2.1 Antarctica

2.1.1 History

Antarctica has been the target of many scientific expeditions for almost a century since its first official discovery in 1911 by a Norwegian explorer named Roald Amundsen. It all started a few centuries before where the existence of an unknown southern "Cold Land" led numerous explorations due south of the globe (Ward, 2001). The first theory of Antarctica came from the ancient Greeks. Knowing about the Arctic, they believed in order to balance the world there must be another cold landmass opposite of it in the far south end of the world (Rubin, 2005). Early in 15^{th} century the penetration into the southern hemisphere began by numerous explorers from different countries.

It was not before early 19^{th} century when the first sighting of the continent Antarctica was made by a Russian naval officer, Captain Thaddeus Bellingshausen. From late 19^{th} century to early 20^{th} century, mainly sealers and whalers lead various expeditions to all parts of Antarctica. Eventually after several science, geography and exploration driven expeditions, Norwegian Roald Amundsen managed to reach the South Pole for the first time (Ward, 2001). For five decades after this discovery, more attempts to cross the continent both aerial and terrestrial were made and continues to this date.

Since Antarctica is the only continent on Earth with no native human population, a treaty came into effect by 1961. This agreement called the Antarctic Treaty System or ATS regulates international relations with respect to Antarctica. As a result, Antarctica, defined as all land and ice shelves south of 60^{o}S latitude, is set aside as a scientific preserve, establishing freedom of scientific investigation and banned military activity in the continent. Originally 12 nations started the international cooperation in Antarctica

and agreed on the non-national treaty. Currently there are 46 nations as treaty members (ATS, 2008).

2.1.2 Geography and Climate Conditions

Antarctica is the southernmost continent in the Antarctic region of southern hemisphere surrounded by the Southern Ocean (Fig. 2.1). It is secluded from the other Southern Hemisphere continents and has one of the most extreme terrestrial environments on Earth. It rates as the fifth largest continent at an area of about 14.0 million km^2 of which 98% is covered in ice sheets averaging 1.6 km in thickness (The World Factbook, retrieved on April 2010). Antarctica contains 90% of world's ice and 70% of its fresh water supply.

Figure 2.1 Map of Antarctica continent (*retrieved from* http://www.map-of-antarctica.us/**,** *on May 3, 2010*)

The overall precipitation rates average about 166mm per year over the entire continent. Mainly, the precipitation falls as snow at different rates, ranging from meters to as low as 50mm per year (Rubin, 2005). With the temperature remaining mostly below freezing point throughout the year, the snow on major parts of the land rarely melts, compressing to become the glacial ice. The dry air combined with extremely low temperatures cause a very low absolute humidity.

Temperature ranges along the land with latitude, elevation and distance from the ocean. There are three distinct climatic regions in Antarctica: the interior being the coldest (average annual temperature -57℃), and the coastal regions and the Antarctic Peninsula being the more moderate with temperatures as high as 15℃. The lowest naturally occurring temperature on Earth was -89.2℃ recorded at the Vostok research station in July, 1983 (Hutchings, 1998).

Environmental stresses including extremely low temperature, desiccation and levels of solar UV radiation are far more severe than any other counterpart in the world. Within the continent, these stress factors increase in severity from northern Antarctic Peninsula southwards, providing a gradient of increasing stress for the biota of terrestrial habitats (Brinkmann *et al.*, 2007).

However, Antarctica was not always this isolated and inhospitable throughout time. According to geological records, about 200 million years ago Antarctica was joined to South America, Africa, India, Australia, and New Zealand in one large continent (Rubin, 2005). The climate was much warmer with trees and large animals flourishing the land. The proof to Antarctica's temperate past lies within the fossils and coal beds found along with other geological formations. Thus, much can be found about the history of Earth by studying this latest hotspot of the planet.

2.1.3 Flora and Fauna

Due to the extreme climate conditions of Antarctica, vegetation is a luxury that cannot be afforded around most of the land. This is a result of combined harsh environmental forces including freezing temperatures, poor soil quality, very low water content, and not enough sunlight for the plant growth (Aislabie *et al.*, 2006). Therefore, the flora of the continent is mainly made up of lichens, algae, fungi, and bryophytes. There are two species of flowering plants found in the Antarctic Peninsula: *Deschampsia antarctica* (Antarctic hair grass) and *Colobanthus quitensis* (Antarctic pearlwort) (Australian Antarctic Division, 2010).

Same scenario applies for the fauna of Antarctica continent. There are few terrestrial vertebrates found in Antarctica, while a variety of marine animals such as penguins, blue whales, orcas, colossal squids and fur seals live and breed there. Some birds including the flightless midge *Belgica antarctica* and the Snow Petrel breed in Antarctica as well (Asselbergs, 1999). In addition, simpler life forms have adapted to the extreme environment of the South Pole, complementing the ecosystem of Antarctica. Microscopic mites like the *Alaskozetes antarcticus*, lice, nematodes, tardigrades, rotifers, krill and springtails are among the invertebrates living in Antarctica. Multicolored snow algae and diatoms are abundant in coastal regions particularly during summer (NERC-BAS, 2007).

As with any other environment, human impact on the natural biota of the continent is inevitable particularly in coastal regions of sub-Antarctic and maritime Antarctica where terrestrial ecosystems are well-developed (Convey, 2010). Other than a modest number of researchers and staff staying in research stations based in several parts of the continent, tourist attraction is estimated to bring in about 30,000 individuals per year (Convey, 2010). This has obviously altered the original biodiversity in target sites (example of natural versus human-assisted colonization given by Convey, 2008).

