Upper Airways Microbiota Profiling in a Case/Control Study between Wheezing and Healthy Children from the Tropics of Ecuador

PAUL ANDRES CARDENAS-ALDAZ

Imperial College London, Faculty of Medicine,
National Heart and Lung Institute

A Thesis Submitted for the Degree of Doctor of Philosophy

Declaration

I herewith certify that this dissertation is my original work and that all material included which is not my own work has been properly acknowledged. The current study was performed by myself including: samples collection, DNA extraction, PCR, DGGE analysis, bacterial cloning, pyrosequencing, bacterial culture and bioinformatical analysis. The Sanger DNA sequencing was performed by Cogenics, UK.

Paul Cardenas-Aldaz

Acknowledgments

I want to acknowledge first to Prof Miriam Moffatt and Prof Bill Cookson for their incredible support during my PhD program. It has been and honour for me to be a part of the Molecular Genetics and Genomics group in the NHLI and to have been mentored by scientists that personally admire.

Also I am especially thankful with Prof Philip Cooper and all the ECUAVIDA team for letting me to be part of that amazing project in Ecuador. Definitely is important to remark my infinite gratitude with all the children and their parents which are part of this project.

My deep gratitude with Markus Hilty, Aristea Binia, Mike Cox, Chi-Yung Chiu, Ghazala Mirza, Katty Sim, Frankie Bolt, Youming Zhang, Kenny Wong, and all the rest of people in molgen for their friendly encouragement and advice. Additionally I want to thank to Prof Simon Kroll and Prof Brian Spratt with their guidance in the development of the streptococci-typing project.

Many thanks to the NHLI Foundation, the Wellcome Trust and the SENESCYT for funding my PhD scholarship and financing the current project.

Finally I want to thank foremost to my beloved family because everything that I obtained is because of you and for you. Thanks to (in inverse order of appearance) Sergio Antonio (for bringing joy and blessings to my life), Sandra (for being my light, the source of my happiness, and my adventures partner), Dudú (for bringing laughter), Lili (for being my confident), Nydia (for being my biggest supporter), Patricio (for being my example), Willer (for always been there) and María. Thanks also to my best friends Juan Carlos, Julio, Agustín for the laughs and their constant encouragement.

Abstract

Background: The relationship between living in rural areas and the acquisition of protective environmental factors against the development of asthma and atopy gave rise to the hygiene hypothesis. Between the environmental factors attributed to asthma, particular airways microbiota patterns have been encountered in adult asthmatics using molecular independent techniques (Hilty *et al.,* 2010) and in children using conventional culture methods (Bisgaard *et al.,* 2007). Here a retrospective time series study has been performed using samples from infants at different ages to study the microbiome variations in a population with very low antibiotic use history and no corticosteroid usage.

Methods: Pyrosequencing of amplicons of the bacterial 16S rRNA gene was performed from oropharyngeal samples from 134 infants with episodic wheezing versus 200 healthy infants (total of 334 infants examined) sampled at different ages (7, 12 and 24 months). Bioinformatic analyses were conducted using QIIME 1.7 software and Phyloseq package on R. Additionally a new culture-independent pyrosequencing approach using the *map* gene was successfully developed to enable discrimination of streptococci at species level.

Results: Significant abundance differences between infants with wheezing history and healthy controls were found for the Fusobacteria, Proteobacteria and Actinobacteria phyla. At genera level a significant increase in potential pathogenic bacteria (*Neisseriacea, Haemophilus, Staphylococcus*) was found in wheezers whilst a higher prevalence of *Veillonella* spp. was seen in controls. In addition, using *map* gene pyrosequencing, *Streptococcus salivarious* was found to be statistically significantly related with wheezing syndrome whilst *Streptococcus mitis* was more prevalent in controls. When age was considered differences were found in the microbiota displayed as species numbers increased (alpha diversity).

Conclusions: The respiratory microbiota is different at phyla, genera and Operational Taxonomic Unit levels when comparing between wheezing and healthy children. A progressive more complex respiratory microbial community develops with age.

Table of Contents

Figures and Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Abbreviations

- 3D (tri-dimensional)
- A (Adenine)
- AM (Asthma marker)
- ANOVA (Analysis of variance)
- APC (Antigen presenting cells)
- APRIL (Proliferation–inducing ligand)
- APS (Ammonium Persulfate)
- BAL (Bronchoalveolar lavage)
- BCG (Bacillus Calmette-Guérin)
- BLAST (Basic Local Alignment Search Tool)
- C (Cytosine)
- CCL5 (Chemokine C-C motif ligand 5)
- CF (Cystic fibrosis)
- CI (Confidence interval)
- COPD (Chronic obstructive pulmonary disease)
- DALYs (Disability-adjusted life years)
- DGGE (Denaturing gradient gel electrophoresis)
- DMSO (Dimethyl sulfoxide)
- DNA (Deoxyribonucleic acid)
- dNTPs (deoxyribonucleotides)
- DOTUR (Distance based OTU and Richness determination program)
- dsDNA (double stranded DNA)
- ECUAVIDA (Ecuador Life birth cohort)
- ELISA (Enzyme-linked immunosorbent assay)

- ERS (European Respiratory Society)
- FcεR1 (IgE receptor 1)
- FDR (False discovery rate)
- fna (Fasta file)
- FWD (Forward primer)
- G (Guanine)
- GINA (Global Initiative for Asthma)
- GM (Gradient marker)
- GPR (G protein-coupled receptor)
- *guaA* (GMP synthase)
- GWASs (Genome-wide association studies)
- HLA (Human lymphocyte antigen)
- HMP (Human Microbiome Project)
- HMPV (Human metapneumovirus)
- HPAB (Hospital "Padre Alberto Buffoni")
- HPIV (Human parainfluenza virus)
- HRV (Human rhinovirus)
- IBD (Inflammatory bowel disease)
- ICAM (Intercellular Adhesion Molecule)
- IFN (Interferon)
- IgE (Immunoglobulin E)
- IL (Interleukin)
- iNOS (induced nitric oxide synthase);
- IPTG (Isopropyl thiogalactoside) (Sigma, UK)
- ISAAC (International Study of Asthma and Allergies in Childhood)
- KC-1 (Keratinocyte chemokine-1)

- km (kilometres)
- LB media (Lysogeny broth)
- LUL (Left upper lobe in the lung)
- *map* (Methionine aminopeptidase)
- $MgCl_2$ (Magnesium chloride)
- MHC (Major histocompatibility complex)
- MIDs (Multiplex identifiers)
- ML (Maximum likelihood)
- MLSA (Multi locus sequence assignation)
- MLST (Multi locus sequence typing)
- MP (Maximum parsimony)
- MSP (Public Health Ministry of Ecuador)
- NCBI (National Center for Biotechnology Information)
- NJ (Neighbour joining)
- NM (Normal marker)
- NMDS (Nonmetric multidimensional scaling)
- NRI (Net relatedness index)
- NTI (Nearest taxon index)
- ºC (Degrees Celsius
- Odds Ratio (OR)
- OP (Oropharynx)
- ORMDL3 (Orosomucoid like gene type 3)
- OTU (Operational taxonomic unit)
- PCoA (Principal Coordinate Analysis)
- PCR (Polymerase Chain Reaction)
- *pfl* (Pyruvate formate lyase)

- PGD2 (prostaglandin D2)
- pGEM-T (plasmid GEM-T)
- *ppaC* (Inorganic pyrophosphatase)
- *pyk* (Pyruvate kinase)
- QIIME (Quantitative Insights Into Microbial Ecology)
- R (The R Project for Statistical Computing)
- RANTES synonym CCL5 (regulated on activation, normal T cell expressed and secreted cytokine)
- RDP (Ribosomal Database Project)
- REV (Reverse primer)
- *rpoB* (RNA polymerase beta subunit)
- rRNA (Ribosomal ribonucleic acid)
- RSV (Respiratory syncytial virus)
- sff (Standard Flowgram Format binary file)
- SFTPA (surfactant protein A)
- SFTPD (surfactant protein D)
- SMAD3 (Decapentaplegic homolog 3 gene)
- *soda* (Superoxide dismutase)
- sp. (specie)
- spp. (species)
- SPT (Skin prick test)
- STAT (Signal Transducer and Activator of Transcription)
- T (Thymine)
- TAE (Tris Acetate EDTA buffer pH 8.0)
- TE (Tris EDTA buffer ph 8.0)
- TEMED (Tetramethylethylenediamine)

- TGF (Transforming growth factor)
- Th2 (T helper type 2 lymphocyte)
- TLR (Toll like receptor)
- TNF (Tumoral necrosis factor)
- tre (Phylogenetic tree file)
- TSLP (thymic stromal lymphoprotein)
- *tuf* (Elongation factor Tu)
- txt (Text file)
- UPGMA (Un-weighted Pair Group Method with Arithmetic mean)
- VCAM (Vascular cell adhesion protein)
- VDR (vitamin D receptor)
- WHO (World Health Organization)
- XGal (5-bromo-4-chloro-3-indolyl-beta-D-galactopyranoside)
- µl (Microlitres)
- µM (Micromolar)

Chapter 1: Introduction

1.1. Asthma and Wheezing

Asthma is a cluster of airways diseases that is manifested by a wide variety of respiratory symptoms. The disease is characterized by non-specific bronchial hyper-responsiveness that causes wheezing, dyspnoea, coughing, and chest rigidity (Busse and Lemanske, 2001, Nelson and Zorc, 2013). It is important to note that many asthmatics exhibit raised total Immunoglobulin E (IgE) and specific IgE to

inhaled allergens, for that reason asthma is commonly stated as an ‘atopic’ disease (Busse and Lemanske, 2001).

Immunohistopathology studies of the airways in asthmatics reveal a characteristic infiltration of inflammatory cells, thickening of the basal membrane, epithelial-cell damage, occlusion of the bronchial lumen by mucus as well as hyperplasia and hypertrophy of bronchial smooth muscle (Haley and Drazen, 1998, Holgate, 2013, Aoshiba and Nagai, 2004).

The aetiological cause of asthma is complex with both genetic and environmental factors contributing and influencing age of onset and the severity of symptoms (Cookson, 2002, Martinez and Vercelli). Typically an asthma exacerbation occurs when inhaled allergens (antigens), such as house dust mite, trigger a strong T helper type 2 lymphocyte (Th2) response with in the airways. Mast cells through activation of their IgE receptors release potent inflammatory mediators including histamine, leukotrienes, RANTES, and interleukins (IL) including IL-4 and IL-5 (Busse and Lemanske, 2001). These cytokines result in the activation of eosinophils that in turn secrete adhesion molecules such as VCAM-1, ICAM-1 and selectins whilst activation of Th2 lymphocytes results in production of Th2 interleukins including 4, 5, 6, 7, 9, 11 and 18. With the recruitment of inflammatory cells, subsequent complement activation results in epithelial cell damage, a cholinergic

stimulus leading to broncho-constriction and B lymphocyte conversion into plasma cells capable of producing more IgE thereby perpetuating this altered inflammatory cycle (Figure 1.1) (Mamessier and Magnan, 2006, Afshar *et al.*, 2008, Kohl and Wills-Karp, 2007, Wenzel, 2012).

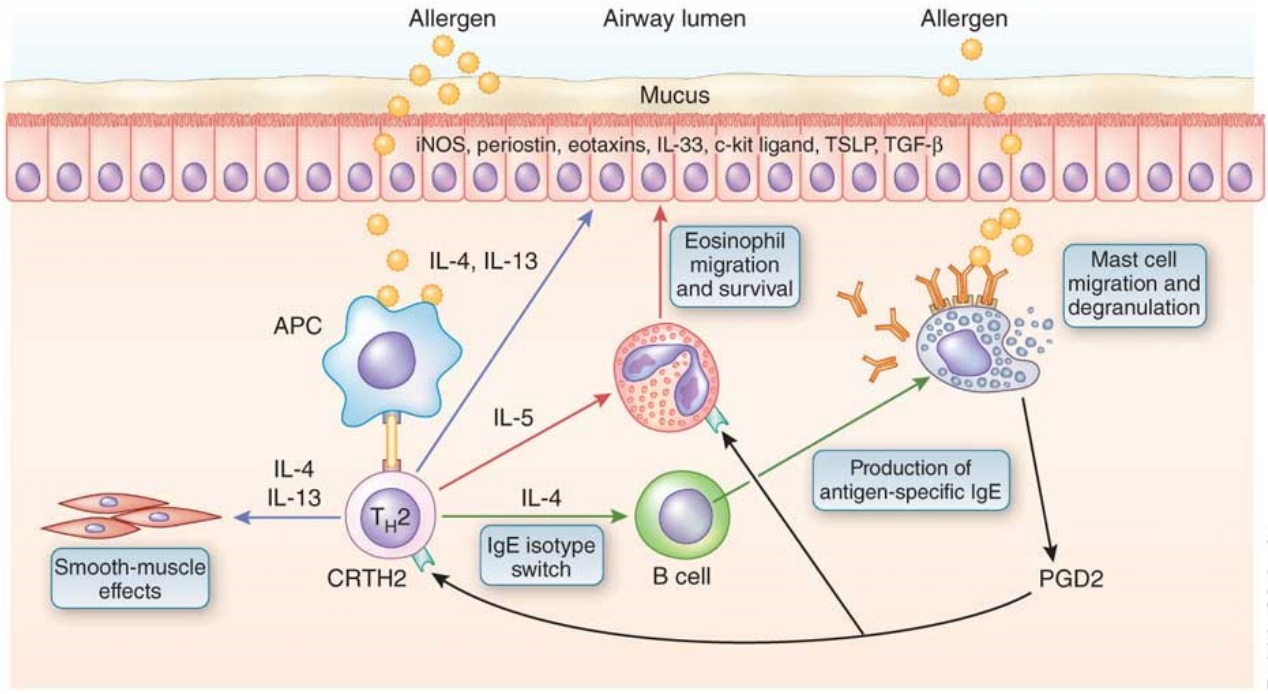


Figure 1.1. The Allergic Th2 Pathway. The allergen penetrates into the organism and is taken up by antigen presenting cells (APC) which migrate to the nearest lymphoid organ where the allergen is processed and presented by the MHC-II molecule to T lymphocytes. A naive T cell recognizes the allergen, expands clonally, and migrates to the site of inflammation. *In situ*, activated Th2 cells produce Th2 cytokines: IL-5 causes eosinophil recruitment and activation leading to tissue damage; IL-4 and IL-13 predominantly activate B cells and allow IgE commutation and synthesis. IgE binds FcεR1 on non-specific immune cells. When contact with allergen occurs again, already bound IgE immediately captures the allergen, resulting in receptor cross-linking and rapid mast cell and basophils activation. These cells are responsible for the release of most of the inflammatory and toxic mediators (such: iNOS, induced nitric oxide synthase; PGD2, prostaglandin D2; TSLP, thymic stromal lymphoprotein) that cause the immediate allergic symptoms. Taken under permission (see appendix) from Wenzel (2012).

1.2. Asthma Diagnosis in Infants

To define a single 'asthma' phenotype is not straight forward as different types and severities of the disease are increasingly being recognized. Diagnosis in children includes clinical history, physical examination, as well as functional and laboratory tests. Clinical diagnosis is crucial in the investigation of apparent asthma symptoms such as wheezing, shortness of breath and cough. In infants, asthma diagnosis is additionally complicated by the fact that before the age of 6 years, lung function tests are not easy to perform and invariably are inaccurate (Szefler *et al.*, 2014). Consequently analysis of recurrent wheezing not related with infections is the most important diagnostic indicator of asthma in this age group (<6 years) (Wechsler, 2009).

Not all children who experience wheezing events will develop asthma, and conversely not all children with asthma wheeze (Just *et al.*, 2010, Wechsler, 2009). It has been reported that around 50% of children have at least one episode of wheezing in the first three years of life; but only 27% of recurrent wheezing infants were found to have wheezing at 6 years of age. Consequently in children that are less than 24 months old rather than a diagnosis of asthma being assigned it is recommended that the diagnosis of recurrent non-infectious wheezing syndrome be given (Just *et al.*, 2010, Wechsler, 2009). The Global Initiative for Asthma (GINA)

and the European Respiratory Society (ERS) have classified this type of wheezing as multiple-trigger wheeze. Multiple-trigger wheeze is characterized by “episodes that occur as episodic exacerbations, but also with symptoms including cough and wheeze occurring between these episodes, during sleep or as a consequence of triggers such as activity, laughing or crying” (Kroegel, 2009, Bousquet *et al.*, 2007).

Asthma is considered an atopic disease with high total and specific IgE levels determined by enzyme-linked immunosorbent assay (ELISA). ELISA is a direct method of measuring proteins using specific antibodies against those proteins. These antibodies have an enzyme linked that produces a specific chemical reaction that can be measured (usually by a fluorometer). Skin prick test (SPT) is another method of measuring specific IgE against particular allergens. This method involves the sub-epidermic inoculation of the allergens and if a positive reaction occurs, a wheal is produced that can be measured in diameter.

1.3. Epidemiology of Asthma and Wheezing Worldwide

GINA has determined that there are 300 million people with asthma worldwide including all ages, ethnicities and socioeconomic status (FitzGerald *et al.*, 2013). When not efficiently controlled asthma causes restriction of an individual’s normal life style invariably hospitalization and even fatalities. The prevalence of asthma notably increases with Western and urbanized lifestyles (Figure 1.2). In 2025, it is anticipated

that from 45% to 59% of the worldwide population will live in urban areas, resulting in a notable increase in asthma prevalence of an additional 100 million people (Masoli *et al.*, 2004). Annually worldwide asthma causes approximately 180,000 deaths. Its fatality rate worldwide is about 1 in every 250 deaths although the majority of the deaths are due to insufficient health care in certain regions of the world (Masoli *et al.*, 2004, Kroegel, 2009).

