.pl
fa2seq.pl	fetch_patric_genomes.pl	faqual_dvec.sh
faqual_dvec_clean.pl	faqual_combine.sh	faqual2fq.pl
file_split.pl	fixgff.pl	fpack.sh
fq_dedup.pl	fq_extract.sh	fq_lfilter.pl
fq_qfilter.pl	fq_score.pl	fq2fa.pl
fq2faqual.pl	fqcat.pl	fqverify.pl
gbf2coords.sh	gbf2gff.sh	genbank2gff.pl
genome_pileup.pl	get_coords_from_acc.pl	get_exon_counts.pl
get_gene_record.pl	get_names_patric.pl	get_ncbi_tid.pl
get_orgdb_ncbi.sh	get_orgdb_patric.sh	get_protein_record.pl
get_ref_full.pl	get_refseq.pl	get_scaffold_from_gene.pl
get_scaffold_from_protein.pl	get_scaffold_record.pl	get_seqs_by_id.pl
get_seqs_by_org.pl	get_sra.pl	get_subseq.pl
get_table_rows.pl	get_tax_by_term.pl	get_tax_names.pl
get_tracedata.sh	gff2patric.pl	gg_org2tax.pl
fgb2circos.pl	fgb2genes.pl	fgb2gff.pl
glimmer2faa.pl	glimmer2fna.pl	glimmer2gff.pl
gmsizer.pl	make_wheel_pairwise.sh	make_wheel.sh
linkgen.sh	links.pl	mk_aliasdb.pl
mk_orx.pl	ntaxon_statements.sh	og_core.pl
og_getgenes.pl	og_intersect.pl	og_uniques.pl
org2table.pl	pairwise_align.pl	pairwise_table_reorder.pl
pick_random.pl	process_sra.pl	protid2fna.pl
protid2gff.pl	query_tracedb.pl	read_batch.pl
read_prep.sh	run_bwa.sh	run_fastq-dump.sh

sam2bam.sh	sra2fq.pl	srafetch.pl
srtabulate.pl	table_add_col.pl	table_append.pl
table_cat.pl	table_colswap.pl	table_extract_accs.pl
table_extract_col_mean.pl	table_extract_col.pl	table_extract_row.pl
table_filter.pl	table_killcols.pl	table_match_related.pl
table_merge_rows.pl	table_merge.pl	table_remap.pl
table_sort.pl	table_subdiv.pl	table_summarize.pl
table_tops.pl	table_transpose.pl	seqConverter.pl
tree_aresibs.pl	tree_idswap.pl	

Table A3. External executables and their versions used in this dissertation.

Software	Version
Blast+	2.2.25+
Blat	34 (stand alone)
BWA	0.6
Circos	0.62-1
EMBOSS	6.4.0
ENCprime	Downloaded March 28, 2013
fastqc	0.10.0
FastTree	(custom)
fastx	0.0.13
fgenesb	Downloaded March 26, 2013
GBlocks	0.91b
hmmbuild	3.0
hmmsearch	3.0
Krona	2.0
mcl	12-068
Mira	3.4
MUSCLE	3.8
PAUP*	4.0b10
Phrap/cross_match	0.990319
PhyloBayes-MPI	1.2e
RAxML	7.4.2
samtools	0.1.16
sra-toolkit	2.0.1

Appendix B. Trace read quality control results for *Ixodes scapularis* study (Chapter 3).

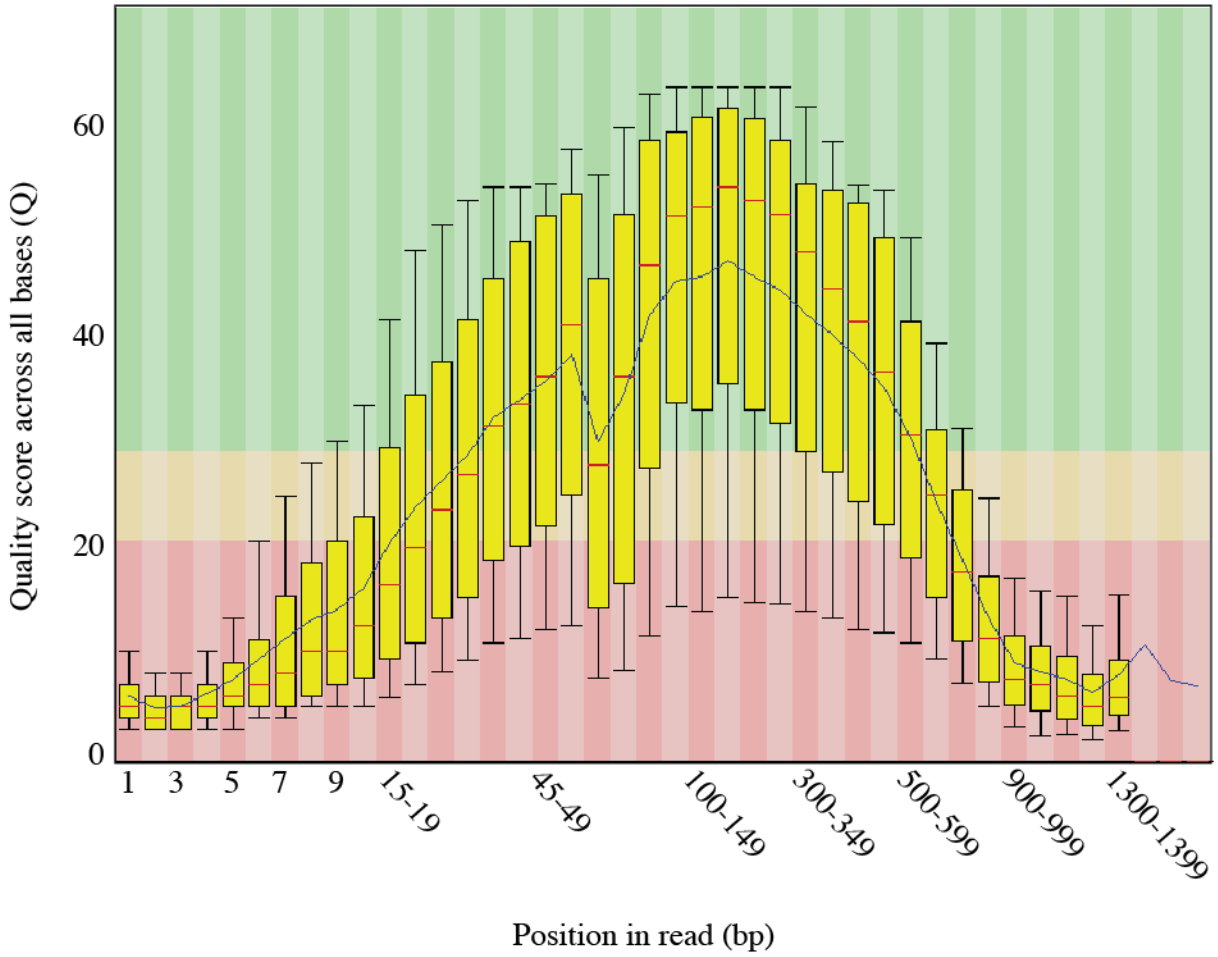


Figure B1. Quality of trace reads from the *Ixodes scapularis* genome project, after decontamination (removal of cloning vector) but before application of the quality control pipeline. This illustrates partial output from the `fastqc` program, showing the mean quality score (Q) across all reads at each position. The results plotted in this graph are used to trim low-quality positions (Q<25) from the ends of all reads.

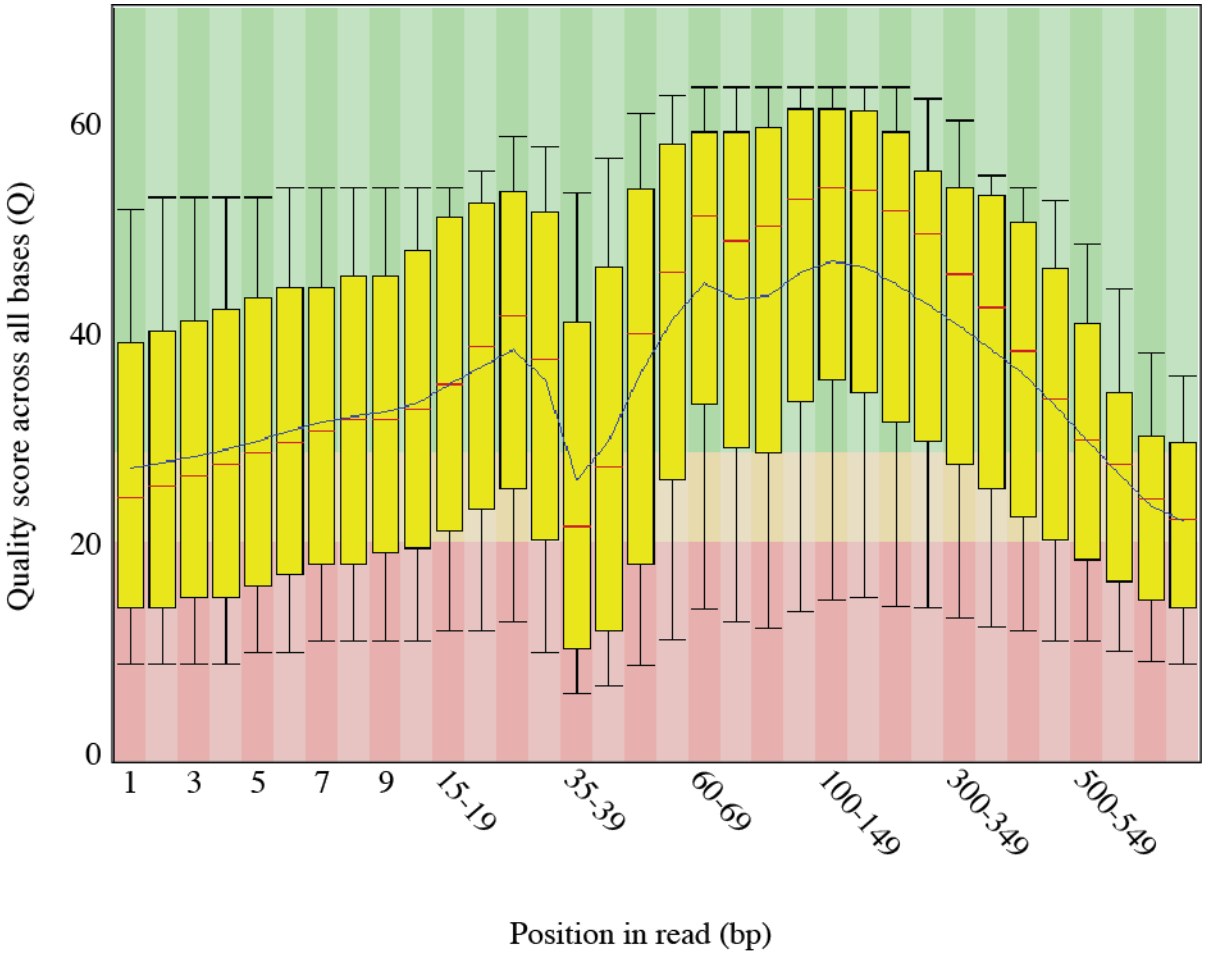
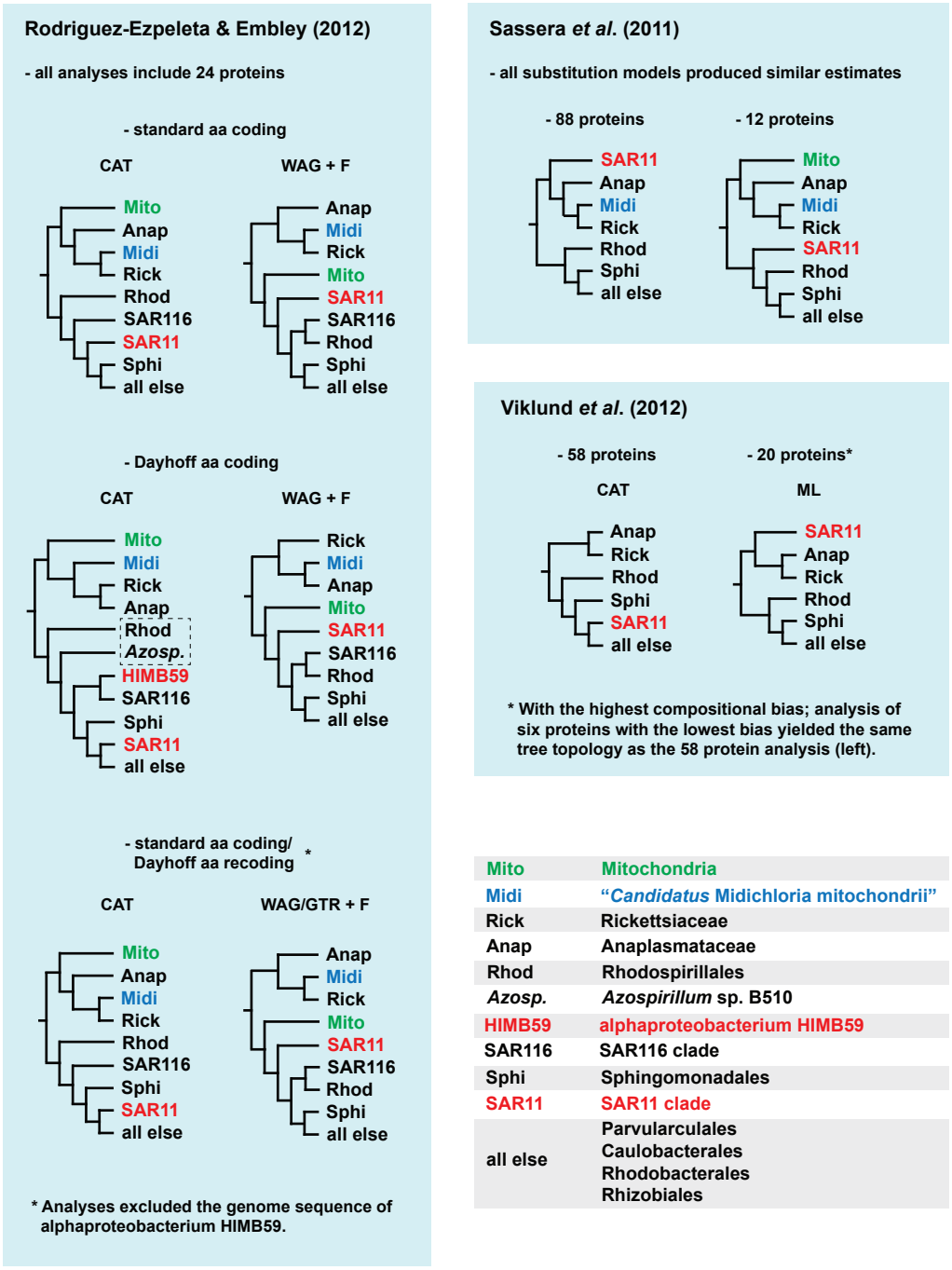


Figure B2. Quality of trace reads from the *Ixodes scapularis* genome project after application of the full quality control pipeline. This illustrates partial output from the `fastqc` program, showing the mean quality score (Q) across all reads at each position.

Appendix C. Supplemental information for Chapter 4 manuscript.