2.1.4 Microbial Diversity

A key component in every ecosystem is its microbial community, knowing that they are the most diverse and the most abundant of all organisms on Earth (Fierer and Jackson, 2006). Antarctica's extremely low temperatures, low humidity, constant freeze-thaw cycles and soil salinity in most parts causes almost no plant diversity, in which case microorganisms become dominant life forms throughout the land (Aislabie *et al.*, 2006). This indicates that majority of biological activities of the ecosystem are controlled by microorganisms (Simmons, 2009).

In the last two decades much work has been done to indicate the microbiota diversity of Antarctic ecosystems, from aquatic environments to simpler terrestrial ecosystems (Aislabie *et al.*, 2008). Basically the main microorganisms introduced so far are a limited community of mosses, lichens, and mats of algae and cyanobacteria near lakes and ephemeral streams. Furthermore, bacteria, yeasts and molds can be found among the mats as well, while even less diverse communities exist in the exposed arid soils occupying most of the ice-free land (Simmons, 2009).

The understanding of Antarctic soil microbiology has been entirely reliant on culture-based community studies until recent years (Smith *et al.*, 2006). Then again, it is argued that less than 10% of microbial species are readily isolated via current culturing techniques. Alternatively, modern molecular techniques come in handy to construct a more thorough outline of microbial diversity, especially in case of Antarctica's peculiar mode of natural selection and a currently scattered set of info on the issue. To begin with, there is one critical topic to look into; what biodiversity is and why it is so important to keep track of microbial diversity in a land "discovered" barely two centuries ago with spectacularly unusual conditions.

2.2 Biodiversity; an Overview

By definition, biodiversity is a term implying existence of a wide range of species or variety of organisms that lies in the area of phylogeny (Colwell *et al.*, 1995). On the other hand, a restricted diversity consisting of functionally key species is of great significance as well. In comprehension, certain species duelling under selective environmental pressures or biogeographical isolation have unique and perhaps extraordinary characteristics allowing them to survive (Wynn-Williams, 1996). Under these circumstances, biodiversity includes a variety of biochemical, physiological and strategic attributes as well as species diversity contributing to extreme environments, i.e. Antarctic ecosystems (Wynn-Williams, 1996).

Before the striking application of molecular techniques to relate organisms, culturing methods were mainly applied to identify bacteria. The methods of identification were solely based on physiological traits and metabolism pathways (Hunter-Cevera, 1998). However, as it is now well understood, this approach is destined to failure in terms of painting the entire picture. Not only most bacteria are unculturable, there are those microorganisms that although culturable in certain conditions, have the tendency to become dormant and unculturable under different environmental conditions (Hunter-Cevera, 1998). Furthermore, the use of physiological traits as an indication tool for different groups of bacterial species has limitations of its own (Woese, 1987).

Consequently, the culturing conditions can only select for a limited subpopulation of bacteria in the environment and thus the resulting diversity measurements are based solely on the characteristics of those isolated bacteria (Øvreås, 2000). The outcome of this short-sighted analysis was the assumption of multicellular eukaryotes as the most diverse life forms on planet until late 60’s (Pace, 1997).

Once the molecular sequence-based phylogenetic tree was established by Carl Woese (Woese and Fox, 1977; Woese, 1987), the main evolutionary diversity of life was found to lie ironically among microorganisms, particularly bacteria (Pace, 1997). This is a very well established fact regarding the high degree of adaptability of microbial communities and their ability to colonize almost any environmental niche (Øvreås, 2000). Among all, soil bacteria attract more attention due to their enormous numbers and vital roles in the terrestrial ecosystems as will be discussed in detail later on.

2.3 Significance of Biodiversity

Perhaps it is best to elucidate more on the benefits of microbial diversity in our daily lives to better justify the need for a constant update of our knowledge about these important yet primarily neglected forms of life.

2.3.1 Story of Evolution

Life began on the Earth with the occurrence of microbial life some 3.8 billion years ago in an environment simply uninhabitable to most organisms today. These conditions included an oxygen-free world with hot and acidic water bodies and an atmosphere incapable of protecting against the damaging ultraviolet radiations (Sogin, 2007). It was a physiologically challenging world, but one in which microbes survived and succeeded, and through their succession they began to change the world into places that are less stressed.

The primary microbes (mostly Archaea) had an extremely important role in creating habitability over billions of years and therefore, an essential role in the evolution of life (Sogin, 2007). Having evolved into enormous diversity during this

extensive period of time, the prokaryotic diversity is believed to be very high with an estimation of 10^6 for bacteria (Tiedje and Stein, 1999).

By looking further into the diversity of bacteria present today, much has been discovered about the history of life on Earth. In words of Carl Woes, who in a sense started the molecular revolution in bacterial evolution, "the cell is basically an historical document, and gaining the capacity to read it (by the sequencing of genes) cannot but drastically alter the way we look at all of biology" (Woese, 1987). In a way, as the oldest occupants of the planet, the bacterial cells simply offer a gateway to tell the story of the starting days of life from a scientific point of view.

Furthermore, discovery of microbial biodiversity expands the frontiers of knowledge about the strategies and limits of life, especially microbes that live at extreme conditions (Tiedje and Stein, 1999). The theory of extra-terrestrial (ET) life can be backed up by case reports such as the Greenland glacier survivor (Loveland-Curtze *et al.*, 2009). If life can exist against all odds in the core of a 120,000 year-old glacier with practically no food, no air and constant freezing temperature, hope remains in turning ET sci-fi into reality specifically where there are potential analogues for environmental conditions on Mars found in Dry Valleys of Antarctica (de la Torre *et al.*, 2003).

2.3.2 The Circle of Life

The microbial diversity has several important values to society and to the Earth's ecosystem. From an Earth-centric perspective, all of the biosphere's diverse multi-cellular life forms including plants and animals are totally dependent upon the microbial world for their continued survival. Not only do they shape the environment, some microbial communities reside within multi-cellular hosts where they contribute

intermediate metabolites, metabolic functions and microbial genes to the host organism (Sogin, 2007).