Asthma is also considered an economic burden with the yearly costs worldwide being approximately 19.7 billion USD, of which 14.7 billion are direct medical costs, and 5 billion are indirect costs (related to losses in productivity and premature death). Drug prescriptions alone have an annual cost of more than 6 billion USD. In developed countries the financial cost of treating a patient with asthma ranges from 300 to 1,300 USD per year. About 15 million disability-adjusted life years (DALYs) per year are lost because of asthma, corresponding to 1% of all DALYs lost (FitzGerald *et al.*, 2013, Bateman and Jithoo, 2007, Masoli *et al.*, 2004). Asthma is therefore not only a disease of importance in terms of human health but has significant financial impact globally.

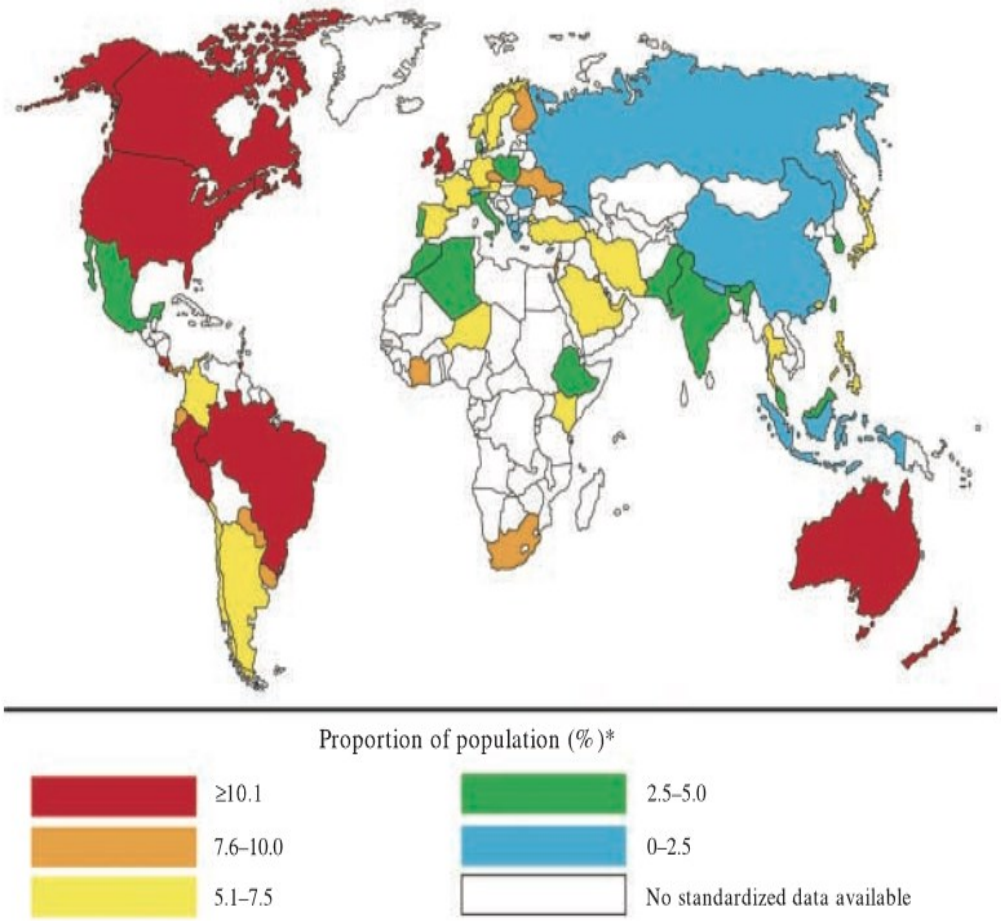


Figure 1.2. Prevalence of Asthma Worldwide. Dark red areas show a higher prevalence (see scale). Note Ecuador has a prevalence of asthma of 8.2% whilst in the United Kingdom it is between 15.3 % (England) and 18.4 % (Scotland). It is important to note that for some country data is not standardized. Taken under permission (see appendix) from Masoli *et al*., 2004.

1.4. Epidemiology of Asthma and Wheezing in Latin America and Ecuador

Latin America is very varied ethnically, geographically and economically. Approximately 25% of the population lives on an income of less than 2

USD/person/day (Lancet Editorial, 2007). As part of the International Study of Asthma and Allergies in Childhood (ISAAC), it has been estimated that diagnosed asthma prevalence ranges from 1.2% to 33.1%, whilst recurrent wheezing ranges from 3.9% to 30.8%. Geographical localization (as determined by latitude, altitude or tropical location) has been found to have no statistical correlation with asthma prevalence (Mallol *et al*., 2010). In Latin America only 11% of asthma appears to be related with the atopic phenotype (Weinmayr *et al*., 2007), with stronger correlations being seen with poverty (dirt, poor diet, obesity and psychosocial stress) (Cooper *et al.*, 2012).

The Latin American country Ecuador has regional differences revealed by studies that have been conducted in rural versus urban areas. According to the GINA criteria the prevalence of asthma in Ecuador is 8.2% (Figure 1.3) (Masoli *et al*., 2004) although the ISAAC study has found the incidence to be slightly higher at 10.9%. The incidence of wheezing however is 16.6% in urban areas, ISAAC centres in Quito and Guayaquil (Mallol *et al.*, 2010, Masoli *et al.*, 2004), whilst in rural areas, the Pichincha province, the rate of current wheezing is only 0.8%, and atopic wheezing 0.2% (Weinmayr *et al.*, 2007, Cooper *et al.*, 2009). In Ecuador, the fatality rate is 2.3 cases per 100,000 asthmatics. According to the World Health Organization (WHO) report 1998 only between the 50% and 80% of asthma patients

in Ecuador have access to appropriate medications (Masoli *et al.*, 2004). Furthermore, the estimated prevalence of current wheeze in relation to SPT reactivity (>3 mm) and allergen-specific IgE (>0.35 kU/L) is 0.1% (Weinmayr *et al.*, 2007). Non atopic asthma has been associated with recurrent chest infections, otitis media, croup, household damp or mould and pet exposures during the first year of life Moncayo *et al* 2010. In addition parasitic infections as *Trichuris trichiura* has been shown to be strongly inversely associated with atopic wheeze (Moncayo *et al.*, 2010), and similarly *Ascaris lumbricoides* or *Ancylostoma duodenale* with atopy (Cooper *et al.*, 2003). People in the tropics of Ecuador usually have high polyclonal IgE levels but low specific IgE (Cooper *et al.*, 2012), therefore atopy has to be redefined in this population. Consequently it would appear that the atopic component of asthma often seen in Western developed countries is not a feature in asthma/wheeze in Ecuador.

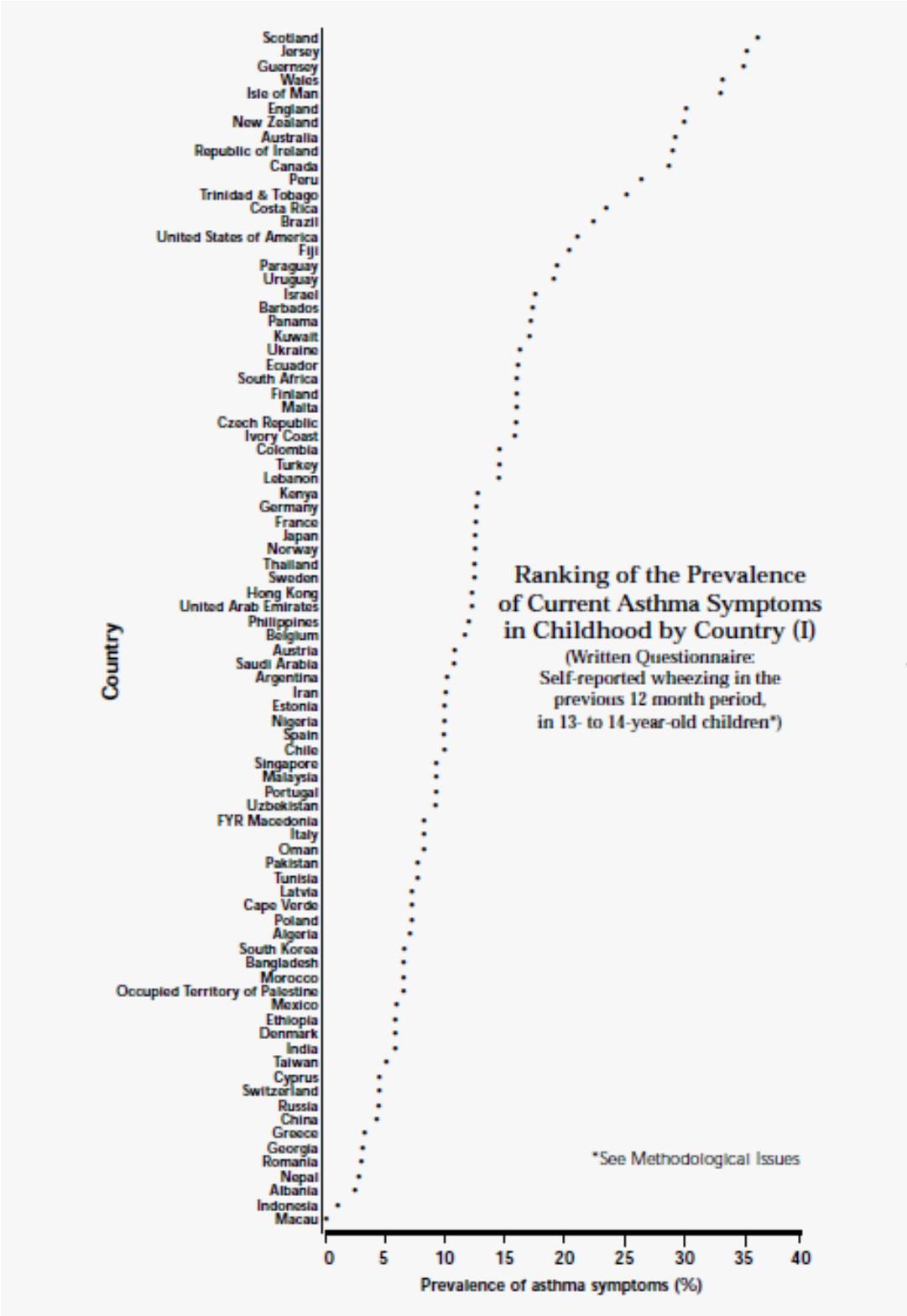


Figure 1.3. Prevalence of Asthma Symptoms by Country. Ecuador has a wheezing prevalence of 16.6 % Taken under permission (see appendix) from ISAAC study published by Masoli *et al*., 2004 .

1.5. Human Genetics and Asthma

Asthma is a complex disease involving both environmental and host-dependant factors (Cookson and Moffatt, 2011), and therefore host genetics plays an important role in asthma development. Caused by the asthma a multi-genic origin, it has been difficult to address a common gene that explains the wide range of symptomatology produced (Cookson *et al.*, 2011), and strikingly possible candidates like the IgE expression associated genes were not been found related (Moffatt *et al.*, 2010, Zhang *et al.*, 2012). For that reason genome-wide association studies (GWASs) have been a successful tool identifying loci related with this disease (Zhang *et al.*, 2012, Cookson *et al.*, 2011).

A large scale asthma consortium in Europe (the GABRIEL Consortium) using the GWAS approach identified a number of key loci related with asthma that involves the genes: IL18R1, IL33, IL1rL1, SMAD3 and ORMDL3 (Moffatt *et al.*, 2010). Several markers in chromosome 17q21 (where are located the genes ORMDL*3* and *GSDMB*) are more commonly related with childhood-onset asthma (Moffatt *et al.,* 2010), and particular variants regulating the *ORMDL3* gene expression have been related with childhood-onset asthma in different ethnic populations (Moffatt *et al.*, 2007, Zhao *et al.*, 2014). Additionally this gene locus also has been related with severe asthma phenotypes (Binia *et al.*, 2011).

1.6. The Hygiene Hypothesis and Asthma

The hygiene hypothesis has attempted to explain why some populations have a higher frequency of an over-reactive immune system in relation with the environmental factors affecting a particular group of people. This hypothesis was originally proposed following a number of epidemiological studies of asthma and atopy (Strachan, 1989), and the original study that included British children, postulated that allergic diseases could be prevented by the development of infection in infancy (there are mentioned in this study the viral infections) that could be transmitted by older siblings or by contact with the mother. Consequently, the reduction of family size and improvement in hygiene that has happened since the modern centuries, has resulted in notable increases in the development of developing atopic disease (Strachan, 1989),. It is noteworthy that the Strachan, 1989 study has subsequently been supported by data from studies involving both functional immunology and genetics (Liu, 2007).

It is interesting to also note that urbanization, migration and modernization are some of the most important factors related to asthma risk and these factors reflect the resultant changes in nutrition, exercise, exposure to allergens, use of antibiotics and vaccines, microbial exposures, effects of pollution, and psychosocial stressors (Cooper *et al.*, 2009, Weinberg, 2000). Poverty exhibits a dual association with

asthma depending on locality. Strikingly poverty in rural areas has been found to not be associated with asthma, whilst in urban areas poverty is an important risk factor for the disease (Benicio *et al.*, 2004).

Several studies suggest that factors including improved hygiene, antibiotics use, vaccinations, etc. and removal of certain microbial contacts during childhood may promote the development of atopy and asthma. These factors cause an imbalance in the immune system that becomes oversensitive to harmless antigens such as those of the house dust mite as reviewed by Liu (2007). Atopy has been shown to be related with asthma in developed countries. In Latin America however, several studies have shown a predominance of non-atopic wheeze in children especially in rural areas (Moncayo *et al.*, 2010).

1.7. Microorganisms and Asthma

Several microorganisms that produce respiratory infections during infancy and childhood have been associated with the development of asthma. Most of the microorganisms associated are viruses which have been related not only with asthma exacerbations (Nicholson *et al.*, 1993, Sebastian *et al.*, 1995) but as an aetiological agent of chronic inflammation on the airways (Tsukagoshi *et al.*, 2013, Openshaw *et al.*, 2003, Moser *et al.*, 2014). The viruses that have most frequently been implicated in the aetiology of asthma are: respiratory syncytial virus (RSV)

(Lotz *et al.*, 2013, Piedimonte, 2013), human rhinovirus (HRV) (Camargo, 2013, Engelmann *et al.*, 2013, Kotaniemi-Syrjanen *et al.*, 2003), human metapneumovirus (HMPV) (Schuster and Williams, 2013, Edwards *et al.*, 2013), human parainfluenza virus (HPIV) (Henrickson, 2003, Henrickson and Savatski, 1997) and human enterovirus (Lu *et al.*, 2013).

The mechanism from which viruses induce asthma is still controversial. Some studies consider that children born with a deficient innate immune response also tend to develop viral infections (Okayama, 2013). This in turn has been related to specific polymorphisms in genes that have been functionally implicated in the innate immune response including: *VDR, JUN, NOS2A* and *IFNA5* (Janssen *et al.*, 2007, Jackson *et al.*, 2012) as well as *SFTPA* (surfactant protein A), *SFTPD* (surfactant protein D), *TLR* 4, *TNF*, IL4, *IL9, IL10, IL8, IL13, IL4RA,* and *CCL5* (Huckabee and Peebles, 2009*)*. There are however notably more publications and evidence suggesting that viral infections tend to skew the immune response (Lemanske *et al.*, 2005, Sigurs *et al.*, 2000, Schauer *et al.*, 2002, Stern *et al.*, 2008). There are two potential mechanisms by which viral infections in early life lead to subsequent chronic inflammation: the immune system response and airways remodelling (Gern *et al.*, 2005). For instance, an increased frequency of viral infections in infancy has been shown to result in increased IFN-γ responses at 1 year of age (Friedlander *et*

al., 2005, Culley *et al.*, 2002). Other studies have shown that children that suffered an RSV infection and developed asthma later on life present a higher production of Th2 cytokines (IFNA13, IL15, IL10, and IL17) as well as cytokine receptor signalling molecules (IL4R, IL9R, and STAT1) (Castro *et al.*, 2008).

Fungal infections have also been implicated with the development of asthma exacerbations. Fungi are specially related with seasonal asthma phenotypes specifically the genders *Alternaria*, *Cladosporium*, and *Didymella* (Pulimood *et al.*, 2007). Additionally more than 20 fungal species have been implicated due to possessing highly immunogenic antigens (Bowyer *et al.*, 2006), that could potentially result in atopic diseases.

Although to a lesser extent than viruses, bacterial pathogens have also been implicated in causing asthma exacerbations (Edwards *et al.*, 2012). Bacterial species that have been associated with wheezing episodes in childhood include S*treptococcus pneumoniae*, *Haemophilus influenza* and *Chlamydia pneumoniae* (Jackson *et al.*, 2011). These in turn have been linked with a higher total immune cell counts, neutrophilia, sputum IL-8 and free radicals concentrations (8-isoprostane) in patients (Wood *et al.*, 2010). Interestingly when patients diagnosed with asthma are challenged with *S. pneumoniae* they show a delayed activation of the immune response with low concentrations of TNF-α and a higher production of

IL-5 (Otero *et al.*, 2013). Recently the importance of bacteria in the development of asthma has been supported by the development of culture independent molecular techniques. This has allowed more accurate characterization of the human microbiome including that of the airways (Brar *et al.*, 2012, Cox *et al.*, 2013a).

1.8. The Human Microbiome

Ninety percent of the cells in the human body are not human but microbial (Savage, 1977). These microorganisms are part of the commensal microbiota (also known as the microbiome, normal microflora or indigenous microbiota) present on body surfaces that are exposed to the external environment e.g. the gastrointestinal tract, respiratory tract and skin (a total of approximately 300m^2) (Tlaskalova-Hogenova *et al.*, 2004). It has therefore been suggested that we should perhaps regard ourselves as 'super-organisms' together with the indigenous microbes (Lederberg, 2000).

