Utility of redesigned *cpn*60 UT primers and novel fungal specific *cpn*60 primers for microbial profiling

A Thesis Submitted to the College of

Graduate Studies and Research

In Partial Fulfillment of the Requirements

For the Degree of Master of Science

In the Department of Microbiology and Immunology

University of Saskatchewan

Saskatoon

By

NEERZA BANSAL

PERMISSION TO USE

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Microbiology and Immunology
2D01, Health Sciences Building
107 Wiggins Rd
University of Saskatchewan
Saskatoon, Saskatchewan (S7N 5E5)
Canada.

Abstract

The *cpn*60 gene is a DNA barcode for bacteria. Recently, the PCR primers that have been used extensively to amplify the *cpn*60 Universal Target (UT) region of bacteria were redesigned to improve their utility for fungal taxa. Additional novel primers were designed to amplify other regions of the *cpn*60 gene, specifically from fungal genomes. Design of the redesigned and novel primers was based on 61 nucleotide full-length *cpn*60 reference sequences available in 2012, including Ascomycota (51), Basidiomycota (5), Chytridiomycota (2), Glomeromycota (1), and Oomycota (2). The research described here investigated the utility of these primers for detecting and identifying fungal taxa and for profiling mixed communities of bacteria and fungi. The redesigned primers were used to discover *cpn*60 UT sequences for Ascomycota (1), Basidiomycota (2), and Chytridiomycota (1). The novel primers were used to discover new *cpn*60 sequence data for Ascomycota (3), Basidiomycota (1), and Zygomycota (1). To be adopted for use in studies of microbial communities that are predominantly bacterial, the redesigned *cpn*60 UT primers must perform at least as well as the original primers for bacterial profiling. Bacterial profiles, created using the original and redesigned primers and two DNA template samples created by pooling DNA extracts from vaginal swabs from individual women, were compared. These included comparisons of diversity indices, rarefaction curve analysis and Operational Taxonomic Unit abundances. Diversity indices and rarefaction curve analysis for bacterial profiles with original and redesigned primers were similar. OTU abundance estimates with the original and redesigned primers were compared at higher and lower taxonomic levels. The overall patterns produced were similar. For one template only, the phylum Bacteroidetes had a greater apparent abundance with the original primers than with the redesigned primers. The greater apparent abundance of Bacteroidetes taxa was balanced by a lesser apparent abundance of taxa that were not assigned to a phylum. These differences may reflect differences in the performance of the two primer sets. At lower taxonomic level, most OTU were represented with apparently equal abundances with redesigned and original primers in same template. Very few OTU were represented with different proportional abundances with redesigned and original primers. Different OTU having same reference *cpn*60 UT sequence as best hit were sometimes represented by different proportional abundance with same primer in same template that made the analysis difficult. On the whole, the redesigned *cpn*60 UT primers behaved at least as good as the original *cpn*60 UT primers. The overall results showed that the redesigned and novel primers used in this study had substantial utility for the identification of fungal samples and mixed microbial communities.

Acknowledgements

It is sometimes difficult to express gratitude in words because feelings go beyond it. My journey to the final destination was accompanied with enjoyment, moments of satisfaction, hardship and sometimes frustration. I would like to express my gratitude to those who were there for me and provided guidance to get to my final destination.

I would humbly like to express my sincere gratitude to my supervisor Dr. Sean Hemmingsen, whose willingness to help and efforts in making things easier for me were without compare. I thank him for sincerely welcoming me into his lab. His guidance and encouragement allowed me to rediscover my potential during frustrating times.

I also wish to extend a special thanks to my committee members, Dr. Hughes Goldie and Dr. Linda Chelico for their great support. I am equally grateful to my external evaluator, Dr. Vladimir Vujanovic, for his time, and for the valuable suggestions to improve the final document.

I deeply appreciate Brenda Haug for her excellent technical assistance and our little conversations. Your presence in the lab was a great moral support at all times. Thank you Brook and Cindy for sharing a constructive work space and making my stay pleasant in lab.

My heartfelt gratitude to Dr. Peter Howard and Dr. Aaron White for giving me encouraging feedback that kept me going.

I am also thankful to Dr. Janet Hill for letting me use the pyrosequencing set up in her lab. Thank you for making me a part of many events in your lab and I wish I had connected more. Thanks are also to Dr. Bonnie Chaban for helping me use her expertise to run my pyrosequencing experiments.

I am thankful to Dr. Levesque and Dr. Dumonceaux for providing the fungal samples. I would like to thank National Research Council, Saskatoon for providing the facility and the Canadian Institute of Health Research for financial assistance.

Thank you Ravi and Jyoti for welcoming us to Saskatoon with open arms. We would not have been here but for you. Also, my thanks to Dr. Umashankar Das and Dr. Swagatika Das for their help and encouragement. A special thanks to Bharti aunty for her constant moral support, great memories, chats, tea and snack sessions, trips and her family for always being there for us.

The deepest gratitude goes to my loving parents and my brother and his family. Their infallible belief was the pillar of my strength. This accomplishment would have simply been impossible without them. I am also thankful to my in-laws for their love and constant support.

Finally, I would like to thank my husband for loving me unconditionally and supporting me throughout the times I felt incapable and demotivated. I cannot thank my lovely daughters enough for letting me steal their time to complete my studies. Lastly, I would like to thank the spirit unknown that watches over me and motivates me to be a good human being.

I dedicate this thesis to my ever dearest husband

List of abbreviatons

Abbreviation	Description
Amp	Ampicillin
AMF	Arbuscular Mycorrhizal Fungi
ARDRA	Amplified Ribosomal DNA Restriction Analysis
BLAST	Basic Local Alignment Search Tool
BV	Bacterial Vaginosis
Cpn	Chaperonin
DGGE	Denaturing Gradient Gel Electrophoresis
DNA	Deoxy-ribo Nucleic Acid
EF	Elongation Factor
FAME	Fatty Acid Methyl Ester
FISH	Fluorescent *In situ* Hybridization
HIV	Human Immunodeficiency Virus
IGS	Inter Genic Spacer
ITS	Internal transcribed spacer
kDa	kilo Dalton
MEGAN	MEtaGenome Analyzer
MID	Multiplexing Identifiers
mPUMA	microbial Profiling Using Metagenomic Assembly
NGS	Next Generation Sequencing
OUT	Operational Taxonomic Unit

PAH	Poly Aromatic Hydrocarbons
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
RAPD	Random Amplified Polymorphic DNA
rRNA	ribosomal Ribo Nucleic Acid
SSCP	Single Strand Conformation Polymorphism
T-RFLP	Terminal- Restriction Fragment length Polymorphism
UT	Universal Target
UV	Ultra Violet
X-gal	5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside
YS	Yeast extract Soluble starch

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

1.0 Introduction and Review of Literature

Although life first emerged on earth 3.8 billion years ago, microorganisms were recognised as life forms only relatively recently by Robert Hooke (in 1660) and Leeuwenhoek in about 1673, where they used simple microscopes that could magnify 50 to 300 times. Development of microscopy set out to be the key to recognition of microorganisms at that time but is non-specific as stained smears of microorganisms do not lead to species identification. Further progress in the study of microorganisms was made when among other discoveries, Louis Pasteur gave his "Germ theory of disease" in 1857 stating that diseases can be caused by microorganisms. His theory was further proved by Robert Koch in 1882. Koch provided Koch`s postulates that could be used to identify the causative agent of an infectious disease. A microbiologist in his laboratory, Julius Richard Petri, invented the indispensable petri-dish in 1887, that led to the rigorous isolation and identification of bacteria in Koch`s laboratory (Bulloch, 1938). This ability to isolate microorganisms and study them in pure cultures was key to development of methods to study their physiology and genetics. The ability to study microbes in pure cultures added much information between 1960s and 1980s to the then existing knowledge of microbiology. But the scientific community soon realized that pure culturing alone cannot give them the full spectra of microbial diversity. It was not sufficient to reveal the unimaginable diversity of 8.7 million (±1.3 million SE) species existing in the microbial communities (Whitman et al., 1998) that are found in nature as it is not possible to culture all microbes using standard methods. Moreover, the actual interactions among microbes in a microbial community cannot be studied under a microscope or in pure cultures. Exploration of this tremendous diversity is necessary for the advancement of microbiology and improving of our understanding of the vast and complex microbial world.

The constraints associated with microscopy and pure cultures, methods that provided us information about the microbes based only on morphological and nutritional criteria, were alleviated, although not fully, by development of molecular tools like gene sequencing. The basic step to this was the discovery of DNA structure by Watson and Crick in 1953. The differentiation between prokaryotes and eukaryotes came later in 1962. Then in 1977, Carl Woese and George Fox reported three domains of life as Eukaryotes, Bacteria and Archaea. This was done by comparing the rRNA gene sequences of organisms and that has become the standard approach to classify and identify organisms. This was followed by development of molecular sequencing methods that are based on comparing the nucleic acid sequences of specific gene targets from different organisms (O'Sullivan, 2000; Hill et al., 2002). These methods exploit the polymerase chain reaction (PCR) for gene isolation and amplification. The most popular gene target for identification of bacteria using molecular methods has been 16S rRNA (Stahl et al., 1984; Lane et al., 1985) and for identification of fungi has been 18S rRNA and ITS regions (Guo et al., 2012; Schmidt et al., 2013). There has been a large scale development of high throughput sequencing methods using different gene targets. The next generation sequencing methods like pyrosequencing can now give up to million sequencing reads with an average read length of ~700 bp with 99.9% accuracy (www.454.com). In spite of such a boom in sequencing methods, a study estimates that out of the ~8.7 million species (±1.3 million SE) predicted to exist on earth, there may be 86% species on Earth and 91% species in oceans that have still not been described (Mora et al., 2011). A combination of culture-dependent and culture independent techniques can prove useful in describing this unrevealed diversity.

1.1 Culture Dependent Techniques for Characterizing Microbial Communities

Standard methods of culturing used to characterize microbial communities involve isolation and culturing of microbes using media like Luria-Bertani medium, nutrient agar, Tryptic Soy agar (Kirk et al., 2004). The limitation of this method is that >99% of the viable microorganisms from any environment observed under microscopes fail to produce visible colonies on plates (Staley and Konopka, 1985; Hugenholtz, 2002). To improve the cultivable fraction in a given microbial sample, several culture media and cultivation conditions are used that imitate the natural conditions for that sample, like nutrient composition, oxygen concentration and availability, pH, temperature requirements, long incubation conditions for slow growing bacteria (da Rocha et al., 2009; Vartoukian et al., 2010). These methods have been further combined with sophisticated high throughput techniques like micro-chip based culturing or encapsulating microbes in agar droplets so that they are physically separated from one another but interact with environment and one another (Ben-Dov et al., 2009). In spite of applying all these methods in combination, the ratio of uncultured to cultured microbes still remains high. This difference between the number of microbes actually present in a particular community to the number of microbes that can be cultured on plates has been termed as "the great plate count anomaly" (Staley and Konopka, 1985).

Taxonomic diversity of small eukaryotes like fungi has been suggested to be around 1.5 million species (Hawksworth, 1991) to around as much as 6 million species (Taylor et al., 2014). Out of the estimate of 6 million fungal species, results by Taylor et al., 2014 suggest that 98% still remain undescribed. One of the major problems in detecting fungi from environmental communities like soil is that they are fastidious in nature and estimates show that only 17% of known fungi can be grown in culture. Moreover, just one type of media does not suffice for the growth of all the fungi. Different taxonomic groups may need different types of media. For example, benomyl or dichloran added to potato dextrose agar, which otherwise is a general purpose media for fungi, is more effective for the general isolation of basidiomycetes (Worrall, 1991). Therefore, culture dependent methods bias the results for diversity towards those fungi in a microbial sample that can be grown in culture. Therefore, scientists shifted focus on culture independent methods for detecting fungi and bacteria in microbial samples.

1.2 Culture Independent Techniques for Characterizing Microbial Communities

The culture independent techniques include polymerase chain reaction based methods (PCR). In these methods, DNA or RNA is extracted from microbial samples and is used as a template for detection of microorganisms. The main source of information from uncultured microorganisms in culture independent techniques is their biomolecules such as nucleic acids, lipids and proteins. PCR amplification of conserved genes using universal primers is widely used for microbial profiling by nucleic acid approach and protein encoding gene approach. Among these, the nucleic acid approach includes the analysis of 16S rRNA and 18S rRNA from prokaryotes and eukaryotes respectively. In the lipid analysis approach, microbial cells in a community have their own distinctive FAME (Fatty acid methyl ester) profiles that can be used for their taxonomic classification. This method has been used to study whole cell FAME profiles of 605 *E.coli* isolates to establish their host specificity (Haznedaroglu et al., 2007). Protein based approaches use the genes encoding proteins as targets for microbial identification. Two protein encoding genes that have been used for microbial profiling are rpoB (Mollet et al., 1997) and *cpn*60 (Goh et al., 1996). The PCR products obtained as a result of amplification can be analysed in one or all of these ways: the clone library method, genetic fingerprinting and DNA microarrays.

1.2.1 Clone Library Method

The original method used to analyze amplified PCR products from environmental microbial samples was to clone and sequence the obtained amplicons for species identification. This method produces a high phylogenetic resolution by direct species identification or comparing the conserved gene sequences from microbial sample to a reference sequence database and finding the closest similarity to a known species. The clone library method revealed phylogenetic diversity in microbial community samples (Singleton et al., 2001). But since this method is labour intensive, expensive and time consuming; therefore, for some types of studies the clone library method has been supplanted by methods based on next generation sequencing. However, the method remains an important approach in many labs.

1.2.2 Genetic Fingerprinting

Genetic fingerprinting techniques are used to compare profiles of microbial communities although they do not provide any direct taxonomic information about the microbes present in those communities. They allow the simultaneous analysis of multiple samples for example, while comparing the genetic diversity of microbial samples from different environments or studying microbial succession in a particular community over time. The analysis is based on the "fingerprints" produced by gene variants for an individual species assumed to be present in a microbial community. These "fingerprints" from different samples are then compared using software packages like GelCompar (Stahl and Capman, 1994; Muyzer, 1999; Rastogi and Sani, 2011). Some of the commonly used genetic fingerprinting techniques are discussed here.

1.2.2.1 Denaturing Gradient Gel Electrophoresis (DGGE)

In DGGE, the PCR products obtained from environmental DNA samples are separated electrophoretically on a polyacrylamide gel that already contains a linear gradient of denaturing agents like urea and formamide. The DNA extracted from a complex group of microorganisms is amplified using primers specific for molecular markers like 16S rRNA. To prevent the complete separation of double strands, a 5'- GC clamped (30-50 nucleotide) forward primer is used during PCR reaction. The denaturants in the gel melt the double helical form of DNA, as a result, the mobility of denatured DNA decreases which is dependent on the nucleotide variation present among DNA from different species in a microbial community sample. Therefore DNA molecules with different sequences stop migrating on the gel at different positions. Migrating patterns formed by different samples on the same gel can be compared to see apparent differences or similarities in the behaviour of those communities. For phylogenetic identification, the bands can be excised from the gel, re-amplified and sequenced. Another variation of the same technique is TGGE, temperature gradient gel electrophoresis, where temperature gradient is used instead of chemical denaturants. The disadvantages of both techniques are that relatively short fragments (~500 bp) can be separated, which provide limited phylogenetic information about the microbes. Different DNA molecules can have similar melting points; sequence variation among multiple RNA copies in same species can produce multiple bands leading to over-estimation of diversity. DGGE was applied to study soils collected from different agricultural fields. One of the soil samples from these fields was highly contaminated with polyaromatic hydrocarbons (PAH). Bacterial and archaeal profiles were generated using 16S rRNA primers. It was found that overall bacterial diversity was much more than archaeal diversity in different soil samples except in samples from soils with high PAH content (Nakatsu et al., 2000).

1.2.2.2 Single Strand Conformation Polymorphism (SSCP)

In SSCP, the DNA amplicons are denatured to single stranded DNA fragments and then are separated electrophoretically in a non-denaturing gel (Schwieger and Tebbe, 1998). The separation of single stranded DNA is based on their nucleotide differences that may be as little as a single base pair. This may lead to different secondary structure conformations and thus mobility in the gel. Unlike DGGE, it is simpler, as it does not require 5'-GC clamped forward primers or gradient gels. Also, the bands can be excised as in DGGE and the DNA can be extracted, re-amplified and sequenced. The technique can be further redesigned for determining the predominant bacterial population in the community by hybridizing the DNA strands with taxon-specific probes. But this technique is suitable for the separation of small fragments (150-400 nucleotides) only (Muyzer, 1999). Another disadvantage is that the DNA strands can reanneal after initial denaturation step during electrophoresis. Although this has been overcome by using phosphorylated primers during PCR, and later on, phosphorylated strands can be specifically digested using lambda exonuclease. SSCP analysis was applied to study the rhizosphere bacterial populations associated with two plants growing in the same soil, *Medicago sativa* and *Chenopodium album*. The analysis showed that both plants had different rhizosphere bacteria inspite of their growth in same soils (Schwieger and Tebbe, 1998).

1.2.2.3 Random Amplified Polymorphic DNA (RAPD)

RAPD uses very short primers (5-10 nucleotides) that randomly anneal at different positions on genomic DNA thereby generating amplicons of variable lengths that are separated on a polyacrylamide gel. The annealing occurs at very low temperature (≤ 35°C) and separation is based on genetic complexity of microbial community sample used. The advantage is that it has a high speed and is easy to use (Franklin et al., 1999). The disadvantages are that it is very sensitive to experimental conditions like $MgCl_2$ concentration, annealing temperatures, differences in quality and quantity of template DNA and primers (Hadrys et al., 1992). Therefore, to reveal the differences or similarities between different microbial communities, several combinations of experimental conditions and primers need to be evaluated. As part of a study, changes in microbial diversity in soil samples that were treated with pesticides and chemical fertilizers were assessed using 14 random primers. The results showed that pesticide treated soils maintained the same level of microbial diversity as uncontaminated soil (control) whereas chemical fertilizer treated soil had decreased levels of microbial diversity than control (Yang et al., 2000). In another study, DNA diversities of soil microbial communities in rhizosphere and non-rhizosphere in plant *Panax ginseng* were evaluated using RAPD. Total genomic DNA from soil samples was amplified by 24 primers. The study revealed that microbial diversity of rhizosphere soil was lower than that of non-rhizosphere soil (Yong et al., 2012).

1.2.2.4 Amplified Ribosomal DNA Restriction Analysis (ARDRA)

In ARDRA, DNA fragments are generated using PCR primers for a molecular target which are then digested with restriction enzymes and separated on an agarose or polyacrylamide gel. The technique is based on the principle that the restriction sites on the RNA operons are conserved according to phylogenetic patterns (Massol-Deya et al., 1995). It has proved useful for studying the changes in microbial communities happening over a course of time, estimating number of OTU in clone libraries or for identifying unique clones (Smit et al., 1997). In one of the studies, the ARDRA technique allowed the recognition of 3-4 *Gardnerella vaginalis* genotypes. Some genotypes were found to be more prevalent in certain areas from which they were collected (Ingianni et al., 1997). In another study, ARDRA was used to study the community composition of normal and bulking activated sludge. It was found that the microbial community composition of normal and bulking sludge was different, although it was not possible to determine the exact species or strains of filamentous bacteria responsible for bulking of sludge (Blaszczyk et al., 2011). Therefore, ARDRA is suitable for comparing the microbial diversities but it provides little or no information about the identity of microorganisms present in sample. Sometimes, the restriction profiles generated from microbial communities are too complex to be resolved on electrophoretic gels.

1.2.2.5 Terminal- Restriction Fragment length Polymorphism (T-RFLP)

T-RFLP is a modification of ARDRA. T-RFLP uses 5' fluorescently labelled primers during the PCR reaction. The amplicons are digested using restriction enzymes and the resulting fragments are separated on an automated DNA sequencer. Only those bands that are fluorescently labelled are detected. This produces a much simplified pattern of bands, thus allowing better analysis of complex microbial communities. A study was conducted to understand how bacterial communities develop in *Apis* species (honey bee) midgut. PCR amplification was done using 16S rRNA primers. T-RFLP analysis resulted in 16 distinct terminal restriction fragments (T-RFs). The T-RFs belonged to Beta and Gammaproteobacteria, Firmicutes and Actinomycetes. Gammaproteobacteria were found to be present in all stages of honey bee and Firmicutes were present in only worker bees additionally (Disayathanoowat et al., 2012). The advantages of using T-RFLP for microbial analysis are that it has a high resolution of separation on automated DNA sequencer, comparison between different samples can be done by using different fluorescent labels on different lanes, the bands or peaks can be quantified directly. The drawbacks of using T-RFLP are that the bands cannot be excised and sequenced and the automated DNA sequencer used for separation of T-RFs is very expensive.

1.2.3 DNA Microarrays

A microarray is an orderly arrangement of molecular probes (with known identity) that can be DNA, cDNA or oligonucleotides and are immobilized on a solid support (a microscope glass slide, silicon chips or nylon membrane). The PCR products amplified from total environmental DNA are fluorescently labelled and can be directly hybridized to the molecular probes. DNA microarrays exploit the fact that complimentary strands of nucleic acid base pair with each other and bind. The unbound molecules are then washed away. Positive or negative signals of hybridization are scored by the use of confocal laser scanning microscopy (Gentry et al., 2006; Rastogi and Sani, 2011). A microarray method was developed to differentiate between two taxonomic neighbours, *Helicobacter* and *Campylobacter* species, and to identify clinically relevant *Helicobacter* species. *Helicobacter* species are responsible for many hepatic, biliary and enteric diseases. Both these species are often misidentified under clinical settings. Amplicons were produced using *cpn*60 and 16S rRNA universal primers from a complex human waste sludge DNA samples spiked with *Helicobacter* species. The amplicons were hybridized to specific *cpn*60 and 16S rRNA fragents from *Helicobacter* and *Campylobacter* species immobilized on plastic chips. The study resulted in accurate *Helicobacter* species identification with no cross-hybridization to either 16S rRNA and *cpn*60 fragments obtained from closely related strains of *Campylobacter* species (Masson et al., 2006). Using DNA microarrays for microbial profiling has many advantages. The samples can be rapidly evaluated with replication. The hybridization signal intensity is directly proportional to the abundance of the target species. The major limitation of using microarrays with environmental samples is cross-hybridization; moreover, it is not helpful in detection of novel taxa that may be present in microbial community samples if there is no matching probe on the array.

1.2.4 Fluorescent *In situ* Hybridization (FISH)

FISH allows *in situ* detection and identification of individual microbial cells with the help of fluorescent oligonucleotide probes that bind to those DNA or RNA sequences in the cells that are highly complementary to the probes. The sample is fixed to stabilize the microbial cells using fixatives like formaldehyde or ethanol, and then cells are permeablized, the protocol for which really depends on the composition of cell wall, for example, use of lysozyme for digestion of peptidoglycan, protease for proteinaceous cell walls, removal of wax by solvents etc. Then the probe is added which is around 18-30 nucleotides long and contains a fluorescent dye at the 5' end. The probe binds to its intracellular targets before the excess probe is washed away. The fluorescent probe bound to its intracellular target is then detected by epifluorescence microscopy (Amann and Fuchs, 2008; Rastogi and Sani, 2011). This method was used to study the dynamics of bacterial communities in crop soils treated with herbicides (Caracciolo et al., 2010). Molecular probes targeted at phylogenetic groups α, β, γ and δ of bacteria were made. The herbicide treated soils were incubated with soil samples for 14 days. It was observed that in comparison to control soil (untreated with herbicide), γ-proteobacteria diminished sharply after 14 days, β-proteobacteria populations remained higher than control and α and δ populations were not really affected by use of herbicides.

1.2.5 Immunological Detection Methods

Immunological detection methods are being increasingly used in microbial ecology for identification of specific organisms and for microbial community analysis. The sensitivity of advanced immunological methods is similar to PCR techniques. The detection of microbes by these methods is based on antigen-antibody interaction, where a particular antibody will bind to its specific antigen. However, for a reliable use of these techniques, the monoclonal antibodies or polyclonal antibodies used have to fulfill several quality criteria. These methods can be used for the identification of specific microbes in samples as well as for the visualization of cells *in situ*. They are fast, specific and can be automated to make them more labor-saving and time-efficient. However, cross-reactivity with closely related antigens is a problem as it may lead to non-specific reactions. Different methods are used for immunological detection, some of them are ELISA (Enzyme –Linked ImmunoSorbent Assay and lateral flow assay (Immunochromatographic assays). The sensitivity and specificity for these methods depends on antibody, for example, the detection limit is usually around 105 bacteria per mL in ELISA and 107 bacteria per mL using a lateral flow assay. The time taken by lateral flow assay is 10 min and by ELISA is several hours (Schloter et al., 1995). Immunological method can be used to detect both bacteria (Law et al., 2014) and fungi (Yeo and Wong, 2002) from environmental samples.

1.3 Gene Targets for Characterizing Bacterial Communities

The most important part of molecular microbial analysis is selection of an appropriate gene target. The gene should have a variable segment that should be common to the group or subgroup of interest being studied and it should be flanked by conserved regions. The conserved regions are the ones on which the DNA sequencing primers are based and they make the gene universal. These primers amplify the variable regions during PCR and generate amplicons that are unique to different species. The amplicon sequences are compared to reference sequences in a database. If the target gene is too long, it is difficult to sequence it completely and if the gene is too short, its sequence may not be enough to decide the genus or species to which it belongs. An ideal target gene should be present as single copy gene for accurate quantification purposes. Multicopy genes may not give accurate quantification and, if there is intragenomic variation among these copies, it may also over-estimate diversity in microbial communities. Some of the gene targets used for microbial analysis are discussed briefly here.

1.3.1 16S rRNA

In 1980s, Woese et al. developed a new method to identify bacteria based on the genes encoding 5S, 16S and 23S rRNA; although, the 16S rRNA gene is the part most commonly used, presently, for taxonomic purposes and microbial community analysis (Olsen et al., 1986; Pace et al., 1986; Woese, 1987; Suau et al., 1999; Hopkins et al., 2001; Matsuki et al., 2002; Salzman et al., 2002; Bartram et al., 2011; Poretsky et al., 2014). 16S rRNA is universal in bacteria and is targeted by sets of broad range PCR primers that can be used for the amplification of large number of variable regions. It also has a large reference database. In spite of these positive features, the comparison of 16S rRNA gene sequences allows differentiation between bacteria at genus level but it has a low phylogenetic power at species level owing to insufficient sequence variation (Fox et al., 1992; Clayton et al., 1995; Goh et al., 1996; Coenye et al., 2003) and poor differentiating power for some genera (Zeigler, 2003; Sundquist et al., 2007). Also, presence of multiple gene copies along with evolutionarily diverged copies of 16S rRNA is one of the disadvantages of using it as gene target (Goh et al., 1996). In an *in silico* study done by Vetrovsky and Baldrian, 7,081 16S rRNA sequences were extracted *in silico* from 1,690 available genomes. It was observed that sequence diversity increases with increasing copy numbers, whereas, in some cases, sequences may be common to multiple species, thereby, complicating the studies involving abundance counts and not providing a clear picture of bacterial community composition (Větrovský and Baldrian, 2013).

1.3.2 Internal Transcribed Spacer Region

The Internal transcribed spacer region in rRNA in bacteria is the region between 16S and 23S rRNA gene. Since variation in 16S rRNA gene is insufficient to identify all bacteria below genus level, 16S-23S ITS region was investigated as a potential alternative target for bacterial identification since it has more extensive sequence variation than 16S rRNA gene. This led to the observation that the spacer region can be good source of species specific sequences. The spacer sequences can be amplified by making oligonucleotide primers based on the 16S and 23S rRNA sequences that flank spacer regions (Barry et al., 1991). ITS region sequences were compared for then known *Bartonella species*. It was observed and confirmed that each species had a single species-specific ITS sequence, thereby confirming usefulness of ITS region for subtyping of *Bartonella* species of human and animal origins and understanding the epidemiology of these bacteria (Houpikian and Raoult, 2001). Hoffman et al., used this region, also called as intergenic spacer region or IGS region for the accurate and rapid identification of different *Vibrio* species. They used capillary gel electrophoresis to analyze the PCR products from IGS regions and IGS-typing patterns for each strain were tested. It was found that each *Vibrio* species had a unique typing pattern that could be used to identify each species in the complex *Vibrio* genus (Hoffmann et al., 2010).

1.3.3 *rpo*B gene

The *rpo*B gene is a protein-encoding gene that encodes the β-subunit of RNA polymerase in bacteria and is used for the phylogenetic analysis and identification of bacteria. The universality of *rpo*B gene was first reported by Morse et al. in 1996. The main advantages of using *rpo*B over 16S rRNA are that due to more sequence variation within *rpo*B gene, it provides higher resolution among closely related species. It is a single copy gene which makes it more useful for quantification of species or measuring relative abundance of different species in a microbial community (Rowland et al., 1993). A study on marine environment microbial community was done to see if use of *rpo*B as gene target could avoid limitations of using 16S rDNA as gene target, like intraspecies heterogeneity. As a part of this study, 16S rRNA and *rpo*B DGGE based comparison of microbial community analysis was done on samples from marine red alga. Eight out of 14 isolates displayed multiple bands by 16S rRNA DGGE analysis whereas *rpo*B did not show any intraspecies heterogeneity on DGGE analysis (Dahllöf et al., 2000). *rpo*B has also been used to profile bacterial diversity from tropical soils (Peixoto et al., 2002), kefir grains (Wang et al., 2006), goat rumen (Shi et al., 2007) etc. The disadvantages associated with use of *rpo*B as a gene marker are that it is not conserved enough to be a universal marker, although it can be used to target a particular subset of microbial community. Taxonomic identification of sequences is a problem due to unavailability of appropriate an database as of 2012 (Vos et al., 2012).

1.3.4 *gyr*B gene

The *gyr*B is another target gene used as DNA probe and has a higher specificity than rRNA based probes. *gyr*B genes encode the subunit B protein of DNA gyrase also known as topoisomerase type 2. It is necessary for DNA replication and it regulates supercoiling of double stranded DNA. It is distributed universally among bacterial species. It has been shown to be a suitable phylogenetic marker for study of taxonomic relationships at species level in microbial communities found in activated sludge (Watanabe et al., 1998), acid mine drainage in copper mines in China (Yin et al., 2008). Although a database of *gyr*B sequences was published (Kasai et al., 1998), it is limited to bacterial sequences and it remains doubtful if it is currently being maintained (Hill et al., 2004).

1.3.5 *rec*A gene

Another alternative to 16S rRNA that can be used as a gene probe is the recombinase A gene (*rec*A). This protein is universally present in bacteria and is also one of the most conserved proteins across bacteria. This protein is required for homologous recombination, DNA repair and the SOS response (Karlin et al., 1995). The *rec*A gene has been used to identify six *Bifidobacteria* species from human intestinal tract isolates (Kullen et al., 1997). The *rec*A based gene analysis was also applied to maize rhizosphere where it revealed novel diversity among *Burkholderia* genus (Payne et al., 2006). It has also been shown to be a useful tool in addition to 16S rRNA for revealing the evolutionary relationships between Rapidly Growing Mycobacterium (RGM) species. Presently, no database is available for *rec*A genes.

1.3.6 *cpn*60, Proposed DNA Barcode for Bacteria

Another target used for microbial profiling methods is the *cpn*60 (*gro*EL) gene that has been recently proposed to be adapted as a barcode for the identification of bacteria (Links et al., 2012). The *cpn*60 gene encodes a 60 kDa protein and is present in all bacteria, and in mitochondria and chloroplasts of eukaryotes. The name ‘chaperonin’ (*cpn*) was proposed for this ubiquitous and conserved protein that assists in the correct post-translational assembly of other polypeptides into oligomeric complexes (Hemmingsen et al., 1988). This gene encodes Group 1 chaperonins, has either a 552, 555 or 558 bp segment that can be amplified with universal PCR primers and is called the “universal target” region. Original universal degenerate primers (H279 and H280) were designed based on highly conserved regions within the *cpn*60 gene (or *hsp*60 or *gro*EL) from different organisms. Inosines were added to these primers at specific locations to decrease the degeneracy of the primers (Goh et al., 1996). The original primers have also been modified to make them suitable for the amplification of difficult templates like those rich in G+C content. These modified primers when used in addition to the regular universal primers, have been able to represent the diversity in microbial communities more accurately (Hill et al., 2006).