Soil microorganisms in particular are of critical importance to the sustainability of life on the planet. According to the authors of "Biodiversity of Microbial Life", soil biodiversity is no less exotic than that of a "rainforest" that lies underground and supports the aboveground systems. The complex microbial community in terrestrial ecosystems plays a central role in catalyzing biochemical reactions in soil as well as a considerable effect on atmospheric chemistry and global climate by influencing gases such as CO_2, CH_4, H_2, N_2O, and NO (Buckley and Schmidt, 2002).

Most of the recycling processes in nature are carried out by the microbial components of Earth's terrestrial biomass which are enormous in numbers. Soil microbes contribute to break down of a wide range of organic compounds from natural to synthetic sources. Carbon and nitrogen linkage seen in natural cycles is partly due to the activity of microbes in soil. The sole players in nitrogen fixation processes are prokaryotes, contributing to the productivity of terrestrial ecosystems. Moreover, soil bacteria mineralize nutrients and supply the carbon-limited environment with carbon input from plant decomposition (Buckley and Schmidt, 2002).

The overall physical and chemical structure of soil depends on the microbial transformations of organic compounds. To be specific, humic compounds and polysaccharides which are the by-products of microbial biomass, play a major role in forming soil aggregates. The microbial activity in soil aggregates influences a series of reactions from methane production to denitrification since it helps oxygen distribution in soils. Additionally, water infiltration and water-holding capacity of the soil is directly related to these aggregates, altering the rate of microbial processes in cycling of nutrients (Buckley and Schmidt, 2002).

2.3.3 Biotechnology Perspective

Biotechnology-based industries have made it their business to search for exploitable biology, or in other terms, genetic resources which is defined as "genetic material of actual or potential value" (Bull, 2004). Here is the basic outline of a search and discovery program in biotechnology: find and collect the appropriate genetic resources, screen for a useful and novel property and ultimately, develop a commercialized product or process (Bull, 2004).

Primarily, the scope of biotechnologically valuable industries revolved around pharmaceuticals and involved only a small range of biota. However in the recent decade, the effort to globalize biotechnology has led to a broader range of industries and thus, embracing more diverse biota with more focus on microorganisms (Hunter-Cevera, 1998; Bull, 2004). Microbial diversity represents the largest unexploited pool of biodiversity for potential discovery of new biotechnology products such as new pharmaceuticals, enzymes, specialty chemicals or even new organisms that carry out novel processes (Tiedje and Stein, 1999).

As mentioned earlier, an estimation of 10^6 has been suggested for bacterial diversity. However, there is only about 5000 species of bacteria identified and described so far (Tiedje and Stein, 1999). The rest of these non-extrapolated orphans are still out there, with many possible advantages in each one of them. The number of golden opportunities hidden in each individual species is of great value, let alone a whole new genera or even families of novel bacteria (Tiedje and Stein, 1999). Therefore, as long as exploitable microbiology is unexhausted, especially in many soil samples (Lazzarini *et al.*, 2000), the hunt is on.

Recently, biotechnologists are searching high and low for microorganisms that can function in unusual conditions to enable the creation of new processes and

industries (Horikoshi, 1995). To find such microorganisms, the researchers have to select a field that would be otherwise totally neglected, meaning a new and unlikely environment with conditions beyond those ever inhabited by any living creature. In such extreme environments it is more likely to find novel species capable of managing to survive by all means necessary and coming up with amazingly innovative ways to best tolerate the harsh conditions they thrive in. From a biotechnological perspective, the goal is to find potential resources of novel microorganisms producing novel products including antibiotics, enzymes, biologically active substances, polysaccharides, etc. (Horikoshi, 1995).

2.4 Scale Estimation

It is quite fascinating to think about how far the horizons of bacterial diversity stretch. Though, the challenge of presenting a unanimous verdict on this vast diversity can be gruelling. The very first problem lies in the search target itself, i.e. the bacterial species, which to this date remains only vaguely and arguably defined (Bull and Stach, 2004; Konstantinidis *et al.*, 2006; Wilkins, 2010). Currently, bacterial speciation has raised a lot of controversy in the microbiological community, with little consensus reached in how to identify and categorize the species present (Kirk *et al.*, 2004).

A few other issues arise when studying microbial diversity, particularly in soil samples. Spatial heterogeneity of soil, inability to culture soil microorganisms, limitations of molecular-based methods, and taxonomic ambiguity of microbes are some of these problems explained in detail by Kirk *et al.* (2004). Yet attempts to assess microbial diversity in any given niche have always been made and new techniques are introduced constantly. These techniques group within two main categories: culture-dependent and culture-independent.

2.4.1 Culture-Dependent Techniques

Traditionally, microbiologists had to revive as much living cells as possible from a given sample in order to study the microbial diversity of the respective environment. These techniques, referred to as culture-dependent, have been in use for over 150 years (Fry, 2004). The concept is rather simple; selective enrichment and growing on agar plates containing general or selective media, which has led to a wide range of pure cultures from various natural habitats as well as artificial environments. After isolation, the bacteria would be further investigated for their physiology, described and named accordingly and put within a taxonomic framework (Fry, 2004). A direct viable count of the colonies formed in the first stage of this process can be used in biodiversity assessment (Kirk *et al.*, 2004).

As simple as this might sound, there are many limitations with this outline for assessing bacterial diversity. The basic problem, aside from unculturable nature of majority of the population, is the fact that more dominant bacteria seem to grow slower compared to their less abundant counterparts on artificial media. This flaw introduces an initial bias in the diversity by isolating only the faster-growing, less numerous populations (Fry, 2004).