Most of the microorganisms in a natural environment, including the microbiome in animals and humans, live in mixed populations (Staley and Konopka, 1985). Culture based techniques are the most frequent method used for bacterial identification but these techniques have many limitations including long culturing times, demanding growth requirements and isolation difficulties. It has been estimated that only approximately 1% of the total bacteria in a sample can be

cultured using standard conditions (Relman, 2002). These limitations in culture based methods means that the full extent of bacterial diversity has the potential to be underestimated. Similarly novel organisms may remain undetected despite the fact they could play a significant role in disease causation.

By using more rapid and accurate bacterial identification techniques morbidity and mortality rates could be significantly improved and there is the potential of streamlining and reducing the use of expensive wide spectrum antibiotics (Relman, 1999). In the diagnosis of airways infections, it is already recognised that culture techniques have limitations. For example it has been estimated that approximately 30-40% of purulent sputum samples from patients with bronchiectasis fail to give a positive result with standard clinical microbial culture (King, 2011). As a result, the most powerful approach to study the microbial diversity in a community would be the use molecular culture-independent techniques.

1.9. The Gut Microbiome and Asthma

The first study of the microbiome that compared asthmatics with healthy individuals was in fact not focused on the airways but instead examined the microbial communities of the gut (Murray *et al.*, 2005), Moreover several studies have encountered that the intestinal microbiome plays a very important role in the development of the immune system from early infancy (Torrazza and Neu, 2011,

Vael and Desager, 2009). Recent studies have shown that the microbiome in turn can modulate the development of the immune response of the host (Kaplan *et al.*, 2011, Vael and Desager, 2009). The immune response produced depends on the balance of particular bacteria in the microbiome and this dictates whether a balanced or non-balanced Th1/Th2 immune response occurs and can result in enhancement of the systemic innate immunity (Kaplan *et al.*, 2011). It has been suggested that the mechanism by which this happens is direct activation of the immune response by bacterial metabolites and antigens (Allan and Devereux, 2011). It has also been postulated that some bacterial metabolites could affect epigenetic gene expression control in the host (Licciardi *et al.*, 2010).

From birth the immune systems displays a Th2 immune profile (Prescott *et al.*, 1998), but as time progresses, depending on exposures and stimuli, one starts to develop the Th1 and cellular immune response. Both the Th1 and Th2 arms of immunity in turn influence the microbial communities that develop in the intestine (the gut microbiome) (Holt *et al.*, 2005). In addition, nutrients, immune cells, cytokines and maternal antibodies in the breast milk can also impact and modulate the gut microbiome of the neonate. This specifically occurs via TLR mediated immune response (LeBouder *et al.*, 2006).

1.10. The Airways Microbiome in Healthy Individuals

The airways anatomy starts in the air entrances corresponding to the nose and the mouth, and continues through the pharynx, larynx, trachea and bronchial tubes until reaching the alveoli of the lungs. The respiratory airways system is a complex interplay between the ciliated mucosal epithelia, the mucus, connective tissue, the immune system and the musculature (Ward *et al.*, 2006). There are different types of epithelia forming the mucosa in the airways, different types of muscle cells, different ratios of connective tissue on the airways course, and moreover different environmental conditions such as temperature, humidity and air pressure. As a consequence such environments mean that each different area harvested contain different types of microorganisms (Hilty *et al.*, 2010), and the bacterial microbiome in the airways varies in abundance and diversity in their anatomical trajectory. There are three main unique established bacterial niches in the airways: nasopharynx, oropharynx and lower airways-alveoli (Hilty *et al.*, 2010).

The healthy lower airways in humans historically were considered sterile using standard culturing methods (Thorpe *et al.*, 1987). With the advancement in sequencing technologies and the capability to conduct more high throughput studies, newer and more sensitive culture-independent methods have revealed that bacterial communities are permanent residents of the lower airways and that different

microbiome configurations are related with the development of disease (Hilty *et al.*, 2010).

Studies have revealed that the nasopharynx (nasal) microbiome has a lower diversity of bacteria (in terms of species number) when compared with that of the oropharynx (Lemon *et al.*, 2010, Huse *et al.*, 2012). The nasal microbiome was found to be dominated by Actinobacteria *(*represented mainly by *Corynebacterium* spp. and *Propionibacterium* spp,), Firmicutes (represented mainly by *Staphylococcus* spp.), and Proteobacteria (represented by *Enterobacter* spp.) (Hilty *et al.*, 2010, Lemon *et al.*, 2010, Frank *et al.*, 2010). More than 50% of the bacteria present in the nasopharynx are Gram-negatives compared with the oropharynx where more than 80% of its community are Gram-positives (Bogaert *et al.*, 2011). The microbiome in the nostril and nasopharynx shows strong similarity to that seen for the skin microbiome due to the high presence of *Staphylococcus* spp. (Lemon *et al.*, 2010, Huse *et al.*, 2012, Costello *et al.*, 2009).

Several species of staphylococci are present in the nasal microbiome principally *S. aureus* and *S. epidermidis* (Frank *et al.*, 2010). The nasal mucosa is known to be a suitable carriage environment for difficult to treat methicillin-resistant *S. aureus* (MRSA) (Gupta *et al.*, 2013). Although *S. aureus* can act as a commensal in the healthy nasal microbiome, some virulent strains can grow at higher rates and

produce an impact on the bacterial community structure, specifically diminishing the abundance of the commensals *Propionibacterium acnes* and *Sthaphylococcus epidermidis (Frank et al., 2010).* Interestingly using a metagenomic approach, a seasonal variability has also been observed in the nasopharynx microbiome. During fall and winter it was found that there was a higher prevalence of Proteobacteria and Fusobacteria, whilst in spring a preponderance of Bacteroidetes and Firmicutes (mainly caused by an increase in *Bacillus* and *Lactobacillus* spp.) existed (Bogaert *et al.*, 2011).

The oropharynx normal microbiome has been shown to be characterized by a high prevalence of Firmicutes, Proteobacteria and Bacteroidetes (Hilty *et al.*, 2010, Cardenas *et al.*, 2012, Huse *et al.*, 2012). The bacterial species within these groups correspond in many instances with those seen in the microbiome of the saliva and are substantially less similar to those of the oral, gingival and esophagic microbiome (Lemon *et al.*, 2010). When alpha diversity was compared between gingiva, tongue, buccal mucosa and saliva, the oropharynx was found to have lower values (Huse *et al.*, 2012). At genera level *Prevotella, Streptococcus, Staphylococcus, Corynebacterium, Veilonella, Haemophilus* and *Neisseria* were found to be the normal commensals in the oropharynx (Hilty *et al.*, 2010, Charlson *et al.*, 2010, Cardenas *et al.*, 2012). If age is taken into consideration, the most abundant genus

seen adults is *Prevotella* whilst in children *Streptococcus* is the most common (Charlson *et al.*, 2010, Hilty *et al.*, 2010),

The lower airways have a lower biomass of bacteria compared with the upper airways (Hilty *et al.*, 2010, Charlson *et al.*, 2011) but their community structure is similar to that of the oropharynx (Hilty *et al.*, 2010) (Figure 1.4). Bacteroidetes, Firmicutes and Proteobacteria were found to be the most abundant phyla in the lower airways microbiome. At genera level the most relevant bacteria were *Prevotella, Haemophilus, Pseudomonas, Streptococcus, Fusobacterium, Veillonella* and *Porphyromonas* (Hilty *et al.*, 2010, Erb-Downward *et al.*, 2011). Interestingly when compared with the oropharynx, *Haemophilus* spp. were found to be more prevalent in the lower airways (Hilty *et al.*, 2010). Whether the left lung varies from the right lung was investigated by Charlson *et al.* (2010), who found no dissimilarities.

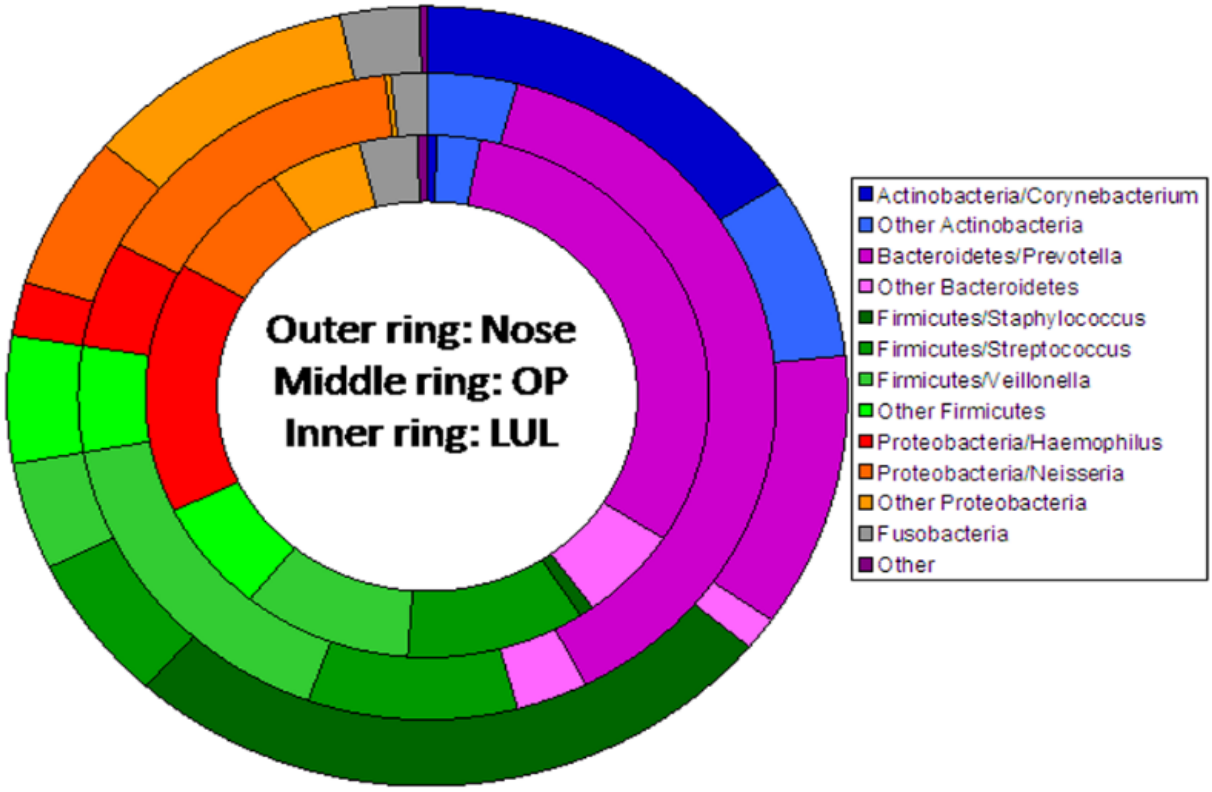


Figure 1.4. The Regional Differences in the Airways Microbiome. Percentage distribution of common phyla and genera at different airway levels (nose, OP and LUL), subdivided into the seven most frequent genera *Croynebacterium*, *Prevotella*, *Staphylococcus*, *Streptococcus*, *Veilonella*, *Haemophilus and Neisseria*) found in the samples (Hilty *et al.*, 2010). Taken under permission (see appendix).

Variations in the oral microbiome related with life-style activities have been found. In patients suffering alcoholism there was a higher prevalence of *Bacteroides, Prevotella, Veilonella, Fusobacterium, Peptostreptococcus, Propionibacterium, Bifidobacterium, Clostridium* species and enterobacteria compared with non-alcoholic adults (Golin *et al.*, 1998). Smoking has also been found to have an effect on the upper airways microbiome, with both the naso and oropharynx bacterial ecology being more diverse in smokers than in non-smokers (Charlson *et al.*, 2010,

Erb-Downward *et al.*, 2011). At genera level smokers showed an increase in the anaerobic Gram-negative *Megasphaera* spp. and other potential pathogenic bacteria such as *Streptococcus, Veillonella, Actinomyces* and *Atopobium* spp. In contrast there were decreased level of species including *Peptostreptococcus, Capnocytophaga, Fusobacterium* and *Neisseria* (Charlson *et al.*, 2010).

Although there is very limited data to draw definitive conclusions about the effect of latitude on the airway microbiome, one could anticipate that climate as well as ethnicity, culture and access to antibiotics could be additional important influences on the airway microbiome (Costello *et al.*, 2009). In a study on cystic fibrosis in which the microbiome of the lower airways was compared between patients from the United Kingdom and United States, both groups were found to cluster separately as well as vary on community diversity and structure (Stressmann *et al.*, 2011).

1.11. The Airways Microbiome in Asthma

The microbiome mediates mechanisms of immunological tolerance and is essential in regulating mucosal inflammation (Yamanaka *et al.*, 2003, Herbst *et al.*, 2011). Consequently it could contribute to the development of atopic disease and asthma in children (Noverr and Huffnagle, 2005). Several studies have found that an early exposure during infancy to infections and antibiotics could cause a higher prevalence of asthma and atopy (McKeever *et al.*, 2002, Droste *et al.*, 2000). A

number of studies have found a relationship between distinct airways microbiome patterns between asthmatics and healthy controls. In a study comparing the airways microbiome of asthmatic adults with that of healthy age matched controls a higher frequency of Proteobacteria was seen in cases with *Haemophilus* spp. more abundant in asthmatics (Hilty *et al.*, 2010). Bactereoidetes, Firmicutes and Actinobacteria were more common in controls (Hilty *et al.*, 2010, Marri *et al.*, 2013). Additionally bacterial DNA quantification has shown that asthmatics have larger bacterial burdens than controls (Huang *et al.*, 2011).

In poorly controlled asthmatics, bacterial genera not commonly found in the airways have been observed to be more abundant (Huang *et al.*, 2011). These included *Nitrosomonas, Oxalobacter formigenes, Comamonadaceae and Sphingomonadaceae*. Interestingly these particular genera have been observed previously in cystic fibrosis (CF) patients and may be related with high bronchial hyper-responsiveness (Huang *et al.*, 2011).

In a birth cohort in children using periodical culture based bacterial identification showed that *Streptococcus pneumoniae, Haemophilus influenzae, Moraxella catarrhalis* and *Staphylococcus aureus* are more common in wheezing children and are related with future development of asthma (Bisgaard *et al.*, 2007). In the first culture-independent study of bronchoalveolar lavage taken from difficult

asthmatic children, a higher prevalence of Proteobacteria was related with asthma development, particularly *Haemophilus* and *Staphylococcus* species whilst *Prevotella* was considered a protective genus (Hilty *et al.*, 2010).

Particular airways microbiome patterns also have been associated with other respiratory diseases including chronic obstructive pulmonary disease (COPD) (Erb-Downward *et al.*, 2011), cystic fibrosis (CF) (Klepac-Ceraj *et al.*, 2010) and non-CF bronchiectasis (Rogers *et al.*, 2014).

1.12. How to Study the Human Microbiome?

Culturing has been since more than a century ago the most commonly used technique for bacterial identification. This method involves the growing of a specific type of bacteria in media (containing specific nutrients for the type of bacteria studied), and the observation of a colony growing on a plate. This approach has fundamentally not changed since Koch presented it at the International Medical Congress in London in 1881 (James, 2009) and has been used for numerous environmental and clinical studies (Davis, 2014, Burns and Rolain, 2014, Joint *et al.*, 2010).

Since a couple of decades ago with the development of culture-independent techniques, it has been determined that between 5 to 50% of bacteria (depending of

the sample type being environmental or human) can be cultured whilst the rest remained unstudied (Hengstmann *et al.*, 1999, Chin *et al.*, 1999). The recent development of culture-independent techniques and the studies in mucosal immunity has resulted in an increased interest in studying the human microbiome. The Human Microbiome Project (HMP) was established in 2008 with its main goal being the identification and characterization of microbial communities associated with healthy and diseased states. The first phase of the HMP has been focused on sequencing microbial reference genomes particularly the most commonly found bacteria in the gut. So far 356 genomes have been sequenced including members of two kingdoms (Bacteria and Archaea), 9 phyla, 18 classes, and 24 orders that are part of the microbiome (Nelson *et al.*, 2010). Other large-scale studies already conducted for the gut microbiome have resulted in the identification of at least 1800 genera (90% ID), 16000 phylotypes at the species level (97% sequence identity) and 36000 phylotypes at the strain level (99% sequence identity) (Frank *et al.*, 2007a). Only 20% of these organisms have however been cultured in the laboratory using conventional techniques (Eckburg *et al.*, 2005).

The majority of the molecular tools used for the characterization of the bacterial microbiome are based on the genomic evolutionary relationships (Cox *et al.*, 2013a) (mainly in the similarities of housekeeping genes) between the bacterial

genomes. Comparisons of the 16S rRNA gene (which encodes for a section of the RNA that constitutes the small subunit of the ribosome) are the most popular techniques to classify bacteria phylogenetically due to the gene possessing both highly conserved sequences as well as highly variable segments and the ease with which evolutionary relationships can be identified (Wang *et al.*, 2005, Turnbaugh *et al.*, 2007, Grice *et al.*, 2008). This gene is relatively short with 1,542 nucleotides of longitude, and is part of the 30S subunit of the bacterial ribosome (Woese and Fox, 1977). It consists of conserved and variable regions (Figure 1.5), the latter allowing differentiation between different organisms with the former being utilised for primer design to allow comprehensive amplification of all types of bacteria (Woese and Fox, 1977).

