Microbial profiling using metagenomic assembly

Submitted to the College of Graduate Studies and Research of the University of Saskatchewan in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Veterinary Microbiology at the University of Saskatchewan.

By

MATTHEW GRAHAM LINKS

PERMISSION TO USE

In agreement with the outlines set out by the College of Graduate Studies and Research at the University of Saskatchewan, I allow the University of Saskatchewan Libraries to make this thesis available to all interested parties. Also in accordance with the College of Graduate Studies and Research, I allow this thesis to be copied "in any manner, in whole or in part, for scholarly purposes". This thesis may not, however, be reproduced or used in any manner for financial gain without my written consent. Any scholarly use of this thesis, in part or in whole, must acknowledge both myself and the University of Saskatchewan.

Any requests for copying or using this thesis, in any form or capacity, should be made to:

Head of Department of Veterinary Microbiology

University of Saskatchewan

Saskatoon, Saskatchewan

S7N 5B4

ABSTRACT

The application of second generation sequencing technology to the characterization of complex microbial communities has profoundly affected our appreciation of microbial diversity. The explosive growth of microbial sequence data has also necessitated advances in bioinformatic methods for profiling microbial communities. Data aggregation strategies should allow the relation of metagenomic sequence data to our understanding of microbial taxonomy, while also facilitating the discovery of novel taxa.

For eukaryotes, a method has been established that links DNA sequences to the identification of organisms: DNA Barcoding. A similar approach has been developed for prokaryotes using target genic regions as markers for species identification and to profile communities. A key difference in these efforts is that within DNA barcoding there is a formalized framework for the evaluation of barcoding targets, whereas for prokaryotes the 16S rRNA gene target has become the *de facto* barcode without formal evaluation. Using the framework developed for evaluating DNA barcodes in eukaryotes, a study was undertaken to formally evaluate 16S rRNA and *cpn*60 as DNA barcodes for Bacteria. Both 16S rRNA and *cpn*60 were found to meet the criteria for DNA barcodes, with *cpn*60 a preferred barcode based on its superior resolution of closely related taxa.

The high resolution of *cpn*60 enabled a method of sequence data aggregation through sequence assembly: microbial profiling using metagenomic assembly (mPUMA). The scoring of metagenomic assemblies in terms of sensitivity and specificity of the operational taxonomic units formed was used to evaluate and optimize the assembly of *cpn*60 barcodes. Using optimized parameters, mPUMA was demonstrated to faithfully

reconstruct a synthetic community in terms of richness and abundance. To facilitate the use of mPUMA, a software package was developed and released under an open source license.

The utility of mPUMA was further examined through the characterization of the epiphytic seed microbiomes of *Triticum* and *Brassica* species. A microbiome shared across both crop genera including fungi and bacteria was detected: a particularly important observation as it implies that seeds may serve as a vector for microbes that could include both pathogenic and beneficial organisms. The relative abundances of taxa identified by mPUMA were confirmed by qPCR for multiple cases of both fungal and bacterial taxa. By culturing isolates of both bacteria and fungi from the seed surfaces it was demonstrated that mPUMA faithfully assembled consensus sequences for OTUs that were 100% identical to isolated fungi and bacteria. Patterns observed in the relative abundances of the shared microbiome OTUs were used to generate the hypothesis that an *Pantoea-like* bacterium and an *Alternaria*-like fungus had an antagonistic relationship, since sequences corresponding to these organisms showed reciprocal abundance patterns on *Triticum* and *Brassica* seeds. Studies of the interactions of cultured isolates revealed fungistatic interactions that could account for their reciprocal abundances. These interactions could be directly relevant to plant health, given that *Alternaria*-like fungi are linked to grain spoilage in wheat, and diseases in canola.

Taken together, results of this thesis demonstrate the superiority of the *cpn*60 universal target as a barcode for Bacteria, forming the basis for an assembly-based strategy for microbial profiling of bacterial and eukaryotic microbial communities that can lead to the discovery of novel taxa and microbial interactions.

ACKNOWLEDGEMENTS

I am very grateful to have been a part of numerous multidisciplinary teams. These teams were instrumental in my ability to complete the works presented in this thesis. These teams also provided me opportunities to be a participant in all aspects of the scientific process and for that I am truly grateful.

Having been part of the Hill lab has been an honor. I am thankful for the support they have given me. I am also thankful of the opportunities I have had to work with them in our mutual pursuit of Science.

My committee has been a resource for me both in terms of this thesis as well as in a larger mentorship role. I am acutely aware and thankful for the guidance that they provided me while pursuing this degree. I also wish to note that their willingness to mentor me in Science is something that I am indebted to them for.

To my supervisor I will forever be measuring myself by the standard that she has and continues to set for a Scientist. I am appreciative that she was willing to take me on as a student and for her patience as the overall theme of my thesis coalesced. Looking back at the times of frustration and joy doing Science I can see that she was always able to strike the right balance in mentoring me through this journey. Thank you.

My late friend Steve, as I submit these final documents for my thesis I do feel my feet.

Thank you to my family, especially my son Aidan and wife Paige. Your support and ability to remind me why I am going through this process was key to my success. I am

glad to be part of a team that measures not whether I am a successful scientist but whether I am a happy one.

Soli Deo Gloria

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

CHAPTER 1 - INTRODUCTION AND LITERATURE REVIEW

Communities of microbes

Estimates have placed the number of prokaryotic cells on earth at 10^{30} and suggested that the amount of carbon within those cells is similar to the amount of carbon found in all plant life on this planet (Whitman *et al.* 1998). Microbes naturally exist in environmental settings where they interact with each other, and in some cases a host. Interactions can be both direct and indirect and may result in dependencies between members of a community. Based on direct morphological observation it is clear that in most cases microbial communities are comprised of multiple distinct organisms. In fact it is largely only through some form of enrichment culture technique that a community of microbes be distilled to a sample comprised of a single discrete type of microbe. It is necessary to study microbes together in order to understand how they can perform specialized functions as a community. This examination of microbes as populations of organisms that are interacting with one another is Microbial Ecology.

The concepts of Microbial Ecology were born out of studies of Plant and Animal systems where a community is a collection of organisms that interact with one another (Konopka 2009). A fundamental question in ecological studies, microbial or otherwise, is *who is there*? This organismal richness is the alpha diversity of the community (Whitaker 1960).

Given that ecological studies aim to understand how a community of microbes functions as a whole, assumptions have to be made that members of the same species will function in the same manner, generally speaking. If each type of microbe within a community can perform a series of functions then the number of functions that need to be understood,

modeled or predicted in an ecological study is largely based in terms of the number of distinct organisms present.

An example of the linkage between community composition and function can be seen in the biosignatures left by microbial communities as they form sedimentary deposits in stromatolites (Baumgartner *et al.* 2009). A biosignature is an organomineral deposit that contains chemical characteristics of microbial origin. These origins could be the microbes themselves (e.g. cellular debris) or extra-cellular polymeric substances (EPS), which are the products of microbial activities. Baumgartner *et al.* demonstrated that changes in the richness of microbial communities could be linked to differences in EPS deposit characteristics (Baumgartner *et al.* 2009). As EPS deposits are derived from the activity of the microbes within the bacterial mats this demonstrates the connection between metabolic functions within a microbial community and the diversity of the community itself. In a microbial setting, the need to first understand the taxonomic richness and thereby composition is paramount to any understanding of community function.

Microbial ecology also plays a crucial role in determining plant health and disease. Plants can affect the properties of the soil through alteration of humidity, pH and oxygen levels. Soil conditions affect survival of microbes, which in turn can affect nutrient availability to the plant and be the source of diseases. Recent work (Lundberg *et al.* 2012) has demonstrated the use of microbial profiling to examine the root microbiome of plants. The work of Lundberg *et al.* (2012) established that specific communities of microbes living in close association with plants could be reproducibly found in association with particular species of plants. Further it was demonstrated that soil type and genotype of the host plant could affect the composition of microbial communities that closely associated

with plant roots. The interactions between a host organism and its associated microbiota need to be well characterized in order to understand how these interactions can modulate the host's health.

Culture independent methods

Prior to the adaptation of molecular methods in microbiology, characterization of organisms relied heavily on the isolation of strains as pure cultures. Once isolated, a microbe could be assessed in terms of phenotypic characteristics (e.g. morphology) and the phenotype derived from these observed characteristics provided systems for classification across isolates (Kampfer and Glaeser 2012). Culture dependent methods are extremely valuable in microbiology as they provide phenotypic information discretely linked to single isolates. Unfortunately, current culture-based methods have proven deficient in terms of their ability to access all microbial life present in complex communities. This discrepancy between numbers of microbial cells present in complex communities and those recoverable within culture has been termed *the great plate count anomaly* (Staley and Konopka 1985).

A partial solution to the limitations of culture-based methods arose from the theory of a molecular clock (Zuckerkandl and Pauling 1965). The basis for the molecular clock theory is that if two organisms shared a common evolutionary ancestor then the sequence for a gene (conserved within both organisms) would have acquired mutations providing a record of evolution from the common ancestor. Thus by deriving the differences between organisms for a conserved gene (through sequence alignment) it is possible to infer a phylogenetic lineage for the evolution of those organisms (Fitch and Margoliash 1967).

From a practical standpoint, the sequence-derived phylogeny could be used as a tool within a taxonomic system of classification.

Carl Woese and George Fox are largely responsible for moving microbiological taxonomy in a molecular direction (Woese and Fox 1977). Their work described life in terms of three domains and was based on phylogenetic analysis of ribosomal RNA genes. It was perhaps the direct way that Woese & Fox demonstrated the connection of nucleic acid sequences to phylogenetic relationships among microbes that empowered a paradigm shift in microbiology to connect the field directly with sequence data. By connecting Fitch's work on molecular evolution (Fitch 1976; Fitch and Margoliash 1967) to an understanding of taxonomic relationships between microbes, Woese and Fox promoted a sequence-based understanding of taxonomy. Sequence-derived phylogenies have fundamentally changed prokaryotic taxonomy in that it has led to the adoption of a 16S rRNA framework within *Bergey's Manual of Systematic Bacteriology* (2001).

While the adoption of sequence information into taxonomic frameworks is now commonplace, there still remains a need for a polyphasic (genotypic, phenotypic, and phylogenetic) approach to bacterial taxonomy (Vandamme *et al.* 1996). By relying on multiple characteristics to determine taxonomic relationships it is hoped that the strengths of one characteristic can compensate for weaknesses in another. For example, when a biochemical property of two organisms overlaps they may still be resolvable on the basis of the sequence of a gene they have in common. While a polyphasic approach to classification can be seen as stronger and more robust than any of the single characteristics on their own, it is crucial to maintain the understanding that any criteria

used for classification will have limits. Therefore it is important to establish classification criteria where limits can be both expressed and tested.

Application of sequence-based identification to complex communities of microbes is largely attributed to the work of Norman Pace and colleagues (Stahl *et al.* 1985). Focusing on a demonstration of the enumeration of microbes present in a hot spring, Stahl *et al.* were able to show the adaptation of modern DNA sequencing advances to the identification of microbes in a complex community. Recognizing that a large fraction of ribosomal RNA could be extracted directly from a gentle lysis, Pace's group was able to demonstrate the acquisition of nucleic acid directly from a naturally occurring microbial community. The group also recognized that through direct sequencing of the rRNAs the richness of the community could be estimated by the number of discrete rRNAs sequenced. Further, they began to explore additional concepts of microbial diversity through the abundance of each distinct rRNA, noting that issues like unequal incorporation efficiencies of the radio-labels would skew abundance estimates. Additional work by Lane *et al.* also examined the use of probes targeting 16S rRNA to directly address quantification and visualization (Lane *et al.* 1985).

With increasing recognition that many microbes lack morphologically informative differences and that enrichment based culture techniques bias the view of a community (Giovannoni *et al.* 1990; Ward *et al.* 1990); there was a need to develop methods that were more universal in terms of their ability to identify and potentially quantify microbes using molecular data. A key development, made possible by the invention of the polymerase chain reaction (PCR), was the use of universal primers to anchor molecular data to a location within the gene of interest. This targeting of specific regions can be

seen in the demonstration of Weller and Ward, in which a conserved region (1392 bp - 1406 bp) of the *Escherichia coli* 16S rRNA gene was used for surveying the community of microbes present in a hot-spring (Weller and Ward 1989).

Microbial profiling using 16S rRNA gene sequences

Weller and Ward's work ushered in a new era of microbial community profiling based on sequencing of marker genes PCR amplified from microbial communities. These studies rely on the use of an informative molecular target, which in order to be useful, must also be highly conserved in the population under study. While some genes may be informative when looking at specific, related taxa (e.g. *gyrB* for identification of *Campylobacter* spp. (Kawasaki *et al.* 2008) and *mcrA* for the identification of methanogenic Archaea (Gagnon *et al.* 2011)), there is a larger need within microbial profiling for the conservation to extend to all possible taxa being studied.

The 16S rRNA gene was the first gene used for microbial profiling (Lane *et al.* 1985) and it remains the most widely used gene today. This is a natural extension from the works of Woese & Fox and Ward & Miller (Ward *et al.* 1990; Weller and Ward 1989; Woese and Fox 1977) as they collectively demonstrated that this gene is informative, could be used to distinguish between the Archaea and Bacteria, and could be amplified and cloned from naturally occurring communities using broad range primers. The 16S rRNA gene is an approximately 1.5 kb (*Escherichia coli* (Brosius *et al.* 1978)), highly conserved, gene that encodes a structural RNA and forms part of the small ribosomal subunit. The secondary structure of the 16S rRNA is crucial for its catalytic function. Encoded within the sequence of the 16S rRNA gene is the information that defines its secondary structure as a series of nine hyper-variable regions alternating with highly conserved regions

(Chakravorty *et al.* 2007). Thus when mutations arise within the 16S rRNA gene they can have a direct effect on the secondary structure by affecting intra-strand base pairing. With respect to its application to microbial profiling, both variable and conserved regions are fundamentally important to the utility of the gene. In order to make the amplification of targeted genic regions feasible from communities of microbes there needs to exist a universal primer set and amplification conditions which allow the recovery of the specific region from all taxa in a mixed population simultaneously. The highly conserved regions allow for "universal" primers to target the same sequence with very minor variation across distinct taxa. When amplification is carried out using primers targeting the conserved regions the resulting amplicon sequence contains one or more hyper-variable regions. The hyper-variable regions provide informative sequence data that can be used to distinguish and identify taxa. There have been numerous primer sets designed for the 16S rRNA gene, varying in their scope in terms of the hyper-variable regions they cover and the taxa to which they have been applied. Recent studies examining published primer sets in terms of their broad applicability across phyla recognized the presence of hundreds (> 500) of primer pairs that could be used in 16S rRNA studies (Klindworth *et al.* 2013). The numerous options for primers actually present a limitation for studies targeting 16S rRNA because the ability to compare across studies is impeded by targeting of different regions within the gene. Additional problems in the use of 16S rRNA can arise due to the level of conservation amongst the highly conserved regions of the gene. Conserved regions can provide sites for chimera formation through aborted extension and mis-priming, which may occur when performing PCR (Haas *et al.* 2011).

Classical bioinformatics tools (e.g. global and local alignments) commonly use scoring strategies that may be inappropriate for the alignment of 16S rRNA data due to it being a structural RNA gene. The relationship between secondary structure and function of the 16S rRNA gene has prompted the derivation of alignment methods that use secondary structure information to calculate similarity (Nawrocki *et al.* 2009). Other methods for alignment of 16S rRNA sequences such as NAST (DeSantis *et al.* 2006) produce alignments against a reference template and are prone to containing large numbers of gaps making the alignments difficult to interrogate visually. Despite these limitations, the 16S rRNA target has been adopted almost universally for microbial profiling, and its status as the first and most commonly used gene for microbial profiling has resulted in a huge wealth of sequence data housed in databases such as RDP (Cole *et al.* 2007).

Protein-coding alternative genes for microbial profiling

In addition to 16S rRNA, protein-coding genes can also be targeted for microbial profiling. The degeneracy of the genetic code means that there is the potential for synonymous mutations within its DNA sequence. Thus protein-coding genes can accumulate some mutations (the synonymous ones), which do not affect the function of the encoded protein. Information about each amino acid is encoded by a codon in the sequence of protein-coding genes. For protein-coding genes there is thus a more direct connection between sequence of the gene and its function than there is for a structural RNA (e.g. 16S rRNA) where the secondary structure is not as easily predicted from the sequence. The sequence divergence rates of protein-coding genes are thereby more appropriate as a molecular clock than structural RNA-coding genes, and are a better choice for reconstruction of phylogenetic lineages. Protein-coding genes can be used in

terms of either their DNA or amino acid sequences, offering two levels of interpretation of sequence relationships. Only two protein-coding genes have been described for use in microbial profiling: *rpo*B (Mollet *et al.* 1997) and *cpn*60 (Goh *et al.* 1996).

*rpo*B is a universally conserved gene that encodes the beta sub-unit of the bacterial RNA polymerase required for RNA synthesis. As a single copy gene (Case *et al.* 2007), *rpo*B was originally proposed as an alternative to 16S rRNA for denaturing gradient gel electrophoresis (DGGE) studies (Dahllof *et al.* 2000). Given that Dahllof *et al.* were focused on the use of *rpo*B for microbial profiling using DGGE, the primers they designed contained no degeneracy and would be predicted to have a limited taxonomic range. In analyses of whole genome data from GenBank, Case *et al.* (2007) were able to illustrate that there are some cases where *rpo*B could identify a monophyletic lineage (e.g. *Firmicutes*) while 16S rRNA could not. The lack of a validated universal PCR system for the amplification of *rpo*B limits its adoption for microbial profiling even though it has been used in conjunction with pyrosequencing to study microbial diversity in soil (Vos *et al.* 2012).

*cpn*60 (also known as *groEL* or *hsp*60) encodes the type-I molecular chaperone, a 60 kDa protein which assists in the folding and stability of protein structures within the cell, and is highly conserved due to its essential function (Hemmingsen *et al.* 1988). A PCR system has been developed for the amplification of a universal target region of the *cpn*60 gene (corresponding to nucleotides 274-828 of the *E. coli cpn*60 gene) (Goh *et al.* 1996). The original primers (H279 and H280) for amplification of the universal target region of *cpn*60 are highly degenerate (2^{19} and 2^{18} respectively) and multiple inosines (9 and 6 respectively. The degenerate nature of the *cpn*60 universal target is a significant

difference from 16S rRNA where primers can be based on highly conserved regions that flank one or more hyper-variable regions. The universal amplification system for *cpn*60 has been optimized for difficult templates (Hill *et al.* 2006b) and the recovery of under-represented taxa (Hill *et al.* 2010).

Initial applications of the universal target region for *cpn*60 were for species level identification (Goh *et al.* 1998; Goh *et al.* 2000; Goh *et al.* 1997; Hill *et al.* 2006a). Additional work found that *cpn*60 routinely provided a higher level of resolution when compared to 16S rRNA and other targets (Verbeke *et al.* 2011). Resolution at the subspecies level has also been demonstrated (Brousseau *et al.* 2001; Paramel Jayaprakash *et al.* 2012; Vermette *et al.* 2010). Identification of samples using *cpn*60 relies heavily on the existence of cpnDB, a database that houses chaperonin sequence data from a broad range of taxa (Hill *et al.* 2004). With a large amount of data available in cpnDB covering a wide breadth of taxa, it is possible to assess cross-reactivity of nucleic acid probes, *in silico* and assess suitability for diagnostic uses (Chaban *et al.* 2009; Chaban *et al.* 2010; Dumonceaux *et al.* 2009; Dumonceaux *et al.* 2011).

Applications of *cpn*60 to microbial profiling were originally based upon di-deoxy terminator sequencing methods. A wide range of microbial communities were surveyed using these methods including industrial settings (Dumonceaux *et al.* 2006c), the intestinal / fecal communities of various animals (Desai *et al.* 2009; Dumonceaux *et al.* 2006b; Hill *et al.* 2005b; Hill *et al.* 2002) including studies of humans (Hill *et al.* 2010) as well as studies relating to the human vaginal microbiome (Hill *et al.* 2005a). With the public availability of pyrosequencing in the form of Roche / 454's GS FLX platform *cpn*60 was used in next-generation sequencing studies of the human vaginal microbiome

(Schellenberg *et al.* 2009; Schellenberg *et al.* 2011a; Schellenberg *et al.* 2011b) and studies in fish (Desai *et al.* 2012) and dogs (Chaban *et al.* 2012).

An advantage *cpn*60 holds over 16S rRNA is that it is conserved across all domains of life. There is conservation of *cpn*60 among bacteria and eukaryotes as noted in the original characterization by Hemmingsen *et al.* (1998). More recently, the conservation has been shown to extend to some Archaea that have a type-I chaperone. However the majority of Archaea only possess a type-II chaperone with a small number containing both a type-I and type-II chaperone (Large *et al.* 2009). Chaban and Hill proposed a universal system for amplification of the type-II chaperone, which is highly analogous to the amplification system for *cpn*60 (Chaban and Hill 2012).

A number of gene targets have potential for microbial profiling (16S rRNA, *rpo*B and *cpn*60), but there is currently no accepted framework for evaluating gene targets for microbial profiling. This leads to a need for a systematic mechanism through which gene targets could be proposed and evaluated in order identify strengths and denote weakness. One example of such a mechanism is found in the DNA Barcoding movement, where a formalized system for the selection and validation of DNA sequences for identification of eukaryotes has been developed.

DNA Barcoding

Hebert *et al.* raised concerns about the growing need for taxonomists to characterize eukaryotic species on the basis of morphology alone (Hebert *et al.* 2003). There was a series of issues that Hebert *et al.* identified with morphological based studies. Chiefly, the use of morphology alone is problematic given that plasticity of morphologic

characteristics can lead to inaccurate classification. Issues with morphological characteristics can arise from developmental linkages of the characteristics themselves or cryptic taxa. Also, increasingly expert-level knowledge was required to recognize taxonomically relevant characteristics as characterization continually refined species. Recognizing the increasing use of genomic characteristics to delineate taxa with application to microbes, Hebert *et al.* set out to define a system that could apply broadly to all life. The proposed solution of Hebert *et al.* was to adopt the use of informative DNA sequences from vouchered samples as barcodes for the organisms. These DNA barcodes could thus be used to identify organisms within a second collection without the need to repeat all of the morphological study present in the original voucher.

From a collections perspective, DNA barcoding is fundamentally liberating. If each organism within a collection were identifiable to the species level using a DNA sequence then dependence on expert knowledge in each and every domain could be removed, or at least mitigated. Thus collections could be initially curated on the basis of a conserved DNA barcode and the sequence data could represent an initial measure of the richness of the collection itself. By no means was DNA barcoding proposed to eliminate the need for expert knowledge, but rather it was meant as a way to free experts from routine comparisons and allow taxonomic experts to focus on the unique aspects of collections or provide an initial starting point for the identification of new specimens.

In the original description of DNA barcodes, Hebert *et al.* identified the cytochrome oxidase I (COI) gene as a candidate barcode. For eukaryotes it was suggested that mitochondrial genes like COI would be better barcodes given the mitochondrial genome accumulates mutations at a faster rate than the nuclear genome. Ribosomal genes were

considered, but discounted due to indels and the complications that these structures introduce to sequence alignments. Of the protein-coding gene options available, Hebert *et al.* acknowledged that any of the genes encoded in the animal mitochondrial genome could be suitable, but pointed out two key factors that make COI advantageous. There was a universal PCR amplification system for COI, and the distance between COI sequences of related taxa was greater than the other genes considered. These advantages have lead to the framework for determining selection of DNA barcodes.

An evaluation scheme is essential since although COI has been widely used as a barcode across many taxa, there are inevitably cases where COI will not perform as an effective barcode. The CBOL (Consortium for the Barcode of Life) recognized cases where an alternative barcode may be considered: when COI alone does not provide sufficient discriminating information to resolve taxa, or when the research community has established *significant* amounts of data on the use of another gene. In order to deal with these situations, the CBOL has established criteria for the selection of a non-COI barcode. In cases where a non-COI barcode is to be proposed through the CBOL, the proponents must provide a documented argument as to why COI is inappropriate for the specific taxa under consideration. Requirements for a non-COI barcode include its accessibility in all taxa under study (usually through a universal PCR protocol for the target region), and that it be able to resolve species level relationships.

An example of where COI is an inadequate barcode can be seen in the Kingdom Viridiplantae (Hollingsworth *et al.* 2011). The substitution rate of the mitochondrial genome in plants is lower than in other eukaryotes and thus makes COI less informative as a barcode for plants. This lower substitution rate within the mitochondrial genome has

led groups to look for alternatives to COI for plants. The best description of the state of DNA barcodes for plants would be *contentious*. Several groups of researchers have made proposals for combinations of genes, many of which overlap across proposals, and there have also been proposals of single genes for barcoding in plants (reviewed in Hollingsworth 2011). Taken together there is significant, ongoing debate around the establishment of a preferred plant barcode. There have been formal, international efforts to assess many of these genes, leading to the discounting of some non-COI candidates. While not formally a consensus, there is a working *majority preference* that a two-gene (*rbcL* + *matK*) strategy be used and augmented as necessary.

There has been a rapid adoption of DNA barcoding across all fields involving biological collections. These adoptions have been highly significant leading to international engagement of 25 nations that have signed agreements to catalogue biodiversity with barcodes and the establishment of the International Barcode of Life Project (iBOL). Additional consortia have been formed around specific thematic areas of DNA barcoding. For example the CBOL is not involved in the process of data acquisition, but rather in the dissemination of knowledge about the methods and procedures for DNA barcoding. CBOL is chiefly responsible for creating working groups to address specific questions arising from the field (e.g. determining a framework for evaluating candidate barcodes).

Data from conventional DNA barcoding is derived from di-deoxy terminator sequencing. These data exist as sequencing traces from a single vouchered specimen whose DNA was subjected to PCR amplification and sequencing. Both conceptually and logistically this is analogous to the use of marker genes for microbial ecology. There are, however, two key differences between these approaches: DNA barcoding employs a formal evaluation

framework to select barcode targets, and microbial ecology is chiefly interested in measuring sample richness directly from complex communities rather than identification of isolates.

For studies of microbial life there has been only a single evaluation of a potential DNA barcode for fungi (Schoch *et al.* 2012). It was demonstrated that the ribosomal internal transcribed spacer (ITS) could be used to identify and differentiate fungi and was thereby a reasonable choice as a DNA Barcode. While not a demonstration of microbial profiling, Schoch *et al.* illustrated the highly similar application of DNA barcoding to the identification of microbial eukaryotes.

Second Generation Sequencing and microbial profiling

Initial applications of DNA sequencing for microbial profiling were very low throughput in terms of the numbers of samples examined and sequencing reads obtained (e.g. < 10 sequences) (Giovannoni *et al.* 1990; Ward *et al.* 1990). The limiting throughput of early studies kept most experimental designs limited to small, descriptive studies of single samples of microbial communities. The adaptation of Capillary Array Electrophoresis to di-deoxy sequencing (CAE) in the 1990's provided a fundamental shift in sequencing technology that enabled significant numbers of sequences to be obtained for a given experiment. The use of an array format of parallel capillaries mated to electronics which could simultaneously resolve fluor-labeled di-deoxy nucleotides meant that the throughput of a single sequencing reaction was expanded by the number of parallel capillaries in the array. Using CAE it was thus possible to capitalize on microtitre plate formats and sequence in units based around a standard 96 well plate. It was this coupling

of automated DNA sequencing (in parallel) to PCR amplicons of informative DNA / RNA genes that enabled microbial ecologists to begin looking deeper into communities.

More recently, a number of "next-generation" technologies for DNA sequencing have become available. By far the most significant of these in terms of its impact on microbial ecology is pyrosequencing. Pyrosequencing is a sequencing methodology based on coupling three biochemical reactions (Margulies *et al.* 2005). DNA dependent DNA polymerase releases pyrophosphate as a result of the incorporation of dNTP into the newly synthesized strand. The free pyrophosphate can then be consumed by ATP sulphurylase along with adenosine 5' phosphosulphate to produce ATP. Lastly, luciferase is capable of consuming ATP in order to convert luciferin to oxyluciferin and results in the production of light. Through this enzymatic cascade, the incorporation of free dNTP can be translated into an observable signal, which is the emission of light. Thus by cycling the availability of single dNTP moieties it becomes possible to determine the sequence of a DNA molecule during its synthesis.

The data derived by pyrosequencing is collected as floating-point numbers, which represent measurements of the light emitted during a single dNTP flow and the name of the moiety that was available to the polymerase. Given that dNTPs are cycled in a repetitive fashion (e.g. A, G, C, T, A, G, C, T, A, …) it is necessary but trivial to exclude the flows that do not result in an incorporation event (i.e. have an amount of emitted light ~ 0). The difficulty in interpreting pyrosequencing data comes from the occurrence of homopolymeric stretches in the DNA template being sequenced. As a contiguous stretch of a single repeating dNTP moiety, homopolymers result in multiple incorporation events within a single flow. Thus the amount of light emitted is > 1 and an interpretation of how

close the floating-point value is to an integer value determines the accuracy of correctly interpreting the length of the homopolymer. The length of pyrosequencing data generated with 454/Roche's GS 20 platform (2005) was < 100 bp. This original read length was so dramatically shorter than conventional di-deoxy based sequencing (700-800bp) it was only appropriate for re-sequencing applications where data was mapped on to some pre-existing reference dataset. Read lengths beyond 200 bp became possible with the release of the FLX, Titanium and FLX+ platforms, and were long enough to enable initial applications to microbial ecology (Schellenberg *et al.* 2009).

Prior to the release of pyrosequencing instruments, there was an increasing application of di-deoxy based sequencing to microbial profiling (Hill *et al.* 2002). While sequencing using di-deoxy terminator chemistries and CAE machines was revolutionary, it limited *high throughput* sequencing studies by being based on 96-well microtitre plate. The picotitre plates commonly used for pyrosequencing feature wells of 50 microns in diameter and yields 1 million reads per sequencing run.

A throughput of 1 million reads / run has enabled significantly more complex experimental designs. Studies including biological and technical replicates, multiple treatment levels and time-courses have become feasible, allowing researchers to move beyond anecdotal, observational studies consisting of mere "snapshots" of communities. All of these possibilities, and the correspondingly large volume of sequence data generated increased the need for bioinformatics methods that aggregate these data in automated ways. In the context of microbial ecology, data aggregation commonly is performed through the formation of Operational Taxonomic Units.

Operational Taxonomic Units

The concept of Operational Taxonomic Units (OTUs) arose from the numerical taxonomy work of Sokal and Sneath, which focused on grouping organisms on the basis of numerical categories (Sneath 2010; Sokal and Sneath 1963). This phenetic approach to taxonomy differed from the existing prevalent phylogenetic view in that a phenetic approach uses a similarity measure to create assemblages of organisms and this grouping is an Operational Taxonomic Unit (OTU). Phenetic approaches are not intrinsically linked to an understanding of evolution as phylogenetic approaches are. At the time of the OTU's inception, sequence data for microbes was far from commonplace and so the creation of OTUs was based on biochemical properties.

With the advent of modern DNA sequencing based on Sanger's (Sanger and Coulson 1975) di-deoxy terminator method, it was feasible to sequence fragments of nucleic acid from bacterial isolates and thus form OTUs from sequence data. Sequence similarity (between genes) and more precisely the complementary idea of distance (or divergence) could be calculated based on sequence alignment. This similarity measurement or distance could also be used to attempt to infer species-level relationships between organisms (Devereux *et al.* 1990).