The next shortcoming of this technique is in the identification procedure applied. It is customary in the culture-dependent techniques to study the morphological and physiological properties of the pure cultures isolated. If one is to look at physical characteristics to correlate individuals, as it is done with higher forms of life, given a single cell with no complicated features is not much of a help. Looking further at the cells' physiology and biochemical/metabolic properties can be advantageous but very misleading since relatives lacking certain traits have been reported. Not to mention

bacterial cells have a tendency to behave differently under different environmental conditions, causing even more trouble with their accurate identification (Woese, 1987).

2.4.2 Culture-Independent Techniques

Limitations of cultivation-based techniques for isolation and identification of bacteria, described as the "great plate anomaly" by Staley and Konopka (1985), forced microbiologists to divert their attention towards molecular biomarkers for identifying and categorizing bacteria (Buckley and Schmidt, 2002; Smalla, 2004). These are usually molecules such as lipids, proteins, or nucleic acids which reveal some phenotypic or genotypic information about the microorganism of origin (Buckley and Schmidt, 2002). Lipids and proteins composition tend to change more vigorously in response to physiologic conditions of the cell and environmental factors, while nucleic acids are less prone to such instability.

2.4.2.1 DNA Extraction

With the rapid development of nucleic acid extraction methods directly from environmental samples and sequencing them to represent available bacterial genome in the sample, new horizons for studying microbial communities were illuminated (Smalla, 2004). Working with genotypic information is advantageous to the classical approach of relating and identifying organisms based on their phenotype (Woese, 1987). As the author explains, three-dimensional phenotypic patterns, as opposed to one-dimensional genotypic sequences, are more complicated to measure and very subjective to interpret. Whereas the elements of a sequence are well-defined and easily quantized mathematically, making their interpretation easier and somewhat foolproof (Woese, 1987).

There are several factors involved in the quality of the extracted DNA including the handling of sample collection and the subsequent treatment of microbial cells for

DNA extraction. If the methods for cell disruption are not sufficient or lead to preferential cell lysis, it can cause a bias in the view of microbial diversity composition. On the other hand, harsh treatments required for lysis of Gram-positive bacteria may result in a highly fragmented nucleic acids from Gram-negative cells. Furthermore, various biotic and abiotic components of environmental ecosystems, such as inorganic particles or organic matter, affect cell lysis efficiency and might interfere with subsequent DNA purification and enzymatic steps (Wintzingerode *et al.*, 1997).

The most commonly applied method for quantifying DNA samples is by routine absorbance reading from a UV spectrophotometer. The photo-detector in the spectrophotometer measures the ultraviolet light density that passes through the sample. The amount of light absorbed by the sample is directly proportional to its concentration; i.e. the higher the concentration, the higher the absorbance value (Tataurov *et al.*, 2008).

Since nucleic acids have a maximum light absorption at 260nm wavelength while proteins and single stranded DNA/RNA absorb maximally at 280nm, it is practical to take both readings and calculate the amount of pure dsDNA in presence of contaminants. Therefore, a ratio of A_{260}/A_{280} is calculated to approximate the purity of the DNA sample. This value is best between 1.8 and 2.0 for a cuvette spectrophotometer (DNA Quantification, 2009).

Another concern regarding interference with DNA absorbance at 260nm is the presence of organic compounds such as phenols (used in DNA extraction procedure) and other aromatic compounds, which absorb light of 230nm wavelength (Luebbehusen, 2004). Likewise, a ratio of A_{260}/A_{230} can be calculated to quantify the presence of such contaminating agents, when best results are achieved with values greater than 1.5 (DNA Quantification, 2009).

Once the nucleic acids are obtained from environmental samples, further information about the structural composition of microbial communities and the available functional genes in that community can be achieved (Smalla, 2004). There are several approaches to study cultivation-independent, molecular microbial diversity. For a detailed discussion on some of these techniques, refer to the review by Kirk *et al.* (2004). A summary of these methods with their advantages and disadvantages is tabulated by the author available in Appendix A (Table 1S).

2.4.2.2 Pursuing the Gene

Amongst all molecular-based techniques for investigating microbial diversity, the PCR-based methods have been favoured by researchers. This is perhaps due to the relatively simple and fast nature of the process. The general layout includes amplifying any desired gene for sequencing and identifying the source organism based on the sequence similarity to previously known organisms; or in case of a novel entity, finding the phylogenetic relationship between that sequence and others in order to categorize it accordingly (Kirk *et al.*, 2004).

It all started when Woese and others showed that phylogenetic relationships of bacteria, and all life forms for that matter, can be conducted by comparing a stable part of the genetic code (Woese and Fox, 1977; Woese *et al.*, 1985; Woese, 1987). This new movement, following an earlier paper by Zuckerkandl and Pauling (1965), successfully launched biology into the world of "molecular chronometers" (a molecule whose sequence changes randomly in time). A molecular chronometer needs to fulfil a few criteria in order to be considered a suitable candidate for phylogenetic studies, and the ultimate choice for this purpose is found to be ribosomal RNA (rRNA) molecules (Woese, 1987).

The genes that encode rRNA have a slow rate of evolutionary change and are highly conserved among all cellular life forms. Furthermore, the nucleic acid sequence and secondary structure of rRNA includes conserved regions found in all living organisms as well as variable domains containing group-specific sequence motifs which act as indicative factors (Buckley and Schmidt, 2002; Clarridge, 2004). Therefore, the rRNA-encoding genes have proved to be very promising as biomarkers for studying microbial diversity in natural environments.

From the several segments of rRNA molecule, the genes that code for the small-subunit (ssu) rRNA (i.e. 5S and 16S) and the 23S rRNA have been used more extensively for identifying bacteria (Clarridge, 2004). 16S and 23S rRNA genes are large enough to portray informative phylogenetic positions for reasonable characterization and identification of bacteria species (Kolbert and Persing, 1999). The most commonly used sequence for bacterial taxonomy studies is the 16S rRNA gene which is comparable among all bacteria as well as the archeobacterial 16S rRNA gene and the eukaryotic 18S rRNA gene (Clarridge, 2004). This has allowed for a molecular sequence-based phylogenetic tree relating all organisms and reconstructing the history of life (Pace, 1997).