Several experiments have been developed using the 16S rRNA gene to identify the bacteria existent in the microbiome. One of the first approaches that allowed identifying several bacterial taxons in a mixed microbial community is denaturing gradient gel electrophoresis (DGGE). This technique uses 16S rRNA PCR, and is based on the differences in the variable regions on this gene between different bacteria (McAuliffe *et al.*, 2005). Such differences produce variation on the migration of the amplicons in a denaturant gel, which produce different band patterns (see Chapter 2: Section 2.6). As there is the need to use a marker for taxonomical

identification, and invariably bands cannot be assigned with 100% confidence to any bacterial genus, the technique is often complemented with cloning and sequencing of the amplicons. Although more labour intensive, this provides a better insight of the bacterial abundance present in the microbiome, which allows a more accurate ecological analysis (Ahmed *et al.*, 2007). Usually using this approach, bacteria can be assigned to a specific genus and, in some cases, to specific species by comparing the obtained sequence to those of already characterized species in the GeneBank and Ribosomal DNA databases (Wang *et al.*, 2005).

Currently, the new next generation sequencing techniques such as 454 Roche pyrosequencing and later Illumina HiSeq and MiSeq have allowed hundreds thousands of sequences to be generated in a few hours avoiding the creation of genomic libraries and subsequent cloning. These new techniques also allow multiplexing of numerous samples at the same time using barcode sequences or MIDs (a unique sequence for each sample is used in the 3' extreme of the reverse primer) in the PCR step. This then permits the discrimination between samples in the subsequent data analysis step. All these advantages have allowed the characterization on a large scale of very complex microbial communities (Cox *et al.*, 2013a).

Bioinformatics tools have also become indispensable in the analysis of the microbiome. Tools have been developed widely to allow a more easy analysis of the huge loads of data produced by next generation sequencing. New algorithm calculations have been set to determine operational taxonomic units (OTUs) (Edgar, 2010a), annealing sequences (Caporaso *et al.*, 2010a), taxonomy assignation (McDonald *et al.*, 2012), phylogeny (Price *et al.*, 2010b) as well as diversity calculations (Lozupone and Knight, 2005b).

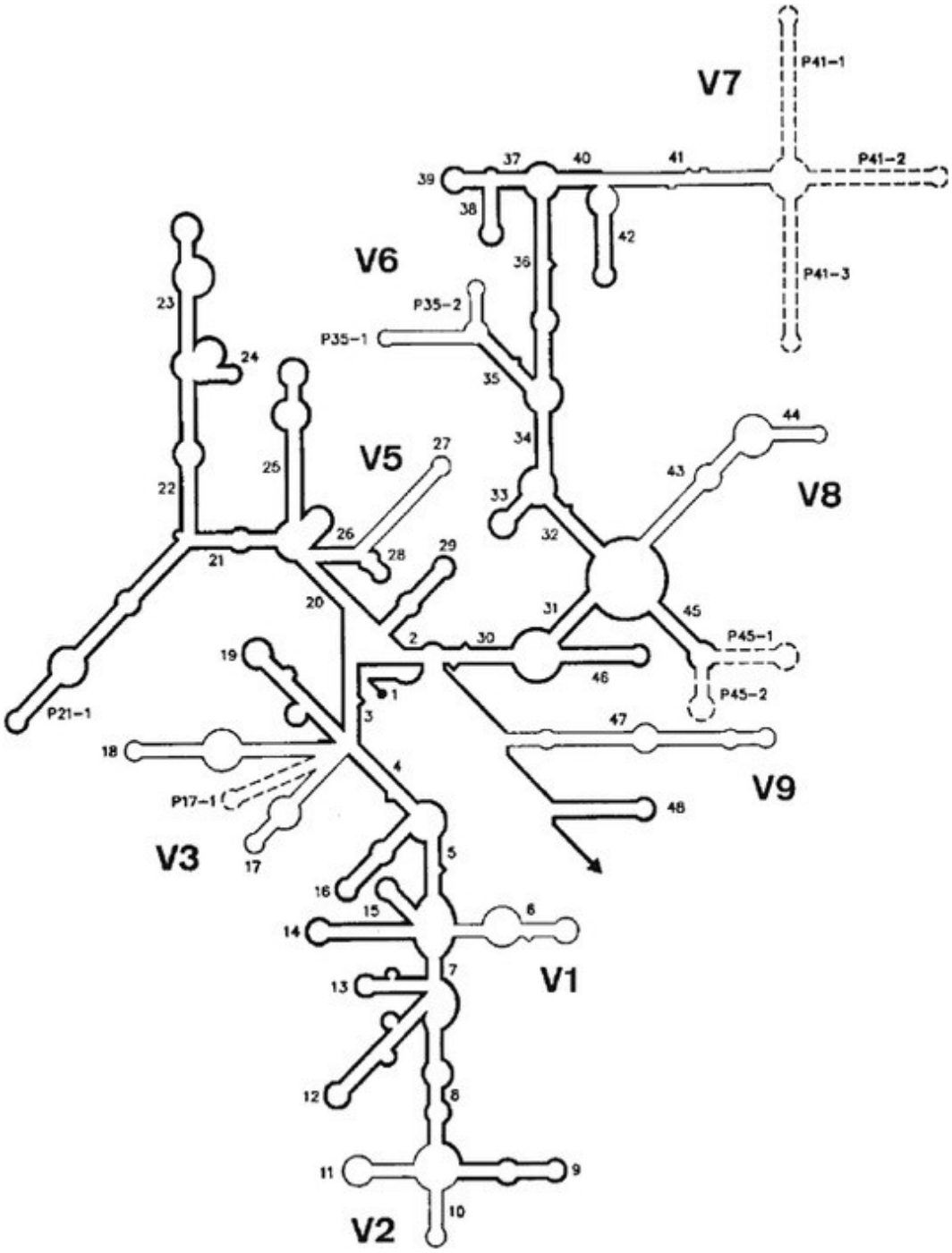


Figure 1.5. The Bacterial 16S rRNA gene. Secondary structure model of the small subunit rRNA, eight variable regions have been identified in 16S rRNA, which are termed V1 to V3 and V5 to V9 (region V4 occurs only in eukaryotic 18S rRNA. These Variable regions are used for Molecular based bacterial identification. Taken under permission (see appendix) from Lawson (2004).

1.13. The Microbiome and the Immune System Development

Many molecular pathways have portrayed the crosstalk between the microbiome and the immune cells. The principal relationships are through the direct contact of microorganisms with the epithelia sited at the mucosal barriers (Kabat *et al.*, 2014). The toll-like receptors (TLR) are involved in the continuous stimuli of the microorganisms to the epithelial cells, which in turn secrete epithelial protective cytokines like TNF-α, keratinocyte chemokine-1 (KC-1), and heat shock proteins (Rakoff-Nahoum *et al.*, 2004). Additionally short chain fatty acids like butyrate produced by the intestinal bacteria, induce IL-18 expression through the G protein-coupled receptor (GPR) 109A and regulates the apoptotic signals on epithelial cells (Kalina *et al.*, 2002), and strikingly bacteria that produce these fatty acids are reduced in chronic inflammatory states in the gut like inflammatory bowel disease (Frank *et al.*, 2007b). Through TLR-2 activation in epithelia and in dendritic cells (DC), the microbiota also stimulates the acquired immune response by secretion of B cell activating factor of the TNF family (BAFF), a proliferation–inducing ligand (APRIL) and transforming growth factor (TGF)-β. These cytokines promote the transformation of B-cells into plasma cells (Hapfelmeier *et al.*, 2010). Consequently alterations in the microbiota patterns could trigger the immune response to produce

a chronic state of inflammation with a Th2-bias (Kabat *et al.*, 2014), which in the airways could be expressed as asthma.

Environmental exposures early in life are critical determinants of host immune development and function (Cooper *et al.*, 2009). Early exposures to microorganisms may occur through an infected mother (i.e. trans-placental transfer of antigens) (Guadalupe *et al.*, 2009), through breast-feeding (transfer through breast milk), and exposures to infections in infancy may 'program' infant immunity to a Th2-bias that is related with atopy and asthma development (Cooper *et al.*, 2006). Interaction between the mucosal microbiome and immune cells is important in the development of the immune system. The microbiome mediates mechanisms of immunological tolerance and is essential in regulating mucosal inflammation. Modifications in microbiome patterns during infancy could therefore alter the subsequent host immune responses, and might be essential in preventing mucosal inflammation (Yamanaka *et al.*, 2003). Consequently they could contribute to the development of atopic disease and asthma in children (Noverr and Huffnagle, 2005, Ege *et al.*, 2011).

Several studies conducted in adults and children have supported the statement that altered gut (Huang, 2013, Russell and Finlay, 2012) or airways bacterial microbiome (Hilty *et al.*, 2010) are related with a higher prevalence of atopic

disease and asthma. This could be one of the reasons why the 'hygiene hypothesis' is influential in allergic disease, explicated by the antibiotic use, the westernized diet and a lack of early microbial stimulation.

For these reasons for this thesis a case/control study was established in order to determine the airways microbiome patterns in infancy and relate them with the development of wheezing. The project hypothesis was done taking into consideration the findings of Hilty *et al*. (2010) for adults:

Project hypothesis

Short and long term colonization of the airways with Proteobacteria is associated with the development of early onset wheezing illness during the first 2 years of life whilst long term colonization with Firmicutes is a protective factor against the development of wheezing in the rural tropics.

Projects Aims

To identify bacteria associated with increased or decreased risk of wheezing in early life.

Determine the upper airways microbiota patterns of infants during an acute episode of wheezing, and to monitor the long-term relationships between wheezing and the microbiota.

Determine what changes occur in the upper airways microbiota over time by characterizing the microbiota present at 7 months, 12 months and 24 months of age in healthy children.

Chapter 2: Materials and Methods

2.1. List of Solutions and Kits used

- 454 Junior Titanium Sequencing kit (Roche Applied Biosystems, UK)
- Acrylamide (Sigma, UK)
- Ampicillin 500mg (Pfizer, UK)
- AMPure XP beads (Roche, UK)
- APS (Ammonium Persulfate) (Sigma, UK)
- Beads Recovery kit (Roche Applied Biosystems, UK)
- Brain Heart Infusion (BHI) broth
- Butanol 100% (Sigma, UK)
- DMSO (Dimethyl sulfoxide) (Sigma, UK)
- dNTPs (0.2 mM each of A, T, C and G) (Roche Applied Biosystems, UK)
- Emulsion PCR kit(Roche Applied Biosystems, UK)
- Ethanol molecular biology gradient (Sigma, UK)
- Ethidium Bromide (Sigma, UK)
- FastStart Hifi Polymerase (Roche Applied Biosystems, UK)
- Formamide 100% (Sigma, UK)
- Gradient marker GM-100 (C.B.S. Scientific, Del Mar, USA)

- High-efficiency competent JM109 *E. coli* cells (Promega, UK)
- HotStart Polymerase (Roche Applied Biosystems)
- IPTG (isopropyl thiogalactoside) (Sigma, UK)
- Isopropanol 100% (Sigma, UK)
- LB media (Lysogeny broth) (Sigma, UK)
- $MgCl_2$1.5 mM (Roche Applied Biosystems, UK)
- MPBio Kit for Soil (Life Technologies, UK)
- Oil and breaking kit (Roche Applied Biosystems, UK)
- PCR buffer 10x (Roche Applied Biosystems, UK)
- PCR sterile water (Qiagen, UK)
- pGEM-T Easy Vectors (Promega, UK)
- QIAmp DNA Mini Kit (Qiagen, UK)
- Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies, UK)
- Ready-Lyse™ Lysozyme Solution (EPICENTRE, UK)
- Sterilin swabs (Fisher, UK)
- TAE 50X (Tris Acetate EDTA buffer pH 8.0) (Sigma, UK)
- TE 10x (Tris EDTA pH 8.0) (Sigma, UK)
- TEMED (Tetramethylethylenediamine) (Sigma, UK)
- Urea (Sigma, UK)

- Wizard PCR Clean Up Kit (Promega, UK)
- XGal (5-bromo-4-chloro-3-indolyl-beta-D-galactopyranoside) (Sigma, UK)

2.2. Study Population

Oropharynx throat swabs samples were collected as part of the Wellcome Trust-funded Birth Cohort study (ECUAVIDA) in Ecuador. The aim of the ECUAVIDA cohort study is the investigation of the effects of early infant infections on the development of immunity, allergic sensitization and allergic disease. This ongoing study is an unselected population-based birth cohort that has recruited 2,403 new-

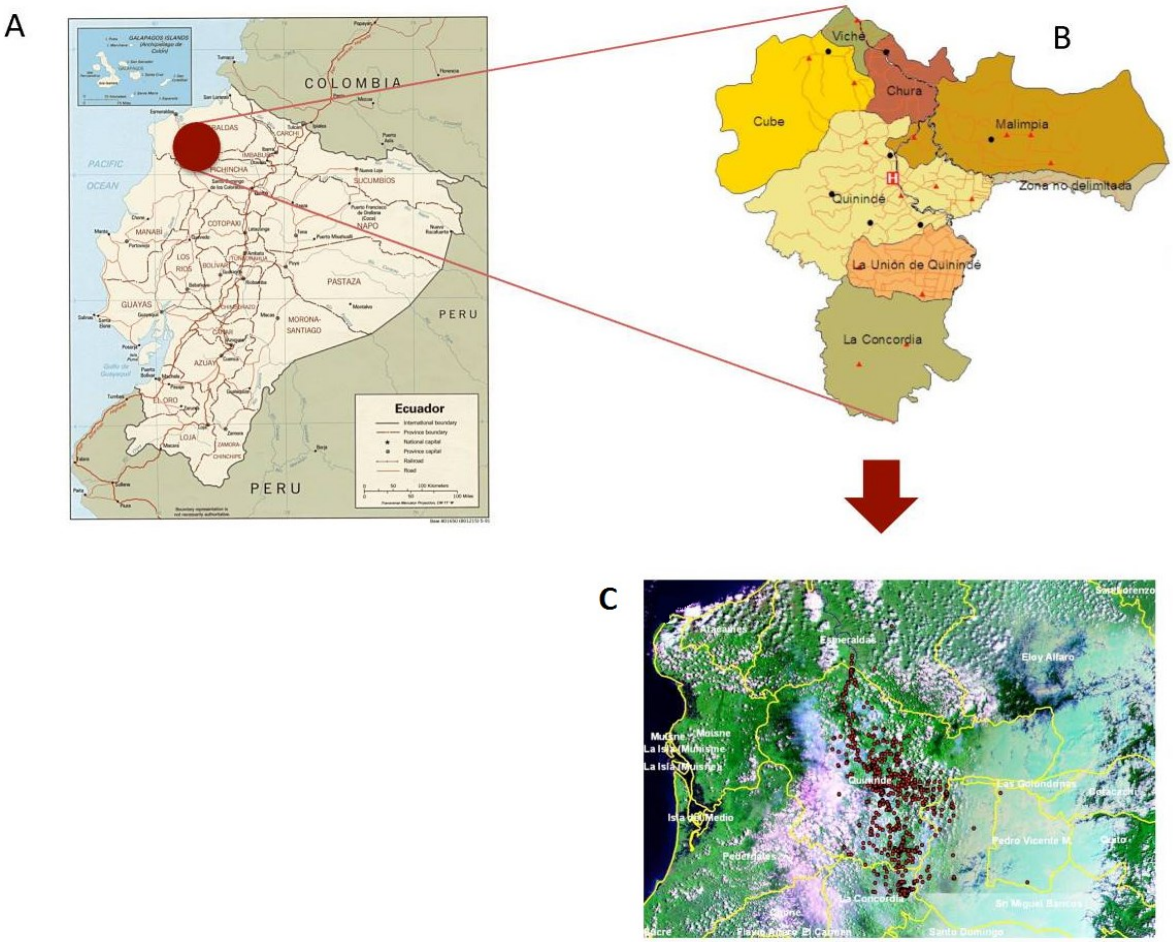


Figure 2.1 Study Site. A) Map of Ecuador showing location of district of Quinindé, Esmeraldas province (red oval). B) Map showing parishes within the district of Quinindé including H-Hospital Padre Alberto Buffoni. C) Geographic location of households of cohort infants. Modified from Cooper *et al.* (2011) under permission (see appendix).

borns in the rural District of Quinindé in the Esmeraldas Province of Ecuador (Figure 2.1) (Cooper *et al.*, 2011).

Esmeraldas Province is located in the North-West coast of Ecuador and is considered one of the poorest provinces of the country having less than USD2,000 per capita income. The rural District of Quinindé includes an extension of 3,471 km^2 and therefore contains an estimated 150,000 people within one urban and six rural counties. The District is located in a Tropical secondary rainforest, with an average annual day temperature of 30°C with 75% humidity, and is located at an altitude of 100m above sea level (Cooper *et al.*, 2011).

The District has an ethnically mixed population with the majority of individuals being mestizos (90%), with the remaining either of Afro-Ecuadorian ancestry (7%) or Amerindians (3%). Seventy-eight percent of the population lives in the rural counties of the District and only 22% in the main town (urban county). In the urban county, approximately 90% of the population have access to electricity, 60% to portable water, 40% to sanitation; and 60% to solid waste disposal services. In contrast for the rural zones, only 10% have access to electricity and none have access to the other basic services. The main Quinindé District sources of income are derived from agriculture, specifically African palm oil and fruit farming, although livestock cattle farming also occurs, and timber extraction.

The ECUAVIDA study is facilitated through the Hospital "Padre Alberto Buffoni" (HPAB) in Quinindé town, which is the only Hospital in the District and therefore the provider of maternity facilities for the region (Cooper *et al.*, 2011).

2.3. Project Design

As part of the ECUAVIDA cohort, detailed data has been collected from the mothers at the time of their first antenatal visit using questionnaires and during periodic health check-ups at different children ages. Additionally, environmental and clinical samples have been taken during respiratory and gastrointestinal diseases, which were stored in -80ºC freezers. Questionnaires data were recorded in the general database of ECUAVIDA. For the present study a sub-cohort of the ECUAVIDA was sub-selected based on the patients' age and clinical characteristics. The ECUAVIDA project has been given full ethical approval by the Universidad San Francisco de Quito and the Ecuadorian Health Ministry ethical committees. The Wellcome Trust funds totally this project. The present study is a case control retrospective study, divided into two stages (1 and 2) according to age. Stage 1 has additionally been subdivided into two parts according to the phase of disease when samples were collected: Part 1a during a current wheezing episode and Part 1b during a past wheezing episode.