The *cpn*60 gene has been initially exploited in many studies for species-specific identification like identification of *Staphylococcus* species and subspecies (Goh et al., 1996; Goh et al., 1997), *Streptococcus suis* serotypes (Brousseau et al., 2001), *Enterococcus* species from phenotypically similar *Lactococcus* and *Vagococcus* species (Goh et al., 2000). Later on, in additional studies, it has proved to be a useful tool in the characterization of microbial communities from different environments like pig intestinal microbial community (Hill et al., 2002), activated sludge communities (Dumonceaux et al., 2006b), vaginal microbial communities (Hill et al., 2005; Schellenberg et al., 2011; Chaban et al., 2014), faecal communities of different animals (Dumonceaux et al., 2006a; Desai et al., 2009).

The *cpn*60 gene provides many advantages over 16S rRNA as gene target. The *cpn*60 based sequencing provides more discriminating and phylogenetically informative data than the 16S rRNA target, especially between closely related species (Goh et al., 1996; Brousseau et al., 2001; Zeigler, 2003). The *cpn*60 gene usually occurs as a single copy gene in bacteria, making it attractive quantitative target. Even if it occurs as multiple a copy gene, the copies are sufficiently different from each other, thus, acting as independent phylogenetic targets. The *cpn*60 UT is of relatively small size, which facilitates the study of microbial communities where large libraries of fragments are sequenced or in pyrosequencing where short sequence read lengths are obtained. Finally, the sequences can be compared with the *cpn*60 database (Hill et al., 2004).

Recently, *cpn*60 was evaluated for its status as a DNA barcode for bacteria (Links et al., 2012). Barcodes are short and specifically designed DNA sequences that can be used to identify organisms by comparing the barcode sequence from an unknown organism to a collection of known sequences from a reference database. The *cpn*60 gene was shown to fulfill the requirements for a gene to be classified as a barcode for bacteria. It is universal among bacteria and universal primers have already been developed that can amplify the universal target from *cpn*60 gene of any bacteria. A huge collection of reference sequences (cpnDB) is available for robust identification of organisms. Species-level and even subspecies identification is provided by *cpn*60 gene in metagenomic studies whereas such identification is not reported with 16S rRNA. However, species level identification is desirable in some studies. For example, human vaginal microbiome is dominated by *Lactobacilli*, and many efforts have been made in research studies to resolve the *Lactobacilli* species using 16S rRNA (Hummelen et al., 2010; Srinivasan et al., 2012) whereas, species resolution of *Lactobacilli* has been easy and rapid using *cpn*60 UT, by sequence comparison with the reference database. In another study on human vaginal microbiota, *cpn*60 UT data has been used to resolve *G. vaginalis* into subspecies (Jayaprakash et al., 2012). Additionally, the inter-specific distance is greater than the intra-specific distance for *cpn*60 sequence, which is one of the important criterias for a gene to be defined as a barcode . The separation between the average interspecific and intraspecific distance for a given locus is called a ‘barcode gap’ Moreover, the absence of length variation in *cpn*60 UT sequences (552, 555 or 558 bp) makes it suitable to use either local or global alignment when comparing sequences.

1.4 Gene Targets for Characterizing Fungal Communities

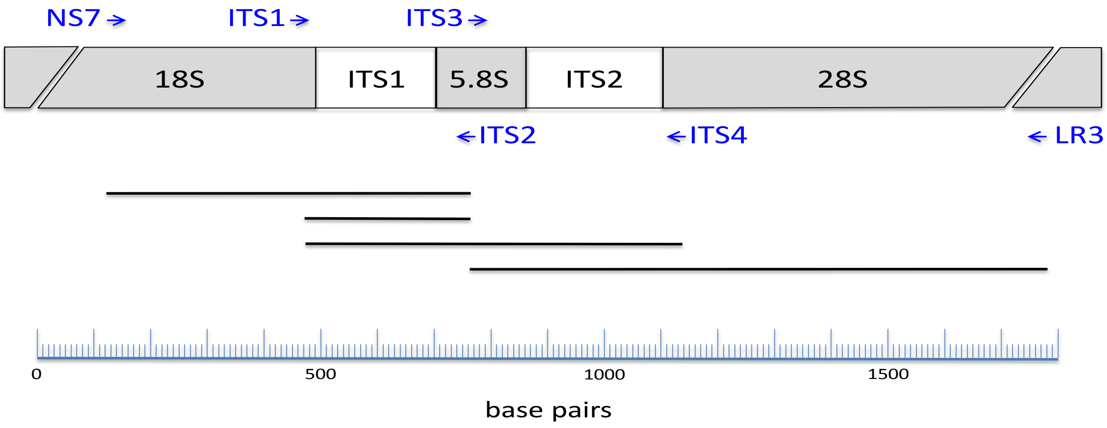
1.4.1 18S rRNA

For fungi, sequences of the nuclear ribosomal RNA genes (nrDNA) are the most commonly used genetic markers for phylogenetic and taxonomic identification (Hibbett et al., 2007). 18S rRNA gene has been the most widely used nuclear ribosomal gene, using both variable and conserved regions (White et al., 1990; Smit et al., 1999; Borneman and Hartin, 2000; Vainio and Hantula, 2000; Zheng et al., 2013; Buse et al., 2014). Taxonomic identification of fungi with 18S rRNA has been limited to genus and family level, due to lack of variation within 18S rRNA gene between closely related fungal species (Hugenholtz and Pace, 1996). An absence of an extensive reference database further adds to its limitation. But the variation in 18S rRNA gene for Glomeromycota has been fairly sufficient to differentiate between species (SCHÜßLER et al., 2001). They have been used to differentiate Arbuscular Mycorrhizal fungi (belong to Glomeromycota) up to species and subspecies level (Vandenkoornhuyse and Leyval, 1998).

1.4.2 Internal Transcribed Spacer Region

The ITS sequences have been proposed to be adapted as the fungal barcode for identification of fungal species at lower levels (Schoch et al., 2012). The ITS region (Figure 1) includes the ITS1 and ITS2 regions, separated by the 5.8S gene and is situated between the small subunit (SSU: 18S) and large subunit (LSU: 28S) genes in the nrDNA gene. Fungal metagenomic studies may target different regions of ITS in parallel for identification of all the constituent species (multi-region approach) as targeting a single region may not be sufficient to reveal all the diversity (Bellemain et al., 2010). ITS primers may be biased towards some taxa, therefore, it has been suggested that they should be used in combination with LSU and SSU primers (Toju et al., 2012). Targeting of ITS1 and ITS2 regions biases the amplification towards Ascomycota whereas targeting only the ITS1 region may lead to bias towards ‘non-dikarya’ fungi and LSU has been adopted as a gene marker for yeasts. The fungal rDNA is present as a multicopy gene in fungal genomes. This increases the sensitivity of PCR assay, but during analysis of a microbial community sample, due to the variability in copy number among different fungal species (from tens to several hundred), the number of sequence reads attributed to any fungal species may be wrongly magnified (Black et al., 2013). Another problem currently faced with profiling fungal communities is the limited availability of reference data for comparison of experimental sequences. As of 2012, there were only ~172,000 full length fungal ITS sequences in Genbank. Although ITS is useful in discriminating phylogenetically distant species, its ability to distinguish closely related fungal species is doubtful because of the substantial intragenomic variability present within the species (Nilsson et al., 2008; Kiss et al., 2012).

Figure 1: Generalized structure of fungal rRNA locus as represented on the fungal rRNA gene in *Serpula himantioides* (AM946630) modified from Bellemain et al., 2010.



(Based on *Serpula himantiodes* sequence, AM946630)

The figure shows positions of primers and expected length of sequences obtained with different primers. Grey boxes show the small subunit (18S), 5.8S and large subunit (28S) regions of rRNA. White boxes are the internal transcribed spacer regions (ITS1 and ITS2). The expected lengths of sequences are depicted by black lines. ITS3-LR3 (1000bp), ITS1-ITS2 (300bp), ITS1-ITS4 (~610bp), NS7-ITS2 (~600bp).

1.4.3 Protein Coding Genes

Protein coding genes are also used for species identification in fungi. In fact, for Ascomycota they have been able to determine the taxonomic levels in a finer way than the rRNA genes (Schoch et al., 2009). RPB1 and RPB2 (RNA polymerase II largest and smallest subunit), Elongation factor 1 alpha (EF1-α) and *cpn*60 are the protein coding genes that have been largely sequenced and used for microbial analysis. When the performance of protein coding genes was tested and compared with that of ribosomal RNA genes, it was found that RPB1 and RPB2 gave the best resolution under most of the situations (Liu et al., 2006; Hofstetter et al., 2007) than rRNA genes. In spite of protein coding genes having better species resolving power than rRNA genes, PCR and sequencing failures limit their use as gene targets for identification of fungi. Recently, the *cpn*60 gene was used for simultaneous profiling of bacteria and fungi associated with seeds (Links et al., 2014).

1.4.4 *cpn*60 as Gene Target for Detection of Fungi

A potential advantage of using *cpn*60 as a gene target for profiling microbial communities is that it can be used simultaneously to identify both bacteria and fungi, unlike other gene targets like 16S rDNA that are limited to bacteria or 18S rDNA and ITS, that are used to identify fungi and other eukaryotes. As vaginal microbiome may have both bacteria and fungi, *cpn*60 can be a useful gene target to profile the same. It has already been used to obtain bacterial profiles from vaginal samples in a variety of studies (Hill et al., 2005; Schellenberg et al., 2011; Chaban et al., 2014). The cpnDB had 19,667 entries as of 4 August, 2014. In 2012, cpnDB had 61 full length fungal *cpn*60 sequences (Hemmingsen, unpublished). These numbers show the obvious lag cpnDB has in the context of fungal sequences. The original *cpn*60 primers are known to amplify the fungal *cpn*60 UT gene from microbial community samples, but they had not been studied systematically to determine their ability to amplify fungal *cpn*60 gene sequences from environmental samples. An *in silico* approach showed that 33 of these 61 sequences should be amplified by the original primers. In the remaining 28 cases, there is a one amino acid difference (serine (S) instead of threonine (T)) in the C-terminal residue of the coding amino acid consensus sequence as compared to the original forward primer (H279) amino acid consensus sequence (Figure 2).

Primer	cpnDB ID	Aligned Amino Acid Sequences										Description
H279		-	E/D	X	A	G	D	G	T	T	T	Original Consensus Sequence
	b1554	N	E	V	A	G	D	G	T	T	T	*S. pombe*
	b198	N	E	S	A	G	D	G	T	T	S	*C. albicans*
	b10353	N	E	A	A	G	D	G	T	T	S	*S. cerevisiae*
H1780		N	E/D	X	A	G	D	G	T	T	-	Revised Consensus Sequence

Figure 2: Region of *cpn*60 amino acid sequences used for design of primers H279 and H1780. The amino acid consensus sequence at the top (in red) was used for design of original degenerate primer H279. The consensus sequence used for primer H1780 is based on sequences for 61 full length *cpn*60 sequences for fungi in cpnDB out of which three are shown here. The C-terminal residue of the amino acid sequence consensus is serine (S) instead of threonine (T) in many fungi including *C. albicans* and *S. cerevisiae*. For H1780, the consensus sequence for primer is moved one codon to left, so that the C-terminal residue of the consensus amino acid is threonine only and so that this does not affect the length of the primer, "N" (asparagine) is included as the first N-terminal amino acid of this primer. "X" can be any amino acid.

The consensus sequence for the H279 primer shows that it should not be able to amplify these 28 sequences including *C. albicans* and *S. cerevisiae*. Hemmingsen redesigned primer H279 to accommodate this difference (unpublished). The resulting primer is H1780. For H1780, the consensus sequence for primer is moved one codon to the left, so that the C-terminal residue of the consensus amino acid sequence is threonine only and so that this does not affect the length of the primer. In addition, he designed novel primers to amplify regions of *cpn*60 from fungal templates. The novel fungal primers may anneal at regions conserved among fungi and they may be able to amplify the UT from any fungal isolate that may be present in environment samples (Hemmingsen, unpublished).

1.5 The Rare Biosphere Concept

The complex microbial communities of mucosa are dominated by a relatively small number of species, a larger number of low abundance species or OTUs also exist that form 'the rare biosphere'. For example in human faeces, the fungal microbiome forms the rare biosphere as compared to the bacterial microbiome (Huffnagle and Noverr, 2013). Some members of this fungal microbiome can become potentially pathogenic if the mucosal environment is disturbed. This holds true for the vaginal mucosa too. Therefore, complete profiling of both the dominating and the low abundance species in a microbial community is essential for complete understanding of the disease. It is easier to identify the dominating taxa in a community but real challenge is posed by the rare microbes.

1.6 Next Generation Sequencing Technologies

The completion of the Human Genome Project in 2003 used the first generation sequencing called the Sanger sequencing (dideoxy chain termination method) which remained the fundamental method of large scale genome sequencing for many years. Another first generation sequencing method by Maxam and Gilbert involved nucleobase specific chemical modification of DNA, followed by cleavage of DNA at site adjacent to redesigned nucleotide. This method used hazardous radioactive materials and was technically complex; therefore, is no longer in widespread use. Except for the Maxam and Gilbert sequencing method, all other methods use sequencing by synthesis. The Human Genome project stimulated improvements in Sanger sequencing like the use of fluorescent dyes, polymerases specifically designed for sequencing and improvements in software packages for sequence analysis and made it a high throughput method of sequencing. It took 13 long years to accomplish the project, and this led to the demand for faster and cheaper sequencing methods. This led to the development of next generation sequencing (NGS) where millions of fragments of DNA from a single sample are sequenced simultaneously. NGS permits massive sequencing with a much higher throughput than Sanger sequencing. The most currently used NGS technologies include 454 sequencing (Roche applied science, Basel, Switzerland), Illumina/Solexa genome analyzer (Illumina, San Diego, CA, USA), SOLiD (Applied Biosystems, Foster City, CA, USA), HeliScope Single Molecular Sequencer (Helicose Biosciences, Cambridge, MA, USA), and the Single Molecule Real Time Technology (SMRT, Pacific Biosciences, Menlo Park, CA, USA). All these platforms perform massive parallel sequencing. The first three platforms sequence clonally amplified products and the last two, sequence single DNA molecules.

1.6.1 Pyrosequencing

Pyrosequencing is a widely used next generation sequencing technology. The Sanger sequencing approach is considered the first generation technology and is associated with high cost and technical difficulties like analyzing large numbers of clones from large numbers of samples.

Also, the method is expected to reveal only the dominant members of the microbial community and the sampling depth is low. An advantage is that it is capable of sequencing 900-1200bp. Pyrosequencing provides large numbers of sequence reads in a single run, giving very large sampling depth and allowing detection of both dominant as well as rare taxa present in a microbial community. Pyrosequencing eliminates the need for cloning the amplicons generated by PCR, thereby removing at least one of the biases. Although the output read length is shorter than that obtained by Sanger sequencing (~250 bp for GS-FLX and ~800 bp for GS-FLX Titanium series and ~900 bp using Sanger) (454.com), for *cpn*60 amplicons, sequences of >150 bp are sufficient to determine organism identities (Schellenberg et al., 2009). The latest addition to the sequencing technology with long-read sequencing performance is the 454 GS-Junior System that gives an average read length of ~400 bp and the accuracy is 99%. The run time is only 10 hours for sequencing and 2 hours for data processing. The number of amplicon reads it gives per run is 70,000 (http://www.454.com/). One of the disadvantages of 454-pyrosequencing is a high error rate in the homopolymer regions (three or more consecutive identical DNA bases). The 454-pyrosequencing has now phased out due to the advent of lower cost and higher throughput sequencing technologies like Illumina-Solexa Miseq is capable of generating 25 million reads with 98% accuracy in ~55h and Life Technologies SOLiD 5500 series can generate 1.2-1.4 billion reads in 1-2 weeks at a low cost (Gilles et al., 2011).

Principle of pyrosequencing: Sequencing is by synthesis and it involves light generation after nucleotides are incorporated in a growing chain of DNA. PCR amplicons are amplified using MID (multiplexing identifier) tagged primers. The use of MIDs enables the simultaneous sequencing of multiple libraries and generates microbial profiles for large number of samples in a single sequencing reaction. Libraries are made by ligating short DNA sequencing adaptors to

MID-tagged amplicons. The DNA libraries thus made are immobilized on DNA capture beads. Each bead has a unique ssDNA oligonucleotide sequence that is complementary to the sequence of adaptors. The bead-bound library is emulsified with amplification reagents in a water-in-oil mixture resulting in micro-reactors containing one bead bound to one DNA fragment. The emulsion PCR (emPCR) amplification takes place inside this microreactor. After amplification, the emulsion is broken. DNA is denatured, beads having ssDNA are transferred to picotitre plates where one bead rests in one picotitre well.

Inside the well, a DNA fragment attached to a bead is the template which is hybridized to the sequencing primers. One of the dNTPs (N=A,T,C,G) is added to the reaction. DNA polymerase catalyses the addition of this dNTP if it is complementary to the base on the template strands and a pyrophosphate (PPi) is released. ATP sulfurylase converts PPi to ATP in the presence of substrate adenosine 5` phosphosulfate (APS). The released ATP provides energy for conversion of luciferin to oxyluciferin catalysed by luciferase. Oxyluciferin produces visible light proportional to amount of ATP which is detected by a CCD chip as a peak in the program output. Apyrase (ATP diphosphatase) degrades ATP and unincorporated dNTP after which another cycle of nucleotide addition starts (http://www.454.com/).

1.6.2 Other Next Generation Sequencing technologies

In the Sequencing by Oligonucleotide Ligation and Detection (SOLiD), sequencing is obtained by measuring ligation of an oligonucleotide to a sequencing primer by a DNA ligase enzyme. DNA fragments are ligated to oligonucleotide adapters that are attached to beads. DNA fragments are then amplified by emulsion PCR until it provides sufficient signal for the sequencing reactions. Beads are deposited on a flow cell surface. Sequencing primers are annealed to the adapter sequences on each amplified fragment and, with this, the ligase mediated sequencing begins. Each ligation step is accompanied by fluorescence detection. A regeneration step prepares the extended primer for the next ligase reaction.

With Illumina sequencing, in each sequencing cycle, a single labelled deoxynucleoside triphosphate is added to nucleic acid chain (the four dNTPs have different labels). The labels, such as a fluorescent dye, acts as a terminator. Dye is imaged to identify the dNTP added and is enzymatically cleaved so that next dNTP can be added.

The HeliScope Biosystem is a single molecule sequencing system. It also utilizes the sequencing by synthesis principle. The DNA sample is fragmented and polyadenylated at the 3` end and final adenine is labelled with Cy3 fluorescent dye. PolyT oligos are immobilized on flow cell surface and polyA template molecules get attached to them by hybridization. The labels can be imaged to identify the DNA molecule and then cleaved. The cycle is repeated by adding each of Cy3 labelled nucleotides to flow cell.

1.7 Metagenomics

Metagenomics literally means 'beyond the genome' (Gilbert and Dupont, 2011). It is the cultivation independent analysis of the collective genomes of microbes within a given environment. Therefore, metagenomics has made it possible to sequence libraries from a mixture of organisms. It is now feasible to conduct sequence based studies on organisms that were previously considered to be inaccessible like obligate pathogens and symbionts, that do not survive outside their hosts; microorganisms in environmental samples that cannot be grown in pure cultures and primitive organisms for which information is only available in their fossilized remains (Tringe and Rubin, 2005). Metagenomic studies are useful in increasing our understanding of structure (gene and species richness) and function of environmental microbial communities. The metagenomic approach involves the extraction and isolation of DNA from environmental samples and the DNA samples should be representative of the population of all organisms present in the environment to be studied. However, DNA obtained from community of microbes may or may not provide the complete genomic picture of the microorganisms in that environment, as this mostly depends on our ability to sample (Wooley et al., 2010). This is because the genomic material from the more abundant organism dominates the sample (Tringe and Rubin, 2005). Study of organisms that make up an acid-mine biofilm was the first environmental metagenomic study (Tyson et al., 2004). The acid-mine biofilms are formed when FeS_2 from mining drainage is exposed to water and sulphuric acid is produced. Microbial communities with low diversity flourish in these biofilms due to extreme acidic conditions. Tyson et al., generated 76.2 million bp of sequence from biofilm bacteria and archeans. Almost two complete genomes and three partial genomes were assembled from this data. In one of the other prominent studies, a project to sequence the entire metagenome of Sargasso Sea surface waters was taken up by Venter et al., and unexpected community diversity and complexity was revealed (Venter et al., 2004). Metagenomics has also been used to study the differences in fungal communities present in healthy and dandruff affected human scalp using 26S rDNA as gene target. The study showed that *Acremonium* (Ascomycete) was abundant on both the healthy and affected individuals, *Cryptococcus* (Basidiomycete) was abundant on healthy scalp and *Filobasidium* sp. (Basidiomycete) was mostly present on dandruff affected scalp (Park et al., 2012).

1.8 Fungi

The fungi are a group of diverse organisms that are characterized by non-motile bodies (thalli) made of elongated walled filaments (hyphae), both sexual and asexual reproduction, heterotrophic nutrition, chitin and glucans as cell wall components. Spindle pole bodies are associated with the nuclear envelope during cell division (Griffin, 1994). Some of the well-known exceptions to these characteristics are chytrids, that have flagella at some stage of their life cycle and have centrioles associated with cell division instead of spindle bodies (Morgan et al., 2007). Some members of Ascomycota, Basidiomycota and Mucoromycotina do not have hyphal growth during part or all of their life cycles. A few species of Ascomycota are characterized by cellulose in their cell walls (Alexopoulos et al., 1996).

The diversity of fungi has been estimated to be 1.5 to 5.1 million species (Taylor et al., 2010; Blackwell, 2011) out of which only about 100,000 have been described. The study of morphological characteristics and advances in the molecular sequencing methods have revolutionized the classification of such a large diversity of fungi. According to O`Brein et al., it may still take about 4000 years to describe all these species using the current approach, so that all of them may be discovered before becoming extinct (O'Brien et al., 2005). Therefore, the need to quicken the process of describing fungi is becoming crucial.

Fungi were initially categorized as a subkingdom in the Kingdom Plantae. The subkingdom had two divisions, Myxomycota (for plasmodial forms) and Eumycota (for non-plasmodial forms) (Ainsworth et al., 1973). Eumycota included subdivisions Mastigomycotina (Chytridiomycetes, Hyphochytridiomycetes, Oomycetes), Zygomycotina (Zygomycetes, Trichomycetes), Ascomycotina, Basidiomycotina and Deuteromycotina. Myxomycota were categorized separately under Kingdom Protista. Later, fungi were considered entirely distinct from plants and were classified into a separate kingdom, Kingdom Fungi (Whittaker and Margulis, 1978). In 1993, Baldauf and Palmer provided evidence that fungi are more closely related to animals than plants by examining sequences from 25 proteins. Among other evidence, it was found by them that four insertions/deletions are uniquely shared by animals and fungi relative to plants, protists, and bacteria (Baldauf and Palmer, 1993). Kirk et al. (2001) further redesigned the classification by accepting phyla Ascomycota, Basidiomycota, Chytridiomycota and Zygomycota within the Kingdom fungi (Ainsworth, 2008) while Myxomycota and Oomycota were excluded from the Kingdom (Berbee and Taylor, 1993; Berbee and Taylor, 1995). Recently, Hibbet et al., proposed a broad phylogenetic classification of Kingdom Fungi. This classification accepts one kingdom, one subkingdom, seven phyla, ten subphyla, 35 classes, 12 subclasses and 129 orders. Dikarya is a sub-kingdom classified into phyla Ascomycota and Basidiomycota. The other five phyla are Chytridiomycota, Neocallimastigomycota, Blastocladiomycota, Microsporidia and Glomeromycota. The traditional phyla Zygomycota and Chytridiomycota have undergone many key changes. The taxa that were originally included in Zygomycota have been distributed between Glomeromycota and four subphyla *incertae sedis* (term used for classifying taxa of uncertain position). Members of Neocallimastigomycota, Blastocladiomycota, and Microsporidia were traditionally placed under Chytridiomycota and have now been elevated to phylum based on their morphology and molecular phylogeny (Hibbett et al., 2007).

1.9 The Human Vaginal Microbiome

The human vaginal microbiome plays an important role in reproductive health and disease. Studies have shown that the dominant bacteria present in vagina are *Lactobacilli* species *L. crispatus*, *L. gasseri, L.jensenni* and *L.iners*, the major component being *L.iners*. Sometimes the healthy vaginal flora may also be replaced by other lactic acid producing bacteria like *Atopobium vaginae*, *Megasphaera* and *Leptotrichia* species. These organisms maintain reproductive health by resisting infection, that may be caused by various pathogens, by producing many factors, for example, they excrete lactate thus reducing the pH of vagina and production of H_20_2 and bacteriocin by some strains is also known to discourage the growth of some bacterial genera like *Streptococcus, Gardnerella vaginalis, Prevotella/Bacteroides species, Peptostreptococcus species, Mycoplasma hominis, Ureaplasma urealyticum and Mobiluncus species* which may be normally present in vagina, but may lead to an abnormal state if they tend to overgrow (Drell et al., 2013). This shift in the vaginal microbiota from a *Lactobacilli* dominated community to a community rich in aerobic/anaerobic potential pathogens present in a dense biofilm leads to a condition called bacterial vaginosis (BV) that is also clinically characterized by a thin, malodorous vaginal discharge. BV additionally leads to many negative impacts on a woman`s health like pelvic inflammatory disease, preterm births, and acquisition of sexually transmitted diseases (Hill et al., 2005). The information regarding the fungal component of the vaginal microbiome is exclusively derived from culture based investigations. The most frequently occurring fungal species in normal vaginal microbiota is *Candida albicans* that can cause vulvovaginal candidiasis in immunocompromised patients. Some non–albicans species are also identified in vaginal cultures like *Candida kefyr, Candida glabrata* and *Candida tropicalis* and are responsible for causing acute recurrent or chronic vulvovaginitis (Drell et al., 2013). Several novel bacterial species have been detected both in normal vaginal flora and in BV using the techniques mentioned above, but no study has concentrated on the detection of fungal species except the already known *Candida albicans* or some *non-albicans Candida* species. Recently, it has been shown that about 101 fungal species are present in the oral mycobiome as against the general perception that only *Candida albicans* or *non albicans Candida* are present in oral microflora. Some unexpected fungal species found were *Fusarium, Aspergillus* and *Cryptococcus* (Ghannoum et al., 2010). A recent culture independent study comparing fungal vaginal flora of healthy women and women suffering from RVC (recurrent vaginal candidiasis) using 18S rDNA as gene target reveals 10 phylotypes of fungi in healthy women and 28 phylotypes in all, although this study was not able to identify them at the species level (Guo et al., 2012). The *cpn*60 gene has not been yet systematically tested for profiling of fungal communities. Although it has been proposed to be the fungal DNA barcode, the ITS region has many drawbacks as a gene target like presence as multiple copies in the genome, intragenomic variability, variable length etc. Whether *cpn*60 can overcome these drawbacks as a fungal gene target, can be determined by using it for analysing fungal DNA samples and subsequently for fungal microbial communities like the vaginal mycobiome.

2.0 Goals

The overall goal of this study was to investigate the utility of the *cpn*60 gene for the detection and identification of Fungi. Already published primers have utility for detection and identification of bacteria and for generation of sequence based profiles of bacterial communities. There were no phylogenetic gaps found in their utility for these purposes except for *Mollicutes* like *Mycoplasma* and *Ureaplasma* which lack the *cpn*60 gene. In this study, it was evaluated if this was also true for fungal taxa. Redesigned versions of the forward UT primers were designed (Hemmingsen, unpublished) to address known deficiencies in the published primers for amplification of the UT from fungal templates. In addition, novel primers putatively specific for fungi were designed (Hemmingsen, unpublished). In section 3, these redesigned and novel primers were tested for their ability to amplify the *cpn*60 gene sequence from diverse fungal taxa. If the redesigned UT primers were to be introduced for use in mixed bacterial and fungal microbial communities, their performance with respect to the bacterial community must be unaffected by the introduced modification. This question is addressed in section 4.

3.0 Assessment of Redesigned and Novel PCR Primers for Amplification of *cpn*60 Gene Sequences from Phylogenetically Diverse Fungal Taxa

3.1 Hypotheses and Experimental Approach

3.1.1 The redesigned *cpn*60 UT primers should be able to amplify the *cpn*60 UT gene present in DNA extracts from pure cultures of *S. pombe* and *S. cerevisiae* and a broad range of fungal phyla from environmental samples.

3.1.2 The novel *cpn*60 primers (based on regions within and flanking the *cpn*60 UT) should be able to amplify the respective sequences flanking the *cpn*60 UT in DNA extracts from pure cultures of *S. pombe* and *S. cerevisiae* and a broad range of fungal phyla from DNA extracts from environmental samples. The following experimental approach was used to frame these hypotheses.

In 2012, cpnDB had 61 full length fungal *cpn*60 sequences (Hemmingsen, unpublished). An *in silico* approach showed that 33 of these 61 sequences should be amplified by the original primers. In the remaining 28 cases, the C-terminal residue is serine (S) instead of threonine (T), which would seriously impair base pairing between the original primers and the coding template at the 3` end of the primers (Figure 2). The consensus sequence for H279 primer shows that it should not be able to amplify these 28 sequences including *C. albicans* and *S. cerevisiae*. Hemmingsen redesigned primer H279 to accommodate this difference (unpublished). The resulting primer is H1780. For H1780, the consensus sequence for primer is moved one codon to left, so that the C-terminal residue of the consensus amino acid is threonine only. This does not affect the length of the primer. In addition, he designed novel primers to amplify regions of *cpn*60 from fungal templates. The novel fungal primers may anneal at regions conserved among fungi and they may be able to amplify the UT from any fungal isolate that may be present in environment samples (Hemmingsen, unpublished).

3.2 Objectives

3.2.1 Assessment of redesigned UT primers for amplification of *cpn*60 UT from fungal taxa represented in cpnDB: *S. pombe* and *S. cerevisiae* were chosen for this study

3.2.2 Ability of redesigned primers and novel primers to amplify *cpn*60 UT or portions of the *cpn*60 gene from DNA extracts from phylogenetically diverse fungal taxa and for other fungal taxa chosen for this study

3.3 Materials and Methods

3.3.1 DNA Extraction from *S. pombe* and *S. cerevisiae* Pure Cultures

YS Media (BIO101 systems) was prepared (500mL broth and 500mL with agar) and autoclaved at 121˚C for 15 min. Two YS plates were inoculated with *S. pombe* strain 922 and two plates were inoculated with *S. cerevisiae* strain 1285. The plates were incubated for 48 h at 30˚C. Isolated colonies from these plates were inoculated in two tubes containing 2mL YS broth for *S. pombe* and two tubes containing 2 mL YS broth for *S. cerevisiae*. The inoculated tubes were incubated in shaker at 30˚C for 24 h. 2 mL of this culture growth was inoculated into 40 mL YS broth, and incubated in shaker at 30˚C for 24 h. The culture was then centrifuged in corning tubes at 4000 rpm for 10 min. The cells were resuspended in 2 tubes each containing 365 µL Buffer B1 (50 mM Tris-Cl+50 mM EDTA with 0.5% Tween 20 and 0.5% Triton X-100) with 40 µL RNaseA. The cell suspensions were divided into two parts in bead beating tubes. Qiagen Genomic DNA buffer kit reagents were used in this experiment. The DNA was extracted from these suspended cells by following the yeast genomic DNA isolation procedure, that uses a combination of chemical, physical and enzymatic treatments to maximize DNA recovery (Apajalahti et al., 1998; Hill et al., 2002). 7.5 µL lysozyme (100mg/ml in water) and 20 µL proteinase K (20 mg/ml in water) was added to each tube and incubated at 37˚C for 30 min. Added 135 µL Buffer B2 (3 M guanidine HCl with 20% Tween 20), mixed and incubated at 50˚C for 30 min and put them at -70˚C for 30 min. In fume hood, put 700 µL 25:24:1 phenol:chloroform:isoamyl alcohol (by volume). After placing tubes on ice, used Bead Beater Fast prep unit (20 s, 5 speed) 3 times and centrifuged at 14,000 rpm for 15 min. Removed top phase in fume hood (~500 µL) in 1.5 mL tubes while avoiding the white interphase. Added 0.1volume (50 µL) of 3M Sodium acetate and 1.1 volume (550 µL) of isopropanol to each tube, mixed and centrifuged for 15min at 1400rpm. Poured off supernatant, washed pellet with 1mL 70% ethanol, and centrifuged at 14,000rpm for 5min. Poured off supernatant, dried pellet for 10-20 min and resuspended pellet in 50-100 µL 10mM TE buffer. Dissolved the pellet and stored DNA at -20˚C until use.