Mito	Mitochondria
Midi	"*Candidatus* Midichloria mitochondrii"
Rick	Rickettsiaceae
Anap	Anaplasmataceae
Rhod	Rhodospirillales
Azosp.	*Azospirillum* sp. B510
HIMB59	alphaproteobacterium HIMB59
SAR116	SAR116 clade
Sphi	Sphingomonadales
SAR11	SAR11 clade
all else	Parvularculales Caulobacterales Rhodobacterales Rhizobiales

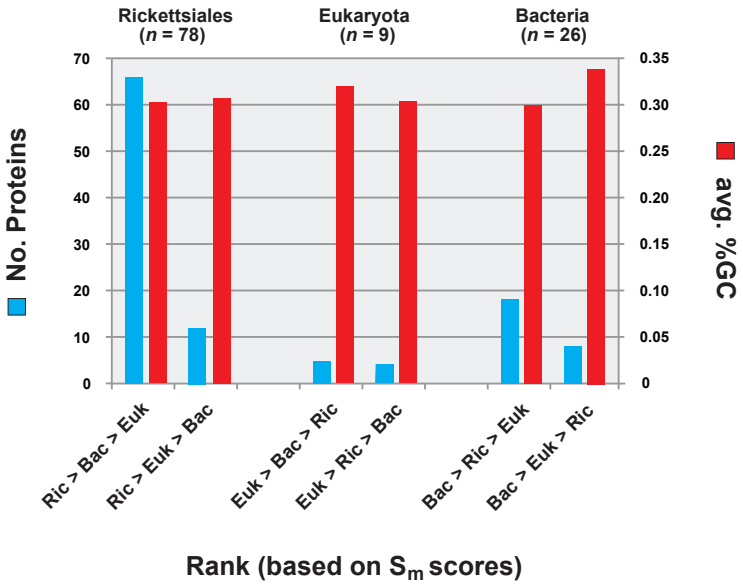
Supplementary Figure S1. Hypotheses of alphaproteobacterial phylogeny summarized from several recent studies. Trees are summarized from three studies **(Sassera, Lo et al. 2011; Rodriguez-Ezpeleta and Embley 2012; Viklund, Ettema et al. 2012)**, with the taxon abbreviations explained in the inset at bottom right.

Rodriguez-Ezpeleta, N. and T. M. Embley (2012). "The SAR11 Group of Alpha-Proteobacteria Is Not Related to the Origin of Mitochondria." PLoS One **7**(1): e30520.

Sassera, D., N. Lo, et al. (2011). "Phylogenomic evidence for the presence of a flagellum and cbb(3) oxidase in the free-living mitochondrial ancestor." Mol Biol Evol **28**(12): 3285-3296.

Viklund, J., T. J. Ettema, et al. (2012). "Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade." Mol Biol Evol **29**(2): 599-615.

a



b

RETA no.	CDS	Top hit (Rickettsiales)	S_m	Top hit (Eukaryota)	S_m	Top hit (Bacteria, sans Rickettsiales)	S_m
RETA0008	tilS	Orientia tsutsugamushi str. Ikeda	47.84	Schizosaccharomyces pombe 972h-	31.68	Candidatus Pelagibacter ubique HTCC1062	29.75
RETA0012	dapF	Rickettsia bellii RML369-C	115.05	Populus trichocarpa	30.87	Sphingobium sp. SYK-6	49.58
RETA0016	murA	Rickettsia bellii RML369-C	286.16	Bombus impatiens	183.37	Sphingobium sp. SYK-6	225.06
RETA0021	ssb	Rickettsia bellii RML369-C	92.22	Nematostella vectensis	34.81	Hirschia baltica ATCC 49814	86.00
RETA0023	htpG	Candidatus Midichloria mitochondrii IricVA	255.19	Xenopus (Silurana) tropicalis	133.88		
RETA0024	fabF	Candidatus Midichloria mitochondrii IricVA	312.82	Botryotinia fuckeliana B05.10	201.05	Parvibaculum lavamentivorans DS-1	279.60
RETA0025	uvrA	Rickettsia prowazekii str. Madrid E	28.90	Physcomitrella patens subsp. patens	10.98	Novosphingobium sp. PP1Y	26.80
RETA0026	lpd	Candidatus Midichloria mitochondrii IricVA	293.34	Paramecium tetraurelia strain d4-2	119.57	Micavibrio aeruginosavorus ARL-13	291.31
RETA0027	pdhC	Wolbachia endosymbiont of D. ananassae	28.81	Schistosoma mansoni	20.94	Psychromonas ingrahamii 37	13.63
RETA0030	argS	Wolbachia endosymbiont of D. melanogaster	94.07	Candida glabrata CBS 138	12.73	Magnetospirillum magneticum AMB-1	87.51
RETA0032	pccB	Candidatus Midichloria mitochondrii IricVA	568.45	Ixodes scapularis	504.50	Rhodospirillum centenum SW	477.61
RETA0034	GTA_TIM	Rickettsia bellii OSU 85-389	133.61	Tetrahymena thermophila	2.19	Roseobacter denitrificans OCh 114	38.20
RETA0035	hslU	Rickettsia felis URRWXCal2	144.99	Bombus impatiens	96.05	Rhodospirillum centenum SW	142.59
RETA0037	qcrA	Candidatus Midichloria mitochondrii IricVA	122.19	Phytophthora infestans T30-4	81.42	Magnetospirillum magneticum AMB-1	112.54
RETA0038	psd	Orientia tsutsugamushi str. Boryong	70.98	Theileria annulata	28.85	Rhodospirillum centenum SW	59.19
RETA0039	uvrB	Candidatus Midichloria mitochondrii IricVA	157.98	Hydra magnipapillata	64.81	Acetobacter pasteurianus IFO 3283-01	116.89
RETA0041	ccmC	Candidatus Midichloria mitochondrii IricVA	77.84	Marchantia polymorpha	56.92	Azospirillum sp. B510	62.02
RETA0042	tyrS	Candidatus Midichloria mitochondrii IricVA	93.94	Thalassiosira pseudonana CCMP1335	64.28	Magnetospirillum magneticum AMB-1	87.75
RETA0044	atpH	Candidatus Midichloria mitochondrii IricVA	25.53	Danio rerio	18.03	Deferribacter desulfuricans SSM1	22.16
RETA0045	atpA		NA				
RETA0049	Porin_4	Candidatus Midichloria mitochondrii IricVA	50.44	Cryptosporidium muris RN66	0.13	Azospirillum sp. B510	24.57
RETA0050	dsbA	Candidatus Midichloria mitochondrii IricVA	41.61	Bombyx mori	5.90	Ochrobactrum anthropi ATCC 49188	33.15
RETA0051a	tolC	Candidatus Midichloria mitochondrii IricVA	23.22	Arabidopsis thaliana	0.95	Sulfurihydrogenibium azorense Az-Fu1	14.76
RETA0051b	sppA	Ehrlichia ruminantium str. Gardel	42.79	Paulinella chromatophora	18.84	Brucella melitensis bv. 1 str. 16M	41.91
RETA0052a	rmuC	Candidatus Midichloria mitochondrii IricVA	96.98	Ostreococcus lucimarinus CCE9901	36.04	Sphingobium sp. SYK-6	72.63
RETA0052b	uvrD	NA	NA	NA	NA	NA	NA
RETA0053	rplB	Candidatus Midichloria mitochondrii IricVA	117.90	Ixodes scapularis	90.57	Granulibacter bethesdensis CGDNIH1	102.61
RETA0054	rpsS	Candidatus Midichloria mitochondrii IricVA	78.66	Hydra magnipapillata	51.24	Zymomonas mobilis subsp. mobilis ZM4	72.03
RETA0057	kdsB	Candidatus Midichloria mitochondrii IricVA	109.70	Phytophthora infestans T30-4	38.20	Rhodopseudomonas palustris CGA009	80.98
RETA0059	mnmG	Candidatus Midichloria mitochondrii IricVA	115.79	Aedes aegypti	75.00	Ruegeria sp. TM1040	89.50
RETA0060	parB	Candidatus Midichloria mitochondrii IricVA	101.60	Hydra magnipapillata	60.26	Thermosediminibacter oceani DSM 16646	98.33
RETA0063	rpoB	Orientia tsutsugamushi str. Boryong	37.67	Ixodes scapularis	24.63	Hyphomonas neptunium ATCC 15444	30.43
RETA0065	ispH	Candidatus Midichloria mitochondrii IricVA	117.58	Amphimedon queenslandica	75.40	Haemophilus parasuis SH0165	107.39
RETA0066	rlmE	Ehrlichia chaffeensis str. Sapulpa	28.32	Saccoglossus kowalevskii	16.21	Acidiphilium multivorum AIU301	13.45
RETA0067	qmk	Candidatus Midichloria mitochondrii IricVA	100.95	Xenopus (Silurana) tropicalis	33.73	Rhodobacter sphaeroides ATCC 17029	81.84
RETA0068	alaS	Wolbachia endosymbiont of D. willistoni	112.28	Arabidopsis lyrata subsp. lyrata	34.63	Zymomonas mobilis subsp. mobilis ZM4	90.01
RETA0069	mpg	Ehrlichia chaffeensis str. Arkansas	111.80	Trichinella spiralis	38.58	Candidatus Liberibacter solanacearum CLso-ZC1	92.19
RETA0070	dxr	Candidatus Midichloria mitochondrii IricVA	153.77	Bombus impatiens	107.14	Zymomonas mobilis subsp. pomaceae ATCC 29192	143.69
RETA0072	rpsP	Rickettsia canadensis str. McKiel	59.86	Arabidopsis lyrata subsp. lyrata	49.72	Azospirillum sp. B510	52.27
RETA0077	murG	Candidatus Midichloria mitochondrii IricVA	93.02	Ricinus communis	49.66	Dinoroseobacter shibae DFL 12	77.82
RETA0081	clpP	Wolbachia endosymbiont of D. ananassae	177.08	Ixodes scapularis	119.22	Haemophilus parasuis SH0165	134.86
RETA0082	tig	Candidatus Midichloria mitochondrii IricVA	131.33	Xenopus (Silurana) tropicalis	32.06	Brevundimonas subvibrioides ATCC 15264	91.81
RETA0083	tqt	Rickettsia bellii RML369-C	164.94	Ixodes scapularis	130.32	Zymomonas mobilis subsp. mobilis ZM4	142.47
RETA0084	atpB	Candidatus Midichloria mitochondrii IricVA	147.91	Malawimonas jakobiformis	81.99	Candidatus Pelagibacter sp. IMCC9063	114.99
RETA0085	atpE	Rickettsia bellii RML369-C	61.05	Saprolegnia ferax	35.87	Tistrella mobilis KA081020-065	46.16
RETA0087	murD	Candidatus Midichloria mitochondrii IricVA	158.39	Paulinella chromatophora	55.92	Hirschia baltica ATCC 49814	124.16
RETA0089	recJ	Wolbachia endosymbiont of D. simulans	65.69	Ixodes scapularis	53.29		
RETA0090	msbA	Candidatus Midichloria mitochondrii IricVA	56.45	Drosophila persimilis	11.31	Magnetospirillum magneticum AMB-1	34.75
RETA0093	ftsI	Wolbachia endosymbiont of D. simulans	43.90	Physcomitrella patens subsp. patens	12.19	Azospirillum sp. B510	28.88
RETA0096	mdh	Candidatus Midichloria mitochondrii IricVA	234.98	Ixodes scapularis	176.85	Roseobacter litoralis Och 149	225.15

b (continued)

RETA no.	CDS	Top hit (Rickettsiales)	S_m	Top hit (Eukaryota)	S_m	Top hit (Bacteria, sans Rickettsiales)	S_m
RETA0099	ppa	Candidatus Midichloria mitochondrii IricVA	158.37	Caenorhabditis remanei	99.10	Parvibaculum lavamentivorans DS-1	124.06
RETA0101	rplF	Candidatus Midichloria mitochondrii IricVA	56.24	Bombus impatiens	39.66	Rhodospirillum centenum SW	55.22
RETA0102	rpsE	Candidatus Midichloria mitochondrii IricVA	118.04	Amphimedon queenslandica	68.60	Pelagibacterium halotolerans B2	95.74
RETA0104	sucA	Candidatus Midichloria mitochondrii IricVA	42.17	Hydra magnipapillata	28.26	Starkeya novella DSM 506	42.13
RETA0105	hemF	Rickettsia felis URRWXCal2	210.88	Cryptococcus gattii WM276	90.69	Micavibrio aeruginosavorus ARL-13	206.50
RETA0106	map	Candidatus Midichloria mitochondrii IricVA	240.69	Bombus impatiens	137.58	Micavibrio aeruginosavorus ARL-13	204.05
RETA0112	proS	Candidatus Midichloria mitochondrii IricVA	185.54	Ixodes scapularis	176.11	Azorhizobium caulinodans ORS 571	178.48
RETA0113	nusA	Candidatus Midichloria mitochondrii IricVA	147.21	Paulinella chromatophora	24.77	Rhodospirillum centenum SW	114.44
RETA0114	cysS	Candidatus Midichloria mitochondrii IricVA	70.41	Arabidopsis thaliana	45.62		
RETA0116	rimN	Ehrlichia ruminantium str. Gardel	42.07	Tribolium castaneum	24.24	Elusimicrobium minutum Pei191	32.43
RETA0118	terC	Candidatus Midichloria mitochondrii IricVA	44.87	Physcomitrella patens subsp. patens	30.86	Calditerrivibrio nitroreducens DSM 19672	44.36
RETA0119	rpsJ	Candidatus Midichloria mitochondrii IricVA	96.13	Sordaria macrospora k-hell	72.00	Gluconacetobacter diazotrophicus PAl 5	88.56
RETA0120	rplC	Rickettsia canadensis str. McKiel	125.75	Micromonas pusilla CCMP1545	88.33	Ruegeria pomeroyi DSS-3	123.49
RETA0122	lepA	Candidatus Midichloria mitochondrii IricVA	307.95	Ixodes scapularis	250.73	Azospirillum sp. B510	259.35
RETA0125	lysS	Candidatus Midichloria mitochondrii IricVA	183.60	Aedes aegypti	1.88	Rhodopseudomonas palustris BisA53	148.47
RETA0128	quaA	Wolbachia endosymbiont of D. willistoni	210.51	Ricinus communis	207.58	Thermodesulfovibrio yellowstonii DSM 11347	178.27
RETA0131	tolB	Candidatus Midichloria mitochondrii IricVA	104.17	Xenopus (Silurana) tropicalis	16.78	Candidatus Pelagibacter ubique HTCC1062	88.08
RETA0133	pccA	Candidatus Midichloria mitochondrii IricVA	18.42	Ixodes scapularis	17.88	Azospirillum sp. B510	12.42
RETA0136	ispG	Candidatus Midichloria mitochondrii IricVA	42.27	Sordaria macrospora k-hell	3.56	Thioalkalivibrio sulfidophilus HL-EbGr7	37.21
RETA0140	grpE	Candidatus Midichloria mitochondrii IricVA	74.38	Ixodes scapularis	35.50	Yersinia enterocolitica subsp. palearctica 105.5R(r)	43.13
RETA0142	pheS	Candidatus Midichloria mitochondrii IricVA	116.20	Xenopus (Silurana) tropicalis	79.03	Azospirillum sp. B510	98.12
RETA0147	COG2945	Candidatus Midichloria mitochondrii IricVA	175.25	Dictyostelium purpureum	15.53	Rhodospirillum centenum SW	133.96
RETA0150	fumC	Wolbachia endosymbiont of D. simulans	82.47	Candida albicans SC5314	69.75	Micavibrio aeruginosavorus ARL-13	78.21
RETA0151	gyrB	Wolbachia endosymbiont of D. ananassae	80.75	Hydra magnipapillata	64.64	Acetobacter pasteurianus IFO 3283-01	62.89
RETA0157	GTA_N	Wolbachia endosymbiont of C. quinquefasciatus JHB	37.33	Naegleria gruberi	1.16	Caulobacter crescentus NA1000	14.77
RETA0158a	kdsA	Candidatus Midichloria mitochondrii IricVA	98.75	Bombus impatiens	54.51	Candidatus Puniceispirillum marinum IMCC1322	84.43
RETA0158b	smpB	NA	NA	NA	NA	NA	NA
RETA0159	glnA	Rickettsia canadensis str. McKiel	51.66	Nematostella vectensis	17.94	Acinetobacter sp. DR1	15.64
RETA0165	sucC	Candidatus Midichloria mitochondrii IricVA	168.07	Hydra magnipapillata	126.51	Acidiphilium multivorum AIU301	167.61
RETA0166	sucD	Wolbachia endosymbiont of D. willistoni	28.16	Nomascus leucogenys	20.31	Cellvibrio japonicus Ueda107	21.23
RETA0168	nrdA	Candidatus Midichloria mitochondrii IricVA	147.96	Hydra magnipapillata	141.73	Micavibrio aeruginosavorus ARL-13	127.10