Stage 1 is focussed on early onset wheezing (cases) and healthy (controls) children between 7 and 12 months of age. The aim of this stage of the project is to profile the upper airways microbiota for the early months of life.

Stage 1a tries to determine the acute changes in upper airways microbiota related with a non-infectious wheezing episode. Consequently, the case samples were collected when infants had a current wheeze episode (not caused by any evident respiratory infection) according to the Global Initiative for Asthma [GINA] 2009, and at least one episode of wheezing that had previously been diagnosed by a physician. Healthy controls were paired by age to cases and had no wheezing history, current respiratory disease, chronic disease or current infections.

Stage 1b was designed to determine the mid-term changes in upper airways microbiota related with non-infectious wheezing episodes. Consequently, the cases were collected when infants had at least two non-infectious wheeze episodes history according to GINA with at least one having been diagnosed by a physician. When sampling was performed all cases were asymptomatic.

Stage 2 was intended to identify the long-term effects of previous non-infectious wheezing episodes. For this reason, subjects and sampling was almost

identical to that of Stage 1b (also using the same infants from stage 1b) with the exception that sampling was done when children were 24 months of age.

For Stage 1a samples were collected from a total of 75 children. One sample failed downstream processing (Section 2.4) and consequently in the final analysis 74 samples were included (24 cases and 50 controls). For Stage 1b, 180 samples were collected of which 172 samples were processed successfully allowing their inclusion in the final analysis (72 cases and 100 controls). As part of Stage 1b, the samples from children collected at 7 months were 30 cases and 50 controls and at 12 months were 42 cases and 50 controls. For Stage 2 although 100 samples were collected, post-processing only 88 were suitable for inclusion in the final analysis (38 cases and 50 controls). Details of selection criteria for Stages 1a, 1b and 2 are summarised in Table 2.1.

All samples were collected when infants (both cases and controls) did not have any evidence of a current airway infection (absence of cold symptoms and fever). The infants did not receive antibiotics and corticosteroids for any reason at least two weeks prior to sampling. All the subjects were of the same mixed ethnical background (mestizos), lived in the same town and had access to the same basic services (electricity, water, sanitation, etc.). All of them received the vaccines recommended by the Ecuadorian Public Health Ministry (BCG, mumps, measles and

rubella, tetanus, diphtheria, hepatitis B and *Haemophilus influenza* B; MSP, 2005). None of the subjects received anti-Streptococcal vaccination (not compulsory for the MSP).

Stage		Number of samples included	Age	Cases		Controls			
				Inclusion Criteria	Exclusion Criteria	Inclusion Criteria	Exclusion Criteria		
1	1a	74 samples (24 cases and 50 controls)	7-12 months	a) A current non-infectious wheezing episode during the sample collection diagnosed by a physician in the medical control	a) Use of antibiotics in the last 2 weeks b) Any respiratory infection in the last 2 weeks	No history of wheezing reported	a) Use of antibiotics in the last 2 weeks b) Any respiratory infection in the last 2 weeks		
	1b	172 samples in total (72 cases and 100 controls) 7 months: 30 cases and 50 controls 12 months: 42 cases and 50 controls.				a) At least 1 non-infectious wheezing episode reported during the first 12 months of life and diagnosed by a physician in the medical control b) No current wheezing episode e	a) Use of antibiotics in the last 2 weeks b) Any respiratory infection in the last 2 weeks	No history of wheezing reported	a) Use of antibiotics in the last 2 weeks b) Any respiratory infection in the last 2 weeks

2	88 (38 cases and 50 controls)	24 months	a) At least 2 wheezing episodes reported during the first 24th months of life and diagnosed by a physician in the medical control b) No current wheezing episode	a) Use of antibiotics in the last 2 weeks b) Any respiratory infection in the last 2 weeks	No history of wheezing reported	a) Use of antibiotics in the last 2 weeks b) Any respiratory infection in the last 2 weeks

Table 2.1. Detailed information about inclusion and exclusion criteria during sample selection.

Controls in all the stages were matched by age to cases and had no history of wheezing episodes or any chronic morbidity, and were therefore defined as ‘healthy’.

2.4. Throat Swabs: Collection and Storage

Throat swabs were collected using sterile cotton swabs (Fisher, UK) and were placed in collection tubes (Fisher, UK). Sampling was performed carefully without touching any surface (i.e. tongue, mouth, and teeth) other than the oropharynx and using a tongue depressor.

Each swab was rubbed approximately five times around the oropharynx, applying an even pressure and rotating the swab without interruption. The swab was immediately placed back into the collection tube and stored at -20^{o}C and subsequently at -80^{o}C. Samples were shipped on dry ice under careful temperature control to the National Heart and Lung Institute, Imperial College London for DNA extraction.

2.5. Bacterial DNA Extraction from Throat Swabs

Bacterial DNA was extracted from the 400 throat swabs using a modified protocol of the commercial QIAmp DNA Mini Kit (Qiagen). Additional steps at the

beginning of the protocol were included to improve the lysis of Gram positive (+ve) bacteria (Hilty *et al.*, 2010) and are detailed below.

Lysozyme (4X concentration = 1000U/µl) was prepared from lysozyme stock (30,000 U/µl Ready-Lyse™ Lysozyme Solution, EPICENTRE, UK). Each swab head was transferred into a 2 ml sterile microcentrifuge tube and 432µl TE (Tris EDTA pH 8.0) plus 18µl 4X lysozyme solution added. Samples were incubated for 1 hour at 37°C to allow improved lysis of Gram +ve bacterial walls. During this hour, samples were vortexed for 20 seconds at intervals of 15 minutes. Next 30µl of Proteinase K and 450µl of Buffer AL were added to the tube and samples were incubated at 56°C for 30 minutes. To terminate the Proteinase K digestion step, samples were incubated for 5 min at 95°C.

Next 450µl of Ethanol (96–100%) was added to the sample and vortexed in order to obtain a homogeneous solution. The solution was then applied to the QIAamp Spin Column as per the manufacturer's protocol.

In the final step 40µl of nuclease free water was added instead of the elution buffer supplied by the kit. If, after extraction, the DNA was not being used immediately samples were stored at -20°C until required.

2.6. Preliminary Microbiome Studies using Denaturing Gradient Gel Electrophoresis (DGGE) and Sanger Sequencing

2.6.1. Amplification of 16S rRNA Gene

Polymerase chain reaction (PCR) was used to amplify the variable region 3 (V3) of the gene that encodes for 16S rRNA in bacteria. A nested PCR approach was implemented for running samples on Denaturing Gradient Gels (Electrophoresis - DGGE). PCR products from the first step of the nested PCR were cloned for the purpose of sequencing.

2.6.2. First Round PCR

For the first round PCR, the V3 region of the 16S rRNA gene was amplified using the following conserved primers:

339F (Forward) $^{5'}$-ACTCCTACGGGAGGCAGCAGT-$^{3'}$

907R (Reverse) $^{5'}$-CCGTCAATTCATTTGAGTTT-$^{3'}$

Reactions with a final volume of 25µl were set up containing: 2µl of template DNA, 2.5µl 10X PCR buffer (Roche Applied Biosystems, UK), 1.5 mM $MgCl_2$, 6% DMSO, 0.5 µM of each primer, 0.8 mM dNTPs (0.2 mM each of A, T, C and G) and

1U HotStart Polymerase (Roche Applied Biosystems). PCR cycling conditions used were:

- Hot start: 95℃ for 5 min,
- Twenty cycles of 95℃ for 30s, 55℃ for 30s and 72℃ for 60s,
- Final extension of 72℃ for 5 min

PCR reactions and cycling conditions had previously been optimized within the Molecular Genetics and Genomics Group at the National Heart and Lung Institute (Hilty *et al.*, 2010).

2.6.3. Second Round PCR

To improve detection of single base pair changes, a GC rich clamp was incorporated in the second round PCR by including it in the design of the forward primer sequence (Sheffield *et al.*, 1989). First round PCR products were diluted with PCR sterile water (Qiagen) 1:10 and 2µl of the dilution was used as a template for the second round PCR amplification. The primers used were:

- 339F (Forward with [in red font] a GC-clamp)

5'-CGCCCGGGGCGCGCCCCGGGCGGGGCGGGGGCACGGGGGGA CTCCTACGGGAGGCAGCAGT-3'

- 539R (Reverse) $^{5'}$-GTATTACCGCGGCTGCTGGCAC-$^{3'}$.

As in the first round PCR, reactions of 25µl were set up containing: 2µl of template DNA, 2.5µl 10X buffer (Roche Applied Biosystems, UK), 1.5 mM $MgCl_2$, 6% DMSO, 0.5 µM of each primer, 0.8 mM total dNTP (0.2 mM each of A, T, C and G) and 1U HotStart Polymerase (Roche Applied Biosystems) (Hilty *et al.*, 2010). PCR cycling conditions used were:

- Hot start: 95℃ for 5 min,
- Twenty cycles of 95℃ for 30s, 55℃ for 30s and 72℃ for 60s,
- Final extension of 72℃ for 5 min

After the second round PCR, PCR products were stored at -20℃ until required.

2.6.4. Denaturing Gradient Gel Electrophoresis (DGGE)

DGGE is a technique based on differences in the intrinsic denaturing behaviour of double stranded DNA fragments which allows separation of fragments that are the same size but that differ in sequence. The electrophoresis is performed in a vertical polyacrylamide gel containing a gradient of denaturant chemicals (Fisher & Lerman, 1979). A denaturant gradient of 30 to 60% was used for the DGGE experiments. The gel was 0.75 mm thick and contained 8% polyacrylamide.

2.6.4.1. Preparation of the Gel

A vertical electrophoresis system PROTEAN II (Bio-Rad, USA) was used for preparation and running of the gels. Both sides of the spacers where covered with a thin layer of silicon grease (the full length of the spacer but only a quarter of the spacer width). The greased side of the spacers was positioned at the outer edges of the glass plates. Sandwich clamps were used to secure the glasses together and to prevent leakage during gel pouring.

The denaturing gradient was created using acrylamide with a 30% and 60% concentration of denaturant. To prepare the 400 ml of 60% denaturant acrylamide-solution, 80 ml of 40% acrylamide, 100.8 g of urea, 96 ml of 100% formamide, 4 ml of 50X TAE (Tris Acetate EDTA buffer pH 8.0) were combined and MilliQ water was added to a final volume of 400 ml. For the 30% denaturant stock solution, 80 ml of 40% acrylamide, 50.4 g of urea, 48 ml of 100% formamide, 4 ml of 50X TAE were combined and MilliQ water was added to a final volume of 400 ml.

High and low solutions of acrylamide (11.5 ml each) were set up for polymerization using 80µl of APS and 5µl of TEMED and the two solutions were placed into the separate chambers of the gradient maker GM-100 (C.B.S. Scientific, Del Mar, USA). Both solutions were mixed in the gradient marker and simultaneously poured into the gel chamber avoiding bubbles. Immediately after pouring, a few ml of

water-saturated butanol were put on top to ensure that a straight surface was obtained. The gel was fully polymerized after 60 minutes.

After polymerization the butanol was flushed from the top of the gel using MilliQ water, the comb inserted, and the stacking gel (consisting of 5 ml 8 % acrylamide 5µl TEMED and 50µl APS) was poured using a Pasteur pipette. Polymerization of the stacking gel took 15 minutes.

An aquarium tank with 1 cm thick glass (length: 44cm, width: 30cm, height: 29 cm) was used. The tank was filled with 23 litres of 0.5X TAE buffer, and set up at 60°C using the immersion circulator and thermostat pump DL-30 (Haake Karlsruhe, Germany).

Eight µl of PCR product was mixed with 2µl of loading buffer (Bio-Rad, UK) and then loaded onto the gel. Additionally two identification markers were loaded in the outer lanes of the gel. These markers were designed by Hilty *et al.* using 16S rRNA amplicons (with the same primers and conditions of the second round PCR). The first marker labelled the ´Asthma Marker´ consists of 16S rRNA amplicons of bacteria that according to previous studies were identified in asthma patients. The marker therefore includes:

- *Haemophilus* sp.

- *Haemophilus influenzae*
- *Finegoldia magna*
- *Moraxella* sp.
- *Prevotella* sp.
- *Streptococcus* sp.
- Moraxellaceae
- *Neisseria* sp.
- *Neisseria meningitidis*
- *Bordetella hinzii*

The ´Normal Marker´ was designed using 16S rRNA amplicons from commensal airway bacteria and includes:

- Bacteroidales
- *Staphylococcus* sp.
- *Haemophilus* sp.
- *Porphyromonas* sp.
- *Prevotella* sp.
- *Neisseria* sp.
- *Veillonella* sp.

The gel chamber was placed in the aquarium and run at 100 Volts (about 20 mA for one gel) for a total of 16 hours. Afterwards, the gel was removed from the chamber and stained in 1 x TBE/Ethidium Bromide at a concentration of 0.5 µg/ml for 60 min and then visualized using an ultraviolet lamp (Hilty *et al*., 2010).

2.6.5. Cloning and Sequencing

2.6.5.1. Ligation/Transformation and Amplicon Libraries Construction

First round PCR products were cloned and sequenced. PCR products were run in a 0.9 % agarose gel. The bands were excised and cleaned using a Wizard PCR Clean Up Kit (Promega, UK). The purified amplicons were re-suspended in 20µl of sterile MilliQ water.

Two µl of purified PCR products were ligated into pGEM-T Easy Vectors (Promega) and the ligated vectors transformed into high-efficiency competent JM109 *E. coli* cells (Promega). Both the ligation and transformation reactions were carried out following the protocol set out in the Promega manual (pGEM-T and pGEMT-T Easy Vector Systems TM042). Clones were grown in LB/Ampicillin/IPTG-XGal medium (Promega) for 24 hours at 37℃.

2.6.5.2. Picking Clones and Sequencing

Ninety-six well plates (well capacity 0.8 ml) were prepared by pipetting 0.2ml LB medium containing ampicillin at 0.2mg/ml into each well. White colonies (positive selection of *lac* –ve colonies that have amplicons inserted) were picked with sterile pipette tips and transferred to an individual well. Plates were shipped to a private sequencing company (Cogenics, UK) and sequenced using the Beckman Coulter Genomics sequencer (USA-UK). The sequencing was performed using the Sanger di-deoxy technology using the SP6 vector primer as the sequencing primer. Post sequencing, files with the sequences (.fasta) and chromatographic traces (.ab1) were received back from Cogenics via email.

2.6.6. Computation Analysis of Preliminary Results

High quality sequences were selected by checking the chromatographic traces using Codon Code Aligner software version 3.5.6 (http://www.codoncode.com/aligner/). Next primer and vector sequence regions were trimmed and full-length sequences ranging from 359 to 906 were analysed. Fasta files were exported and uploaded to the Greengenes project online program (http://greengenes.lbl.gov). Sequences were aligned using NAST and ClustalW (Thompson *et al.*, 2002). Alignments were manually verified and checked for

chimeras using part of the Greengenes project; Bellerophon Version 3 (DeSantis *et al.*, 2006). Chimeras identified were excluded from subsequent analyses.

The DNADist tool (Greengenes) was used to calculate triangular distance matrices. Operational taxonomic units (OTUs) were determined using DOTUR (Distance based OTU and richness determination program) using the furthest-neighbour algorithm with a cut off of 97% sequence identity. Chao 1 richness index and Shannon diversity index was also estimated in DOTUR (Schloss and Handelsman, 2005). Evolutionary distances were calculated with the Jukes-Cantor (tree-Markov) model.

The phylogenetic tree was created using the Neighbour-Joining algorithm (Saitou and Nei, 1987) within Mega 6 (Tamura *et al.*, 2013a). The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura, Nei, & Kumar, 2004). Bootstrapping was performed using 1000 tree replicates and the confidence levels of over 50% were shown at the internal nodes.

The identity of OTUs was defined by comparing the sequence of interest to those within the ribosomal database (20 best match sequences define taxonomic rank) and these were further compared to NCBI BLAST to determine the percentage of sequence identity.

Microbial community comparisons were made using the Bonferroni correction test which corrected the *P*-values for multiple comparisons in UNIFRAC (Lozupone and Knight, 2005a). This is achieved by multiplying the raw *P*-value by the number of permutations. In the single phylum comparison tests, *P*-values were calculated for association between the groups using Two-tailed Fisher's exact tests and by the application of a conservative Bonferroni correction for multiple testing (Durrant *et al.*, 2004).

2.7. Pyrosequencing Analysis of the Microbiome

2.7.1. Amplification of the 16S rRNA Gene

Polymerase chain reaction (PCR) was used to amplify the variable region 3 to 5 of the gene that encodes for 16S rRNA in bacteria. To minimize the PCR nucleotide insertion mistakes, samples were amplified in quadruplicate reactions with 20 cycles each and then pooled.

For the PCR, the V3 region of the 16S rRNA gene was amplified using the following conserved primers:

336F (Forward) $^{5'}$-CCATCTCATCCCTGCGTGTCTCCGACTCAG

CCTACGGGAGGCAGCAG-$^{3'}$

926R (Reverse) 5'-CCATCTCATCCCTGCGTGTCTCCGACTCAG

CCGTCAATTCMTTTRAGT-3'

The Adaptor Sequence in both the reverse and forward primers (highlighted in red) are used to link the biotinylated 5' extreme of the amplicon with a single Enzyme bead in the Emulsion PCR (Section 2.6.4 below).