3.3.2 Quantification of DNA using Quant iT dsDNA kit (Qubit dsDNA BR assay; Life Technologies, Burlington, Canada). Set up the required number of 0.5 mL thin-wall, clear Qubit® assay tubes for standards and samples. The Qubit® dsDNA BR assay requires 2 standards. Made the Qubit® working solution by diluting the Qubit® dsDNA BR reagent 1:200 in Qubit® dsDNA BR buffer so that the final volume in each assay tube was 200 µL. Prepared sufficient Qubit® working solution to accommodate all standards and samples (for 2 standards and 4 samples in this case, 1200 µL of working solution (6 µL of Qubit® reagent plus 1994 µL of Qubit® buffer). Loaded 190 µL of Qubit® working solution into each of the tubes used for standards. Added 10 µL of each Qubit® standard to the appropriate tube, then mixed by vortexing 2–3 s. Loaded the Qubit® working solution into individual assay tubes so that the final volume in each tube after adding sample was 200 µL. Added each of samples to assay tubes containing the correct volume of Qubit® working solution (prepared in step 6), then mixed by vortexing 2–3 s. Allowed all tubes to be incubated at room temperature for 2 min. Noted the reading for each tube on the home screen for Qubit fluorometer. Concentration of sample was calculated as reading on screen (QF) multiplied by 200 and divided by number of microlitres of sample added to the Qubit assay tube.

3.3.3 Polymerase Chain Reaction

For a 50 µL reaction, 10X PCR buffer (5 µL), 50 mM MgCl2 (2.5 µL), 10 µM primer1 (2 µL), 10 µL primer2 (2 µL), 10 mM dNTPs (1 µL), 5 U/ µL Taq (0.5 µL), water (36 µL), template (1 µL) were added. The thermocycling parameters used were 5 min at 94˚C, 40 cycles of 30 s at 94˚C, 30 s at annealing temp, 45 s at 72˚C, and 10 min at 72˚C. Primers H1780, H1781, H280, H1613, 1786, H1787, 1788, 1789 were used in this experiment and their sequences are given in Table 1a and the regions they amplify are explained in Table 1b. The primer site positions and the amplicon sizes they amplify are shown in Figure 3. The redesigned and novel primers were designed by Dr. Sean Hemmingsen.

To know the range of temperatures over which the primers amplify *S. pombe* and *S. cerevisiae* DNA templates, temperature gradient experiments were performed on the DNA extracts with all the different primer sets. The PCR was run for 40 cycles and at temperatures 42, 43.3, 45.5, 48.6, 53.3, 56.5, 58.7 and 60˚C. The PCR product was observed for positive or negative amplification by running the PCR products on 1% agarose gel at 90 V and observing under UV light. In previous studies cpn60 primers have been found to work well between temperatures range of 46˚C to 50˚C. The temperatures at which these primers were effective were tested by studying the behaviour of these primers at a range of temperatures around 46˚C to 50˚C,   so 42˚C to 60˚C was chosen. Results were interpreted on the basis of visibility of the resulting amplicon under UV light on agarose gel. This is an end-point assay where the results depend on whether or not expected PCR products are obtained. Therefore, this is not the most sensitive assay for knowing the effect of temperatures.

To confirm that the sequences amplified by these primers were the same as we had anticipated, PCR products from each template were agarose gel purified and ligated into pGEM T-easy vector (Kobs, 1997). Ligation mixtures were used to transform *E. coli* strain JM109 (Messing et al., 1981). 100 µl of transformed cells were plated on Luria broth agar plates (X-gal/Amp) and 4 colonies were picked up for each primer set and inoculated into 5 mL LB overnight/300 rpm. The resulting cultures were prepared using a Qiagen miniprep kit. The DNA so obtained was quantified using Quant iT dsDNA kit and sent for sequencing (3.4.2).

Table 1a: Sequences of oligonucleotide primers used for PCR amplification of regions of *cpn*60.

Primer	Function Table 1b	[1]Primer sequence (5`-3`)	*Reference
H279	A	GAIIIIGCIGGIGAYGGIACIACIAC	2
H1780	B	AAYGAIIIIGCIGAYGGIACIAC	3
H1782	C	ACGAGTGCGTAAYGAIIIIGCIGAYGGIACIAC	3
H1784	D	ACGCTCGACAAAYGAIIIIGCIGGIGAYGGIACIAC	3
H280	E	YKIYKITCICCRAAICCIGGIGCYTT	2
H1786	F	CCIAARATHACIAARGAYGGIGTIACIGTIGC	3
H1787	G	GCIATGGARIIIGTIGGIAARGARGGIGTIAT	3
H1788	H	ATIACICCYTCYTTICCIACIIIYTCCATIGC	3
H1789	I	GCIACICCICCIIIIARYTTIGCIARICKYTC	3
H1612	J	GAIIIIGCIGGYGACGGYACSACSAC	4
H1613	K	CGRCGRTCRCCGAAGCCSGGIGCCTT	4
H1781	L	AAYGAIIIIGCIGGYGACGGYACSAC	3
H1783	M	ACGAGTGCGTAAYGAIIIIGCIGGYGACGGYACSACSAC	3
H1785	N	ACGCTCGACAAAYGAIIIIGCIGGYGACGGYACSAC	3

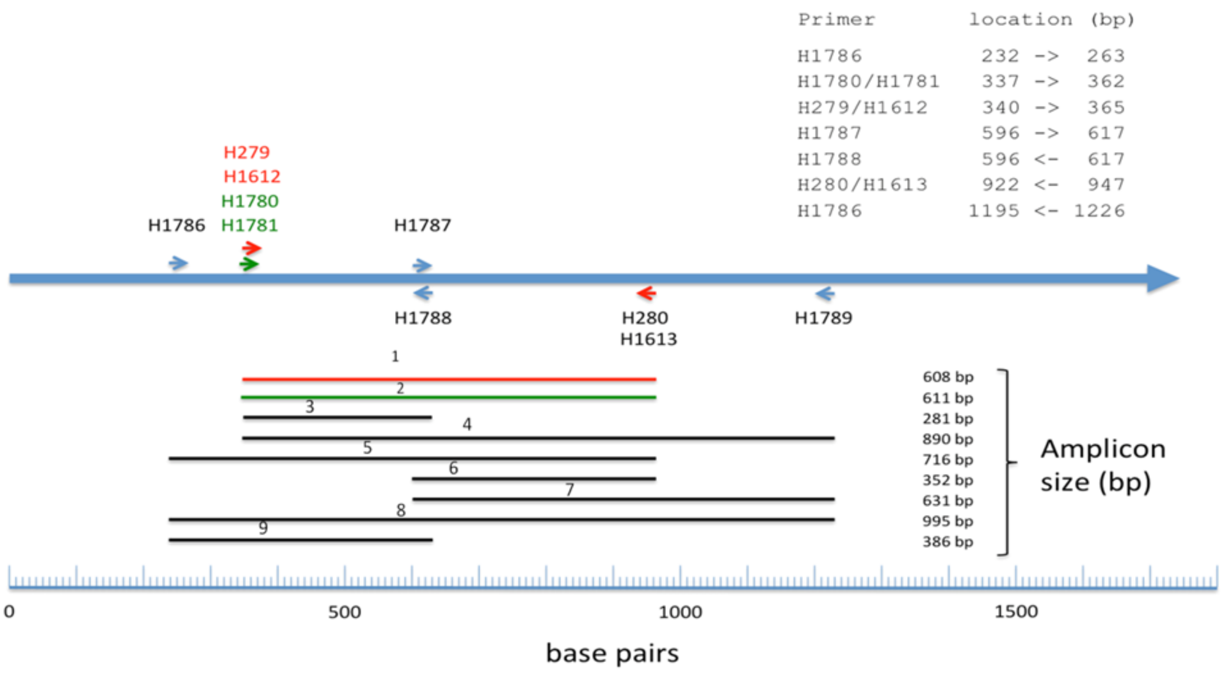
[1]Nucleotide code	Name of Base
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
R	A or G
Y	C or T
S	G or C
K	G or T
I	Inosine

*References

2	Goh et al., 1996
3	Hemmingsen, 2012 (unpublished)
4	Hill et al., 2006

Table 1b: Description of functions for *cpn*60 primers given in Table 1a.

Function	Description of function
A	Original forward UT primer. Amplification of region 1. (Fig.3)
B	Redesigned version of H279 to include more fungal taxa. Amplification of regions 2,3,4 (Fig.3)
C	H1780 with MID 1 for multiplex sequencing. Amplification of regions 2, 3, 4. (Fig.3)
D	H1780 with MID 2 for multiplex sequencing. Amplification of regions 2, 3, 4. (Fig.3)
E	Reverse UT primer. Amplification of regions 1,2,5,6. (Fig.3)
F	Fungal primer upstream to UT forward primer. Amplification of regions 5, 8, 9. (Fig.3)
G	Fungal forward primer between UT forward and UT reverse primer sites. Amplification of regions 6, 7. (Fig.3)
H	Fungal reverse primer between UT forward and UT reverse primer. Amplification of regions 3, 9. (Fig.3)
I	Fungal primer downstream of UT reverse primer. Amplification of regions 4, 7, 8. (Fig.3)
J	"Strong" version of H279 primer to enable amplification of GC rich templates. Amplification of region 1. (Fig.3)
K	"Strong" version of H280 primer to enable amplification of GC rich templates. Amplification of region 1,2,5,6. (Fig.3)
L	"Strong" version of H1780 UT primer to enable amplification of GC rich templates. Amplification of regions 2, 3, 4. (Fig.3)
M	"Strong" version of H1780 primer with MID1 to enable amplification of GC rich templates. Amplification of regions 2, 3, 4. (Fig.3)
N	"Strong" version of H1780 primer with MID2 to enable amplification of GC rich templates. Amplification of regions 2, 3, 4. (Fig.3)



Primer	location (bp)
H1786	232 -> 263
H1780/H1781	337 -> 362
H279/H1612	340 -> 365
H1787	596 -> 617
H1788	596 <- 617
H280/H1613	922 <- 947
H1786	1195 <- 1226

(Based on *Schizosaccharomyces pombe* sequence, cpnDB b1554)

Figure 3: Fungal *cpn*60 primer site positions and amplicon sizes.
Original *cpn*60 universal primer names (H279 and H1612) are shown in red and their position with a red arrow. Universal reverse primer (H280 and H1613) position is shown with a reverse red arrow. Redesigned *cpn*60 universal forward primer is shown in green and its position is shown shifted a little to left owing to its modification from original forward primer. Expected amplicon with original primers is shown with red line. Expected amplicon with redesigned universal primers is shown with green line. Expected amplicons produced by novel fungal specific primers are shown with black lines. The numbers (1-9) above the lines refer to the regions mentioned in Table 1b.

3.3.4 Protocol for Ligations Using the pGEM®-T Easy Vectors and Transformation using JM109 High Efficiency Competent Cells (Promega, Madison, Wisconsin)

Ligation Protocol and Transformation Protocol were followed as recommended in the technical manual of p-GEM T easy vector systems (www.promega.com/protocols/)

Six fungal DNA extract samples, sent by Dr. Andre Levesque from Agriculture and Agri Food Canada, Ottawa, which had already been amplified using the ITS4 and ITS5 primers in his lab, were amplified using our redesigned and novel primers. The amplicons were cloned and the DNA extracts were sent for sequencing and results were analyzed. Three fungal DNA extract samples, sent by Dr. Tim Dumonceaux, Agriculture and Agri Food Canada Saskatoon, were also amplified using our redesigned and novel primers. The amplicons were cloned and the DNA extracts were sent for sequencing and results were analyzed. Diversity of fungal genomes tested for amplification of *cpn*60 gene sequences is shown in a cladogram in Figure 4.

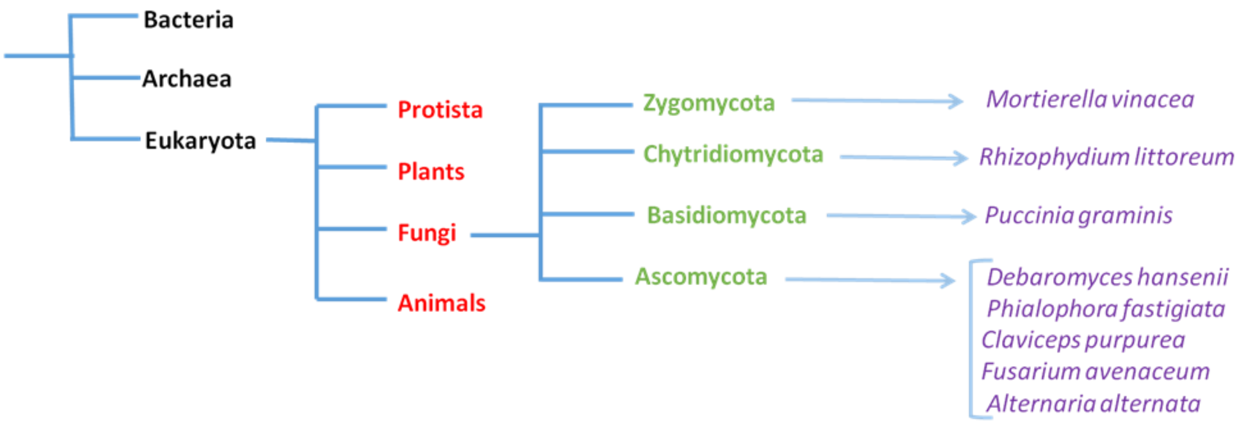


Figure 4: Diversity of fungal genomes tested for amplification of *cpn*60 gene sequences. The cladogram shows the three domains of life (in black), four groups of eukaryotes (in red) and four fungal phyla (in green). Species tested in this study are in purple.

3.3.5 Sequence Analysis

The DNA was sequenced in the NRC-PBI sequencing lab using Sanger sequencing. The sequences were analyzed using FASTA and BLASTP (after converting nucleotide to peptide sequence using Transeq in EMBOSS) in cpnDB and blastn and blastx in NCBI under default parameters.

3.4 Results

3.4.1 Amplification of *S. pombe* and *S. cerevisiae cpn*60 Gene Sequences Using Redesigned and Novel Primer Sets.

The redesigned UT primers were tested with respect to their phylogenetic reach, that is, for their ability to amplify phylogenetically divergent fungal *cpn*60 sequences. *S. pombe* and *S. cerevisiae* are yeasts with an ancient last common ancestor. The *S. pombe cpn*60 peptide sequence was in the original alignment that formed the basis for the design of the original primers. In 2012, cpnDB had 61 full length fungal *cpn*60 sequences. An *in silico* approach showed that in 33 of the 61 cases, the *cpn*60 UT sequences should be amplifiable by the original primers including that of *S. pombe*. In the remaining 28 cases, the C-terminal amino acid residue of the consensus sequence is serine (S) as compared to threonine (T) (Figure 2). Therefore, these 28 sequences would not be expected to amplify using the original primers. These included *C. albicans* and *S. cerevisiae* sequences. *Candida* species were of specific interest because of the focus on the human vaginal mycobiome. For reasons of convenience, *S. cerevisiae* was used as a stand in for *Candida albicans,* which is of more direct interest with respect to the human vaginal microbiome. Therefore, *S. pombe* and *S. cerevisiae* DNA extracts were used as genomic templates for the following experiments. Various primer combinations were used in PCR to amplify *cpn*60 gene regions for each template. Amplicons were cloned in an *E. coli* plasmid vector and sequenced. However, some sequences were not obtained because of the technical problems like failure of cloning or failing of ligation of amplicon in spite of repeated attempts; although, the templates produced a PCR product of expected size on gel.

Redesigned primers designed to amplify the *cpn*60 UT sequences (region 2, Figure 3) were used on both yeast templates. PCR products of the expected sizes were observed in both cases. The identity of the *S. pombe* and *S. cerevisiae* products was confirmed by cloning and sequencing. Amplification may have been less productive for the *S. pombe* template at higher temperatures. Annealing temperature appeared to have little effect for the *S. cerevisiae* template (Figure 5-A1,B1).

The novel primers designed to amplify regions of *cpn*60 from fungal templates were also tested with respect to their ability to amplify regions of the *cpn*60 gene from these two templates. One of these novel fungal primer pairs (H1780,H1781/H1788) was designed to specifically amplify a part of fungal *cpn*60 region (region 3, Figure 3) in both yeast templates. A PCR product of the expected size was the predominant band observed in each case. Amplicon production may have decreased with increase in annealing temperature for both yeast templates. Cloning and sequencing of amplicons for both yeasts produced vector sequences only. This was not pursued further (Figure 5-A2,B2).

The redesigned *cpn*60 UT primer and novel fungal reverse primer H1789 designed for the amplification of region 4 (Figure 3) produced two major PCR products with the *S. pombe* template. The larger product was confirmed to be correct by sequencing. With *S. cerevisiae* template, multiple bands were observed, one of these was of the expected size. Sequencing of selected clones produced only vector sequence. This was not pursued further. The amplification appeared to be more productive at lower temperatures in both cases (Figure 5-A3,B3).

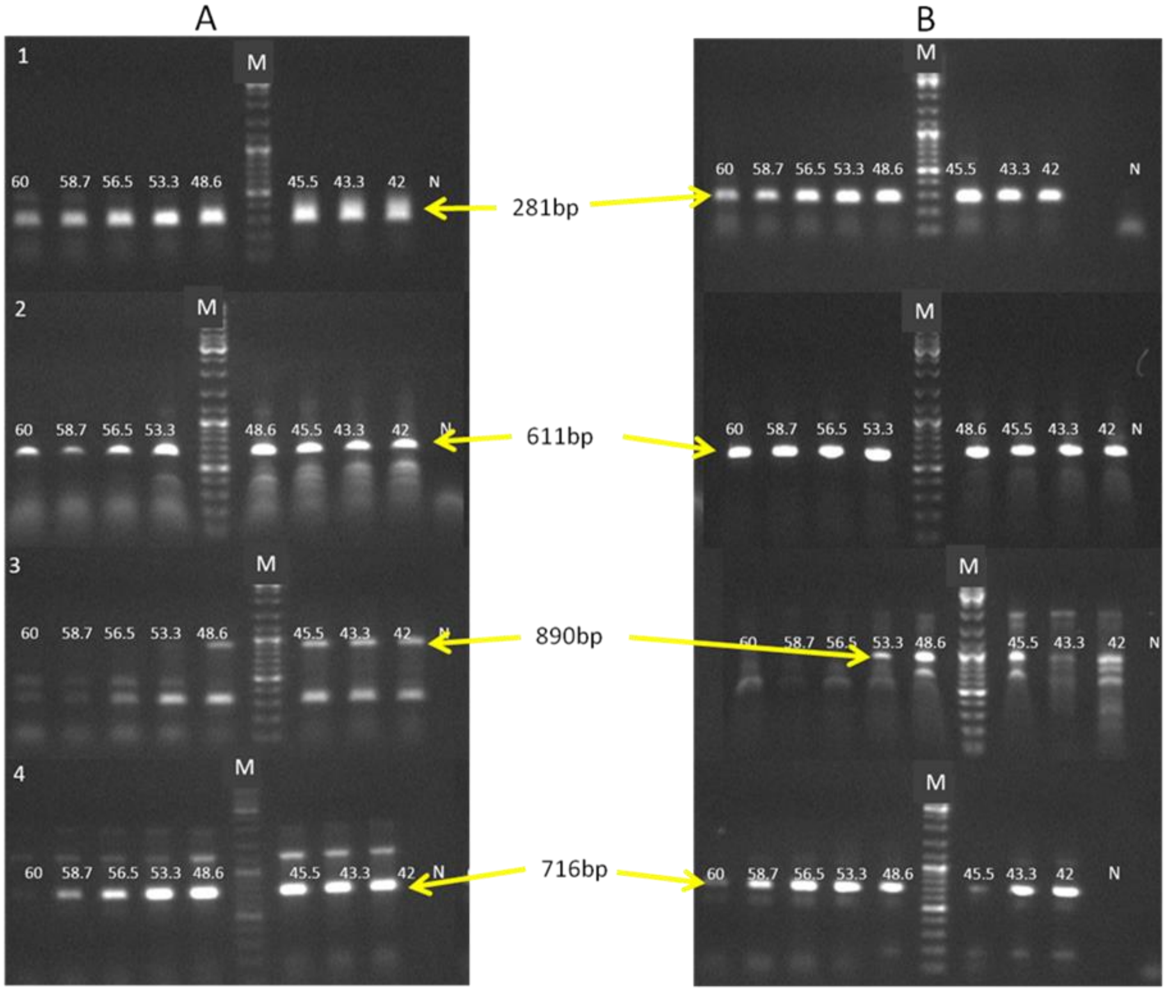
Region 5 of both the yeast templates was amplified with another novel fungal forward primer H1786 and reverse primers H280 and H1613. Amplification of region 5 produced two major PCR products with the *S. pombe* template. The smaller product was confirmed to be correct by sequencing. The corresponding *S. cerevisiae* sequencing produced vector sequence only. The annealing temperatures did not seem to influence amplicon productivity of both yeast templates except for the absence of any PCR product at highest temperature in the given gradient (60˚C). Absence of a PCR product at one of the temperatures for *S. cerevisiae* was attributed to the accidental loss of amplicon while loading the gel (Figure 5-A4,B4).

The novel fungal primer H1787 was used along with reverse primers H280 and H1613 on both the yeast templates to amplify region 6 (Figure 3). PCR products of expected size were observed for both *S. pombe* and *S. cerevisiae.* The identity of the *S. pombe* product was confirmed by cloning and sequencing. The corresponding *S. cerevisiae* sequencing produced vector sequence only. The amplification was observed to be more productive at lower annealing temperatures for both the yeast templates (Figure 5-A5,B5).

The novel fungal primer pair H1787/H1789 was used to amplify the *cpn*60 region 7 (Figure 3) in both the *S. pombe* and *S. cerevisiae* templates. The size of the PCR products observed were as expected for both the yeast templates. The sequence of the respective *cpn*60 region of the *S. pombe* template was confirmed by cloning and sequencing. The corresponding region of the *S. cerevisiae* amplicon generated only vector sequence. Amplification products were clearly visible at lower annealing temperatures whereas higher temperature gradients failed to produce any visible amplicon in either of the cases (Figure 5-A6,B6).

The novel fungal primer pair H1786/H1789 was tested to amplify the *cpn*60 region 8 (Figure 3) in both the *S. pombe* and *S. cerevisiae* templates. PCR products of expected size were observed for the *S. pombe* templates. PCR product for *S. cerevisiae* was of the expected size but the visible productivity of the amplicon was very low or negligible. Cloning and sequencing in either case produced only vector sequence. Amplification may have been less productive at higher annealing temperatures for *S. pombe* whereas for *S. cerevisiae* it was negligible at higher and very low at lower annealing temperatures (Figure 5-A7,B7).

The novel fungal primer pair H1786/H1788 was tested to amplify the *cpn*60 region 9 (Figure 3) in both the *S. pombe* and *S. cerevisiae* templates. PCR products of expected size were observed for both the yeast templates. Cloning and sequencing in both the cases produced only vector sequence. Amplification may have been less productive for both the templates at the higher annealing temperatures (Figure 5-A8,B8).


A
B
1
M
60 58.7 56.5 53.3 48.6 45.5 43.3 42 N
281bp
60 58.7 56.5 53.3 48.6 45.5 43.3 42 N
2
M
60 58.7 56.5 53.3 48.6 45.5 43.3 42 N
611bp
60 58.7 56.5 53.3 48.6 45.5 43.3 42 N
3
M
60 58.7 56.5 53.3 48.6 45.5 43.3 42 N
890bp
60 58.7 56.5 53.3 48.6 45.5 43.3 42 N
4
M
60 58.7 56.5 53.3 48.6 45.5 43.3 42 N
716bp
60 58.7 56.5 53.3 48.6 45.5 43.3 42 N

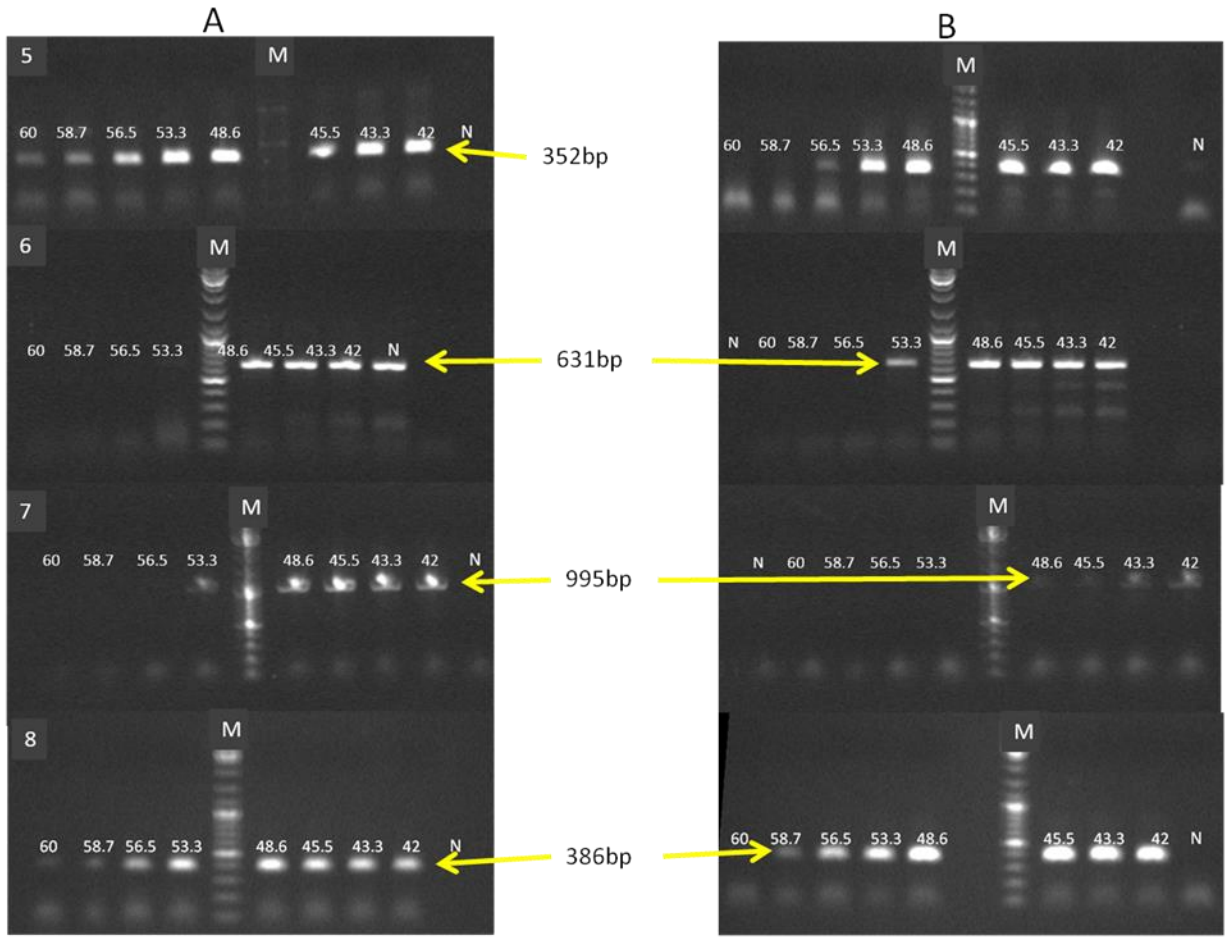

A
B
5
6
7
8
M
60 58.7 56.5 53.3 48.6 45.5 43.3 42 N
N 60 58.7 56.5 53.3 48.6 45.5 43.3 42
352bp
631bp
995bp
386bp

Figure 5: Electrophoretic analysis of *S. pombe* and *S. cerevisiae* PCR products. Ethidium Bromide fluorescence images showing electrophoresis of (A) *S. pombe* and (B) *S. cerevisiae* PCR products (See Figure 3) amplified with primers (1) H1780, H1781 and H1788 (2) H1780, H1781 and H280, H1613 (3) H1780, H1781 and H1789 (4) H1786 and H280 (5) H1787 and H280 (6) H1787 and H1789 (7) H1786 and H1789 (8) H1786 and H1788 at temperatures (60°C to 42°C). Gel was made with 1% agarose. (M) is the DNA molecular weight markers loaded at 100 ng. Ten µL of PCR reaction was loaded in each well for each of the templates. (N) is the no-template control. The number of base pairs indicated by yellow arrows indicate migration of correct PCR amplicon.

3.4.2 Utility of Redesigned *cpn*60 UT Primers and Novel Primers for Phylogenetically Diverse Fungal Taxa

Dr. A.Lèvesque provided us with DNA samples from a number of fungi from phylogenetic groups that were poorly represented in cpnDB. The identities of these fungi had been determined by phenotypic methods and confirmed by analysis of their ITS region sequences. The Genbank accession numbers for these fungi based on the ITS region sequences sent by Dr.A.Lèvesque were *D. hansenii* (KP132002.1), *M. vinacea* (EF434083.1), *P. fastigiata* (FM999988.1), *P .graminis* (DQ417378.1) and *R. littoreum* (DQ485604.1).

The redesigned UT primers and combinations of novel *cpn*60 primers and redesigned UT primers were used to amplify *cpn*60 sequences from these templates. Amplicons were cloned and clones were subjected to Sanger sequencing. The experimental sequences were compared to known fungal *cpn*60 sequences to determine if they were derived from the fungi identified by Dr. Levesque or if they were derived from contaminating templates.

Samples studied are listed in Table 2. The redesigned UT primers were used with each DNA template. For four of the seven templates, amplicons of the expected size were generated. These were for *F. avenaceum* (ascomycota), *D. hansenii* (ascomycota), *P. graminis* (basidiomycota), and *R. littoreum* (chytridiomycota). In the cases of the first three templates, sequence analysis of the cloned amplicons confirmed their expected identities (Table 2). In the case of *R. littoreum*, determination that the amplified product represented the target fungus rather than a contaminant was not as obvious. A putative intron was found in the experimental sequence that is discussed below. If this putative intron sequence is removed and resulting sequence is compared to known *cpn*60 fungal sequences, best hits were to *N. patriciarum* and *B. dendrobatidis* (73% and 72% respectively). These numbers are low, however, both of these fungi belong to Chytridiomycota suggesting that the experimental sequence is from the *R. littoreum* genome. Furthermore, the amino acid sequence identity was 81% to *B.dendrobatidis* that further supported the analysis. Fungal primers H1787/1789 were also used with each of the DNA templates. For five of these seven DNA templates, amplicons of the expected size were generated. These were for *A. alternata* (Ascomycota), *C. purpurea* (Ascomycota), *D. hansenii* (Ascomycota), *F. avenaceum* (Ascomycota) and *M. vinacea* (Zygomycota). Except for *F. avenaceum*, sequence analysis of the other four cloned amplicons confirmed their expected identities. Sequence analysis of *Fusarium* sample generated vector sequence only. Fungal primers H1786/H1788 were tried on each of seven DNA templates and amplicons of expected size were generated for three of these seven DNA samples, *C. purpurea, D. hansenii and Phialophora fastigiata*. Sequence analysis confirmed the expected identities of *C. purpurea*, *D. hansenii* DNA samples. The *C. purpurea* sample was amplified by two sets of primers H1786/H1788 and H1787/1789. When the sequences amplified by both the primer sets were put together, there was a 32 bp gap formed where primer landing sites for 1787 and 1788 primers overlapped (Figure 3). Reference sequence for *P. fastigiata* is present neither in cpnDB nor in NCBI database. The closest nucleotide hit for its experimental nucleotide sequence was *Marssonina brunnea* (NCBI:XM_007293454.1), also an ascomycete and the identity was 87%.

Table 2: PCR amplification of the *cpn*60 UT or portions of the *cpn*60 gene from DNA extracts from phylogenetically diverse fungal taxa not represented in cpnDB using redesigned or novel *cpn*60 primers.