2.4.2.3 Primers and Polymerase Chain Reaction (PCR)

Thus the first step of the process is to isolate the 16S rRNA gene from the total DNA extracted from the environmental samples. The highly conserved nature of the 16S rRNA gene makes it possible to design universal or specific primers for PCR amplification of the ribosomal RNA genes from practically all kinds of organisms in the sample (Smalla, 2004). The choice of primers for any PCR amplification is a critical step for achieving reliable and accurate results with high confidence level. When the starting material is a heterogeneous mixture of species DNA from all three domains of

life, the primer has to be bacteria-specific, leading to near full length DNA with little or no bias, and ideally representing the entire bacterial community (Hongoh *et al.*, 2003).

With the target of choice being the highly conserved bacterial 16S rRNA gene, there are quite a number of primers to select from, including forward primers 27, 39, 41, 63, 64F and reverse primers 1387, 1389, 1392, 1492, 1522R (Suzuki and Giovannoni, 1996; Hongoh *et al.,* 2003). The most widely used primer set is 27F (bacteria-specific) and 1492R (universal) designed from the corresponding base positions of *E. coli* 16S rRNA gene (Lane, 1991; Hongoh *et al.*, 2003; Frank *et al.*, 2008). The binding positions complementing the primers 27F and 1492R are highly conserved in bacterial 16S rRNA gene, making them ideal candidates for covering a vast group of bacterial species present in the DNA mixture obtained from environmental samples (Frank *et al.*, 2008).

The most important step in a successful and stringent PCR is the annealing of primers to correct regions. There are a few factors contributing to efficiency of primer annealing, such as the chemical components of the PCR buffer, concentration of the primer, Mg^{2+} concentration, and last but certainly not least, annealing temperature (Sipos *et al.*, 2007). The former factors are not readily modified since they require extensive molar calculations and are hard to measure accurately, whereas temperature can be easily measured and adjusted making it a good variable for modification (Sipos *et al.*, 2007).

Generally, if the annealing temperature is too low, there might be unspecific binding of the primer called mismatch (Sipos *et al.*, 2007). On the other hand, if the temperature is too high, wobble primer-template hybridization will fail to initiate enzymatic polymerization. The rule of thumb is to optimize the PCR amplification to the lowest annealing temperature such that while the reaction is specific, no unspecific products are observed (Sipos *et al.*, 2007). Overall, 27F has been shown to have the

lowest rate of mismatch with a wide range of target sequences and a high product yield (Suzuki and Giovannoni, 1996).

2.4.2.4 Separation via Cloning

There will be numerous copies of the target gene originating from various organisms in the final product that need to be differentiated. The most widely used technique for separating the DNA fragments in the mixture is by cloning them into a vector and transforming them into *E. coli* cells.

An important feature of the vector system used in most cloning procedures is the antibiotic resistance gene introduced into the plasmid. This allows for selective growth of transformed cells as opposed to non-transformed ones and is an essential key for immediately eliminating undesired growth by addition of antibiotic of choice (i.e. Ampicillin) to the media. The transformed cells encompassing the plasmid are now resistant to the fatal effect of antibiotic and are able to grow (Brown, 2001).

However the primary characteristic of the vector which makes it ideal for real-time screening of cloning efficiency is the presence of a multiple cloning sequences (MCS) within the *lacZ* gene coding for the β-galactosidase enzyme. This enzyme is in charge of breaking up disaccharide lactose to simpler sugar molecules in absence of glucose as primary energy source (Juers *et al.*, 2001). Looking closer at the biochemistry of the process, the operator region of *lac* operon is under constant repression by a repressor molecule. To start transcription of the gene, an inducer molecule is required to remove this repression, a term called derepression or induction of enzyme synthesis (Eron and Block, 1971).

Naturally, allolactose has the role of inducing this reaction in presence of lactose. Alternatively, molecules such as IPTG (isopropylthiogalactose) as inducer along with X-gal (5-bromo-4-chloro-3-indolyl *β*-D-galactopyranoside) as substrate can

be substituted to offer a colorimetric assay for rapid visual detection of cloned colonies (Juers *et al.*, 2001). In comprehension, a disruption of *lacZ* gene due to insertion of DNA into the vector will halter the synthesis of β-galactosidase enzyme. On the other hand, those intact plasmids which did not effectively undergo insertion process will retain the ability to synthesize a fully functional enzyme. Therefore, they are able to metabolize lactose, or in this case the modified galactose molecule called X-gal and turn it into a blue-coloured insoluble product (5-bromo-4-chloro-indigo) whereas the cloned cells with a non-functional enzyme are unable to break this molecule down and thus remain clear (i.e. white) (Karcher, 1995).

It is noteworthy that this process only happens in absence of glucose as primary source of energy, meaning the cells will preferentially exhaust the available glucose in the media before they turn to less favourable nutrients, i.e. X-gal (Karcher, 1995). This is due to the fact that RNA polymerase requires a positive regulation by cAMP-CAP complex to initiate transcription. The cAMP levels are very low when glucose is being metabolized by the cell, but starts to build up once the glucose level is diminished. Only then the regulatory complex is formed and the transcription of *lacZ* gene begins, leading to activation of X-gal metabolism. Eventually, the blue colour starts to form after a while and gradually continues to become a deeper shade of blue once cells containing the intact plasmid start breaking down the X-gal molecules present in the media.