Currently the bacterial species concept is widely agreed to be based upon sharing of one or more distinguishing phenotypic traits as well as a 70% DNA-DNA hybridization cutoff. Thus if two or more isolates are phenotypically similar and exhibit ≥70% DNA-DNA hybridization then they should be considered the same species (Wayne *et al.* 1987). There are pragmatic reasons why a unification of the species concept with OTU formation is sought. Chiefly there is no reasonable way to perform DNA-DNA

hybridizations unless an isolate exists in pure culture. Therefore in order to describe uncultured, complex microbial communities in terms of species some connection is required between the metagenomic sequence data (and OTU) derived from the community and microbial species. The purpose of this *translation* of OTU to species is the harmonization of sequence-based microbial profiling and conventional microbiology, an idea captured best in the writing of Konstantinidis *et al.*

> "the purpose of the species is to be soundly predictive of the phenotypic potential of a strain (as the greater public assumes it to be)"(Konstantinidis *et al.* 2006)

No formal connection is currently accepted between species and OTU derived from microbial profiling experiments based on marker genes such as 16S rRNA and *cpn*60. Some researchers have attempted to connect OTU formation (and the distances / similarities defined therein) to the species concept through the proposal of similarity cutoffs for 16S rRNA studies (Devereux *et al.* 1990; Schloss and Handelsman 2005; Stackebrandt and Ebers 2006). However, a rigid cutoff is necessarily problematic since the taxonomic framework to which sequence-based OTU are being compared was not arrived at using consistent measurements, but rather through a variety of qualitative and quantitative approaches that have evolved over centuries of study.

Before his death, Sneath published some reflections on microbial systematics and gave an excellent description of how an OTU can be related to identification and specifically speciation.

> "One can imagine the various species as globes in space. An unknown strain is represented by a point in this space. It is possible to measure the distance of the unknown to the spheres, and find which it is nearest to. That will be the most likely identity. But in addition one can say whether the unknown is

> within the envelope of that sphere – which implies that it is extremely likely to be correctly identified. If the unknown is just outside the envelope the identification is less certain, and it may be an atypical strain. If it is midway between two spheres it may be a hybrid of the two species. And if it is a long way from any sphere it is a strain that cannot be identified from the existing data. Such strains, when further work is done, commonly turn out to be new species." page 81(Sneath 2010)

Therefore it is reasonable to connect OTU formation with a definition of bacterial species, however there is a need to establish robust criteria to define a species in terms of an OTU.

For microbial profiling using marker genes (e.g. 16S rRNA and *cpn*60) it has been commonplace to adopt some form of clustering to form OTUs. The sequence based OTUs formed through clustering can be used, as envisioned by Sneath, for identification. In addition, the use of bioinformatic methods for clustering serves as a general approach for data aggregation. Typically sequences are compared pairwise through an alignment method. Using criteria such as percent identity and length, cutoffs are used to determine whether or not sequences are similar on the basis of these cutoffs. Once all pairs of sequences have been compared a transitive closure is performed across the similarity results, producing a set of clusters, which are defined as OTUs. In many cases a maximum distance constraint for a cluster (or OTU) is used to limit membership. These distance constraints can be based on single linkage (nearest neighbour), average linkage (un-weighted pair-group method with arithmetic mean UPGMA), or complete linkage (furthest neighbour). Single linkage is the most straightforward extension of the transitive closure approach that has been widely used in the analysis of cDNA library data. Under single linkage a sequence is placed in a cluster if it shares similarity with any member of that cluster. Average linkage relies on iteratively joining sequences into the cluster

beginning with the two most similar and successively adding the next most similar sequence until some maximum average similarity constraint forces the creation of a new cluster. Complete linkage relies on a sequence being similar to every other member of the cluster.

In the field of microbial ecology there have been a number of clustering implementations demonstrated specifically for OTU formation based on marker gene sequence data. DOTUR (Schloss and Handelsman 2005), more recently implemented within MOTHUR (Schloss *et al.* 2009), is an example of software that has the capacity to form OTUs through clustering by these various methods. In MOTHUR, a maximum distance cutoff (e.g. 3%) is used in conjunction with any of these clustering methods. Thus the OTUs formed by MOTHUR are simply the groupings of reads that are clustered together based on a linkage rule and a distance cutoff. These OTUs consist of a list of sequences, and leave the issue of finding a sequence to represent the OTU as a post-OTU-formation problem.

Given that sequence based clustering is meant to find clusters of related sequences present in a dataset, it is common that implementations of these methods will require some all *vs.* all comparison of the dataset, typically sequence alignments. One feature of these data sets recognized by Edgar is that similar sequences will tend to have many sub-sequences in common (Edgar 2004). Exploiting the similarity of multiple short sub-sequences allowed Edgar to suggest that database type comparisons (the single sequence *vs.* all others in a clustering approach) could exploit this property. Using an ordered list of the database records, Edgar implemented a method to search a query sequence against a database only until the hit / matching falls below a specific level (e.g. 3% distance) and

then knowingly avoid searching the remainder of the database. This heuristic has been demonstrated to increase the performance of clustering (memory consumption and execution time) and is implemented within UCLUST (Edgar 2010). There are some costs associated with the use of such a heuristic, including that they may lead to an increase in error associated with detecting distant relationships.

Clustering of sequences to form OTUs does achieve a level of data aggregation, but it alone does not provide a representative sequence for each OTU formed. Commonly the use of clustering approaches for OTU formation will rely on some *post hoc* step through which a representative sequence for each OTU is chosen. These choices could be based on the length of the sequences within the OTU, nearest sequence chosen from a reference database, or selecting a sequence at random (Caporaso *et al.* 2010).

Selecting a representative sequence based on length could be appropriate when a distance of 0 was used as a cutoff for OTU formation. In this case an OTU would contain only sequences that were identical, thus making the longest one the best representative since other, shorter sequences in the cluster would be subsequences of the representative. Likewise the choice of a representative sequence for an OTU from a reference database could be appropriate if the representative OTU sequence was identical to the reference sequence. Lastly, the choice of a representative sequence from the OTU itself at random could be reasonable if all members of the OTU are equally likely to occur. Where all of these approaches fail is when one considers that a major goal of applying molecular techniques to complex microbial communities is the identification and characterization of unknown organisms. If the thing that defines the OTU (the representative sequence) does not capture all of the variation within the OTU (because the OTU was built with a

distance criterion > 0) then the character of the OTU is in doubt, and its suitability for numerical taxonomy should be called into question. This is perhaps a subtle but critical issue because it goes directly back to Sneath & Sokal's original description of the OTU for numerical taxonomy.

> "Problems may arise if a taxon used as an OTU proves to be variable for one or more characters." Page 121 (Sokal and Sneath 1963)

In Sokal's terminology, an OTU's character (or set of characters) is the criterion used to define it. These characters could involve phenotypic, genotypic observations, or combinations thereof, which are used to classify the organisms being studied. Ideally all organisms grouped into an OTU need to be invariable in the characters that define the OTU.

With modern, sequence-based OTU formation the primary character of an OTU is the sequence chosen to represent the OTU. This sequence becomes the OTU *ipso facto* and so Sokal's comment about problems arising can be interpreted in terms of sequence data: *all sequences within an OTU must be equally represented by the OTU's sequence*. This realization was a key motivator of this thesis, and the argument that OTU formation needs to be accomplished using sequence assembly methods not simply clustering methods. Clustering methods only group sequences together based on some distance criterion. At best, an OTU formed using a distance cutoff of 0 (where all sequences are identical) would be equivalent to the representative sequence produced through assembly.

Assembly methods build the thing (i.e. the consensus sequence or rather the OTU) to which all of the component sequences belong. Sequence assembly for OTU formation has a major advantage over clustering: *assembly supports discovery*. By relying on

sequence assembly it is possible to build novel OTU that have no similarity to a reference database sequence. Further, the sequence that is built through a sequence assembly process is inherently more representative of the OTU than any individual sequence chosen to represent a cluster-based OTU. This means in practice, that assembly can reveal completely novel OTUs and provide the most representative biomarker possible for the OTU from the sequence data: the consensus sequence itself.

It was suggested that as little as 1% of microbial species have been characterized in pure cultures (Staley and Konopka 1985). Of those microbes that have been characterized and identified, estimates place most of the isolates into as few as four phyla. Hugenholtz demonstrated this phenomenon using all isolates from the Australian culture collection to show that 97% of the collection is from the phyla Proteobacteria (54%), Actinobacteria (23%), Firmicutes (14%) and Bacteroidetes (6%) (Hugenholtz 2002) while estimating that there likely are ~45 distinct phyla present within taxonomic outlines or current culture collections but most phyla are represented by extremely small number of examples. In essence this means that we know much about very few phyla (n=4) of a vanishingly small fraction (1%) of the world's bacterial population (10^{30} cells). As the significance of prokaryotic life is on an equal footing with plant life (Whitman *et al.* 1998) but is largely unknown, microbial ecology is in desperate need of bioinformatics methods that enable novel discovery of OTU and produce robust biomarkers for further study of those novel taxa.

OBJECTIVES

1. Apply the Barcode of Life's framework for barcode evaluation to Bacteria.

2. Demonstrate that *de novo* assembly of DNA barcodes from complex microbial communities can faithfully construct the DNA barcodes.

3. Develop a computational process for microbial profiling using metagenomic assembly (mPUMA) that is capable of forming Operational Taxonomic Units (OTUs) through assembly and providing outputs suitable for analysis of OTUs.

4. Demonstrate the use of mPUMA to characterize a previously unstudied microbial community, and derive hypotheses as to the interactions of the community members.

CHAPTER 2 - The chaperonin-60 universal target is a barcode for Bacteria that enables *de novo* assembly of metagenomic sequence data

Copyright

Citation

Links MG, Dumonceaux TJ, Hemmingsen SM, Hill JE. The chaperonin-60 universal target is a barcode for bacteria that enables *de novo* assembly of metagenomic sequence data.PLoS One. 2012;7(11):e49755.

Authors Contributions

Conceived and designed the experiments: MGL TJD SMH JEH. Performed the experiments: MGL JEH. Analyzed the data: MGL JEH. Contributed reagents/materials/analysis tools: MGL TJD. Wrote the paper: MGL TJD SMH JEH.

Abstract

Barcoding with molecular sequences is widely used to catalogue eukaryotic biodiversity. Studies investigating the community dynamics of microbes have relied heavily on gene-centric metagenomic profiling using two genes (16S rRNA and *cpn*60) to identify and track Bacteria. While there have been criteria formalized for barcoding of eukaryotes, these criteria have not been used to evaluate gene targets for other domains of life. Using the framework of the International Barcode of Life we evaluated DNA barcodes for Bacteria. Candidates from the 16S rRNA gene and the protein-coding *cpn*60 gene were evaluated. Within complete bacterial genomes in the public domain representing 983 species from 21 phyla, the largest difference between median pairwise inter- and intra-specific distances ("barcode gap") was found from *cpn*60. Distribution of sequence diversity along the ~555 bp *cpn*60 target region was remarkably uniform. The barcode gap of the *cpn*60 universal target facilitated the faithful *de novo* assembly of full-length operational taxonomic units from pyrosequencing data from a synthetic microbial community. Analysis supported the recognition of both 16S rRNA and *cpn*60 as DNA barcodes for Bacteria. The *cpn*60 universal target was found to have a much larger barcode gap than 16S rRNA suggesting *cpn*60 as a preferred barcode for Bacteria. A large barcode gap for *cpn*60 provided a robust target for species-level characterization of data. The assembly of consensus sequences for barcodes was shown to be a reliable method for the identification and tracking of novel microbes in metagenomic studies.

Introduction

Molecular barcoding is a strategy for cataloging biodiversity through identification and differentiation of organisms using DNA sequencing. Barcodes are relatively short,

specifically defined DNA sequences used to identify organisms by comparing the barcode sequence from an unknown sample to a collection of sequences from known reference samples. In order to facilitate molecular barcoding across life the International Barcode of Life (iBOL) project has developed a framework for evaluating potential barcodes (Hebert *et al.* 2003).

Criteria for barcodes include that they must be universal in the taxa of interest, allowing the development of broad-range ("universal") PCR assays. The existence of reference sequence data from vouchered samples derived from curated specimen collections is required in order to permit robust identification of sequences. For discrimination of taxa it is also essential that inter-specific sequence distance for the barcode sequence be greater than intra-specific distance. This separation between the average intra-specific and inter-specific distance for a given locus defines the "barcode gap"(Meyer and Paulay 2005). Gap size is a critical characteristic for any proposed barcode since it is the key determinant of confident resolution of taxa. For example, Schoch *et al.* (2012) recently demonstrated that the ITS region offers a superior barcode target for fungi compare to the commonly used 18S rRNA target since it features a larger barcode gap, making species discrimination more robust.

For prokaryotes, it follows that a barcode locus that meets the iBOL criteria, and is suitable for cataloging biodiversity through the examination of individual specimens, would be a powerful tool for barcoding in communities of microorganisms. Confident resolution of taxa is paramount in either application. The gene encoding the small subunit (16S) ribosomal RNA has been used extensively in gene-centric metagenomic studies of microbial communities. Despite positive features that have led to its status as the *de facto*

barcode for Bacteria (universality, many sets of broad-range PCR primers targeting different variable regions, large reference database), the 16S rRNA sequence often fails to provide sufficient information for species-level identification (Zeigler 2003), and the occurrence of multiple, identical or nearly identical copies per genome complicates its use as a target for quantification. 16S rRNA sequence based metagenomic studies are commonly limited to reporting of taxa at the genus level or above (Sundquist *et al.* 2007).

Protein-coding genes have long been recognized as providing superior resolution of closely related bacterial taxa compared to 16S rRNA (Case *et al.* 2007; Verbeke *et al.* 2011; Zeigler 2003), and despite statements to the contrary (Schloss *et al.* 2011), one of these protein-coding genes has been demonstrated to provide an alternative "universal target" for bacteria. The gene encoding the 60 kDa chaperonin protein (*cpn*60) found in Bacteria and Eukaryotes has been established as a target for the detection, identification and quantification of microorganisms (Brousseau *et al.* 2001; Chaban *et al.* 2010; Dumonceaux *et al.* 2006c; Goh *et al.* 1998; Goh *et al.* 2000; Goh *et al.* 1996; Goh *et al.* 1997; Hill *et al.* 2006a), as well as for gene-centric metagenomic profiling of microbial communities (Chaban *et al.* 2012; Desai *et al.* 2012; Desai *et al.* 2009; Dumonceaux *et al.* 2006c; Hill *et al.* 2005a; Hill *et al.* 2005b; Hill *et al.* 2002; Oliver *et al.* 2008; Schellenberg *et al.* 2009; Schellenberg *et al.* 2011b). A set of broad-range PCR primers that amplify a region of the gene (the Universal Target, UT) that is generally 552-558 bp (Goh *et al.* 1996; Hill *et al.* 2006b), and cpnDB, a curated sequence database (Hill *et al.* 2004), enhance its utility and contribute to its status as a potentially preferred barcode for Bacteria.

We performed a barcode analysis of the *cpn*60 UT and several regions of the 16S rRNA gene that are widely exploited in systematics and microbial ecology. We show that the barcode gap for the *cpn*60 UT is largest of those examined, and that combined with the length of the target region, facilitate the use of a *de novo* assembly strategy for the formation of operational taxonomic units (OTU) that include the entire length of the target sequence. Finally, we present an approach for the evaluation and optimization of metagenomic assemblies, and demonstrate its application in an examination of the results of assembly of a synthetic community of cloned *cpn*60 UT sequences. The results of this work are discussed in terms of their implications for overcoming some of the limitations of 16S rRNA based sequencing for high resolution profiling of microbial communities, and the characterization of communities where novel taxa are likely to be encountered.

Materials and Methods

16S rRNA and *cpn*60 sequences

A list of completed BioProjects for bacterial genomes (Pruitt *et al.* 2012) was obtained from the NCBI GenBank RefSeq FTP site (circa 12 April 2012). Each GenBank file was processed to extract DNA sequence data as FASTA, Taxon ID, and gene annotations. The resulting GFF files were parsed in order to identify 16S rRNA and *cpn*60 genes, and the DNA sequences of the genes were extracted. Taxon IDs were used to look up the species name and lineage using the NCBI Taxonomy database. Identification of gene annotations for 16S rRNA and *cpn*60 were based on a list of possible annotations ranging from InterPro annotation to explicit keyword sequences. In cases where multiple gene copies were annotated within a single genome all copies were extracted and used in the subsequent analyses.

Definition of putative bacterial barcode regions

Predicted annealing sites of PCR primer pairs for amplification of commonly targeted variable regions were used to delineate putative barcode regions within the 16S rRNA gene. Primers used to amplify V1-V3 and V3-V5 were from the Human Microbiome Project (16S 454 Sequencing Protocol version 4.2.2, http://www.hmpdacc.org/doc/16S_Sequencing_SOP_4.2.2.pdf). Two additional regions (V2-V4, and variable region V6) were delineated using established primers (Sundquist *et al.* 2007). An alternate, shorter version of the V6 region was identified using primers from Hummelen *et al.* (2010). For *cpn*60, the universal target (UT) region corresponding to nucleotides 274-828 of the *E. coli* 60 kDa chaperonin gene was used for barcode gap analysis (Hill *et al.* 2006b). Primer sequences and corresponding gene regions are shown in Table 2-1.

Table 2-1 Definition of barcode regions based on established PCR primers.

Gene	Target region	*E. coli* nucleotides	Primer sequence (5'-3')	Reference
16S rRNA	V1-V3	27-534	27F AGAGTTTGATCCTGGCTCAG 534R ATTACCGCGGCTGCTGG	HMP 16S 454Sequencing Protocol version 4.2.2
	V2-V4	101-806	AGYGGCGIACGGGTGAGTAA GGACTACARGGTATCTAAT	(Sundquist *et al.* 2007)
	V3-V5	357-926	357F CCTACGGGAGGCAGCAG 926R CCGTCAATTCMTTTRAGT	HMP 16S 454Sequencing Protocol version 4.2.2
	V6	907-1073	AAACTCAAAKGAATTGACGG ACGAGCTGACGACARCCATG	(Sundquist *et al.* 2007)
	V6-alternate	985-1078	L-V6 CAACGCGARGAACCTTACC R-V6 ACAACACGAGCTGACGAC	(Hummelen *et al.* 2010)
*cpn*60	UT	274-828	H279 GAIIIIGCIGGIGAYGGIACIACIAC H280 YKIYKITCICCRAAICCIGGIGCYTT	(Goh *et al.* 1996)

Extraction of barcode sequences from whole genome sequences

Full length sequences for 16S rRNA and *cpn*60 were aligned with the RDP aligner (Cole *et al.* 2007) or ClustalW (Thompson *et al.* 2002), respectively. Primer annealing sites were identified manually in each alignment using the eBioX alignment viewer (http://www.ebioinformatics.org/ebiox/) and the predicted amplicon sequence between the annealing sites was extracted. Extracted sequences were processed to remove gaps and then subjected to a second round of multiple sequence alignment by the same algorithm to ensure the best alignment of the putative barcode region for distance calculations. Multiple sequence alignments were converted to PHYLIP format and analyzed with DNADIST to calculate pairwise distances (F84) between all sequence pairs (Felsenstein 1989). DNADIST output was parsed to partition intra-specific and inter-specific distances, and histograms were plotted using Excel.

Synthetic community sequencing

A previously described mixture of 20 cloned *cpn*60 UT sequences of human vaginal bacteria with pairwise nucleotide sequence identities of 56-96% was used as a synthetic microbial community for sequence assembly experiments (Dumonceaux *et al.* 2009). An equimolar mixture of the 20 plasmids was subjected to *cpn*60 universal primer PCR and pyrosequencing on the Roche GS-FLX Titanium platform. Preparation of amplicon libraries for sequencing was done using established protocols (Schellenberg *et al.* 2011a).

Results

Identification of vouchered 16S rRNA and *cpn*60 sequences

Using all RefSeq bacterial genomes in GenBank as a starting point, 1,394 bacterial genomes were identified where both 16S rRNA and *cpn*60 genes could be identified based on annotation. BioProject descriptions provided voucher information for the strain sequenced. A total of 983 species, including at least one representative from each of 21 phyla were included, with the majority (92%) of the genomes belonging to Proteobacteria (48%), Firmicutes (21%), Actinobacteria (12%), Bacteroidetes (5%), Cyanobacteria (3%), or Spirochaetes (3%). Nine records had Taxon IDs corresponding to "unknown" phylum in the NCBI taxonomy (Table 2-2). Numbers of annotated 16S rRNA genes per genome ranged from 1 to 15 (median = 3), and the number of *cpn*60 genes per genome ranged from 1 to 7 (median = 1). All annotated paralogs of *cpn*60 and 16S rRNA genes were identified and used in distance calculations.

Table 2-2 Taxonomic affiliations of the bacterial genomes used in the study.

Phylum*	Number of genomes (%)
Proteobacteria	661 (47.4)
Firmicutes	285 (20.4)
Actinobacteria	161 (11.5)
Bacteroidetes	69 (4.9)
Cyanobacteria	40 (2.9)
Spirochaetes	35 (2.5)
Chlamydiae	22 (1.6)
Chloroflexi	15 (1.1)
Thermotogae	13 (1)
Deinococcus-Thermus	13 (1)
Chlorobi	11 (0.8)
Tenericutes	10 (0.7)
Aquificae	10 (0.7)
Acidobacteria	7 (0.5)
Synergistetes	5 (0.4)
Planctomycetes	5 (0.4)
Fusobacteria	5 (0.4)
Verrucomicrobia	4 (0.3)
Deferribacteres	4 (0.3)
Thermodesulfobacteria	2 (0.1)
Nitrospirae	2 (0.1)
Dictyoglomi	2 (0.1)
Gemmatimonadetes	1 (0.1)
Fibrobacteres	1 (0.1)
Elusimicrobia	1 (0.1)
Chrysiogenetes	1 (0.1)
Unknown/Unassigned	9 (0.6)
TOTAL	1394 (100

*Based on TaxID lookup in the NCBI Taxonomy database

Extraction of the 16S rRNA records from the GenBank annotation was relatively straightforward based on matching of rRNA genes with "16S" in their annotated names. By contrast there was no single annotation characteristic sufficient to recognize the *cpn*60 genes. Thus it was necessary to use keyword matching on gene names ("cpn60", "groEL", "hsp60", "60 kDa chaperonin", "chaperonin 60", etc.) in order to extract *cpn*60 sequences.

Barcode gap analysis

The barcode gap analysis of all 16S rRNA regions and the *cpn*60 UT is summarized in Figure 2-1 and Table 2-3. The data from 1,394 complete RefSeq bacterial genomes allowed for thousands of intraspecific comparisons for each target and nearly 2 million and 16 million interspecific comparisons for *cpn*60 and 16S rRNA, respectively. Data shown is from distance calculation using the F84 method. Other methods for calculating distance (Kimura's 2-parameter model, and the Jukes and Cantor model) yielded similar observations both between and within barcodes, and did not affect the conclusions (data not shown). Barcode gaps for 16S rRNA ranged from 0.26 (V6) to 0.35 (V1-V3), with the exception of the ~75 bp V6-alternate region (Hummelen *et al.* 2010), which had a gap of 0.59. The *cpn*60 UT gap was 0.61 (Table 2-3). The intra-specific distance distributions for 16S rRNA V1-V3, V2-V4 and V3-V5 were the most narrow, with more than 50% of pairwise comparisons in the range of 0.00 to 0.01 (Figure 2-1). This was particularly true for V2-V4 and V3-V5 where >80% of the intra-specific comparisons were 0.00-0.01. The intra-specific distance distributions for both V6 targets, and the *cpn*60 UT were relatively enriched in their right-hand tails. The *cpn*60 UT had the highest median intra-specific (0.07) and inter-specific (0.68) distances (Table 2-3).

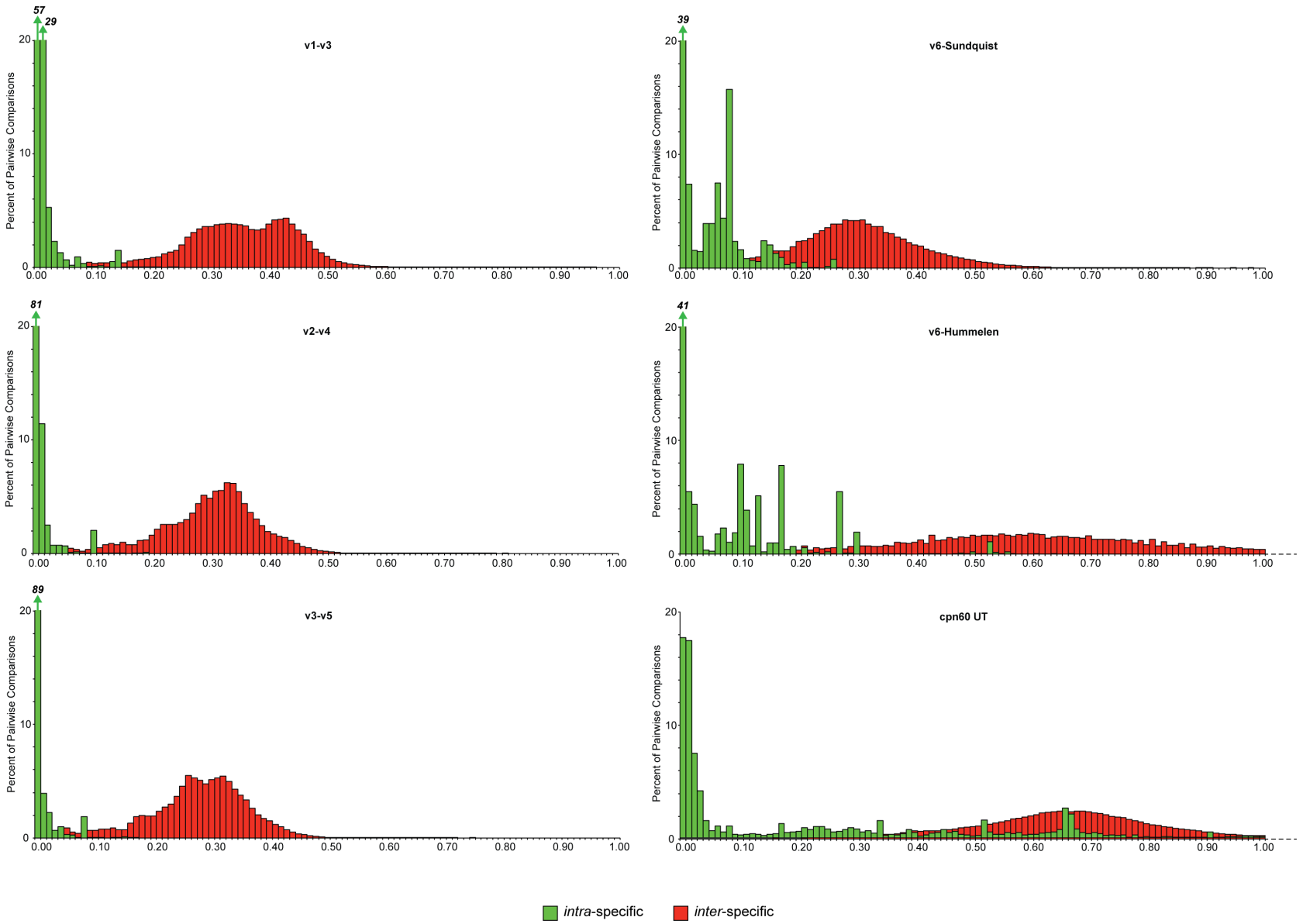

v1-v3
v6-Sundquist
v2-v4
v6-Hummelen
v3-v5
cpn60 UT
Percent of Pairwise Comparisons
intra-specific
inter-specific

Figure 2-1 Barcode gaps for candidate targets. Barcode gap analysis of potential barcodes derived from the 16S rRNA and *cpn*60 genes. Each panel shows the distribution of inter- (red) and intra-specific (green) distances in terms of percent of the total number of comparisons made (see Table 2-3). In cases where percent values exceed 20, the actual value is indicated above an arrow on the relevant bar in the chart. For both V6-alternate and *cpn*60 UT, only distances up to 1.00 are plotted.

Table 2-3 Barcode gap analysis for 16S rRNA and cpn60 targets.

				Intraspecific				Interspecific			
Gene target	Region	Average length (bp)[1]	Barcode gap[2]	# comparisons	Min. distance[3]	Max. distance[3]	Median distance[3]	# comparisons (Millions)	Min. distance[3]	Max. distance[3]	Median distance[3]
16S rRNA	V1-V3	490	0.35	81247	0	0.30	0.00	15.8	0	0.97	0.35
	V2-V4	666	0.31	81247	0	0.22	0.00	15.8	0	0.83	0.31
	V3-V5	551	0.28	81247	0	0.17	0.00	15.8	0	0.80	0.28
	V6	127	0.26	81247	0	0.31	0.04	15.8	0	2.88	0.30
	V6-alternate	75	0.59	81241	0	0.78	0.02	15.8	0	5.91	0.61
*cpn*60	UT	556	0.61	3803	0	5.57	0.07	1.7	0	5.89	0.68

[1]Median length of the target region, between amplification primer annealing sites.
[2]Barcode gap is the difference between the median inter-specific distance and median intra-specific distance.
[3]Distance is expressed in terms of substitutions / site

The distributions of inter-specific diversity in the *cpn*60 UT and 16S rRNA gene were determined by calculating the percent identity of each sequence to the next closest sequence. Figure 2-2 shows the average (median) percent identity between sequences in all 120 bp windows across the targets. Most of the *cpn*60 sequences (87%) were 552-558 bp in length. Diversity along the target length was also remarkably uniform, with median percent identities ranging from 82 to 92%, with most (82%) below 90% identity. Variable and conserved regions of the 16S rRNA gene were visible in the distribution of diversity across the full length of the gene (Figure 2-2) with conserved regions appearing as stretches of windows with a median identity at or near 100% between sequences. The lowest median inter-sequence identities for the 16S rRNA gene were approximately 96%, and were observed near the 5' end of the gene, corresponding to variable regions V1 and V2.

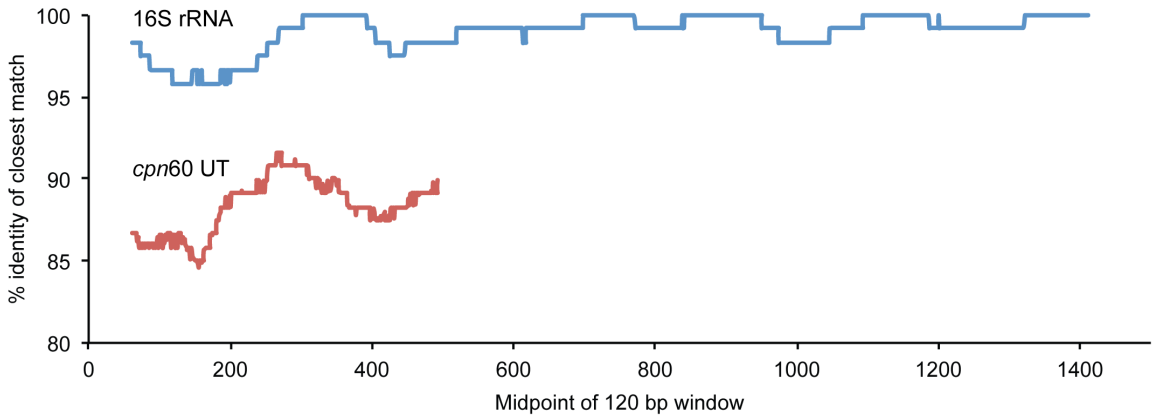


Figure 2-2 Sequence diversity across the 16S rRNA gene and *cpn*60 UT. Median percent identity of each of sequence to its nearest neighbour among the 16S rRNA *cpn*60 UT sequences from 1,394 bacterial genomes. Median percent identity was calculated for each 120 bp window along the length of the targets and identity values are plotted for the midpoint of each window. Due to target length variation, particularly among 16S rRNA genes, data is shown for windows for which at least 95% of the genomes could be included.