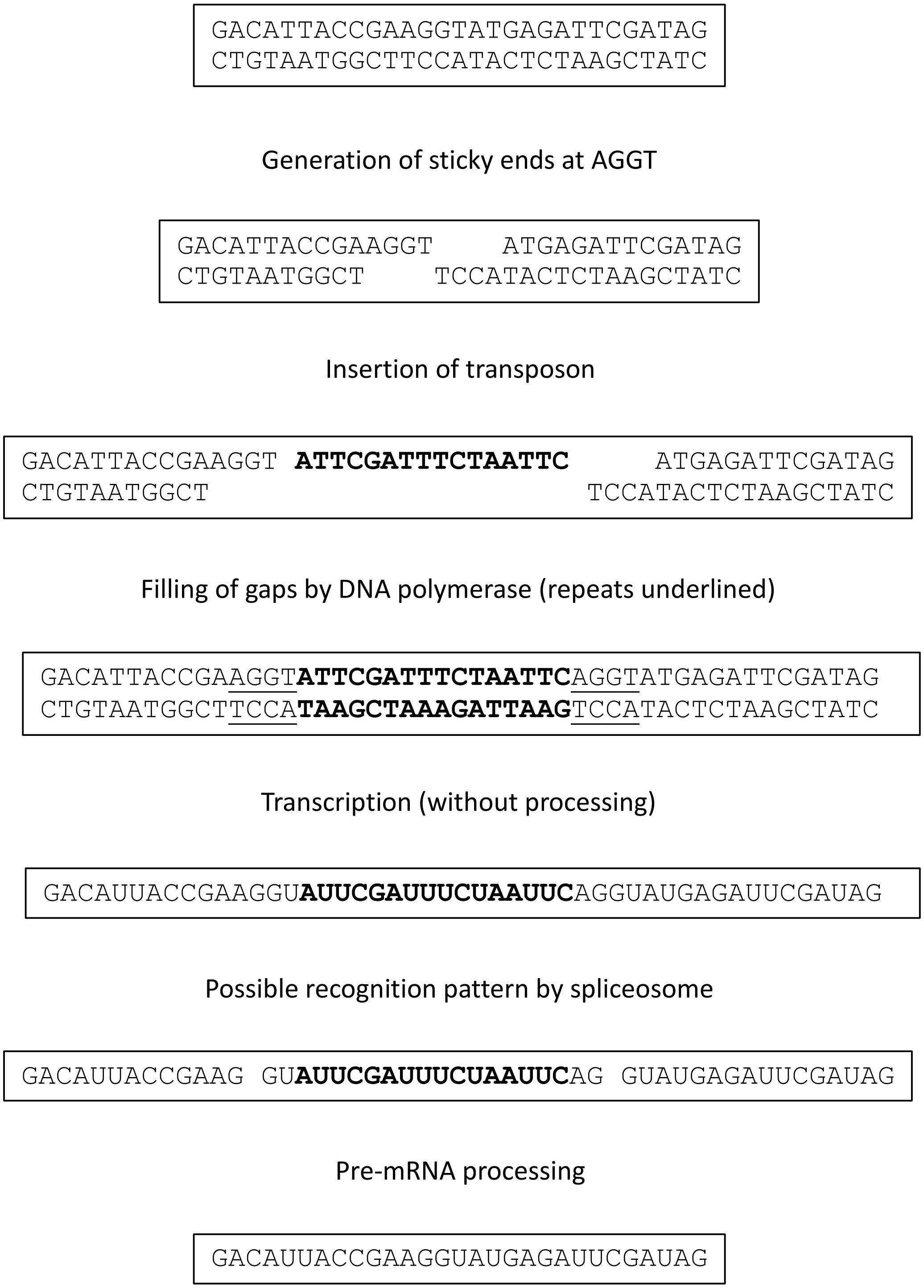
Sample ID	Taxonomy	*cpn*60 regions sequenced (Fig.3)	Comparison of experimental sequences to reference sequences
Alternaria alternate	Ascomycota *A. alternata*	7	Experimental nucleotide sequence was 99% identical to *A. alternata* (NCBI:EU285274.1).
Ergot	Ascomycota *Claviceps purpurea*	7, 9	Experimental nucleotide sequence was 95% identical to *C. purpurea* reference nucleotide sequence (NCBI: XM_003716778.1) *and* translated experimental sequence was 96% similar to *C. purpurea* (NCBI:CCE28256.1).
Fusarium	Ascomycota *F. avenaceum*	2	Experimental nucleotide sequence was 97.7% identical to *Gibberella avenaceum* (cpnDB :b7306), an ascomycete. *F. avenaceum* is an anamorphic form of *G. avenaceum* (Cook, 1967).
KS-81	Ascomycota *Phialophora fastigiata*	9	Experimental nucleotide sequence was 87% identical to *Marssonina brunnea* (NCBI:XM_007293454.1) that is an ascomycete. The translated experimental sequence was 100% identical to *M. brunnea* (NCBI:XP_007293516.1) an ascomycete. cpnDB and NCBI do not have *P. fastigiata* reference sequence (as of 2012).
KS-45	Ascomycota *Debaryomyces hansenii*	2, 7, 9	Experimental nucleotide sequence was 100% identical to *D. hansenii* (cpnDB:b5730).
RSA1924 B-3B	Basidiomycota *Puccinia Graminis*	2	Experimental nucleotide sequence was 99% identical to *P. graminis* (NCBI: XM_003334157.2).
LEV5712	Chytridiomycota *Rhizophydium littoreum*	2	Experimental nucleotide sequence was 73% identical to *Neocallimastix patriciarum* (cpnDB: b4156) and 72% identical to *Batrachochytrium dendrobatidis* (NCBI: XM_006682509.1) and translated sequence was 81% identical to *B. dendrobatidis* (NCBI: XP_006682572.1), a chytridiomycete. cpnDB and NCBI do not have *R. littoreum cpn*60 sequence (as of 2012).
LEV1641	Zygomycota *Mortierella vinacea*	7	The experimental nucleotide sequence was 88% identical to *Mucor circinelloides* (NCBI:KE124010.1) and translated sequence was 88% identical to *M. circinelloides* (NCBI:EPB85507.1), a zygomycete. cpnDB does not include *M. vinacea* sequence (as of 2012).

3.4.3 Putative Intron in *cpn*60 Sequence of *Rhizophydium littoreum* (LEV5712).

An alignment of the experimental *R. littoreum* UT nucleotide sequence and its best hit match *N. patriciarum* was produced (Figure 6). The *R. littoreum* sequence appeared to include a 20 base internal addition relative to the reference sequence *N. patriciarum* which would produce a frame-shift in the experimental sequence. The alignment in Figure 7 was adjusted to maximize the amino acid sequence similarity between the experimental sequence and reference sequence. The 20 base addition is discussed further in Figure 7 and in 3.6.

```
A  T  V  L  T  R  A  I  F  T  E  G  L  K  N  V  S  A  G  V

gctactgtcttgactcgtgctatctttaccgaaggtttaaagaacgtctctgccggtgtc  60
||||||||  |  |  | || || ||  | |||||||||||||| |||||||| |||||
gctactgttcttgccagagccattttcgctgaaggtttaaagaatgtctctgctggtgtt  60
A  T  V  L  A  R  A  I  F  A  E  G  L  K  N  V  S  A  G  V

N  P  N  D  L  R  R  G  V  Q  Q  A  V  E  L  V  V  A  Y  L
aacccaaatgacttgagacgcggtgttcaacaagcggtagaactcgttgttgcctactta 120
||||||  |||  | ||| | ||||||||| | || || ||  | |||||||  | | |
aacccagttgaacttagaagaggtgttcaaaaggctgttgatgttgttgttgatttcctt 120
N  P  V  E  L  R  R  G  V  Q  K  A  V  D  V  V  V  D  F  L

K  A  N  A  Q  P  I  T  T  S  Q  E  I  A  Q  V  A  T  I  S
aaggcaaatgctcaaccaatcactaccagtcaagaaattgctcaagttgccaccatctct 180
|| | | | ||||| ||||| | |||   | |||||||||||||||| |  ||||| |||
aaagaacaagctcatccaattagtacttttgaagaaattgctcaagtcggtaccatttct 180
K  E  Q  A  H  P  I  S  T  F  E  E  I  A  Q  V  G  T  I  S

A  N  G  D  K  H  V  G  E  M  I  A  K  A  M  D  K  V  G  K
gccaacggtgacaagcatgtcggtgaaatgattgcaaaggccatggacaaggttggcaaa 240
|| || ||||| |||||| | ||||   |  | ||  | |||||| | |||||||| ||
gctaatggtgataagcatattggtggtcttttagctgaagccatgaaaaaggttggtaag 240
A  N  G  D  K  H  I  G  G  L  L  A  E  A  M  K  K  V  G  K

E  G  V  I  T  C  Q  E  G  K  T  L  V  D  E  L  D  I  T  E
gaaggtgtcattacctgccaagaaggaaagactcttgttgatgaattggacattaccgaA 300
|| ||||| |||| |   || ||||| || |||||||  ||||||||   |||||| ||
gatggtgttattaacattcatgaaggtaaaactcttgaagatgaattaaccattactga. 299
D  G  V  I  N  I  H  E  G  K  T  L  E  D  E  L  T  I  T  E

                    G  M  R  F  D  R  G  F  I  S  P  Y  F  M
GGTATTCGATTTCTAATTCaggtatgagattcgatagaggtttcatttctccatacttta 360
                   |||||||| ||||||||  |||||| | |||| | |||| |
...................aggtatgaaattcgataacggtttcttatctccacacttca 340
                    G  M  K  F  D  N  G  F  L  S  P  H  F  I

  T  N  N  K  S  Q  K  V  E  F  E  K  P  L  V  L  L  S  E  G
tgaccaacaacaagtcccaaaaggttgaatttgaaaagcctttggttttgctttccgagg 420
| ||  | || |||    | ||   |||| | ||||| || |   ||||  || |||| |
ttactgataataagggtaagaaatgtgaactcgaaaatccatacattttaattaccgaag 400
  T  D  N  K  G  K  K  C  E  L  E  N  P  Y  I  L  I  T  E  E

  K  I  S  Q  L  Q  D  L  L  P  A  M  E  I  A  A  Q  S  R  R
gaaagatctctcaattgcaagatttgcttcctgccatggaaattgctgctcaatcccgtc 480
 ||| || |||    | |||||| |  |||| ||  | |||||||||||| |   ||||
aaaaaatttctgctgttcaagatattgttccagctttagaaattgctgctaacaaccgta 460
  K  I  S  A  V  Q  D  I  V  P  A  L  E  I  A  A  N  N  R  R

  P  L  L  I  I  A  E  D  V  D  G  E  A  L  A  A  C  I  L  N
gtccattgttgattattgctgaagatgttgatggtgaagctttggctgcttgtatcctca 540
| ||| | || |||||||||||| |||||||| ||||| ||||| ||| ||||| | || |
gaccacttttaattattgctgatgatgttgaaggtgatgctttagctacttgtgttctta 520
  P  L  L  I  I  A  D  D  V  E  G  D  A  L  A  T  C  V  L  N

  K  L  R  G  Q  L  Q  V  A  C  V
acaagcttagaggacaattgcaagtcgcttgtgta 575
||||| || | || ||| | |||||   |||  |
acaagattcgtggtcaagtccaagtttgttgtatt 555
  K  I  R  G  Q  V  Q  V  C  C  I
```

Figure 6: Aligned *R. littoreum* (upper sequence) and *N. patriciarum* (lower sequence) *cpn*60 UT sequence.



GACATTACCGAAGGTATGAGATTCGATAG

CTGTAATGGCTTCCATACTCTAAGCTATC

Generation of sticky ends at AGGT

GACATTACCGAAGGT   ATGAGATTCGATAG

CTGTAATGGCT   TCCATACTCTAAGCTATC

Insertion of transposon

GACATTACCGAAGGT **ATTCGATTTCTAATTC**   ATGAGATTCGATAG

CTGTAATGGCT   TCCATACTCTAAGCTATC

Filling of gaps by DNA polymerase (repeats underlined)

GACATTACCGAAGGT**ATTCGATTTCTAATTC**AGGTATGAGATTCGATAG

CTGTAATGGCTTCCA**TAAGCTAAAGATTAAG**TCCATACTCTAAGCTATC

Transcription (without processing)

GACAUUACCGAAGGU**AUUCGAUUUCUAAUUC**AGGUAUGAGAUUCGAUAG

Possible recognition pattern by spliceosome

GACAUUACCGAAG GU**AUUCGAUUUCUAAUUC**AG GUAUGAGAUUCGAUAG

Pre-mRNA processing

GACAUUACCGAAGGUAUGAGAUUCGAUAG

Figure 7: Proposed origin of putative intron in *R. littoreum cpn*60 UT sequence. The figure is superimposed on the original figure by Yenerall and Zhou, 2012, showing one of the modes of intron gain called transposon insertion.

3.5 Discussion

The universality of the *cpn*60 gene, its demonstrated utility to differentiate closely related species and subspecies because of the sufficient sequence difference present in *cpn*60 UT from closely related species (Hill et al., 2004) and its recent evaluation as a DNA barcode for bacteria (Links et al., 2012), led us to investigate whether the redesigned and novel *cpn*60 primers amplify *cpn*60 gene from a broad range of fungal taxa.

The redesigned *cpn*60 universal primers amplified UT sequence both from *S. pombe* and *S. cerevisiae* as expected and DNA sequence analysis evidence was available for both the templates. With novel primers, based on amplicon sizes, all appeared to have worked on both the templates. DNA sequence evidence was available for some templates. Multiple bands were observed in few cases, out of these, DNA sequence analysis evidence was obtained for *S. pombe* templates.

For taxa not represented in cpnDB, little reference data was available for comparison to experimental data. The experimental sequences were probably of the same taxonomic group as identified in Dr.Lèvesque`s lab unless and until there was another fungal contaminant in these samples. Low level contamination by commensal fungi or environmental spores can be a problem when using universal primers. Previously, in the Hemmingsen lab, DNA extracts prepared from spores of Arbuscular Mycorrhizal (AM) Fungi were analyzed. In some cases, *cpn*60 sequences were amplified from these extracts that were consistent with a fungal source but not consistent with an AM fungal source. In these cases the most abundant template in the extract (AM fungal genomic DNA) failed to produce an amplicon while a template representing a minor contaminant did. That means, great caution must be taken while analyzing sequence data to avoid false positive results. Some cases where reference *cpn*60 sequence was not available in databases, identification of fungal taxa will get more specific as more of fungal sequences are deposited in cpnDB or NCBI database.

The *cpn*60 universal target sequence was amplified from *F. avenaceum* (ascomycota), *D. hansenii* (ascomycota), *P. graminis* (basidiomycota), and *R. littoreum* (chytridiomycota). Three other ascomycetes and a zygomycete did not amplify with UT primers. Other *cpn*60 gene parts were also amplified from different samples using novel *cpn*60 primers that were designed specifically for fungal *cpn*60, although some regions were not amplified in these different samples. Novel fungal primers H1787/1789 generated DNA sequences for *A. alternata* (Ascomycota), *C. purpurea* (Ascomycota), *D. hansenii* (Ascomycota), and *M. vinacea* (Zygomycota), except for *F. avenaceum* (Ascomycota). The *cpn*60 UT region seemed to be more often amplified (in seven out of eight samples) than other regions (three out of eight samples on 5` end and four out of eight samples on 3` end). Sequence analysis identified *C. purpurea* and *D. hansenii* as expected when novel fungal primers H1786/H1788 were used. With same primers, *P. fastigiata,* an ascomycete, was identified to be 87% identical to another ascomycete in the absence of any reference sequence.

In cases where expected sequences were obtained upon analysis of experimental sequence, it was possible to obtain the exact sequence of the degenerate primers. As an example, if expected sequence is generated by primers H1787/H1789 for DNA sample of *A. alternata,* the exact sequence of primer H1780 can be known from it and can be used to make more specific primers for *A. alternata* and this can be helpful to know another part of its *cpn*60 sequence. This study produced substantial evidence that redesigned *cpn*60 UT primers and novel primers specific for fungi have utility for detecting and identifying fungal taxa from phylogenetically diverse fungi. Therefore, the *cpn60* UT can be useful for the detection of both bacterial and fungal taxons unlike other gene targets for detection of micro-organisms that can detect either fungi or bacteria.

The chytrid sample, LEV5712 was identified as *Rhizophydium littoreum* by ITS. Chytrids belong to phylum Chytridiomycota. They are characterized by the formation of zoospores and a posterior flagellum at some stage of their lifecycle. They are mostly parasites on marine algae, other chytrids and invertebrates. The interest in chytrids was heightened in 1998, when a vertebrate parasite *Batrachochytrium dendrobatidis* (Bd) was discovered which was devastating populations of amphibians (Longcore et al., 1999). There were only three chytrid sequences in cpnDB (2 *Piromyces* and one *Neocallismatix*) and although there were 6 *Rhizophydium* sequences in Genbank, none were *cpn*60. Redesigned *cpn*60 UT primers were able to amplify the UT part of the *Rhizophydium* DNA and the best nucleotide hit was *Neocallimastix patriciarum* (74%) and the best peptide hit was Bd (81%) in NCBI. The sequence on analysis showed a 20 bp insertion as compared to the cpnDB entries (Figure 6). Since the number of inserted nucleotides is not a multiple of three, this insertion should cause a frameshift mutation in the gene and render it non-functional in which case it may be a pseudogene. But if it is not making the gene non-functional, the insertion may be an intron occurring as a result of transposon insertion, a type of intron gain (Figure 7). In this type of intron creation, a transposon sequence inserts itself into sequence AGGT which is believed to be the preferential site for intron gain and the coding sequence of the gene is not altered (Yenerall and Zhou, 2012). Whether this insertion is an intron can be demonstrated using a simple experiment of amplifying the cDNA, obtained by reverse transcription of mRNA extracted from chytrid sample, with *cpn*60 primers. On further cloning and sequencing, the insertion believed to be an intron will no longer be present in the final sequence.

The presence of a putative intron in the *Rhizophydium* sequence was interesting because intron was short i.e. 20 bp. Minimum length of introns in two of the most studied fungi, *S. cerevisiae* is 52 bp (Spingola et al., 1999) and in *S. pombe* is 35 bp. The intron size distribution in fungi is biased towards shorter introns with 33% of the introns being shorter than 100bp (41 to 60 nt). The intron observed in the *Rhizophydium* sample was 20 bp long. In one of the studies on *Rhizophydium* tubulin genes, 4 introns were found with lengths of 22, 23, 25 and 37 bp. From these findings, it suggests that *Rhizophydium* chytrid seems to have very short introns. This is the first time that *cpn*60 gene has been systematically studied as a target for fungal identification; therefore, not much literature is available for the same. According to standard protocol followed for amplification of bacterial templates, amplicons obtained from PCR are run on ethidium bromide gel and bands of appropriate size are cut. In case of eukaryotic templates, the size of required PCR products may not appear to be of appropriate size on gel due to the presence of insertions in them, as a result, the product can be discarded although it is the right PCR product. Therefore, the standard protocols should be redesigned for the appropriate detection of eukaryotic PCR products on gel. Chytrid amplicon in our lab had a 20 base internal addition that was very short and the difference between size of PCR product with and without this addition was not very visible on gel, and therefore the PCR product was extracted expecting it to be of the expected *cpn*60 size.

3.6 Conclusions

The redesigned *cpn*60 UT fungal primers amplified *cpn60* UT from fungal DNA extracts of both *S. pombe* and *S. cerevisiae* that represent the sequences present in cpnDB as confirmed by sequencing. Based on the observation of PCR products of expected sizes in all cases and confirmation by sequencing in few cases, we have reasonable evidence to show that the novel fungal primers specific for fungi, amplified most parts of the *cpn*60 gene from fungal DNA extracts of both *S. pombe* and *S. cerevisiae*. The redesigned *cpn60* UT and novel fungal primers also amplified many parts of the *cpn*60 gene including *cpn60* UT from fungal DNA extracts from fungal species that are not represented in cpnDB (diverse phylogeny in the fungal kingdom). Although more number and wider diversity of fungal samples could have been tested using the redesigned and novel cpn60 primers in this study, it is worthwhile to say that these primers can be useful for the amplification of *cpn*60 from a wide if not all diversity of fungi. Therefore, here we show that the *cpn60* UT region is useful for the detection of both bacterial and fungal sequences. This also gives it the advantage over 16S rRNA encoding gene that can be used just for bacterial detection or 18S rRNA encoding gene that can be used just for the fungal detection. The cpnDB is a sparsely populated database with regard to fungal kingdom, we hope to expand it using the redesigned and novel fungal *cpn*60 primer sets. The results from temperature studies for redesigned primers and novel primers may not provide very useful information for future studies. The amplifications may have been carried out by choosing one or two temperatures already used successfully in many previous studies to save time, resources and labour. As in cases where no amplification was obtained, other factors like presence of secondary structures in templates, number of PCR cycles, Mg^{2+} concentration may have been involved.

4.0 Utility of Redesigned *cpn*60 UT PCR primers for Microbiome Profiling

4.1 Hypotheses and Experimental Approach

To be useful, for a given microbial community comprised of both fungal and bacterial taxa, the redesigned *cpn*60 UT primers should produce similar bacterial profiles, and similar or improved fungal profiles as compared to the original *cpn*60 UT primers. To test these hypotheses, bacterial and fungal profiles were generated using the original and redesigned *cpn*60 UT primers and two independent vaginal metagenomic DNA templates. Two DNA templates with distinct community structures were produced by pooling aliquots of selected DNA samples obtained from 100 individual women.

4.2 Objectives

To test the efficacy of redesigned *cpn60* UT primers on vaginal microbiome, DNA templates from healthy and unhealthy women as representative complex microbial communities with different profiles and compare the profile so obtained on same templates with original *cpn60* UT primers.

4.3 Material and Methods

4.3.1 Vaginal Sample Pools (metagenomic DNA templates)

The current study was a part of a larger ongoing study of the vaginal microbiome (Vaginal Microbiome Group initiative). In the larger study, vaginal samples were collected from 100 women who were classified as either HIV negative or HIV positive. It should be noted that as a result of treatment with retroviral drugs, the woman in the latter category were largely healthy. DNA was extracted from these samples using Magmax^{TM} Total Nucleic Acid Isolation Kit (http://tools.lifetechnologies.com/content/sfs/manuals/cms_055603.pdf). DNA extraction was done by Dr. Bonnie Chaban in the laboratory of Dr. Janet Hill, University of Saskatchewan. This method included a bead beating step to shear/tear open cells. Supposedly it should have extracted DNA from all types of cells including fungal cells if present in the vaginal samples. 300 µL of sterile 1X Phosphate Buffered Saline (PBS) Buffer (pH 7.4) was initially added to the vaginal swab in the dry swab container. The swab was vortexed for 30 seconds and 200 µL of the sample solution was removed from the swab container and placed into a 1.5 mL tube. At this step, the original swab and swab container were discarded. 235 µL of MagMAX Lysis/Binding Solution Concentrate was added to a prepared tube of zirconia beads in a guanidinium thiocyanate-based solution. 175 µL of the sample solution was then transferred from the 1.5 mL tube and added to the prepared tube of zirconia beads. This tube was then vortexed for 15 minutes and then centrifuged for 3 minutes at 16,000 x *g* using the Eppendorf Centrifuge 5430. This procedure allowed the zirconia beads to mechanically disrupt the cells, releasing nucleic acid content. Guanidinium thiocyanate was present to inactivate the nucleases present in sample solution. 115 µL of this sample supernatant was transferred to a well of the processing plate (96-well plate). 65 µL of 100% isopropanol was added to each sample in processing plate and the plate was shaken for 1 min on the orbital multi-well plate shaker. 20 µL of freshly vortexed bead mix was added to the sample. It was shaken for 5 min so that nucleic acid could bind to the nucleic acid binding beads in the bead mix. The plate was then moved to the magnetic stand and left there for 5 min. When the beads formed a pellet in the magnetic stand, the supernatant was aspirated and discarded without disturbing the bead pellet and the processing plate was removed from the magnetic stand. 150 µL of Washing solution 1 (12mL 100% isopropanol added to bottle labelled Washing solution 1) was added and the plate was shaken until mixture was clear (~1 min). Supernatant was again aspirated and discarded without disturbing the beads. Washing with washing solution 1 was repeated. The next washing was done with 150 µL Washing Solution 2 (32mL 100% ethanol added to bottle labelled Washing Solution 2) twice in the same way. The beads were dried by shaking the plates until all the alcohol had evaporated. Elution buffer was brought to 65°C and ~30 µL was added to sample and shaken vigorously for ~3min, so that beads are evenly suspended in solution. The beads were captured by placing the plate on magnetic stand. The supernatant containing DNA was transferred to a nuclease free container. Two DNA pools were created from these extracted samples, an HIV negative pool (V1A) and HIV positive pool (V1B) with 10 µL aliquots from 12 women each (personal communication Dr.Bonnie Chaban). The resulting 2 pools had distinct metagenomic profiles. The prepared pools were stored at -80°C until further use.

4.3.2 Amplicon Libraries for Next-Generation Sequencing

The PCR amplification was done with *cpn*60 MID-tagged UT primers. A MID (Multiplex IDentification) tag is a member of a set of unique 10 bp sequences that is added to primer sets to be used in the amplification of DNA templates. The MID allows for the differentiation of unique samples in future processing steps. The *cpn*60 UT was amplified from the pools using 5'MID-tagged *cpn*60 UT original and redesigned primers on each of DNA template V1A and V1B resulting in four libraries. PCR was done on an Eppendorf Mastercycler EP gradient thermocycler. For PCR amplification, a separate master mix solution was created for each of the subsequent four libraries that will be made as a result of PCR amplification. This master mix consisted of: 477.4 µL of Ultrapure Water, 70 µL of 10 x PCR Buffer (In vitrogen), 35 µL of 50 mM $MgCl_2$ (In vitrogen), 14 µL of 10 mM dNTP, and 5.6 µL of 5 U/ µL Platinum Taq (In vitrogen) so that the final concentrations of reagents in the master mix were PCR buffer 1X, $MgCl_2$ 2.5 mM, dNTPs 200 µM and Platinum Taq 2.5 U/reaction and final volume of the master mix was 602 µL. The MID-tagged primer stocks were made as follows: For V1A, original primers with MID20 i.e. 3 µL of 100 mM H279, 3 µL of 100 mM H280, 9 µL of 100 mM H1612, 9 µL of 100 mM H1613 and 276 µL of Ultrapure Water were mixed and 70 µL of this MID-primer mix added to master mix later. For V1B, original primers with MID4, i.e. 3 µL of 100 mM H279, 3 µL of 100 mM H280, 9 µL of 100 mM H1612, 9 µL of 100 mM H1613 and 276 µL of Ultrapure Water were mixed and 70 µL of this MID-primer mix added to master mix later. For V1A, redesigned primers with MID1, i.e. 3 µL of 100 mM H1782, 3 µL of 100 mM H280, 9 µL of 100 mM H1783, 9 µL of 100 mM H1613 and 276 µL of Ultrapure Water were mixed and 70 µL of this MID-primer mix added to master mix later. For V1B, redesigned primers with MID2, i.e. 3 µL of 100 mM H1784, 3 µL of 100 mM H280, 9 µL of 100 mM H1785, 9 µL of 100 mM H1613 and 276 µL of Ultrapure Water were mixed and 70 µL of this MID-primer mix added to master mix later. The sequence and function of these primers have been explained in Table 1a. The primers H279 and H280 fail to amplify GC rich templates such as *Bifidobacteria* from complex mixture of templates. The reason for the inclusion of the primer set of H1612 and H1613 was that this primer set has proven to improve the representation of templates with high GC contents when used with previously developed degenerate *cpn*60 primers (Hill et al., 2006). A No Template Control tube or “NTC,” was also set up to test for any potential contamination that may occur in course of study protocol as well as to ensure that reagents were free of contamination. The master mix solution, the primer working stock solution and the sterile PCR tubes were then placed under an ultraviolet (UV) light in a “Cleanspot” UV cabinet for 10 minutes to allow for the inactivation of any DNA products through the formation of thymine dimers (Schreier et al., 2007). 70 µL of the MID-primer mix was then added to the tube of master mix solution. 48 µL of this complete master mix solution was then added into the

NTC tube. 24 µL of the vortexed sample solution was then added to this complete master mix solution. 50 µL of this mixed solution was aliquoted into each of the 12 PCR tubes. These 12 PCR tubes were then added to the top row of the Eppendorf Mastercycler EP Gradient Thermal Cycler over temperature gradient. Annealing temperatures were 41.9, 42.3, 43.4, 45.1, 47.2, 49.6, 52.0, 54.4, 56.5, 58.3, 59.5 and 60.1°C. The NTC tube was added to column 12 of the second row of the thermal cycler. PCR conditions were: 95°C – 5min, 40 cycles (95°C – 30s, 41.9-60.1°C Gradient – 30s, 72°C – 30s), 72°C – 2min, 10°C – hold. After the PCR program was complete, the amplified *cpn60* target samples in all 12 of the PCR tubes were pooled together into a single microfuge tube. The names of the resulting four libraries were: V1A with original primers, V1A with redesigned primers, V1B with original primers and V1B with redesigned primers. In order to check for any contamination that may have occurred during following the protocol, the pooled PCR samples were run on 1% agarose gel along with NTC. 1 µL of ethidium bromide was added to the gel wells for its DNA visualization under UV light. 5 µL PCR sample and NTC was mixed with 2 µL DNA electrophoresis sample buffer and run on the gel. A DNA ladder was also added to one of the wells to indicate the size of the DNA in the sample. The gel was run at 100 volts for ~35 min. An image of the exposed gel was captured with Alpha Innotech AlphaImager instrument. The NTC lanes were blank with no visible bands showing absence of any contamination. The samples were ready for concentration and purification.

The amplified samples were concentrated using Amicon Ultra 0.5 Centrifugal Filters Units with Ultracel-30 membranes. This concentrated PCR product was then purified by gel purification by using a rainbow tracking dye. Rainbow tracking dye composition was: 0.5 mL of 0.5 M EDTA (pH 8.0), 12 g Sucrose, 0.06 g Bromophenol Blue, 0.07 g Xylene Cyanol FF, 0.06

g Cresol Red, 0.11 g Orange G, and Ultrapure Water to a total volume of 25 mL. 5 µL of this Rainbow tracking dye was added to ~30 µL of each amplified sample. The gel was run at 100 V for ~30min. To get the relevant *cpn*60 amplicons, the entire red band (has the 600-900 bp region) and the top part of the purple band (has the 300-600 bp region) were cut out. Each gel fragment was then purified using Qiagen`s Q1AEX II gel extraction kit (catalog no. 20021). Each purification resulted in ~20 µL PCR product. The amount of DNA present in each tube was quantified using Quant iT dsDNA kit (Qubit dsDNA BR assay; Life Technologies, Burlington, Canada) as described in section 3.4.2. The concentration of samples was: V1A with original primers-365 µg/mL, V1A with redesigned primers-53.8 µg/mL, V1B with original primers-151 µg/mL and V1B with redesigned primers-196 µg/mL.

4.3.3 Pyrosequencing of *cpn*60 UT Amplicons

The four libraries were pooled and a single pool was created so that each sample contributed 1250 ng DNA and the concentration of the sample was 28.3 µg/ µL. The 3 major stages involved in the preparation of these samples included: fragment end repair and adaptor ligation, emulsion PCR and bead enrichment, and PicoTiterPlate preparation. The 3 manuals followed for this processing can be found at (http://454.com/downloads/my454/documentation/gs-junior.pdf).

In the first step, Rapid Library preparation was performed using the GS_Junior Titanium series Rapid Library preparation method. In the manual in section 3.2 for fragment end repair at step 2, 16 µL of pooled 28.3 µg/ µL sample was used (500 ng of DNA). AMPure bead preparation was done using steps 5-8 from section 3.3 and then adaptor ligation was done using section 3.4, unligated adaptor removed in section 3.5 and library quality assessed in section 3.6.2 using Agilent bioanalyzer. The assessment showed that the average fragment length was between 600~900 bp (715 bp) and the lower size cut off was less than 10% below 350 bp as expected.

In the second step, Emulsion PCR (emPCR) amplification was performed by following the GS-Junior emPCR Amplification method. Here ssDNA is annealed to excess of DNA capture beads. Then DNA capture beads and PCR reagents were emulsified in water-in-oil microreactors where amplification took place. The method was followed from section 2.1. In Section 3.1, Live Amp Mix was prepared according to Table 1a of manual in section 3.1.2. In section 3.2, step 6, 50 µL of 2.7x10^7 copies/ µL of adaptor ligated library was put into a PCR tube and heat denatured and in step 8, 10 µL of DNA library was added to the tube of washed capture beads. And then steps were thoroughly followed till section 3.7.

The third step is the sequencing step, for which the sequencing method manual was thoroughly followed. The emPCR amplicons were sequenced on a picotiter plate (PTP). The PTP is loaded into the GS junior sequencer for sequencing. The resulting sequence data was then sorted by the unique multiplexing ID (MID).

4.3.4 Data Analysis using Microbial Profiling of Metagenomic Samples.

Pyrosequencing data was analysed using a bioinformatic pipeline called mPUMA (Links et al., 2013) (Figure 8). Sequence assembly and chimera checking was performed with gsAssembler (Grabherr et al., 2011) and Bowtie2 (Langmead and Salzberg, 2012) was used for reference mapping to map each experimental read on to reference OTU sequences assembled with gsAssembler. Removal of PCR primer sequences was done with seqclean (sourceforge). Non-chimeric OTU and non-redundant peptide sequences were clustered at 100% identity by CD-hit (Li and Godzik, 2006) to remove redundant sequences. BLASTX (Altschul et al., 1997) was used to identify the correct reading frame for translation of OTU and then translate it to corresponding peptide OTU. Libraries were compared in a taxonomic context using classifier results loaded into MEGAN (Huson et al., 2007). Abundance files, rarefaction curves and indices of diversity for OTU were created using MOTHUR (Schloss et al., 2009).