Multiplexing was achieved using molecular identifiers (MIDs) also known as nucleotide barcodes, which are specific sequences of 12 nucleotides long. There were used the MIDs with the numbers 1 to 300 recommended by the US Human Microbiome Project (http://www.hmpdacc.org).

PCR reactions with a final volume of 25µl were set up containing: 1µl of template DNA, 2.5µl 10X PCR buffer (Roche Applied Biosystems, UK), 1.3% DMSO, 1 µM of each primer (µM), 0.5µl dNTPs (10 mM each of A, T, C and G) and 1.25 U FastStart Hifi Polymerase (Roche Applied Biosystems). PCR cycling conditions used were:

- Hot start: 92°C for 2 min,
- Thirty cycles of 94°C for 20s, 50°C for 30s and 72°C for 5min,
- Final hold 4°C for 5 minutes and then products moved to fridge.

Quadruplicate PCRs were conducted for each sample in order to minimise PCR nucleotide incorporation errors, which would have an unfavourable impact on the downstream sequencing. PCR reactions and cycling conditions had previously been optimized within the Molecular Genetics and Genomics Group at the National Heart and Lung Institute.

2.7.2. Amplicon Cleaning

After PCR, amplicons were purified using the AMPure XP beads (Roche, UK) to eliminate any primer dimer products and the PCR reagent remnants. Purification was performed using a modified and optimized protocol of DNA cleaning using 96 well plates (Magnetic plate Ambion® #AM10027 in combination with 96 well round bottom plates Costar® 3367) instead of tubes. To purify the desired Amplicon the ratio of AMPure beads volume and PCR product was also modified from a 1:1 ratio to a 0.7:1 ratio.

Products from the quadruplicate PCRs were pooled and the resultant 100µl volumes pipetted into a well of the round bottom plate. The AMPure bead bottle was vortexed to ensure all the AMPure beads were completely re-suspended (approximately 20 seconds) and 70µl of beads transferred to each well. Beads and sample were mixed thoroughly by pipetting up and down at least 12 times thereby ensuring a homogeneous was obtained.

Plates were incubated for 10 min at room temperature and then placed on to the 96-well magnetic stand for 5 min at room temperature until the beads were located and fixed to the bottom of the wells. The supernatant was removed and discarded by manual pipetting. The beads were then washed twice with addition of 100µl of 70% ethanol to each well. Plates were then incubated at room temperature placed in the magnetic stand until all pellets were completely dry (10-20 min). After this time, the pellets were re-suspended by addition of 50µl of 1x TE to each well after which the supernatant with the DNA was transferred into a new plate. The latter was stored at -20 ºC until the quantification step.

2.7.3. DNA quantification

Quantification of DNA was performed twice during the protocol. First after the 16S rRNA gene PCR and purification, and the second time after sample pooling (library creation). On both occasions the Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies, UK) was used.

Using the DNA standard (100 ng/µl) provided in the kit and diluting with 1x TE, a full set of standards, including a blank 1x TE control, were created: 100 ng, 50 ng, 25 ng, 12.5 ng, 6.25 ng, 3.13 ng, 1.56 ng. Standards were plated in duplicate into 96-well black fluorometer plates.

To the remaining 80 wells of each 96 well plate, 99µl of TE was added plus one microliter of each sample to be quantified, and mixed by up and down pipetting. To all wells 100µl of a 1:200 dilution of PicoGreen reagent was added resulting in a final total volume of 200µl in each well. The samples were excited at 480 nm and the fluorescence emission intensity was measured at 520 nm using a spectrofluorometer.

For each plate measured, it was verified that the R^2 value of the standard curve was at least 0.98 thereby ensuring accuracy of the quantification. To calculate each sample's DNA concentration it was used the curve equation (y = ax + b). From each fluorometry value it was calculated the DNA concentration using the 'a' and 'b' values.

2.7.4. Pyrosequencing:

The Amplicon libraries that were generated were quantified using the Quant-it Picogreen dsDNA Assay Kit according to the manufacturer protocols (See Section 2.6.3 and www.454.com). The number of molecules per µl was calculated using the formula:

$$\text{Molecules/µl} = \frac{\text{sample conc}[\text{ng}/\mu\text{l}] \times 6.022 \times 10^{23}}{656.6 \times 10^{9} \times \text{amplicon length}[\text{bp}]}$$

Amplicon libraries were then diluted separately to $1x10^9$ molecules/µl in 1x TE Buffer. The libraries were then pooled to create a working solution with every sample containing the same number of molecules per µl. To avoid over-enrichment during the Emulsion PCR prior to pyrosequencing, a maximum of 0.5 molecules of amplicon per bead was calculated to using the formula:

$$\text{µl of DNA library per tube} = \frac{\text{desired molecules per bead x 10 million beads}}{\text{library concentration (in molecules/µl)}}$$

Emulsion PCR was carried out using the DNA libraries previously constructed. There were combined the amplicons with the enzyme beads using their linked adaptor sequence. Prior optimisation within the Molecular Genetics Group of the emulsion PCR meant that a slightly modified master mix, to that recommended by Roche, was used consisting of molecular biology grade water 397.5µl, additive 515µl, amp mix 282.5µl, amp primer 80µl, enzyme mix 70µl, PPiase 2µl on a total volume of 1347µl (part of the 454 Junior Titanium Sequencing kit).

Post emulsion PCR beads were captured using a vacuum pump and cleaned from the emulsion oil (beads recovery kit) using isopropanol, ethanol and finally the enhancing buffer provided on the kit. Only the beads successfully ligated to an amplicon were recovered using magnetic filtering. The success of this step was validated quantifying the number of beads that had DNA ligated. The result was

positive if were counted around 500,000 beads (using the beads counter provided by Roche).

The 454 Roche technology uses enzyme beads (containing sulfurylase and luciferase) (part of the 454 Junior Titanium Sequencing kit) placed onto a PicoTiterPlate device. This device is necessary to ensure that the DNA remains positioned in the wells during the sequencing reaction. On the wells plate occurs a de novo synthesis of the complementary DNA template ligated to the beads. The sequencing reaction occurs when a nucleotide successfully is ligated to the daughter chain as each time PPi is liberated, which reacts with PPiase and produces a luminescent signal. If there is not for a certain nucleotide a successful reaction during amplification, they are washed out and no signal is produced. The luminescent signals are then interpreted as nucleotide sequences on each well.

2.7.5. Data Analysis

2.7.5.1. Sequences Analysis in QIIME

From the raw sequences obtained, data analysis was performed using the microbiological communities' program QIIME http://qiime.sourceforge.net/ (Caporaso *et al.*, 2010b). In order to use this multiplatform pipeline, it was initially necessary to create the mapping file containing the name of each sample, the barcode sequence

used for each sample, the linker/primer sequence used to amplify the sample's desired gene. In addition all the metadata related to the sample was included e.g. age, gender, type of sample, Case/Control, wheezing status, treatment, etc. For the data analysis the sequences text file (fasta file *.fna), the sequencing quality file (*.qual), and the Standard Flowgram Format binary file (*.sff) were used.

As the mapping file is generated manually, it was important to check if it had been formatted correctly and this check was performed using the script:

check_id_map.py -m Fasting_Map.txt -o mapping_output

-m = mapping file

-o = output directory name

Read errors were removed if there were <200 and >800 nucleotide (nt), mismatches in the barcodes or primers, ambiguous nucleotides and if the quality score was <25 recommended by Brockman *et al.* (2008) for the Titanium platform. The Ampliconoise algorithm (Reeder and Knight, 2010) was used to avoid overestimates in diversity as well as chimera removal using the following command with appropriate modification of the files name pending on the stage of the project being analysed:

ampliconnoise.py -i stage1b_1to36.sff.txt -m Map_stage1b_1to36.txt -o anoised_stage1b_1to36.fna --platform titanium -n 4

-i = input file

-m = mapping file

-o = output file

--platform = platform used for pyrosequencing

-n = cores of CPU used on the analysis

As many sequencing runs were analysed together, it was necessary to merge the data produced from different runs. The following command was used to achieve this:

cat seqs1.fna seqs2.fna >> combined_seqs.fna

After denoising and ensuring, by using the sequence barcode, that each sequence was assigned to a sample, OTUs were defined using the UCLUST algorithm (Edgar, 2010b) through the pick_otus.py command at 97% sequence identity:

pick_otus.py -i seqs.fna -r refseqs.fasta -m uclust_ref -C

-i = input file

-r = Path to reference sequences

-m = otus picking method

-C = suppress creation of new clusters using seqs that don't match reference

After OTU picking, a representative sequence was selected for the posterior analysis using the following command:

pick_rep_set.py -i seqs_otus.txt -f seqs.fna -o rep_set1.fna

Sequences were aligned to the Greengenes Core reference alignment (DeSantis *et al.*, 2006) using a Python based tool, by means of the Nearest Alignment Space Termination (NAST) algorithm called PyNAST (Caporaso *et al.*, 2010a), and the following command in QIIME:

align_seqs.py –i unaligned.fna -t core_set_aligned.fasta.imputed -o pynast_aligned_defaults/

-i = input file from alignment

-t = Greengenes core reference alignment

-o = output file

Phylogenetic classification was assigned using the Greengenes reference database version 10.2 (McDonald *et al.*, 2012) and the command line:

assign_taxonomy.py -i repr_set_seqs.fasta -r ref_seq_set.fna -t id_to_taxonomy.txt

The alignment produced in the previous step was filtered prior to phylogenetic tree construction. It was used the default lanemask template file on QIIME lanemask_in_1s_and_0s.txt:

filter_alignment.py -i seqs_rep_set_aligned.fasta -m lanemask_in_1s_and_0s -o filtered_alignment/

-i = input file from alignment

-m = lanemask template file located in QIIME /Users/caporaso/data/greengenes_core_sets/lanemask_in_1s_and_0s.txt

In order to build a maximum-likelihood phylogenetic tree using Fasttree version 2.1.3 (Price *et al.*, 2010b) the command line:

make_phylogeny.py -i aligned.fasta -o rep_phylo.tre –r midpoint

-i = input file filtered alignment

-o = output file name

-r = rooted method midpoint was used.

Using the taxonomy assignation obtained and the abundance of sequences per OTU, an OTU summary table could be constructed:

make_otu_table.py -i otu_map.txt -t tax_assignments.txt -o otu_table.biom

-i = input file from the ampliconnoise command

-t = taxonomy assignations

-o = output file name in format .biom

To determine the number of reads per sample obtained in the OTU table, the command used was:

biom summarize-table -i otus/otu_table.biom -o otus/otu_table_summary.txt

-i = the input OTU table file

-o = output file summary

An interactive heatmap showing the number of reads per OTU and per sample was initially constructed using:

make_otu_heatmap_html.py -i otu_table.biom -o heatmap/

-i = input OTU table

-o = output directory path

The heatmap obtained was subsequently modified using Microsoft Excel 2013® to change the colours and to group the samples by age and into cases and controls.

Sequences were rarefied (to remove the heterogeneity in the number of sequences per sample) prior to calculation of alpha diversity. The latter was calculated using a multiple rarefaction pipeline:

multiple_rarefactions.py -i otu_table.biom -m 10 -x 140 -s 10 -n 2 -o rarefied_otu_tables/

-i = input out table.

-m = Minimum number of seqs/sample for rarefaction.

-x = Maximum number of seqs/sample for rarefaction.

-s = Size of each steps between the min/max of seqs/sample.

-n = The number of iterations at each step.

-o = output directory.

Alpha diversity calculations provide a comparison intra-group of the number of species found and their distribution on each group. The diversity indexes like Shannon and Inverse Simpson are an average calculation of evenness and richness. Evenness shows how uniform is the distribution of the bacterial OTUs within each group, and richness calculate the number of species per group (Figure 2.2).

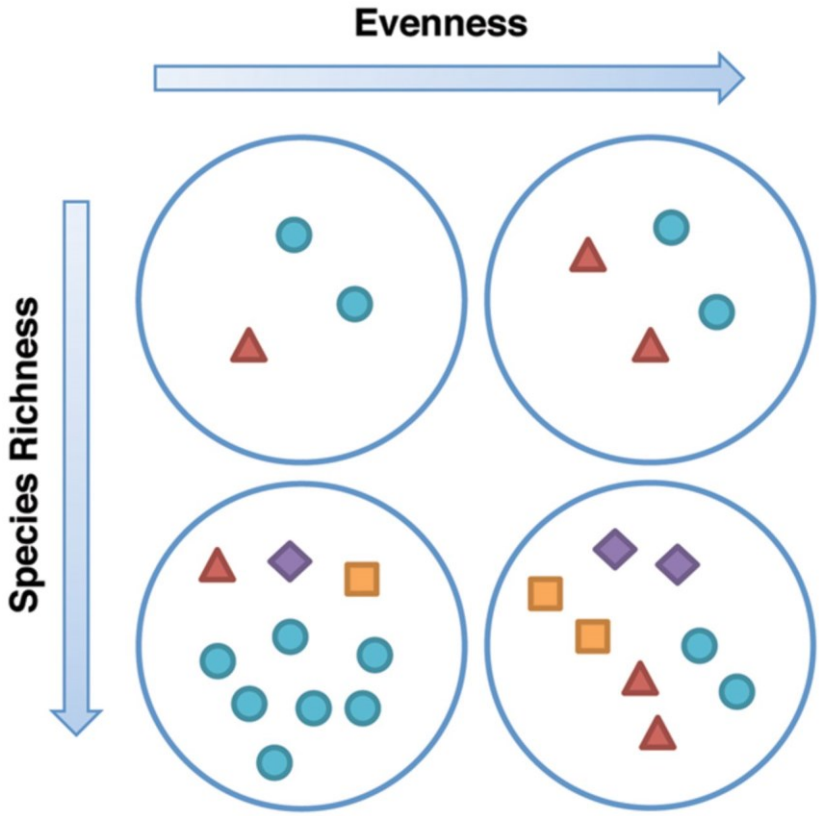


Figure 2.2. Alpha Diversity Calculations on Mixed Microbial Communities showing Evenness and Richness. Each shape represents an individual and the colour and nature of the shape represents a different type of organism. Increased numbers of different types of organism is described as increased species richness. When no one organism is dominant, the community is described as even. Taken under permission (see appendix) from Cox *et al.* (2013b)

The multiple rarefaction OTU tables, alpha diversity coefficients, Shannon; OTUs observed per sample; evenness and richness statistics (Chao1) were computed:

alpha_diversity.py -i otu_tables/ -t rep_set.tre -m shannon,chao1,eveness,observed_species -o adiv_pd.txt

-i = multiple rarefactions directory path.

-t = phylogenetic tree.

- m = alpha diversity metrics.

-o = output directory.

Results from each statistic for each rarefied OTU table were collated in a single file using the following command:

collate_alpha.py -i alpha_div/ -o collated_alpha/

-i = alpha diversity files.

-o = output directory.

Using the collated alpha diversity files, plots from each statistic were constructed by:

make_rarefaction_plots.py -i alpha_div_collated/ -m Fasting_Map.txt -d 300 -g pdf

-i = input directory.

-m = mapping file.

-d = image resolution.

-g = image format output

Beta diversity statistics are performed to provide in a numerical matrix the diversity differences compared between two or more groups. They were executed using UNIFRAC metrics (Lozupone and Knight, 2005a, Chang *et al.*, 2011) on single rarefaction OTU tables. The value of reads per sample on the single rarefaction tables was selected using the minimum number of sequences obtained in a sample with good enough diversity (seen both on the heatmap and the OTU table).

single_rarefaction.py -i otu_table.biom -o otu_table_even100.biom -d 100

Two methods were implemented for beta diversity statistics. The weighted UNIFRAC (that takes into account the abundance and the phylogenetic differences of each OTU between samples), and the un-weighted UNIFRAC (which takes account of the phylogenetic differences of each OTU between samples).

beta_diversity.py -i otu_table.biom -m weighted_unifrac,unweighted_unifrac -o beta_div/ -t rep_set.tre

-i = single rarefied OTU table.

-m = metrics used.

-o = output directory path.

-t = phylogenetic tree.

Using the beta diversity distance matrices obtained, a principal coordinates analysis (PCoA) was used to compare groups of samples based on phylogenetic or count-based distance metrics:

principal_coordinates.py -i beta_div.txt -o beta_div_coords.txt

-i = input beta diversity metrics

-o = output file

Results from the PCoA were visualized using EMPeror, for the analysis and visualization of results on 3D plots (Vazquez-Baeza *et al.*, 2013). The following command was used:

make_emperor.py -i unweighted_unifrac_pc.txt -m mapping_file.txt -b "Age,Treatment" -o emperor/

-i = beta diversity PCoA file.

-m = mapping file.

-b = categories used for comparisons.

-o = output directory path.

In addition to the 3D beta diversity plots produced previously, bi-plots that include taxa summary were constructed using the command lines:

make_emperor.py -i wf_bdiv_even146/unweighted_unifrac_pc.txt -m Fasting_Map.txt -t wf_taxa_summary/otu_table_L3.txt --n_taxa_to_keep 5 -o 3d_biplot

-i = beta diversity PCoA file.

-m = mapping file.

-t = taxa summary produced from the OTU tables.

--n_taxa_to_keep = number of most abundant phylum-level taxa to be included.

-b = categories used for comparisons.

-o = output directory path.

In addition to the UNIFRAC metrics used for the beta diversity analysis, Jackknifed Beta Diversity and Hierarchical Clustering were performed. This method uses the phylogenetic distances obtained in a tree by dividing the sum of "unshared" branch lengths over the sum of all tree branch lengths. As consequence phylogenetic distances between two communities are represented in a distance matrix (Lozupone *et al.*, 2006). From the rarefied distance matrices an Un-weighted Pair Group Method with Arithmetic mean (UPGMA or average linkage) tree was performed:

upgma_cluster.py -i beta_div.txt -o beta_div_cluster.tre

-i = beta diversity calculations input file.