Assembly of *cpn*60 UT amplicon sequences

Based on the large barcode gap, target length and uniformity of sequence diversity within the *cpn*60 barcode, optimization of *de novo* assembly of OTU from 454 pyrosequencing data was investigated. A synthetic community comprising 20 cloned *cpn*60 universal targets ranging from 56% - 96% pairwise identity was subjected to universal primer amplification and sequencing with the 454 Titanium pyrosequencing platform. Sequence data was obtained from both ends of the amplicon. A total of 3,437 reads were obtained for the synthetic community with a median read length of 394 (NCBI Sequence Read Archive accession SRR531430). The resulting Standard Flowgram Format (SFF) file was used as input for the gsAssembler in cDNA mode (v2.3, 454 Life Sciences, Branford CT) to form an initial set of OTU. For sequence assembly we focused on the effects of two key parameters: minimum overlap length (-ml) and minimum overlap identity (-mi). Assemblies of the data were conducted, using combinations of minimum overlap length settings of 50, 100, 150, 200, 250, 300, 350 and 400 nucleotides and minimum overlap identity values from 90-99%.

For each assembled OTU, the consensus sequence was evaluated in terms of the extent to which it represented all of the component sequences that were assembled into the OTU. The results of this evaluation were expressed in the form of sensitivity and specificity metrics as follows. Each component sequence and the consensus sequence were compared using wateredBLAST (Schellenberg *et al.* 2009) to the *cpn*60 UT sequences of the clones that comprised the synthetic community. "True positives" were individual sequences from the OTU that matched the same reference sequence as the consensus sequence assembled for the OTU, whereas "false positives" were those sequences that

were incorrectly placed in the OTU being evaluated (i.e. they matched a different reference sequence than the OTU consensus). "True negatives" were defined as those component sequences that were correctly placed into OTUs other than the one being evaluated, and "false negatives" were identified as sequences that matched the same reference as the consensus for the OTU but had been assembled into other OTUs. Thus for each OTU the specificity was calculated using Equation 2-1 and the sensitivity for each OTU using Equation 2-2.

Equation 2-1. Specificity of an OTU consensus sequence.

$$Sp = \frac{TrueNegatives}{TrueNegatives + FalsePositives}$$

Equation 2-2. Sensitivity of an OTU consensus sequence.

$$Sn = \frac{TruePositives}{TruePositives + FalseNegatives}$$

By definition, both sensitivity and specificity of an OTU are values between 0 and 1, with a perfectly assembled OTU having Sp = 1 and Sn = 1 (i.e. no false positives or false negatives). By representing the accuracy of each OTU as a point in a 2-dimensional plane where one axis represented specificity and the other axis sensitivity it was possible to describe the error for a single OTU in terms of Euclidean distance from its coordinates to the optimal coordinates of (1,1) (Equation 2-3). The total error for an assembly was then calculated by summing the error associated with each OTU in the assembly (Equation 2-4).

Equation 2-3 Residual error associated with an OTU consensus sequence.

$$ErrorOTU_i = \sqrt{(1 - Sn)^2 + (1 - Sp)^2}$$

Equation 2-4 Total error of an assembly.

$$ErrorAssembly = \sum_{1}^{i} ErrorOTU_i$$

In general, minimum identity values of >94% resulted in over-splitting of OTU, regardless of the minimum length parameter. For example, setting the minimum identity parameter at 98% resulted in the assembly of 21 to 25 OTU across the range of minimum overlap settings. For minimum identity values ≤94%, the total error for each assembly varied with the minimum length parameter. Figure 2-3 shows the results of assemblies of the synthetic community data over a range of minimum overlap lengths with minimum identity 92%. The number of reads identified by the gsAssembler as singletons increased consistently with increasing minimum overlap length, to a maximum of 21% of the reads at -ml = 400, a value that exceeded the median read length of 394 (Figure 2-3A).

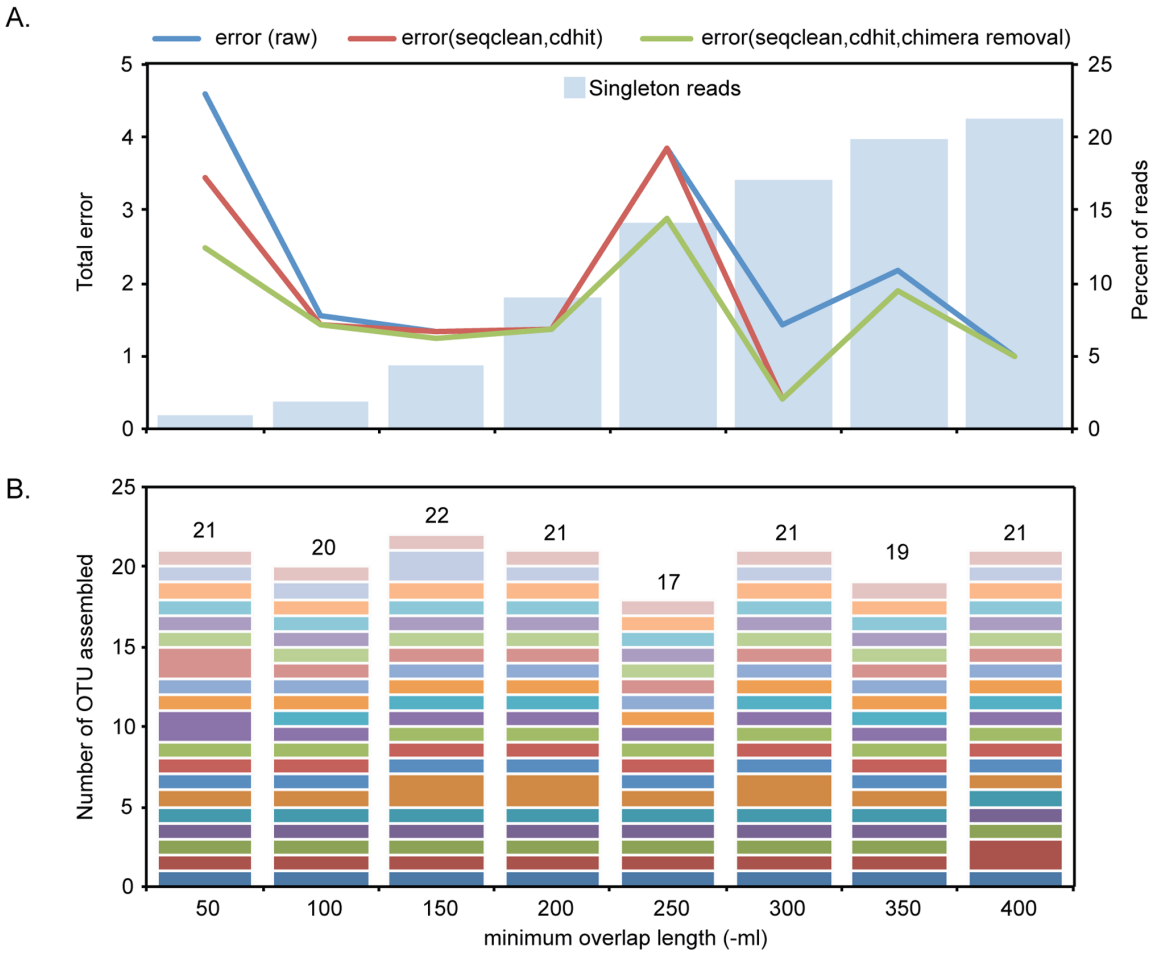


Figure 2-3 Error trade-offs in OTU assembly optimization. A. Total error (left ordinate) for *de novo* assemblies of *cpn*60 UT sequence reads from a synthetic community of 20 cloned targets, using a minimum identity value of 92% and a range of minimum overlap lengths (50-400 nucleotides). Raw total error (blue line), as well as error remaining after post-assembly primer trimming and clustering (red line), and after chimera removal (green line). Light blue bars indicate the percent of sequence reads identified as singletons in each assembly (right ordinate). B. Number of OTU assembled at each minimum overlap length. Each coloured segment of the stacked bar indicates a different member of the panel of 20 community members. The total number of OTU assembled is indicated on the top of each stack.

The impact of post-assembly trimming of amplification primer sequences was also investigated. Following assembly, the universal PCR primer sequences were removed using SeqClean (http://compbio.dfci.harvard.edu/tgi/software/) and the sequence data was clustered at 100% identity using CD-hit (Li and Godzik 2006) to combine OTU that were identical once primer sequences were removed. Chimeric OTU were identified as those where the 5'- 150 bp and 3'- 150 bp matched different sequences in the reference data set. This post-assembly clean up never increased total error, and routinely reduced it, although the amount of error reduction accounted for by primer trimming, clustering and chimera removal was less than that resulting from the optimization of assembly parameters (Figure 2-3A).

A consistent, small amount of total error was observed across minimum overlap lengths of 100, 150 and 200 (1.44, 1.24 and 1.37 respectively), with almost no reduction in error following post-assembly clean up. The lowest error value observed was for -ml = 300 (0.41 total error), although in this assembly, 17% of the reads were lost as singletons. The total numbers of OTU resulting from the assemblies is shown in Figure 2-3B.

Assembly of only the 20 expected OTU sequences was achieved with -ml = 100. This included the correct assembly of OTU corresponding to templates *Lactobacillus gasseri* and *Lactobacillus johnsonii* N2, which are 96% identical over their *cpn*60 UT sequences. Eighteen of the 20 OTU sequences included the full-length *cpn*60 UT. OTU corresponding to synthetic community members *Lactobacillus johnsonii* N2 and *Streptococcus* sp. N1 were incomplete (305 and 214 bp respectively). Most of the error observed in the assemblies shown in Figure 2-3 was due to low sensitivity (i.e. high false

negatives). All assembled OTU consensus sequences were 100% identical to the corresponding input *cpn*60 UT template sequences.

With a minimum overlap length of 150, a similarly low total error was observed, but 22 OTU were generated including 2 OTU corresponding to each of *L. gasseri* and *S. gallolyticus*. A comparison of these OTU pairs showed that in both cases there was a full length OTU 100% identical to the reference sequence formed, however in both cases the full length OTU still had a *cpn*60 UT primer sequence on its 5' end. The shorter sequences were each variant in a single position. In the case of *S. gallolyticus* the shorter sequence had an incorrect single nucleotide deletion and for *L. gasseri* the last nucleotide of the shorter sequence was incorrectly a T.

Discussion

*cpn*60 (also known by synonyms GroEL and Hsp60) is a molecular chaperone conserved in Bacteria and in Eukaryotes (Hemmingsen *et al.* 1988). The *cpn*60 UT sequence, accessible by PCR using degenerate broad-range ("universal") primers has been shown to provide resolution of closely related taxa at the species and subspecies level (Blaiotta *et al.* 2008; Brousseau *et al.* 2001; Goh *et al.* 2000; Goh *et al.* 1996; Hill *et al.* 2006a; Sakamoto and Ohkuma 2010; Sakamoto *et al.* 2010; Vermette *et al.* 2010). The utility of the *cpn*60 UT sequence as a robust tool for detection, identification and quantification of microorganisms is well established, and it has already been implemented in the development of diagnostic tools based on a variety of technologies including quantitative PCR (Chaban *et al.* 2009; Chaban *et al.* 2010), hybridization on solid substrates (Goh *et al.* 2000; Goh *et al.* 1997; Masson *et al.* 2006), and suspension arrays (Dumonceaux *et al.* 2009). The success of *cpn*60-based diagnostics is a direct result of the sequence diversity

of the barcode, and its length, which provides an abundance of informative sequence differences evenly distributed along the length of the target. Even closely related taxa have sufficient sequence differences to allow their discrimination with confidence. Recently, Verbeke *et al.* (Verbeke *et al.* 2011) demonstrated that unlike 16S rRNA sequences, *cpn*60 UT sequence identities alone are strong predictors of whole genome sequence relationships.

The demonstrated utility of the *cpn*60 UT led us to investigate whether it could be evaluated as a DNA barcode using the iBOL framework. In order to fulfill the requirement for vouchered reference data, we limited our analysis to bacteria for which complete sequences were available in NCBI BioProjects. The first significant difference between 16S rRNA and *cpn*60 targets was encountered at the sequence alignment stage. Since *cpn*60 is a protein-coding gene, all classical bioinformatic methods that evaluate DNA evolution in terms of point mutation frequencies are directly applicable to the analysis of *cpn*60 sequences, and sequence alignments are rapidly accomplished with established tools such as ClustalW (Thompson *et al.* 2002). Additionally, the lack of significant length variation in *cpn*60 UT sequences (185 amino acids +/- 1 codon) makes it appropriate to use either global or local alignment methods when comparing sequences. By contrast, 16S rRNA genes encode structural RNA, necessitating the evaluation of mutations in the context of secondary structure using specialized algorithms such as INFERNAL (Nawrocki *et al.* 2009). Multiple sequence alignment tools such as the RDP Aligner and NAST (DeSantis *et al.* 2006) exploit methods to generate alignments based on comparison of input sequences to a reference alignment template, resulting in alignments that are not generally "human readable" due to large numbers of gaps. This

sort of bioinformatic advantage is an important, but perhaps under-recognized aspect of a preferred barcode sequence.

The separation between distributions of intra- and inter-specific distances was originally termed a barcode gap (Meyer and Paulay 2005), and while "the simplest test is whether genetic distances within species are less than those between species"(Kerr *et al.* 2007), there is continuing debate about the best way to measure the barcode gap. Barcode gaps have been expressed in various ways including the difference between the smallest values in either distribution, the average distances, or the ratio of inter- and intra-specific distances (Meier *et al.* 2008). Some authors have gone as far as recommending the establishment of defined cutoffs for barcode gaps (Hebert *et al.* 2004). Use of the difference between minimum distances, or ratio of inter- and intra-specific distances, was inappropriate in our case since the minimum distance in all of the intra- and inter-specific distributions was zero (Table 2-3). Instead we opted to compare the average intra- and inter-specific distances, using median values rather than mean since the distributions, especially the intra-specific distributions, were not normal. The use of the median value as the parameter of comparison between distributions has the additional advantage of reducing the influence of extreme values within the distributions due to factors discussed below.

While on first viewing, the presence of inter-specific zero distances may seem surprising, it is less so when one considers that the genome sequences examined included those from "problematic" taxa such as *Brucella* and *Bacillus*. It is known for these genera and others that historical definitions of species based on phenotypic properties are not always

congruent with comparisons of phylogenetic markers such as 16S rRNA and *cpn*60 (Janda and Abbott 2007).

We included all annotated paralogs of 16S rRNA (median = 3 copies / genome) and *cpn*60 (median = 1 copy / genome) in our analysis since our interest in application of the barcoding concept in bacteria extends beyond the examination of isolates to characterization of complex microbial communities, where practitioners cannot select which paralogs of 16S rRNA or cpn60 are sequenced. In the case of 16S rRNA, paralogs are generally highly similar if not identical to one another (Pei *et al.* 2010), which tends to shift the intra-specific distance distribution toward zero (Figure 2-1). For *cpn*60, multiple copies per genome are the exception rather than the rule, and these paralogs are generally highly divergent, leading to the opposite effect of shifting the intra-specific distance distribution away from zero.

We were able to obtain sufficient data for thousands of intra-specific and millions of inter-specific comparisons from bacteria with complete genome sequences. The results of the barcode gap analysis (Figure 2-1, Table 2-3) revealed that among the longer 16S rRNA loci (those including 3 variable regions), V1-V3 had the largest gap. At 0.35, it is consistent with that of the ITS locus, recently proposed as a preferred barcode for fungi (Schoch *et al.* 2012), and indicates that these loci do exhibit a barcode gap, albeit a small one compared to the other targets examined.

The difference between the two versions of the V6 target was striking, with the shorter (average 75 bp) version having a substantially larger barcode gap (0.59 *vs.* 0.26 for the 125 bp locus). This difference is likely accounted for by the fact that the shorter version

of the locus is defined by PCR primers designed to exclude most of one of the adjacent conserved regions (Hummelen *et al.* 2010), which accounts for a substantial proportion of the 125 bp amplicon defined by Sundquist (Sundquist *et al.* 2007). Short target regions such as the V6 regions of 16S rRNA have recently become more popular for gene-centric metagenomic studies that exploit short-read methods such as Illumina (Hummelen *et al.* 2010; Post *et al.* 2011; Sylvan *et al.* 2012). The disadvantage of these short targets is that they provide relatively few informative positions and may thus be substantially affected by PCR and sequencing error. Their short length also makes them limited in utility for the development of diagnostic methods for the detection of the corresponding organisms in complex samples.

Both the 75 bp V6 locus and the *cpn*60 UT had broad intra-specific distance distributions with long right-hand tails. In the case of the *cpn*60 UT, some of this is accounted for by the occurrence in some taxa of multiple *cpn*60 paralogs with highly divergent sequences. Although the median number of *cpn*60 copies per genome in our study was 1, and the norm in Bacteria is for a single copy per genome, the occurrence of multiple copies that are widely divergent in sequence is well known in some taxa including *Chlamydia,* some Rhizobia and some Actinobacteria (Lund 2009), and these taxa were represented in the genome sequence collection examined in this study. Another contributing factor to the long right-hand tail in the intra-specific distance distribution for the *cpn*60 UT is the inclusion of some non- *cpn*60 but *cpn*60-related sequences as a result of the necessity of using multiple search terms to identify *cpn*60 gene annotations. Although there has been significant effort to standardize chaperonin nomenclature (Coates *et al.* 1993; Hemmingsen *et al.* 1988; Lund 2009), current bacterial genome annotations often do not

conform to these recommendations. An advantage to having *cpn*60 recognized as a barcode for Bacteria would be the standardization of annotations for this gene in bacterial genome sequences.

The *cpn*60 UT had the highest median intra-specific (0.07) and inter-specific (0.68) distances and the largest barcode gap of the loci examined (0.61, Table 2-3), clearly meeting the barcode evaluation criteria. However, an additional criterion is that the barcode be accessible with broad-range PCR primers. Although we exploited published sequence data rather than directly amplifying the target from bacterial isolates, there is a wealth of published studies of targeted analysis of particular taxa and un-targeted metagenomic studies to provide evidence of the efficacy of the broad-range ("universal") PCR primers for the *cpn*60 UT. A review of the data within cpnDB that has been generated through application of the broad-range PCR primers shows that the distribution of the >150 distinct taxonomic lineages closely resembles the distribution in Table 2-2.

In addition to offering robust differentiation of bacterial species based on the examination of isolates, the *cpn*60 UT barcode can be exploited in high resolution profiling of microbial communities. Species level identifications are not often reported in 16S rRNA based metagenomic studies as a direct consequence of its frequent failure to differentiate bacterial species, and widely used tools such as the RDP classifier only provide identification to the genus level (Wang *et al.* 2007). However, in some environments, species level resolution is desirable. For example, the human vaginal microbiome is dominated by *Lactobacillus* species and in some cases, special effort has been dedicated to resolving the common lactobacilli based on partial 16S rRNA sequences (Hummelen *et al.* 2010; Srinivasan *et al.* 2012). In contrast, in *cpn*60 UT-based studies of the vaginal

microbiome, species resolution was accomplished based on comparison of OTU sequences to a reference database using simple, rapid sequence comparisons (Schellenberg *et al.* 2011b).

In this study, we have demonstrated that the features of the *cpn*60 UT enable *de novo* assembly of OTUs, a process that has some important differences from more common clustering methods employed in gene-centric metagenomic sequence analysis. Current popular methods for OTU aggregation *via* clustering (Schloss *et al.* 2009) form clusters of related sequences but do not yield a consensus sequence directly. Instead, clustering methods identify a representative sequence for each OTU by selecting either the nearest neighbour sequence in a reference database, the most common experimental sequence, or a sequence selected at random from the OTU constituents. These existing methods of OTU formation by clustering are useful, unsupervised methods to aggregate large amounts of experimental data but they do not empower discovery of novel OTUs.

The methods for OTU formation using sequence assembly we have described provide a framework for the assembly of full-length OTU consensus sequences in an unsupervised manner. We were able to reconstruct a synthetic community of 20 *cpn*60 UT sequences faithfully, and evaluate the quality of the results of different assembly strategies using an objective, quantitative measure (Figure 2-3). An examination of the results of assemblies of a single data set using a range of minimum overlap length and minimum identity values showed that there is a series of trade-offs involving the various types of error that may result from adjusting these settings. As shown in Figure 2-3, increasing the minimum overlap length can reduce the amount of total error in the final assembly, but there is a corresponding loss of raw data as the median read length is approached.

Decreasing the minimum overlap length and/or the minimum identity could result in an increased likelihood of reads from closely related templates being inappropriately assembled into a single OTU. In the case of the 92% minimum identity assemblies, minimum overlap lengths of 100 and 150 result in low total error, and less than 5% data loss due to singletons, suggesting that this may represent a "sweet spot" for assembly parameters.

Optimal parameters could vary with different microbial population compositions, but the sensitivity and specificity metrics allow an objective assessment of the results of any assembly, even in the absence of knowledge of the actual composition of the community as we had with the synthetic community. Given that the OTU assembly procedure can be optimized to yield robust full-length barcode sequences with high specificity and sensitivity, it becomes possible to trust that if the assembly procedure yields novel sequences these can be relied upon as real biomarkers for an uncharacterized microbe (i.e. not represented among existing reference sequence data). Furthermore, the values for Sn and Sp for any individual OTU provide a potentially useful tool for the evaluation of the quality of a particular OTU of interest. Application of this concept has already resulted in the characterization of distinct subspecies groups within *Gardnerella vaginalis* that were originally identified based on assembly of metagenomic *cpn*60 UT data from human vaginal microbiota (Paramel Jayaprakash *et al.* 2012; Schellenberg *et al.* 2011b).

The assemblies presented here are generated from 454 Titanium sequence data with an average read length of 394, which is typical of the 454 Titanium chemistry on the FLX and Junior platforms. It is anticipated the average read lengths for 454 pyrosequencing will consistently exceed 700 bp with the introduction of the FLX+ chemistry (Roche /

454), which will most likely further improve *cpn*60 UT sequence assembly. We have not yet experimented with *cpn*60 UT sequencing on the Illumina platform, where average read length is commonly lower than that obtained with pyrosequencing. However current forecasts for read length suggest that Illumina's MiSeq platform may reach an average of 400 bp soon. The fact that sequence diversity is evenly distributed along the length of the *cpn*60 target suggests that even existing technologies that produce shorter reads would provide good discrimination of closely related taxa (Figure 2-2), even if full-length OTU assembly would not be possible. *De novo* assembly of 16S rRNA gene sequences would be significantly more difficult as the average sequence difference between species is in the range of technical errors, which may arise from PCR and sequencing protocols. The most informative regions of the 16S rRNA gene (corresponding to the V1-V3 regions where sequences have an average 96% identity to their closest match in the database) are less informative that the most conserved segments of the *cpn*60 UT, for which average sequence identity does not exceed 92% (Figure 2-2).

The results of our study indicate that the *cpn*60 UT provides a preferred barcode for Bacteria compared to the regions of the 16S rRNA gene we examined. The breadth of complete bacterial genome sequence data currently available is influenced by factors such as the cultivability of various taxa, and their relevance to human and animal health, and other well-explored environments. As this spectrum expands beyond what is currently available due to efforts to generate genome sequences for currently under-represented taxa (Wu *et al.* 2009), there will be continuing opportunities to evaluate barcoding potential of the *cpn*60 UT for these new taxa. However, based on the evidence to date, it is clear that the *cpn*60 UT barcode offers significant advantages for cataloging

bacterial biodiversity through the analysis of isolates or in the context of microbial ecology studies. We suggest that *de novo* assembly of metagenomic sequence data from the *cpn*60 UT, or from any appropriate barcode sequence, is a useful approach, especially in cases where resolution beyond the genus level and the confident identification of potentially novel taxa is desirable. To support these activities, we are preparing to release a software package for metagenomic profiling using metagenomic assembly that provides a pipeline for the analysis of microbial profiling data using sequence assembly of barcodes, including the calculation of sensitivity and specificity.

Our results demonstrate that the Barcode of Life's framework has relevance for a domain of life other than Eukaryota. Thus it is reasonable to consider the use of this framework for evaluating barcoding targets for Archaea, including 16S rRNA and the Type II chaperonin (ortholog of *cpn*60) (Chaban and Hill 2012).

Acknowledgments

The authors are grateful to John Schellenberg for his work with cloning of the templates for the synthetic community and Alberto Severini for the cloned products.

CHAPTER 3 - mPUMA: a computational approach to microbiota analysis by *de novo* assembly of OTUs based on protein-coding barcode sequences

Citation

Links MG, Chaban B, Hemmingsen SM, Muirhead K, Hill JE. mPUMA: a computational approach to microbiota analysis by *de novo* assembly of OTUs based on protein-coding barcode sequences. (Accepted).

Authors contributions

MGL designed mPUMA. MGL and KM developed the mPUMA codebase. JEH, BC and SMH contributed to the design and validation of mPUMA, and data analysis. BC designed the C3 chimera checker and generated the *cpn*60 amplicon library. MGL and JEH drafted the manuscript and figures. All authors read and approved the final manuscript.

Abstract

Background

Formation of operational taxonomic units (OTU) is a common approach to data aggregation in microbial ecology studies based on amplification and sequencing of individual gene targets. The *de novo* assembly of OTU sequences has been recently demonstrated as an alternative to widely used clustering methods, providing robust information from experimental data alone, without any reliance on an external reference database.

Results

Here we introduce mPUMA (microbial Profiling Using Metagenomic Assembly, http://mpuma.sourceforge.net), a software package for identification and analysis of protein-coding barcode sequence data. It was developed originally for *cpn*60 universal target sequences (also known as *gro*EL or *hsp*60). Using an unattended process that is independent of external reference sequences, mPUMA forms OTUs by DNA sequence assembly and is capable of tracking OTU abundance. mPUMA processes microbial profiles both in terms of the direct DNA sequence as well as in the translated amino acid sequence for protein-coding barcodes. By forming OTUs and calculating abundance through an assembly approach, mPUMA is capable of generating inputs for several popular microbiota analysis tools. Using SFF data from sequencing of a synthetic community of *cpn*60 sequences derived from the human vaginal microbiome, we demonstrate that mPUMA can faithfully reconstruct all expected OTU sequences and produce compositional profiles consistent with actual community structure.

Conclusions

mPUMA enables analysis of microbial communities while empowering the discovery of novel organisms through OTU assembly.

Keywords

operational taxonomic unit; assembly; automated sequence analysis pipeline; 60 kDa chaperonin; *cpn*60; barcode; metagenomics; microbial profiling microbiota; microbiota analysis

Background

A common approach to the profiling of complex microbial communities is the amplification and sequencing of 'universal' genes, such as *cpn*60 (also known as *gro*EL or *hsp*60) or 16S rRNA, as DNA barcodes for the genomes in which they reside. Barcodes are defined by the International Barcode of Life Project as short, phylogenetically informative sequences from standardized regions of the genome that can be used for species identification and discovery (Hebert *et al.* 2003), and preferred barcodes for microbes including fungi (Schoch *et al.* 2012) and bacteria (Links *et al.* 2012) have been proposed recently. In microbial community studies, broad-range 'universal' PCR primers are used to amplify regions of the target genes and amplicon sequences are determined directly using next-generation sequencing methods. These gene-targeted methods arguably fall under the umbrella of "metagenomics" along with whole genome sequencing approaches, since these are methods based on the analysis of total genomic content of a community of organisms rather than individual isolates (Schloss and Handelsman 2003). The number of individual sequences generated is typically in the

order of 10^6 and can be much greater. Thus, some form of data aggregation is required to reduce the complexity of the raw sequence data, and facilitate interpretation. Data aggregation is focused on the *in silico* steps following sequence data acquisition, and not issues which arise from methods of DNA extraction and possible biases in PCR amplification. The key challenge in aggregation is ensuring that the resulting "profiles" (list of sequences and their abundances), are faithful to the raw sequence data that was aggregated.

Currently, the most widely used method for data aggregation is the formation of operational taxonomic units (OTU) with clustering approaches such as those of MOTHUR (Schloss *et al.* 2009) or UCLUST (Edgar 2010) as implemented within packages such as QIIME (Caporaso *et al.* 2010). Clustering procedures culminate in the selection of a representative sequence for each OTU, which may be selected from the experimental data according to various rules: longest sequence in the cluster, most abundant sequence in the cluster, or random selection. However, representative sequences selected from the experimental data may not include full-length coverage of the target, depending on its length. This in turn limits information content, and the ability to conduct multiple sequence alignments and phylogenetic analysis for characterization of novel OTU sequences. Alternatively, the closest sequence from a reference database may be used to represent the OTU (Schloss *et al.* 2009). A limitation common to all of these approaches is apparent when the community under study contains novel sequences not represented in reference databases. In these cases, novel sequences in the experimental data may be ignored or pooled together as “unclassified” since they do not closely resemble the reference sequences. The end result is that the aggregated

description of the community may not reflect the input sequence data generated in the experiment.

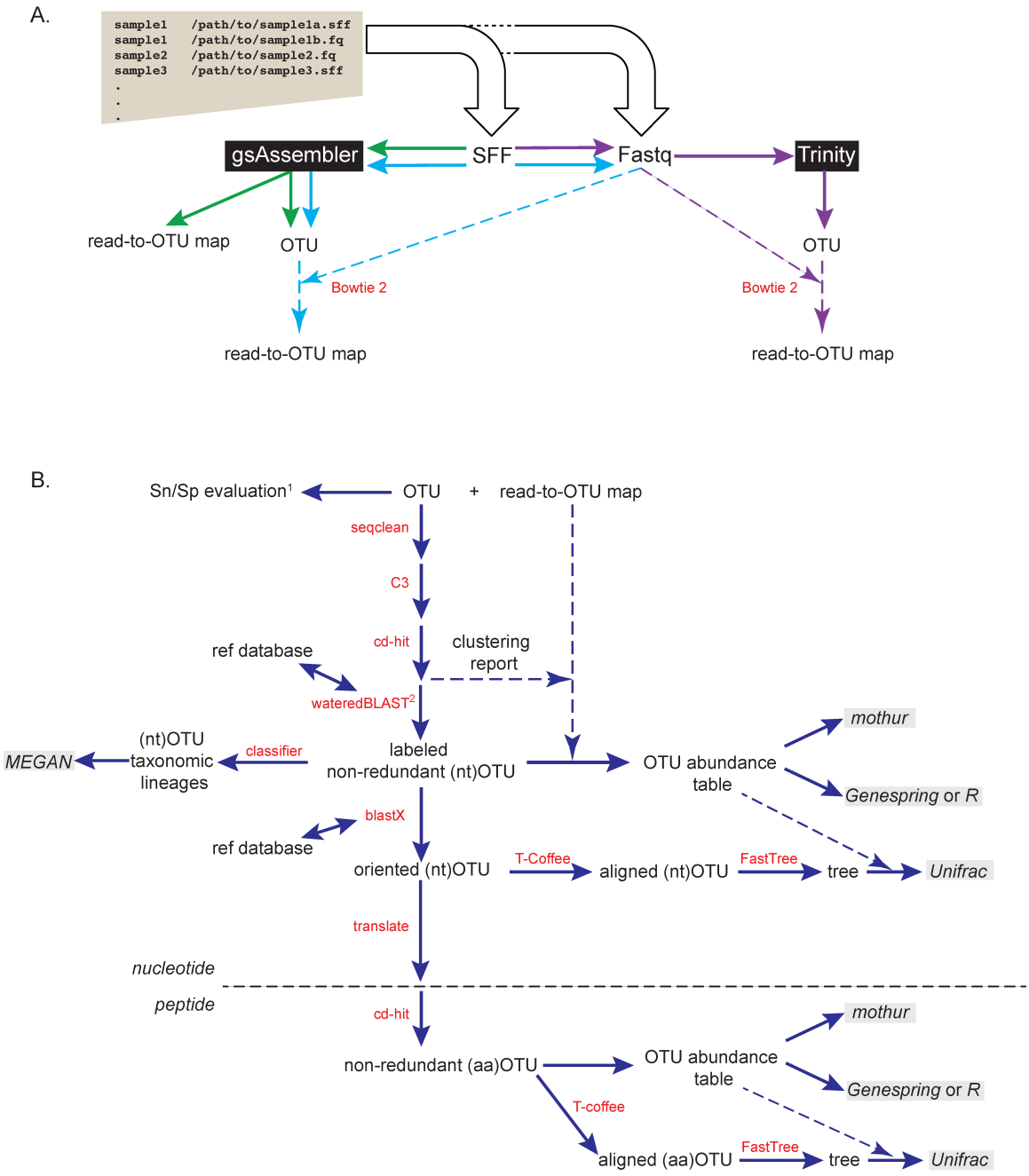
We have demonstrated recently that *de novo* assembly of OTU sequences is an alternative strategy for sequence data aggregation that provides robust information from experimental data alone (Links *et al.* 2012). In this approach, OTU sequences are consensus sequences derived from the experimental data, without any reliance on an external reference database. This strategy has been used successfully in producing high resolution profiles of a variety of complex microbial communities (Chaban *et al.* 2012; Desai *et al.* 2012; Schellenberg *et al.* 2011b) and has led to the resolution of subspecies level diversity within previous established bacterial "species" (Paramel Jayaprakash *et al.* 2012). However, until now there has been no computational pipeline available for this work, requiring practitioners to attend to each step of the assembly and post-assembly analysis individually. Here, we introduce mPUMA (microbial profiling using metagenomic assembly), a computational pipeline for the automated assembly and analysis of OTU sequences from protein-coding gene sequence data derived from microbial communities.