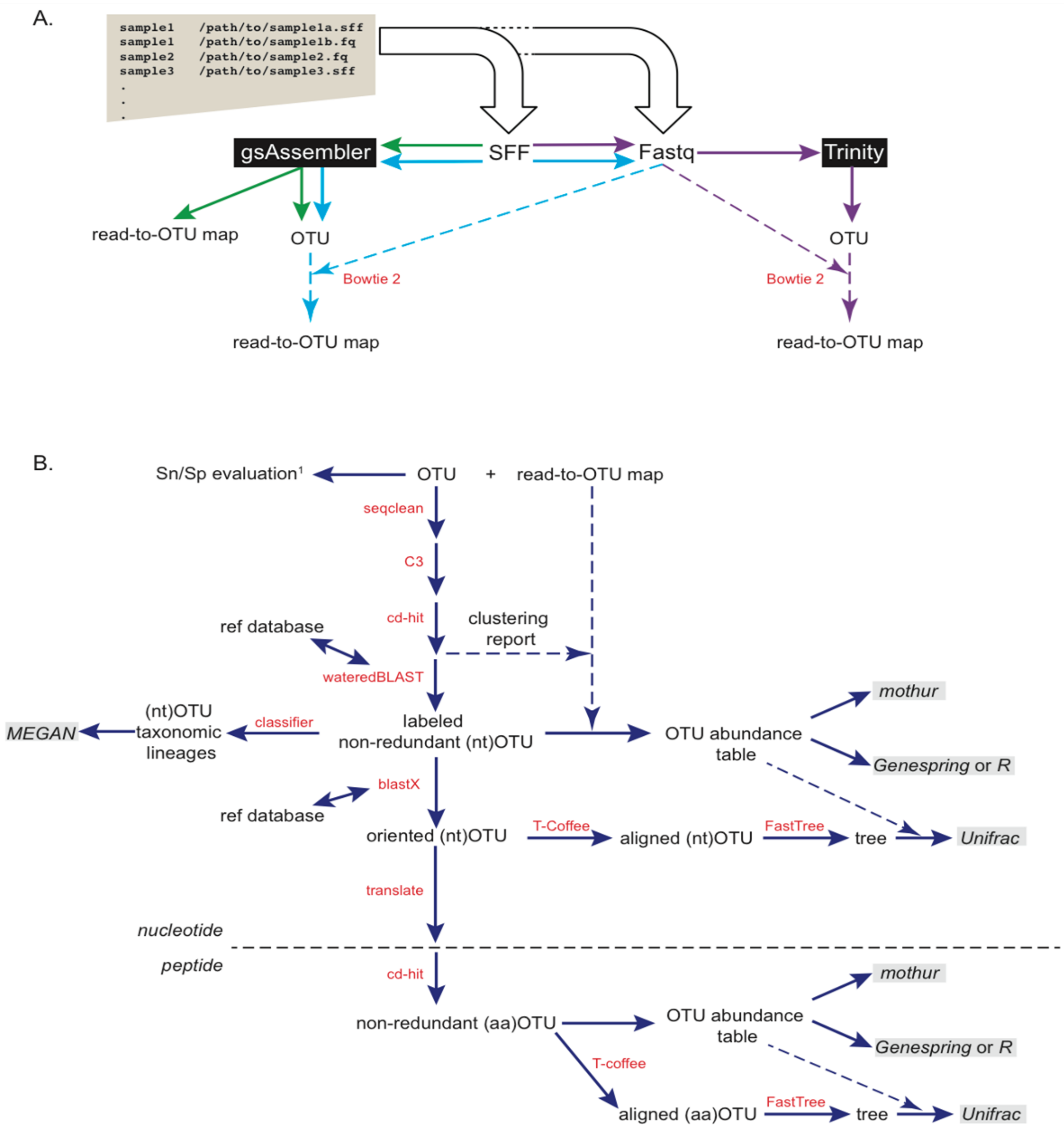


Figure 8: Microbial profiling of metagenomic assemblies pipeline (Links et al., 2013) (with permission from authors) showing mPUMA workflow.

mPUMA workflow. Programs used at each step in the pipeline are shown in red. **A**. User-defined protocol options for assembly and read-to-operational taxonomic unit (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. Gray 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). WateredBLAST is a combination of BLAST and Smith-Waterman alignments (Links et al., 2013).

4.3.5 Phylogenetic Trees

*cpn*60 reference sequences that were identified as best hits for experimental sequences were used to generate a phylogenetic tree. Experimental sequences having the same best hit were clustered for this analysis. Thus, a number distinct sequences, each sequence being most similar to a given reference sequence were clustered. Sequences for this tree were aligned using ClustalW (gap opening penalty=10, gap extension penalty=0.10) (Thomopson et al., 1994), followed by utilization of the Phylip software package (Felsenstein, 1989) to calculate a distance matrix using dnadist and construct a tree using neighbor. The final tree was obtained from the bootstrapped consensus of 300 trees and was visualized using Treeview (Page, 1996). The abundance of experimental sequences were represented using http://itol.embl.de/ (Letunic and Bork, 2007).

The tree for *G. vaginalis* sequences was made using all the *G. vaginalis* reference sequences from cpnDB including those used in (Jayaprakash et al., 2012) along with all the experimental *G. vaginalis* sequences obtained in our study. Sequences for both the trees were aligned using ClustalW (gap opening penalty=10, gap extension penalty=0.10), followed by utilization of the Phylip software package to calculate a distance matrix using dnadist and construct a tree using neighbor. The final tree was obtained from the bootstrapped consensus of 300 trees and was visualized using Treeview. The abundance of experimental *G. vaginalis* sequences were represented using http://itol.embl.de/ (Letunic and Bork, 2007).

4.4 Results

4.4.1 Bacterial Profiles for Vaginal Samples.

A total of 71,552 reads was generated from four amplicon libraries. The number of reads in each library were: V1A with original primers-15274, V1A with redesigned primers-9163, V1B with original primers-34878 and V1B with redesigned primers-12237. The reads were assembled into

504 OTU where each OTU was a unique *cpn*60 UT nucleotide sequence. Two OTU may have differed by as little as one nucleotide over the UT.

4.4.1.1 Rarefaction Curves for Amplicons produced using Original and Redesigned *cpn*60 UT Primers and Vaginal Metagenomic DNA Templates. To assess if the sampling of each vaginal sample was thorough and well represented, rarefaction curves were generated using MOTHUR (Figure 9). Subsampling was performed to normalize sequence reads as each library had different number of sequence reads. This was done to avoid biases introduced by unequal sampling effort (Gihring et al., 2012). To accomplish this, OTU abundance data for each sample was sub-sampled at random to the size of the smallest library. In other words, random selection of number of sequence reads from each sample was done that pertained to lowest sequence abundance among all samples. Here the number of reads ranged from 9163 to 34878 and the number was normalized to 8900 sequence reads. The rarefaction curves were generated using MOTHUR by plotting the number of OTUs as a function of the number of sequence reads. The number of normalized sequence reads sampled were plotted on the X-axis of graph and the no. of OTUs observed for that number of sequence reads were plotted on Y-axis. As the number of sequences analysed leads to completion, the curve will flatten if all the species present in the samples have been discovered and further analysis will not give additional taxa. In the present results, pyrosequencing of the *cpn60* UT resulted in nearly complete sampling of the taxonomic richness of the samples meaning that most unique taxa were identified by sampling effort applied. In Figure 9, for DNA template V1A, the curve for redesigned primers exactly followed the curve generated for original primers, indicating that both primers produced equal species richness. The same was the case with curves generated for template V1B with original and redesigned primers, where both of them showed nearly equal species richness (labelled 'difference between primers'). If one primer set had amplified more taxa than the other primer set, than the curves would have been far apart. Also, the large difference between the curves from samples V1A and V1B (labelled 'difference between templates') showed that we used samples from two microbial communities with very distinct profiles. The difference between the DNA templates (V1A and V1B) from two microbial communities is obvious and greater than the difference between the working of redesigned and original primers on the same microbial community.

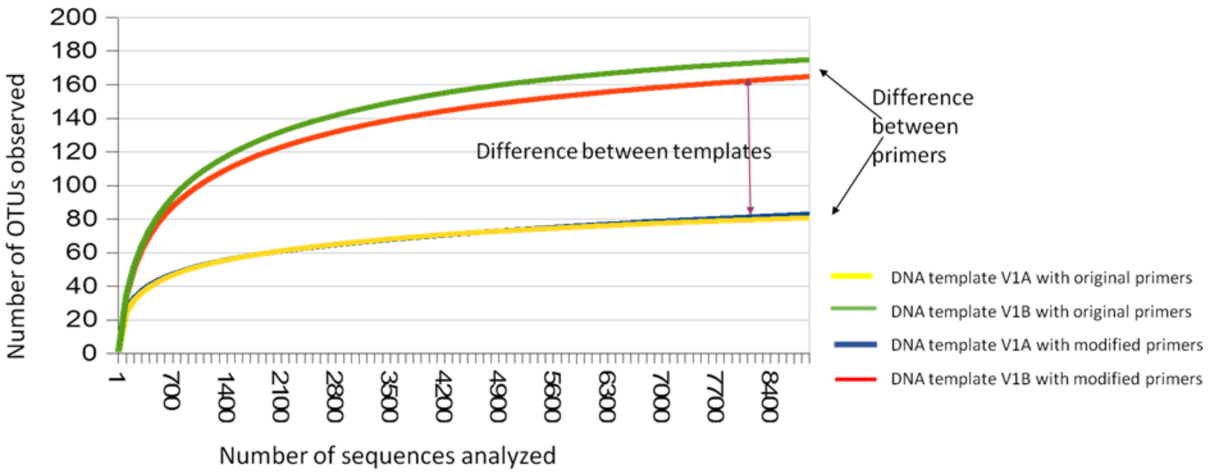


Figure 9: Rarefaction curves for amplicons produced using original and redesigned *cpn*60 UT primers and vaginal metagenomic DNA templates. The figure shows that the difference between diversity of templates V1A and V1B is more than the difference between the behaviour of original and redesigned primers

4.4.1.2 Comparison of Diversity Indices for *cpn*60 UT Amplicon Sequences produced using Original and Redesigned Primers and Vaginal Metagenomic DNA Templates.

There are a handful of indices for looking at diversity of samples. One of these is Shannon`s diversity index the value for which falls between 1.5-3.5 (Shannon, 1948). It takes into account both the number and relative evenness of OTU in a given sample. A greater number of species and a more even distribution of species both increase the Shannon`s diversity. Simpson`s dominance index value ranges from 0 (all taxa are equally present) to 1.0 (one taxon dominates the community completely) (Simpson, 1949). Diversity indices were generated using MOTHUR. The Simpson and Shannon indices were similar with both the original and redesigned *cpn*60 UT primers (Table 3). This indicates that the primers behaved similarly for detecting bacterial diversity.

Table 3: Comparison of diversity indices for *cpn*60 UT amplicon sequences produced using original and redesigned primers and vaginal metagenomic DNA templates.

Primers	DNA template	*Number of sequences	$^{\alpha}$Shannon index	$^{\beta}$Simpson index
Original[1]	V1A	9030	2.9	.10
Redesigned[2]		8934	2.8	.13
Original[1]	V1B	8979	3.3	.11
Redesigned[2]		8966	3.2	.13

*Number of sequences- randomly downsampled to model equal sampling effort.

$^{\alpha}$Simpson diversity index range 0 to 1.0

$^{\beta}$Shannon diversity index range 1.5 to 3.5

Original[1] H279,H1612/H280,1613

Redesigned[2] H1780,H1781/H280,H1613

4.4.1.3 Comparison of Bacterial Profiles produced by Original and Redesigned *cpn*60 UT Primers and Vaginal Metagenomic DNA Templates. As a starting point for comparison of bacterial profiles produced from each DNA template with the original or redesigned *cpn*60 UT primers, comparisons were made after clustering experimental sequences according to phyla. The percentage abundance profiles for these phyla were obtained from mPUMA in the classifier profiles directory which includes text files for each library that describe the library in terms of its taxonomic composition. The taxonomic distribution of bacterial phyla in all 4 libraries is summarized in Table 4 and represented graphically in Figure 10. The graphical representation is included here since it has been used in published studies. In one of such studies by Schellenberg et al., a variation of ~2 fold was considered within normal range for a given species where pyrosequencing was done on technical replicates of *cpn*60 amplicons from a vaginal sample from an individual (Schellenberg et al., 2009).

Table 4: Comparison of bacterial profiles at the phylum level produced by original and redesigned *cpn*60 UT primers and vaginal metagenomic DNA templates.

Primers	DNA template	Proportion of reads in phylum (%)				
		Bacteroidetes	Firmicutes	Actinobacteria	Proteobacteria	Unknown
Original[1]	V1A	15.06	30.47	52.76	0	1.72
Redesigned[2]		14.62	25.72	57.35	0	2.31
Original[1]	V1B	24.86	35.74	15.42	.04	23.93
Redesigned[2]		15.48	39.0	12.31	.08	33.13

Original[1] H1782,H1783/H280,1613

Redesigned[2] H1784,H1785/H280,H1613

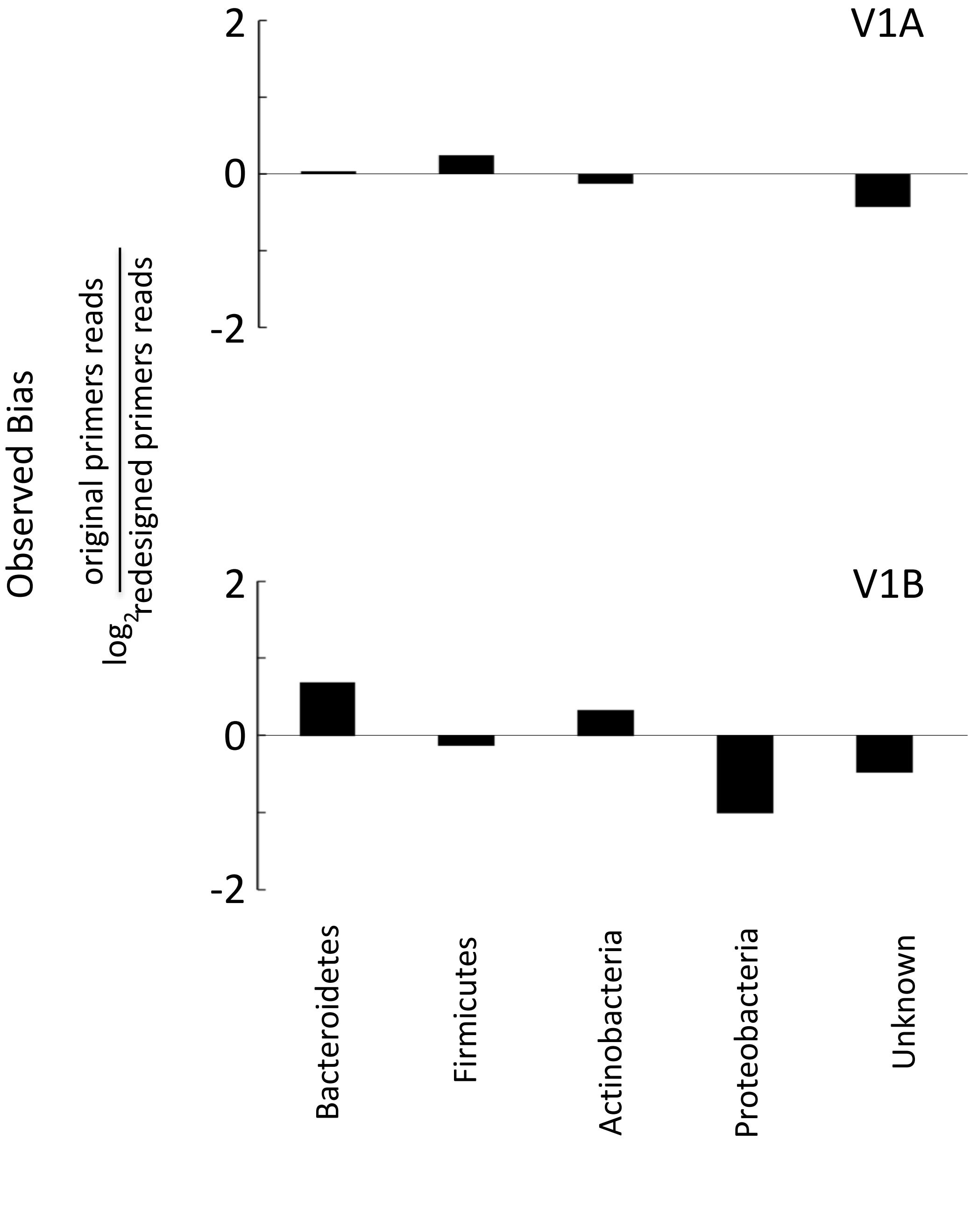


Figure 10: Graphical comparison of bacterial profiles at the phylum level produced by original and redesigned *cpn*60 UT primers and vaginal metagenomic DNA templates. Y-axis – Primer bias observed using ratio of bacterial phylum abundance estimates using original and redesigned *cpn*60 UT primers. X-axis - phylum. This graphical comparison is based on the data used to produce Table 4.

The proportion of reads in each phylum were similar with the original and redesigned primers for template V1A. When used on template V1B, the proportion of bacteroidetes was greater with the original primers than with the redesigned primers and the proportion of unknown sequences was greater with the redesigned primers than with the original primers. These differences may reflect differences in the performance of the two primer sets. A possibility for this difference is that for V1B template, the sequences obtained for bacteroidetes had no reference sequences present in cpnDB and that the bacteroidetes sequences in V1B are actually the unknown sequences that could not be identified. For template V1A, the proportion of unknown sequences was only 2% and similar for each primer set tested. In contrast, the unknown reads for template V1B represented a significant proportion of the total. Unknown sequences were those sequences which were *cpn*60 but had no match in the cpnDB reference database. No statistical tests could be done in absence of replicates.

The above analysis was done by defining OTU to be phylum. Thus each of the originally defined 504 unique nucleotide sequences were assigned to one of the four phyla or as unknown. To analyse the OTU at a lower phylogenetic level, each of the 504 original OTU were assigned the identities of their closest reference sequence match. Therefore for this analysis, OTU was defined as the *cpn*60 reference sequence that was the closest hit to given experimental sequences. As shown in figure 11, the number of OTU observed with both primers in the same template are the same. 25 OTU were observed for V1A and 35 OTU for V1B and these OTU were represented proportionally on log2 graph (Figure 11) to compare our observations with published studies (Schellenberg et al., 2009) already mentioned. The proportional abundance of majority of OTU showed maximum of or less than 4 fold variation with both primers. Only one OTU in V1A and 3 OTU in V1B are more than the four fold variation. With V1A, *Atopobium* *vaginae* (b13654) showed more than 32 fold abundance with original primers than with redesigned primers. With V1B, *Mobiluncus mullieris* (b13762) was more than 8 times and *Prevotella amnii* (b17632) was 16 times more abundant with redesigned primers and *Atopobium vaginae* (b13654) was 8 times more abundant with original primers. We compared our results with a similar study where results from pooled samples from 4 individuals were compared between *cpn*60 and 16S rRNA GS-FLX sequencing and then the results were represented in a graphical way (Schellenberg et al., 2009). In our study, the proportional difference between the taxa amplified using original and redesigned primers was somewhere between these two studies, one showing maximum proportional abundance between technical replicates of same sample as ~4 fold and other showing maximum proportional abundance between 4 pooled samples amplified using different gene targets as ~128 folds (for 2 taxa) and ~16 folds for remaining 13 taxa. The differential representation of some of the templates may be due to the efficiency with which various species are amplified by the universal primers.

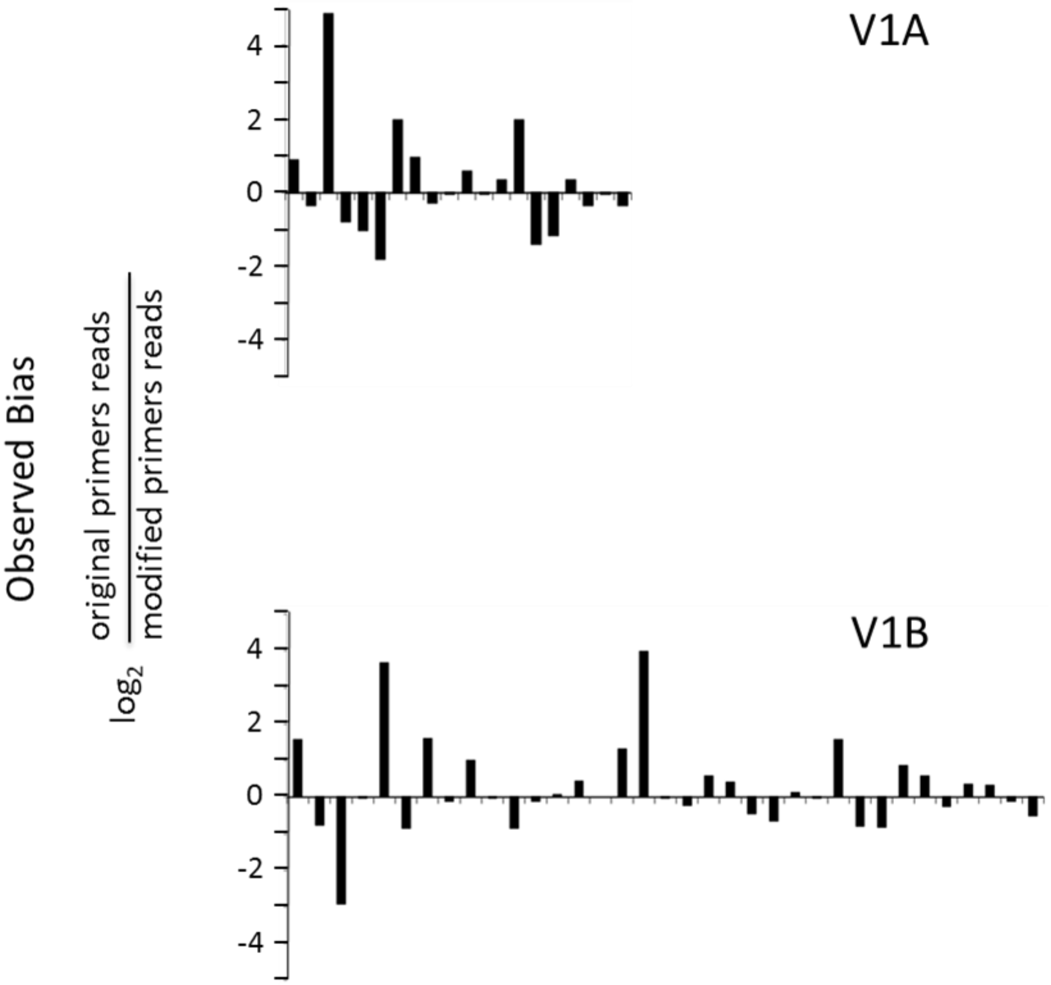


Figure 11: Graphical comparison of bacterial profiles produced by original and redesigned *cpn*60 UT primers with vaginal metagenomic DNA templates. Y-axis – Primer bias observed using ratio of bacterial OTU abundance estimates observed using original and redesigned *cpn*60 UT primers*. X-axis – OTU (OTU- defined as *cpn*60 UT reference sequences) (Appendix 1a and b).

*The OTU showing no bar on the graph is either due to same value of % reads with original and redesigned primers or because one of these values was zero and the log2 ratio did not give valid result for the ratio.

OTU may also be defined as a unique *cpn*60 UT sequence where two OTU may differ by as little as one nucleotide over the UT. For this type of analysis, reads were assembled into OTU in mPUMA using gsAssembler and read mapping was done using Bowtie 2. The relative proportions of each OTU were quantified in each library while normalizing for library size and the log2 of the ratio of the percentage of each OTU in its corresponding library was determined. A value of "0" means that both the libraries have same proportions of that OTU. A positive value means the OTU is more abundant with original primer and a negative value indicates the OTU is less abundant with original primer. In case of template V1A, for ~50% of OTUs (33 out of 65), the variation was less than or equal to 2 fold. As shown in figure 12, proportional abundance of majority of OTU is concentrated within the 2 fold variation range. The rest of the 32 OTU, showed more than 2 fold variation which for one of the OTU was 32 fold (*Atopobium vaginae*) (Figure 12). Different OTU having the same reference *cpn*60 UT sequence as best hit were sometimes represented by different proportional abundance with the same primer in the same template, for example OTU108, OTU110 and OTU123 were identified as *G. vaginalis*, when template V1A was amplified with original and redesigned primers. OTU108 was ~16 fold less abundant with original primers, OTU110 was less than 2 fold abundant with original primers and OTU123 was more than 4 fold abundant with original primers. In case of template V1B, for majority of OTU (95 out of 131), the variation was less than or equal to 2 fold. The rest of the 36 OTU, showed more than 2 fold variation which for one of the OTU was 32 fold (*Atopobium vaginae*). In this template also, same primers behaved differently on different OTU that had same *cpn*60 reference sequence as best hit, therefore, the testing of hypothesis is not addressed very closely here that the original and redesigned primers produced indistinguishable profiles. Some OTU produced no bar on the graph, either due to same value of % reads with original and redesigned primers or because one of these values was zero and the log2 ratio did not give valid result for the ratio. Therefore, the log2 ratios with invalid (undefined) values were not shown in graph at all.

The OTU analysis above was done from higher to lower taxonomic level. The analysis at phylum level showed the variation between the relative abundance of four phyla and unknown sequences to be within a normal variation as 2 fold was the normal variation between technical replicates in a similar study with same primers. Therefore, the primers behaved similarly at higher taxonomic levels. For lower taxonomic levels, the proportional abundance was within variation of 4 fold for majority of OTU in both the templates showing that primers behaved similarly for them. The species that showed a proportional difference of more than fourfold were the ones that represented real differences in their representation in the two templates, since this is the maximal variation that was seen in the technical replicates in study by Schellenberg et al. in 2009, but the number of such species showing greater variability were very few in our data.

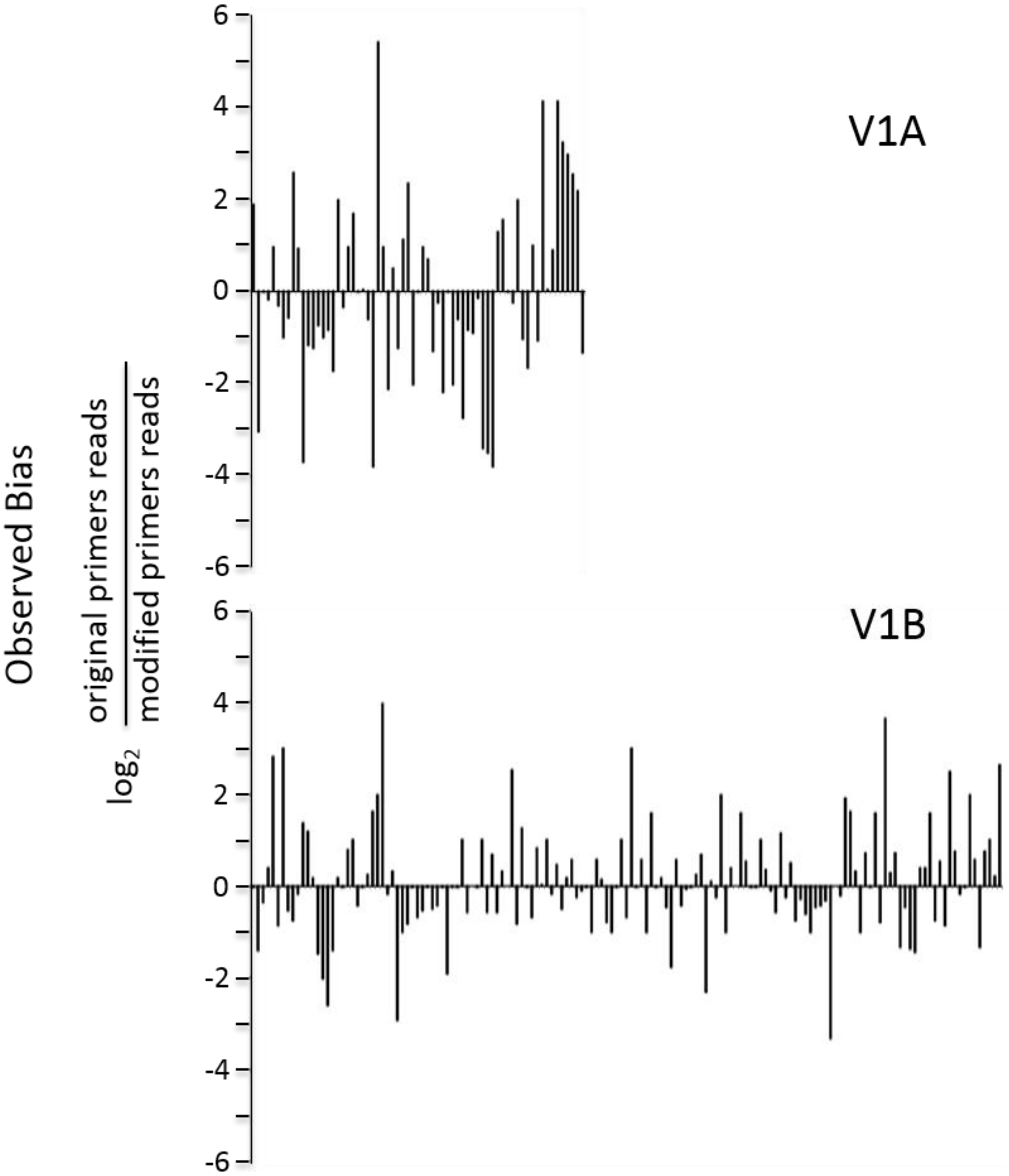


Figure 12: Graphical comparison of bacterial profiles at OTU level (OTU=*cpn*60 UT sequence reads) produced by original and redesigned *cpn*60 UT primers and vaginal metagenomic DNA templates. Y-axis –Primer bias observed using ratio of bacterial OTU abundance estimates observed using original and redesigned *cpn*60 UT primers. X-axis - OTU. (For description of OTU see Appendix 2a and 2b).

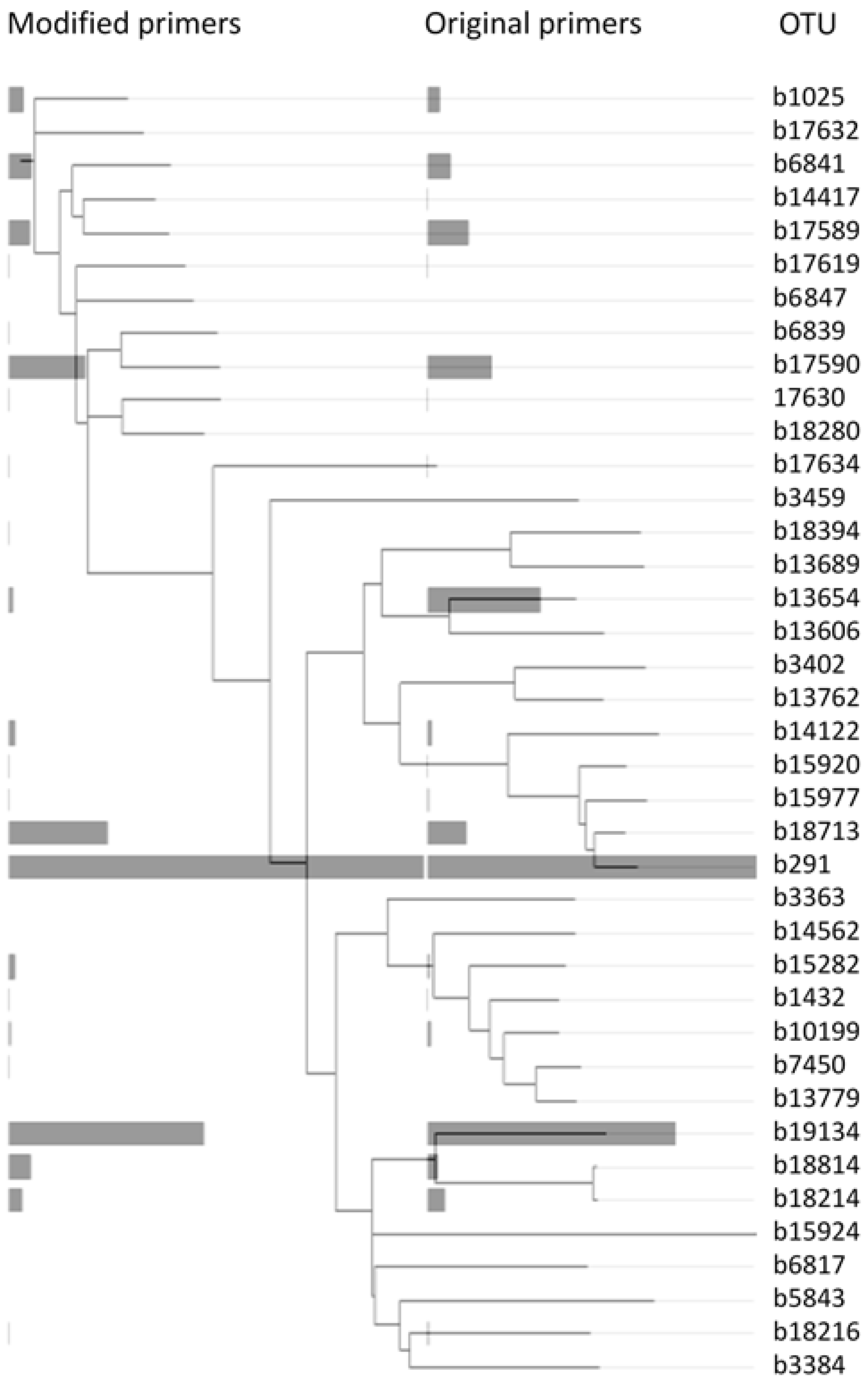
4.4.1.4 Phylogenetic Representation of Bacterial Profiles.