An example of vector system used in experiments is a high copy number plasmid containing SP6 and T7 RNA polymerase promoters. These two sequences are flanking the MCS where the DNA fragment of interest will be inserted. This is another useful feature for further inspection of white colonies for the correct type of insert. It is mandatory to confirm at this stage whether or not the correct insertion has occurred. It takes a very little piece of DNA to disturb the *lacZ* gene and produce a clone which cannot metabolize the X-gal molecule. However this does not mean that every white

colony includes the desired fragment. This is where the SP6 and T7 promoters come in handy. Using primers designed for their respective sequences, the inserted DNA will be isolated via direct colony PCR amplification and tested for the correct size with a routine agarose gel electrophoresis.

Initially, most of molecular-based techniques involved cloning and sequencing of the target gene isolated from environmental samples (Pace, 1997; Kirk *et al.,* 2004). However, sequencing thousands of clones is unnecessary, especially when most of the results turn out identical. Alternatively, there are a number of techniques to separate the products of the amplified DNA segment.

2.4.2.5 Molecular Fingerprinting

Some of molecular fingerprinting techniques include denaturing or temperature gradient gel electrophoresis (DGGE; TGGE), single strand conformation polymorphism (SSCP), and restriction fragment length polymorphism (RFLP) of 16S rRNA gene amplified from total community DNA.

DGGE method involves separation of DNA strands on a polyacrylamide gel with a gradient of increasing concentration of denaturants based on the melting behaviour of double-stranded DNA. TGGE follows the same protocol except that the gradient is temperature instead of chemical denaturant (Muyzer, 1999). Specific bands can be excised from gels, re-amplified and sequenced for identification purposes. These techniques are reliable, reproducible, rapid and fairly inexpensive, though sample handling is labour-intensive and might influence the microbial community [refer to the review by Kirk *et al.,* (2004)].

SSCP is another technique based on electrophoretic separation, which relies on difference in mobility of single-stranded DNA caused by their folded secondary structure. This folding pattern is dependent on the DNA sequence as long as DNA

fragments are of equal size and no denaturant is present (Lee *et al.*, 1996). Thus, the conformation polymorphism of different DNA molecules provides some information about the organism from which it originates. However, it must be noted that certain single-stranded DNA are capable of forming more than one conformation and therefore, will be represented by more than one band on the gel (Tiedje *et al.*, 1999).

Restriction fragment length polymorphism (RFLP), aka amplified ribosomal DNA restriction analysis (ARDRA), is another powerful tool for discriminating bacterial individuals. The mechanics of the process include PCR amplification of the rRNA gene and digesting the resulting molecules with restriction enzymes. Different fragment lengths are then detected by routine agarose or non-denaturing polyacrylamide gel electrophoresis for community analysis (Liu *et al.*, 1997). Indeed straightforward, this technique, combined with clone library construction, is one of the most promising tactics for tackling bacterial diversity. Nonetheless, like the previous techniques, this method is not 100% flawless. For one, banding patterns in diverse communities might turn out too complex to analyze using RFLP (Øvreås and Torsvik, 1998; Tiedje *et al.*, 1999). A summary of advantages and disadvantages of general molecular fingerprinting techniques is presented in Appendix A (Table 1S).

Overall, profiling by means of DGGE, SSCP, and RFLP is ideal for analyzing PCR products amplified from genes such as 16S rRNA gene which is composed of both conserved and variable regions. By reviewing DNA fingerprint profiles of the microbial community, immediately dominant members are revealed and representatives of each community member are selected for DNA sequencing. This will provide coherent information on their phylogenetic affiliations to correctly identify the bacterial diversity of corresponding environments (Smalla, 2004).

2.4.2.6 DNA Sequencing

As mentioned, it is now routine practice to amplify 16S rRNA gene from total community DNA using either universal or bacteria-specific primers and clone the fragments to construct a clone library. The clones are then screened by certain fingerprinting techniques such as RFLP to roughly distinguish different types of bacteria. Later on, DNA sequencing is carried out to confirm the estimated biodiversity.

It all comes down to the sequence patterns at hand. Nucleic acid sequencing methods are evolving tremendously to meet with the high demands for them. The quality of sequence data are improving with the rising speed and technology every day (Clarridge, 2004). There are several methods for determining DNA sequence. The classical approach is similar to PCR in that it uses the purified PCR products as DNA template in a process called cycle sequencing. Forward or reverse primers can be used in separate reactions if both forward and reverse sequences are desired, although the use of only the forward primer is as effective (Bosshard *et al.*, 2003; Mignard and Flandrois, 2006). By adding specially labelled bases called dye terminators as well as normal bases to the reaction, the generated sequences will be terminated at different lengths once the labelled bases get randomly inserted. Since each of the four terminator bases has a unique fluorescent dye, the terminal base of each DNA fragment can be determined by a fluorometer (Clarridge, 2004).

The basic idea behind DNA sequencing is to reveal the order of nucleotides in a DNA molecule which is unique for every individual. This information is critical for correct identification of the organism and will be the key to unravel the phylogenetic relationships between all living creatures. For this purpose, sequences from different organism are aligned and compared two by two. The differences of the sequences are then counted and considered as a measure of “evolutionary distance” between the two

organisms (Pace, 1997). Woese (1987) developed the universal evolutionary tree of life based on pair-wise comparison of ribosomal RNA genes from several organisms and came up with three main domains: eubacteria (now known as Bacteria), archaebacteria (now known as Archaea), and eukaryotes (or Eukarya).

Contemporary microbiology has benefitted dramatically from the features of molecular-phylogenetic package offered as means of identifying microorganisms, particularly unculturable bacteria. With the help of the sequence-based taxonomic framework presented by molecular phylogenetic trees, it only takes a single gene sequence (such as that of 16S rRNA molecule) to identify the phylogenetic type of unknown organisms (Pace, 1997). Nowadays, phylogenetic types (i.e. phylotypes) of any given organism and their distribution pattern in natural habitats can be inspected by sequencing of 16S rRNA gene amplified from total community DNA extracted directly from the respective environment (Pace, 1997).