Methods

mPUMA workflow

mPUMA was written in PERL using BioPerl (Stajich *et al.* 2002) and is maintained as a sourceforge project (http://mpuma.sourceforge.net/). It was developed originally for assembly of cpn60 universal target sequences (Goh *et al.* 1996; Hill *et al.* 2004) since the characteristics of this target make it a preferred sequence barcode for resolution of

bacterial taxa (Links *et al.* 2012). However, mPUMA is applicable to any other suitable molecular barcode. mPUMA assembles OTU from PCR amplicon sequence libraries generated from any number of samples, starting from a set of SFF or Fastq files, and a text file explaining how the files relate to experimental samples. Following assembly, the abundance of each OTU is determined and files for downstream analysis using several common microbial ecology and phylogeny tools are generated. The mPUMA workflow is illustrated in (Figure 3-1).



[1]Quality of assembly can be evaluated by assessing Sensitivity/Specificity of each OTU as defined in [3].
[2]wateredBLAST is a combination of BLAST and Smith-Waterman alignments, described in detail in [15].

Figure 3-1 mPUMA workflow. Programs used at each step in the pipeline are shown in red. A. User-defined protocol options for assembly and read-to-OTU tracking include gsAssembler for both processes (green arrows), gsAssembler plus Bowtie 2 for read tracking (blue arrows), and Trinity assembly plus Bowtie 2 for read tracking (purple arrows). B. Post-assembly analysis of OTU and abundance data. Grey boxes indicate possible downstream analysis tools for which input is generated by mPUMA. The horizontal broken line indicates the transition from analysis of nucleotide OTU [(nt)OTU] and translated peptide OTU [(aa)OTU].

Sequence assembly

Sequence assembly within mPUMA can be performed by two methods: gsAssembler (Roche/454, Branford, CT) in cDNA mode, or Trinity (Grabherr *et al.* 2011). Abundance per OTU can be calculated by mPUMA from a read-to-OTU map produced in one of two ways (Figure 3-1). For gsAssembler assemblies, the internal read tracking of the assembly process can be used as the basis for the read tracking. Alternatively, reference mapping with Bowtie 2 (Langmead and Salzberg 2012) can be used to map each experimental read onto reference OTUs assembled with either gsAssembler or Trinity. Considerations for the optimal assembly and read tracking strategy for any particular project are discussed below. Regardless of the strategy used, the quality of the assembly and read tracking result is assessed in terms of the specificity and sensitivity of each OTU as described previously (Links *et al.* 2012).

Post-assembly analysis of OTU

Removal of PCR primer sequences is accomplished with seqclean (http://sourceforge.net/projects/seqclean/files/). Identification and removal of chimeric sequences is performed by two strategies implemented within mPUMA. First, gsAssembler identifies chimeras resulting from the assembly process. Second, the Chaban Chimera Checker (C3) identifies putative chimeras that may be removed from subsequent analyses. In C3 the 5' and 3' ends of each OTU (150 bp) are extracted, compared to a reference set of sequences (e.g. non-redundant set of sequences from cpnDB (Hill *et al.* 2004)) and evaluated to see if both ends match the same reference sequence in the expected orientations. Putative chimeras are identified as assembled OTU

that fail this test. In novel environments where taxa are not well represented in the reference database, it may be appropriate to forego the use of C3 because the novelty of the experimental sequences could lead to an increased false positive rate in chimera identification.

Non-chimeric OTU are clustered at 100% identity by CD-hit (Li and Godzik 2006) to remove redundant sequences. For protein-coding barcode sequences, mPUMA implements BLASTX (Altschul *et al.* 1997) to identify the correct reading frame for translation of OTU, and then translates the nucleotide OTU to their corresponding peptide OTU sequences. Redundant peptide sequences are also collapsed using CD-hit (Li and Godzik 2006) at 100% identity. mPUMA calculates the abundance of each non-redundant peptide OTU for each library, resulting in a peptide OTU abundance table.

Nucleotide and peptide OTU and abundance data are formatted for use with additional tools, which are run automatically where appropriate. Prior to generating input files for these applications, mPUMA carries out a down-sampling process where reads are sampled at random to the depth of the smallest library to address the concerns raised by Gihring *et al.* related to the effects of unequal sampling effort on calculation and comparison of ecological parameters such as richness, diversity and evenness (Gihring *et al.* 2012). Abundance files for OTU are used to create input for MOTHUR (Schloss *et al.* 2009). Using t-coffee (Notredame *et al.* 2000) for multiple sequence alignments and FastTree (Price *et al.* 2009), a phylogenetic tree of the OTU is calculated, which can be used in conjunction with abundance data to analyze libraries in Unifrac (Hamady *et al.* 2010; Lozupone and Knight 2005). A naïve Bayesian classifier trained on cpn60 universal target sequences from cpnDB (Hill *et al.* 2004) has been developed using the

RDP classifier framework (Wang *et al.* 2007). Classifier results can be loaded into MEGAN (Huson *et al.* 2007) for comparison of multiple libraries in a taxonomic context. All of the output files generated by mPUMA for secondary analyses are generated both for the nucleotide and the amino acid OTU sequences.

Computational platform

Demonstrations of mPUMA running in an unattended fashion were performed using a previously published dataset (Chaban *et al.* 2012) that included 711 MB of data in SFF files. Analyses were carried out on a Dell R910 equipped with 128 GB of RAM and 2 x Intel Xeon 6-core E7530 processors running CentOS 5.8.

Results & Discussion

To validate the primary function of mPUMA (OTU formation and abundance calculation), we tested its performance in the analysis of sequence data generated by amplification and sequencing of *cpn*60 universal target sequences from a synthetic community containing cloned *cpn*60 universal target sequences from 20 human vaginal bacteria with pairwise sequence identity values of 60-96% (Dumonceaux *et al.* 2009). PCR from this template mixture and pyrosequencing of the resulting amplicon library on a Roche GS FLX instrument was performed using established protocols (Schellenberg *et al.* 2011a), resulting in 9877 sequence reads from either the 5' or 3' end of the target sequence. The SFF data is accessible through the mPUMA sourceforge site (http://mpuma.sourceforge.net/). We verified that all 20 target sequences were represented in the results by using Bowtie 2 to map all reads on to the reference sequences for the synthetic community ("Target" in Figure 3-2)

OTU formation and abundance calculations were performed on the dataset using all three options available within the mPUMA pipeline (gsAssembler OTU assembly/gsAssembler read-to-OTU mapping, gsAssembler OTU assembly/Bowtie 2 read-to-OTU mapping and Trinity OTU assembly/Bowtie 2 read-to-OTU mapping) and the resulting microbial profiles were evaluated for number of OTU generated, number of reads unmapped, amount of total error generated and comparison of the profile to the known "Target" synthetic community profile (Figure 3-2).

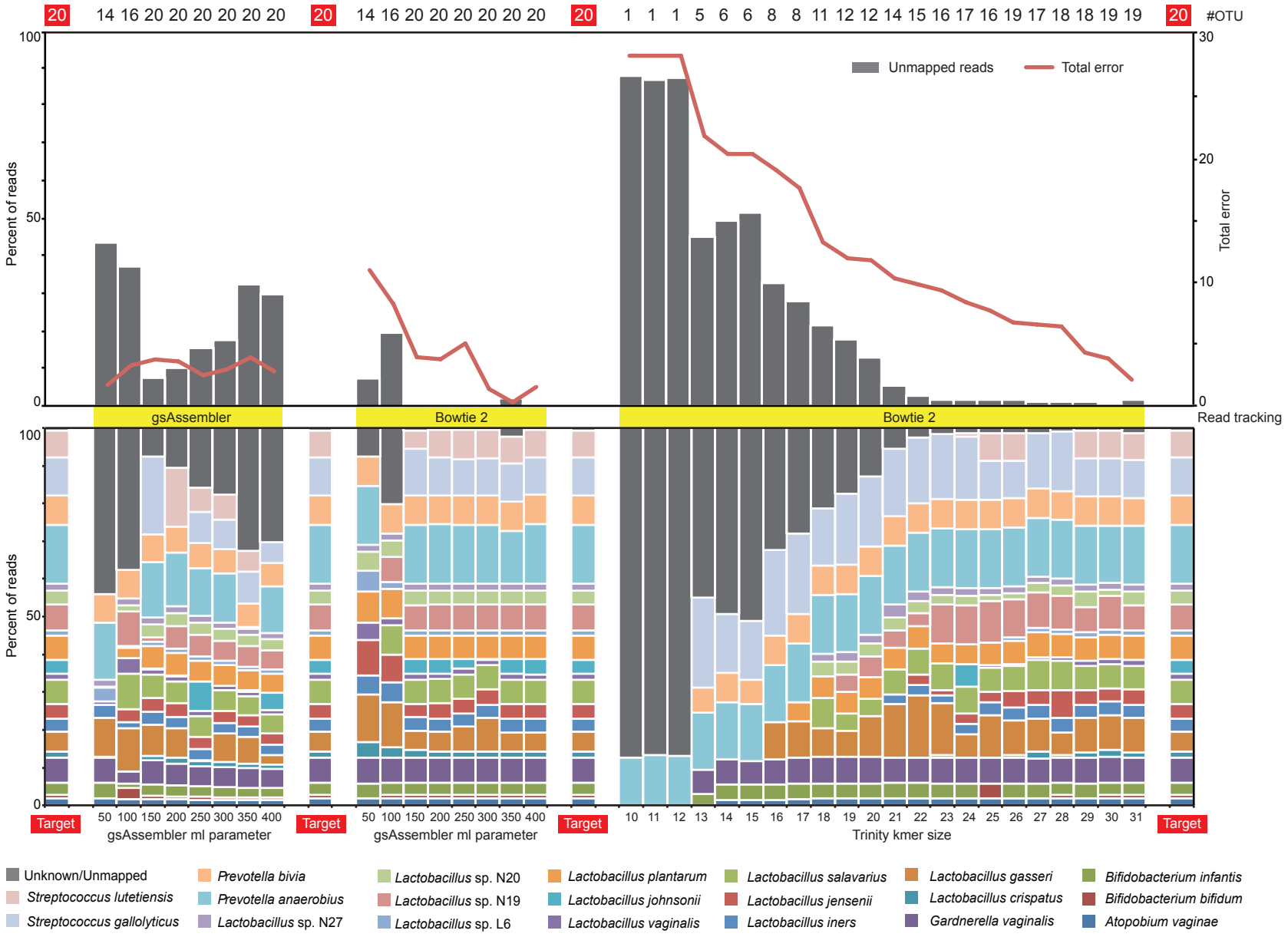

#OTU
Unmapped reads
Total error
Percent of reads
Total error
gsAssembler
Bowtie 2
Bowtie 2
Read tracking
Target
gsAssembler ml parameter
gsAssembler ml parameter
Trinity kmer size
Unknown/Unmapped
Streptococcus lutetiensis
Streptococcus gallolyticus
Prevotella bivia
Prevotella anaerobius
Lactobacillus sp. N27
Lactobacillus sp. N20
Lactobacillus sp. N19
Lactobacillus sp. L6
Lactobacillus plantarum
Lactobacillus johnsonii
Lactobacillus vaginalis
Lactobacillus salavarius
Lactobacillus jensenii
Lactobacillus iners
Lactobacillus gasseri
Lactobacillus crispatus
Gardnerella vaginalis
Bifidobacterium infantis
Bifidobacterium bifidum
Atopobium vaginae

Figure 3-2 Comparison of methods for both assembly and abundance calculation using a synthetic community of 20 cloned *cpn*60 universal target sequences. Three different scenarios were investigated for the generation of a microbial profile (left-to-right): gsAssembler alone, gsAssembler plus Bowtie 2 for abundance, and Trinity plus Bowtie 2 for abundance. The number of community members recovered is shown across the top (out of 20). The major parameter affecting the accuracy of assembly is varied across the lower x-axis. For gsAssembler the minimum identity of overlaps was held constant at 90 while the minimum length parameter was varied. In the case of Trinity, the k-mer length was varied from 10 to 31 bp. The upper panel shows the percentage of reads which were un-trackable (left ordinate) and the total error associated with each assembly (right ordinate). In the lower panel, microbial profiles are plotted as stacked bars with each element colored by organism according to the legend. Profiles marked as "Target" indicate the actual composition of the amplicon library determined by Bowtie 2 mapping of all reads on to the 20 reference sequences.

gsAssembler was able to reconstruct all 20 expected OTU with minimum length parameter settings of >100 bp (Figure 3-2). However, despite accurately describing the richness of the sample (20 OTUs), read tracking within gsAssembler failed to place a substantial proportion of data in any OTU. The proportion of sequence reads unmapped increased steadily from 8% to a maximum of 33% as the minimum length parameter was increased from 150 through 350 bp (Figure 3-2). There are several possible explanations for this unplaced data: the reads could be short or of low quality, or the assembly process may not have completely accounted for the placement of each read to an OTU. In our experience, situations in which a study contains samples with extreme differences in richness can lead to incomplete mapping when utilizing gsAssembler which cannot be resolved using the available command line options (-ig, -it, and -icc). The occurrence of such "thresholding" problems is recorded in the 454IsotigsLayout.txt files generated by gsAssembler. Given that we confirmed that gsAssembler had correctly resolved all 20 of the expected OTU for this synthetic community, we were left with the possibility that either there was a proportion of the data which was of insufficient quality and/or length to be placed in the OTUs at higher stringencies (i.e. greater minimum overlap length requirement) or the placement was incomplete. To determine which of these phenomena were occurring we employed Bowtie 2 (Langmead and Salzberg 2012) as a method to independently assess the read to OTU mapping.

When read mapping was performed using Bowtie 2 to place reads onto a gsAssembler assembly, there was a dramatic reduction in the proportion of unmapped data and in total error of the assembly coincident with all 20 members of the synthetic community being resolved (Figure 3-2). The results of assembly using gsAssembler with a minimum

overlap >100 bp followed by read mapping with Bowtie 2 served to construct a microbial profile indistinguishable from the actual profile of the synthetic community at both the nucleotide and peptide levels, with the 20 expected nucleotide OTU and 19 corresponding peptide OTU (peptide sequences for *Lactobacillus gasseri* and *L. johnsonii* are identical). This result confirmed that the reads were of sufficient length and quality for inclusion, and thus the more likely explanation for the relatively large proportion of data that is not placed by gsAssembler read tracking is that the assembler had failed to completely assign all reads to the OTU assembled (the thresholding problem described above).

gsAssembler uses an Overlap-Layout-Consensus (OLC) strategy for assembly, which is dramatically affected by coverage depth (Li *et al.* 2012). The dominant alternative approach for assembly is the use of a de Bruijn graph (DBG) to analyze sequence composition in terms of k-mers. The total length of sequence being assembled, independent of coverage depth, governs the size of a de Bruijn graph. Being unaffected by coverage depth is the chief computational advantage of DBG approaches. We explored whether Trinity, a DBG method (Grabherr *et al.* 2011), offers a valid alternative to gsAssembler in cDNA mode for the analysis of microbial barcode data. Within Trinity, the parameter most likely to affect the accuracy of assembly results is k-mer size. We examined all possible k-mer lengths supported by Trinity (k-mer ranging from 10 to 31, inclusive). Bowtie 2 was then used to map the individual reads onto the non-redundant set of OTU formed by Trinity for calculating abundance because the reductive process of distilling sequences to component k-mers eliminates the ability of tracking reads directly within DBG approaches.

As can be seen in Figure 3-2, increasing k-mer length resulted in the formation of more of the expected OTU, reduction of the proportion of unmapped reads and a corresponding reduction in total error of the assembly. However, in no case did Trinity resolve all 20 OTUs from the synthetic community. Trinity assemblies with a k-mer of 30 or 31 were nearly complete, failing only to resolve an OTU for *L. johnsonii*. This was perhaps not surprising since *L. johnsonii* and *L. gasseri* are the two most similar members of the community (96% identical) and have similar abundances, being the 11th and 9th most abundant in this dataset, respectively. The *L. johnsonii* reads were placed in the *L. gasseri* OTU when an *L. johnsonii* OTU was not formed.

Resource usage by mPUMA can vary significantly depending on the size and complexity of the datasets being analyzed. In our experience the use of Trinity over gsAssembler can be necessary for computational constraints (memory and cpu time) when dealing with datasets that are extremely rich or diverse. mPUMA is suitable for the assembly and analysis of OTU from other suitable targets besides *cpn*60, such as the universal archaeal type-II chaperone (also known as Thermosome or TCP1 or CCT) (Chaban and Hill 2012), and *rpo*B (Vos *et al.* 2012). Pyrosequencing data from both have been processed through mPUMA, confirming its utility for other protein-coding targets. To date, we have applied mPUMA to the analysis of amplicon sequence data from the 454 GS FLX, Titanium and Junior platforms. We encourage the microbial ecology community to investigate the application of mPUMA to other sequence data types and gene targets of interest.

Conclusions

The *de novo* assembly of OTUs from barcode sequence data can be optimized to reduce error and accurately reflect the richness of a microbial community, presenting possible advantages over clustering methods that may mask diversity or inhibit discovery of novel sequences. The mPUMA pipeline was developed to facilitate the use of assembly in microbial ecology studies where both accurate descriptions of richness and calculation of OTU abundance are desired. Based on our examination of a synthetic community, optimal resolution of OTU sequence barcodes and calculation of their abundance can be achieved through use of gsAssembler with a minimum overlap length parameter > 100 bp followed by Bowtie 2 read tracking for determining OTU abundance. In cases where computational performance is limiting, Trinity assembly followed by read tracking with Bowtie 2 should produce near-optimal results with only exceptionally similar barcodes remaining unresolved. In choosing the most appropriate strategy for assembly and abundance calculations from among the options available in mPUMA, researchers will need to balance the computational performance of the assembly approach with the precision of OTU formation.

The mPUMA software package is available from sourceforge and it is covered by an open-source license (http://mPUMA.sourceforge.net). At present, mPUMA is distributed on its own but it is possible that in the future it may become incorporated into a Virtual Machine image. Since it is as an open-source platform, mPUMA can be extended by anyone interested in utilizing *de novo* assembly for the analysis of microbial profiling data.

Availability of Supporting Data

The SFF data used in the validation and demonstration of mPUMA is available through the mPUMA sourceforge site (http://mpuma.sourceforge.net/).

Acknowledgements

This work was supported by funding from the Canadian Institutes for Health Research, the Natural Sciences and Engineering Research Council of Canada, and Agriculture and AgriFood Canada. We are grateful to the members of the Hill Lab and the *cpn*60 research collaboratorium for their valuable feedback, and contributions to testing mPUMA.

CHAPTER 4 - Simultaneous profiling of seed-associated bacteria and fungi reveals antagonistic interactions between microorganisms within a shared epiphytic microbiome on *Triticum* and *Brassica* seeds

Citation

Links MG, Demeke T, Gräfenhan T, Hill JE, Hemmingsen SM, Dumonceaux TJ. Simultaneous profiling of seed-associated Bacteria and Fungi reveals a core microbiome on *Triticum* and *Brassica* seeds.

Authors contributions

MGL, TD, TG, SMH and TJD conceived of the study. MGL and TJD performed the experiments. MGL and TJD analyzed the data. MGL, JEH, SMH, and TJD wrote the manuscript and drafted the figures. All authors read and approved the final manuscript.

Supporting Data

The sequences reported in this manuscript have been submitted to GenBank (PRJNA203419 and JX909334-JX909350).

Abstract

We characterized the prokaryotic and eukaryotic microbiota associated with healthy crop seed surfaces by microbial profiling with *cpn*60. Over 400,000 sequences derived from independent *Triticum* spp. (n=6) and *Brassica* spp. (n=5) seed washes were assembled into 5,477 operational taxonomic units (OTU). Total epiphytic bacterial load, as measured by the number of 16S rRNA-encoding gene copies/g seeds, was not significantly different between the seed types, nor were community diversity parameters (richness and evenness). Analysis of the sample prevalence of OTU revealed a shared microbiota between the *Triticum* and *Brassica* samples, with 578 OTU found commonly in these crops at a variety of abundances. Hierarchical clustering of these shared OTU revealed that 203 OTU were significantly different in abundance on *Triticum* seeds compared to *Brassica*. This was confirmed for selected OTU by quantitative PCR. Microorganisms were isolated from seeds corresponding to 5 bacterial and 4 fungal OTU, showing 99-100% identity between the *cpn*60 sequences of the isolates and the assembled OTU sequences. Bacterial strains identified as *Pantoea agglomerans* were found to have antagonistic properties toward one of the fungal isolates (*Alternaria* sp.), providing a possible explanation for their reciprocal abundances on *Triticum* and *Brassica* seeds. Use of the *cpn*60 universal target enabled the simultaneous profiling of prokaryotic and eukaryotic microbiota and revealed previously unrecognized microbial interactions that could be exploited to protect seeds from spoilage and reduce pathogen burden.

Keywords

seed microbiota, seed microbiome, *cpn*60 universal target, biocontrol, core microbiome/barcode

Introduction

The seeds of crops such as wheat (*Triticum* spp.) and canola (*Brassica napus*) are products of the agricultural enterprise and the source of the next generation of plants. Healthy, high quality seeds are critically important for the stability of the world's food supply and the economic success of farmers. Crop seeds, like other parts of the plant, are colonized by epiphytic microbiota consisting of synergistic, commensal, and potentially pathogenic microbes that play a crucial role in health and susceptibility to disease (Critzer and Doyle 2010; Hashidoko 2005). Since the plant-associated microbiota clearly plays a role in plant fitness (Hallmann *et al.* 1997), different crops might be expected to harbor distinct microbiota on their seed surfaces and the constituents of these microbial communities are likely to have functional relevance during plant growth, development and seed storage. For example, specific microorganisms such as *Penicillium verrucosum* and *Alternaria alternata* in stored crop seeds can cause spoilage, decrease crop value, or produce mycotoxins that have a direct effect on human health (Duarte *et al.* 2010; Magan and Aldred 2007; Magan *et al.* 2010). On the other hand, commonly utilized crop rotations, such as canola-wheat, are known to have positive benefits for yields and for pathogen control (Bushong *et al.* 2012; Harker *et al.* 2012; Zegada-Lizarazu and Monti 2011). Microorganisms that associate with each crop may influence the growth and development of the subsequent crop in the rotation. The potential impact of crop-based

microbial communities on yields and on pest control demands that a comprehensive knowledge of microbiota associated with seed surfaces be elucidated.

Culture-independent methods for characterizing microbes associated with an environment involve the PCR amplification of taxonomic gene markers with universal primers, sequencing amplicons, and comparison of sequences to a reference database for taxonomic assignment. More recently, next-generation sequencing techniques have substantially increased the volume of DNA sequence data for this type of analysis; however, data analysis methods are the focus of ongoing development (Hamady and Knight 2009).

One approach for data analysis is to cluster sequences based on a similarity criterion (e.g. 97% identity). Software packages such as mothur (Schloss *et al.* 2009) and QIIME (Caporaso *et al.* 2010) implement clustering methods to form Operational Taxonomic Units (OTUs) based on a rule for cluster membership (single, average or complete linkage). In order to carry out downstream analyses on OTUs, a representative sequence must be chosen for each OTU. Clustering methods are a robust approach for OTU formation but they are limited, as clustering does not prescribe how to select a representative sequence for each OTU. Choices for a representative sequence from clustered OTUs vary from the longest, the closest sequence from a reference database to one chosen at random. These choices for representative sequence selection can allow for additional downstream analyses but they are poorly suited for identifying and tracking novel microbes.

In order to detect novel microbes it would be ideal that the sequence chosen for an OTU be the most representative sequence possible. For clustering methods, the only case where a perfect representative sequence could be identified is when clustering was performed at 100% identity and at least one sequence spans the full length of the OTU. This scenario is unlikely given the variable read lengths and error rates of current next-generation sequencing technologies.

An alternative method for OTU formation using *de novo* assembly has recently been described, and demonstrated for microbial profiling (Links *et al.* 2012)(Chapter 3). Assembly provides consensus sequences for OTU that are inherently the most representative sequences possible. Thus, when an OTU is assembled for an uncharacterized or novel organism, the assembled consensus sequence is a discrete biomarker that can identify the organism.

The chaperonin 60 gene (*cpn*60, also known as *hsp*60 or *gro*EL) encodes a protein that functions as a molecular chaperone assisting in the formation and maintenance of protein structures in cells (Hemmingsen *et al.* 1988). Determination of microbial community composition based on the amplification and sequencing of a portion of the *cpn*60 gene, the universal target (*cpn*60 UT) (Hill *et al.* 2004), offers a protein coding alternative to 16S rRNA based approaches. Cpn60-encoding genes are found in essentially all prokaryotes and eukaryotes, and the *cpn*60 UT is accessible with a set of universal PCR primers (Hill *et al.* 2006b). The *cpn*60 UT has been exploited for characterizing microbial communities using both traditional (Dumonceaux *et al.* 2006c; Hill *et al.* 2005a) and next-generation (Chaban *et al.* 2012; Desai *et al.* 2012; Schellenberg *et al.* 2009; Schellenberg *et al.* 2011b) technologies, and it provides a convenient molecular target

with higher taxonomic resolution than 16S rRNA for microbial profiling (Paramel Jayaprakash *et al.* 2012; Schellenberg *et al.* 2011b; Verbeke *et al.* 2011; Vermette *et al.* 2010). In addition, *cpn60* has recently been shown to possess a larger "barcode gap" for *Bacteria* compared to 16S rRNA, and therefore is a preferred barcode for the domain *Bacteria* (Links *et al.* 2012).

Using the *cpn*60 UT as a DNA barcode we tested the hypothesis that the seed-associated epiphytic microbiota of *Triticum* spp. and *Brassica* spp. are distinguishable. Furthermore, by comparing the assembled OTU sequences with those from bacteria and fungi isolated from these samples we demonstrate that OTU can be assembled accurately for microbes in complex samples. Finally, we examined the interactions of microorganisms that were originally identified based on sequence analysis of *cpn*60 amplicons.

Materials and Methods

Seed sources.

Crop seeds of diverse geographic origins within Canada were chosen for analysis. Seeds of *Brassica* (*B. juncea*, *B. rapa*, *B. napus*), and *Triticum* (*T. aestivum*, *T. durum*) were used for the study (Table 4-1). All seeds were assessed as healthy by the Canadian Grain Commission with their respective grades denoted in Table 4-1. The seeds were from the 2009 harvest and were stored separately in plastic bags at room temperature.

Table 4-1 Description of samples.

Sample name	Sample source	Sample description[1]	Geographic origin[2]
Wheat-1	*Triticum durum*	CWAD, grade 2	Western Canada
Wheat-2	*Triticum durum*	CWAD, grade 3	Western Canada
Wheat-3	*Triticum aestivum*	CESRW, grade 2	Eastern Canada
Wheat-4	*Triticum aestivum*	CWRS, grade 1	Western Canada
Wheat-5	*Triticum aestivum*	CWRS, grade 2	Western Canada
Wheat-6	*Triticum aestivum*	CWRS, grade 3	Western Canada
Brassica-1	*Brassica juncea*	Brown mustard, grade 1	Western Canada
Brassica-2	*Brassica napus*	Canola B	Western Canada
Brassica-3	*Brassica napus*	Canola A	Western Canada
Brassica-4	*Brassica juncea*	Oriental mustard, grade 1	Western Canada
Brassica-5	*Brassica rapa*	Brown mustard	Western Canada

[1]abbreviations: CWRS, Canada Western Red Spring wheat; CESRW, Canada Eastern Soft Red Winter wheat; CWAD, Canada Western Amber Durum wheat

[2]seeds were sourced from different geographic locations in Eastern Canada (Ontario or Quebec) or Western Canada (Manitoba, Saskatchewan, Alberta, or British Columbia)

DNA extraction from seed-associated epiphytic microbiota.

A 10 g sample of each seed lot was soaked in a solution of 45 ml buffered peptone water (10 g peptone, 5 g NaCl, 3.5 g Na_2HPO_4, 1.5 g KH_2PO_4 $liter^{-1}$ (Kim *et al.* 2006) containing 0.05% Triton X-100 (Sigma, St. Louis, MO) in a 250 ml Erlenmeyer flask at room temperature with shaking (150 rpm) for 1 hour. The liquid fractions were centrifuged at 4000 × *g* for 15 minutes and the supernatant discarded. Pellets were resuspended in 200 µl of TE buffer and subjected to DNA extraction using the previously described bead-beating protocol (Hill *et al.* 2005b). DNA was quantified using a Quant-IT DNA quantification kit and Qubit fluorometer (Invitrogen, Burlington, Ontario).

Quantification of bacterial 16S rRNA-encoding genes.

To determine the total number of bacterial 16S rRNA-encoding genes associated with each seed lot, quantitative PCR was employed using universal primers SRV3-1 and SRV3-2 targeting nucleotides 330-533 (numbered according to *E. coli*) of the 16S rRNA gene (Lee *et al.* 1996). Reactions were prepared using SsoFast EvaGreen supermix (Bio-Rad, Mississauga, Ontario) with 400 nM of each primer in a final volume of 20 µl. Amplification conditions were: 95°C, 3 min (1x); followed by 30 cycles of 95°C, 15 sec, 62°C, 15 sec, 72°C, 15 sec. Data collection was set at the extension step. Results were expressed as 16S rRNA gene copies g^{-1} seeds by considering the weight of seeds used for extraction and the template volume used for qPCR.

cpn60 UT amplicon generation and sequencing.

Amplicons were generated from each sample using multiplexing ID (MID)-adapted universal primers as described previously (Schellenberg *et al.* 2009; Schellenberg *et al.* 2011b). Purified, concentrated amplicon from all seed samples was pooled on an equimolar basis prior to emPCR adaptor ligation and pyrosequencing using Titanium chemistry (Roche/454).

Assembly of Operational Taxonomic Units.

The de-multiplexing of pyrosequencing data was done as described previously (Chaban *et al.* 2012). OTUs were derived from the pyrosequencing data using sequence assembly, and OTU abundances were determined using the mPUMA software package (Links *et al.* 2012)(Chapter 3).

α-diversity measures.

To avoid biases introduced by unequal sampling effort (Gihring *et al.* 2012), OTU abundance data for each sample was sub-sampled at random to the size of the smallest library (3,606 reads). Calculation of community parameters including Chao1 richness, Simpson's index D, the Shannon-Weiner index (H'), and Good's coverage estimator was performed using mothur (Schloss *et al.* 2009).