c

RETA no.	CDS	Top hit (Rickettsiales)	S_m	Top hit (Eukaryota)	S_m	Top hit (Bacteria, sans Rickettsiales)	S_m
RETA0017	prfA	Orientia tsutsugamushi str. Ikeda	128.63	Nematostella vectensis ●	182.29	Bartonella henselae str. Houston-1	129.51
RETA0048	COG1485	Rickettsia bellii RML369-C	23.47	Strongylocentrotus purpuratus	37.83	Hyphomicrobium sp. MC1	25.08
RETA0058	obgE	Rickettsia peacockii str. Rustic	214.12	Ixodes scapularis ●	214.62	Sphingomonas wittichii RW1	173.53
RETA0095	gltA	Candidatus Midichloria mitochondrii IricVA	32.58	Sordaria macrospora k-hell ●	54.86	Burkholderia rhizoxinica HKI 454	32.85
RETA0110	COG2902	Rickettsia endosymbiont of Ixodes scapularis	124.86	Ixodes scapularis ●	147.94	Polaromonas naphthalenivorans CJ2	75.93
RETA0117	dnaK	Rickettsia slovaca 13-B	22.24	Sordaria macrospora k-hell ●	35.41	Bartonella quintana str. Toulouse	17.95
RETA0137	secA_N	Candidatus Midichloria mitochondrii IricVA	48.78	Chlamydomonas reinhardtii	86.33	Micavibrio aeruginosavorus ARL-13	45.02
RETA0029	secA_C	NA	NA	NA	NA	NA	NA
RETA0138	clpX	Candidatus Odyssella thessalonicensis L13	105.66	Xenopus (Silurana) tropicalis ●	126.71	Candidatus Pelagibacter ubique HTCC1062	117.83
RETA0143	adk	Candidatus Midichloria mitochondrii IricVA	29.99	Schistosoma mansoni ●	38.77	Taylorella asinigenitalis MCE3	37.93

● = nuclear encoded mitochondrial protein
● = bacterial gene in eukaryotic assembly

d

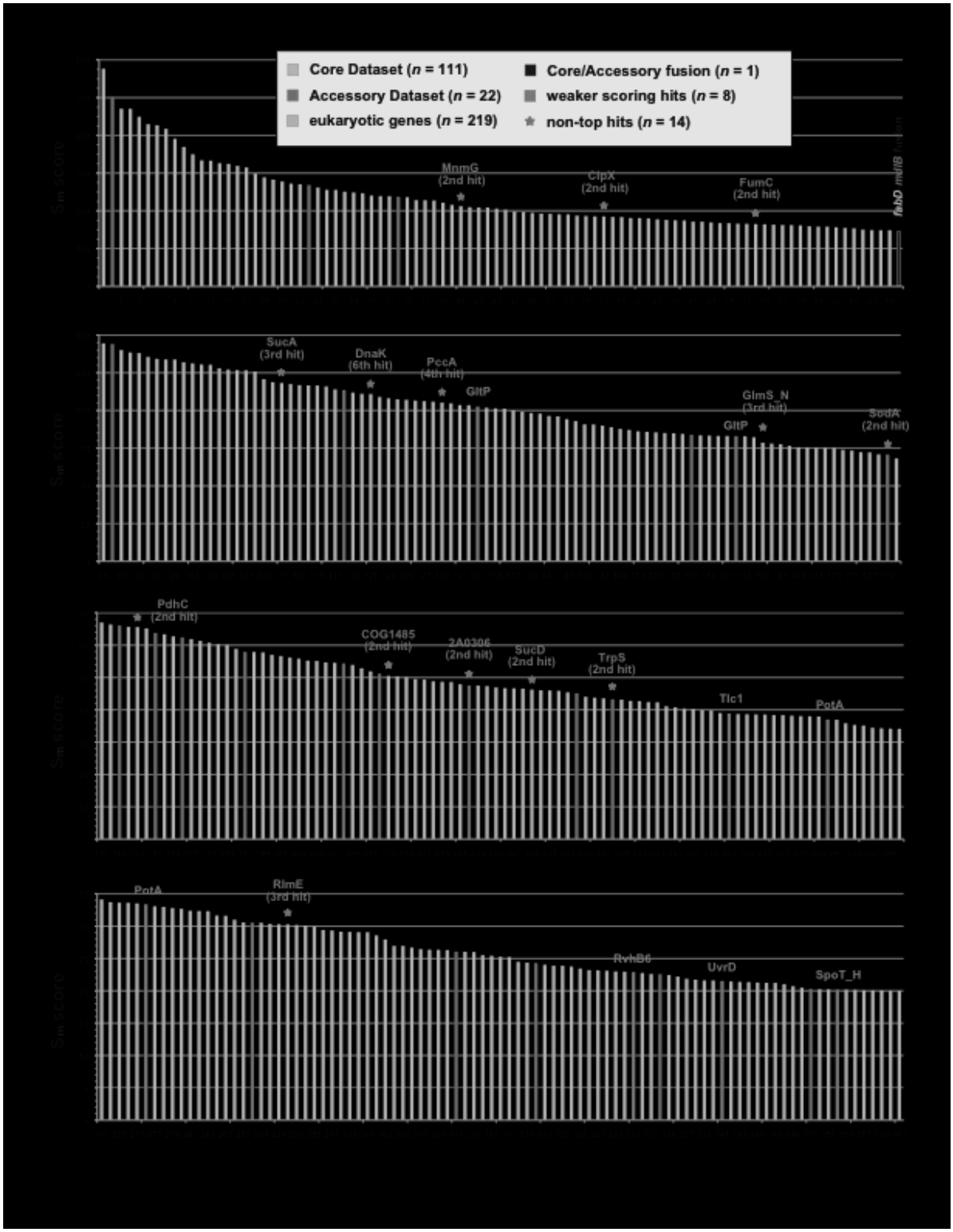
RETA no.	CDS	Top hit (Rickettsiales)	S_m	Top hit (Eukaryota)	S_m	Top hit (Bacteria, sans Rickettsiales)	S_m
RETA0004	leuS	Neorickettsia risticii str. Illinois	43.006	Bombus terrestris	56.308	Sulfurihydrogenibium azorense Az-Fu1	59.689
RETA0006	mutS	Candidatus Midichloria mitochondrii IricVA	125.03	Dictyostelium purpureum	115.5	Parvibaculum lavamentivorans DS-1 ●	133.47
RETA0015	rpmA	Rickettsia prowazekii str. Madrid E	33.917	Dictyostelium purpureum	34.472	Bartonella tribocorum CIP 105476 ●	46
RETA0018	rplQ	Candidatus Odyssella thessalonicensis L13	40.857	Thalassiosira pseudonana CCMP1335	47.911	Symbiobacterium thermophilum IAM 14863	58.145
RETA0020	yccA	Candidatus Odyssella thessalonicensis L13	37.037	Ixodes scapularis	20.774	Magnetospirillum magneticum AMB-1 ●	37.779
RETA0028	groS	Candidatus Midichloria mitochondrii IricVA	98.165	Xenopus (Silurana) tropicalis	69.411	Candidatus Puniceispirillum marinum IMCC1322 ●	100.73
RETA0036	recN	Rickettsia bellii RML369-C	20.228	Ixodes scapularis	7.8274	Rhodomicrobium vannielii ATCC 17100 ●	24.583
RETA0040	rpsK	Candidatus Odyssella thessalonicensis L13	83.144	Cyanophora paradoxa	89.067	Maricaulis maris MCS10 ●	114.51
RETA0046	fabD	Rickettsia felis URRWXCal2	135.75	Bombus impatiens	74.227	Bartonella henselae str. Houston-1 ●	150.04
RETA0047	fabG	Ehrlichia chaffeensis str. Arkansas	155.95	Bombus impatiens	129.54	Asticcacaulis excentricus CB 48 ●	157.61
RETA0055	rpsC	Anaplasma phagocytophilum HZ	138	Bombus impatiens	138.87	Magnetococcus sp. MC-1 ●	156.38
RETA0056	rplN	Candidatus Midichloria mitochondrii IricVA	94.53	Cuscuta reflexa	75.737	Lawsonia intracellularis PHE/MN1-00	102.06
RETA0071	COG0217	Candidatus Midichloria mitochondrii IricVA	79.04	Ixodes scapularis	74.195	Rhodopseudomonas palustris BisB5 ●	81.792
RETA0075	ampD	Rickettsia canadensis str. McKiel	23.985	Micromonas pusilla CCMP1545	4.9413	Chelativorans sp. BNC1 ●	37.906
RETA0088	yidD	Rickettsia felis URRWXCal2	28.885	Micromonas sp. RCC299	20.266	Lactobacillus crispatus ST1	35.832
RETA0103	mnmA_N	Candidatus Midichloria mitochondrii IricVA	52.131	Schistosoma mansoni	31.765	Azospirillum sp. B510 ●	57.162
RETA0169	mnmA_C	NA	NA	NA	NA	NA	NA
RETA0108	ppsA	Candidatus Midichloria mitochondrii IricVA	87.526	Oryza sativa Japonica Group	90.453	Deferribacter desulfuricans SSM1	105.16
RETA0127	lpxA	Candidatus Midichloria mitochondrii IricVA	124.65	Arabidopsis thaliana	66.31	Micavibrio aeruginosavorus ARL-13 ●	135.78
RETA0130	tpiA	Wolbachia endosymbiont strain TRS of B. malayi	52.822	Nectria haematococca mpVI 77-13-4	54.612	Brachyspira murdochii DSM 12563	66.39
RETA0132	DUF3576	Rickettsia endosymbiont of Ixodes scapularis	34.846	Saccoglossus kowalevskii	1.7379	Rhodospirillum rubrum ATCC 11170 ●	43.087
RETA0141	rplT	Ehrlichia ruminantium str. Gardel	67.548	Physcomitrella patens subsp. patens	50.459	Magnetospirillum magneticum AMB-1 ●	69.885
RETA0144	secY	Candidatus Midichloria mitochondrii IricVA	68.241	Xenopus (Silurana) tropicalis	50.055	Xanthobacter autotrophicus Py2 ●	68.696
RETA0145	rpoD	Rickettsia bellii OSU 85-389	24.896	Bombus impatiens	1.3431	Dinoroseobacter shibae DFL 12 ●	27.708
RETA0156	grxC2	Neorickettsia risticii str. Illinois	85.571	Neosartorya fischeri NRRL 181	56.271	Candidatus Liberibacter solanacearum CLso-ZC1 ●	87.369
RETA0161	acpP	Rickettsia peacockii str. Rustic	35.007	Hydra magnipapillata	48.821	Herminiimonas arsenicoxydans	54.525
RETA0170	glmS_N	Candidatus Odyssella thessalonicensis L13	181.51	Bos taurus	118.32	Azospirillum sp. B510 ●	197.78
RETA0121	glmS_C	NA	NA	NA	NA	NA	NA

● = *Alphaproteobacteria*

- Best S_m score across three BLASTP analyses.
- Results determined using phylogeny estimation.
- ***atpA*** was identified after the core dataset was parsed into subsets for analysis.
- ***tolC*** and ***sppA*** regions were separated and analyzed as different proteins.
- Only one region (***rmuC***, ***kdsA***) from these chimeric gene models was analyzed.
- The larger NT protein domains from these split genes (***secA***, ***mnmA***, ***glmS***) were analyzed.