The above analysis was done to compare and analyze the results among broader categories of bacterial classification. Here, the same data analysis and comparison on a finer scale for different species of bacteria found in this study was done and the phylogenetic relationship among clustered experimental sequences was explored. The *cpn*60 reference sequences that were identified as best hits for experimental sequences were used to generate a phylogenetic tree (Figure 13) as explained in Materials and Methods 4.3.5. Experimental sequences having the same best hit were clustered for this analysis. Thus, a number of distinct sequences, each sequence being most similar to a given reference sequence were clustered. For example, b291 identifies 24 distinct sequences with range of sequence identities to b291 (92-99%). 86% of reads were in 3 OTU that were 98-99% identical to b291 (Appendix 3c). A graphical representation of the relative abundance of each clustered OTU was superimposed on the tree using an online program called iTol (Letunic and Bork, 2007). The size of the bar for an OTU in the graph is directly proportional to the normalized count of that OTU. See Appendix 3a and 3b for values of normalized counts. The bacterial abundance profiles produced by both original and redesigned primers for both templates V1A and V1B were observed to be similar. Except for *Atopobium vaginae* b13654, which was more abundant with original primers in template V1A.

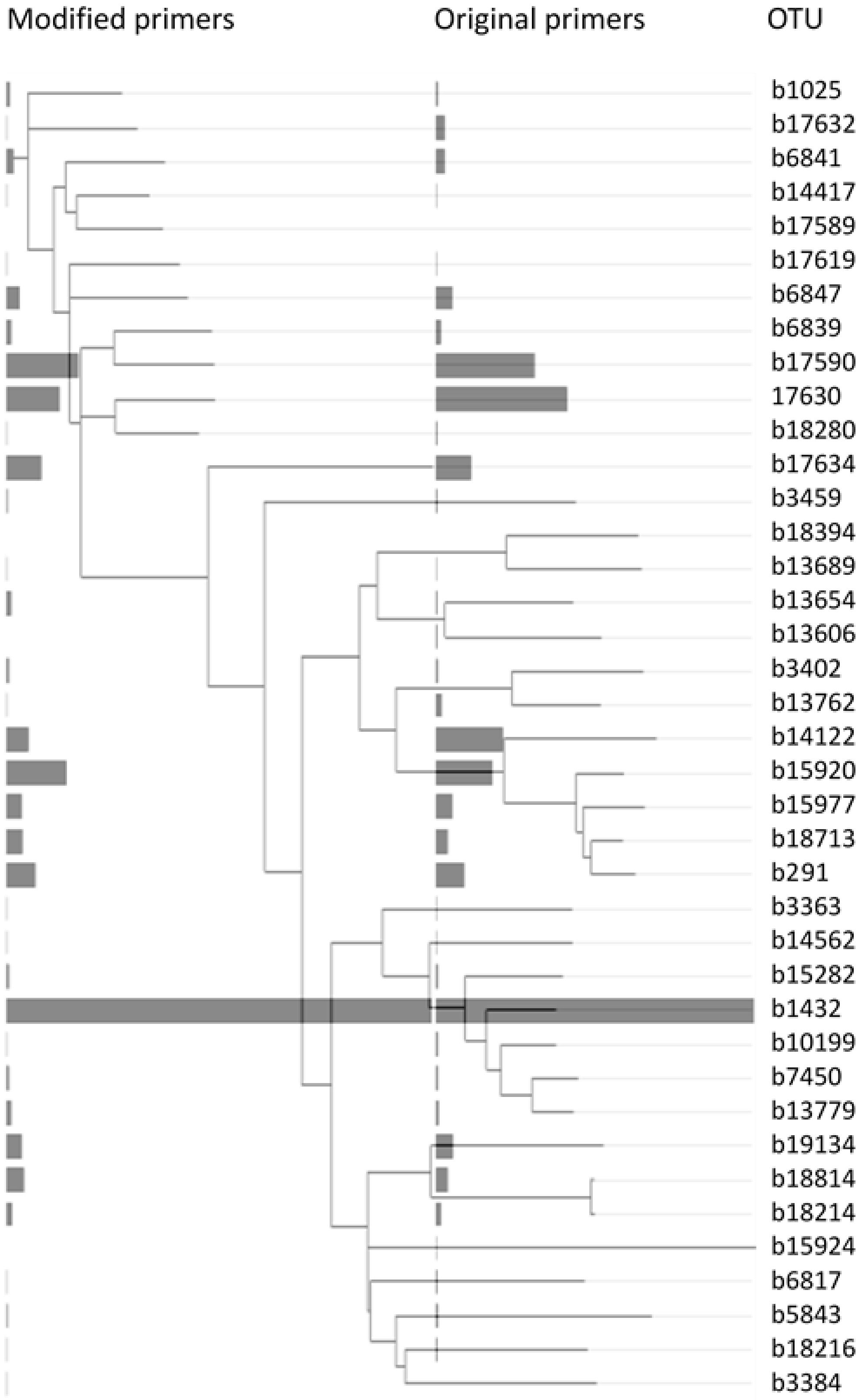
Figure 13: Phylogenetic representation of bacterial profiles.

A phylogenetic tree was generated from *cpn*60 reference sequences that were the best hits for experimental sequences observed in the library. The experimental sequences were produced using either redesigned or original *cpn*60 UT primers on vaginal templates V1A and V1B. Experimental sequences having the same best hit were clustered for this analysis. Thus, a number of distinct sequences, each sequence being most similar to a given reference sequence were clustered. For example, b291 identifies 24 distinct sequences with range of sequence identities to b291 (92-99%). 86% of reads were in 3 OTU that were 98-99% identical to b291 (Appendix 3c). A graphical representation of the relative abundance of each clustered OTU was superimposed on the tree (Bacteria V1A and Bacteria V1B). The size of the bar for an OTU in the graph is directly proportional to the normalized count of that OTU. See Appendix 3a and 3b for values of normalized counts.

Bacteria V1A



Bacteria V1B



4.4.1.5 Phylogenetic Representation of *G. vaginalis* Profiles.

Recently, *cpn*60 UT has been proved to be a robust tool to resolve the available *G. vaginalis* strains into four sub-groups. There has also been evidence shown that may eventually lead to reclassification of these four sub-groups into four species (Jayaprakash et al., 2012). A phylogenetic tree of *G. vaginalis* reference and experimental sequences was overlain with a graphical representation of the relative abundances of each *G. vaginalis* OTU observed using either redesigned or original *cpn*60 UT primers (Figure 14). Reference sequences were from cpnDB including those described in Jayaprakash et al. 2012. It is a rooted tree with *Alloscardovia omnicolens* as outgroup (b10027A.om). The size of the bar for an OTU in graph is directly proportional to the normalized count of that OTU in its library. The values of normalized counts is shown in Appendix 4a and 4b. Four subgroups similar to those observed in Jayaprakash et al., 2012 were observed in this study. The *G. vaginalis* abundance profiles with original and redesigned primers were observed to be similar. Some OTU observed to be showing extreme variation with original primers were balanced by similar sequences observed to be showing opposite trend with original primers. In template V1A, OTU 010b291 was observed to be most abundant with redesigned primers whereas OTU 049b291 showed very low abundance with same primers. The abundance trend for both these OTU was opposite with original primers. Both these OTU were 98% similar to *G. vaginalis* reference sequence (cpnDBID:b291).

As 010b291 and 049b291 OTU were most abundant in DNA template V1A and V1B (see figure 14a and 14b), another phylogenetic tree was made where 010b291 and 049b291 were not included to see whether they are masking any difference in abundance comparisons among other *G. vaginalis* OTU. It was observed that the abundance profiles without 049b291 and 010b291 on bar graph looked similar for other sequences as well (Figure 15).

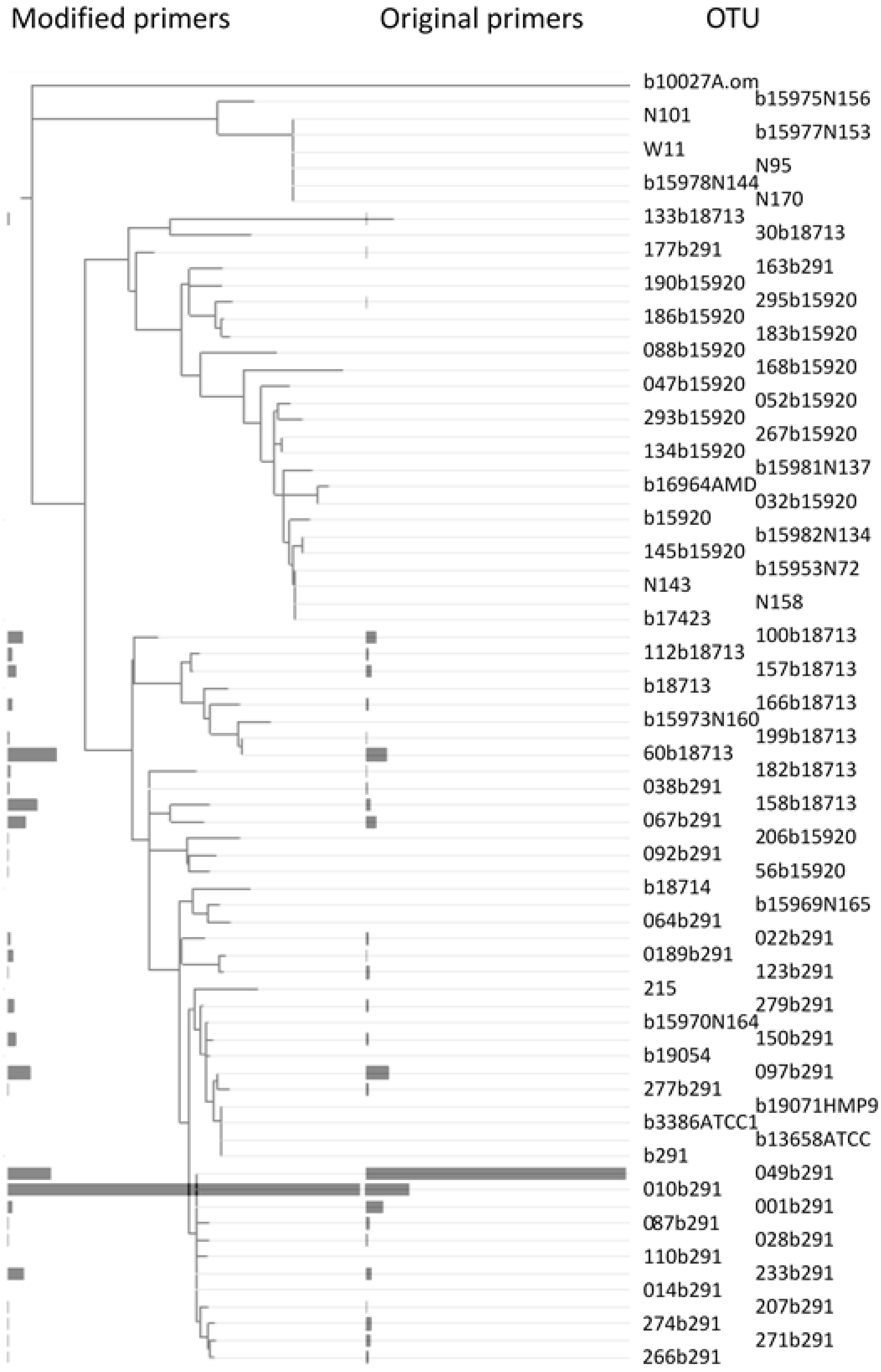
4.4.2 Fungal Profiles for Vaginal Samples

In the metagenomic templates used in our study, no fungal sequences were observed in the vaginal samples either with original or redesigned primers in both the V1A and V1B templates.

Figure 14: Phylogenetic representation of *G. vaginalis* profiles.

A phylogenetic tree was generated from *G. vaginalis* like experimental sequences observed in the study and reference *G. vaginalis* sequences. A graphical representation of the relative abundances of each *G. vaginalis* OTU was superimposed on the tree (*G. vaginalis* V1A and *G. vaginalis* V1B). The experimental sequences were produced using either redesigned or original *cpn*60 UT primers on vaginal templates V1A and V1B. Reference sequences were obtained from cpnDB including those described in Jayaprakash et al., 2012. Names of experimental sequences in the tree are represented by their OTU numbers followed by the cpnDB ID of the reference sequence it looks like. Reference sequences are highlighted in the tree. It is a rooted tree with *Alloscardovia omnicolens* as outgroup (b10027A.om). The size of the bar for an OTU in the graph is directly proportional to the normalized count of that OTU in its library. See Appendix 4a and 4b for values of normalized counts.

Gardnerella vaginalis V1A



Gardnerella vaginalis V1B

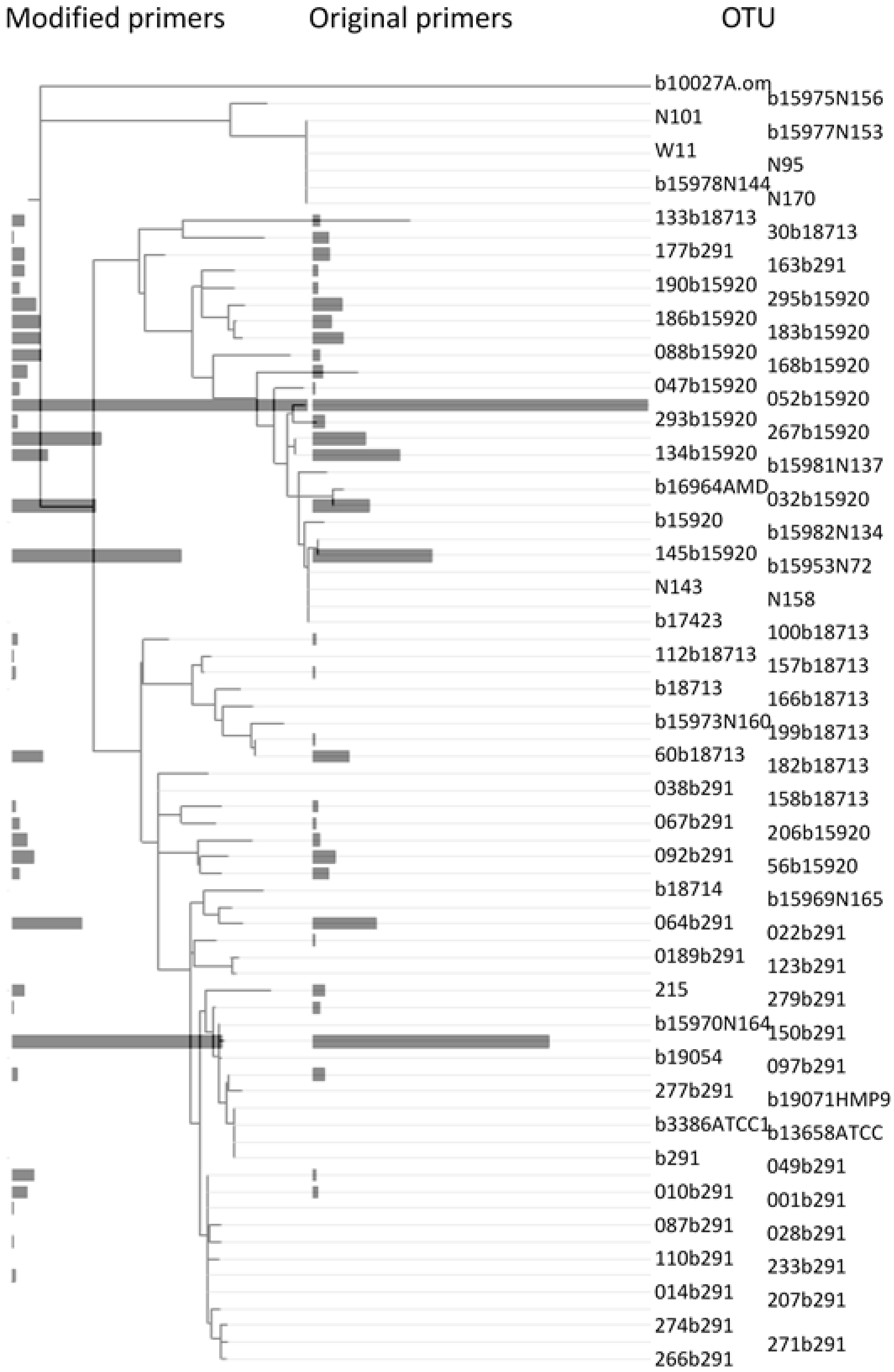
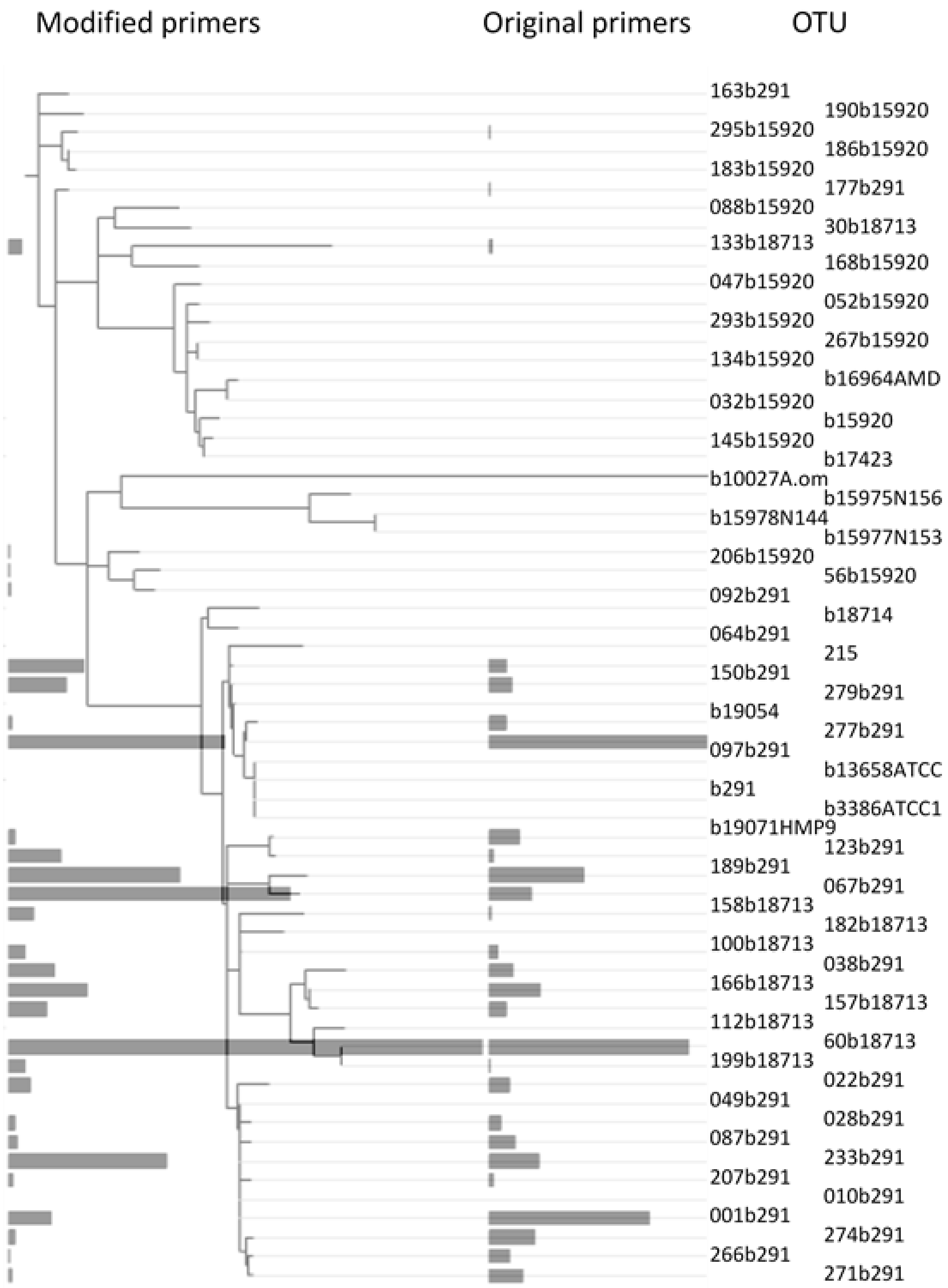


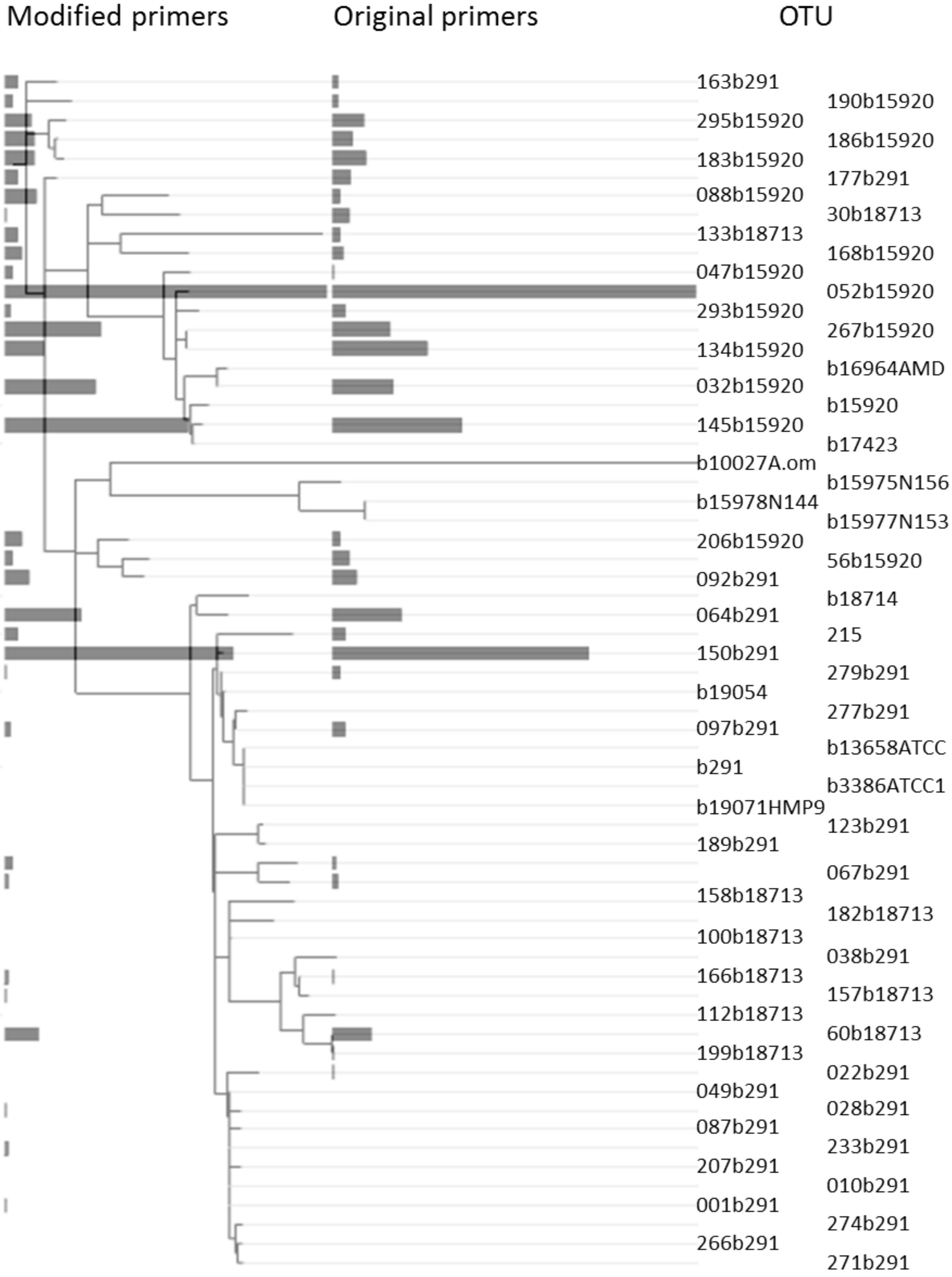
Figure 15: Phylogenetic representation of selected *G. vaginalis* profiles.

A phylogenetic tree of *G. vaginalis* reference and experimental sequences was overlain with a graphical representation of the relative abundance of each *G. vaginalis* OTU (except 010b291 and 049b291) observed using either redesigned or original *cpn*60 UT primers (*G. vaginalis* (minor)V1A and *G. vaginalis* (minor) V1B). The experimental sequences were produced using either redesigned or original *cpn*60 UT primers on vaginal templates V1A and V1B. Reference sequences were from cpnDB including those described in Jayaprakash et al. 2012. It is a rooted tree with *Alloscardovia omnicolens* as outgroup (b10027A.om). The size of the bar for an OTU in the graph is directly proportional to the normalized count of that OTU in its library. See Appendix 4a and 4b for values of normalized counts. As 010b291 and 049b291 OTU were most abundant in DNA template V1A and V1B (see figure 14a and 14b), they were not included in this tree to see whether they are masking any abundance comparisons between other *G. vaginalis* OTU.

Gardnerella vaginalis (minor) V1A



Gardnerella vaginalis (minor) V1B



4.5 Discussion

Most of the taxonomic markers available presently are either suitable to identify prokaryotes or eukaryotes. A gene marker that can identify both prokaryotes and eukaryotes from the same microbial sample can be of an immense advantage. The *cpn*60 gene is such a phylogenetic marker that can be used simultaneously to profile both bacteria and fungi from microbial communities (Links et al., 2014) using its *cpn60* UT region that can be amplified by a set of PCR primers (Hill et al., 2006). It is a protein coding gene that is present in all prokaryotes (except Mollicutes) and eukaryotes. It helps in the formation and maintenance of protein structures acting as a molecular chaperone, hence, the name, chaperonin (Hemmingsen et al., 1988). The *cpn*60 UT sequence identities are also strong interpreters of genome-scale sequence identities (Verbeke et al., 2011). The *cpn*60 UT has almost always a uniform length of 555bp±1 codon that makes sequence alignments an easy task (Hill et al., 2004). The *cpn*60 gene has been recently proposed to be the preferred barcode for bacteria (Links et al., 2012). The *cpn*60 gene also has the following disadvantages. The primers are degenerate and the targets have to be amplified with a cocktail of primers and a range of annealing temperatures have to be used so that all the community members are amplified (Hill et al., 2006). Moreover, the Ribosomal Database is quiet vast and includes reference sequences from diverse environments and taxa as compared to cpnDB that still needs to be expanded.

In this experiment, the taxonomic profiles and the relative proportions of taxa generated by original or redesigned *cpn*60 primers at phylum level, were nearly identical as indicated by rarefaction curve, the diversity indices and log2 graph for relative abundance. The purpose of this study was to compare the performance of primers and not the difference in microbial profiles among HIV-ve and HIV+ve women. The four phyla (Bacteriodetes, Firmicutes, Actinobacteria, Proteobacteria) are well known to be present in vagina in studies using *cpn*60 (Hill et al., 2005; Schellenberg et al., 2009; Schellenberg et al., 2011) and using 16S rRNA (Ling et al., 2010). For majority of OTUs, the variation was less than 2 fold. In the absence of any statistical tests, we wanted a rationale, based on which we could interpret our results. We interpreted our results on a similar study where pyrosequencing was done on technical replicates of *cpn*60 amplicons from a vaginal sample from an individual and where variation of ~2 fold was considered within normal range for a given species. In same study, results from pooled samples from 4 individuals were compared between *cpn*60 and 16S rRNA GS-FLX sequencing and then the results were represented in a graphical way (Schellenberg et al., 2009). The maximal variation extended ~2 fold between these datasets. The proportional difference between the relative abundance of OTU in template V1A produced by original and redesigned primers in our study was somewhere between these two studies for majority of OTU, one showing proportional abundance between technical replicates of same sample that was < 2 fold and other showing proportional abundance between 4 pooled samples amplified using different gene targets that was ~2 fold. The differential representation of some of the templates may be due to the efficiency with which various species are amplified by the universal primers. In conclusion, although a few species were represented with different proportional abundances within their respective datasets, most of them were represented with almost equal abundances as is apparent from figures 11 and 12, the majority of OTU relative abundance bars are concentrated around 2 fold change. Our study does not address the hypothesis very conclusively, but the results still point to the absence of any differences between the performance of primers.

Fungal sequences have been observed in the vaginal mucosa using 18S rRNA (Guo et al., 2012) and ITS (Drell et al., 2013) sequencing. The inability to observe any fungal sequences in this study may be due various reasons discussed below, although we found them successful when used on pure fungal DNA extracts. Since *cpn*60 has the advantage of simultaneous profiling of both prokaryotes and eukaryotes from microbial communities (Links et al., 2014), the major reason for unobserved fungal sequences may be their mere absence in the given samples.

The other probable reasons for the un-observed fungal species in the vaginal samples using the redesigned primers are discussed here. Vaginal mycobiome exists as a 'rare biosphere' (Huffnagle and Noverr, 2013) or low abundant species in the vaginal mucosa that is primarily dominated by *Lactobacilli* bacteria. Such low abundance taxa are under-represented when microbial communities are profiled using PCR-based methods. Reason being that the PCR reaction reagents are diminished rapidly by the dominant species that have high template abundance in the high diversity samples (Amend et al., 2010). The same concept can be applied to the representative microbial sample from vaginal mucosa used in our study. We used a 12 sample pool with 10 µL of each sample. As fungal sequences are already rare in the vagina, we are further diminishing their presence by using only 10 µL of one sample, thereby, increasing the chances of missing out on the rarer sequences.

In another experiment (Gonzalez et al., 2012), two experimental artificial microbial communities from the human oral cavity were prepared. Both experimental communities were sequenced by shotgun sequencing (no PCR) and also after their PCR amplification by 16S rRNA. Since shotgun or direct sequencing involved no PCR amplification, it was assumed to represent the closest data available to the oral microbial community. 33 OTUs (664reads) and 28 OTUs (1302 reads) were obtained by direct sequencing of experimental communities 1 and 2. 17 OTUs (230 reads) and 15 OTUs (2056 reads) were obtained by sequencing 16S rRNA amplicons. Although the experiment is based on 16S rRNA primers, the idea here is that some of the OTUs present in low abundance remained undetected after PCR amplification.

PCR bias due to secondary structures in DNA templates

Presence of secondary structures in DNA templates can lead to folding of the DNA template which may have resulted in failure of the primers to amplify the UT in the vaginal samples. This has already been observed in another experiment in our lab (unpublished, Hemmingsen lab) where redesigned primers were used. Our lab received samples of total DNA purified from AMF spores from another collaborating lab. When these samples were amplified using *cpn*60 UT redesigned primers, bacterial and other fungal *cpn*60 UT sequences were obtained but there were no AMF sequences in the sequence results. But, when novel *cpn*60 1786-1788 primers were used, AMF sequences were obtained, pointing to the presence of secondary structures in the *cpn*60 UT of AMF DNA which may have impeded the amplification of AMF UT by *cpn*60 UT primers. In this study, only *cpn*60 UT primers were used, that may have failed to amplify the fungal sequences present in the vaginal samples because of the presence of secondary structures in the UT part.