Analysis of OTU abundance across crops.

Prior to analysis in R the OTU abundances were scaled to a library size of 10^7 to approximate the community size as measured by 16S rRNA copies g^{-1} for these samples

(Figure 4-1). Clustering and statistical tests based on OTU abundance were performed in R (version 2.15.1) on a Linux server (CentOS 5.8). Hierarchical clustering was performed using an average linkage method based on the Euclidean distance of both OTU and samples. OTU with significantly differential abundances were identified using an unpaired Mann-Whitney test followed by a Benjamini-Hochberg correction for multiple hypothesis testing at an alpha = 5% level of significance.

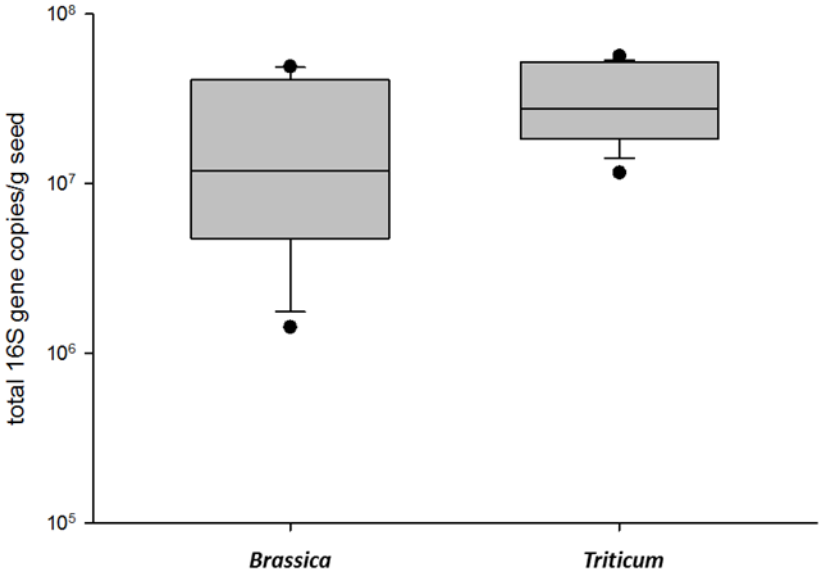


Figure 4-1 Total bacterial 16S rRNA gene counts as measured by quantitative PCR for *Brassica* and *Triticum* seed washes. The lower and upper edges of each box correspond to the 25th and 75th percentiles, while the whiskers correspond to the 10th and 90th percentiles. The median value is indicated by a horizontal line.

Quantitative PCR targeting specific microbes.

Primers designed to target specific OTU were designed using sigoligo (Zahariev *et al.* 2009), Beacon Designer 7 (Premier Biosoft, Palo Alto, CA, USA), and primer3 (Rozen and Skaletsky 2000). Primers targeting bacterial OTU00845 were 5'-CGG TAT TGA CCA GGC TGT TAT C-3' and 5'-AGT TCA ATC GCA CCG GTT T-3' (271 bp product). Amplification conditions used were 95°C, 3 min followed by 40 cycles of 95°C, 15 sec, 60°C, 15 sec, 72°C, 30 sec. Primers targeting fungal OTU03024 were 5'-GCT TGA GGT TAC CGA AGG-3' and 5'-GGA GAG GAG GAT CAG AGG-3' (112 bp product). Amplification conditions were 95°C, 3 min followed by 40 cycles of 95°C, 15 sec, 63°C, 15 sec, 72°C, 30 sec. For both assays, data collection was at the extension step (72°C). Quantitative PCR with SsoFast Eva Green Supermix (Bio-Rad) and primer concentrations of 400 nM each was used to determine the apparent genome number of each organism in each seed extract as described (Dumonceaux *et al.* 2006a).

Isolation and identification of microbes.

To isolate fungi from *Brassica* or *Triticum* seeds, a 4 g sample of seeds was incubated in 50 ml of Taylor minimal medium (Taylor 1993) or malt extract broth (Difco, Houston, TX), each containing antibiotics: tetracycline (100 μg ml^{-1}); streptomycin (100 μg ml^{-1}); and penicillin (1000 units ml^{-1}). Seed samples were incubated with shaking (150 rpm) at room temperature (20-23°C) for 4 days, then 100 μl of serial dilutions of the broth were plated on Taylor minimal medium or malt extract agar plates with antibiotics until colonies appeared. Some samples showed outgrowth in broth culture of large mycelial agglomerates; these were blended in a sterile Eberbach blender cup for 10 seconds prior

to dilution and plating. A similar strategy was used to isolate bacteria from *Triticum* seeds, except that 50 ml of antibiotic-free trypticase soy broth (Difco) was used as a culture medium and the cultures were incubated overnight at room temperature prior to dilution and plating on trypticase soy agar plates. DNA was extracted from each fungal strain using a miniprep method (Wendland *et al.* 1996) and from each bacterial strain using a Wizard genomic DNA extraction kit (Promega, Madison, WI). The *cpn*60 UT sequences of bacterial isolates were determined by direct sequencing of amplicons using M13-adapted universal primers H729/H730 as described previously (Goh *et al.* 2000). Sequences of the nuclear ribosomal internal transcribed spacer (ITS) were determined for each fungal isolate using PCR primers and amplification conditions as described (Schoch *et al.* 2012).

Phylogenetic analysis.

Full-length assembled OTU sequences were aligned with the *cpn60* sequences determined from the isolates as described above and with selected reference strains from cpnDB (Hill *et al.* 2004) using clustalw (Thompson *et al.* 1994). Phylogenetic trees were constructed using the neighbor-joining method (Saitou and Nei 1987) with bootstrapping of 500 replicates. Distances were calculated using the maximum composite likelihood method. Alignments were performed and trees were calculated using MEGA v5.05 (Tamura *et al.* 2007).

Biological interaction assays.

Triticum seeds (Canada Western Red Spring wheat, grade 3) were sterilized by submerging 30 g of seeds within a nylon bag in 250 ml of 95% ethanol for 20 seconds,

followed by 250 ml of 20% commercial bleach for 15 minutes with shaking. Seeds were then washed in 7×250 ml of sterile water (3 minutes for the first three washes and 10 minutes for the final four washes). Sterilized seeds were dried overnight in a sterile Petri dish. Seeds were re-colonized with the desired strains by diluting overnight cultures of each strain 1:100 in 5 ml of sterile peptone water, then adding ~50 seeds to each dilution. This inoculum corresponded to approximately $1.6x10^7$ cfu g^{-1} seeds. Control seeds were added to sterile peptone water without bacterial culture. The seeds were incubated at room temperature for 15 minutes with gentle agitation, and then placed in the center of plates containing Czapek-Dox agar medium (containing 30 g sucrose, 2 g sodium nitrate, 1 g dipotassium phosphate, 0.5 g each of $MgSO_4$ and KCl, and 0.01 g of $FeSO_4$ $liter^{-1}$). A 5 mm punchout from the edge of a colony of *Leptosphaeria maculans* strain WA51 (Yu *et al.* 2005) or of fungal isolate 15 was placed within 3 cm of the seeds and the plates were incubated at 25°C for one week. Inhibition of fungal growth was scored using previously described methods (Chakraborty *et al.* 1994).

Results

Total 16S rRNA-encoding gene counts.

The total 16S rRNA gene copy number associated with each *Triticum* seed type varied over a range of approximately 4-fold, with Wheat-4 (CWRS grade 1) being the lowest and Wheat-2 (CWAD grade 3) the highest (Figure 4-1). The range was somewhat wider (approximately 9-fold) within the *Brassica* seeds, with Brassica-5 (*B. rapa*) being the lowest and Brassica-4 (oriental mustard) the highest. Although the *Triticum* samples tended to have higher 16S rRNA gene counts than the *Brassica* samples, no statistically

significant differences were detected at a significance level of 0.01 (p=0.018, Mann-Whitney rank sum test).

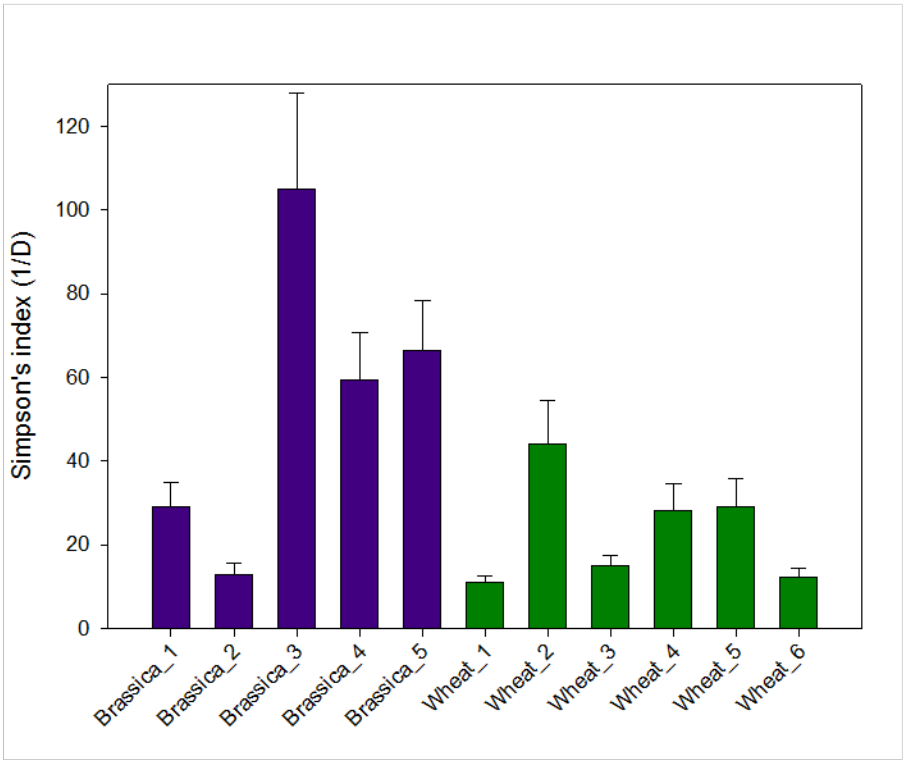
Pyrosequencing of *cpn*60 UT amplicons.

A total of 408,658 reads was generated from the 11 amplicon libraries. The median library size was 34,594 with a range of 3,606 reads (*B. rapa*) to 96,834 reads (CWRS grade 3). These reads were assembled into 5,477 distinct OTU.

Microbial community diversity.

Community richness (Chao1, expressed as the projected total number of OTU in each sample), evenness (Simpson's index, D) and the Shannon index H' (Hill *et al.* 2003) was calculated for each sample. No correlation was observed between community richness or evenness, and total bacterial 16S rRNA gene copy numbers (Spearman rank correlation). Comparing the microbial communities associated with *Triticum* and *Brassica* seeds revealed no significant differences in the diversity parameters by Mann-Whitney test and one-way ANOVA (Figure 4-2).

A


Simpson's index (1/D)
0
20
40
60
80
100
120
Brassica_1
Brassica_2
Brassica_3
Brassica_4
Brassica_5
Wheat_1
Wheat_2
Wheat_3
Wheat_4
Wheat_5
Wheat_6

B

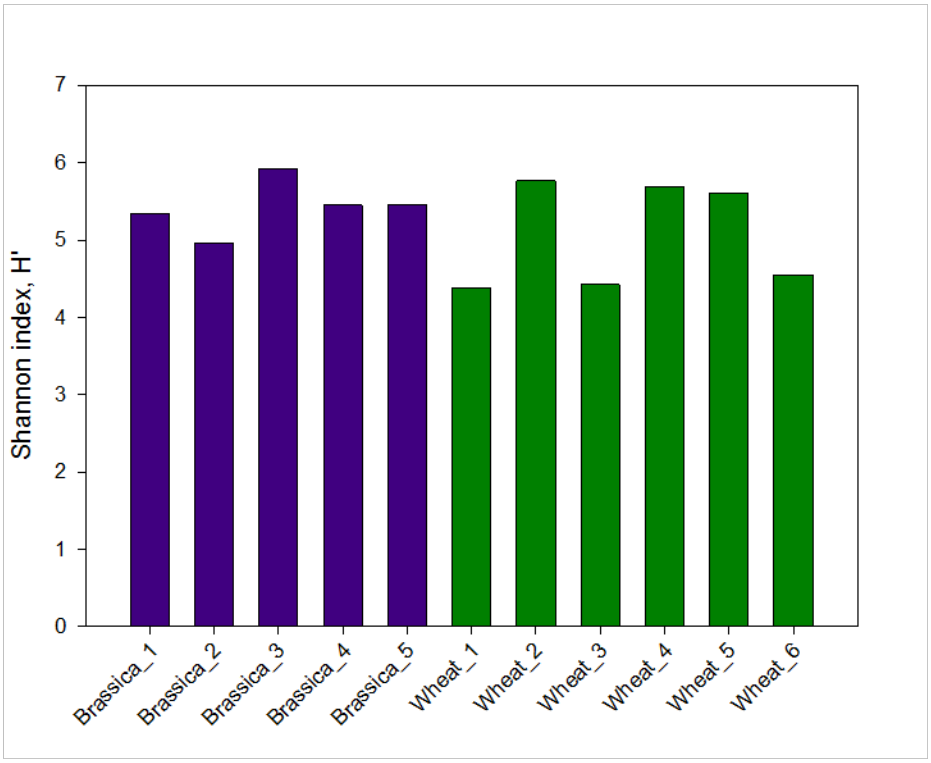

Shannon index, H'
0
1
2
3
4
5
6
7
Brassica_1
Brassica_2
Brassica_3
Brassica_4
Brassica_5
Wheat_1
Wheat_2
Wheat_3
Wheat_4
Wheat_5
Wheat_6

Figure 4-2 Community statistics for *Triticum* and *Brassica* seed samples. A. Simpson's index (1/D), which measures community evenness. Error bars represent the 95% confidence intervals of the data. B. Shannon index (H'). No statistically significant differences were detected between *Brassica* and *Triticum* samples for any of these community diversity measures (Mann-Whitney rank sum test, $p > 0.05$).

The shared epiphytic microbiota of *Triticum* and *Brassica* seeds.

Microbial profiles determined for *Triticum* and *Brassica* samples were compared, resulting in the identification of a core microbiome for each host plant genus. All *Triticum* (n=6) samples had 262 OTU in common while all *Brassica* (n=5) samples had 215 OTU in common. In order to identify the microbiota shared between seeds of *Brassica* and *Triticum* we established a sample prevalence of at least 7 / 11 as a lower limit for an OTU to be considered shared. This would ensure that any OTU identified as shared was observed in at least one sample of each host genus. There were 578 OTU identified with a sample prevalence of 7 / 11 or higher. Additionally we determined whether there were any OTU found in all samples. Across host plant species 64 OTU were detected in every sample. We examined the effect of sample size (per host plant genus) on the number of OTU identified as shared. The number of shared OTU was calculated for each combination of *Brassica* and *Triticum* samples (from 1 to 5 samples for each host plant genus). The number of shared OTU diminished as sample size increased in a non-linear fashion, suggesting an asymptote around 60 OTUs (Figure 4-3). These results are consistent with the identification of a shared microbiome at the sample size used in this study and suggest that larger sample numbers would not substantially decrease the size of the shared microbiome.

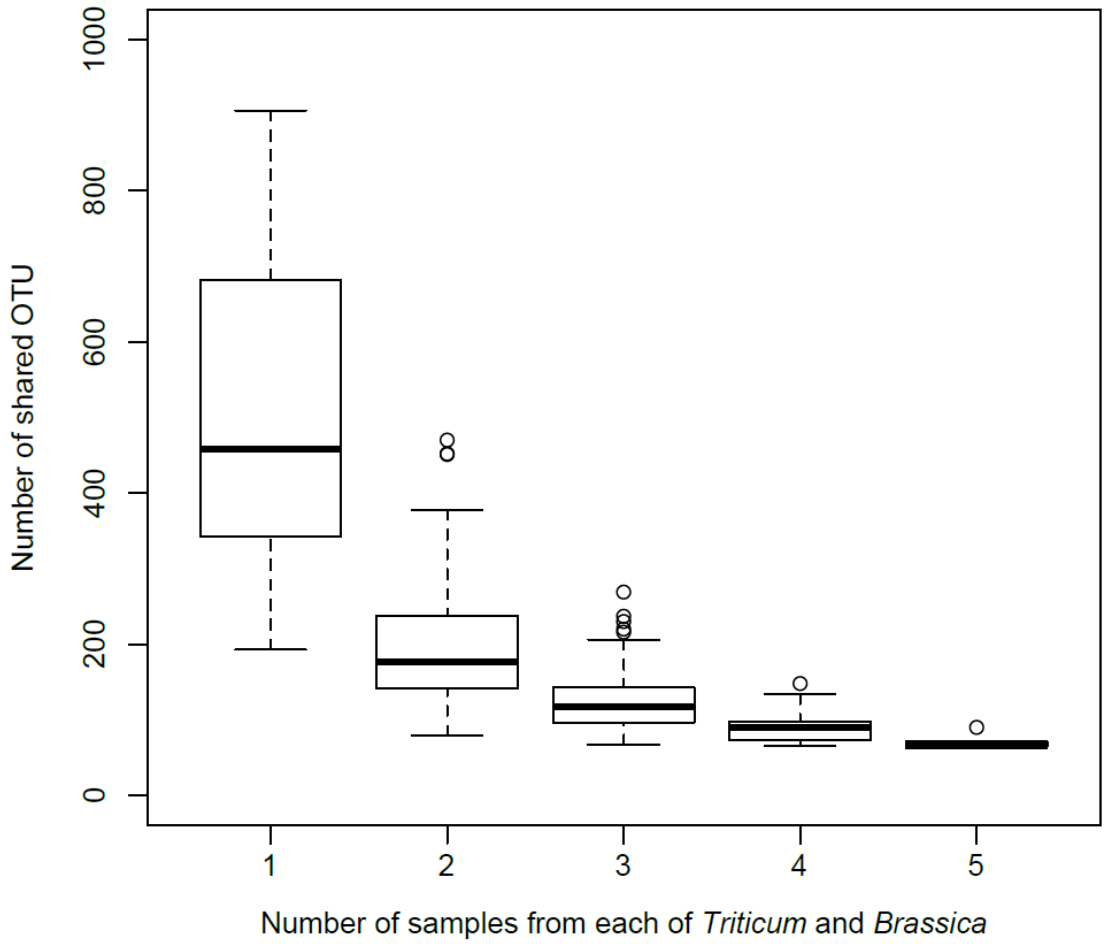


Figure 4-3 Determination of the *Triticum*/*Brassica* seed-associated shared OTU with increasing paired sample size. All possible combinations of *Triticum* and *Brassica* libraries were compared at each sample size (e.g. 1 *Triticum*-1 *Brassica*; 2 *Triticum*-2 *Brassica*, *etc*.) and the number of OTU that were observed in all samples of both *Triticum* and *Brassica* was determined. The median value for each data point is shown by a horizontal line, and the outer edges represent the 10^{th} and 90^{th} percentiles of the data.

Sequences for these 64 shared OTU were similar but not identical (88-99% identity) to records from cpnDB that included matches to *Pantoea agglomerans* (99%), *Massilia timonae* (93%), *Pantoea stewartii* (93%), *Porphyrobacter sanguineus* (88%), *Pseudomonas fluorescens* (97%), *Pseudomonas syringae* (95%), *Pyrenophora tritici-repentis* (93%), *Sphingobium japonicum* (90%), *Sphingomonas wittichii* (90%), *Telluria mixta* (93%), *Xanthomonas axonopodis* (94%), *Xanthomonas fuscans* (95%) and some novel sequences.

Differential abundance within the epiphytic microbiota of *Triticum* and *Brassica* seeds.

Hierarchical clustering of the microbial profiles showed that the *Triticum* and *Brassica*-derived samples could be separated on the basis of the 578 shared OTU (Figure 4-4). A Mann-Whitney test identified 203 of these OTU that were significantly differentially abundant between *Triticum* and *Brassica,* including all 64 OTU with a sample prevalence of 11 / 11 (Table 4-2;Table 4-3).

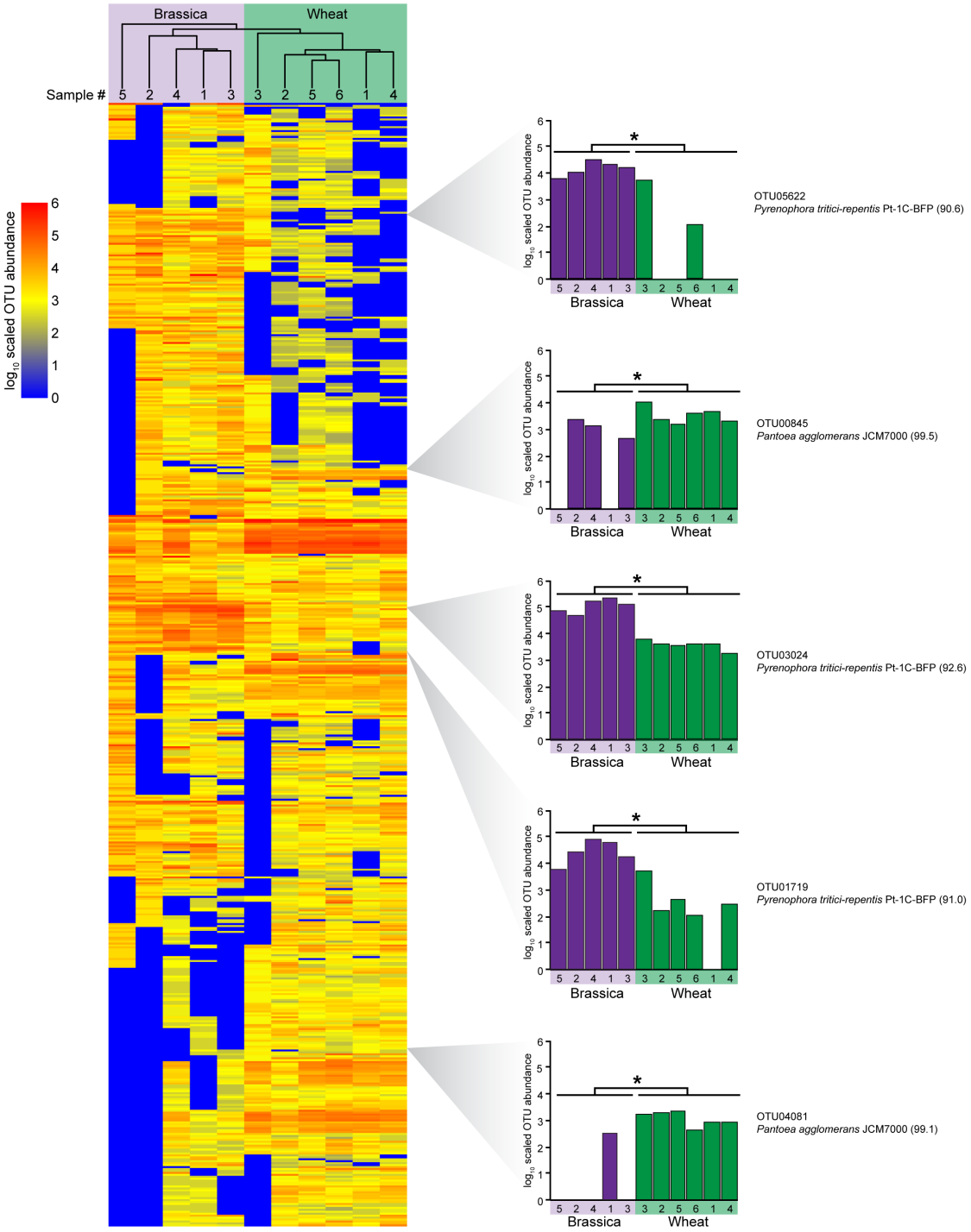


Figure 4-4 Hierarchical clustering of samples and OTU from crop seeds. These 578 OTU were found in at least 7/11 samples from the two seed types. Libraries are represented by columns while OTU are represented by rows. Abundances of each OTU are presented as a heat map (blue, less abundant to red, more abundant). Specific OTU corresponding to cultured isolates are identified along with their corresponding read abundances in each library and cpnDB nearest neighbor (with percent identities indicated in parentheses).

Table 4-2 OTU found to have a significantly higher abundance on seeds *Brassica* spp. vs. *Triticum* spp.

OTU	*p* value	Best match from cpnDB (Hill *et al.* 2004)	Identity (%)	Length (bp)
00868	0.028	b12096 XM_001931520 *Pyrenophora tritici-repentis* Pt-1C-BFP	91.5	577
00947	0.028	b12096 XM_001931520 *Pyrenophora tritici-repentis* Pt-1C-BFP	91.0	555
01356	0.028	b12096 XM_001931520 *Pyrenophora tritici-repentis* Pt-1C-BFP	91.2	612
01719[1]	0.028	b12096 XM_001931520 *Pyrenophora tritici-repentis* Pt-1C-BFP	91.0	586
02679	0.028	b12096 XM_001931520 *Pyrenophora tritici-repentis* Pt-1C-BFP	91.2	555
02750	0.028	b12096 XM_001931520 *Pyrenophora tritici-repentis* Pt-1C-BFP	92.6	555
02863	0.028	b12096 XM_001931520 *Pyrenophora tritici-repentis* Pt-1C-BFP	89.2	471
03024[2]	0.028	b12096 XM_001931520 *Pyrenophora tritici-repentis* Pt-1C-BFP	92.6	580
03573	0.028	b12096 XM_001931520 *Pyrenophora tritici-repentis* Pt-1C-BFP	87.2	555
03644	0.028	b12096 XM_001931520 *Pyrenophora tritici-repentis* Pt-1C-BFP	86.9	580
05457	0.028	b12096 XM_001931520 *Pyrenophora tritici-repentis* Pt-1C-BFP	92.3	467
05622[3]	0.028	b12096 XM_001931520 *Pyrenophora tritici-repentis* Pt-1C-BFP	90.6	555
06305	0.028	b12523 NC_011144 *Phenylobacterium zucineum* HLK1	85.4	552
02815	0.028	b12837 NC_012791 *Variovorax paradoxus* S110	92.4	554
01363	0.028	b14162 NC_012778 *Eubacterium eligens* ATCC 27750	41.0	384
01435	0.028	b14498 NZ_ACVD01000045 *Acidithiobacillus caldus* ATCC 51756	43.0	492
04904	0.028	b16011 CP001854 *Conexibacter woesei* DSM 14684	88.5	452
05240	0.028	b16011 CP001854 *Conexibacter woesei* DSM 14684	85.1	433
00636	0.028	b16615 AP010803 *Sphingobium japonicum* UT26S NBRC 101211	89.4	558
01178	0.028	b16615 AP010803 *Sphingobium japonicum* UT26S NBRC 101211	89.5	465
01227	0.028	b16615 AP010803 *Sphingobium japonicum* UT26S NBRC 101211	89.4	555
01345	0.028	b16615 AP010803 *Sphingobium japonicum* UT26S NBRC 101211	90.2	486
01502	0.028	b16615 AP010803 *Sphingobium japonicum* UT26S NBRC 101211	89.5	555
02199	0.028	b16615 AP010803 *Sphingobium japonicum* UT26S NBRC 101211	89.5	555
02495	0.028	b16615 AP010803 *Sphingobium japonicum* UT26S NBRC 101211	89.9	554
02528	0.028	b16615 AP010803 *Sphingobium japonicum* UT26S NBRC 101211	89.0	555
02643	0.028	b16615 AP010803 *Sphingobium japonicum* UT26S NBRC 101211	90.0	505
05162	0.028	b16615 AP010803 *Sphingobium japonicum* UT26S NBRC 101211	88.6	566
04161	0.028	b17405 GG774665 *Pseudomonas savastanoi* pv. *savastanoi* NCPPB	94.1	554
01373	0.028	b17482 XM_001840049 *Coprinopsis cinerea* okayama7#130	69.0	622
01342	0.028	b17665 AB547563 *Butyricimonas virosa* JCM 15149	45.0	449
06099	0.028	b18163 AP010968 *Kitasatospora setae* KM-6054	87.9	576
03642	0.028	b18946 AEAO01000548 *Pseudomonas syringae* pv. *aceris* str.	95.3	555
00301	0.028	b19041 CP002727 *Pseudomonas fulva* 12-X	44.4	335
01754	0.028	b19232 CP002897 *Paracoccus denitrificans* SD1	87.9	570
04791	0.028	b19627 JF745945 *Tetrasphaera vanveenii* DSM 17518	48.3	497
00080	0.028	b19629 JF745943 *Tetrasphaera duodecadis* DSM 12806	52.2	160
06861	0.028	b20425 NZ_CAGB01000016 *Wolbachia pipientis* wAlbB	42.7	481
03434	0.028	b20442 NC_016887 *Nocardia cyriacigeorgica* GUH-2	83.0	553
02318	0.028	b21604 AMQP01000036 *Pseudomonas viridiflava* UASWS0038	97.8	585
01073	0.028	b21890 AGZI01000020 *Massilia timonae* CCUG 45783	92.9	445
02542	0.028	b21890 AGZI01000020 *Massilia timonae* CCUG 45783	91.7	458
02748	0.028	b21890 AGZI01000020 *Massilia timonae* CCUG 45783	92.8	558
02869	0.028	b21890 AGZI01000020 *Massilia timonae* CCUG 45783	92.1	552
03058	0.028	b21890 AGZI01000020 *Massilia timonae* CCUG 45783	91.5	458
01377	0.028	b21893 AGZU01000014 *Sphingobium yanoikuyae* ATCC 51230	88.3	471
03149	0.028	b21893 AGZU01000014 *Sphingobium yanoikuyae* ATCC 51230	89.3	475
01809	0.028	b22217 ANIU01000043 *Rhodococcus wratislaviensis* IFP 2016	87.5	556

OTU	*p* value	Best match from cpnDB (Hill *et al.* 2004)	Identity (%)	Length (bp)
02808	0.028	b3420 AY123661 *Pseudomonas fluorescens* ATCC 13525	95.1	555
03647	0.028	b3420 AY123661 *Pseudomonas fluorescens* ATCC 13525	92.8	555
06166	0.028	b5482 NC_002937 *Desulphovibrio vulgaris* subsp. *vulgaris* strain	40.4	357
00061	0.028	b6878 AADS00000000 *Phanerochaete chrysosporium* RP-78	40.2	379
00929	0.028	b6878 AADS00000000 *Phanerochaete chrysosporium* RP-78	41.2	392
00738	0.028	b7280 AY837539 *Phoma pomorum* DAOM172382	38.7	438
03118	0.028	b7361 AJ716085 *Botrytis hyacinthi* MUCL442	39.8	508
04877	0.028	b8480 CP000781 *Xanthobacter autotrophicus* Py2	41.4	599
06892	0.028	v6236 j0614 *Fusarium equiseti* 6	100.0	555
01011	0.035	b12423 EU790571 *Porphyrobacter sanguineus* ATCC 25659	88.4	402
01848	0.035	b16615 AP010803 *Sphingobium japonicum* UT26S NBRC 101211	89.6	578
00524	0.035	b18946 AEAO01000548 *Pseudomonas syringae* pv. *aceris* str.	94.2	728
00525	0.035	b18946 AEAO01000548 *Pseudomonas syringae* pv. *aceris* str.	94.2	707
04621	0.035	b18985 AB627073 *Methylobacterium marchantiae* JT1	96.0	555
02566	0.035	b21890 AGZI01000020 *Massilia timonae* CCUG 45783	92.8	578
02536	0.036	b12429 EU790577 *Telluria mixta* ATCC 49108	92.1	555
02926	0.036	b12837 NC_012791 *Variovorax paradoxus* S110	91.3	428
04856	0.036	b16615 AP010803 *Sphingobium japonicum* UT26S NBRC 101211	88.8	465
01523	0.036	b13298 NC_011988 *Agrobacterium vitis* S4	92.6	554
00516	0.036	b20091 BAED01000033 *Gordonia amarae* NBRC 15530	44.3	362
01734	0.049	b10265 EF685238 *Rhodococcus fascians* ATCC 12974	99.5	457
03097	0.049	b10265 EF685238 *Rhodococcus fascians* ATCC 12974	99.6	577
02008	0.049	b10915 AM849034 *Clavibacter michiganensis* subsp. *spedonicus*	90.2	552
01349	0.049	b12429 EU790577 *Telluria mixta* ATCC 49108	93.0	553
06404	0.049	b1263 AY263151 *Renibacterium salmoninarum* ATCC 33209	43.3	646
01947	0.049	b12784 NS_000195 *Candidatus Cloacamonas acidaminovorans*	45.5	657
00869	0.049	b15304 NC_013521 *Sanguibacter keddieii* DSM 10542	99.8	567
01803	0.049	b16615 AP010803 *Sphingobium japonicum* UT26S NBRC 101211	89.2	555
03028	0.049	b16615 AP010803 *Sphingobium japonicum* UT26S NBRC 101211	89.4	768
03402	0.049	b16615 AP010803 *Sphingobium japonicum* UT26S NBRC 101211	89.4	555
02196	0.049	b17885 FN298395 *Erwinia rhapontici* WMR127	94.4	553
03415	0.049	b17885 FN298395 *Erwinia rhapontici* WMR127	91.7	556
06254	0.049	b21890 AGZI01000020 *Massilia timonae* CCUG 45783	93.0	555
00582	0.049	b3420 AY123661 *Pseudomonas fluorescens* ATCC 13525	93.9	618
01222	0.049	b3420 AY123661 *Pseudomonas fluorescens* ATCC 13525	94.6	558
04236	0.049	b415 AB008150 *Pantoea agglomerans* JCM7000	82.8	364
00813	0.049	b9562 CP000699 *Sphingomonas wittichii* RW1	89.9	580

[1] 100% identical to fungal isolate #6
[2] 100% identical to fungal isolate #15
[3] 100% identical to fungal isolate #9

Table 4-3 OTU found to have a significantly higher abundance on seeds *Triticum* spp. vs. *Brassica* spp.