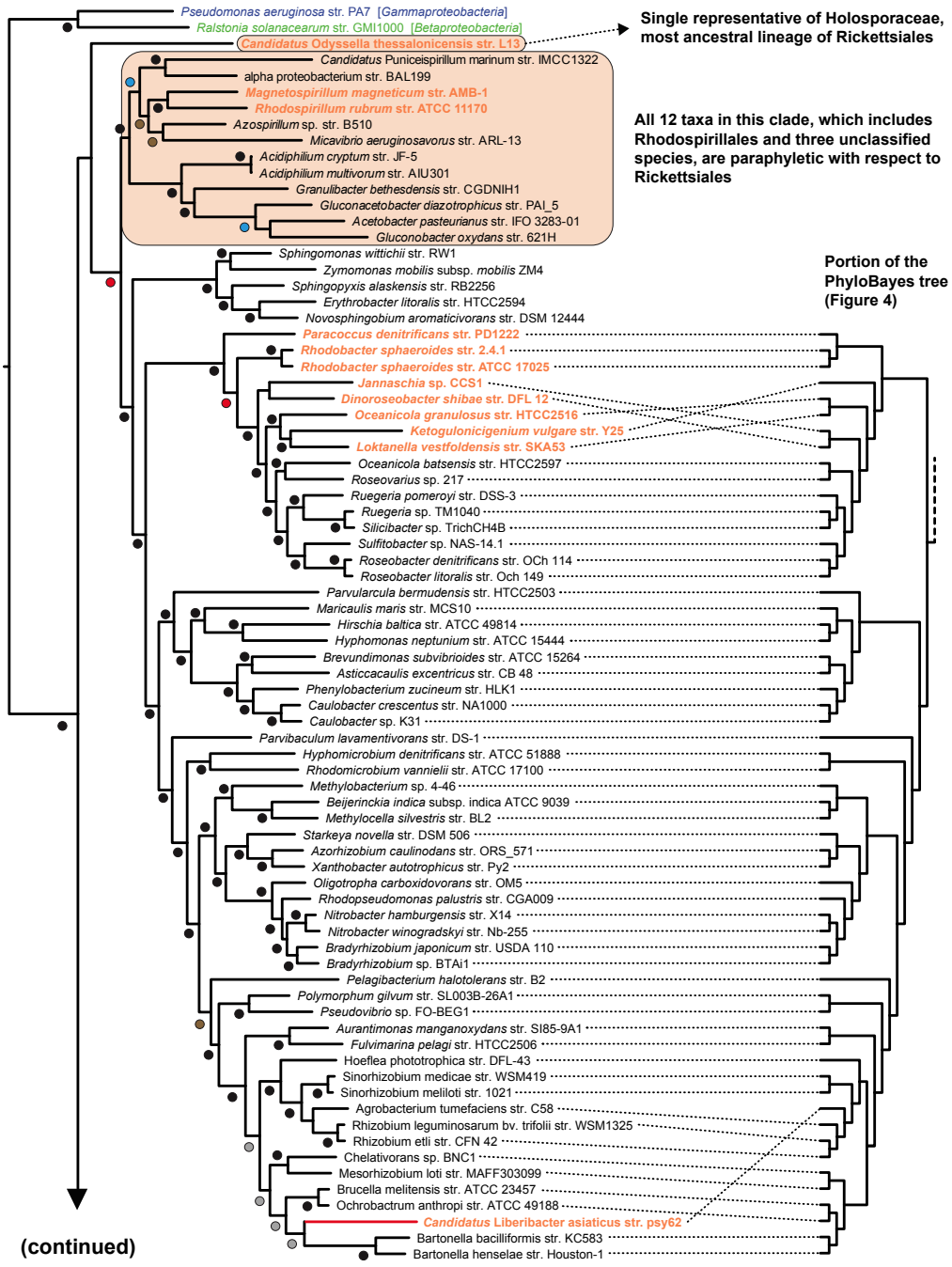
Supplementary Figure S2. Subset parsing of the Core Dataset proteins. The Rickettsiales endosymbiont of *T. adhaerens* (RETA) Core Dataset proteins were blasted against three databases: 'Rickettsiales' (Ric), 'Eukaryota' (Euk) and 'Bacteria minus Rickettsiales' (Bac). **(a)**

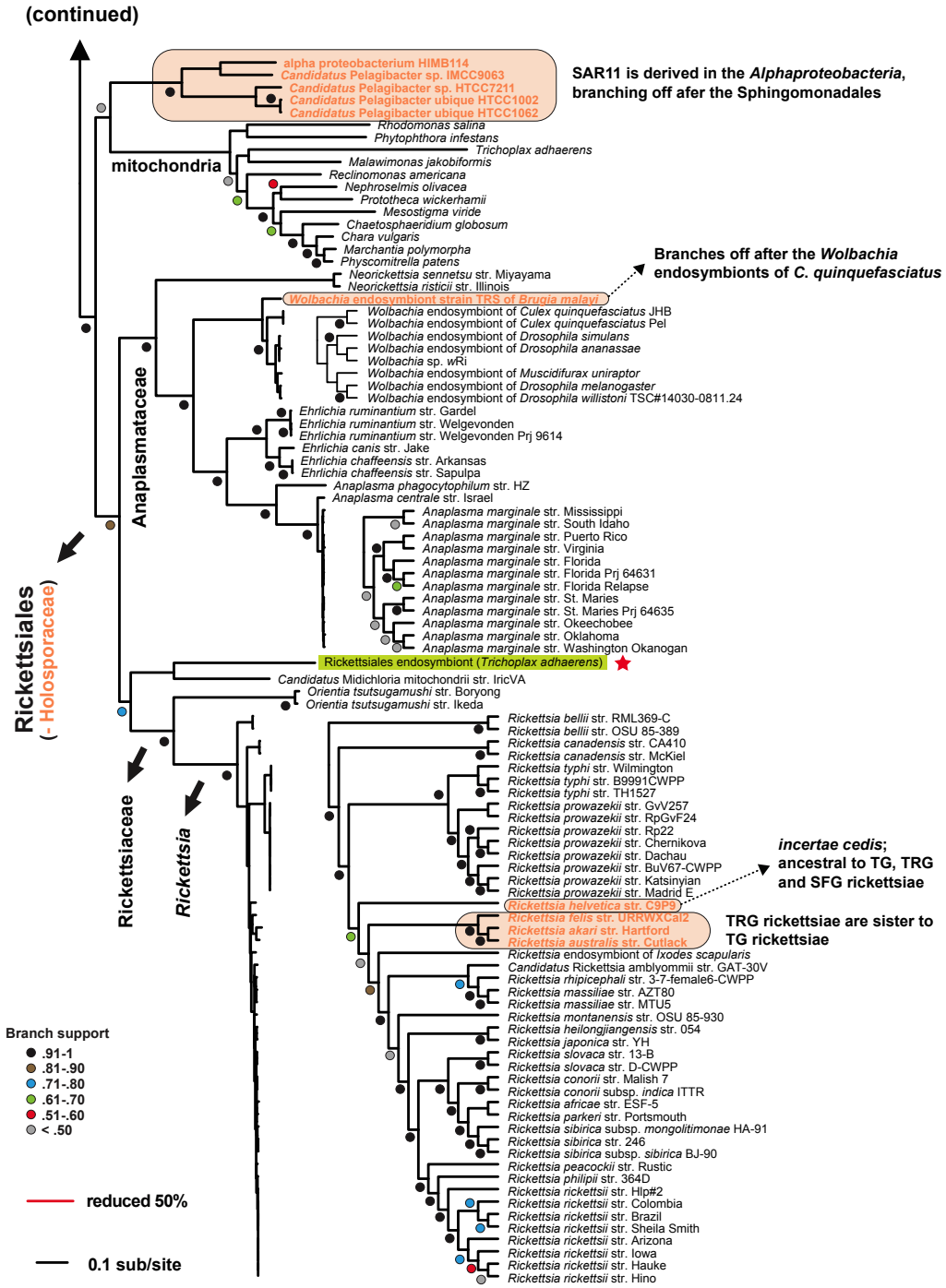
Division of 113 RETA Core Dataset proteins into Ric (78), Euk (9) and Bac (26) sub-datasets based on top S_m scores. Total number of proteins (blue) and avg. %GC (red) for each sub-dataset are provided, with each sub-dataset further divided into the two possible ranks of top S_m scores (described at bottom of *x*-axis). **(b)** RETA Core Dataset proteins of the Ric sub-dataset. **(c)** RETA Core Dataset proteins of the Euk sub-dataset. **(c)** RETA Core Dataset proteins of the Bac sub-dataset.



Supplementary Figure S3. Graphical depiction of the results from an all-against-all BLASTP analysis between the genomes of *T. adhaerens* Grell-BS-1999 and "*Candidatus* Midichloria mitochondrii" str. IricVA. Results are shown in less detail in ring 3 of the genome comparison

provided in the text (**Figure 4a**). Briefly, the analysis of proteins from the *T. adhaerens* (*n* = 11,540) and *M. mitochondrii* (*n* = 1,211) genomes resulted in 347 *T. adhaerens-M. mitochondrii* matches with an S_m score > 20. An additional 14 cases, in which the predicted match between the Rickettsiales endosymbiont of *T. adhaerens* (RETA) and *M. mitochondrii* was not the top hit, were manually added to bring the total matches to 361. These 361 BLASTP matches are graphed along the *x*-axis by decreasing S_m score (*y*-axis). Inset at top describes the different categories for the protein matches: yellow, RETA Core Dataset protein; blue, RETA Accessory Dataset protein; gray, eukaryotic protein (not RETA); black, *fabD-mdlB* fusion protein; green, *M. mitochondrii* protein matching a eukaryotic protein (not RETA) since a paralog (or a protein with an analogous domain) had a higher hit to a RETA protein. Red stars depict the 14 proteins added manually (described above). NOTE: based on annotation (and subsequent phylogeny estimation) we also detected five matches between RETA and *M. mitochondrii* that were below the cutoff (S_m score > 20). These are not shown in the histogram, but are depicted on ring 3 of **Figure 4a** and highlighted yellow. In sum, a total of 138 RETA CDS were determined to have homologs in the *M. mitochondrii* genome (**Figure 4b**).


Pseudomonas aeruginosa str. PA7 [Gammaproteobacteria]
Ralstonia solanacearum str. GMI1000 [Betaproteobacteria]
Candidatus Odyssella thessalonicensis str. L13
Single representative of Holosporaceae, most ancestral lineage of Rickettsiales
Candidatus Puniceispirillum marinum str. IMCC1322
alpha proteobacterium str. BAL199
Magnetospirillum magneticum str. AMB-1
Rhodospirillum rubrum str. ATCC 11170
Azospirillum sp. str. B510
Micavibrio aeruginosavorus str. ARL-13
Acidiphilium cryptum str. JF-5
Acidiphilium multivorum str. AIU301
Granulibacter bethesdensis str. CGDNIH1
Gluconacetobacter diazotrophicus str. PAl_5
Acetobacter pasteurianus str. IFO 3283-01
Gluconobacter oxydans str. 621H
All 12 taxa in this clade, which includes Rhodospirillales and three unclassified species, are paraphyletic with respect to Rickettsiales
Portion of the PhyloBayes tree (Figure 4)
Sphingomonas wittichii str. RW1
Zymomonas mobilis subsp. mobilis ZM4
Sphingopyxis alaskensis str. RB2256
Erythrobacter litoralis str. HTCC2594
Novosphingobium aromaticivorans str. DSM 12444
Paracoccus denitrificans str. PD1222
Rhodobacter sphaeroides str. 2.4.1
Rhodobacter sphaeroides str. ATCC 17025
Jannaschia sp. CCS1
Dinoroseobacter shibae str. DFL 12
Oceanicola granulosus str. HTCC2516
Ketogulonicigenium vulgare str. Y25
Loktanella vestfoldensis str. SKA53
Oceanicola batsensis str. HTCC2597
Roseovarius sp. 217
Ruegeria pomeroyi str. DSS-3
Ruegeria sp. TM1040
Silicibacter sp. TrichCH4B
Sulfitobacter sp. NAS-14.1
Roseobacter denitrificans str. OCh 114
Roseobacter litoralis str. Och 149
Parvularcula bermudensis str. HTCC2503
Maricaulis maris str. MCS10
Hirschia baltica str. ATCC 49814
Hyphomonas neptunium str. ATCC 15444
Brevundimonas subvibrioides str. ATCC 15264
Asticcacaulis excentricus str. CB 48
Phenylobacterium zucineum str. HLK1
Caulobacter crescentus str. NA1000
Caulobacter sp. K31
Parvibaculum lavamentivorans str. DS-1
Hyphomicrobium denitrificans str. ATCC 51888
Rhodomicrobium vannielii str. ATCC 17100
Methylobacterium sp. 4-46
Beijerinckia indica subsp. indica ATCC 9039
Methylocella silvestris str. BL2
Starkeya novella str. DSM 506
Azorhizobium caulinodans str. ORS_571
Xanthobacter autotrophicus str. Py2
Oligotropha carboxidovorans str. OM5
Rhodopseudomonas palustris str. CGA009
Nitrobacter hamburgensis str. X14
Nitrobacter winogradskyi str. Nb-255
Bradyrhizobium japonicum str. USDA 110
Bradyrhizobium sp. BTAi1
Pelagibacterium halotolerans str. B2
Polymorphum gilvum str. SL003B-26A1
Pseudovibrio sp. FO-BEG1
Aurantimonas manganoxydans str. SI85-9A1
Fulvimarina pelagi str. HTCC2506
Hoeflea phototrophica str. DFL-43
Sinorhizobium medicae str. WSM419
Sinorhizobium meliloti str. 1021
Agrobacterium tumefaciens str. C58
Rhizobium leguminosarum bv. trifolii str. WSM1325
Rhizobium etli str. CFN 42
Chelativorans sp. BNC1
Mesorhizobium loti str. MAFF303099
Brucella melitensis str. ATCC 23457
Ochrobactrum anthropi str. ATCC 49188
Candidatus Liberibacter asiaticus str. psy62
Bartonella bacilliformis str. KC583
Bartonella henselae str. Houston-1
(continued)



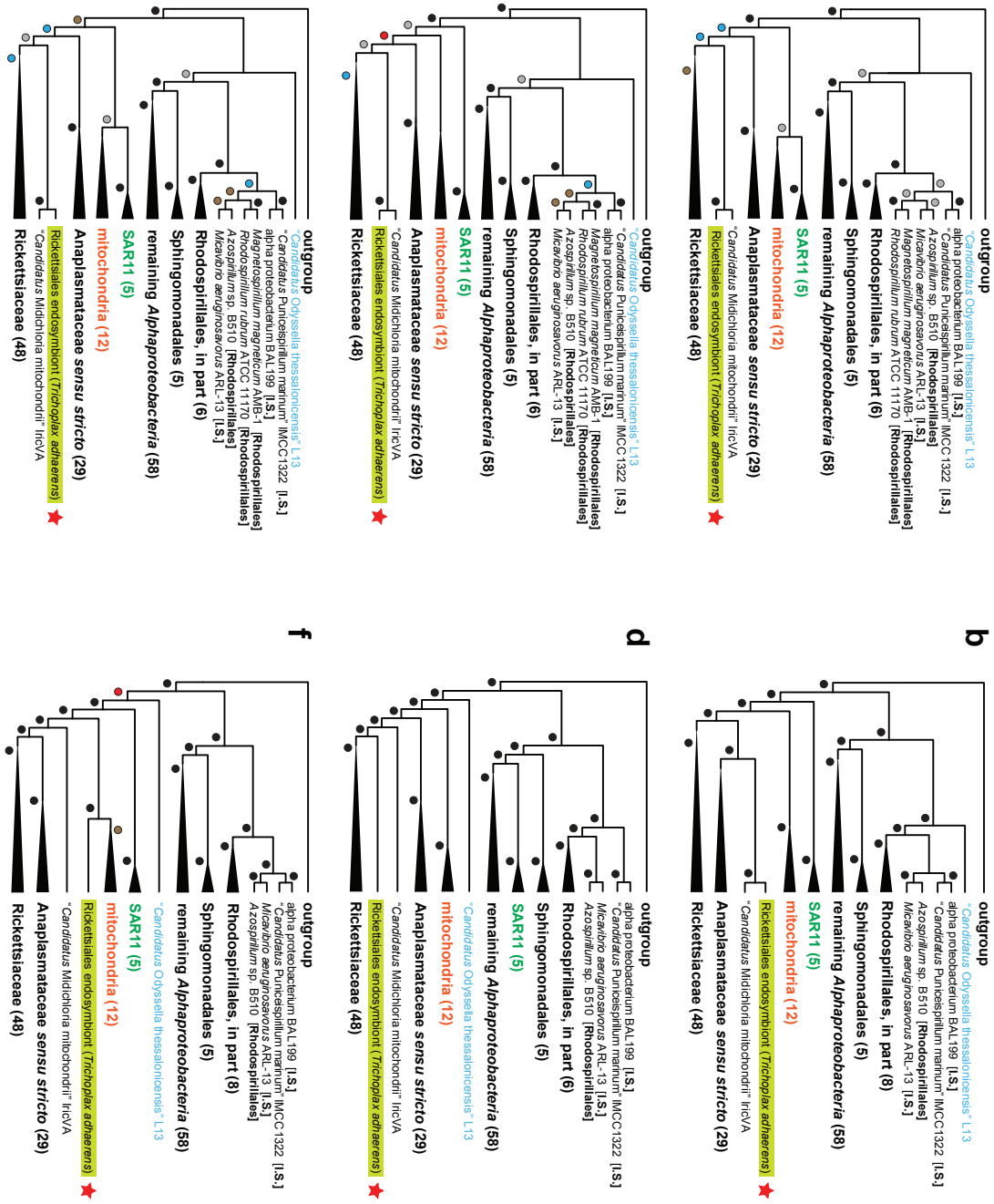
Supplementary Figure S4. Genome-based phylogeny (FastTree) estimated for Rickettsiales endosymbiont of *T. adhaerens* (RETA), 162 alphaproteobacterial taxa, twelve mitochondria, and two outgroup taxa. RETA core proteins (n = 113) were included in the phylogenetic pipeline

that entails ortholog group (OG) generation, OG alignment (and masking of less conserved positions), and concatenation of aligned OGs (see text). Tree was estimated using FastTree **(Price, Dehal et al. 2010)**. Branch support was estimated using a modified bootstrapping procedure, with 100 pseudoreplications sampling only half of the aligned protein sets per replication. RETA is boxed green and noted with a red star. Classification scheme for *Rickettsia* spp. follows previous studies **(Gillespie, Beier et al. 2007; Gillespie, Williams et al. 2008)**. Taxon names, PATRIC genome IDs (bacteria) and NCBI accession numbers (mitochondria) for the 176 genomes are provided in **Supplementary Table S3**. Areas highlighted in orange depict differences between this tree and the tree generated using the CAT-GTR model of substitution (**Figure 5**). Red branch is reduced 50%.