Inhibition of amplification by humic acids

Co-extraction of humic acids along with DNA from microbial samples may decrease amplification efficiency by inhibiting PCR. Humic acids inhibit amplification by binding to the polymerase, target DNA or co-factor magnesium ions (Wilson, 1997; Roose-Amsaleg et al., 2001). When it binds to target DNA, it does sequence specific binding and in this study it may have been the fungal DNA to which it bound and made it unavailable for PCR (Opel et al., 2010). The protocol used to extract DNA from vaginal samples in our study, although not done in our lab, used the Magmax^{TM} total DNA extraction kit which had no additional step to remove humic acids or other inhibitors. A combination of different DNA extraction procedures may have been used to get a more realistic view of the microbial diversity present in vaginal samples.

There was a possibility to include a positive control in our study in the form of "mock community" having known sequences of fungal species that were expected to be found in vagina. We could also have used ITS or 18S rRNA primers to see if the issue is cpn60 primers or not. This could have revealed whether the fungal species were not observed in our samples due to other reasons or because the fungi were just absent in the samples.

4.6 Conclusions

Original *cpn*60 UT primers have already been extensively used to profile microbial communities from diverse environments like pig faeces, dog faeces, vaginal samples and intestinal communities. Most of the analysis of the above mentioned microbiomes has been limited to bacteria. In this study, the original *cpn*60 primers were redesigned to amplify universal target and some parts from any *cpn*60 gene, including most of the *cpn*60 genes from fungal species present in cpnDB. The redesigned primers were to be considered useful if they produced the same bacterial profiles and abundances as produced by original primers and additionally, if the redesigned primers also amplify the fungal templates if present in given microbial samples. When these primers were used on microbial samples in the present study, the bacterial profiles and abundances obtained with original and redesigned primers were the same, possibly indistinguishable. Also, these primers were successful in amplifying *cpn*60 gene parts from diverse fungal phylogeny like Ascomycota, Basidiomycota, Chytridiomycota and Zygomycota. Although in this study fungal sequences in the complex microbial sample were not observed using *cpn*60, the utility of *cpn*60 gene to successfully and simultaneously detect both bacteria and fungi cannot be denied (Links et al., 2014). Also, the fungal sequences in microbial community samples may have stayed undetected due to the co-extraction of humic-acids along with DNA from vaginal samples or presence of secondary structures in the templates or may be due to the rarity of fungal templates in the vaginal samples which were further diminished by the few µL (10 µL each from 12 samples) used to make synthetic pool sample.

There were some limitations to our study which, if overcome, could have addressed our hypothesis more conclusively. We could have used technical replicates and it would have helped in testing the statistical significance of our results. In the absence of replicates, we interpreted our results based on a similar study where pyrosequencing was done on technical replicates of *cpn*60 amplicons from a vaginal sample from an individual (Schellenberg *et al*., 2009). We could have chosen a different and more diverse microbial sample rich in fungal templates for getting better interpretable results.

Any set of primers, ribosomal RNA based (like 16S, 18S) or ITS or protein based (like *cpn*60) are not self-sufficient for complete profiling of microbial communities and some of them may also be biased towards amplification of particular phylum. Therefore, they have to be used in combination with each other for best results. In future, the redesigned or original *cpn*60 primers may also be used in combination with other gene targets where the *cpn*60 gene may provide an increased level of resolution as compared to structural rRNA encoding genes and at the same time facilitating simultaneous detection of prokaryotes and eukaryotes from same microbial samples. Since only about 5% of fungal species have been described out of the estimated 1.5 million species, it is becoming more urgent to describe the remaining species before they become extinct, since habitats of many species including those of fungi are destroyed each year (L HAWKSWORTH, 2001; Blackwell, 2011). Therefore, a gene target that can identify both eukaryotes and prokaryotes can be of immense utility to reveal the diversity of microbes present in our environment.

References

AINSWORTH, G. C. 2008. *Ainsworth & Bisby's dictionary of the fungi*, Cabi.

AINSWORTH, G. C., SPARROW, F. K. & SUSSMAN, A. S. 1973. The Fungi. An advanced treatise. Vol. IV B. A taxonomic review with keys: Basidiomycetes and lower fungi. *The Fungi. An advanced treatise. Vol. IV B. A taxonomic review with keys: Basidiomycetes and lower fungi.*

ALEXOPOULOS, C., MIMS, C. W. & BLACKWELL, M. 1996. Introductory Mycology. John Willey and Sons. *Inc., New York,* 868.

ALTSCHUL, S. F., MADDEN, T. L., SCHÄFFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W. & LIPMAN, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research,* 25**,** 3389-3402.

AMANN, R. & FUCHS, B. M. 2008. Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques. *Nature Reviews Microbiology,* 6**,** 339-348.

AMEND, A. S., SEIFERT, K. A. & BRUNS, T. D. 2010. Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Molecular Ecology,* 19**,** 5555-5565.

APAJALAHTI, J. H., SÄRKILAHTI, L. K., MÄKI, B. R., HEIKKINEN, J. P., NURMINEN, P. H. & HOLBEN, W. E. 1998. Effective recovery of bacterial DNA and percent-guanine-plus-cytosine-based analysis of community structure in the gastrointestinal tract of broiler chickens. *Applied and Environmental Microbiology,* 64**,** 4084-4088.

BALDAUF, S. L. & PALMER, J. D. 1993. Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proceedings of the National Academy of Sciences,* 90**,** 11558-11562.

BARRY, T., COLLERAN, G., GLENNON, M., DUNICAN, L. K. & GANNON, F. 1991. The 16s/23s ribosomal spacer region as a target for DNA probes to identify eubacteria. *Genome Research,* 1**,** 51-56.

BARTRAM, A. K., LYNCH, M. D., STEARNS, J. C., MORENO-HAGELSIEB, G. & NEUFELD, J. D. 2011. Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. *Applied and environmental microbiology,* 77**,** 3846-3852.

BELLEMAIN, E., CARLSEN, T., BROCHMANN, C., COISSAC, E., TABERLET, P. & KAUSERUD, H. 2010. ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases. *Bmc Microbiology,* 10**,** 189.

BEN-DOV, E., KRAMARSKY-WINTER, E. & KUSHMARO, A. 2009. An in situ method for cultivating microorganisms using a double encapsulation technique. *FEMS microbiology ecology,* 68**,** 363-371.

BERBEE, M. & TAYLOR, J. 1993. Ascomycete relationships: dating the origin of asexual lineages with 18S ribosomal RNA gene sequence data. *The fungal holomorph: mitotic, meiotic and pleomorphic speciation in fungal systematics***,** 67-78.

BERBEE, M. L. & TAYLOR, J. W. 1995. From 18S ribosomal sequence data to evolution of morphology among the fungi. *Canadian Journal of Botany,* 73**,** 677-683.

BLACKWELL, M. 2011. The Fungi: 1, 2, 3… 5.1 million species? *American Journal of Botany,* 98**,** 426-438.

BLASZCZYK, D., BEDNAREK, I., MACHNIK, G., SYPNIEWSKI, D., SOLTYSIK, D., LOCH, T. & GALKA, S. 2011. Amplified Ribosomal DNA Restriction Analysis (ARDRA) as a Screening Method for Normal and Bulking Activated Sludge Sample Differentiation. *Polish Journal of Environmental Studies,* 20.

BORNEMAN, J. & HARTIN, R. J. 2000. PCR primers that amplify fungal rRNA genes from environmental samples. *Applied and environmental microbiology,* 66**,** 4356-4360.

BROUSSEAU, R., HILL, J. E., PRÉFONTAINE, G., GOH, S.-H., HAREL, J. & HEMMINGSEN, S. M. 2001. Streptococcus suis serotypes characterized by analysis of chaperonin 60 gene sequences. *Applied and environmental microbiology,* 67**,** 4828-4833.

BULLOCH, W. 1938. THE HISTORY OF BACTERIOLOGY. *The American Journal of the Medical Sciences,* 196**,** 868.

BUSE, H. Y., LU, J., LU, X., MOU, X. & ASHBOLT, N. J. 2014. Microbial diversities (16S and 18S rRNA gene pyrosequencing) and environmental pathogens within drinking water biofilms grown on the common premise plumbing materials unplasticized polyvinylchloride and copper. *FEMS microbiology ecology,* 88**,** 280-295.

CARACCIOLO, A. B., BOTTONI, P. & GRENNI, P. 2010. Fluorescence in situ hybridization in soil and water ecosystems: a useful method for studying the effect of xenobiotics on bacterial community structure. *Toxicological & Environmental Chemistry,* 92**,** 567-579.

CHABAN, B., LINKS, M. G., JAYAPRAKASH, T. P., WAGNER, E. C., BOURQUE, D. K., LOHN, Z., ALBERT, A. Y., VAN SCHALKWYK, J., REID, G. & HEMMINGSEN, S. M. 2014. Characterization of the vaginal microbiota of healthy Canadian women through the menstrual cycle. *Microbiome,* 2**,** 23.

CLAYTON, R. A., SUTTON, G., HINKLE, P. S., BULT, C. & FIELDS, C. 1995. Intraspecific variation in small-subunit rRNA sequences in GenBank: why single sequences may not adequately represent prokaryotic taxa. *International journal of systematic bacteriology,* 45**,** 595-599.

COENYE, T., VANDAMME, P. & LIPUMA, J. J. 2003. Ralstonia respiraculi sp. nov., isolated from the respiratory tract of cystic fibrosis patients. *International Journal of Systematic and Evolutionary Microbiology,* 53**,** 1339-1342.

DA ROCHA, U. N., VAN OVERBEEK, L. & VAN ELSAS, J. D. 2009. Exploration of hitherto-uncultured bacteria from the rhizosphere. *FEMS microbiology ecology,* 69**,** 313-328.

DAHLLÖF, I., BAILLIE, H. & KJELLEBERG, S. 2000. rpoB-Based Microbial Community Analysis Avoids Limitations Inherent in 16S rRNA Gene Intraspecies Heterogeneity. *Applied and Environmental Microbiology,* 66**,** 3376-3380.

DESAI, A. R., MUSIL, K. M., CARR, A. P. & 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**,** 120-128.

DISAYATHANOOWAT, T., YOUNG, J. P. W., HELGASON, T. & CHANTAWANNAKUL, P. 2012. T-RFLP analysis of bacterial communities in the midguts of Apis mellifera and Apis cerana honey bees in Thailand. *FEMS microbiology ecology,* 79**,** 273-281.

DRELL, T., LILLSAAR, T., TUMMELEHT, L., SIMM, J., AASPÕLLU, A., VÄIN, E., SAARMA, I., SALUMETS, A., DONDERS, G. G. & METSIS, M. 2013. Characterization of the vaginal micro-and mycobiome in asymptomatic reproductive-age Estonian women. *PLoS One,* 8**,** e54379.

DUMONCEAUX, T. J., HILL, J. E., BRIGGS, S. A., AMOAKO, K. K., HEMMINGSEN, S. M. & 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 microbiological methods,* 64**,** 46-62.

DUMONCEAUX, T. J., HILL, J. E., PELLETIER, C. P., PAICE, M. G., VAN KESSEL, A. G. & HEMMINGSEN, S. M. 2006b. 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**,** 494-500.

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

FOX, G. E., WISOTZKEY, J. D. & JURTSHUK, P. 1992. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *International Journal of Systematic Bacteriology,* 42**,** 166-170.

FRANKLIN, R. B., TAYLOR, D. R. & MILLS, A. L. 1999. Characterization of microbial communities using randomly amplified polymorphic DNA (RAPD). *Journal of Microbiological Methods,* 35**,** 225-235.

GENTRY, T., WICKHAM, G., SCHADT, C., HE, Z. & ZHOU, J. 2006. Microarray applications in microbial ecology research. *Microbial ecology,* 52**,** 159-175.

GHANNOUM, M. A., JUREVIC, R. J., MUKHERJEE, P. K., CUI, F., SIKAROODI, M., NAQVI, A. & GILLEVET, P. M. 2010. Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. *PLoS pathogens,* 6**,** e1000713.

GIHRING, T. M., GREEN, S. J. & 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**,** 285-290.

GILBERT, J. A. & DUPONT, C. L. 2011. Microbial metagenomics: beyond the genome. *Annual Review of Marine Science,* 3**,** 347-371.

GILLES, A., MEGLÉCZ, E., PECH, N., FERREIRA, S., MALAUSA, T. & MARTIN, J.-F. 2011. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC genomics,* 12**,** 245.

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. 2000. Identification of Enterococcus Species and Phenotypically Similar Lactococcus andVagococcus Species by Reverse Checkerboard Hybridization to Chaperonin 60 Gene Sequences. *Journal of clinical microbiology,* 38**,** 3953-3959.

GOH, S. H., POTTER, S., WOOD, J. O., HEMMINGSEN, S. M., REYNOLDS, R. P. & CHOW, A. W. 1996. HSP60 gene sequences as universal targets for microbial species identification: studies with coagulase-negative staphylococci. *Journal of Clinical Microbiology,* 34**,** 818-823.

GOH, S. H., SANTUCCI, Z., KLOOS, W. E., FALTYN, M., GEORGE, C. G., DRIEDGER, D. & 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**,** 3116-3121.

GONZALEZ, J. M., PORTILLO, M. C., BELDA-FERRE, P. & MIRA, A. 2012. Amplification by PCR artificially reduces the proportion of the rare biosphere in microbial communities. *PLoS One,* 7**,** e29973.

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. & REGEV, A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotech,* 29**,** 644-652.

GRIFFIN, D. 1994. Fungal physiology.

GUO, R., ZHENG, N., LU, H., YIN, H., YAO, J. & CHEN, Y. 2012. Increased diversity of fungal flora in the vagina of patients with recurrent vaginal candidiasis and allergic rhinitis. *Microbial ecology,* 64**,** 918-927.

HADRYS, H., BALICK, M. & SCHIERWATER, B. 1992. Applications of random amplified polymorphic DNA (RAPD) in molecular ecology. *Molecular ecology,* 1**,** 55-63.

HAWKSWORTH, D. L. 1991. The fungal dimension of biodiversity: magnitude, significance, and conservation. *Mycological research,* 95**,** 641-655.

HAZNEDAROGLU, B. Z., YURTSEVER, D., LEFKOWITZ, J. R. & DURAN, M. 2007. Phenotypic characterization of Escherichia coli through whole-cell fatty acid profiling to investigate host specificity. *Water research,* 41**,** 803-809.

HEMMINGSEN, S. M., WOOLFORD, C., VAN DER VIES, S. M., TILLY, K., DENNIS, D. T., GEORGOPOULOS, C. P., HENDRIX, R. W. & ELLIS, R. J. 1988. Homologous plant and bacterial proteins chaperone oligomeric protein assembly.

HIBBETT, D. S., BINDER, M., BISCHOFF, J. F., BLACKWELL, M., CANNON, P. F., ERIKSSON, O. E., HUHNDORF, S., JAMES, T., KIRK, P. M. & LÜCKING, R. 2007. A higher-level phylogenetic classification of the Fungi. *Mycological research,* 111**,** 509-547.

HILL, J. E., GOH, S. H., MONEY, D. M., DOYLE, M., LI, A., CROSBY, W. L., LINKS, M., LEUNG, A., CHAN, D. & HEMMINGSEN, S. M. 2005. Characterization of vaginal microflora of healthy, nonpregnant women by chaperonin-60 sequence-based methods. *American journal of obstetrics and gynecology,* 193**,** 682-692.

HILL, J. E., PENNY, S. L., CROWELL, K. G., GOH, S. H. & HEMMINGSEN, S. M. 2004. cpnDB: a chaperonin sequence database. *Genome research,* 14**,** 1669-1675.

HILL, J. E., SEIPP, R. P., BETTS, M., HAWKINS, L., VAN KESSEL, A. G., CROSBY, W. L. & HEMMINGSEN, S. M. 2002. Extensive profiling of a complex microbial community by high-throughput sequencing. *Applied and environmental microbiology,* 68**,** 3055-3066.

HILL, J. E., TOWN, J. R. & HEMMINGSEN, S. M. 2006. 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**,** 741-746.

HOFFMANN, M., BROWN, E. W., FENG, P. C., KEYS, C. E., FISCHER, M. & MONDAY, S. R. 2010. PCR-based method for targeting 16S-23S rRNA intergenic spacer regions among Vibrio species. *BMC microbiology,* 10**,** 90.

HOFSTETTER, V., MIADLIKOWSKA, J., KAUFF, F. & LUTZONI, F. 2007. Phylogenetic comparison of protein-coding versus ribosomal RNA-coding sequence data: a case study of the Lecanoromycetes (Ascomycota). *Molecular phylogenetics and evolution,* 44**,** 412-426.

HOPKINS, M., SHARP, R. & MACFARLANE, G. 2001. Age and disease related changes in intestinal bacterial populations assessed by cell culture, 16S rRNA abundance, and community cellular fatty acid profiles. *Gut,* 48**,** 198-205.

HOUPIKIAN, P. & RAOULT, D. 2001. Molecular phylogeny of the genus Bartonella: what is the current knowledge? *FEMS Microbiology Letters,* 200**,** 1-7.

HUFFNAGLE, G. B. & NOVERR, M. C. 2013. The emerging world of the fungal microbiome. *Trends in microbiology,* 21**,** 334-341.

HUGENHOLTZ, P. 2002. Exploring prokaryotic diversity in the genomic era. *Genome Biol,* 3**,** 1-0003.8.

HUGENHOLTZ, P. & PACE, N. R. 1996. Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. *Trends in biotechnology,* 14**,** 190-197.

HUMMELEN, R., FERNANDES, A. D., MACKLAIM, J. M., DICKSON, R. J., CHANGALUCHA, J., GLOOR, G. B. & REID, G. 2010. Deep sequencing of the vaginal microbiota of women with HIV. *PloS one,* 5**,** e12078.

HUSON, D. H., AUCH, A. F., QI, J. & SCHUSTER, S. C. 2007. MEGAN analysis of metagenomic data. *Genome research,* 17**,** 377-386.

INGIANNI, A., PETRUZZELLI, S., MORANDOTTI, G. & POMPEI, R. 1997. Genotypic differentiation of Gardnerella vaginalis by amplified ribosomal DNA restriction analysis (ARDRA). *FEMS Immunology & Medical Microbiology,* 18**,** 61-66.

JAYAPRAKASH, T. P., SCHELLENBERG, J. J. & HILL, J. E. 2012. Resolution and characterization of distinct cpn60-based subgroups of Gardnerella vaginalis in the vaginal microbiota. *PLoS One,* 7**,** e43009.

KARLIN, S., WEINSTOCK, G. M. & BRENDEL, V. 1995. Bacterial classifications derived from recA protein sequence comparisons. *Journal of bacteriology,* 177**,** 6881-6893.

KASAI, H., WATANABE, K., GASTEIGER, E., BAIROCH, A., ISONO, K., YAMAMOTO, S. & HARAYAMA, S. 1998. Construction of the gyrB database for the identification and classification of bacteria. *Genome Informatics,* 9**,** 13-21.

KIRK, J. L., BEAUDETTE, L. A., HART, M., MOUTOGLIS, P., KLIRONOMOS, J. N., LEE, H. & TREVORS, J. T. 2004. Methods of studying soil microbial diversity. *Journal of microbiological methods,* 58**,** 169-188.

KISS, L., SCHOCH, C. L., SEIFERT, K. A., CALDEIRA, K., MYHRVOLD, N. P., ALVAREZ, R. A., PACALA, S. W., WINEBRAKE, J. J., CHAMEIDES, W. L. & HAMBURG, S. P. 2012. Limits of nuclear ribosomal DNA internal transcribed spacer (ITS) sequences as species barcodes for Fungi. *Proc Natl Acad Sci USA,* 109**,** 10741-10742.

KOBS, G. 1997. Cloning blunt-end DNA fragments into the pGEM®-T Vector Systems. *Promega Notes,* 62**,** 15-18.

KULLEN, M. J., BRADY, L. J. & O'SULLIVAN, D. J. 1997. Evaluation of using a short region of the recA gene for rapid and sensitive speciation of dominant bifidobacteria in the human large intestine1. *FEMS microbiology letters,* 154**,** 377-383.

L HAWKSWORTH, D. 2001. The magnitude of fungal diversity: the 1· 5 million species estimate revisited. *Mycological research,* 105**,** 1422-1432.

LANE, D. J., PACE, B., OLSEN, G. J., STAHL, D. A., SOGIN, M. L. & PACE, N. R. 1985. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proceedings of the National Academy of Sciences,* 82**,** 6955-6959.

LANGMEAD, B. & SALZBERG, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods,* 9**,** 357-359.

LAW, J. W.-F., AB MUTALIB, N.-S., CHAN, K.-G. & LEE, L.-H. 2014. Rapid methods for the detection of foodborne bacterial pathogens: principles, applications, advantages and limitations. *Frontiers in Microbiology,* 5**,** 770.

LETUNIC, I. & BORK, P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics,* 23**,** 127-128.

LI, W. & GODZIK, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics,* 22**,** 1658-1659.

LING, Z., KONG, J., LIU, F., ZHU, H., CHEN, X., WANG, Y., LI, L., NELSON, K. E., XIA, Y. & XIANG, C. 2010. Molecular analysis of the diversity of vaginal microbiota associated with bacterial vaginosis. *BMC genomics,* 11**,** 488.

LINKS, M. G., CHABAN, B., HEMMINGSEN, S. M., MUIRHEAD, K. & HILL, J. E. 2013. mPUMA: a computational approach to microbiota analysis by de novo assembly of operational taxonomic units based on protein-coding barcode sequences. *Microbiome,* 1**,** 1-7.

LINKS, M. G., DEMEKE, T., GRÄFENHAN, T., HILL, J. E., HEMMINGSEN, S. M. & DUMONCEAUX, T. J. 2014. Simultaneous profiling of seed-associated bacteria and fungi reveals antagonistic interactions between microorganisms within a shared epiphytic microbiome on Triticum and Brassica seeds. *New phytologist,* 202**,** 542-553.

LINKS, M. G., DUMONCEAUX, T. J., HEMMINGSEN, S. M. & 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**,** e49755.

LIU, Y. J., HODSON, M. C. & HALL, B. D. 2006. Loss of the flagellum happened only once in the fungal lineage: phylogenetic structure of kingdom Fungi inferred from RNA polymerase II subunit genes. *BMC Evolutionary Biology,* 6**,** 74.

LONGCORE, J. E., PESSIER, A. P. & NICHOLS, D. K. 1999. Batrachochytrium dendrobatidis gen. et sp. nov., a chytrid pathogenic to amphibians. *Mycologia***,** 219-227.

MASSOL-DEYA, A. A., ODELSON, D. A., HICKEY, R. F. & TIEDJE, J. M. 1995. Bacterial community fingerprinting of amplified 16S and 16–23S ribosomal DNA gene sequences and restriction endonuclease analysis (ARDRA). *Molecular microbial ecology manual.* Springer.

MASSON, L., MAYNARD, C., BROUSSEAU, R., GOH, S.-H., HEMMINGSEN, S. M., HILL, J. E., PACCAGNELLA, A., ODA, R. & KIMURA, N. 2006. Identification of pathogenic Helicobacter species by chaperonin-60 differentiation on plastic DNA arrays. *Genomics,* 87**,** 104-112.

MATSUKI, T., WATANABE, K., FUJIMOTO, J., MIYAMOTO, Y., TAKADA, T., MATSUMOTO, K., OYAIZU, H. & TANAKA, R. 2002. Development of 16S rRNA-gene-targeted group-specific primers for the detection and identification of predominant bacteria in human feces. *Applied and Environmental Microbiology,* 68**,** 5445-5451.

MESSING, J., CREA, R. & SEEBURG, P. H. 1981. A system for shotgun DNA sequencing. *Nucleic acids research,* 9**,** 309-321.

MOLLET, C., DRANCOURT, M. & RAOULT, D. 1997. rpoB sequence analysis as a novel basis for bacterial identification. *Molecular microbiology,* 26**,** 1005-1011.

MORA, C., TITTENSOR, D. P., ADL, S., SIMPSON, A. G. & WORM, B. 2011. How many species are there on Earth and in the ocean? *PLoS biology,* 9**,** e1001127.

MORGAN, J. A., VREDENBURG, V. T., RACHOWICZ, L. J., KNAPP, R. A., STICE, M. J., TUNSTALL, T., BINGHAM, R. E., PARKER, J. M., LONGCORE, J. E. & MORITZ, C. 2007. Population genetics of the frog-killing fungus Batrachochytrium dendrobatidis. *Proceedings of the National Academy of Sciences,* 104**,** 13845-13850.

MUYZER, G. 1999. DGGE/TGGE a method for identifying genes from natural ecosystems. *Current opinion in microbiology,* 2**,** 317-322.

NAKATSU, C. H., TORSVIK, V. & ØVREÅS, L. 2000. Soil community analysis using DGGE of 16S rDNA polymerase chain reaction products. *Soil Science Society of America Journal,* 64**,** 1382-1388.

NILSSON, R. H., KRISTIANSSON, E., RYBERG, M., HALLENBERG, N. & LARSSON, K.-H. 2008. Intraspecific ITS variability in the kingdom Fungi as expressed in the international sequence databases and its implications for molecular species identification. *Evolutionary bioinformatics online,* 4**,** 193.

O'BRIEN, H. E., PARRENT, J. L., JACKSON, J. A., MONCALVO, J.-M. & VILGALYS, R. 2005. Fungal community analysis by large-scale sequencing of environmental samples. *Applied and environmental microbiology,* 71**,** 5544-5550.

O'SULLIVAN, D. J. 2000. Methods for analysis of the intestinal microflora. *Current issues in intestinal microbiology,* 1**,** 39-50.

OLSEN, G. J., LANE, D. J., GIOVANNONI, S. J., PACE, N. R. & STAHL, D. A. 1986. Microbial ecology and evolution: a ribosomal RNA approach. *Annual Reviews in Microbiology,* 40**,** 337-365.

OPEL, K. L., CHUNG, D. & MCCORD, B. R. 2010. A Study of PCR Inhibition Mechanisms Using Real Time PCR*,†. *Journal of forensic sciences,* 55**,** 25-33.

PACE, N. R., OLSEN, G. J. & WOESE, C. R. 1986. Ribosomal RNA phylogeny and the primary lines of evolutionary descent. *Cell,* 45**,** 325-326.

PAGE, R. D. 1996. TreeView. *An application to display phylogenetic trees on personal computer. Comp Appl Biol Sci,* 12**,** 357-358.

PARK, H. K., HA, M.-H., PARK, S.-G., KIM, M. N., KIM, B. J. & KIM, W. 2012. Characterization of the fungal microbiota (mycobiome) in healthy and dandruff-afflicted human scalps. *PLoS One,* 7**,** e32847.

PAYNE, G. W., RAMETTE, A., ROSE, H. L., WEIGHTMAN, A. J., JONES, T. H., TIEDJE, J. M. & MAHENTHIRALINGAM, E. 2006. Application of a recA gene-based identification approach to the maize rhizosphere reveals novel diversity in Burkholderia species. *FEMS microbiology letters,* 259**,** 126-132.

PEIXOTO, R., DA COSTA COUTINHO, H., RUMJANEK, N., MACRAE, A. & ROSADO, A. 2002. Use of rpoB and 16S rRNA genes to analyse bacterial diversity of a tropical soil using PCR and DGGE. *Letters in applied microbiology,* 35**,** 316-320.

PORETSKY, R., RODRIGUEZ-R, L. M., LUO, C., TSEMENTZI, D. & KONSTANTINIDIS, K. T. 2014. Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PloS one,* 9**,** e93827.

RASTOGI, G. & SANI, R. K. 2011. Molecular techniques to assess microbial community structure, function, and dynamics in the environment. *Microbes and microbial technology.* Springer.

ROOSE-AMSALEG, C., GARNIER-SILLAM, E. & HARRY, M. 2001. Extraction and purification of microbial DNA from soil and sediment samples. *Applied Soil Ecology,* 18**,** 47-60.

ROWLAND, G., ABOSHKIWA, M. & COLEMAN, G. 1993. Comparative sequence analysis and predicted phylogeny of the DNA-dependent RNA polymerase beta subunits of Staphylococcus aureus and other eubacteria. *Biochemical Society transactions,* 21**,** 40S-40S.

SALZMAN, N. H., DE JONG, H., PATERSON, Y., HARMSEN, H. J., WELLING, G. W. & BOS, N. A. 2002. Analysis of 16S libraries of mouse gastrointestinal microflora reveals a large new group of mouse intestinal bacteria. *Microbiology,* 148**,** 3651-3660.

SCHELLENBERG, J., LINKS, M. G., HILL, J. E., DUMONCEAUX, T. J., PETERS, G. A., TYLER, S., BALL, T. B., SEVERINI, A. & PLUMMER, F. A. 2009. Pyrosequencing of the chaperonin-60 universal target as a tool for determining microbial community composition. *Applied and environmental microbiology,* 75**,** 2889-2898.

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. 2011. Molecular definition of vaginal microbiota in East African commercial sex workers. *Applied and environmental microbiology,* 77**,** 4066-4074.

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. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology,* 75**,** 7537-7541.

SCHLOTER, M., AßMUS, B. & HARTMANN, A. 1995. The use of immunological methods to detect and identify bacteria in the environment. *Biotechnology Advances,* 13**,** 75-90.

SCHMIDT, P.-A., BÁLINT, M., GRESHAKE, B., BANDOW, C., RÖMBKE, J. & SCHMITT, I. 2013. Illumina metabarcoding of a soil fungal community. *Soil Biology and Biochemistry,* 65**,** 128-132.

SCHOCH, C. L., SEIFERT, K. A., HUHNDORF, S., ROBERT, V., SPOUGE, J. L., LEVESQUE, C. A., CHEN, W., BOLCHACOVA, E., VOIGT, K. & CROUS, P. W. 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences,* 109**,** 6241-6246.

SCHOCH, C. L., SUNG, G.-H., LÓPEZ-GIRÁLDEZ, F., TOWNSEND, J. P., MIADLIKOWSKA, J., HOFSTETTER, V., ROBBERTSE, B., MATHENY, P. B., KAUFF, F. & WANG, Z. 2009. The Ascomycota tree of life: a phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. *Systematic biology***,** syp020.

SCHREIER, W. J., SCHRADER, T. E., KOLLER, F. O., GILCH, P., CRESPO-HERNÁNDEZ, C. E., SWAMINATHAN, V. N., CARELL, T., ZINTH, W. & KOHLER, B. 2007. Thymine dimerization in DNA is an ultrafast photoreaction. *Science,* 315**,** 625-629.

SCHÜßLER, A., SCHWARZOTT, D. & WALKER, C. 2001. A new fungal phylum, the Glomeromycota: phylogeny and evolution. *Mycological research,* 105**,** 1413-1421.

SCHWIEGER, F. & TEBBE, C. C. 1998. A New Approach To Utilize PCR–Single-Strand-Conformation Polymorphism for 16S rRNA Gene-Based Microbial Community Analysis. *Applied and Environmental Microbiology,* 64**,** 4870-4876.

SHANNON, C. 1948. A mathematical theory of communication, bell System technical Journal 27: 379-423 and 623–656. *Mathematical Reviews (MathSciNet): MR10, 133e*.

SHI, P., BAI, Y., YUAN, T., YAO, B. & FAN, Y. 2007. [Use of rpoB and 16S rDNA genes to analyze rumen bacterial diversity of goat using PCR and DGGE]. *Wei sheng wu xue bao= Acta microbiologica Sinica,* 47**,** 285-289.

SIMPSON, E. 1949. Measurement of Diversity Nature 163.•.

SINGLETON, D. R., FURLONG, M. A., RATHBUN, S. L. & WHITMAN, W. B. 2001. Quantitative comparisons of 16S rRNA gene sequence libraries from environmental samples. *Applied and environmental microbiology,* 67**,** 4374-4376.

SMIT, E., LEEFLANG, P., GLANDORF, B., VAN ELSAS, J. D. & WERNARS, K. 1999. Analysis of fungal diversity in the wheat rhizosphere by sequencing of cloned PCR-amplified genes encoding 18S rRNA and temperature gradient gel electrophoresis. *Applied and Environmental Microbiology,* 65**,** 2614-2621.

SMIT, E., LEEFLANG, P. & WERNARS, K. 1997. Detection of shifts in microbial community structure and diversity in soil caused by copper contamination using amplified ribosomal DNA restriction analysis. *FEMS Microbiology Ecology,* 23**,** 249-261.

SPINGOLA, M., GRATE, L., HAUSSLER, D. & ARES, M. 1999. Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae. *Rna,* 5**,** 221-234.

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. 2012. Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. *PloS one,* 7**,** e37818.

STAHL, D. & CAPMAN, W. 1994. Application of molecular genetics to the study of microbial communities. *In:* STAL, L. & CAUMETTE, P. (eds.) *Microbial Mats.* Springer Berlin Heidelberg.