A word of caution to this tale is the effect of the sequence quality on accurate identification. The quality of the sequence is critical for its correct interpretation, leaving it to the microbiologist to check the sequence chromatograms one by one for any ambiguities such as double peaks (Mignard and Flandrois, 2006). Deciding what region and how much of the ssu RNA gene to sequence is another major factor contributing to final results. Partial sequences of only 200bp are adequate for phylogenetic analysis (Priest, 2004).

2.4.2.7 Identification

Samuel T. Cowan cracked the code of identification by breaking it down into three clear steps: classification, nomenclature, and identification (Cowan, 1965). It seems reasonable to engage the unknown organism to one that has been previously characterized and placed within a classification framework. At present, classifications

are mostly kept as computer databases, available online via the World Wide Web, enabling their rapid and accurate interrogation for identification purposes (Priest, 2004).

Bacterial phylogenetic classification now mainly relies on sequence analysis of the 16S rRNA gene. There are over one million full length 16S rRNA gene sequences deposited in various online databases such as GenBank at National Centre for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/), Ribosomal Database Project (RDP-II) (http://rdp.cme.msu.edu/html/), the Ribosomal Database Project European Molecular Biology Laboratory (http://www.ebi.ac.uk/embl/), and others (Clarridge, 2004).

Once the ribosomal RNA gene sequence has been obtained, it is submitted to a BLAST (Basic Local Alignment Search Tool) query search at any of the mentioned websites. The server then aligns the query to all the known sequence data in the database according to preferred algorithm and returns the results in order of similarity score, with the most similar sequence on top of the list. Scores are based on the number of identical positions between the two sequence patters, meaning the more number of bases they have in common, the higher the similarity score and thus, the closer they are. Typically, above 97% similarity to an entry in the database can be a good indicative of identification (Priest, 2004).

Sequences can be assigned to operational taxonomic units (OTUs) based on the genetic distances between them (Schloss and Handelsman, 2005). There are a few methods available, some of which can be publicly accessed online such as FastGroup (Seguritan and Rohwer, 2001). Although there is no universal agreement on the matter, there are several cut-off similarity levels suggested from 95% to 99% (Martin, 2002; Mignard and Flandrois, 2006). At species level, sequences showing ≥97% similarity in their base patterns are considered belonging to the same OTU and treated as same. This

is a key point in assessing bacterial diversity since all downstream calculations are influenced by the way one defines OTU (Martin, 2002; Schloss and Handelsman, 2005).

Once the sequences have been discriminated, a multiple sequence alignment can be created. There are many sequence-comparing software packages available for this purpose (http://evolution.genetics.washington.edu/phylip/software.html website includes a complete list of websites offering bioinformatics tools). A hierarchical cluster analysis is performed to show how distant the sequences are from one another in a tree-like plot (i.e. dendrogram). The most commonly used methods for generating dendrograms are the NJ (Neighbour-Joining) method, the UPGMA (Unweighted Pair Group Method with Arithmetic average), and WPGMA (Weighted Pair Group Method with Arithmetic average) (Clarridge, 2004). Eventually, a phylogenetic tree is constructed and the unknown sequences will be linked to their closet relatives, providing them with some sort of classification.

It should be noted that sequence analysis is not the most straightforward part of the process. It requires careful interpretation of phylogenetic trees, similarity coefficients, and a well-trained eye in bioinformatics as well as ecological knowledge to validate the identification. Not to mention the primary taxonomic assignment is entirely dependent on the database used (Mignard and Flandrois, 2006). At any rate, there is no such thing as a “push-button” technique at this point and the final call is up to the researcher.

2.5 Biodiversity Assessment

There are two main components to species diversity; species richness or the total number of species present (*S*), and evenness (or equitability *J*) or the distribution of individuals among the species (relative abundance) (Hsieh and Li, 1998; Øvreås, 2000). Even though species richness is the milestone of biodiversity measurements, it remains an elusive quantity to measure. Despite the application of the most modern molecular techniques, only a small fraction of the original richness can be captured (Gotelli and Colwell, 2001).

Therefore, the number of species in most communities is estimated statistically based on a small subset of samples and extrapolating to a larger area (Chazdon *et al.,* 1998; Hong *et al.,* 2006). Typically, a comparison of different communities from characteristically different environments illustrates the significance of diversity for a better grasp of the issue. Several studies have used diversity indices, estimated species richness, and rarefaction curves along with other diversity statistics specific to microbial samples (Hughes *et al.,* 2001).

2.5.1 Diversity Indices

Numerous diversity indices have been used in ecology to measure proportional abundance of species in a given environment. The most extensively employed diversity indices in literature are Shannon diversity (*H*), Simpson diversity (*D*), and Fisher's α (Colwell, 2009). Shannon index combines the two components of biodiversity, i.e. richness and evenness into one single value and is sample size dependent (Hsieh and Li, 1998). Generally, the higher the sampling effort, the more number of organism types are identified and thus, a higher value of Shannon index.

Simpson diversity index is considered a dominance index and weights towards the abundance of the most common species. It is more sensitive to evenness of species

distribution than species richness (Colwell, 2009). Fisher's α is used for predicting the number of unseen species in the sample set. An index contributing exclusively to evenness is the Berger-Parker index, which is simply the reciprocal of the proportion of individuals belonging to the most abundant species. Together these indices can be very informative about how diverse and how even the population is, but is not sufficient (Colwell, 2009).

2.5.2 Species Richness Estimation

As mentioned, statistical approaches have been developed to compare and estimate species richness from samples of microorganisms. To avoid the limitations of diversity indices, rarefaction diversity measurements have been introduced which scale down all collections to the same sample size. Rarefaction is a statistical method estimating the expected number of species in a randomly selected sample of individuals from a collection (Hsieh and Li, 1998). A rarefied curve will result from averaging randomizations of the observed accumulation curve. The observed richness among samples can be compared by applying the variance around the repeated randomizations, but it is far from a measure of confidence about the actual richness in the communities (Hughes *et al.*, 2001).