Gillespie, J. J., M. S. Beier, et al. (2007). "Plasmids and rickettsial evolution: insight from *Rickettsia felis*." PLoS ONE **2**(3): e266.

Gillespie, J. J., K. Williams, et al. (2008). "*Rickettsia* Phylogenomics:Unwinding the Intricacies of Obligate Intracellular Life." PLos One **3**(4): e2018.

Price, M. N., P. S. Dehal, et al. (2010). "FastTree 2--approximately maximum-likelihood trees for large alignments." PLoS One **5**(3): e9490.



Supplementary Figure S5. Phylogeny estimations for the three sub-datasets (Ric, Bac or Euk) parsed from the Core Dataset proteins. Briefly, all Rickettsiales endosymbiont of *T. adhaerens*

(RETA) Core Dataset proteins were blasted against three databases: Rickettsiales, Bacteria (excluding Rickettsiales) and Eukaryota. Proteins were then binned into three 'sub-datasets' (Ric, Bac or Euk) based on the highest S_m score against each database. The phylogenetic pipeline (see text) was implemented, resulting in trees estimated with FastTree **(a, c, e)** and PhyloBayes **(b, d, f)**. RETA is boxed green and noted with a red star. Taxon names, PATRIC genome IDs (bacteria) and NCBI accession numbers (mitochondria) for the 176 genomes are provided in **Supplementary Table S3**. **(a, b)** Trees estimated on the Ric-78 sub-dataset. **(c, d)** Trees estimated on the Bac-26 sub-dataset. **(e, f)** Trees estimated on the Euk-9 sub-dataset.

Supplementary Figure S6 is omitted here due to space considerations; it can be accessed in its entirety in the online manuscript (http://gbe.oxfordjournals.org/content/5/4/621).

Supplementary Figure S6. Phylogeny estimation and bioinformatic analysis of 27 rickettsial-like CDSs mined from the *Trichoplax adhaerens* genome assembly. These proteins of the Rickettsiales endosymbiont of *T. adhaerens* (RETA) Accessory Dataset are highly similar to Rickettsiales signature proteins (**Supplementary Table S5**). CDSs have an avg. %GC that is typical of rickettsial genomes and similar to the CDSs of the RETA Core Dataset (**Supplementary Figure S2**). Information pertaining to each RETA CDS is provided in **Supplementary Table S2**. See text for complete dataset construction, alignment and tree-building methods. Trees shown here are from estimations under maximum likelihood using RAxML v.7.2.8 **(Stamatakis, Hoover et al. 2008)**. A gamma model of rate heterogeneity was used with estimation of the proportion of invariable sites. Branch support was assessed with 1000 bootstrap pseudoreplications. Trees were also estimated under parsimony using PAUP* v4.0b10 (Altivec) **(Wilgenbusch and Swofford 2003)** (data not shown), and while differing in some instances with the RAxML trees, the phylogenetic position of RETA was unchanged. Within each panel, RETA is boxed green and noted with a red star. Gray boxes depict Rickettsiales taxa, with rickettsial lineages containing RETA within dashed boxes.

Stamatakis, A., P. Hoover, et al. (2008). "A rapid bootstrap algorithm for the RAxML Web servers." Syst Biol **57**(5): 758-771.

Wilgenbusch, J. C. and D. Swofford (2003). "Inferring evolutionary trees with PAUP*." Curr Protoc Bioinformatics **Chapter 6**: Unit 6 4.

Supplementary Figure S7 is omitted here due to space considerations; it can be accessed in its entirety in the online manuscript (http://gbe.oxfordjournals.org/content/5/4/621).

Supplementary Figure S7. Phylogeny estimation and bioinformatic analysis of 18 bacterial-like CDSs mined from the *Trichoplax adhaerens* genome assembly. These proteins of the Rickettsiales endosymbiont of *T. adhaerens* (RETA) Accessory Dataset are present in some (or all) Rickettsiales genomes yet divergent in sequence and phylogenetic signal. CDSs have an avg. %GC that is higher than that of rickettsial genomes, and higher (39% vs. 30%) than the CDSs of the RETA Core Dataset (**Supplementary Figure S2**). Information pertaining to each RETA CDS is provided in **Supplementary Table S2**. See text for complete dataset construction, alignment and tree-building methods. Trees shown here are from estimations under maximum likelihood using RAxML v.7.2.8 **(Stamatakis, Hoover et al. 2008)**. A gamma model of rate heterogeneity was used with estimation of the proportion of invariable sites. Branch support was assessed with 1000 bootstrap pseudoreplications. Trees were also estimated under parsimony using PAUP* v4.0b10 (Altivec) **(Wilgenbusch and Swofford 2003)** (data not shown), and while differing in some instances with the RAxML trees, the phylogenetic position of RETA was unchanged. Within each panel, RETA is boxed green and noted with a red star. Gray boxes depict Rickettsiales taxa.

Stamatakis, A., P. Hoover, et al. (2008). "A rapid bootstrap algorithm for the RAxML Web servers." Syst Biol **57**(5): 758-771.

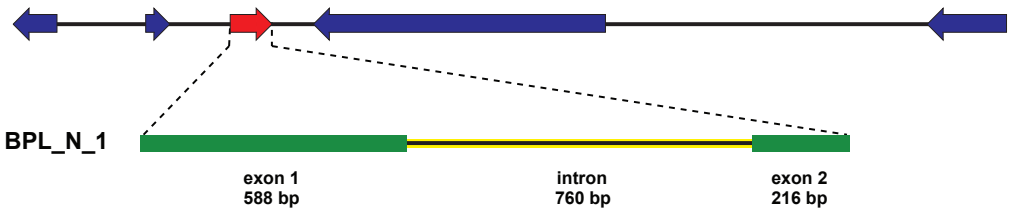
Wilgenbusch, J. C. and D. Swofford (2003). "Inferring evolutionary trees with PAUP*." Curr Protoc Bioinformatics **Chapter 6**: Unit 6 4.

Supplementary Figure S8 is omitted here due to space considerations; it can be accessed in its entirety in the online manuscript (http://gbe.oxfordjournals.org/content/5/4/621).

Supplementary Figure S8. Phylogeny estimations for the 18 bacterial-like proteins determined to be present within large *Trichoplax adhaerens* scaffolds. These proteins of the Rickettsiales endosymbiont of *T. adhaerens* (RETA) Core and Accessory Datasets are arrayed with eukaryotic genes on large scaffolds (**Figure 7**). RETA IDs and protein names, scaffold information, and fgenesb predictions are further described in **Supplementary Table S2**. See text for complete dataset construction, alignment and tree-building methods. Trees shown here are from estimations under maximum likelihood using RAxML v.7.2.8. A gamma model of rate heterogeneity was used with estimation of the proportion of invariable sites. Branch support was assessed with 1000 bootstrap pseudoreplications. Trees were also estimated under parsimony using PAUP* v4.0b10 (Altivec) (data not shown), and while differing in some instances with the RAxML trees, the phylogenetic position of RETA was unchanged. Within each panel, RETA is boxed green and noted with a red star. Gray boxes depict Rickettsiales taxa. Accession numbers colored red and blue depict annotations that implicate the protein to be imported to the mitochondria and chloroplast, respectively. Potential bacteria-to-*T. adhaerens* LGT products are noted for PhrB_1-4, GNAT_1-2, RsmE, MhpC, DapF, BPL_N_1-2, and MurA.

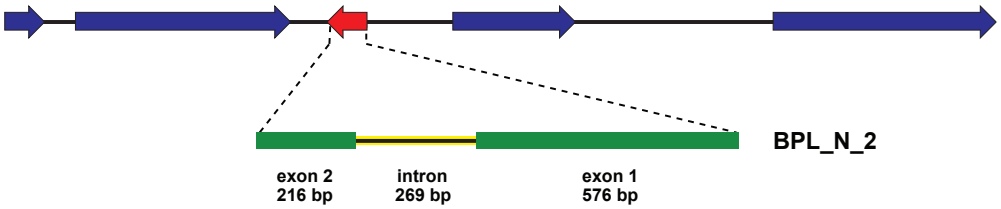
a

EDV27249	EDV26701	EDV26702	EDV27250	EDV27251
8,683,252-8,684,967	8,688,615-8,689,543	8,692,064-8,693,627	8,695,455-8,707,163	8,720,302-8,723,340
DUF1762 - 1,716 bp - 4 exons - 329 aa	PA28_beta - 929 bp - 6 exons - 146 aa	BPL_N - 1,564 bp - 2 exons - 267 aa	Adcy5 - 11,709 bp - 23 exons - 1,044 aa	MBOAT - 3,039 bp - 11 exons - 411 aa



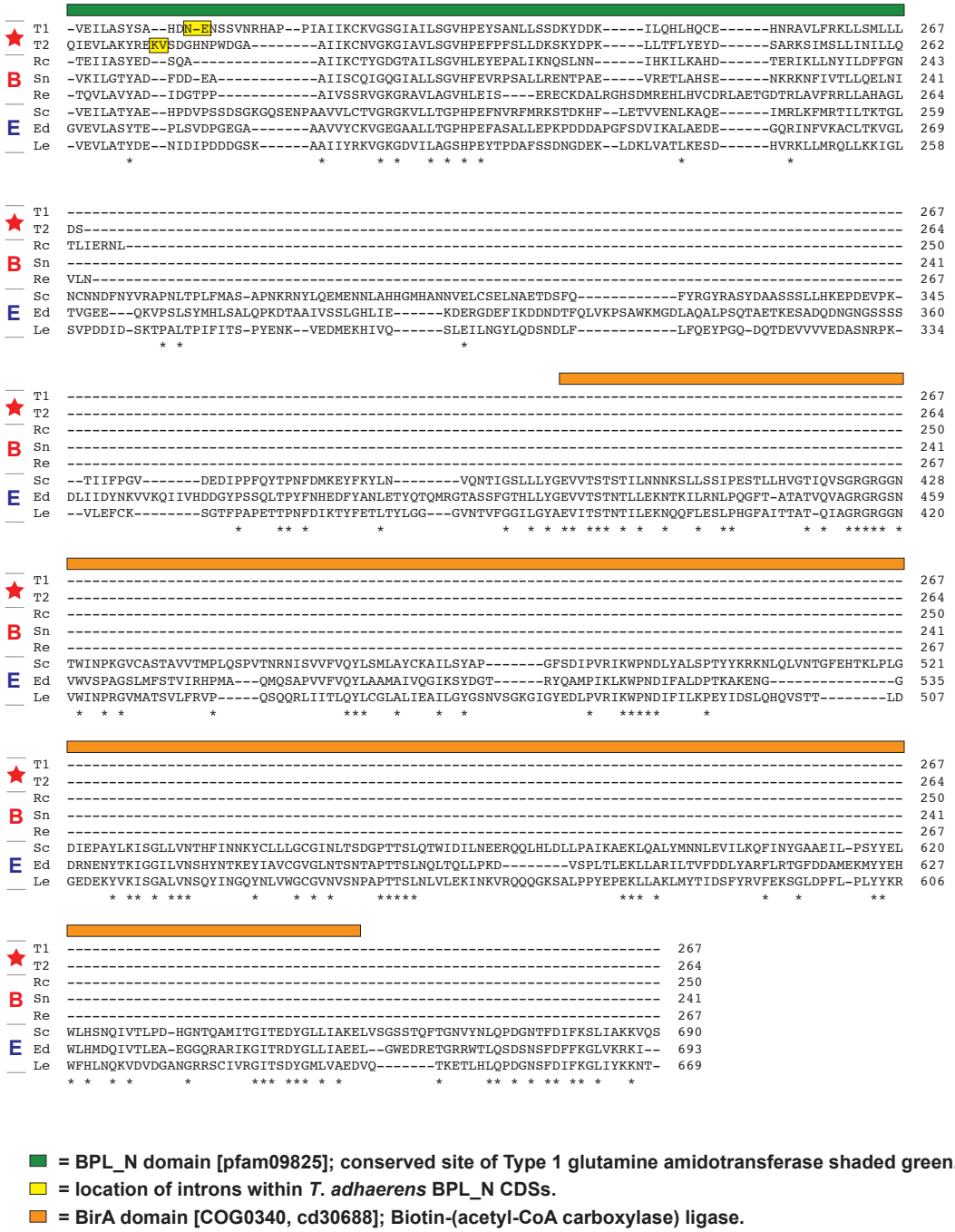
b

EDV23818	EDV23819	EDV24178	EDV23820	EDV23821
2,692,045-2,693,002	2,693,951-8,699,614	2,700,642-2,701,705	2,704,037-2,707,264	2,712,599-2,718,523
HP - 958 bp - 4 exons - 111 aa	Armc3 - 5,664 bp - 24 exons - 800 aa	BPL_N - 1,064 bp - 2 exons - 264 aa	HP - 3,228 bp - 3 exons - 294 aa	HP - 5,925 bp - 8 exons - 403 aa



c

c (continued)



= BPL_N domain [pfam09825]; conserved site of Type 1 glutamine amidotransferase shaded green.
= location of introns within *T. adhaerens* BPL_N CDSs.
= BirA domain [COG0340, cd30688]; Biotin-(acetyl-CoA carboxylase) ligase.