STAHL, D. A., LANE, D. J., OLSEN, G. J. & PACE, N. R. 1984. Analysis of hydrothermal vent-associated symbionts by ribosomal RNA sequences. *Science,* 224**,** 409-411.

STALEY, J. T. & KONOPKA, A. 1985. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Reviews in Microbiology,* 39**,** 321-346.

SUAU, A., BONNET, R., SUTREN, M., GODON, J.-J., GIBSON, G. R., COLLINS, M. D. & DORÉ, J. 1999. Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Applied and environmental microbiology,* 65**,** 4799-4807.

SUNDQUIST, A., BIGDELI, S., JALILI, R., DRUZIN, M. L., WALLER, S., PULLEN, K. M., EL-SAYED, Y. Y., TASLIMI, M. M., BATZOGLOU, S. & RONAGHI, M. 2007. Bacterial flora-typing with targeted, chip-based Pyrosequencing. *BMC microbiology,* 7**,** 108.

TAYLOR, D. L., HERRIOTT, I. C., STONE, K. E., MCFARLAND, J. W., BOOTH, M. G. & LEIGH, M. B. 2010. Structure and resilience of fungal communities in Alaskan boreal forest soils This article is one of a selection of papers from The Dynamics of Change in Alaska's Boreal Forests: Resilience and Vulnerability in Response to Climate Warming. *Canadian Journal of Forest Research,* 40**,** 1288-1301.

TAYLOR, D. L., HOLLINGSWORTH, T. N., MCFARLAND, J. W., LENNON, N. J., NUSBAUM, C. & RUESS, R. W. 2014. A first comprehensive census of fungi in soil reveals both hyperdiversity and fine-scale niche partitioning. *Ecological Monographs,* 84**,** 3-20.

THOMOPSON, J., HIGGINS, D. G. & GIBSON, T. 1994. ClustalW. *Nucleic Acids Res,* 22**,** 4673-4680.

TOJU, H., TANABE, A. S., YAMAMOTO, S. & SATO, H. 2012. High-coverage ITS primers for the DNA-based identification of ascomycetes and basidiomycetes in environmental samples. *PLoS One,* 7**,** e40863.

TRINGE, S. G. & RUBIN, E. M. 2005. Metagenomics: DNA sequencing of environmental samples. *Nature reviews genetics,* 6**,** 805-814.

TYSON, G. W., CHAPMAN, J., HUGENHOLTZ, P., ALLEN, E. E., RAM, R. J., RICHARDSON, P. M., SOLOVYEV, V. V., RUBIN, E. M., ROKHSAR, D. S. & BANFIELD, J. F. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature,* 428**,** 37-43.

VAINIO, E. J. & HANTULA, J. 2000. Direct analysis of wood-inhabiting fungi using denaturing gradient gel electrophoresis of amplified ribosomal DNA. *Mycological research,* 104**,** 927-936.

VANDENKOORNHUYSE, P. & LEYVAL, C. 1998. SSU rDNA sequencing and PCR-fingerprinting reveal genetic variation within Glomus mosseae. *Mycologia***,** 791-797.

VARTOUKIAN, S. R., PALMER, R. M. & WADE, W. G. 2010. Strategies for culture of 'unculturable'bacteria. *FEMS microbiology letters,* 309**,** 1-7.

VENTER, J. C., REMINGTON, K., HEIDELBERG, J. F., HALPERN, A. L., RUSCH, D., EISEN, J. A., WU, D., PAULSEN, I., NELSON, K. E. & NELSON, W. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *science,* 304**,** 66-74.

VERBEKE, T. J., SPARLING, R., HILL, J. E., LINKS, M. G., LEVIN, D. & 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**,** 171-179.

VĚTROVSKÝ, T. & BALDRIAN, P. 2013. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS One,* 8**,** e57923.

VOS, M., QUINCE, C., PIJL, A. S., DE HOLLANDER, M. & KOWALCHUK, G. A. 2012. A comparison of rpoB and 16S rRNA as markers in pyrosequencing studies of bacterial diversity. *PLoS One,* 7**,** e30600.

WANG, Y., LI, H., JIA, S., WU, Z. & GUO, B. 2006. [Analysis of bacterial diversity of kefir grains by denaturing gradient gel electrophoresis and 16S rDNA sequencing]. *Wei sheng wu xue bao= Acta microbiologica Sinica,* 46**,** 310-313.

WATANABE, K., TERAMOTO, M., FUTAMATA, H. & HARAYAMA, S. 1998. Molecular detection, isolation, and physiological characterization of functionally dominant phenol-degrading bacteria in activated sludge. *Applied and Environmental Microbiology,* 64**,** 4396-4402.

WHITE, T. J., BRUNS, T., LEE, S. & TAYLOR, J. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *PCR protocols: a guide to methods and applications,* 18**,** 315-322.

WHITMAN, W. B., COLEMAN, D. C. & WIEBE, W. J. 1998. Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences,* 95**,** 6578-6583.

WHITTAKER, R. H. & MARGULIS, L. 1978. Protist classification and the kingdoms of organisms. *Biosystems,* 10**,** 3-18.

WILSON, I. G. 1997. Inhibition and facilitation of nucleic acid amplification. *Applied and environmental microbiology,* 63**,** 3741.

WOESE, C. R. 1987. Bacterial evolution. *Microbiological reviews,* 51**,** 221.

WOOLEY, J. C., GODZIK, A. & FRIEDBERG, I. 2010. A primer on metagenomics. *PLoS computational biology,* 6**,** e1000667.

WORRALL, J. J. 1991. Media for selective isolation of hymenomycetes. *Mycologia***,** 296-302.

YANG, Y.-H., YAO, J., HU, S. & QI, Y. 2000. Effects of agricultural chemicals on DNA sequence diversity of soil microbial community: a study with RAPD marker. *Microbial Ecology,* 39**,** 72-79.

YENERALL, P. & ZHOU, L. 2012. Identifying the mechanisms of intron gain: progress and trends. *Biol Direct,* 7**,** 29.

YEO, S. F. & WONG, B. 2002. Current Status of Nonculture Methods for Diagnosis of Invasive Fungal Infections. *Clinical Microbiology Reviews,* 15**,** 465-484.

YIN, H., CAO, L., QIU, G., WANG, D., KELLOGG, L., ZHOU, J., LIU, X., DAI, Z., DING, J. & LIU, X. 2008. Molecular diversity of 16S rRNA and gyrB genes in copper mines. *Archives of microbiology,* 189**,** 101-110.

YONG, L., YIXIN, Y., DONGYUE, Z. & WANLONG, D. 2012. Microbial community diversity analysis of Panax ginseng rhizosphere and non-rhizosphere soil using randomly amplified polymorphic DNA method. *Open Journal of Genetics,* 2012.

ZEIGLER, D. R. 2003. Gene sequences useful for predicting relatedness of whole genomes in bacteria. *International journal of systematic and evolutionary microbiology,* 53**,** 1893-1900.

ZHENG, N.-N., GUO, X.-C., LV, W., CHEN, X.-X. & FENG, G.-F. 2013. Characterization of the vaginal fungal flora in pregnant diabetic women by 18S rRNA sequencing. *European journal of clinical microbiology & infectious diseases,* 32**,** 1031-1040.

Appendices

Appendix 1: Primer bias produced by original and redesigned *cpn*60 UT primers with vaginal metagenomic DNA templates (OTU is defined as the best *cpn*60 UT reference sequence) (Figure 11).

(a) V1A

OTU	Bias Log_2(Original primers/ Redesigned primers)
b10199	0.89
b1025	-0.32
b13654	4.85
b14122	-0.77
b1432	-1.01
b15282	-1.79
b15977	1.98
b17589	0.95
b17590	-0.26
b17619	-0.02
b17630	0.57
b17634	-0.01
b18214	0.36
b18216	1.98
b18713	-1.37
b18814	-1.16
b19134	0.35
b291	-0.34
b6841	-0.002
UNKNOWN	-0.34

(b) V1B

OTU	Bias Log$_2$(Original primers / Redesigned primers)
b10199	1.58
b1025	-0.76
b13654	-2.91
b13689	-0.004
b13762	3.66
b13779	-0.85
b14122	1.61
b1432	-0.11
b14417	0.99
b14562	-0.01
b15282	-0.83
b15920	-0.09
b15977	0.07
b17590	0.43
b17619	0
b17630	1.30
b17632	3.98
b17634	-0.01
b18214	-0.23
b18216	0.58
b18280	0.41
b18713	-0.45
b18814	-0.64
b19134	0.13
b291	-0.009
b3363	1.58
b3402	-0.79
b3459	-0.81
b5843	0.84
b6817	0.58
b6839	-0.26
b6841	0.34
b6847	0.33
b7450	-0.10
UNKNOWN	-0.49

Normalized - Due to differences in library size, the actual read counts (used to calculate log2 values) are not comparable between libraries, therefore, comparisons are based on normalized counts.

Appendix 2a: Log2 values produced by original and redesigned *cpn*60 UT primers with vaginal metagenomic DNA templates. (Figure 12).

a) V1A

Isotig	Bias Log$_2$	isotig	Bias Log2
00001	1.58	00136	0.69
00010	-0.76	00142	-1.31
00014	-2.91	00143	-0.25
00022	-0.004	00150	-2.21
00028	3.66	00156	-0.01
00033	-0.85	00157	0.69
00038	1.61	00158	-1.31
00039	-0.11	00159	-0.25
00049	0.99	00166	-0.91
00050	-0.01	00170	-0.15
00051	-0.83	00182	-3.41
00053	-0.09	00189	-3.51
00060	0.07	00199	-3.82
00061	0.43	00201	1.31
00065	0	00203	1.55
00067	1.30	00207	-0.02
00069	3.98	00216	-0.24
00070	-0.01	00222	1.98
00071	-0.23	00223	-1.04
00080	0.58	00233	-1.68
00087	0.41	00239	0.99
00090	-0.45	00240	-1.08
00097	-0.64	00244	4.14
00100	0.13	00247	0.005
00103	-0.009	00250	0.89
00105	1.58	00266	4.15
00107	-0.79	00271	3.26
00108	-0.81	00274	2.98
00110	0.84	00275	2.57
00112	0.58	00277	2.21
00119	-0.26	00279	-1.34
00123	0.34		
00126	0.33		
00128	-0.10		
00131	-0.49		

b) V1B

Isotig	Bias Log2	isotig	Bias Log2
00008	0.00	00076	0.00
00010	-1.42	00080	-0.50
00012	-0.38	00082	-0.43
00016	0.41	00084	-0.01
00020	2.80	00088	-1.91
00026	-0.87	00090	0.01
00030	2.99	00092	0.00
00032	-0.55	00097	1.00
00033	-0.76	00100	-0.59
00039	-0.19	00102	Error
00041	1.37	00108	-0.01
00044	1.19	00115	0.99
00045	0.18	00118	-0.59
00046	-1.48	00120	0.70
00047	-2.01	00126	-0.59
00049	-2.58	00127	0.33
00051	-1.40	00128	0.00
00052	0.17	00131	2.52
00053	0.00	00133	-0.83
00055	0.78	00134	1.27
00056	1.00	00135	-0.01
00058	-0.42	00136	-0.69
00059	0.00	00137	0.84
00060	0.25	00138	0.03
00061	1.62	00142	0.99
00062	1.99	00143	-0.19
00063	3.98	00144	0.47
00064	-0.17	00145	-0.50
00065	0.32	00150	0.17
00066	-2.91	00152	0.58
00067	-1.00	00154	-0.26
00069	-0.83	00155	-0.11
00070	0.00	00156	0.00
00071	-0.70	00157	-1.01
00072	-0.55	00158	0.58

isotig	Bias log2	isotig	Bias log2	isotig	Bias log2
00159	0.14	00220	0.59	00273	2.48
00161	-0.81	00221	1.17	00275	0.74
00163	-1.00	00223	-0.24	00276	-0.17
00164	0.00	00224	0.51	00278	-0.01
00165	0.99	00225	-0.74	00279	1.99
00168	-0.68	00226	-0.30	00281	0.58
00172	2.99	00228	-0.60	00288	-1.32
00174	0.00	00229	-1.03	00289	0.75
00177	0.58	00230	-0.46	00293	1.00
00179	-1.00	00231	-0.42	00295	0.20
00180	1.58	00232	-0.32	00296	2.62
00181	0.00	00234	-3.33	00273	2.48
00183	0.19	00238	0.00	00275	0.74
00186	-0.49	00240	-0.21	00276	-0.17
00187	-1.75	00243	1.91	00278	-0.01
00188	0.58	00244	1.63	00279	1.99
00190	-0.42	00247	0.32	00281	0.58
00192	-0.07	00248	-1.00	00288	-1.32
00194	0.00	00249	0.73	00289	0.75
00195	0.25	00250	-0.01	00293	1.00
00197	0.68	00251	1.58	00295	0.20
00198	-2.32	00252	-0.79	00296	2.62
00200	0.12	00253	3.66		
00203	-0.25	00254	0.30		
00204	2.00	00255	0.73		
00206	-1.00	00256	-1.33		
00208	0.39	00258	-0.46		
00209	0.00	00259	-1.37		
00210	1.58	00260	-1.45		
00211	0.55	00262	0.41		
00214	-0.01	00263	0.41		
00215	0.00	00265	1.58		
00216	0.99	00267	-0.75		
00217	0.35	00269	0.55		
00218	-0.10	00270	-0.85		

Appendix 2b: Description of OTU used as labels in Figure 12.

isotig	cpnDB ID	Genbank	Genus	Species
isotig00001	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00006	b19134	NZ_AFIJ01000035	*Megasphaera*	*sp.*
isotig00007	b19134	NZ_AFIJ01000035	*Megasphaera*	*sp.*
isotig00008	b17590	AB547633	*Prevotella*	*timonensis*
isotig00010	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00011	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00012	UNKNOWN			
isotig00014	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00016	b1432	AY608421	*Lactobacillus*	*jensenii*
isotig00020	UNKNOWN			
isotig00022	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00026	UNKNOWN			
isotig00028	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00030	b18713	AEJD00000000	*Gardnerella*	*vaginalis*
isotig00032	b15920	CP001849	*Gardnerella*	*vaginalis*
isotig00033	b1025	AF440233	*Prevotella*	*bivia*
isotig00036	UNKNOWN			
isotig00038	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00039	b18214	NZ_AENT01000025	*Dialister*	*microaerophilus*
isotig00041	b17630	AB547593	*Prevotella*	*bergensis*
isotig00044	b17630	AB547593	*Prevotella*	*bergensis*
isotig00045	b15977	FJ577599	*Actinobacteria*	*sp.*
isotig00046	b1432	AY608421	*Lactobacillus*	*jensenii*
isotig00047	b15920	CP001849	*Gardnerella*	*vaginalis*
isotig00048	b17590	AB547633	*Prevotella*	*timonensis*
isotig00049	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00050	b17589	AB547634	*Prevotella*	*veroralis*
isotig00051	b18814	AFBB01000007	*Dialister*	*micraerophilus*
isotig00052	b15920	CP001849	*Gardnerella*	*vaginalis*
isotig00053	b19134	NZ_AFIJ01000035	*Megasphaera*	*sp.*
isotig00055	b17590	AB547633	*Prevotella*	*timonensis*
isotig00056	b15920	CP001849	*Gardnerella*	*vaginalis*
isotig00058	b1432	AY608421	*Lactobacillus*	*jensenii*
isotig00059	b1432	AY608421	*Lactobacillus*	*jensenii*
isotig00060	b18713	AEJD00000000	*Gardnerella*	*vaginalis*
isotig00061	b14122	ACCG01000008	*Bifidobacterium*	*breve*
isotig00062	b15977	FJ577599	*Actinobacteria*	*sp.*
isotig00063	b17632	AB547591	*Prevotella*	*amnii*
isotig00064	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00065	UNKNOWN			
isotig00066	b13654	ACGK01000047	*Atopobium*	*vaginae*
isotig00067	b291	AF240579	*Gardnerella*	*vaginalis*

isotig00068	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00069	b15282	NZ_ACLN01000008	*Lactobacillus*	*iners*
isotig00070	b15977	FJ577599	*Actinobacteria*	*sp.*
isotig00071	UNKNOWN			
isotig00072	UNKNOWN			
isotig00075	UNKNOWN			
isotig00076	b13689	CP001682	*Cryptobacterium*	*curtum*
isotig00080	b18814	AFBB01000007	*Dialister*	*micraerophilus*
isotig00082	UNKNOWN			
isotig00084	b18216	NZ_AENP01000025	*Peptoniphilus*	*harei*
isotig00085	UNKNOWN			
isotig00086	b15924	ADFR01000007	*Bulleidia*	*extructa*
isotig00087	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00088	b15920	CP001849	*Gardnerella*	*vaginalis*
isotig00089	b3384	AY123736	*Finegoldia*	*magna*
isotig00090	b17634	AB547589	*Porphyromonas*	*uenonis*
isotig00092	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00093	UNKNOWN			
isotig00097	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00100	b18713	AEJD00000000	*Gardnerella*	*vaginalis*
isotig00102	b17630	AB547593	*Prevotella*	*bergensis*
isotig00103	b18814	AFBB01000007	*Dialister*	*micraerophilus*
isotig00105	b13654	ACGK01000047	*Atopobium*	*vaginae*
isotig00107	b17590	AB547633	*Prevotella*	*timonensis*
isotig00108	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00110	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00112	b18713	AEJD00000000	*Gardnerella*	*vaginalis*
isotig00115	b17590	AB547633	*Prevotella*	*timonensis*
isotig00118	UNKNOWN			
isotig00119	b13654	ACGK01000047	*Atopobium*	*vaginae*
isotig00120	b15977	FJ577599	*Actinobacteria*	*sp.*
isotig00122	b1025	AF440233	*Prevotella*	*bivia*
isotig00123	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00126	b18713	AEJD00000000	*Gardnerella*	*vaginalis*
isotig00127	b6847	AY691286	*Prevotella*	*oris*
isotig00128	b18814	AFBB01000007	*Dialister*	*micraerophilus*
isotig00130	b15282	NZ_ACLN01000008	*Lactobacillus*	*iners*
isotig00131	b17630	AB547593	*Prevotella*	*bergensis*
isotig00133	b18713	AEJD00000000	*Gardnerella*	*vaginalis*
isotig00134	b15920	CP001849	*Gardnerella*	*vaginalis*
isotig00135	b18214	NZ_AENT01000025	*Dialister*	*microaerophilus*
isotig00136	b19134	NZ_AFIJ01000035	*Megasphaera*	*sp.*
isotig00137	b5843	AY691313	*Eubacterium*	*dolichum*
isotig00138	b17634	AB547589	*Porphyromonas*	*uenonis*
isotig00142	b17590	AB547633	*Prevotella*	*timonensis*

isotig00143	b17590	AB547633	*Prevotella*	*timonensis*
isotig00144	b17590	AB547633	*Prevotella*	*timonensis*
isotig00145	b15920	CP001849	*Gardnerella*	*vaginalis*
isotig00150	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00151	UNKNOWN			
isotig00152	b14122	ACCG01000008	*Bifidobacterium*	*breve*
isotig00154	b6839	AY691278	*Prevotella*	*buccalis*
isotig00155	b1432	AY608421	*Lactobacillus*	*jensenii*
isotig00156	b17590	AB547633	*Prevotella*	*timonensis*
isotig00157	b18713	AEJD00000000	*Gardnerella*	*vaginalis*
isotig00158	b18713	AEJD00000000	*Gardnerella*	*vaginalis*
isotig00159	b19134	NZ_AFIJ01000035	*Megasphaera*	*sp.*
isotig00161	b3459	AB071388	*Campylobacter*	*rectus*
isotig00162	UNKNOWN			
isotig00163	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00164	UNKNOWN			
isotig00165	b14417	CP002122	*Prevotella*	*melaninogenica*
isotig00166	b18713	AEJD00000000	*Gardnerella*	*vaginalis*
isotig00167	b17590	AB547633	*Prevotella*	*timonensis*
isotig00168	b15920	CP001849	*Gardnerella*	*vaginalis*
isotig00169	b1432	AY608421	*Lactobacillus*	*jensenii*
isotig00170	b19134	NZ_AFIJ01000035	*Megasphaera*	*sp.*
isotig00172	b17630	AB547593	*Prevotella*	*bergensis*
isotig00173	UNKNOWN			
isotig00174	b1432	AY608421	*Lactobacillus*	*jensenii*
isotig00176	b1432	AY608421	*Lactobacillus*	*jensenii*
isotig00177	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00179	UNKNOWN			
isotig00180	b17630	AB547593	*Prevotella*	*bergensis*
isotig00181	b17634	AB547589	*Porphyromonas*	*uenonis*
isotig00182	b18713	AEJD00000000	*Gardnerella*	*vaginalis*
isotig00183	b15920	CP001849	*Gardnerella*	*vaginalis*
isotig00184	b1432	AY608421	*Lactobacillus*	*jensenii*
isotig00185	b15977	FJ577599	*Actinobacteria*	*sp.*
isotig00186	b15920	CP001849	*Gardnerella*	*vaginalis*
isotig00187	UNKNOWN			
isotig00188	b1432	AY608421	*Lactobacillus*	*jensenii*
isotig00189	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00190	b15920	CP001849	*Gardnerella*	*vaginalis*
isotig00191	b6839	AY691278	*Prevotella*	*buccalis*
isotig00192	UNKNOWN			
isotig00194	b1432	AY608421	*Lactobacillus*	*jensenii*
isotig00195	b1432	AY608421	*Lactobacillus*	*jensenii*
isotig00197	b1432	AY608421	*Lactobacillus*	*jensenii*
isotig00198	b18814	AFBB01000007	*Dialister*	*micraerophilus*

isotig00199	b18713	AEJD00000000	*Gardnerella*	*vaginalis*
isotig00200	b1432	AY608421	*Lactobacillus*	*jensenii*
isotig00201	UNKNOWN			
isotig00203	b18214	NZ_AENT01000025	*Dialister*	*microaerophilus*
isotig00204	b10199	EF571590	*Lactobacillus*	*gasseri*
isotig00205	UNKNOWN			
isotig00206	b15920	CP001849	*Gardnerella*	*vaginalis*
isotig00207	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00208	b1432	AY608421	*Lactobacillus*	*jensenii*
isotig00209	b17619	AB547604	*Prevotella*	*disiens*
isotig00210	UNKNOWN			
isotig00211	b1432	AY608421	*Lactobacillus*	*jensenii*
isotig00212	UNKNOWN			
isotig00214	UNKNOWN			
isotig00215	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00216	b6841	AY691280	*Prevotella*	*denticola*
isotig00217	b17590	AB547633	*Prevotella*	*timonensis*
isotig00218	b7450	AY562570	*Lactobacillus*	*crispatus*
isotig00219	b13606	NC_013203	*Atopobium*	*parvulum*
isotig00220	b15977	FJ577599	*Actinobacteria*	*sp.*
isotig00221	b1432	AY608421	*Lactobacillus*	*jensenii*
isotig00222	b18216	NZ_AENP01000025	*Peptoniphilus*	*harei*
isotig00223	b18814	AFBB01000007	*Dialister*	*micraerophilus*
isotig00224	b1432	AY608421	*Lactobacillus*	*jensenii*
isotig00225	UNKNOWN			
isotig00226	b1432	AY608421	*Lactobacillus*	*jensenii*
isotig00227	b15977	FJ577599	*Actinobacteria*	*sp.*
isotig00228	UNKNOWN			
isotig00229	UNKNOWN			
isotig00230	UNKNOWN			
isotig00231	b17634	AB547589	*Porphyromonas*	*uenonis*
isotig00232	UNKNOWN			
isotig00233	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00234	UNKNOWN			
isotig00235	b17590	AB547633	*Prevotella*	*timonensis*
isotig00238	UNKNOWN			
isotig00239	b19134	NZ_AFIJ01000035	*Megasphaera*	*sp.*
isotig00240	b19134	NZ_AFIJ01000035	*Megasphaera*	*sp.*
isotig00241	UNKNOWN			
isotig00242	b13654	ACGK01000047	*Atopobium*	*vaginae*
isotig00243	b17590	AB547633	*Prevotella*	*timonensis*
isotig00244	b19134	NZ_AFIJ01000035	*Megasphaera*	*sp.*
isotig00247	b6841	AY691280	*Prevotella*	*denticola*
isotig00248	b15977	FJ577599	*Actinobacteria*	*sp.*
isotig00249	b18216	NZ_AENP01000025	*Peptoniphilus*	*harei*

isotig00250	b10199	EF571590	*Lactobacillus*	*gasseri*
isotig00251	b3363	AY123698	*Aerococcus*	*urinae*
isotig00252	b3402	AY123679	*Mobiluncus*	*curtisii*
isotig00253	b13762	NZ_ACKW01000052	*Mobiluncus*	*mulieris*
isotig00254	b19134	NZ_AFIJ01000035	*Megasphaera*	*sp.*
isotig00255	UNKNOWN			
isotig00256	b1432	AY608421	*Lactobacillus*	*jensenii*
isotig00258	UNKNOWN			
isotig00259	UNKNOWN			
isotig00260	UNKNOWN			
isotig00262	UNKNOWN			
isotig00263	b18280	AEPD01000033	*Prevotella*	*buccae*
isotig00265	UNKNOWN			
isotig00266	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00267	b15920	CP001849	*Gardnerella*	*vaginalis*
isotig00268	UNKNOWN			
isotig00269	UNKNOWN			
isotig00270	b13779	NZ_ACGU01000113	*Lactobacillus*	*ultunensis*
isotig00271	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00273	b17630	AB547593	*Prevotella*	*bergensis*
isotig00274	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00275	b17590	AB547633	*Prevotella*	*timonensis*
isotig00276	b1432	AY608421	*Lactobacillus*	*jensenii*
isotig00277	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00278	b14562	GG698804	*Lactobacillus*	*coleohominis*
isotig00279	b291	AF240579	*Gardnerella*	*vaginalis*
isotig00280	UNKNOWN			
isotig00281	b6817	AY691256	*Eubacterium*	*ventriosum*
isotig00284	UNKNOWN			
isotig00286	b18394	ACWN01000067	*Eggerthella*	*sp.*
isotig00288	UNKNOWN			
isotig00289	b17590	AB547633	*Prevotella*	*timonensis*
isotig00293	b15920	CP001849	*Gardnerella*	*vaginalis*
isotig00295	b15920	CP001849	*Gardnerella*	*vaginalis*
isotig00296	UNKNOWN			

Appendix 3a: Normalized read counts associated with OTU used for production of bacterial profiles shown in Figure 13a.

OTU (cpnDB ID)	Number of reads (normalized)	
	Redesigned primers	Original primers
b291	45.49	35.88
b17590	8.31	6.94
b19134	21.27	27.05
b18713	10.74	4.15
b6841	2.41	2.41
b17589	2.26	4.37
b18814	2.33	1.03
b1025	1.57	1.25
b18214	1.41	1.81
b14122	0.66	0.39
b15282	0.63	0.19
b13654	0.43	12.28
b10199	0.18	0.33
b17619	0.03	0.03
b7450	0.02	0
b18394	0.02	0
b17630	0.02	0.03
b15920	0.02	0.01
b1432	0.01	0.01
b17634	0.01	0.01
b6839	0.01	0
b15977	0.01	0.04
b18216	0.01	0.04
b6847	0	0
b13779	0	0
b3402	0	0
b3459	0	0
b5843	0	0
b6817	0	0
b17632	0	0
b3384	0	0
b13762	0	0
b13689	0	0
b18280	0	0
b14562	0	0
b3363	0	0
b14417	0	0.01
b13606	0	0
b15924	0	0

Appendix 3b: Normalized read counts associated with OTU used for production of bacterial profiles shown in Figure 13b.

OTU cpnDB ID	Number of reads (normalized)	
	Redesigned primers	Original primers
b291	2.34	2.29
b17590	5.89	8.07
b19134	1.25	1.36
b18713	1.31	0.96
b6841	0.55	0.69
b17589	0	0
b18814	1.44	0.92
b1025	0.25	0.14
b18214	0.46	0.39
b14122	1.78	5.44
b15282	0.18	0.11
b13654	0.39	0.08
b10199	0.03	0.1
b17619	0.02	0.02
b7450	0.17	0.16
b18394	0	0
b17630	4.37	10.68
b15920	4.91	4.61
b1432	34.84	32.24
b17634	2.86	2.84
b6839	0.41	0.36
b15977	1.26	1.33
b18216	0.04	0.07
b6847	1.04	1.3
b13779	0.4	0.22
b3402	0.21	0.12
b3459	0.08	0.04
b5843	0.06	0.1
b6817	0.04	0.07
b17632	0.04	0.7
b3384	0.04	0
b13762	0.03	0.42
b13689	0.03	0.03
b18280	0.03	0.04
b14562	0.01	0.01
b3363	0.01	0.03
b14417	0.01	0.02
b13606	0	0.07

b15924	0	0.01

Appendix 3c: Normalized read counts and range of identity associated with b291 like OTU, used to represent clustered b291 in Figure 13a and 13b.

Isotigs clustered as b291-like	Normalized count	% identity to b291
isotig00010	3298.63	98
isotig00049	400.72	98
isotig00097	208.19	99
isotig00067	164.54	95
isotig00233	152.23	98
isotig00150	71.64	99
isotig00279	55.97	98
isotig00189	50.37	97
isotig00001	41.41	98
isotig00022	21.27	97
isotig00038	15.67	98
isotig00108	14.55	96
isotig00110	10.07	98
isotig00087	7.84	98
isotig00028	5.6	98
isotig00123	5.6	98
isotig00274	5.6	99
isotig00207	4.48	98
isotig00271	3.36	98
isotig00277	3.36	99
isotig00014	2.24	98
isotig00092	2.24	94
isotig00068	1.12	92
isotig00266	1.12	98

Appendix 4a: Normalized read counts associated with OTU used for production of *G. vaginalis* profiles shown in Figure 14a and 15a.

OTU	Number of reads (normalized)	
	Redesigned primers	Original primers
001b291	0.41	1.54
010b291	32.99	3.99
022b291	0.21	0.19
028b291	0.06	0.11
032b15920	0	0
038b291	0.16	0.08
047b15920	0	0
049b291	4.01	24.13
052b15920	0	0
064b291	0	0
067b291	1.65	0.91
087b291	0.08	0.25
088b15920	0	0
092b291	0.02	0
097b291	2.08	2.09
100b18713	1.35	0.89
112b18713	0.37	0.16
123b291	0.06	0.29
133b18713	0.13	0.03
134b15920	0	0
145b15920	0	0
150b291	0.72	0.16
157b18713	0.75	0.49
158b18713	2.71	0.4
163b291	0	0
166b18713	0.44	0.23
168b15920	0	0
177b291	0	0.01
182b18713	0.24	0.02
183b15920	0	0
186b15920	0	0
189b291	0.5	0.04
190b15920	0	0
199b18713	0.16	0.01
206b15920	0.01	0
207b291	0.04	0.04
215	0	0
233b291	1.52	0.48

266b291	0.01	0.2
267b15920	0	0
271b291	0.03	0.32
274b291	0.06	0.44
277b291	0.03	0.16
279b291	0.56	0.22
293b15920	0	0
295b15920	0	0.01
30b18713	0	0
56b15920	0.01	0
60b18713	4.57	1.92

Appendix 4b: Normalized read counts associated with OTU used for production of *G. vaginalis* profiles shown in Figure 14b and 15b.

OTU	Number of reads (normalized)	
	Redesigned primers	Original primers
001b291	0.01	0
010b291	0.09	0.03
022b291	0	0.01
028b291	0.01	0
032b15920	0.49	0.33
038b291	0	0
047b15920	0.04	0.01
049b291	0.13	0.02
052b15920	1.74	1.96
064b291	0.41	0.37
067b291	0.04	0.02
087b291	0	0
088b15920	0.17	0.04
092b291	0.13	0.13
097b291	0.03	0.07
100b18713	0.03	0.02
112b18713	0.01	0
123b291	0	0
133b18713	0.07	0.04
134b15920	0.21	0.51
145b15920	0.99	0.7
150b291	1.23	1.38
157b18713	0.02	0.01
158b18713	0.02	0.03
163b291	0.07	0.03
166b18713	0	0
168b15920	0.09	0.06
177b291	0.07	0.1
182b18713	0	0
183b15920	0.16	0.18
186b15920	0.16	0.11
189b291	0	0
190b15920	0.04	0.03
199b18713	0	0.01
206b15920	0.09	0.04
207b291	0	0
215	0.07	0.07

233b291	0.02	0
266b291	0	0
267b15920	0.52	0.31
271b291	0	0
274b291	0	0
277b291	0	0
279b291	0.01	0.04
293b15920	0.03	0.07
295b15920	0.14	0.17
30b18713	0.01	0.09
56b15920	0.04	0.09
60b18713	0.18	0.21