-o = output tree file.

2.7.5.2. Sequences Analysis in R

The statistical analysis to compare between groups was performed in R version 3.02 (Team, 2014) using Rstudio version 0.98.497 and the following packages: phyloseq (McMurdie and Holmes, 2013), picante (R tools for integrating

phylogenies and ecologie), ggplot2 (graphical tool), plyr (tools for splitting, applying and combining data), Biostrings (string objects representing biological sequences and matching algorithms) (Pages *et al.*), vegan (community ecology package) (Dixon, 2003), ape (analyses of phylogenetics and evolution) (Paradis *et al.*, 2004), ade4 (analysis of Euclidean data) (Dray *et al.*, 2007).

OTU tables were imported together with the rooted phylogenetic tree produced on QIIME using the following command:

ecu.biom = import_biom("otus/otu_table.biom", treefilename = "rooted.tre")

The metadata from samples was uploaded into R using:

ecu.env = import_qiime(mapfilename = "../mapping_file.txt")

The reference sequences from each OTU were imported using:

refseqs = readDNAStringSet("otus/rep_set/anoise_out _rep_set.fasta")

The OTU table, phylogeny, metadata and reference sequences were merged in a single metafile:

ecu.biom = merge_phyloseq (ecu.biom, ecu.env, refseqs)

A function called “multiplot” was created to allow plotting multiple graphics side by side:

```
multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {

require(grid)

# Make a list from the ... arguments and plotlist

plots <- c(list(...), plotlist)

numPlots = length(plots)

# If layout is NULL, then use 'cols' to determine layout

if (is.null(layout)) {

# Make the panel

# ncol: Number of columns of plots

# nrow: Number of rows needed, calculated from # of cols

layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),

ncol = cols, nrow = ceiling(numPlots/cols))

}

if (numPlots==1) {

print(plots[[1]])
```

```
} else {

# Set up the page

grid.newpage()

pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))

# Make each plot, in the correct location

for (i in 1:numPlots) {

# Get the i,j matrix positions of the regions that contain this subplot

matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))

print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,

layout.pos.col = matchidx$col))

}

}
```

As the number of sequences obtained varied between samples (575 to 4395) and to remove this sample heterogeneity, sequences were rarefied to 575 sequences per sample. To do this:

plot(sort(sample_sums(ecu.biom)), pch = 16, cex = .5, ylab = "Number of reads", xlab = "Ranked abundance", main = "Ranked abundance of reads per sample", sub = "Selection of random re-sampling (rarefaction) level to optimise both reads and samples")

abline(h = 575, col = "red")

ecu.biom = rarefy_even_depth(bx.biom, sample.size = 575, rngseed = TRUE, replace = FALSE, trimOTUs = TRUE)

As the default taxa levels on the OTU table are not very informative for studies of this nature, the text was replaced with the phylogeny ranks:

colnames(tax_table(ecu.biom)) = c("Phylum", "Class", "Order", "Family", "Genus", "Species")

Per each sample alpha diversity was additionally calculated using the command:

sample_data(ecu.biom)$Shannon = diversity(t(otu_table(ecu.biom)))

sample_data(ecu.biom)$InvSimp = diversity(t(otu_table(ecu.biom)), "inv")

sample_data(ecu.biom)$Rich = specnumber(t(otu_table(ecu.biom)))

sample_data(ecu.biom)$Even = specnumber(t(otu_table(ecu.biom)))

sample_data(ecu.biom)$Shannon/log(sample_data(ecu.biom)$Rich)

Alpha diversity obtained per sample was plotted using the software GraphPrism version 5.0 (http://www.graphpad.com). The alpha diversity values influenced by phylogeny were calculated using the net relatedness index (NRI) and nearest taxon index (NTI) (Swenson *et al.*, 2006):

MPD = ses.mpd(t(otu_table(ecu.biom), cophenetic(phy_tree(ecu.biom)), null.model="taxa.labels")

MNTD = ses.mntd(t(otu_table(ecu.biom), cophenetic(phy_tree(ecu.biom)), null.model="taxa.labels")

To determine normality in the alpha diversity values from samples grouped on quantiles, the Shapiro-Wilkinson normality test (SHAPIRO and WILK, 1965) was used:

shapiro.test(sample_data(ecu.biom)$Rich)

shapiro.test(sample_data(ecu.biom)$Even)

shapiro.test(sample_data(ecu.biom)$InvSimp)

shapiro.test(sample_data(ecu.biom)$Shannon)

Results from this were then plotted using:

```(r fig.width=7, fig.height=6)

par(mfrow = c(2,2))

qqnorm(sample_data(bx.biom)$Rich, main = "Richness")

qqnorm(sample_data(bx.biom)$Even, main = "Evenness")

qqnorm(sample_data(bx.biom)$InvSimp, main = "InvSimp")

qqnorm(sample_data(bx.biom)$Shannon, main = "Shannon")
```

Results that showed a normal distribution were analysed with t test statistics for comparisons between two groups and ANOVA (analysis of variance) when considering > than two groups.

```
t.test(sample_data(ecu.biom)$Rich ~ sample_data(ecu.biom)$Gender)

fit = aov(sample_data(ecu.biom)$Rich ~ sample_data(ecu.biom)$Diagnosis)

summary(fit)
```

Whilst non-parametric results were analysed using the Wilcoxon test for two group comparison and the Kruskal test for > than two groups:

wilcox.test(sample_data(ecu.biom)$Shannon ~ sample_data(ecu.biom)$Gender)

kruskal.test(sample_data(ecu.biom)$Even ~ sample_data(ecu.biom)$Diagnosis)

Correlation for ordinal and continuous variables was also tested for by:

cor.test(sample_data(ecu.biom)$Shannon, sample_data(ecu.biom)$Age, method = "kendall")

Beta diversity was calculated between groups at OTU level using Adonis R package on vegan (Dixon, 2003), and subsequent statistical analysis performed using: Canberra, Bray-Curtis, UNIFRAC and Weighted UNIFRAC.

ecu.dist.cross.canberra = distance(ecu.biom, method = "canberra")

ecu.dist.cross.bray = distance(ecu.biom, method = "bray")

ecu.dist.cross.unifrac = distance(ecu.biom, method = "unifrac")

ecu.dist.cross.wunifrac = UniFrac(ecu.biom, weighted = TRUE)

ecu.dist.canberra = distance(ecu.biom, method = "canberra")

ecu.dist.bray = distance(ecu.biom, method = "bray")

ecu.dist.unifrac = distance(ecu.biom, method = "unifrac")

ecu.dist.wunifrac = UniFrac(ecu.biom, weighted = TRUE)

To determine how the cases/controls status explains a different proportion of the variance each time there were used Nonmetric multidimensional scaling (NMDS) stress calculations:

ecu.biom.ord.canberra = ordinate(ecu.biom, "NMDS", "canberra")

ecu.biom.ord.bray = ordinate(ecu.biom, "NMDS", "bray")

ecu.biom.ord.unifrac = ordinate(ecu.biom, "NMDS", "unifrac")

ecu.biom.ord.wunifrac = ordinate(ecu.biom, "NMDS", "unifrac", weighted = TRUE)

And results were plotted:

r1 = plot_ordination(ecu.biom, ecu.biom.ord.bray, type = "taxa", color = "Rank2", title = ecu.biom.ord.bray$Diagnosis) + theme(legend.position = "none")

r2 = plot_ordination(ecu.biom, ecu.biom.ord.canberra, type = "taxa", color = "Rank2", title = ecu.biom.ord.canberra$Diagnosis) + theme(legend.position = "none")

r3 = plot_ordination(ecu.biom, ecu.ord.canberra, type = "taxa", color = "Rank2", title = ecu.ord.canberra$Diagnosis) + facet_wrap(~Rank2) + theme(legend.position = "none")

r4 = plot_ordination(ecu.biom, ecu.ord.bray, type = "taxa", color = "Rank2", title = ecu.ord.bray$Diagnosis) + facet_wrap(~Rank2) + theme(legend.position = "none")

multiplot(r1, r2, r3, r4, layout = matrix(c(1, 2, 3, 4), nrow = 2, byrow = TRUE))

To show a clustering per group the following script was implemented:

plot_ordination(bx.biom, bx.ord.unifrac, type = "samples", color = "Diagnosis")

Adonis was used to partition variance between variables and to indicate whether variance was significant:

ecu.even.group = subset_samples(ecu.biom, Diagnosis != "<NA>")

ecu.dist.bray.group = distance(ecu.even.group, method = "bray")

adonis(ecu.dist.bray.group ~ Diagnosis, as(sample_data(ecu.even.group),

"data.frame"))

ecu.dist.unifrac.group = distance(ecu.even.group, method = "unifrac") adonis(ecu.dist.unifrac.group ~ Diagnosis, as(sample_data(ecu.even.group), "data.frame"))

ecu.dist.canberra.group = distance(ecu.even.group, method = "canberra") adonis(ecu.dist.canberra.group ~ Diagnosis, as(sample_data(ecu.even.group), "data.frame"))

ecu.dist.wunifrac.group = distance(ecu.even.group, method = "unifrac", weighted = TRUE) adonis(ecu.dist.wunifrac.group ~ Diagnosis, as(sample_data(ecu.even.group), "data.frame"))

2.7.5.3. Microbial Community Comparisons

Microbial community comparisons were performed using parametric statistics in METASTATS, with *P* values corrected for multiple hypothesis testing by using the false discovery rate (FDR) (White *et al.*, 2009). As previously for the Sanger Sequencing Analysis (Section 2.6.7), multiple groups comparisons were made using the Bonferroni correction test which corrected the *P*-values for multiple comparisons in UNIFRAC (Lozupone and Knight, 2005a).

2.7.5.4. Phylogenetic Analysis of Bacterial Species

Neighbour joining with nearest neighbour interchange phylogenetic trees were created using the representative sequences of each OTU and FastTree version 2.1.3 (Price *et al.*, 2010a). The heatmap to represent the abundance of sequences was constructed in iTOL (Letunic and Bork, 2011).

Representative sequences from significantly different OTUs of further interest were investigated using more intensive phylogenetic approaches in order to maximize the quality of the identification. These test sequences were aligned using the online SINA aligner (http://www.arb-silva.de/aligner/version 1.29 (Pruesse *et al.*, 2012)) and then imported into the ARB phylogenetic software (version 5.1, http://www.arb-home.de/ (Ludwig *et al.*, 2004b)) running on Biolinux 6.0 (http://nebc.nerc.ac.uk/tools/bio-linux/bio-linux-6.0, (Field *et al.*, 2006b)). The aligned SILVA reference database SSU_REF108 of 618,442 high quality 16S rRNA gene sequences was downloaded and merged with the aligned test sequences. All *Haemophilus* spp. (or *Streptococcus* spp.) sequences within the database were selected and the SINA alignment individually checked for each test sequence in the ARB alignment editor. The length of the alignments used depended on the length of the available reference reads for each OTU. Thus, for the *Haemophilus* spp. alignment, the 522 bp region corresponding to the region between positions 384 and

908 of the *Escherichia coli* reference were selected, and for *Streptococcus* spp. 470 bp between positions 470 and 908. Columns of the alignment containing uninformative positions (gaps) were masked from the phylogenetic analysis. Three trees were constructed for each of the genera, an ARB neighbour joining (NJ) tree with 1000 bootstraps, a Maximum Parsimony (MP) tree with 500 bootstraps and a RAxML Maximum Likelihood (ML) tree (version 7.0.3, (Stamatakis *et al.*, 2005a)) with GTR substitution model in rapid hill-climbing mode. Trees were rooted with sequences from near neighbours outside the genus of interest. Tree topology was compared between the three methods and bootstrap values for the NJ and MP trees were used to determine stability of the phylogeny.

2.8. DNA extraction from Bacterial DNA after culture

Before starting and 24 hours prior extraction, a single bacterial strain was inoculated in 500 µl of BHI broth and incubated at 37 degrees Celsius. After this time bacterial growth was verified using spectrophotometry at light wave length of 590nm verifying the absorbance between 0.2 and 0.4 (to determine turbidity and bacterial growth presence) (Pommerville, 2011).

Two hundred µl from the single bacterial strain culture (BHI broth) were taken and added to the lysing matrix E tube (LME). Then 778µl of sodium phosphate buffer were poured into the LME tube and 122µl of MT buffer (lysis solution). The LME

tubes were homogenized twice in the FastPrep Instrument for 40 seconds at 6.0, and centrifuged at 14,000 xg for 10 minutes to pellet debris.

Two hundred and 50µl volumes of Sodium Phosphate buffer (PPS) were aliquotted into a number of clean 2.0ml tubes. From the LME tubes previously homogenized and centrifuged, the supernatant was transferred to the 2.0ml tubes and mixed by shaking tube by hand 10 times, and again centrifuged at 14,000 xg for 5 minutes to precipitate the pellet.

The binding matrix suspension was re-suspended and 1.0ml was added to clean 15ml tubes together with the supernatant from the previous stage. These tubes were placed on a rotator for 2 minutes to allow binding of DNA, and later transferred to a tube rack for 3 minutes to allow settling of silica matrix.

One ml of supernatant was discarded carefully to avoid touching the settled binding matrix, and the rest of volume was transferred into a 2ml tube. The tubes were centrifuged for 2 minutes at 14,000 xg, and was removed the supernatant. The binding matrix was re-suspended in 500µl SEWS-M & transferred to the column. The columns were centrifuged at 14,000 xg for 1 minute, and the catch tube replaced.

Without any addition of liquid, the columns were centrifuged a second time at 14,000 xg for 2 minutes to ‘dry’ the matrix of residual wash solution, and again the catch tube was discarded and replaced with a new, clean catch tube.

The columns were air dried for 5 minutes at room temperature, and gently re-suspended in 100µl of Tris-low EDTA (10mM Tris-Cl + 0.1mM EDTA). Then the tubes were incubated for 5 minutes at 55^{o}C in a heating block, and after that centrifuged at 14,000 xg for 1 minute to bring eluted DNA into the clean catch tube. Lastly the DNA suspended in the Tris-low EDTA was transferred into a screw-cap tube and stored at -80^{o}C.

Chapter 3: Current Wheezing Microbiota Study - ECUAVIDA Stage 1a

3.1. Introduction

Stage 1a of the ECUAVIDA microbiota study, a case control retrospective study (Chapter 2: Section 2.3), was established in order to investigate the changes in the upper airways microbiota that occur during a current non-infectious wheezing episode. Consequently throat swab samples were collected from infants (cases), ages ranging from 7 to 12 months that at the time of sampling had a current wheezing episode and had had a history of multiple-triggered non-infectious wheezing episodes. Additionally, healthy children; age, sex and place of living matched to the cases; were also recruited and sampled (Chapter 2: Section 2.3).

The microbiota analysis at this stage started as an initial pilot study using 25 samples (12 cases and 13 controls) to determine if differences did or did not exist between cases and controls. At the time of the pilot the molecular techniques available for the investigations were Denaturing Gradient Gel Electrophoresis (DGGE) and Sanger sequencing. Subsequent to the pilot and when next generation sequencing of amplicons of the desired length became a reality, a further more extensive study was carried out with 48 infants using both the new technology as well as a more comprehensive bioinformatics analysis (Chapter 3: Section 3.2).

3.2. Pilot Study

3.2.1. Denaturing Gradient Gel Electrophoresis (DGGE)

PCR-DGGE is a useful technique to evaluate bacterial diversity and since the end of the last century was widely used to characterize bacterial species on mixed microbial communities (Muyzer and Smalla, 1998). Nowadays, sequencing technology has replaced PCR-DGGE as the gold-standard identification method of bacterial species; nonetheless DGGE analysis remains an important initial first step technique to determine diversity of microbial communities (Tzeneva *et al.*, 2008, Muhling *et al.*, 2008).

3.2.1.1. Materials and Methods

DGGE is a technique based on differences in the intrinsic denaturing behaviour of double stranded DNA fragments that allows separation of fragments that are the same size but that differ in sequence. It allows using samples of mixed microbial communities (and therefore different sequences), and generates distinct 16S band patterns (Tzeneva *et al.*, 2008). This technique also provides a relatively accurate taxonomic identification awareness of the principal bacterial groups. DGGE can be a good first step to analyse the microbiota and can provide insight into how

diverse samples are (Green *et al.*, 2004). Consequently it can also inform the depth of sequencing that may be subsequently required.

The DGGE that was conducted (Chapter 2: Section 2.6) included the use of two different markers; the Asthma Marker and the Normal Marker. The constituents of these markers (their makeup determined by previous studies carried out in the Molecular Genetics and Genomics group (Hilty *et al.*, 2010)) allowed an approximation of the taxonomy of every sample run on the DGGE to be to determined. Additionally it was feasible, by counting the number of bands visible on the DGGE to estimate the diversity of each sample. However a band can contain multiple amplicons of similar sequences that run at the same level on the DGGE thus the real potential to underestimate diversity by band counting exists. Due to the somewhat subjective nature of these measures, it was necessary to err on caution in making substantive conclusions but they nonetheless served as an initial tool for the Stage 1a pilot.

A total of 25 samples were used for the pilot study (12 cases and 13 controls). The samples were obtained from infants that are part of the ECUAVIDA cohort in the tropics of Ecuador (Cooper *et al.*, 2011) (Chapter 2: Section 2.2). DNA was extracted from the throat swabs (Chapter 2: Section 2.4) and used as template in 25µl PCR reactions amplifying the V3-V5 region of the 16s rRNA gene (Chapter 2: Sections

2.6.1-2.6.3) after which 8µl aliquots of the PCR products were run on DGGE (Chapter 2: Sections 2.6.4).