OTU	*p* value	Best match from cpnDB (Hill *et al.* 2004)	Identity (%)	Length (bp)
06191	0.028	b12430 EU790578 *Pseudomonas vancouverensis* ATCC 700688	40.7	337
04490	0.028	b12771 NZ_ABYT01000057 *Eubacterium biforme* DSM 3989	42.6	500
01591	0.028	b13605 NC_013530 *Xylanimonas cellulosilytica* DSM 15894	46.0	547
03501	0.028	b14106 NZ_ACIP02000002 *Shuttleworthia satelles* DSM 14600	42.8	652
07027	0.028	b18763 XM_003226995 *Anolis carolinensis*	44.3	526
02563	0.028	b18946 AEAO01000548 *Pseudomonas syringae* pv. aceris str. M302273PT	99.1	556
04911	0.028	b18946 AEAO01000548 *Pseudomonas syringae* pv. aceris str. M302273PT	98.6	556
01371	0.028	b18982 NZ_AFGG01000024 *Sphingomonas sp.* S17	39.3	332
01867	0.028	b19675 AGFC01000008 *Thiocystis violascens* DSM 198	44.8	466
00630	0.028	b21985 GACK01004998 *Rhipicephalus pulchellus*	45.1	431
00047	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	99.1	345
00294	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.4	554
00424	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	99.3	591
00589	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.9	584
00657	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	99.3	443
00859	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.7	463
00963	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	99.3	437
01183	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.8	402
01187	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.5	406
01453	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.4	575
01575	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	99.3	663
01700	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.3	484
02231	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.7	452
02235	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.5	389
02331	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	99.3	558
02347	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	96.2	556
02397	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.5	342
02456	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.0	412
02457	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.3	563
02508	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	99.3	555
02625	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	99.3	458
02702	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	99.1	458
02703	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	99.1	430
02737	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.2	434
02856	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	97.9	429
02983	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	99.1	560
03382	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.7	444
03396	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	99.0	394
03408	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	97.8	673
03692	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.2	433
03734	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.9	567
03806	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	99.1	568
04062	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.7	444
04081[1]	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	99.1	555
04119	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	99.5	456
04145	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.6	584
04256	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.8	363

OTU	*p* value	Best match from cpnDB (Hill *et al.* 2004)	Identity (%)	Length (bp)
04326	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	99.3	434
04519	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.7	383
05148	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	99.1	454
05833	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	96.2	555
06129	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.0	551
06438	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.8	340
06542	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.0	553
06543	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.9	562
06558	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.5	341
06591	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	98.1	365
06594	0.028	b415 AB008150 *Pantoea agglomerans* JCM7000	99.0	400
03082	0.028	b7564 NC_007963 *Chromohalobacter salexigens* DSM 3043	42.2	336
03366	0.028	b7564 NC_007963 *Chromohalobacter salexigens* DSM 3043	44.7	341
02123	0.028	b7834 AF429666 *Lactobacillus acetotolerans* ATCC 43578	43.7	337
03768	0.028	b7834 AF429666 *Lactobacillus acetotolerans* ATCC 43578	41.5	474
04210	0.028	b7834 AF429666 *Lactobacillus acetotolerans* ATCC 43578	40.5	373
05844	0.028	b7834 AF429666 *Lactobacillus acetotolerans* ATCC 43578	42.8	337
00411	0.028	b858 AY004281 *Scardovia inopinata* DSM10107	45.8	336
05486	0.028	b9351 NC_009620 *Sinorhizobium medicae* WSM419	42	352
00814	0.028	v4741 j0500 *Sinorhizobium meliloti* ATCC 9930	41.3	355
02120	0.028	v4741 j0500 *Sinorhizobium meliloti* ATCC 9930	43.8	336
03556	0.028	v5248 j0528 *Cylindrocarpon destructans*	41.9	578
03335	0.035	b12430 EU790578 *Pseudomonas vancouverensis* ATCC 700688	40.5	343
03315	0.035	b18946 AEAO01000548 *Pseudomonas syringae* pv. aceris str. M302273PT	98.6	557
02448	0.035	b415 AB008150 *Pantoea agglomerans* JCM7000	99.5	554
02871	0.035	b415 AB008150 *Pantoea agglomerans* JCM7000	98.6	360
04109	0.035	b415 AB008150 *Pantoea agglomerans* JCM7000	98.5	389
04471	0.035	b415 AB008150 *Pantoea agglomerans* JCM7000	99.2	387
06423	0.035	b415 AB008150 *Pantoea agglomerans* JCM7000	99.5	399
04918	0.035	b7584 CP000356 *Sphingopyxis alaskensis* RB2256	44.5	550
00785	0.035	b7834 AF429666 *Lactobacillus acetotolerans* ATCC 43578	41.1	335
03054	0.035	b862 AY004277 *Bifidobacterium merycicum* JCM8219	41.5	273
05522	0.035	b862 AY004277 *Bifidobacterium merycicum* JCM8219	42.3	484
00235	0.035	b9351 NC_009620 *Sinorhizobium medicae* WSM419	40	337
00512	0.035	b9351 NC_009620 *Sinorhizobium medicae* WSM419	42.5	335
00758	0.035	b9351 NC_009620 *Sinorhizobium medicae* WSM419	42.4	333
00423	0.035	v4741 j0500 *Sinorhizobium meliloti* ATCC 9930	42.8	355
06331	0.036	b19637 JF745935 *Beggiatoa alba* DSM 1416	45.8	722
00685	0.036	b415 AB008150 *Pantoea agglomerans* JCM7000	99.3	555
01208	0.036	b415 AB008150 *Pantoea agglomerans* JCM7000	98.6	430
01237	0.036	b415 AB008150 *Pantoea agglomerans* JCM7000	98.5	399
02513	0.036	b415 AB008150 *Pantoea agglomerans* JCM7000	99.1	553
02872	0.036	b415 AB008150 *Pantoea agglomerans* JCM7000	98.7	378
04166	0.036	b415 AB008150 *Pantoea agglomerans* JCM7000	98.9	554
06516	0.036	b415 AB008150 *Pantoea agglomerans* JCM7000	98.6	428
06590	0.036	b415 AB008150 *Pantoea agglomerans* JCM7000	98.2	429
03319	0.036	b7834 AF429666 *Lactobacillus acetotolerans* ATCC 43578	43.9	346
00244	0.036	b415 AB008150 *Pantoea agglomerans* JCM7000	99.3	554

OTU	*p* value	Best match from cpnDB (Hill *et al.* 2004)	Identity (%)	Length (bp)
00854	0.036	b415 AB008150 *Pantoea agglomerans* JCM7000	98.7	552
03520	0.036	b415 AB008150 *Pantoea agglomerans* JCM7000	98.6	416
03968	0.036	b415 AB008150 *Pantoea agglomerans* JCM7000	99.3	554
00407	0.036	v4741 j0500 *Sinorhizobium meliloti* ATCC 9930	42.7	458
02858	0.038	b415 AB008150 *Pantoea agglomerans* JCM7000	99.3	668
05044	0.049	b16950 CP002049 *Truepera radiovictrix* DSM 17093	39.4	395
03327	0.049	b22216 CAPJ01000179 *Xanthomonas translucens* pv. translucens DSM	98.9	692
01355	0.049	b3420 AY123661 *Pseudomonas fluorescens* ATCC 13525	95.7	553
00409	0.049	b415 AB008150 *Pantoea agglomerans* JCM7000	99.3	555
00845[2]	0.049	b415 AB008150 *Pantoea agglomerans* JCM7000	99.5	555
00958	0.049	b415 AB008150 *Pantoea agglomerans* JCM7000	96.3	267
01030	0.049	b415 AB008150 *Pantoea agglomerans* JCM7000	97.7	340
01163	0.049	b415 AB008150 *Pantoea agglomerans* JCM7000	99.1	430
01374	0.049	b415 AB008150 *Pantoea agglomerans* JCM7000	99.5	630
02189	0.049	b415 AB008150 *Pantoea agglomerans* JCM7000	99.3	428
03480	0.049	b415 AB008150 *Pantoea agglomerans* JCM7000	98.3	398
03683	0.049	b415 AB008150 *Pantoea agglomerans* JCM7000	99.1	573
03710	0.049	b415 AB008150 *Pantoea agglomerans* JCM7000	98.9	433
04198	0.049	b415 AB008150 *Pantoea agglomerans* JCM7000	98.6	432
04477	0.049	b415 AB008150 *Pantoea agglomerans* JCM7000	98.6	419
00220	0.049	b7760 CP000453 *Alkalilimnicola ehrlichei* MLHE-1 ATCC BAA-1101	44.5	336
00488	0.049	b7834 AF429666 *Lactobacillus acetotolerans* ATCC 43578	41.8	618

[1] 99% identical to bacterial isolate #8 (2 bp different)
[2] 100% identical to bacterial isolates #1-7 and #9

Approximately 40% of the significantly differentially abundant OTU (79 / 203, including OTU00845) were closely related (96-99% nucleotide identity) to *Pantoea agglomerans*. *Triticum* seeds were found to have significantly more sequences from *Pantoea-like* OTU than *Brassica* seeds (Table 4-2;Table 4-3). Fungal OTU were also identified as significantly different in abundance between *Brassica* and *Triticum* samples with 12 OTU (including OTU03024) more abundant on *Brassica* seeds as compared to the *Triticum* seeds. These fungal OTU were more similar (up to 99% identity) to a truncated *cpn*60 UT sequence from *Alternaria alternata* (GenBank GQ871196).

Quantitative PCR assays targeting OTU00845 and OTU03024 validated the sequence read abundance patterns seen in the microbial profiles of *Triticum* and *Brassica* (Figure 4-5). Within *Triticum* samples, the *Pantoea-like* OTU00845 was more abundant than the *Alternaria*-like OTU03024 with the inverse relationship observed in *Brassica* samples. Between crops, the *Pantoea-like* OTU00845 was significantly more abundant on *Triticum* compared to *Brassica*. Consistent with the sequencing read counts, the *P. agglomerans* OTU were significantly more abundant on *Triticum* seeds than *Brassica,* while the *Alternaria*-like OTU03024 exhibited an inverse pattern, being more abundant on *Brassica* seeds.

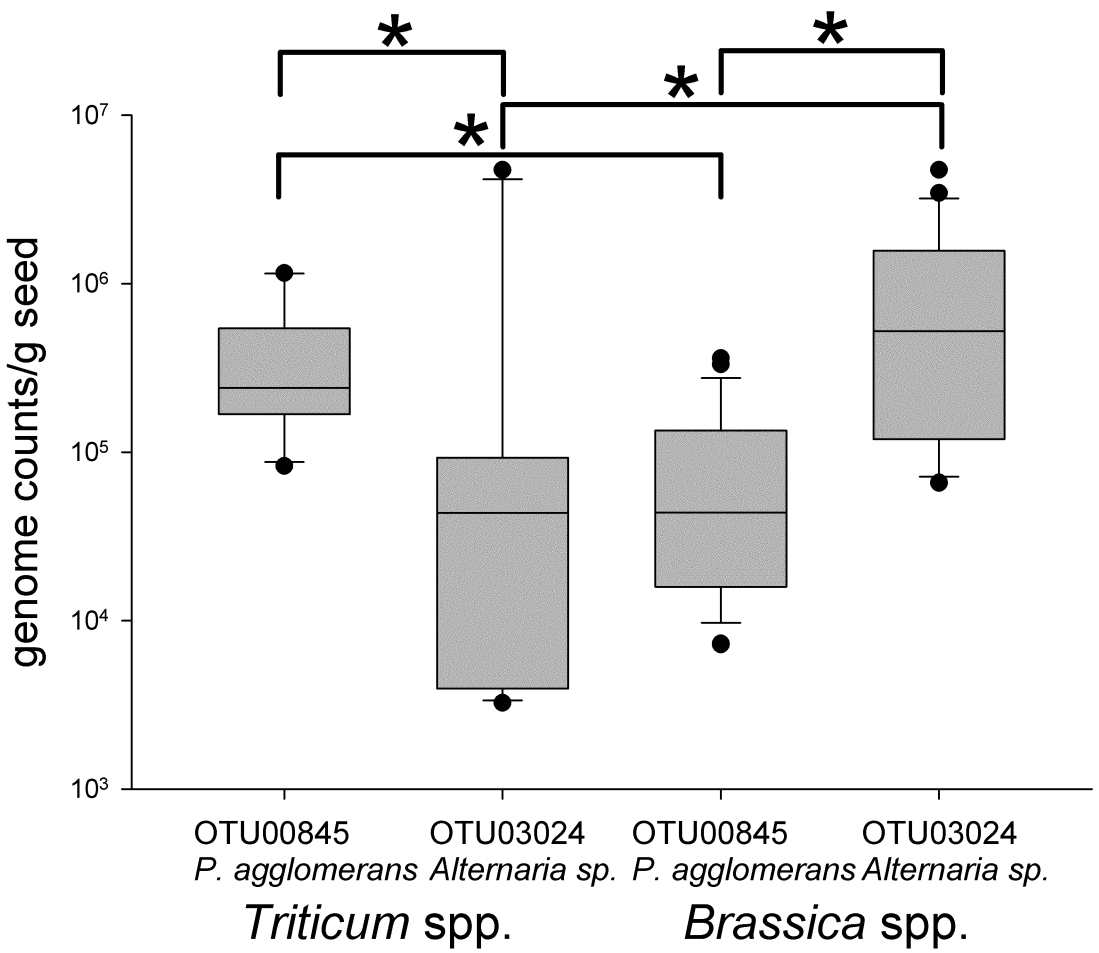
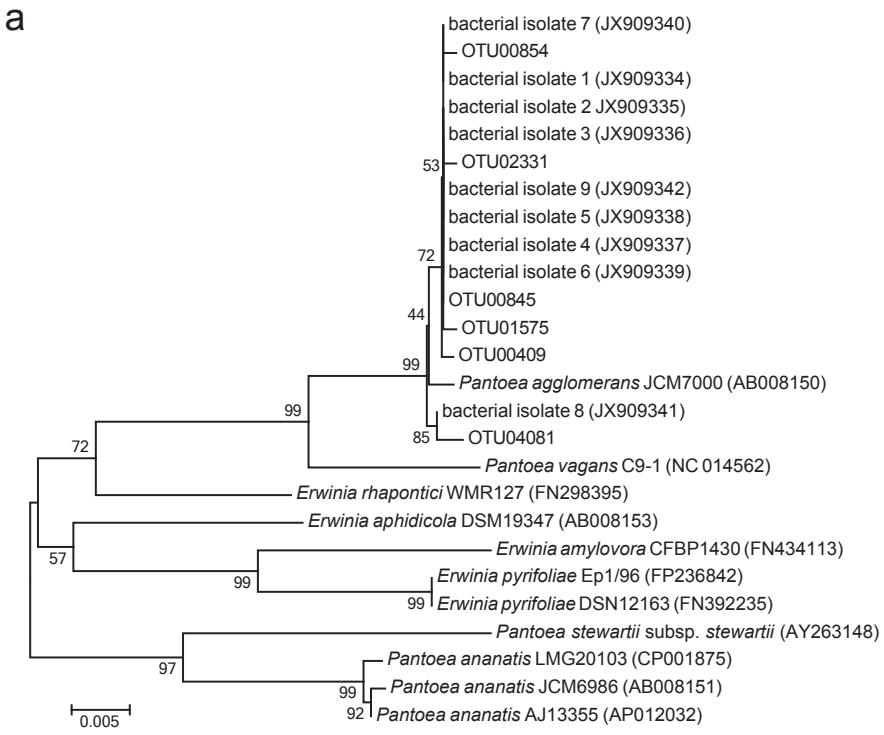


Figure 4-5 Quantification by qPCR of OTU00845 (*Pantoea agglomerans*) and OTU03024 (*Alternaria* sp.) on seeds of *Triticum* (n=12) and *Brassica* (n=24). qPCR results are from at least 2 biological replicates (DNA extractions) and 2 technical replicates per sample. Significant differences in the median values measured by the Mann-Whitney rank-sum test ($p < 0.01$) are indicated (*).

Isolation of bacteria and fungi from *Triticum* and *Brassica* seeds.

To assess potential interactions between members of the shared microbiota, we undertook efforts to culture bacteria corresponding to *Pantoea-like* OTU00845 and fungi corresponding to *Alternaria*-like OTU03024. These specific OTU were targeted due to their reciprocal patterns of abundance in both the microbial profiling and qPCR results (i.e. on seeds where *Pantoea-like* OTU were abundant, *Alternaria*-like OTU were rare and *vice versa* - Figure 4-5 and Table 4-2;Table 4-3). Multiple bacterial colony morphologies were observed when *Triticum* seeds were incubated in trypticase soy broth. Nine yellow colonies were picked from these plates, and these yielded a band with the *P. agglomerans cpn*60 UT-specific primer set and were sub-cultured to purity. Microscopic analysis revealed that the organisms were Gram-negative rods and they formed yellow colonies on the trypticase soy agar plates, consistent with previous reports for *Pantoea agglomerans* (Lee and Liu 1991). Determination of the *cpn*60 UT sequences for all 9 isolates revealed > 99% sequence identity to *P. agglomerans* JCM7000 and that 8 of these were 100% identical with each other, and OTU00845. The isolate sequences clustered together with the OTU sequences and all were distinct from the *P. agglomerans* reference strain as well as from other *Pantoea* species (Figure 4-6).


a
bacterial isolate 7 (JX909340)
OTU00854
bacterial isolate 1 (JX909334)
bacterial isolate 2 JX909335)
bacterial isolate 3 (JX909336)
OTU02331
bacterial isolate 9 (JX909342)
bacterial isolate 5 (JX909338)
bacterial isolate 4 (JX909337)
bacterial isolate 6 (JX909339)
OTU00845
OTU01575
OTU00409
Pantoea agglomerans JCM7000 (AB008150)
bacterial isolate 8 (JX909341)
OTU04081
Pantoea vagans C9-1 (NC 014562)
Erwinia rhapontici WMR127 (FN298395)
Erwinia aphidicola DSM19347 (AB008153)
Erwinia amylovora CFBP1430 (FN434113)
Erwinia pyrifoliae Ep1/96 (FP236842)
Erwinia pyrifoliae DSN12163 (FN392235)
Pantoea stewartii subsp. stewartii (AY263148)
Pantoea ananatis LMG20103 (CP001875)
Pantoea ananatis JCM6986 (AB008151)
Pantoea ananatis AJ13355 (AP012032)
0.005

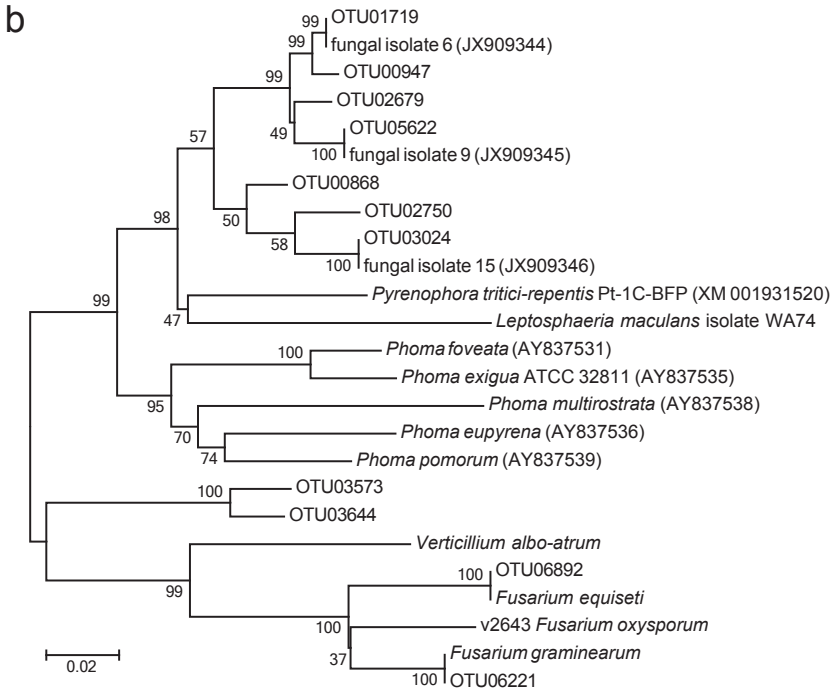

b
OTU01719
fungal isolate 6 (JX909344)
OTU00947
OTU02679
OTU05622
fungal isolate 9 (JX909345)
OTU00868
OTU02750
OTU03024
fungal isolate 15 (JX909346)
Pyrenophora tritici-repentis Pt-1C-BFP (XM 001931520)
Leptosphaeria maculans isolate WA74
Phoma foveata (AY837531)
Phoma exigua ATCC 32811 (AY837535)
Phoma multirostrata (AY837538)
Phoma eupyrena (AY837536)
Phoma pomorum (AY837539)
OTU03573
OTU03644
Verticillium albo-atrum
OTU06892
Fusarium equiseti
v2643 Fusarium oxysporum
Fusarium graminearum
OTU06221
0.02

Figure 4-6 Phylogenetic analysis of the *cpn*60 UT sequences of selected OTU assembled from pyrosequencing data along with reference strains from cpnDB and isolates from *Triticum* and *Brassica* seeds. In both **a** and **b**, the robustness of each node is indicated by the percentage of 500 trees in which the associated taxa cluster together and is presented next to the branches (Tamura *et al.* 2004). The scale bar represents units of base substitutions per site. Sequences corresponding to the *cpn*60 UT of reference strains were retrieved from cpnDB with the nucleotide accession number (ncbi.nlm.nih.gov) for each strain indicated in parentheses. **a.** *P. agglomerans*-related OTU, reference strains, and isolate sequences. **b.** Fungal isolates and OTU along with reference strain sequences from cpnDB.

Fungi were also isolated from *Brassica* and *Triticum* seeds, with a wide range of colony morphologies observed, including yeasts, molds, and filamentous phenotypes. Fungal isolates 6, 9 and 15 had *Alternaria*-like colony morphology and their *cpn*60 UT sequences were identical to OTU01719, OTU05622 and OTU03024, respectively (Figure 4-6) The *cpn*60 UT sequence of a fourth isolate with similar morphology (fungal isolate 5) was 1 bp different from non-significant OTU02724 in a short homopolymer (not shown). The *cpn*60 UT sequences of fungal isolates 5, 6, and 9 (and of OTU 02724, 05622, and 01719) shared 96-99% identity over 483 bp with a truncated *cpn*60 UT sequence from *Alternaria alternata* (GenBank: GQ871196). The *cpn*60 UT sequence of fungal isolate 15 was distinct from the other isolates and identical to OTU03024. Examination of the ITS sequences of these isolates suggested that isolates 5, 6, and 9 were most closely related to *A. alternata* while isolate 15 clustered with *Alternaria infectoria* and *Alternaria triticina* (Figure 4-7). These observations were consistent with the morphological features of the fungal conidia, which were also typical of *Alternaria* spp. (data not shown).

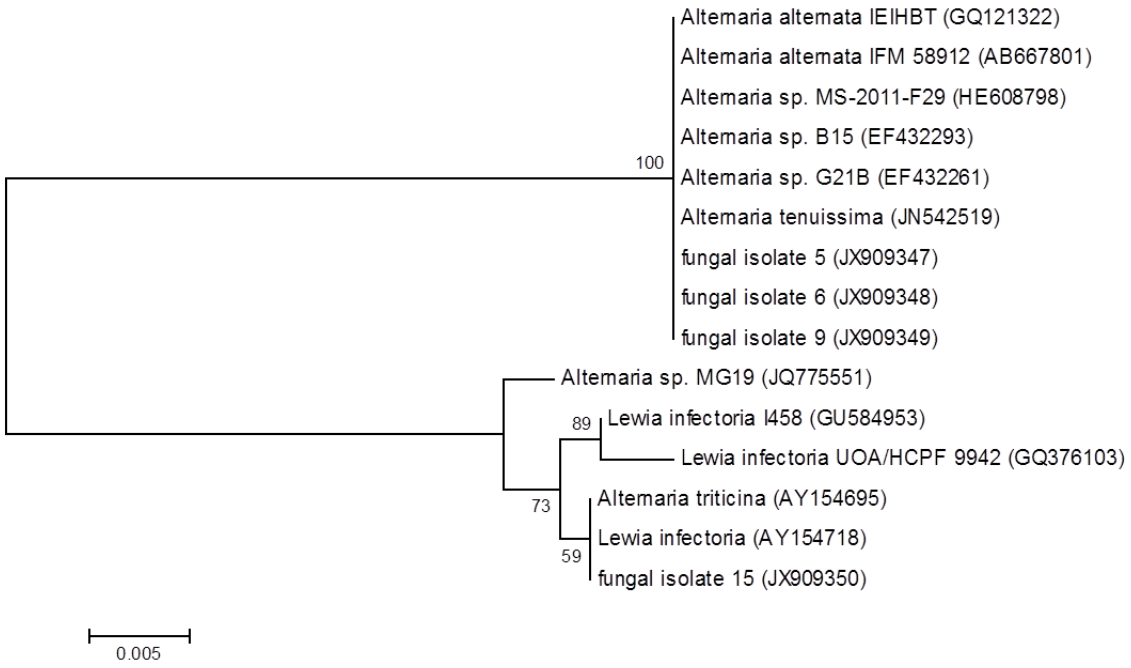


Figure 4-7 Phylogenetic analysis of ITS sequences of fungal isolates compared to reference sequences. Sequences were trimmed to the same length (570-600 bp) prior to alignment. The tree was constructed as described in Materials and Methods, with the percentage of 500 replicates showing identical branching/clustering patterns shown next to the nodes. Reference sequences were obtained from GenBank, with the accession numbers indicated.

Interactions between bacterial and fungal isolates.

The 9 *P. agglomerans* isolates from *Triticum* seeds showed a spectrum of growth suppression against fungal isolate 15 (*Alternaria* sp.; identical to OTU03024) (Figure 4-8) as well as the canola blackleg pathogen *L. maculans* (Figure 4-9). *P. agglomerans* isolate 4 (identical to OTU00845) showed the strongest inhibition while other strains (isolates 3, 6 and 8) as well as unsterilized and sterilized seeds showed no inhibition on both wheat and canola (Table 4-4; Figure 4-8; Figure 4-9). Some of the strains resulted in growth cessation of *L. maculans* at the point of contact, but fungal growth continued away from the bacterial colony (Table 4-4 and Figure 4-9). In general the inhibition of *L. maculans* growth was stronger than of fungal isolate 15 (*Alternaria* sp.) by several of the isolates, but isolate 4 (identical to OTU00845) was quite effective against both fungi (Table 4-4; Figure 4-8; Figure 4-9). Fungal isolate 15 produced a dark pigment upon interaction with bacterial isolate 4 and limited growth continued only in the direction opposite the bacterial challenge (Figure 4-8). Colonization of wheat and canola seeds with bacterial isolate 4 protected both seed types from overgrowth of both the *Alternaria*-like strain (Figure 4-8) and *L. maculans* (Figure 4-9) in these assays.

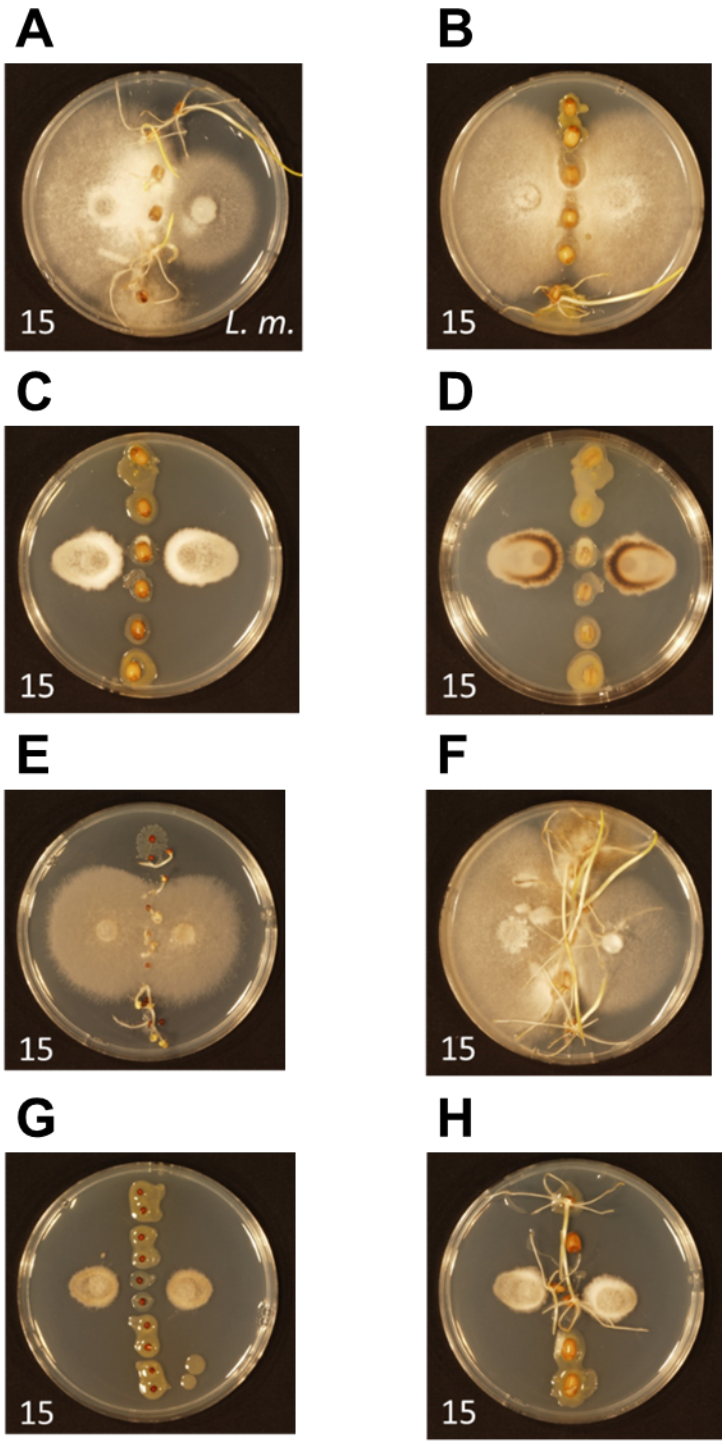

A
15
L. m.
B
15
C
15
D
15
E
15
F
15
G
15
H
15

Figure 4-8 Interactions between selected bacterial isolates and fungal isolate 15. In some instances (A,B,E,F,H), the seeds have begun to germinate, producing shoots on the plates. A. Sterilized wheat seeds not re-colonized with bacteria. Fungal isolate 15 is inoculated on the left while *L. maculans* (*L.m.*) is on the right. In all of the remaining panels, fungal isolate 15 is inoculated on the left and right sides of the seeds. **B.** Sterilized wheat seeds colonized with bacterial isolate 1 (homogenous). **C.** Same as B, but with bacterial isolate 4 (aversion) – top view. **D.** Same as C – bottom view. **E.** Nonsterilized canola seeds. **F.** Nonsterilized wheat seeds. **G.** Nonsterilized canola seeds colonized with bacterial isolate 4. **H.** Nonsterilized wheat seeds colonized with bacterial isolate 4.