Richness estimators allow for an estimation of total richness of the community from a sample which can then be compared across all samples (Hughes *et al.*, 2001). Chazdon *et al.* (1998) categorized richness estimators into three main classes: extrapolation from species-accumulation curves, parametric estimators, and nonparametric richness estimators (Colwell and Coddington, 1994; Colwell, 2009).

In case of accumulation curves (aka collector's curve), the cumulative number of different bacterial types (phylotypes) is plotted versus sampling effort (Chazdon *et al.*, 1998). The resulting curve is primarily linear as more common species are added to the

collection almost constantly but begins to plateau when rarer ones slowly appear. The curve will eventually reach an asymptote when all the species have been sampled and nothing new is discovered (Hughes *et al.*, 2001; Colwell, 2009). This is achieved by extrapolating the curve to an asymptotic value or to sample sizes larger than the observed (Chazdon *et al.*, 1998).

The second class of estimation methods fits data on the relative abundance of species in a sample to parametric distribution (Chazdon *et al.*, 1998). Among the models used are log-series, log-normal, and Poisson log-normal. One way of plotting the species abundance data is by counting the number of species in each abundance category and plotting the frequencies against abundance categories. If the abundance categories are chosen in powers of two, the resulting plot is called a log abundance plot. Lognormal (normal bell-shaped) statistical distribution in such plots is a good indicative of well-inventoried natural communities and confirms the adequacy of sampling effort (Colwell, 2009). By fitting the sample data to lognormal distribution, the parameters of the curve can be evaluated to estimate the number of unobserved species and eventually, the total richness of the community (Hughes *et al.*, 2001).

An initial realization of how well the community has been sampled allows for a more confident estimation of total species richness. On the contrary, if the sample set does not maximally represent the original community due to high diversity or undersampling, the richness estimators must be used with caution. Either accumulation curves or log abundance plots can be employed to test whether the sampling effort was sufficient and the population was inventoried exhaustively (Colwell, 2009).

The last class of estimation methods, nonparametric estimators, shows the most promise for microbial studies. Richness estimation is an alternative to rarefaction for comparing richness among incompletely inventoried communities by extrapolating

beyond what has been observed to estimate the unknown asymptote of a species accumulation curve (Colwell, 2009). Consequently, these methods are dependent on frequencies of the rarest classes of observed species to estimate the number of unseen species (i.e. present but not detected). For example, the simplest nonparametric estimator, *Chao1*, depends on the observed number of singletons (species observed just once) and doubletons (species observed exactly twice) in addition to total number of species observed (Chao, 1984; Colwell, 2009).

Ultimately these estimators consider the proportion of species that have been observed before to those that are observed only once. The probability of "catching" the same species more than once in a community is dependent on how diverse the community is; the higher the diversity, the lower the possibility of having more than one individual per species and vice versa (Hughes *et al.*, 2001). There are various richness estimators introduced for mathematically expressing the species richness in the source environment, each with their pros and cons given a particular task (Colwell and Coddington, 1994; Hortal *et al.*, 2006). It is once more the researcher's decision to select from a wide range of formulas depending on what data is present and what needs to be figured out.

A main problem at this stage is the use and abuse of richness estimators in assessing biodiversity, especially when it comes to bacterial communities from complex ecosystems where diversity is extremely high and the techniques for measuring that diversity are only starting to evolve. Underestimation, and indeed overestimation, of bacterial species richness is inevitable with the current statistical tools since they have shown unrealistic results as well as inconsistency in estimated numbers of species from comparable environments (Hong *et al.*, 2006). It must be kept in mind that even with constant attempts to formulate new ways to estimate species richness, it still remains just that; mere estimations and speculations of total numbers of species. Even then the

numbers do not represent the original sample since at the moment all the estimations are based on the 16S rRNA genes in a clone library (Hong *et al.*, 2006).

2.6 Final Remark

Obviously, the present status of bacterial diversity measurement techniques is not 100% ideal to reflect the original environment's true biodiversity. Nonetheless, the path is rapidly being paved towards more accurate and bias-free approaches for identification and quantification of bacterial species from natural environments. In the meantime, the current advances in molecular microbiology and computational biology can be employed in favour of studying bacterial diversity throughout different environments for all sorts of purposes.

In particular, extreme environments provide a unique source of often highly adapted and tolerant organisms. Research on organisms in these habitats has led to the discovery of novel and useful compounds and may assist in understanding the impact of global change on biodiversity. Valuable indeed, the survey of microbial communities in previously remote areas illuminates the true nature of human impact on the environment and the response to these changes for better or for worse.

Among the extreme environments, the polar regions of the Earth, especially Antarctica, have been grabbing a lot of attention for a possibly more remote access to undisturbed ecology. Researchers are flying in from all over the world to carry out experiments in the abandoned land in hopes of better understanding ecological relationships of living organisms without the interruption of mankind. This, however, has proven to be an almost impossible task for the presence of the researchers at first place is breaking the first code of conduct. As the level of human activities in Antarctica has increased, the relative importance of humans as vectors for the introduction of

exogenous microbiota to the region has also increased. Several studies have contributed to the possible outcomes of such disturbances in the past and present, with the most recent review done by Convey (2010). This area requires constant inspection and should not be easily forgotten.

All in all, studying bacterial diversity in natural environments demands high throughput culture-independent technologies from DNA extraction to PCR and cloning. Each of the steps involved in the process need to be optimized in order to minimize the bias associated with them. Indeed the modern methods applied in the process meet pitfalls down the road, leaving it to the researcher to interpret the results correctly considering all the facts.