Supplementary Figure S9. Evidence for introns within the *Trichoplax adhaerens* genes that encode bacterial-like BPL_N proteins. **(a, b)** Location of BPL_N encoding genes within two *T. adhaerens* scaffolds. Schemas are redrawn from the NCBI genome browser. Descriptions for

each gene are provided above the schema, with eukaryotic-like genes colored blue and bacterial-like genes (BPL_N) colored red. Below the schema, BPL_N encoding genes models are shown, with CDSs containing the BPL_N domain colored green, and introns colored yellow. **(a)** Gene order on *T. adhaerens* scaffold NW_002060945 (coordinates 8,683,252-8,723,340). Eukaryotic genes flanking BPL_N_1: DUF1762 (XP_002111245), proteasome activator complex subunit beta (XP_002110697), adenylate cyclase type 5 (XP_002111246), sterol O-acyltransferase 1 (XP_002111247). **(b)** Gene order on *T. adhaerens* scaffold NW_002060948 (coordinates 2,692,045-2,718,523). Eukaryotic genes flanking BPL_N_2: HP (XP_002113344), Armadillo repeat-containing protein 3 (XP_002113345), HP (XP_002113346), HP (XP_002113347). **(c)** Multiple sequence alignment of BPL_N_1, BPL_N_2, three bacterial BPL_N homologs, and three eukaryotic BirA proteins. Above the alignment, the BPL_N domains are colored green, with the predicted intron insertion sites highlighted in yellow on the sequences, and the BirA domains of the eukaryotic sequences are colored orange. Under the alignment, asterisks indicate conserved positions across all eight BPL_N and BirA sequences, or just the BirA domains. To the left of the alignment, stars indicate the *T. adhaerens* BPL_N sequences, while B and E indicate the bacterial BPL_N proteins and eukaryotic BirA proteins, respectively. Sequences were selected from the dataset used to estimate phylogeny of BPL_N proteins (see **Supplementary Figure S6**). Abbreviations as follows: T1, *Trichoplax adhaerens* hypothetical protein TRIADDRAFT_55000 (XP_002110698); T2, *Trichoplax adhaerens* hypothetical protein TRIADDRAFT_57402 (XP_002113704); Rc, *Rickettsia conorii* str. Malish 7 N-terminal of biotin-protein ligase (NP_360418); Sn, *Simkania negevensis* str. Z hypothetical protein SNE_A06480 (YP_004671016); Re, *Ralstonia eutropha* str. H16 biotin apo-protein ligase (YP_841268); Sc, *Saccharomyces cerevisiae* YJM789 biotin:apoprotein ligase (EDN60216); Ed, *Exophiala dermatitidis* NIH/UT8656 hypothetical protein HMPREF1120_07008 (EHY59008); Le, *Lodderomyces elongisporus* NRRL YB-4239 hypothetical protein LELG_05238 (XP_001523392).

Supplementary Table S1. Sequences used to estimate SSU rDNA phylogeny.

Taxon [1]	Accession [2]
Outgroup (5)	
Uncultured *Ralstonia* sp. clone F-17 (*Betaproteobacteria*)	HQ132432
Halomonas sp. JL1044 (*Gammaproteobacteria*)	DQ985041
Rhodospirillum rubrum ATCC 11170	**VBIAlpPro78664_r028**
Pelagibacterium halotolerans B2	**VBIPelHal211702_r039**
Parvularcula bermudensis HTCC2503	**VBIParBer119301_r044**
Holosporaceae (16)	
"*Candidatus* Caedibacter acanthamoebae"	AF132138
Caedibacter caryophilus str. 221	X71837
Uncultured deep-sea bacterium clone Ucm1520	AM997318
Holospora obtusa	X58198
"*Candidatus* Paraholospora nucleivisitans"	EU652696
Uncultured bacterium clone Oh3123O11E	EU137369
Uncultured bacterium clone HOClCi16	AY328565
"*Candidatus* Captivus acidiprotistae"	AF533508
Uncultured bacterium clone R2-FM3	GU477314
Uncultured bacterium clone R2-Liz3	GU477316
Endosymbiont of *Acanthamoeba* sp. KA/E23	EF140636
"*Candidatus* Odyssella thessalonicensis" str. L13	**VBICanOdy184137_r038**
"*Candidatus* Odyssella thessalonicensis" str. L13	AF069496
Holosporaceae bacterium str. Serialkilleuse_9403403	HM138368
Uncultured "*Candidatus* Odyssella sp." clone 5-F	EU305601
Endosymbiont of *Acanthamoeba* sp. TUMK-23	AY102614
Incertae sedis	
Uncultured Rickettsiales bacterium clone PRTBB8516	HM798949
Mitochondria (10)	
Nephroselmis olivacea	NC_008239
Mesostigma viride	NC_008240
Chaetosphaeridium globosum	NC_004118
Physcomitrella patens	NC_007945
Chara vulgaris	NC_005255
Reclinomonas americana	AF007261
Prototheca wickerhamii	X15435
Malawimonas jakobiformis	NC_002553
Rhodomonas salina	NC_002572
Phytophthora infestans	NC_002387
Rickettsiaceae (18)	
Uncultured bacterium clone ELB16-030	DQ015802
Uncultured bacterium clone R1-FM1	GU477308
Uncultured bacterium clone R1-Liz1	GU477312
"*Candidatus* Cryptoprodotis polytropus" isolate PSM1	FM201293
"*Candidatus* Occidentia massiliensis" (Rickettsiaceae bacterium str. Os18)	GU937608

Taxon [1]	Accession [2]
Secondary symbiont of *Sitobion miscanthi* clone Hefei	HM156647
Orientia tsutsugamushi str. Ikeda	**VBIOriTsu129072_r006**
Rickettsia endosymbiont of *Deronectes semirufus*	FM955311
Rickettsia endosymbiont of *Torix tagoi*	AB066351
Rickettsia endosymbiont of *Ixodes scapularis*	**VBIRicEnd40569_r031**
Secondary endosymbiont of *Curculio sikkimensis*	AB545027
Rickettsia bellii str. RML369-C	**VBIRicBel102610_r012**
Uncultured bacterium clone SHFG464	FJ203077
Uncultured Rickettsiales bacterium clone Ho(lakePohlsee)_4	EF667896
Bacterial symbiont of *Diophrys appendiculata*	AJ630204
Rickettsiaceae endosymbiont of *Carteria cerasiformis*	AB688628
Rickettsiaceae endosymbiont of *Pleodorina japonica*	AB688629
Uncultured Rickettsiales bacterium clone EV221H2111601SAH71	DQ223223

Incertae sedis

Uncultured alphaproteobacterium clone SM1B06	AF445655

Anaplasmataceae (12)

"*Candidatus* Xenohaliotis californiensis"	AF069062
Neorickettsia helminthoeca	U12457
Neorickettsia risticii str. Illinois	**VBINeoRis104330_r001**
Uncultured *Neorickettsia* sp. isolate 184	EU780451
Wolbachia endosymbiont str. TRS of *Brugia malayi*	**VBIWolEnd7741_r015**
Endosymbiont of *Rhinocyllus conicus*	M85267
"*Candidatus* Neoanaplasma japonica" (Anaplasmataceae bacterium str. IS136)	AB190771
Ehrlichia ruminantium str. Gardel	**VBIEhrRum72196_r011**
"*Candidatus* Neoehrlichia mikurensis" (Uncultured "*Candidatus* Neoehrlichia sp." clone 2)	GQ501090
Anaplasma marginale str. St Maries	**VBIAnaMar46146_r010**
Aegyptianella pullorum	AY125087
Wolbachia endosymbiont of *Kleidocerys resedae* clone KrWlbOkn1	JQ726769

"Midichloriaceae" (30)

Uncultured Rickettsiales bacterium clone Hv(lakePohlsee)_25	EF667921
"*Candidatus* Cyrtobacter comes"	FN552697
"*Candidatus* Anadelfobacter veles"	FN552695
Uncultured alphaproteobacterium clone MD3.55	FJ425643
Uncultured proteobacterium clone PEACE2006/237_P3	EU394580
Uncultured Rickettsiales bacterium clone Ho_lab_2.5	EF667892
Uncultured bacterium clone RNA62799	JF328639
Uncultured bacterium clone RPR28	FJ824004
Uncultured bacterium clone Gven_P15	GU118498
Rickettsiales endosymbiont of *Trichoplax adhaerens*	**see below** [A]
Uncultured bacterium clone Cc045	AY942762
Uncultured bacterium clone Mfav_P11	GU118616
Uncultured bacterium clone Mfav_B15	GU118630
Uncultured bacterium clone Mfav_F04	GU118640
Endosymbiont of *Acanthamoeba* sp. UWC36	AF069962
Uncultured alpha proteobacterium clone sw-xj63	GQ302530
Uncultured Rickettsiales bacterium clone ID25L	EU555284

Uncultured bacterium clone Hw124	AF497583
Candidatus Midichloria sp. Ixholo1	FM992372
Taxon [1]	**Accession [2]**
"*Candidatus* Nicolleia massiliensis" str. France	DQ788562
"*Candidatus* Midichloria mitochondrii"	AJ566640
"*Candidatus* Midichloria mitochondrii str. IricVA"	**VBICanMid156609_r008**
Rickettsiales bacterium Huangshan-1	AB297807
"*Candidatus* Lariskella arthropodarum" clone KrLaSpr	JQ726734
"*Candidatus* Lariskella arthropodarum" clone AmLaKka1	JQ726736
Uncultured Rickettsiales bacterium 'Montezuma'	AF493952
Rickettsiales bacterium endosymbiont of *Nysius plebeius*	AB624350
"*Candidatus* Lariskella arthropodarum" clone DbLaKnz	JQ726746
Uncultured bacterium clone XC1	FJ981659
Uncultured bacterium clone CF2	FJ981673

[1] Taxa are listed from the top to the bottom as shown in the tree (**Figure 3**). The term "*Candidatus*" is used as originally suggested (Murray and Stackebrandt 1995). *Incertae sedis*, "of uncertain placement", refers to taxa with unknown or undefined broader relationships.

[2] PATRIC (Gillespie, Wattam et al. 2011) accession numbers (bold) are given for rDNA sequences retrieved from genome projects; all other numbers are NCBI nucleotide accession numbers.

[A] The 16S rDNA sequence for Rickettsiales endosymbiont of *Trichoplax adhaerens* (5'-3'):

AGAGTTTGATCCTGGCTCAGAGTGAACGCTAGTGGCGTGCTTAACACATGCAAGTCGAACGGACGATATTTGTGCTTGCACAAATAAAGTTAGTG
GCAAACGGGTGAGTAATACATGGGAATTTTCCTTGCAGTACGGAATAACTATTGGAAACAGTAGCTAATACCGTATATTGCCGAGAGGTGAAAGA
TTTATCGCTGCAAGATAAGCCCATGCAAGATTAGCTTGTTGGTAAGGTAATGGCTTACCAAGGCTACGATCTTTAGCTGGTTTGAGAGGATGATC
AGCCACACTGGAACTTAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGAAAGCCTGATCCAGCGACGTCAC
GTGAATGGTGAAAGCCTTAGGGTTGTAAAATTCTTTTAGTGGAGAAGATAATGACGGTATCCACAGAAAAAGTCCCGGCTAACTTCGTGCCAGCA
GCCGCGGTAATACGAAGGGGGCAAGCGTTACTCGGAATTATTGGGCGTAAAGCGTGCGTAGGCGGTTTTATAAGTTGAAAGTGAAAGCCTTTGGC
TCAACCAAAGAATTGCTTAC

Gillespie, J. J., A. R. Wattam, et al. (2011). "PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species." Infect Immun **79**(11): 4286-4298.

Murray, R. G. and E. Stackebrandt (1995). "Taxonomic note: implementation of the provisional status Candidatus for incompletely described procaryotes." Int J Syst Bacteriol **45**(1): 186-187.

Supplementary Table S2. Bacterial genes mined from the *Trichoplax adhaerens* assembly. Characteristics are provided for all 181 bacterial-like genes in the RETA Core and Accessory Datasets (see Text). Column descriptions: **A)** NCBI accession id of the corresponding gene from the public *T. adhaerens* assembly; **B)** assigned RETA id; **C)** annotated RETA gene symbol; **D)** RETA Dataset membership; **E)** categorization of the RETA gene model, either single, fusion (the *T. adhaerens* entry represents two fused RETA genes), or split (the *T. adhaerens* entry represents one part of a larger RETA gene); **F)** total number of exons in the *T. adhaerens* gene; **G)** length (nt) of the *T. adhaerens* coding region (excludes introns); **H)** percent GC of the *T. adhaerens* coding region (excludes introns); **I)** NCBI id of the *T. adhaerens* scaffold that contains the gene; **J)** total length (nt) of the *T. adhaerens* scaffold; **K)** percent GC across the entire *T. adhaerens* scaffold; **L-O)** total number of RETA, Rickettsial-like, bacterial-like (excluding Rickettsia), and non-bacterial genes, respectively, located on the given scaffold; **P)** total number of contigs that comprise the scaffold; **Q)** internal id of the closest gene model from re-calling gene models on the scaffold using FGENESB; **R)** total length (nt) of a global (Needleman-Wunsch) alignment between the FGENESB model from column Q and the *T. adhaerens* gene; **S)** total number of identical positions, and **T)** percent identity across the alignment; **U)** total number of similar positions, and **V)** percent similarity across the alignment; **W)** total number of gaps in the alignment; and **X)** score of the alignment, assigned by the NEEDLE algorithm.

Supplementary Table S2 is omitted here due to space considerations; it can be accessed in its entirety in the online manuscript (http://gbe.oxfordjournals.org/content/5/4/621).

Supplementary Table S3. Genomes used to estimate Core Dataset phylogeny.