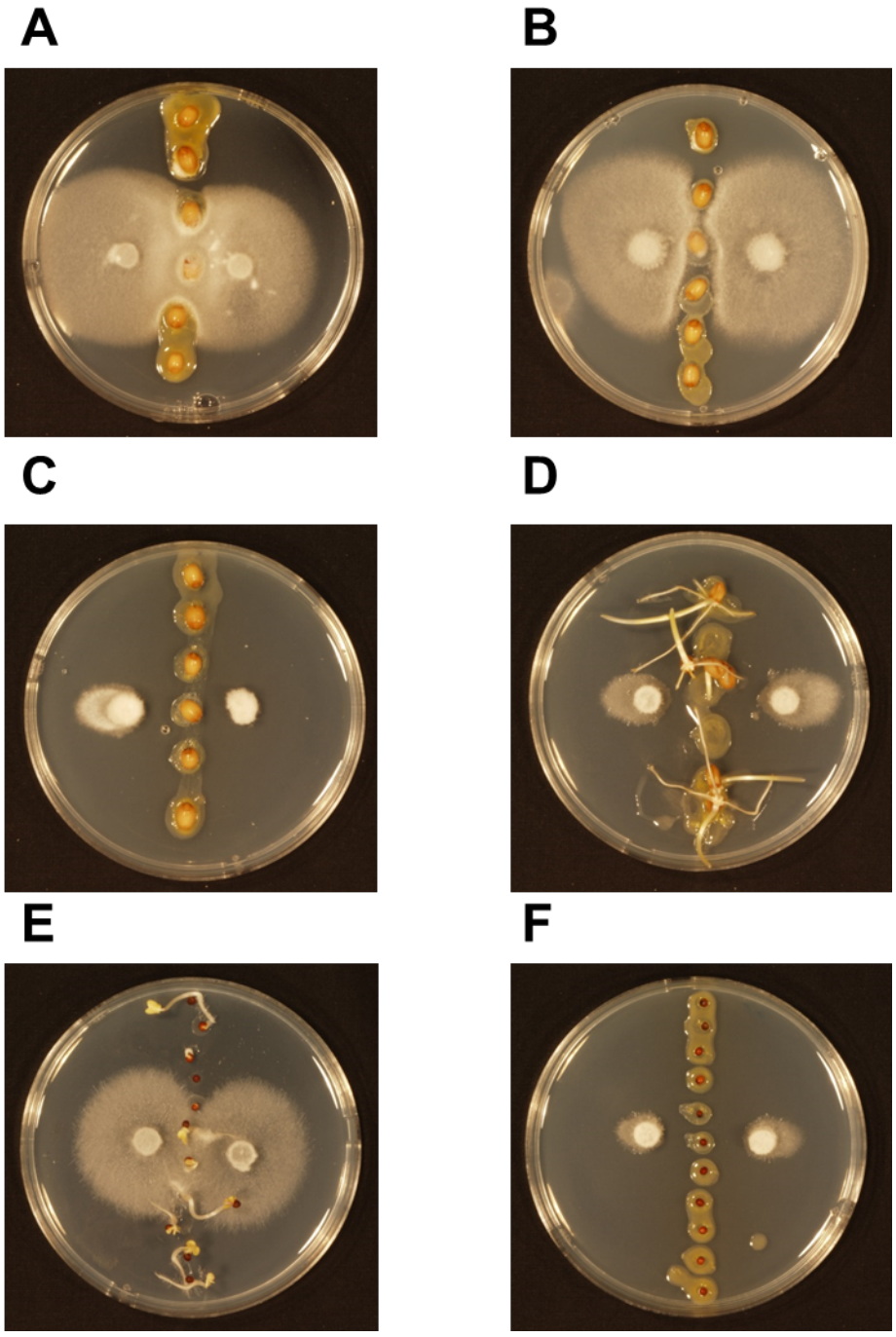


Figure 4-9 Interactions of selected bacterial isolates with the fungal pathogen *Leptosphaeria maculans*. A. Sterilized wheat seeds colonized with bacterial isolate #8 (homogenous). **B.** Same as A., but with bacterial isolate #9 (growth cessation at point of contact). **C.** Same as A, but with bacterial isolate #4 (aversion). **D.** Nonsterilized wheat seeds colonized with bacterial isolate #4. **E.** Nonsterilized canola seeds. **F.** Nonsterilized canola seeds colonized with bacterial isolate #4.

Table 4-4 Interactions between bacterial isolates and *Leptosphaeria maculans* or fungal isolate 15. Colony diameter measurements are the mean of 4 or 6 measurements ± standard deviation. Interactions were scored according to (Chakraborty *et al.* 1994).

Bacterial isolate	Colony diameter increase/day, mm		Interaction	
	L. maculans	Fungal isolate 15	*L. maculans*	Fungal isolate 15
Isolate 1	4.87 ± 0.47	6.84 ± 1.85	growth cessation	homogenous
Isolate 2	4.31 ± 0.12	6.63 ± 1.78	growth cessation	homogenous
Isolate 3	3.83 ± 0.35	6.65 ± 2.06	homogenous	homogenous
Isolate 4	0.24 ± 0.38	1.20 ± 0.36	aversion	aversion
Isolate 5	4.69 ± 0.79	6.74 ± 1.88	growth cessation	homogenous
Isolate 6	4.64 ± 0.43	6.63 ± 2.18	homogenous	homogenous
Isolate 7	4.76 ± 0.70	5.63 ± 1.21	growth cessation	homogenous
Isolate 8	4.58 ± 0.50	7.28 ± 1.92	homogenous	homogenous
Isolate 9	4.32 ± 0.49	6.84 ± 1.41	growth cessation	homogenous
Sterile seed	4.90 ± 0.87	6.81 ± 1.95	homogenous	homogenous
Nonsterile seed	5.08 ± 0.48	7.42 ± 2.13	homogenous	homogenous

Discussion

The microbial complement that is naturally associated with multicellular organisms plays an important role in host health and disease. The microbiota of humans (Turnbaugh *et al.* 2007) and agricultural animals (Dumonceaux *et al.* 2006b; Hill *et al.* 2005b) has been and continues to be extensively studied. While excellent literature exists on the epiphytic and endophytic microbiota associated with a variety of plant surfaces (Lucero *et al.* 2011; Rastogi *et al.* 2012; Tikhonovich and Provorov 2011), there has been little to no characterization of the epiphytic microbiota associated with plant reproductive tissues, including seeds.

In this work, our objective was to describe similarities and differences in the microbiota associated with *Brassica* and *Triticum* seed surfaces. Microbes isolated from the surface of these seeds served to validate the consensus sequences formed for OTU. In thirteen separate cases, including both bacteria and fungi, the sequences assembled were identical (n=11) or essentially identical (99%; n=2) to those obtained from isolates. This demonstrates that OTU assembly yields biologically relevant sequence barcodes that can be used for specific molecular diagnostic assays to detect and quantify microorganisms using established techniques (Dumonceaux *et al.* 2006a; Dumonceaux *et al.* 2009). This is a particularly significant advantage in cases where an OTU sequence is assembled with little similarity to available reference sequences. The simultaneous identification of both prokaryotes and eukaryotes is an advantage of microbial profiling using *cpn*60*,* as opposed to gene targets that are limited to one domain, such as the 16S rRNA gene for Bacteria, or the 18S rRNA and ITS regions commonly used for fungi and other

eukaryotic microbes. As seeds are known to be colonized by both bacteria and fungi, *cpn*60 offers a natural choice for characterizing these microbiomes.

Examination of the microbial communities associated with seeds of these diverse plant species revealed a total epiphytic microbial load of approximately 10^6-10^8 bacterial genomes/g seeds. While this is within the range of what is observed by total aerobic plate counts on other crops such as bean and pea sprouts (Deb and Joshi 2007), there is no baseline data on total epiphytic microbial load of healthy *Triticum* and *Brassica* crop seeds. An endophytic bacterial load in this same range has been reported for *B. napus* seeds (Granér *et al.* 2003). For some related crops, such as buckwheat, customers may set limits on total aerobic plate counts that are considerably lower (5.5 $\log_{10}$ CFU/g seeds) than we observed for *Triticum* and *Brassica* (Dhillon *et al.* 2012), and lower total microbial loads are generally seen as desirable (Olaimat and Holley 2012). The molecular methods used to estimate bacterial genomes g^{-1} seeds are unable to distinguish between live and dead microbes, so estimates of total bacterial load by aerobic plate counts may be considerably lower than is determined using molecular methods. Nevertheless, our results establish a baseline epiphytic microbial load for healthy, high grade seeds of *Triticum* and *Brassica*.

An overall total of 5,477 OTU was associated with all *Brassica* and *Triticum* samples. Core microbiota were identified for all *Brassica* samples (215 OTU) as well as all *Triticum* samples (262 OTU), but remarkably, we also identified a shared microbiome among these seeds from distinct host plant genera harvested from a range of geographic locales, separated by thousands of kilometers. The existence of a shared microbiome conserved across plant genera illustrates that the seed-associated microbiome is not a

casually associated surface contamination but rather a selected, host-specific community, intimately associated with the host, and with potentially profound effects on seed health. These observations are consistent with previous studies of the seed-associated endophytic bacteria within *Zea* spp. (corn), wherein a microbiota was identified that is conserved in various teosinte progenitor species grown in an array of geographical locations (Johnston-Monje and Raizada 2011). Despite these commonalities, the *Triticum* and *Brassica* seed microbiota could be distinguished based on the relative abundances of shared OTU (Figure 4-4).

Studies of *Zea* seed endophytes revealed a preponderance of Gammaproteobacteria including *Enterobacter*, *Pantoea*, and *Pseudomonas* spp. (Johnston-Monje and Raizada 2011). Similarly, Weiss *et al*. examined the microbiota associated with alfalfa, radish, and bean sprouts and found the same genera represented, along with *Lactobacillus* (Weiss *et al.* 2007). The majority of bacterial taxa that we observed in the *Brassica-Triticum* shared microbiome included OTU that were closely related to these genera (Table 4-2;Table 4-3). Many of the microorganisms we identified on the seed surface are also found in soil, suggesting a possible relationship between soil microbiota and seed-borne microorganisms. This is consistent with the fact that *Triticum* and *Brassica* seeds are sown into soils, commonly in rotation with one another. The seed microbiome included a relatively large proportion of OTU that were closely related (95-99% sequence identity) to *P. agglomerans*, including 78 that were significantly differentially abundant on *Triticum* compared to *Brassica* seeds. Among the fungal OTU were several with similarity to yeasts and Ascomycetes, including *Fusarium*. While certain species of *Fusarium* are wheat pathogens, no sequences identical to known pathogens were detected

on these seeds; however, given the ability for the seeds to be associated with microbes closely related to pathogens there is a clear need to monitor seed health. In addition, 18 OTU were identified that clustered with microorganisms such as *Pyrenophora*, *Alternaria*, and *Leptosphaeria,* of which there are related pathogenic species that can cause grain spoilage. While all seeds in this study were healthy, these findings demonstrated that the seed microbiome is crucial as it may harbor both beneficial and potentially pathogenic organisms.

Our data also indicate that observations of OTU abundance patterns can lead to the recognition of interactions between microbes with significant implications for the host. Relatively high levels of *Pantoea-like* OTU and significantly lower levels of *Alternaria*-like OTU were detected on *Triticum* seeds, while this relationship was reversed on *Brassica* seeds (Figure 4-4, Figure 4-5). The reciprocal abundances of *P. agglomerans* and *Alternaria* sequences on *Triticum* and *Brassica* seeds, validated by quantitative PCR, suggested a potential antagonistic relationship between these microbes. It is well known that *P. agglomerans* can be antagonistic to *L. maculans* and other pathogens (Braun-Kiewnick *et al.* 2000; Bryk *et al.* 1998; Chakraborty *et al.* 1994; Kearns and Hale 1996; Kempf and Wolf 1989), but inhibition of the growth of *Alternaria* spp. by *P. agglomerans* has not been described. The fact that we identified this organism within the epiphytic microbiota of healthy *Triticum* and *Brassica* seeds suggests that organisms with pathogen-protective effects naturally associate with seeds. In contrast, *Alternaria* spp., distinct from those detected on the healthy seeds within this study, can cause grain safety concerns in storage due to the production of mycotoxins by specific species (Greco *et al.* 2012). These observations suggest that the *P. agglomerans* strain we identified in this

study has potential as a biocontrol agent, and if applied to seeds may act to protect them from storage-associated spoilage or colonization with pathogenic microorganisms.

We have identified a remarkably conserved epiphytic microbiome on the seeds of geographically and ecologically diverse samples of two important crops. Reproducible differences in the abundances of constituents of this microbiota were used to identify patterns associated with each crop type. Furthermore, this work has shown that differences in OTU abundance within and between microbiomes can be valuable clues and indicators of biological interactions among microorganisms. Finally, we demonstrated a method for simultaneous profiling of the prokaryotes and eukaryotes within the epiphytic microbiota of crop seeds. These results provide a system for understanding the microorganisms associated with crop seeds, and highlight the need for a thorough understanding of these microbial communities and their importance to production and storage of healthy, high quality seeds.

Acknowledgements

We thank Russell Hynes for help with the biocontrol assays, and Teenus Paramel Jayaprakash for assistance with photography of the bioassay results. This work was funded by the genomics program of Agriculture and Agri-Food Canada through the Genomics Research and Development Initiative (GRDI).

CHAPTER 5 - Conclusions and discussion

Summary and limitations of these works

*cpn*60 is the preferred barcode for bacteria

The formal framework developed for DNA barcoding of eukaryotes can be used to evaluate gene targets for microbial profiling (Links *et al.* 2012), and based on this framework, both 16S rRNA and *cpn*60 meet the criteria to be effective barcodes for Bacteria. On average, the distance between *cpn*60 sequences of different species of bacteria is larger than for 16S rRNA, suggesting *cpn*60 is a more robust choice for numerical taxonomy, and thus is the preferred barcode. The results presented in this thesis, based on all complete bacterial genomes in the public domain, also demonstrated that *cpn*60 has an additional advantage over 16S rRNA. *cpn*60 has on average 1 copy per genome as opposed to 16S RNA having 3, which has direct relevance to the inference of organismal abundance from DNA sequencing. While outside the scope of this work, it would be interesting to see formal proposals for both 16S rRNA and *cpn*60 to the Barcode of Life project. The proposal of DNA barcodes for Bacteria would expand the Barcode of Life to a second domain and it would establish within microbiology a rigorous method for evaluating barcodes.

It was also shown that the distance between closely related *cpn*60 sequences could be exploited through methods of sequence assembly to form OTUs. Using sequence assembly to form OTUs requires the optimization of assembly parameters demonstrated initially in Chapter 2 and to a larger extent in Chapter 3. The use of *de novo* assembly for OTU formation ensures that the DNA sequence for each OTU is the most representative

one supported by the data. Clustering of sequence data is currently the most common approach to OTU formation. If all sequences were full length and clustering was performed at 100% identity then OTU clustering and assembly would produce equivalent results. However, it is common that experimental data will not be full length and that there will be errors within the data. Thus clustering approaches will tend to be implemented at less than 100% identity and this will result in problems when choosing a representative sequence for the OTU. In contrast, sequence assembly methods produce a consensus sequence that is by its nature the longest and most representative sequence possible to derive from the data. When an assembly method is used for OTU formation it produces a reliable consensus sequence that is suitable as a biomarker for downstream studies (e.g. isolation of an unknown organism). As a whole, the analyses presented here serve to formally establish suitable DNA barcodes for Bacteria, propose a method for OTU formation through sequence assembly, and provide a mechanism through which OTU assembly can be evaluated and optimized.

While every effort was made to have the most taxonomic breadth possible when assessing DNA barcodes for Bacteria, there are inherent limitations in this work. Working with all complete genomes in the public domain did enable large numbers of comparisons for inter and intra specific distance calculations. However, the analysis may be biased by factors responsible for the organisms being chosen for sequencing and deposition in the public domain in the first place. Most of the data within public databases is highly biased to a limited number of phyla, which is a result of studies being primarily focused on pathogens and microbes that relate to human health. Therefore it is important to recognize that the observations on these data may have some limitations

when applied to novel or poorly described taxa. As future efforts continue to expand the content of public databases these biases will diminish.

A related but much more minor limitation with the use of resources such as GenBank is that it is inherently difficult to extract barcode regions when the region is not yet an accepted barcode. One of the side effects of the acceptance of a DNA Barcode is that sequence repositories will attempt to identify barcodes from complete genome records systematically, and annotate them accordingly. However when evaluating a potential barcode there is a somewhat self-referential problem in that they cannot be identified until they are defined. This limitation manifests itself in that there may be a proportion of the test data arising from mis-annotation of the whole genomes when deposited into the public domain. These mis-annotations could be cases where the annotation (*cpn*60/*gro*EL/*hsp*60) has been incorrectly ascribed to a gene. Conversely there may be annotations that are missed entirely. Both of these mis-annotations will affect systematic studies by either introducing false barcodes into the experimental dataset or missing examples and will affect the results.

Unsupervised OTU formation is possible with mPUMA

mPUMA is an unsupervised pipeline for the assembly of OTU from microbial profiling data. Using a synthetic community it was shown that mPUMA can reliably assemble the OTU present and estimate their abundance in the absence of any information other than DNA sequencing data itself. This enables the discovery of novel, unknown OTU and the tracking of their abundance across experiments. Using the procedures established in

Chapter 2 it was possible to evaluate the effects of changing both the method of assembly and read tracking on the performance of mPUMA (Chapter 3).

The use of mPUMA was tested extensively and optimized in terms of its performance when applied to profiling a synthetic community of *cpn*60 sequences. There have been tests performed to confirm that mPUMA can process data from studies using 16S rRNA or *rpo*B but each non-*cpn*60 barcode warrants a direct investigation into the optimal procedures within mPUMA. Given that both of the assembly methods implemented within mPUMA (gsAssembler and Trinity) support RNA transcript assemblies, there is an opportunity to use multiple DNA barcodes simultaneously (e.g. *cpn*60 and 16S rRNA, or *cpn*60 and the type-II archaeal chaperonin). Obviously, any such application to a natural microbial community should be preceded by performance optimization studies using mPUMA to determine suitable parameters.

Microbial profiles derived through mPUMA can generate testable hypotheses

The application of mPUMA to the epiphytic microbiota of plant seeds demonstrated that multiple domains of life could be profiled simultaneously (Eukaryotes and Bacteria). Tools such as 16S rRNA commonly used for Bacteria and the ITS region used in fungal studies could be used together but would need to be performed independently. Microbial profiles derived through the use of mPUMA were used to identify core microbiomes within a genus and also shared between genera of different hosts. The existence of a core microbiome common within a genus (both for *Triticum* and *Brassica*) is particularly important as it suggests that the microbes, which associate with seeds of these plants, are conserved, and thus predictable. This observation suggests that there is something unique

about the interactions between the host and its seed microbiome. The additional finding that there exists a microbiome shared across these genera is also important. Crops of these genera are commonly grown in rotation with one another within a common field. If there are microbes that can survive on the seeds of both crops, then there may be some form of cycling between microbes on seeds, the soil they are seeded into, and the seeds produced in subsequent generations or crops. This further suggests that the microbes present on seeds could affect subsequent crops in the rotation. These crop-to-crop interactions should therefore be investigated in terms of the ability of the epiphytic seed microbiome to function as a vector through which new pathogens or probiotics could be introduced to a field.

By observing differential patterns in OTU abundance amongst the inter-genera core microbiome it was hypothesized that these OTU may interact. Following up on this hypothesis it was shown that these OTU are functionally relevant to one another with *Pantoea-like* bacterial isolates exhibiting fungistatic properties to *Alternaria*-like fungal isolates. This provides a concrete example of how an understanding of the taxonomic composition of a microbial community can guide functional studies.

While the microbial profiling was by no means the end point, it was a crucial component used to identify possible interactions and guide culture conditions in order to isolate examples of the relevant OTU. For the *Pantoea-like* and *Alternaria*-like isolates there were discrete culture conditions or morphologies known for these organisms that assisted in their isolation. When one considers application to novel or unknown OTU, the use of methods such as mPUMA are important. For a novel OTU there would not necessarily be a single prescriptive culture condition to try. However, when OTU are formed through

assembly, as is the case for mPUMA, there is the creation of a consensus sequence. This consensus sequence is a discrete biomarker for a specific OTU. If one were to try and culture out a completely novel OTU the consensus sequence could be a crucial tool used for PCR or screening assays while determining optimal culture conditions to isolate that organism.

The study of the epiphytic seed microbiome has inherent limitations given the sample size (6 *Triticum* and 5 *Brassica* samples). Additional follow up experiments could investigate large collections of seed from the Plant Gene Resources of Canada, which contain 11435 accessions of *Triticum* and 2096 accessions from *Brassica* (as of July 19th, 2013). Larger studies will be essential to understanding the significance of both the intra-genus core and the inter-genera shared microbiomes that are reported in this thesis.

Discussion of future prospects

Third generation DNA sequencing will not presently disrupt metagenomics

The advent of 3rd generation sequencing brings to light methods which carry out single molecule sequencing (SMS) (Schadt *et al.* 2010). The key advantage of SMS is that it has no reliance on PCR to amplify the original DNA sample, thus overcoming one of the major technical challenges with current methods. PCR amplification from a complex template necessarily results in a distortion of the proportional abundances of sequences in the original sample. In its most extreme manifestation, this results in some taxa being completely absent from the sequence library despite their prevalence in the community (Hill *et al.* 2010). Given that organismal abundance is a critical parameter in microbial ecology, there is an obvious desire to use the most robust estimates possible. Third

generation sequencing results in longer read lengths (1-10kb) but has higher per read error (>5%) than second generation methods (1%)(Gilles *et al.* 2011). Additional sequencing approaches such as strobing or circular consensus provide intriguingly different data types from 3rd generation technologies .

Strobing is a method where data capture is turned on and off according to some pattern while the polymerase is functioning. When the processivity of the polymerase is known, an estimate can be placed on the distance between each data acquisition cycle. In the simplest form, an on-off-on strobe pattern would result in reads that are paired ends. The ability of the PacBio system to change the strobing patterns actually means that any pattern of on and off could be combined and used to capture data. The flexibility of strobing patterns is interesting, but will required new bioinformatics methods to handle these datasets.

Circular consensus is a sequencing approach where multiple reads across a region are generated sequentially. In the sample preparation for the PacBio systems, fragments of DNA are double stranded and adapted to form a barbell structure with a loop at either end. During circular consensus sequencing the polymerase will use each of the barbell loops to complete a cycle through the DNA template and its complementary strand. Data generated from circular consensus would thus be in the form of [Adapter 1, DNA template, Adapter 2, Reverse complement of DNA template, Adapter 1, DNA template, Adapter 2, Reverse complement of DNA template, …]. The bioinformatic challenges of processing circular consensus data are less daunting than for strobing since the data would be trimmed for the two barbell adapters and then aligned.

While it might be tempting to say that 3rd generation sequencing will usher in a revolution for metagenomics, I am not currently convinced.

The longer read lengths (1-10kb) from 3rd generation approaches will likely make tractable the resolution of larger repeat elements. However the associated reduction in accuracy makes the interpretation of such data questionable. The ability to acquire strobed reads which could bridge large gaps is also interesting, but the success of this technique depends on high molecular weight DNA and I do not think it will be feasible to extract DNA from a metagenomic sample while preserving its integrity to a high enough level. Lastly, the throughput of 3rd generation technologies is not currently at the scale where it would be disruptive to the metagenomics field. Currently the chief example of 3rd generation sequencing is the Pacific Biosciences RS. The RS sequences in a chamber called a zero-mode waveguide (ZMW). Each single molecule, real-time (SMRT) cell of the RS is comprised of 75k ZMWs and there are 8 SMRT cells packaged together. So if it were possible to generate 1 kb reads on a Pacific Biosciences RS the machine could yield (1 kb/ZMW × 75,000 ZMWs/cell) × 8 SMRT cells = 600 Mbp of data. Assuming a 5% error rate (and there would be no way to tell which calls were inaccurate) that would result in a theoretical maximum yield of 570 Mbp of high quality data, (or the equivalent of a 1× coverage through a community comprised of 100 organisms each with a 5 Mb genome and existing in equimolar concentrations). Therefore in my opinion the throughput of the current technologies and their corresponding error profiles suggest third generation sequencing is not yet suitable for metagenomic studies.

A *run-until* sequencing paradigm may disrupt microbial profiling

Oxford nanopore has introduced a *run-until* paradigm for bioinformatic analyses linked in real-time sequencing on their nanopore platforms. The key facet to *run-until* is that the user of the DNA sequencer can develop a rule to determine when enough sequence data has been acquired. The simplest example is one of genome re-sequencing for SNP detection. A bioinformatics workflow can be used in real time to assess each sequence data in terms of a critical read depth at a specific location within a genome. When enough data has been acquired to robustly call the sample allele at that SNP, the bioinformatics procedure can signal that sequencing should cease. While these sequencers are not generally available at the time of writing, the inherent concept of a *run-until* paradigm is particularly interesting for microbial profiling. As described by Gihring *et al.*, there is a potential for ecological parameters to be affected by un-equal sampling (Gihring *et al.* 2012). The use of a *run-until* bioinformatics pipeline could enable the use of a constraint based on a target number of sequencing reads or a rarefaction of the distinct OTUs found in the sequencing run in order to produce equalized sampling efforts directly.

Comparisons to presumed *Gold Standards* are problematic

There is a common suggestion that novel bioinformatics methods for metagenomics should be subjected to comparison with some *Gold Standard*. Conceptually this is arising from the need to establish new methods in the context of prior art. Where this becomes problematic is when one considers the growing belief in the microbial bioinformatics community that methods have reached a state where there is in fact *a Gold Standard*, which is to target the 16S rRNA gene and form OTUs through clustering.

Commonly there is a perception that when a large or highly impactful study is published that it establishes a standard. Publication is by no means the establishment of a standard. If there were ever a field of science that should recognize that, it would be microbiology with its rich history in the area of microbial taxonomy. Standards take the engagement of knowledgeable and interested parties (e.g. Barcode of Life project), and usually are the result of discussion and agreement. Certainly the primary scientific literature should form the basis for a standard but the establishment of a standard is a much larger collective aim. There have been a number of large projects initiated (e.g. Human Microbiome Project) which have needed to adopt standards internally, but there has not been a discipline wide acceptance of a single method or procedure.

Previously published studies may serve to provide insight into microbial communities but it is important to recognize that the choices made in each study from experimental design through data generation and interpretation, provide a lens through which the community is observed. If one were to exchange the use of *cpn*60 for 16S rRNA and analyze a previously studied community it is important to recognize that neither the 16S rRNA nor the *cpn*60-based study would generate *the correct* observation. Each study will have some limitations and so it is important to recognize that comparisons cannot be phrased in terms of one study being correct and the second needing to duplicate the findings of the first.

Only a fraction of the world's bacterial diversity has been characterized (Staley and Konopka 1985). Of those isolates that have been characterized there is a clear bias to a few phyla (Hugenholtz 2002). If the volume of bacterial life on Earth is on the same scale as plant life (Whitman *et al.* 1998) and only an exceptionally small fraction of the

diversity has been characterized then there is a clear need for novel approaches to explore microbial diversity. Continued efforts to apply a variety of approaches, challenge establish methods and evaluate new ones such as those presented in this thesis will be crucial to enhance our knowledge of microbial diversity on Earth.

REFERENCES

Consortium for the Barcode of Life. Non-COI Barcode Regions — Guidelines for CBOL Approval. Available from http://barcoding.si.edu/PDF/Guidelines for non-CO1 selection FINAL.pdf accessed 2013-07-17.

Lott, S.T. Revealing SMRT Biology. Available from http://www.gqinnovationcentre.com/documents/rendezVous/2-Steven_T_Lott.pdf accessed 2013-07-25.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research **25**(17): 3389-3402.

Baumgartner, L.K., Dupraz, C., Buckley, D.H., Spear, J.R., Pace, N.R., and Visscher, P.T. 2009. Microbial species richness and metabolic activities in hypersaline microbial mats: insight into biosignature formation through lithification. Astrobiology **9**(9): 861-874.

Blaiotta, G., Fusco, V., Ercolini, D., Aponte, M., Pepe, O., and Villani, F. 2008. *Lactobacillus* strain diversity based on partial hsp60 gene sequences and design of PCR-restriction fragment length polymorphism assays for species identification and differentiation. Applied and Environmental Microbiology **74**(1): 208-215.

Boone, D.R., Castenholz, R.W., and Garity, G.M. 2001. Bergey's Manual of Systematic Bacteriology. Bergey's Manual Trust, New York.

Braun-Kiewnick, A., Jacobsen, B.J., and Sands, D.C. 2000. Biological control of *Pseudomonas syringae* pv. *syringae*, the causal agent of basal kernel blight of barley, by antagonistic *Pantoea agglomerans*. Phytopathology **90**(4): 368-375.

Brosius, J., Palmer, M.L., Kennedy, P.J., and Noller, H.F. 1978. Complete nucleotide sequence of a 16S ribosomal RNA gene from *Escherichia coli*. Proceedings of the National Academy of Sciences of the United States of America **75**(10): 4801-4805.

Brousseau, R., Hill, J.E., Prefontaine, G., Goh, S.H., Harel, J., and Hemmingsen, S.M. 2001. *Streptococcus suis* serotypes characterized by analysis of chaperonin 60 gene sequences. Applied and Environmental Microbiology **67**(10): 4828-4833.

Bryk, H., Dyki, B., and Sobiczewski, P. 1998. Antagonistic effect of *Pantoea agglomerans* on *in vitro* spore germination and germ tube elongation of *Botrytis cinerea* and *Penicillium expansum*. BioControl **43**(1): 97-106.

Bushong, J.A., Griffith, A.P., Peeper, T.F., and Epplin, F.M. 2012. Continuous winter wheat versus a winter canola - winter wheat rotation. Agronomy Journal **104**(2): 324-330.

Cannone, J.J., Subramanian, S., Schnare, M.N., Collett, J.R., D'Souza, L.M., Du, Y., Feng, B., Lin, N., Madabusi, L.V., Muller, K.M., Pande, N., Shang, Z., Yu, N., and Gutell, R.R. 2002. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics **3**: 2.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods **7**(5): 335-336.

Case, R.J., Boucher, Y., Dahllof, I., Holmstrom, C., Doolittle, W.F., and Kjelleberg, S. 2007. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Applied and Environmental Microbiology **73**(1): 278-288.

Chaban, B., and Hill, J.E. 2012. A 'universal' type II chaperonin PCR detection system for the investigation of Archaea in complex microbial communities. ISME Journal **6**(2): 430-439.

Chaban, B., Links, M.G., and Hill, J.E. 2012. A molecular enrichment strategy based on cpn60 for detection of epsilon-proteobacteria in the dog fecal microbiome. Microbial Ecology **63**(2): 348-357.

Chaban, B., Musil, K.M., Himsworth, C.G., and Hill, J.E. 2009. Development of cpn60-based real-time quantitative PCR assays for the detection of 14 *Campylobacter* species and application to screening of canine fecal samples. Applied and Environmental Microbiology **75**(10): 3055-3061.

Chaban, B., Ngeleka, M., and Hill, J.E. 2010. Detection and quantification of 14 *Campylobacter* species in pet dogs reveals an increase in species richness in feces of diarrheic animals. BMC Microbiology **10**: 73.

Chakraborty, B.N., Chakraborty, U., and Basu, K. 1994. Antagonism of *Pantoea agglomerans* towards *Leptosphaeria maculans* causing blackleg disease of *Brassica napus*. Letters in Applied Microbiology **18**(2): 74-76.

Chakravorty, S., Helb, D., Burday, M., Connell, N., and Alland, D. 2007. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J Microbiol Methods **69**(2): 330-339.

Coates, A.R., Shinnick, T.M., and Ellis, R.J. 1993. Chaperonin nomenclature. Molecular Microbiology **8**(4): 787.

Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam-Syed-Mohideen, A.S., McGarrell, D.M., Bandela, A.M., Cardenas, E., Garrity, G.M., and Tiedje, J.M. 2007. The ribosomal

database project (RDP-II): introducing myRDP space and quality controlled public data. Nucleic Acids Research **35**(Database issue): D169-172.

Critzer, F.J., and Doyle, M.P. 2010. Microbial ecology of foodborne pathogens associated with produce. Current Opinion in Biotechnology **21**(2): 125-130.

Dahllof, I., Baillie, H., and Kjelleberg, S. 2000. rpoB-based microbial community analysis avoids limitations inherent in 16S rRNA gene intraspecies heterogeneity. Applied and Environmental Microbiology **66**(8): 3376-3380.

Deb, M.P., and Joshi, P.A. 2007. Microbiological analysis of sprouts and effect of spices on microbial load. Journal of Food Science and Technology **44**(5): 545-547.

Desai, A.R., Links, M.G., Collins, S.A., Mansfield, G.S., Drew, M.D., Van Kessel, A.G., and Hill, J.E. 2012. Effects of plant-based diets on the distal gut microbiome of rainbow trout (*Oncorhynchus mykiss*). Aquaculture **350-353**: 134-142.

Desai, A.R., Musil, K.M., Carr, A.P., and Hill, J.E. 2009. Characterization and quantification of feline fecal microbiota using cpn60 sequence-based methods and investigation of animal-to-animal variation in microbial population structure. Veterinary Microbiology **137**(1-2): 120-128.

DeSantis, T.Z., Jr., Hugenholtz, P., Keller, K., Brodie, E.L., Larsen, N., Piceno, Y.M., Phan, R., and Andersen, G.L. 2006. NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acids Res **34**(Web Server issue): W394-399.

Devereux, R., He, S.H., Doyle, C.L., Orkland, S., Stahl, D.A., LeGall, J., and Whitman, W.B. 1990. Diversity and origin of *Desulfovibrio* species: phylogenetic definition of a family. Journal of Bacteriology **172**(7): 3609-3619.

Dhillon, B., Wiesenborn, D., Sidhu, H., and Wolf-Hall, C. 2012. Improved microbial quality of buckwheat using antimicrobial solutions in a fluidized bed. Journal of Food Science **77**(4): E98-E103.

Duarte, S.C., Pena, A., and Lino, C.M. 2010. A review on ochratoxin A occurrence and effects of processing of cereal and cereal derived food products. Food Microbiology **27**(2): 187-198.

Dumonceaux, T.J., Hill, J.E., Briggs, S.A., Amoako, K.K., Hemmingsen, S.M., and Van Kessel, A.G. 2006a. Enumeration of specific bacterial populations in complex intestinal communities using quantitative PCR based on the chaperonin-60 target. Journal of Microbiol Methods **64**(1): 46-62.

Dumonceaux, T.J., Hill, J.E., Hemmingsen, S.M., and Van Kessel, A.G. 2006b. Characterization of intestinal microbiota and response to dietary virginiamycin supplementation in the broiler chicken. Applied and Environmental Microbiology **72**(4): 2815-2823.

Dumonceaux, T.J., Hill, J.E., Pelletier, C., Paice, M.G., Van Kessel, A.G., and Hemmingsen, S.M. 2006c. Molecular characterization of microbial communities in Canadian pulp and paper activated sludge and quantification of a novel *Thiothrix eikelboomii*-like bulking filament. Canadian Journal of Microbiology **52**(5): 494-500.

Dumonceaux, T.J., Schellenberg, J., Goleski, V., Hill, J.E., Jaoko, W., Kimani, J., Money, D., Ball, T.B., Plummer, F.A., and Severini, A. 2009. Multiplex detection of bacteria associated with normal microbiota and with bacterial vaginosis in vaginal swabs by use of oligonucleotide-coupled fluorescent microspheres. Journal of Clinical Microbiology **47**(12): 4067-4077.

Dumonceaux, T.J., Town, J.R., Hill, J.E., Chaban, B.L., and Hemmingsen, S.M. 2011. Multiplex detection of bacteria in complex clinical and environmental samples using oligonucleotide-coupled fluorescent microspheres. The Journal of Visualized Experiments (56).

Edgar, R.C. 2004. Local homology recognition and distance measures in linear time using compressed amino acid alphabets. Nucleic Acids Research **32**(1): 380-385.

Edgar, R.C. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics **26**(19): 2460-2461.

Felsenstein, J. 1989. PHYLIP-Phylogeny Inference Package (Version 3.2). Cladistics **5**: 164-166.

Fitch, W.M. 1976. The molecular evolution of cytochrome c in eukaryotes. Journal of Molecular Evolution **8**(1): 13-40.

Fitch, W.M., and Margoliash, E. 1967. Construction of phylogenetic trees. Science **155**(3760): 279-284.

Gagnon, N., Barret, M., Topp, E., Kalmokoff, M., Masse, D., Masse, L., and Talbot, G. 2011. A novel fingerprint method to assess the diversity of methanogens in microbial systems. FEMS Microbiology Letters **325**(2): 115-122.

Gihring, T.M., Green, S.J., and Schadt, C.W. 2012. Massively parallel rRNA gene sequencing exacerbates the potential for biased community diversity comparisons due to variable library sizes. Environmental Microbiology **14**(2): 285-290.

Gilles, A., Meglecz, E., Pech, N., Ferreira, S., Malausa, T., and Martin, J.F. 2011. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics **12**: 245.

Giovannoni, S.J., Britschgi, T.B., Moyer, C.L., and Field, K.G. 1990. Genetic diversity in Sargasso Sea bacterioplankton. Nature **345**(6270): 60-63.

Goh, S.H., Driedger, D., Gillett, S., Low, D.E., Hemmingsen, S.M., Amos, M., Chan, D., Lovgren, M., Willey, B.M., Shaw, C., and Smith, J.A. 1998. *Streptococcus iniae*, a

human and animal pathogen: specific identification by the chaperonin 60 gene identification method. Journal of Clinical Microbiology **36**(7): 2164-2166.

Goh, S.H., Facklam, R.R., Chang, M., Hill, J.E., Tyrrell, G.J., Burns, E.C., Chan, D., He, C., Rahim, T., Shaw, C., and Hemmingsen, S.M. 2000. Identification of *Enterococcus* species and phenotypically similar *Lactococcus* and *Vagococcus* species by reverse checkerboard hybridization to chaperonin 60 gene sequences. Journal of Clinical Microbiology **38**(11): 3953-3959.

Goh, S.H., Potter, S., Wood, J.O., Hemmingsen, S.M., Reynolds, R.P., and Chow, A.W. 1996. HSP60 gene sequences as universal targets for microbial species identification: studies with coagulase-negative staphylococci. Journal of Clinical Microbiology **34**(4): 818-823.

Goh, S.H., Santucci, Z., Kloos, W.E., Faltyn, M., George, C.G., Driedger, D., and Hemmingsen, S.M. 1997. Identification of *Staphylococcus* species and subspecies by the chaperonin 60 gene identification method and reverse checkerboard hybridization. Journal of Clinical Microbiology **35**(12): 3116-3121.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology **29**(7): 644-652.

Granér, G., Persson, P., Meijer, J., and Alström, S. 2003. A study on microbial diversity in different cultivars of Brassica napus in relation to its wilt pathogen, *Verticillium longisporum*. FEMS Microbiology Letters **224**(2): 269-276.

Greco, M., Patriarca, A., TerminielloL, Fernández Pinto, V., and Pose, G. 2012. Toxigenic *Alternaria* species from Argentinean blueberries. International Journal of Food Microbiology **154**(3): 187-191.

Haas, B.J., Gevers, D., Earl, A.M., Feldgarden, M., Ward, D.V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S.K., Sodergren, E., Methe, B., DeSantis, T.Z., Human Microbiome, C., Petrosino, J.F., Knight, R., and Birren, B.W. 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Research **21**(3): 494-504.

Hallmann, J., Quadt-Hallmann, A., Mahaffee, W.F., and Kloepper, J.W. 1997. Bacterial endophytes in agricultural crops. Canadian Journal of Microbiology **43**(10): 895-914.

Hamady, M., and Knight, R. 2009. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. Genome Research **19**(7): 1141-1152.

Hamady, M., Lozupone, C., and Knight, R. 2010. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. ISME Journal **4**(1): 17-27.

Harker, K.N., O'Donovan, J.T., Turkington, T.K., Blackshaw, R.E., Lupwayi, N.Z., Smith, E.G., Klein-Gebbinck, H., Dosdall, L.M., Hall, L.M., Willenborg, C.J., Kutcher, H.R., Malhi, S.S., Vera, C.L., Gan, Y., Lafond, G.P., May, W.E., Grant, C.A., and McLaren, D.L. 2012. High-yield no-till canola production on the Canadian prairies. Canadian Journal of Plant Science **92**(2): 221-233.

Hashidoko, Y. 2005. Ecochemical studies of interrelationships between epiphytic bacteria and host plants via secondary metabolites. Bioscience, Biotechnology and Biochemistry **69**(8): 1427-1441.

Hebert, P.D., Cywinska, A., Ball, S.L., and deWaard, J.R. 2003. Biological identifications through DNA barcodes. Proceedings of the Royal Society B **270**(1512): 313-321.

Hebert, P.D., Stoeckle, M.Y., Zemlak, T.S., and Francis, C.M. 2004. Identification of Birds through DNA Barcodes. PLoS Biology **2**(10): e312.

Hemmingsen, S.M., Woolford, C., van der Vies, S.M., Tilly, K., Dennis, D.T., Georgopoulos, C.P., Hendrix, R.W., and Ellis, R.J. 1988. Homologous plant and bacterial proteins chaperone oligomeric protein assembly. Nature **333**(6171): 330-334.

Hill, J.E., Fernando, W.M., Zello, G.A., Tyler, R.T., Dahl, W.J., and Van Kessel, A.G. 2010. Improvement of the representation of bifidobacteria in fecal microbiota metagenomic libraries by application of the cpn60 universal primer cocktail. Applied and Environmental Microbiology **76**(13): 4550-4552.

Hill, J.E., Goh, S.H., Money, D.M., Doyle, M., Li, A., Crosby, W.L., Links, M., Leung, A., Chan, D., and Hemmingsen, S.M. 2005a. Characterization of vaginal microflora of healthy, nonpregnant women by chaperonin-60 sequence-based methods. American Journal of Obstetrics and Gynecology **193**(3 Pt 1): 682-692.

Hill, J.E., Hemmingsen, S.M., Goldade, B.G., Dumonceaux, T.J., Klassen, J., Zijlstra, R.T., Goh, S.H., and Van Kessel, A.G. 2005b. Comparison of ileum microflora of pigs fed corn-, wheat-, or barley-based diets by chaperonin-60 sequencing and quantitative PCR. Applied and Environmental Microbiology **71**(2): 867-875.

Hill, J.E., Paccagnella, A., Law, K., Melito, P.L., Woodward, D.L., Price, L., Leung, A.H., Ng, L.K., Hemmingsen, S.M., and Goh, S.H. 2006a. Identification of *Campylobacter* spp. and discrimination from *Helicobacter* and *Arcobacter* spp. by direct sequencing of PCR-amplified cpn60 sequences and comparison to cpnDB, a chaperonin reference sequence database. Journal of Medical Microbiology **55**(Pt 4): 393-399.

Hill, J.E., Penny, S.L., Crowell, K.G., Goh, S.H., and Hemmingsen, S.M. 2004. cpnDB: a chaperonin sequence database. Genome Research **14**(8): 1669-1675.

Hill, J.E., Seipp, R.P., Betts, M., Hawkins, L., Van Kessel, A.G., Crosby, W.L., and Hemmingsen, S.M. 2002. Extensive profiling of a complex microbial community by high-throughput sequencing. Applied and Environmental Microbiology **68**(6): 3055-3066.

Hill, J.E., Town, J.R., and Hemmingsen, S.M. 2006b. Improved template representation in cpn60 polymerase chain reaction (PCR) product libraries generated from complex templates by application of a specific mixture of PCR primers. Environmental Microbiology **8**(4): 741-746.

Hill, T.C.J., Walsh, K.A., Harris, J.A., and Moffett, B.F. 2003. Using ecological diversity measures with bacterial communities. FEMS Microbiology Ecology **43**(1): 1-11.

Hollingsworth, P.M., Graham, S.W., and Little, D.P. 2011. Choosing and using a plant DNA barcode. PLoS ONE **6**(5): e19254.

Hugenholtz, P. 2002. Exploring prokaryotic diversity in the genomic era. Genome Biology **3**(2): reviews0003-reviews0003.0008.

Hummelen, R., Fernandes, A.D., Macklaim, J.M., Dickson, R.J., Changalucha, J., Gloor, G.B., and Reid, G. 2010. Deep sequencing of the vaginal microbiota of women with HIV. PLoS ONE **5**(8): e12078.

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. 2007. MEGAN analysis of metagenomic data. Genome Research **17**(3): 377-386.

Janda, J.M., and Abbott, S.L. 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. Journal of Clinical Microbiology **45**(9): 2761-2764.

Johnston-Monje, D., and Raizada, M.N. 2011. Conservation and diversity of seed associated endophytes in Zea across boundaries of evolution, ethnography and ecology. PLoS ONE **6**(6).

Kampfer, P., and Glaeser, S.P. 2012. Prokaryotic taxonomy in the sequencing era--the polyphasic approach revisited. Environmental Microbiology **14**(2): 291-317.

Kawasaki, S., Fratamico, P.M., Wesley, I.V., and Kawamoto, S. 2008. Species-specific identification of Campylobacters by PCR-restriction fragment length polymorphism and PCR targeting of the gyrase B gene. Applied and Environmental Microbiology **74**(8): 2529-2533.

Kearns, L.P., and Hale, C.N. 1996. Partial characterization of an inhibitory strain of *Pantoea agglomerans* with potential as a biocontrol agent for *Erwinia amylovora*, the fire blight pathogen. Journal of Applied Bacteriology **81**(4): 369-374.

Kempf, H.J., and Wolf, G. 1989. *Pantoea agglomerans* as a biocontrol agent of *Fusarium culmorum* and *Puccinia recondita* f. sp. tritici on wheat. Phytopathology **79**(9): 990-994.

Kerr, K.C., Stoeckle, M.Y., Dove, C.J., Weigt, L.A., Francis, C.M., and Hebert, P.D. 2007. Comprehensive DNA barcode coverage of North American birds. Molecular Ecology Notes **7**(4): 535-543.

Kim, J., Demeke, T., Clear, R.M., and Patrick, S.K. 2006. Simultaneous detection by PCR of *Escherichia coli*, *Listeria monocytogenes* and *Salmonella typhimurium* in artificially inoculated wheat grain. International Journal of Food Microbiology **111**(1): 21-25.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., and Glockner, F.O. 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Research **41**(1): e1.

Konopka, A. 2009. What is microbial community ecology? ISME Journal **3**(11): 1223-1230.

Konstantinidis, K.T., Ramette, A., and Tiedje, J.M. 2006. The bacterial species definition in the genomic era. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences **361**(1475): 1929-1940.

Lane, D.J., Pace, B., Olsen, G.J., Stahl, D.A., Sogin, M.L., and Pace, N.R. 1985. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proceedings of the National Academy of Sciences of the United States of America **82**(20): 6955-6959.

Langmead, B., and Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods **9**(4): 357-359.

Large, A.T., Goldberg, M.D., and Lund, P.A. 2009. Chaperones and protein folding in the archaea. Biochemical Society Transactions **37**(Pt 1): 46-51.

Lee, D.H., Zo, Y.G., and Kim, S.J. 1996. Nonradioactive method to study genetic profiles of natural bacterial communities by PCR-single-strand-conformation polymorphism. Applied and Environmental Microbiology **62**(9): 3112-3120.

Lee, L.Y., and Liu, S.T. 1991. Characterization of the yellow-pigment genes of *Pantoea agglomerans*. Molecular Microbiology **5**(1): 217-224.

Li, W., and Godzik, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22**(13): 1658-1659.

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B., and Fan, W. 2012. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. Briefings in Functional Genomics **11**(1): 25-37.

Links, M.G., Dumonceaux, T.J., Hemmingsen, S.M., and Hill, J.E. 2012. The chaperonin-60 universal target is a barcode for bacteria that enables de novo assembly of metagenomic sequence data. PLoS ONE **7**(11): e49755.

Lozupone, C., and Knight, R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. Applied and Environmental Microbiology **71**(12): 8228-8235.

Lucero, M.E., Unc, A., Cooke, P., Dowd, S., and Sun, S. 2011. Endophyte microbiome diversity in micropropagated *Atriplex canescens* and *Atriplex torreyi* var *griffithsii*. PLoS ONE **6**(3).

Lund, P.A. 2009. Multiple chaperonins in bacteria--why so many? FEMS Microbiology Reviews **33**(4): 785-800.

Lundberg, D.S., Lebeis, S.L., Paredes, S.H., Yourstone, S., Gehring, J., Malfatti, S., Tremblay, J., Engelbrektson, A., Kunin, V., del Rio, T.G., Edgar, R.C., Eickhorst, T., Ley, R.E., Hugenholtz, P., Tringe, S.G., and Dangl, J.L. 2012. Defining the core Arabidopsis thaliana root microbiome. Nature **488**(7409): 86-90.

Magan, N., and Aldred, D. 2007. Post-harvest control strategies: minimizing mycotoxins in the food chain. International Journal of Food Microbiology **119**(1-2): 131-139.

Magan, N., Aldred, D., Mylona, K., and Lambert, R.J.W. 2010. Limiting mycotoxins in stored wheat. Food Additives and Contaminants - Part A Chemistry, Analysis, Control, Exposure and Risk Assessment **27**(5): 644-650.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., and Rothberg, J.M. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature **437**(7057): 376-380.

Masson, L., Maynard, C., Brousseau, R., Goh, S.H., Hemmingsen, S.M., Hill, J.E., Paccagnella, A., Oda, R., and Kimura, N. 2006. Identification of pathogenic *Helicobacter* species by chaperonin-60 differentiation on plastic DNA arrays. Genomics **87**(1): 104-112.

Meier, R., Zhang, G., and Ali, F. 2008. The use of mean instead of smallest interspecific distances exaggerates the size of the "barcoding gap" and leads to misidentification. Systems Biology **57**(5): 809-813.

Meyer, C.P., and Paulay, G. 2005. DNA barcoding: error rates based on comprehensive sampling. PLoS Biology **3**(12): e422.

Mollet, C., Drancourt, M., and Raoult, D. 1997. rpoB sequence analysis as a novel basis for bacterial identification. Molecular Microbiology **26**(5): 1005-1011.

Nawrocki, E.P., Kolbe, D.L., and Eddy, S.R. 2009. Infernal 1.0: inference of RNA alignments. Bioinformatics **25**(10): 1335-1337.

Notredame, C., Higgins, D.G., and Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of Molecular Biology **302**(1): 205-217.

Olaimat, A.N., and Holley, R.A. 2012. Factors influencing the microbial safety of fresh produce: A review. Food Microbiology **32**(1): 1-19.

Oliver, K.L., Hamelin, R.C., and Hintz, W.E. 2008. Effects of transgenic hybrid aspen overexpressing polyphenol oxidase on rhizosphere diversity. Applied and Environmental Microbiology **74**(17): 5340-5348.

Paramel Jayaprakash, T., Schellenberg, J.J., and Hill, J.E. 2012. Resolution and characterization of distinct cpn60-based subgroups of *Gardnerella vaginalis* in the vaginal microbiota. PLoS ONE **7**(8): e43009.

Pei, A.Y., Oberdorf, W.E., Nossa, C.W., Agarwal, A., Chokshi, P., Gerz, E.A., Jin, Z., Lee, P., Yang, L., Poles, M., Brown, S.M., Sotero, S., Desantis, T., Brodie, E., Nelson, K., and Pei, Z. 2010. Diversity of 16S rRNA genes within individual prokaryotic genomes. Applied and Environmental Microbiology **76**(12): 3886-3897.

Post, A.F., Penno, S., Zandbank, K., Paytan, A., Huse, S.M., and Welch, D.M. 2011. Long term seasonal dynamics of *Synechococcus* population structure in the Gulf of Aqaba, Northern Red Sea. Frontiers in Microbiology **2**: 131.

Price, M.N., Dehal, P.S., and Arkin, A.P. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Molecular Biology and Evolution **26**(7): 1641-1650.

Pruitt, K.D., Tatusova, T., Brown, G.R., and Maglott, D.R. 2012. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Research **40**(Database issue): D130-135.

Rastogi, G., Sbodio, A., Tech, J.J., Suslow, T.V., Coaker, G.L., and Leveau, J.H.J. 2012. Leaf microbiota in an agroecosystem: Spatiotemporal variation in bacterial community composition on field-grown lettuce. ISME Journal **6**(10): 1812-1822.

Rozen, S., and Skaletsky, H. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods in Molecular Biology **132**: 365-386.

Saitou, N., and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution **4**(4): 406-425.

Sakamoto, M., and Ohkuma, M. 2010. Usefulness of the hsp60 gene for the identification and classification of Gram-negative anaerobic rods. Journal of Medical Microbiology **59**(Pt 11): 1293-1302.

Sakamoto, M., Suzuki, N., and Benno, Y. 2010. hsp60 and 16S rRNA gene sequence relationships among species of the genus *Bacteroides* with the finding that *Bacteroides suis* and *Bacteroides tectus* are heterotypic synonyms of *Bacteroides pyogenes*. International Journal of Systematic and Evolutionary Microbiology **60**(Pt 12): 2984-2990.

Sanger, F., and Coulson, A.R. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology **94**(3): 441-448.

Schadt, E.E., Turner, S., and Kasarskis, A. 2010. A window into third-generation sequencing. Human Molecular Genetics **19**(R2): R227-240.

Schellenberg, J., Links, M.G., Hill, J.E., Dumonceaux, T.J., Peters, G.A., Tyler, S., Ball, T.B., Severini, A., and Plummer, F.A. 2009. Pyrosequencing of the chaperonin-60 universal target as a tool for determining microbial community composition. Applied and Environmental Microbiology **75**(9): 2889-2898.

Schellenberg, J., Links, M.G., Hill, J.E., Hemmingsen, S.M., Peters, G.A., and Dumonceaux, T.J. 2011a. Pyrosequencing of chaperonin-60 (cpn60) amplicons as a means of determining microbial community composition. Methods in Molecular Biology **733**: 143-158.

Schellenberg, J.J., Links, M.G., Hill, J.E., Dumonceaux, T.J., Kimani, J., Jaoko, W., Wachihi, C., Mungai, J.N., Peters, G.A., Tyler, S., Graham, M., Severini, A., Fowke, K.R., Ball, T.B., and Plummer, F.A. 2011b. Molecular definition of vaginal microbiota in East African commercial sex workers. Applied and Environmental Microbiology **77**(12): 4066-4074.

Schloss, P.D., Gevers, D., and Westcott, S.L. 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS ONE **6**(12): e27310.

Schloss, P.D., and Handelsman, J. 2003. Biotechnological prospects from metagenomics. Current Opinion in Biotechnology **14**(3): 303-310.

Schloss, P.D., and Handelsman, J. 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Applied and Environmental Microbiology **71**(3): 1501-1506.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Van Horn, D.J., and Weber, C.F. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology **75**(23): 7537-7541.

Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., and Chen, W. 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proceedings of the National Academy of Sciences of the United States of America **109**(16): 6241-6246.

Sneath, P.H.A. 2010. Reflections on microbial systematics. *In* The Bulletin of BISMiS. *Edited by* J. T. Staley, Athens Georgia, USA. pp. 77 - 83.

Sokal, R.R., and Sneath, P.H. 1963. Principles of Numerical Txonomy. W.H. Freeman and Company, San Francisco & London.

Srinivasan, S., Hoffman, N.G., Morgan, M.T., Matsen, F.A., Fiedler, T.L., Hall, R.W., Ross, F.J., McCoy, C.O., Bumgarner, R., Marrazzo, J.M., and Fredricks, D.N. 2012. Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. PLoS ONE **7**(6): e37818.

Stackebrandt, E., and Ebers, J. 2006. Taxonomic parameters revisited: tarnished gold standards. Microbiology Today **33**: 152-155.

Stahl, D.A., Lane, D.J., Olsen, G.J., and Pace, N.R. 1985. Characterization of a Yellowstone hot spring microbial community by 5S rRNA sequences. Applied and Environmental Microbiology **49**(6): 1379-1384.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C.J., Osborne, B.I., Pocock, M.R., Schattner, P., Senger, M., Stein, L.D., Stupka, E., Wilkinson, M.D., and Birney, E. 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Research **12**(10): 1611-1618.

Staley, J.T., and Konopka, A. 1985. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annual Review of Microbiology **39**: 321-346.

Sundquist, A., Bigdeli, S., Jalili, R., Druzin, M.L., Waller, S., Pullen, K.M., El-Sayed, Y.Y., Taslimi, M.M., Batzoglou, S., and Ronaghi, M. 2007. Bacterial flora-typing with targeted, chip-based Pyrosequencing. BMC Microbiology **7**: 108.

Sylvan, J.B., Toner, B.M., and Edwards, K.J. 2012. Life and death of deep-sea vents: bacterial diversity and ecosystem succession on inactive hydrothermal sulfides. MBio **3**(1): e00279-00211.

Tamura, K., Dudley, J., Nei, M., and Kumar, S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Molecular Biology and Evolution **24**(8): 1596-1599.

Tamura, K., Nei, M., and Kumar, S. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proceedings of the National Academy of Sciences of the United States of America **101**(30): 11030-11035.

Taylor, J.L. 1993. A simple, sensitive, and rapid method for detecting seed contaminated with highly virulent *Leptosphaeria maculans*. Applied and Environmental Microbiology **59**(11): 3681-3685.

Thompson, J.D., Gibson, T.J., and Higgins, D.G. 2002. Multiple sequence alignment using ClustalW and ClustalX. Current Protocols in Bioinformatics **Chapter 2**: Unit 2 3.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research **22**(22): 4673-4680.

Tikhonovich, I.A., and Provorov, N.A. 2011. Microbiology is the basis of sustainable agriculture: An opinion. Annals of Applied Biology **159**(2): 155-168.

Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., and Gordon, J.I. 2007. The Human Microbiome Project. Nature **449**(7164): 804-810.

Vandamme, P., Pot, B., Gillis, M., de Vos, P., Kersters, K., and Swings, J. 1996. Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiological Reviews **60**(2): 407-438.

Verbeke, T.J., Sparling, R., Hill, J.E., Links, M.G., Levin, D., and Dumonceaux, T.J. 2011. Predicting relatedness of bacterial genomes using the chaperonin-60 universal target (cpn60 UT): application to *Thermoanaerobacter* species. Systematic and Applied Microbiology **34**(3): 171-179.

Vermette, C.J., Russell, A.H., Desai, A.R., and Hill, J.E. 2010. Resolution of phenotypically distinct strains of *Enterococcus* spp. in a complex microbial community using cpn60 universal target sequencing. Microbial Ecology **59**(1): 14-24.

Vos, M., Quince, C., Pijl, A.S., de Hollander, M., and Kowalchuk, G.A. 2012. A comparison of rpoB and 16S rRNA as markers in pyrosequencing studies of bacterial diversity. PLoS ONE **7**(2): e30600.

Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology **73**(16): 5261-5267.

Ward, D.M., Weller, R., and Bateson, M.M. 1990. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. Nature **345**(6270): 63-65.

Wayne, L.G., Brenner, D.J., Colwell, R.R., Grimont, P.A.D., Kandler, O., Krichevsky, M.I., Moore, L.H., Moore, W.E.C., Murray, R.G.E., Steckebrandt, E., Starr, M.P., and Truper, H.G. 1987. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. International Journal of Systematic Bacteriology **37**(4): 463-463.

Weiss, A., Hertel, C., Grothe, S., Ha, D., and Hammes, W.P. 2007. Characterization of the cultivable microbiota of sprouts and their potential for application as protective cultures. Systematic and Applied Microbiology **30**(6): 483-493.

Weller, R., and Ward, D.M. 1989. Selective recovery of 16S rRNA sequences from natural microbial communities in the form of cDNA. Applied and Environmental Microbiology **55**(7): 1818-1822.

Wendland, J., Lengeler, K.B., and Kothe, E. 1996. An instant preparation method for nucleic acids of filamentous fungi. Fungal Genetics Newsletter **43**: 54-55.

Whitaker, R.H. 1960. Vegatation of the Siskiyou Mountains, Oregon and California. Ecological Monographs **30**(3): 279-338.

Whitman, W.B., Coleman, D.C., and Wiebe, W.J. 1998. Prokaryotes: the unseen majority. Proceedings of the National Academy of Sciences of the United States of America **95**(12): 6578-6583.

Woese, C.R., and Fox, G.E. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences of the United States of America **74**(11): 5088-5090.

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N.N., Kunin, V., Goodwin, L., Wu, M., Tindall, B.J., Hooper, S.D., Pati, A., Lykidis, A., Spring, S., Anderson, I.J., D'Haeseleer, P., Zemla, A., Singer, M., Lapidus, A., Nolan, M., Copeland, A., Han, C., Chen, F., Cheng, J.F., Lucas, S., Kerfeld, C., Lang, E., Gronow, S., Chain, P., Bruce, D., Rubin, E.M., Kyrpides, N.C., Klenk, H.P., and Eisen, J.A. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature **462**(7276): 1056-1060.

Yu, F., Lydiate, D.J., and Rimmer, S.R. 2005. Identification of two novel genes for blackleg resistance in *Brassica napus*. TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik **110**(5): 969-979.

Zahariev, M., Dahl, V., Chen, W., and Levesque, C.A. 2009. Efficient algorithms for the discovery of DNA oligonucleotide barcodes from sequence databases. Molecular Ecology Resources **9 Suppl s1**: 58-64.

Zegada-Lizarazu, W., and Monti, A. 2011. Energy crops in rotation. A review. Biomass and Bioenergy **35**(1): 12-25.

Zeigler, D.R. 2003. Gene sequences useful for predicting relatedness of whole genomes in bacteria. International Journal of Systematic and Evolutionary Microbiology **53**(Pt 6): 1893-1900.

Zuckerkandl, E., and Pauling, L. 1965. Molecules as documents of evolutionary history. Journal of Theoretical Biology **8**(2): 357-366.