Taxon [1]	ID [2]
Outgroup (2)	
Ralstonia solanacearum GMI1000	70888
Pseudomonas aeruginosa PA7	80442
Sphingomonadales (5)	
Zymomonas mobilis subsp. *mobilis* ZM4	102260
Sphingomonas wittichii RW1	55028
Sphingopyxis alaskensis RB2256	23391
Erythrobacter litoralis HTCC2594	102657
Novosphingobium aromaticivorans DSM 12444	50627
SAR11, *incertae sedis* (5)	
"*Candidatus* Pelagibacter sp. HTCC7211"	29514
"*Candidatus* Pelagibacter sp. IMCC9063"	201843
"*Candidatus* Pelagibacter ubique HTCC1002"	113578
"*Candidatus* Pelagibacter ubique HTCC1062"	5618
alpha proteobacterium HIMB114	140191
Rhodobacterales (20)	
Paracoccus denitrificans PD1222	97112
Rhodobacter sphaeroides 2.4.1	57909
Rhodobacter sphaeroides ATCC 17025	94549
Jannaschia sp. CCS1	43325
Roseobacter denitrificans OCh 114	86677
Roseobacter litoralis Och 149	16390
Ruegeria pomeroyi DSS-3	114501
Ruegeria sp. TM1040	69653
Ketogulonicigenium vulgare Y25	170732
Dinoroseobacter shibae DFL 12	9476
Silicibacter sp. TrichCH4B	32447
Sulfitobacter sp. NAS-14.1	100835
Roseovarius sp. 217	94902
Oceanicola batsensis HTCC2597	2886
Oceanicola granulosus HTCC2516	115923
Loktanella vestfoldensis SKA53	90427
Maricaulis maris MCS10	77530
Hyphomonas neptunium ATCC 15444	17450
Hirschia baltica ATCC 49814	9878
Pseudovibrio sp. FO-BEG1	224584
Parvularculales (1)	
Parvularcula bermudensis HTCC2503	119301
Caulobacterales (5)	
Asticcacaulis excentricus CB 48	23432
Brevundimonas subvibrioides ATCC 15264	37974

Taxon [1]	ID [2]
Caulobacter crescentus NA1000	52860
Caulobacter sp K31	18104
Phenylobacterium zucineum HLK1	44517
Rhizobiales (31)	
Agrobacterium tumefaciens str. C58	91616
Azorhizobium caulinodans ORS 571	17976
Bartonella bacilliformis KC583	6912
Beijerinckia indica subsp. *indica* ATCC 9039	21058
Bradyrhizobium japonicum USDA 110	65052
Brucella melitensis ATCC 23457	14466
"*Candidatus* Liberibacter asiaticus str. psy62"	64949
Mesorhizobium (*Chelativorans*) sp. BNC1	72577
Hyphomicrobium denitrificans ATCC 51888	91677
Mesorhizobium loti MAFF303099	2464
Methylobacterium sp. 4-46	32184
Methylocella silvestris BL2	55537
Nitrobacter hamburgensis X14	61822
Oligotropha carboxidovorans OM5	134280
Parvibaculum lavamentivorans DS-1	90819
Rhizobium leguminosarum bv. *trifolii* WSM1325	48398
Rhodomicrobium vannielii ATCC 17100	113057
Rhodopseudomonas palustris CGA009	84835
Sinorhizobium meliloti 1021	96828
Starkeya novella DSM 506	45716
Xanthobacter autotrophicus Py2	29526
Ochrobactrum anthropi ATCC 49188	73124
Sinorhizobium medicae WSM419	134228
Bartonella henselae str. Houston-1	29080
Rhizobium etli CFN 42	108884
Nitrobacter winogradskyi Nb-255	102302
Bradyrhizobium sp. BTAi1	29847
Aurantimonas manganoxydans SI85-9A1	112117
Pelagibacterium halotolerans B2	211702
Fulvimarina pelagi HTCC2506	16688
Hoeflea phototrophica DFL-43	121037
***Incertae sedis* (5)**	
"*Candidatus* Puniceispirillum marinum IMCC1322"	78664
Polymorphum gilvum SL003B-26A1	182274
Micavibrio aeruginosavorus ARL-13	161757
alpha proteobacterium BAL199	7182
Rhodospirillales (9)	
Magnetospirillum magneticum AMB-1	129836
Rhodospirillum rubrum ATCC 11170	82919
Acidiphilium cryptum JF-5	6074
Acidiphilium multivorum AIU301	178205
Granulibacter bethesdensis CGDNIH1	83793
Gluconobacter oxydans 621H	81109
Gluconacetobacter diazotrophicus PAl 5	122154

Azospirillum sp. B510	82869
Acetobacter pasteurianus IFO 3283-01	139226
Taxon [1]	**ID [2]**
Rickettsiales (81)	
Holosporaceae	
"*Candidatus* Odyssella thessalonicensis L13 "	184137
Anaplasmataceae	
Neorickettsia risticii Illinois	104330
Neorickettsia sennetsu Miyayama	119815
Wolbachia (*Culex quinquefasciatus*) JHB	42581
Wolbachia (*Culex quinquefasciatus*) Pel	95846
Wolbachia (*Drosophila ananassae*)	62175
Wolbachia (*Drosophila melanogaster*)	21207
Wolbachia (*Drosophila simulans*)	67463
Wolbachia (*Drosophila willistoni*) TSC#14030-0811.24	100560
Wolbachia (*Muscidifurax uniraptor*)	74720
Wolbachia strain TRS (*Brugia malayi*)	7741
Wolbachia sp. wRi	98304
Ehrlichia canis Jake	118076
Ehrlichia chaffeensis Arkansas	103583
Ehrlichia chaffeensis Sapulpa	122199
Ehrlichia ruminantium Gardel	72196
Ehrlichia ruminantium Welgevonden	92411
Ehrlichia ruminantium Welgevonden (Prj:9614)	203725
Anaplasma centrale Israel	103588
Anaplasma marginale Florida	82315
Anaplasma marginale Florida (Prj:64631)	203711
Anaplasma marginale Florida Relapse	189448
Anaplasma marginale Mississippi	26748
Anaplasma marginale Okeechobee	187681
Anaplasma marginale Oklahoma	188137
Anaplasma marginale Puerto Rico	90594
Anaplasma marginale South Idaho	188390
Anaplasma marginale St. Maries	46146
Anaplasma marginale St. Maries (Prj:64635)	203712
Anaplasma marginale Virginia	66823
Anaplasma marginale Washington Okanogan	187280
Anaplasma phagocytophilum HZ	602
Incertae sedis	
"Candidatus Midichloria mitochondrii IricVA"	156609
Rickettsiales endosymbiont of *Trichoplax adhaerens*	**NA [A]**
Rickettsiaceae	
Scrub Typhus Group	
Orientia tsutsugamushi Boryong	83812
Orientia tsutsugamushi Ikeda	129072
Ancestral Group	

Rickettsia bellii OSU 85-389	35792
Rickettsia bellii RML369-C	102610
Taxon [1]	**ID [2]**
Rickettsia canadensis McKiel	89738
* *Rickettsia canadensis* CA410	238964

Incertae sedis

* *Rickettsia helvetica* C9P9	217856

Typhus Group

Rickettsia typhi Wilmington	34752
* *Rickettsia typhi* B9991CWPP	187203
* *Rickettsia typhi* TH1527	188083
Rickettsia prowazekii Rp22	89493
Rickettsia prowazekii Madrid E	72556
* *Rickettsia prowazekii* RpGvF24	237744
* *Rickettsia prowazekii* BuV67-CWPP	229008
* *Rickettsia prowazekii* Dachau	238803
* *Rickettsia prowazekii* Katsinyian	230392
* *Rickettsia prowazekii* Chernikova	240232
* *Rickettsia prowazekii* GvV257	231244

Transitional Group

Rickettsia felis URRWXCal2	64634
Rickettsia akari Hartford	50705
* *Rickettsia australis* Cutlack	231019

Spotted Fever Group

Rickettsia endosymbiont of *Ixodes scapularis*	40569
* "*Candidatus* Rickettsia amblyommii GAT-30V"	232978
* *Rickettsia rhipicephali* 3-7-female6-CWPP	233851
Rickettsia massiliae MTU5	83254
* *Rickettsia massiliae* AZT80	238520
Rickettsia heilongjiangensis 054	193551
Rickettsia japonica YH	83739
* *Rickettsia parkeri* Portsmouth	233447
* *Rickettsia montanensis* OSU 85-930	232555
Rickettsia peacockii Rustic	48268
Rickettsia rickettsii 'Sheila Smith'	5337
Rickettsia rickettsii Iowa	59104
* *Rickettsia rickettsii* Brazil	239281
* *Rickettsia rickettsii* Hlp#2	236174
* *Rickettsia rickettsii* Hauke	233768
* *Rickettsia rickettsii* Hino	235723
* *Rickettsia rickettsii* Arizona	236594
* *Rickettsia rickettsii* Colombia	228121
Rickettsia slovaca 13-B	180092
* *Rickettsia slovaca* D-CWPP	233215
* *Rickettsia philipii* 364D	124131
Rickettsia conorii Malish 7	45613
* *Rickettsia conorii* subsp. *indica* ITTR	229600

Rickettsia sibirica 246	27963
* *Rickettsia sibirica* subsp. *mongolitimonae* HA-91	225156
* *Rickettsia sibirica* subsp. *sibirica* BJ-90	238733
Taxon [1]	**ID [2]**
Rickettsia africae ESF-5	6986

Mitochondria (12)

Marchantia polymorpha	**NC_001660**
Malawimonas jakobiformis	**NC_002553**
Rhodomonas salina	**NC_002572**
Nephroselmis olivacea	**NC_008239**
Prototheca wickerhamii	**NC_001613**
Chaetosphaeridium globosum	**NC_004118**
Chara vulgaris	**NC_005255**
Mesostigma viride	**NC_008240**
Reclinomonas americana	**NC_001823**
Physcomitrella patens	**AB251495**
Phytophthora infestans	**AY894835**
Trichoplax adhaerens	**NC_008151**

[1] Taxonomic groups are listed from the top to the bottom as shown in the tree (**Figure 5**). The term "*Candidatus*" is used as originally suggested (Murray and Stackebrandt 1995). *Incertae sedis*, "of uncertain placement", refers to taxa with unknown or undefined broader relationships. *Rickettsia* genomes annotated with RAST (Aziz, Bartels et al. 2008) ahead of the PATRIC release on May 31st, 2012 are noted with an asterisk.

[2] PATRIC (Gillespie, Wattam et al. 2011) genome IDs are given for all bacterial genomes; NCBI accession numbers (bold) are provided for mitochondrial genomes.

[A] The CDS for Rickettsiales endosymbiont of *Trichoplax adhaerens* are provided in **Supplementary Table S2**.

Aziz, R. K., D. Bartels, et al. (2008). "The RAST Server: rapid annotations using subsystems technology." BMC Genomics **9**: 75.

Gillespie, J. J., A. R. Wattam, et al. (2011). "PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species." Infect Immun **79**(11): 4286-4298.

Murray, R. G. and E. Stackebrandt (1995). "Taxonomic note: implementation of the provisional status Candidatus for incompletely described procaryotes." Int J Syst Bacteriol **45**(1): 186-187.

Supplementary Table S4. Sequencing reads from the *Trichoplax adhaerens* genome sequencing project that map to 16S rRNA genes in the Greengenes database (DeSantis, Hugenholtz et al. 2006).

Best Classification	OTU	No. Reads
Host (*T. adhaerens*)		
Trichoplax (Metazoa; Placozoa)	-	255
Alphaproteobacteria		
Rhodobacteraceae (Rhodobacterales)	otu_2666	1
Marivita spp. (Rhodobacterales)	otu_2687	5
Phyllobacteriaceae (Rhizobiales)	otu_2578	1
Rickettsiales	otu_2823	2
Gammaproteobacteria		
Alteromonas marina (Alteromonadales)	otu_3413	1
Alteromonadaceae (Alteromonadales)	otu_3406	1
Colwellia psychrerythraea (Alteromonadales)	otu_3423	1
Oceanospirillaceae (Oceanospirillales)	otu_3776	1
Betaproteobacteria		
Limnobacter spp. (Burkholderiales)	otu_2997	3
Deltaproteobacteria		
Lawsonia intracellularis (Desulfovibrionales)	otu_3222	1
Spirochaetes		
Borrelia spp. (Spirochaetales)	otu_4107	2
Firmicutes		
Ruminococcaceae (Clostridales; Clostridia)	otu_2159	1
Plantomycetes		
CL500-15 (agg27)	otu_2405	1
Cyanobacteria *		
Haptophyceae (haptophyte algae)	otu_1404	7
Straemenopiles (heterokonts)	otu_1406	5
Cryptophyta (cryptomonad algae)	otu_1401	1

***** Chloroplast rDNA of probable cyanobacterial origin.

DeSantis, T. Z., P. Hugenholtz, et al. (2006). "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB." Appl Environ Microbiol **72**(7): 5069-5072.

Supplementary Table S5. Rickettsiales signature genes mined from the *Trichoplax adhaerens* assembly.

Supplementary Table S5 is omitted here due to space considerations; it can be accessed in its entirety in the online manuscript (http://gbe.oxfordjournals.org/content/5/4/621).

Appendix D. Trace read quality control results for *Trichoplax adhaerens* study (Chapter 4).

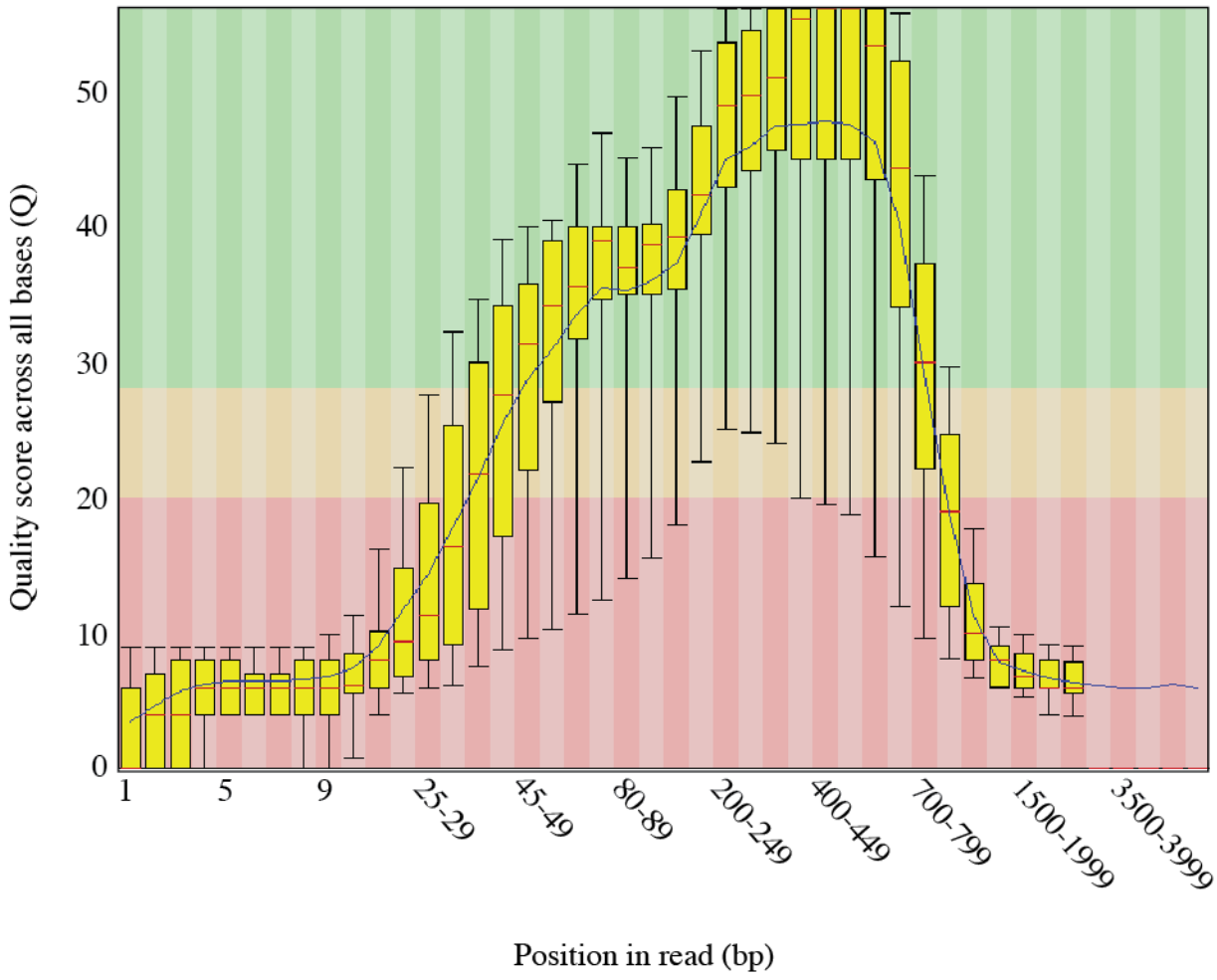


Figure D1. Quality of trace reads from the *Trichoplax adhaerens* genome project, after decontamination (removal of cloning vector) but before application of the quality control pipeline. This illustrates partial output from the `fastqc` program, showing the mean quality score (Q) across all reads at each position. The results plotted in this graph are used to trim low-quality positions (Q<25) from the ends of all reads.

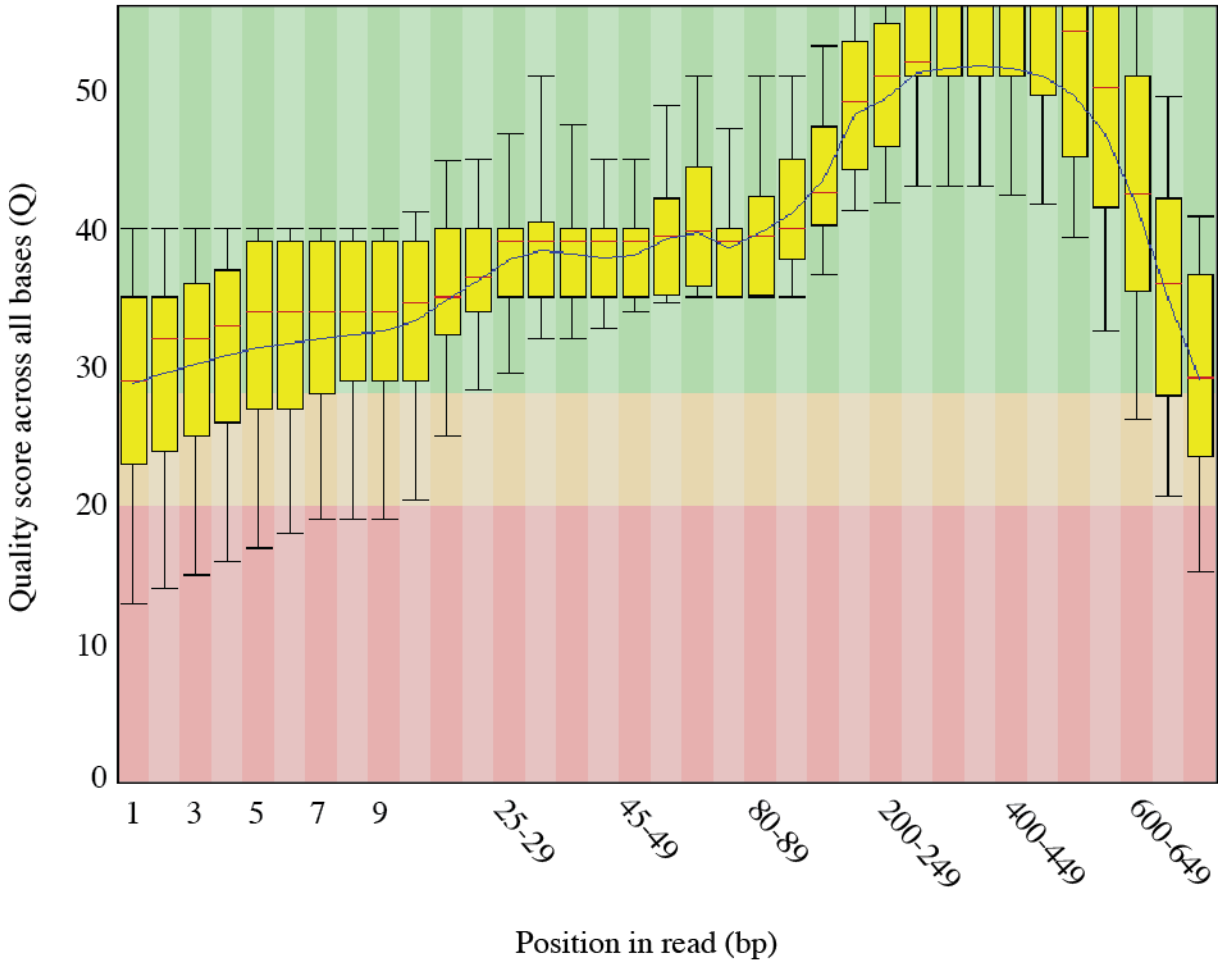


Figure D2. Quality of trace reads from the *Trcihoplax adhaerens* genome project after application of the full quality control pipeline. This illustrates partial output from the `fastqc` program, showing the mean quality score (Q) across all reads at each position.

Appendix E. Supplemental information for *Xenopus tropicalis* study (Chapter 5).

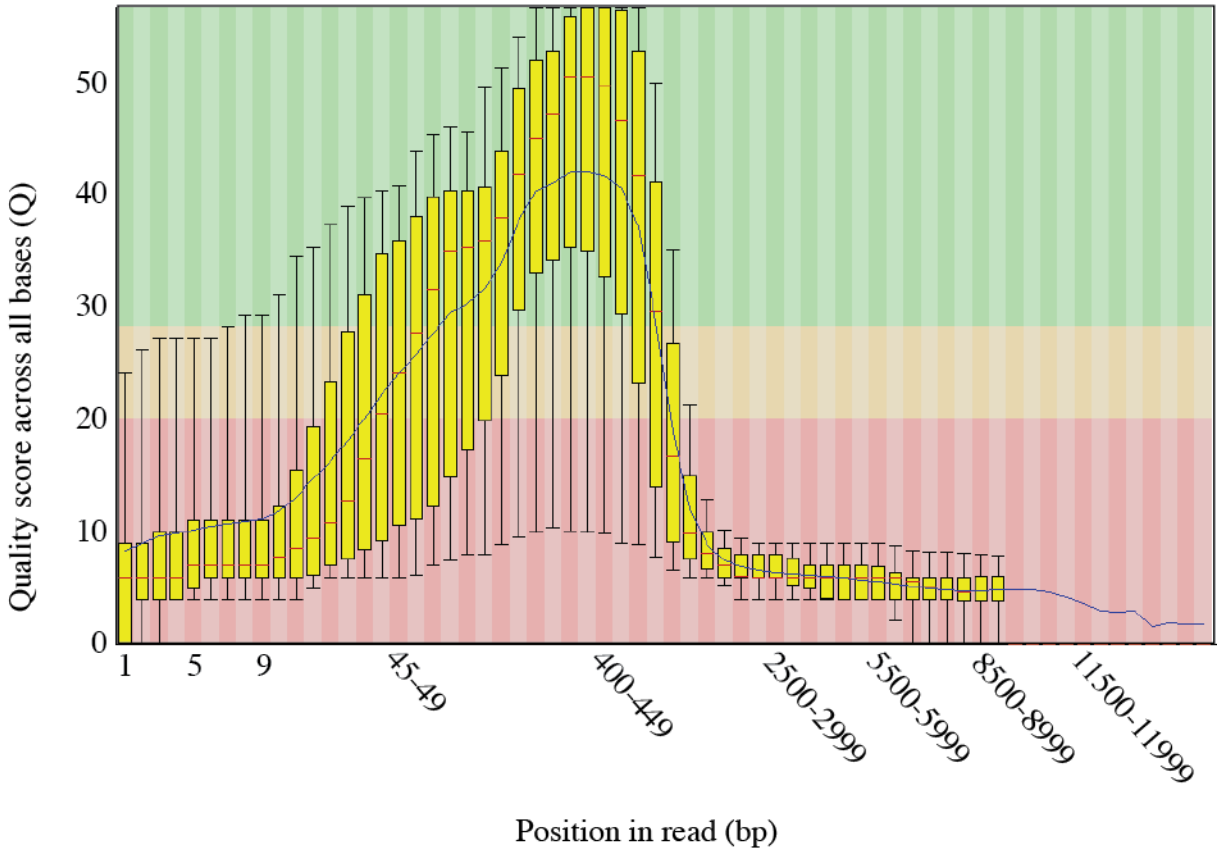


Figure E1. Quality of trace reads from the *Xenopus tropicalis* genome project, after decontamination (removal of cloning vector) but before application of the quality control pipeline. This illustrates partial output from the `fastqc` program, showing the mean quality score (Q) across all reads at each position. The results plotted in this graph are used to trim low-quality positions (Q<25) from the ends of all reads.

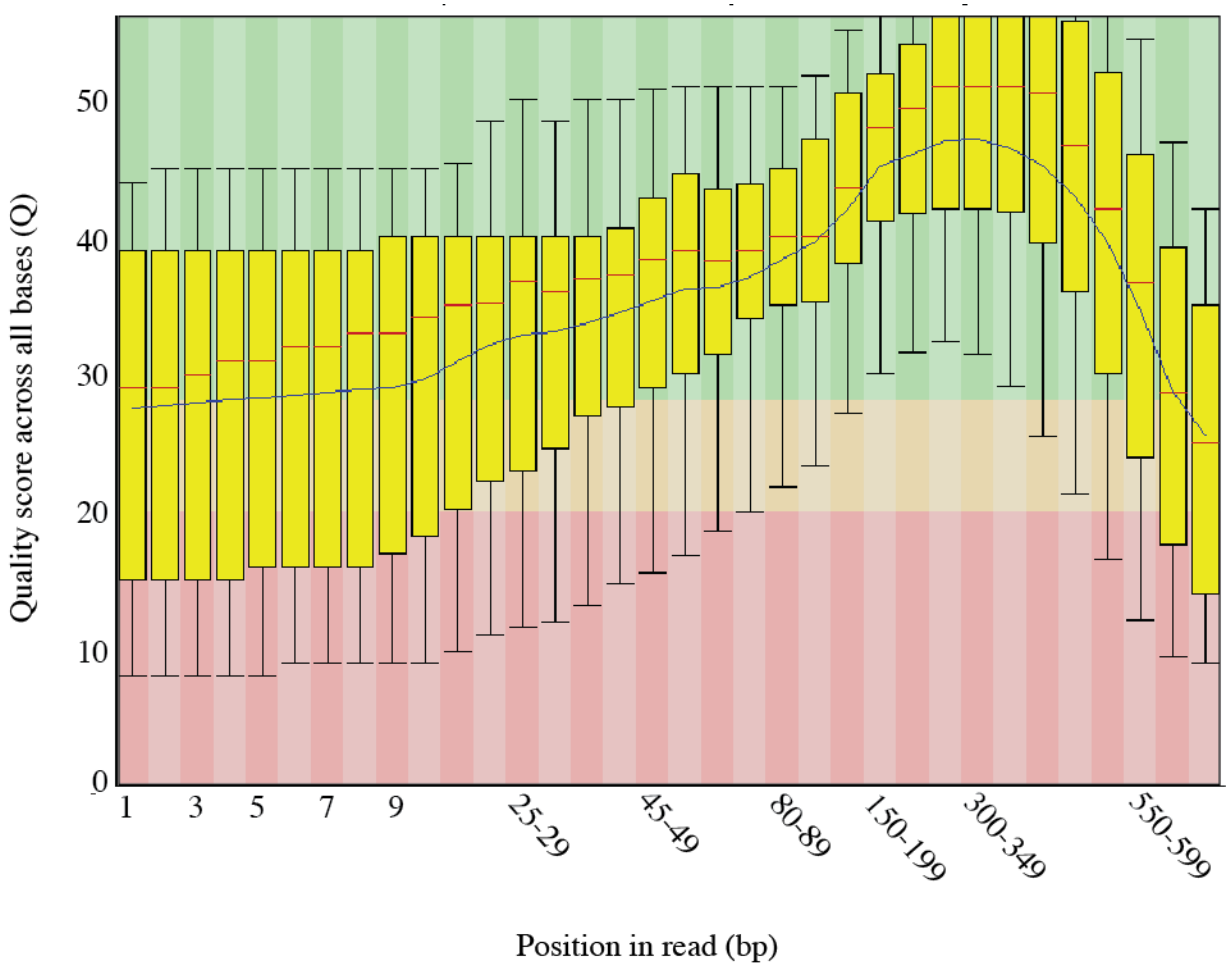


Figure E2. Quality of trace reads from the *Xenopus tropicalis* genome project after application of the full quality control pipeline. This illustrates partial output from the `fastqc` program, showing the mean quality score (Q) across all reads at each position.

GCTCATGAGAAGAGAAGTGAAGTTCACTTCAATTCCGTTTTATGAGTGGAGTCGAGAGACTTTA
AATTTTCAAGATCGAACTGTAGAGTTTGATCCTGGCTCAGATTGAACGCTGGCGGCATGCCTTA
CACATGCAAGTCGAACGGTAACAGGTCTTCGGACGCTGACGAGTGGCGAACGGGTGAGTAACAC
ATCGGAACGTACCCAGACGTGGGGGATAACGAGGCGAAAGCTTTGCTAATACCGCATGATATCT
GAGGATGAAAGCAGGGGACCGCAAGGCCTTGCGCGTTTGGAGCGGCCGATGGCAGATTAGGTAG
TTGGTGAGATAAAAGCCCACCAAGCCGACGATCTGTAGCTGGTCTGAGAGGACGACCAGCCACA
CTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTGGGGAATTTTGGACAATGGGC
GCAAGCCTGATCCAGCAATGCCGCGTGCAGGATGAAGGCCCTCGGGTTGTAAACTGCTTTTGTA
CGGAACGAAAAGGCTCTGATGAACAATTGGGGTTTCTGACGGTACCGTAAGAATAAGCACCGGC
TAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGTGCGAGCGTTAATCGGAATTACTGGGCGT
AAAGCGTGCGCAGGCGGTTTTGTAAGACAGAGGTGAAATCCCTGGGCTCAACCTGGGAACTGCC
TTTGTGACTGCAAAGCTGGAGTGCGGCAGAGGGGGATGGAATTCCGCGTGTAGCAGTGAAATGC
GTAGATATGCGGAGGAACACCGATGGCGAAGGCAATCCCCTGGGCCTGCACTGACGCTCATGCA
CGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCCTAAACGATGTCAACT
GGTTGTTGGGAATTTGCTTTCTCAGTAACGAAGCTAACGCGTGAAGTTGACCGCCTGGGGAGTA
CGGCCGCAAGGTTGAAACTCAAAGGAATTGACGGGGACCCGCACAAGCGGTGGATGATGTGGTT
TAATTCGATGCAACGCGAAAAACCTTACCCACCTTTGACATGTACGGAATCCTGAAGAGATTTA
GGAGTGCTCGAAAGAGAGCCGTAACACAGGTGCTGCATGGCTGTCGTCAGCTCGTGTCGTGAGA
TGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTGCCATTAGTTGCTACGAAAGGGCACTCTAA
TGGGACTGCCGGTGACAAACCGGAGGAAGGTGGGGATGACGTCAAGTCCTCATGGCCCTTATAG
GTGGGGCTACACACGTCATACAATGGCTGGTACAAAGGGTTGCCAACCCGCGAGGGGGAGCCAA
TCCCACAAAGCCAGTCGTAGTCCGGATCGCAGTCTGCAACTCGACTGCGTGAAGTCGGAATCGC
TAGTAATCGCGAATCAGAACGTCGCGGTGAATACGTTCCCGGGTCTTGTACACACCGCCCGTCA
CACCATGGGAGCGGGTTCTGCCAGAAGTAGTTAGCCTAACCGCAAGGAGGGCGACTACCACGGC
AGGGTTCGTGACTGGGGTGAAGTCGTAACAAGGTAGCCGTATCGGAAGGTGCGGCTGGATCACC
TCCTTTCTGGAAAACAGCTGCGCAAAATTAAACGCCCACACTTATCGGCTGTAAACACA

Figure E3. Complete nucleotide sequence (1,659 bases) of the XTAB 16S rDNA sequence mined from *Xenopus tropicalis* trace reads using MetaMiner.