3.2.1.2. DGGE Results

The DGGE for the 25 samples (12 cases and 13 controls) is shown in Figure 3.1). Looking at diversity (by counting the number of bands present for each sample) there was encountered a broad variability across the samples. Within the cases, individual 17 was the most diverse (with 15 identifiable bands), whilst the least variable was individual 23 (having just a single band). In the controls, individual 15 was the most diverse (with 13 bands), and individual 24 has the lowest diversity (with only 2 bands). From the PCR-DGGE neither group (cases or controls) were more diverse than the other (average 7 bands in cases and 8.33 in controls) (Figure 3.1).

According to the ´Asthma Marker´ (AM), band 7 (*Prevotella* sp. and *Streptococcus* sp.) was present in all the samples, and was the most intense band seen in 24 out of 25 samples the single exception being individual 15.

Analysis of band patterns showed that samples shared bands in several positions for example:

- Band number 1 of the ´Normal Marker´ (NM) (corresponding to Bacteroides or *Haemophylus sp.*) was present in individuals 16, 17, 18, 20 and 21.
- Band number 9 in the AM (corresponding to *Streptococcus sp.* or *Neisseria sp.*) was observed in individuals 9, 13, 6 and 20.
- Bands 5 and 6 in the AM (corresponding to *Haemophilus* sp. or *Moraxella* sp.) were present in individuals 4, 17, 3, 15, 7, 10 and 21.
- Band number 12 in the AM (corresponding to *Bordetella hinzii*) was only present in individual 7.
- Band number 3 of the NM (corresponding to *Porphyromonas* sp.) was observed in individuals 6, 8, 9 and 17.

The exact position for the remaining bands observed in individuals was difficult to establish in the equivalent asthma and normal markers. In conclusion the band patterns are very mixed across the cases and controls with no clear segregation by this method. As most of bands cannot be attributed to a bacterial taxon because they do not correspond to any band in the markers and that some bands have in the same position two types of bacteria, therefore the true content of the bands really needs to be determined by a more specific method. For that reason,

the next step was to clone the PCR products generated for each individual and perform Sanger sequencing.

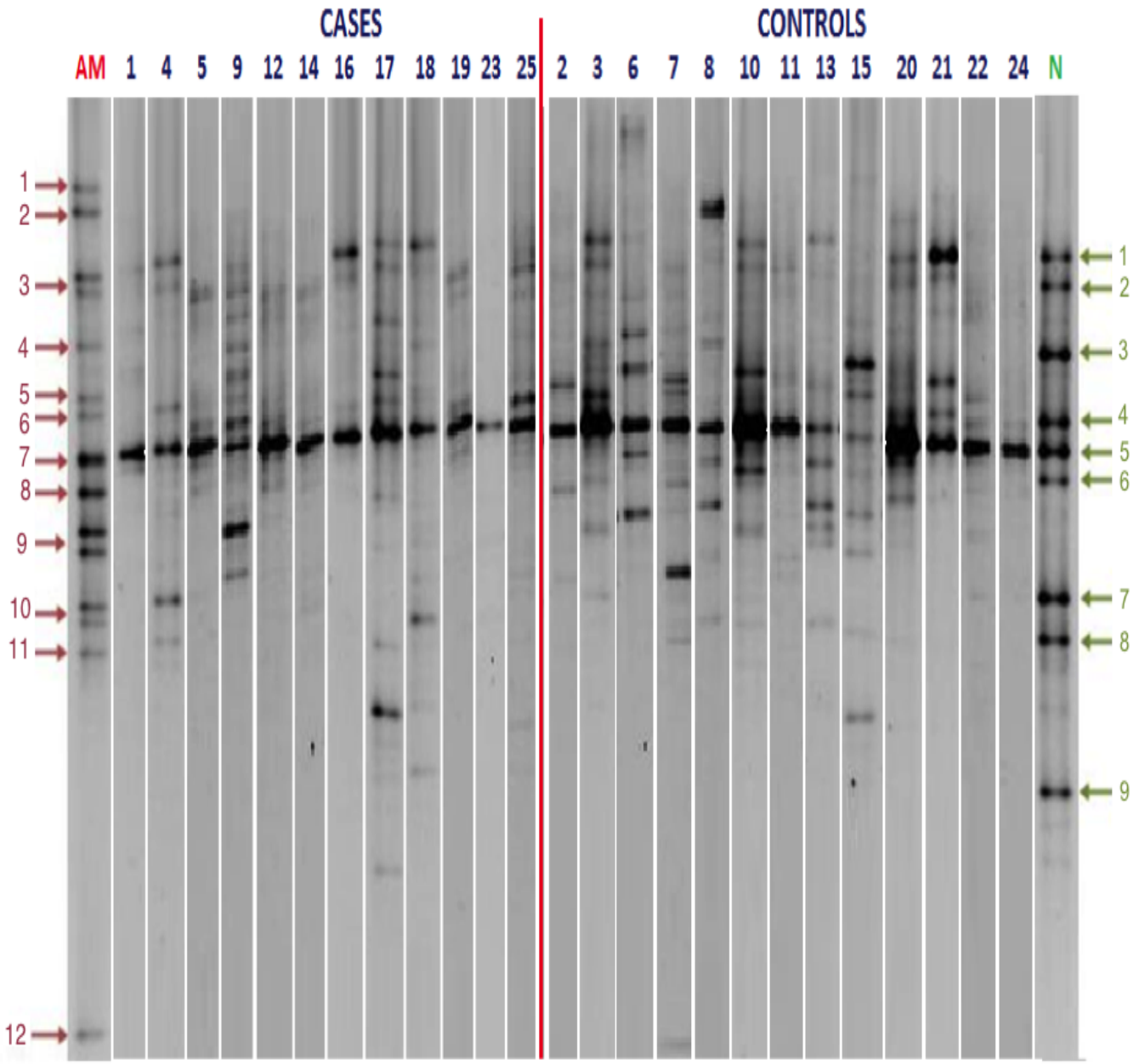


Figure 3.1 Denaturant Gradient Gel Electrophoresis (DGGE) of Cases and Controls in the Pilot Study. On the left extreme is located the "asthma marker" (AM) using bands of bacteria previously observed to be present in asthma patients, and on the right extreme is located the "normal marker" (N) using bands from bacteria previously observed to be present in the healthy lung (controls)(Hilty et al., 2010).

3.2.2. Cloning and Sequencing

The cloning and sequencing approach is a more robust technique to determine the diversity and abundance of microbial communities in a quantitative and qualitative form. The technique's efficiency is directly correlated with the quality of the PCR product and the transformation/cloning efficiency. Excessive PCR cycles can produce modifications in the sequence introducing mutations caused by errors made by the polymerase (Thornhill *et al.*, 2007). For this reason, a limited number of cycles were implemented in this present study. In theory it is believed that variable size in the PCR product obtained could influence and create bias in cloning step. In order to control this, the V3-V5 region of the 16S rRNA gene was selected for amplicon design (Ahmed *et al.*, 2007). This ensured that the size range of PCR products would only vary at most by 200bp (450-650 bp).

3.2.2.1. Materials and Methods

First round PCR products for each sample (Chapter 2: Section 2.6.2) were cloned into pGEM-T vectors. Following successful cloning 50 clones per sample were picked at random for subsequent Sanger sequencing (Chapter 2: Section 2.6.5.2).

Sequences obtained underwent alignment and statistical analysis using software specially designed for microbial metagenomic assessment (Greengenes, RDP (Cole *et al.*, 2009) and UNIFRAC (Lozupone *et al.*, 2006)). After every step, the sequence files and alignments were revised and curated manually to avoid mismatches (Chapter 2: Section 2.6.6).

3.2.2.2. Sanger Sequencing Results

After removal of chimeras, 2362 high quality sequences were obtained from the 25 samples analysed (12 cases and 13 controls). Cladistic classification of the ungrouped sequences showed a high prevalence of the phylum Firmicutes (90.90%) followed by Proteobacteria (3.26%), Bacteroidetes (4.02%), Fusobacteria (1.48%) and Actinobacteria (0.30%).

Members of the Firmicutes phylum occurred significantly more often in controls (P=0.0007) compared to cases. In contrast, members of the phylum Proteobacteria (P<0.0001) were more common in wheezing infants compared with controls. No difference was seen between cases and controls for either of the phyla Bacteriodetes and Fusobacteria (both P values >0.05) (Figure 3.2).

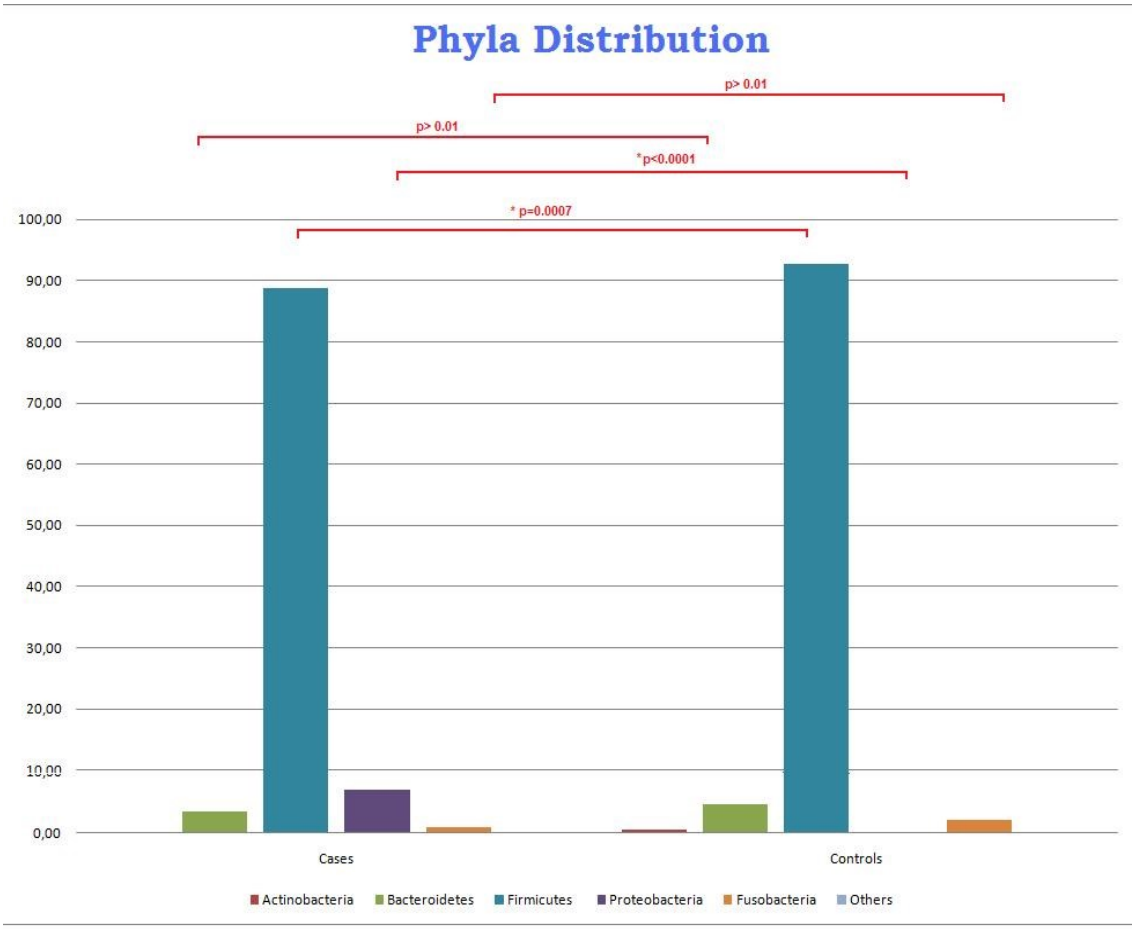
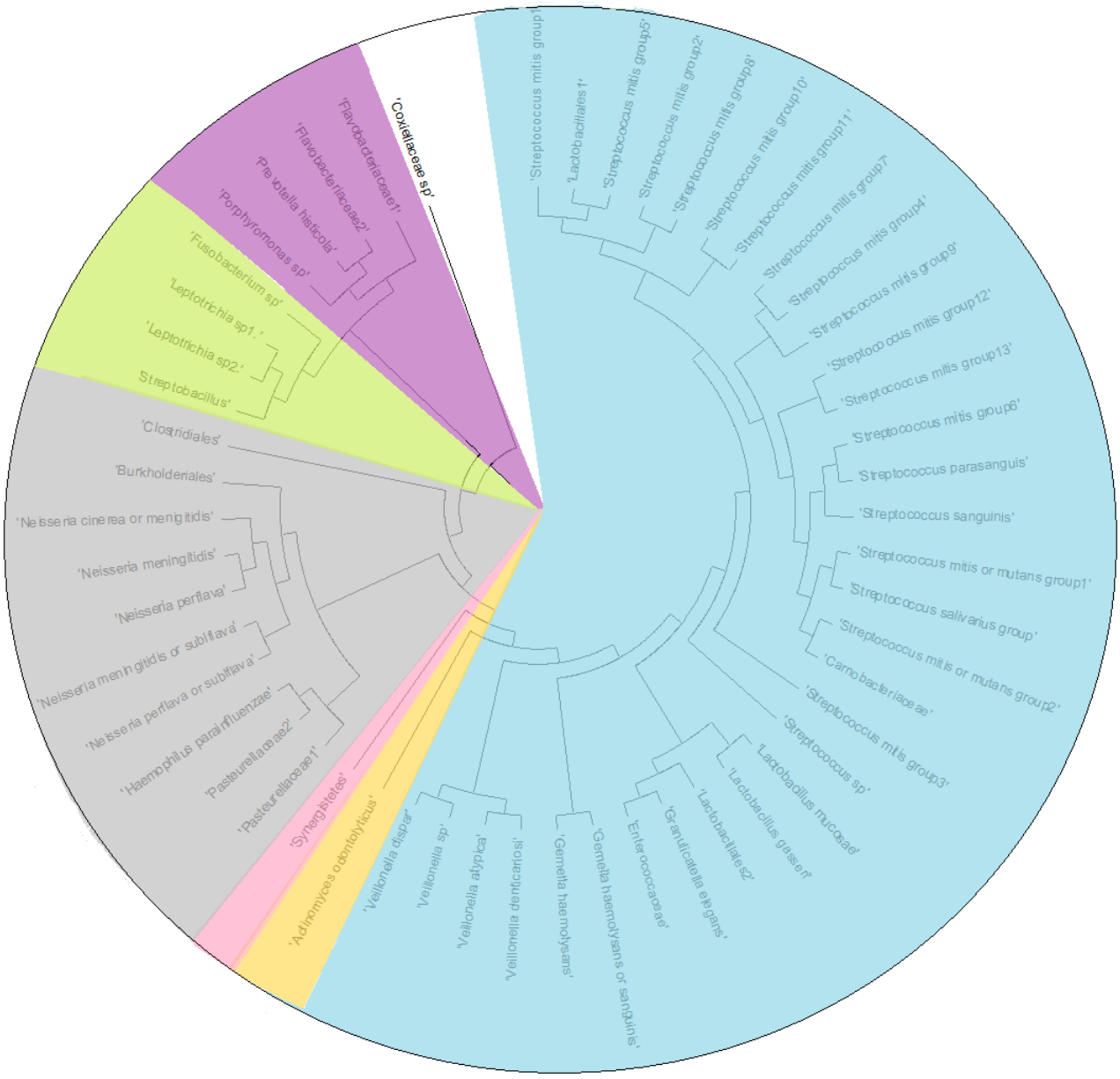


Figure 3.2. Phyla distribution of the Sequences for Cases and Controls. Fisher exact test was used for statistical analysis between both groups. * indicates significant *P* values.

After alignment, chimera removal and distance matrix batch, using DOTUR (Schloss and Handelsman, 2005) high quality sequences were found to cluster into 52 operational taxonomic units (OTUs). Diversity analysis of sequences showed that the phylum Firmicutes was the most diverse with 32 OTUs (62.3% of the found OTUs), followed by Proteobacteria with 10 OTUs (17%), then Fusobacteria and

Bacteroidetes with 4 OTUs each (7.5%) and finally Actinobacteria, Synergistetes and unclassified bacteria with 1 OTU each (1.9%) (Figure 3.3).



phylum	%	Library
"Synergistetes"	1.9	
"Fusobacteria"	7.5	
"Bacteroidetes"	7.5	
"Firmicutes"	62.3	
"Proteobacteria"	17.0	
"Actinobacteria"	1.9	
unclassified_Bacteria	1.9	

Figure 3.3. Neighbour-Joining Circular Phylogenetic Tree. It shows the diversity of sequences grouped by OTUs (each colour corresponds to a different phylum). The table shows the relative percentages of OTUs found.

Additionally a heatmap was created using a phylogenetic tree produced in MEGA6 (Hall, 2013). There were used the sequences from the 52 OTUs found after eliminating all sequences containing gaps and missing data, and an 'out group' sequence was added for rooting. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method (Yong *et al.*, 2012) and are in the units of the number of base substitutions per site (figure 3.4). Firmicutes was the most diverse and frequent phyla observed with *Streptococcus* the most frequent genus. According to Kawamura *et al.* sub-classification of *Streptococcus* sp. according to the phylogenetic relationships in their 16S rRNA gene sequence (Kawamura *et al.*, 1999), there were found thirteen OTUs identified as belonging to the *Streptococcus mitis* group with *Streptococcus mitis* Group 11 being the most prevalent OTU followed by *Streptococcus* Group 6 (Figure 3.4). Two *Streptococcus* sp. OTUs failed to be asigned to a unique group and were classified as being either *Streptococcus mitis* or *mutans*. Six OTUs could not to be asigned to a group or a species and were asigned as Lactobacillales 1 and 2, Carnobacteriacea, Enterococcacea, *Streptococcus sp.* and *Veilonella sp.* respectively. Taking account the number of sequences across all controls versus all the cases, the most marked differences

were encountered in the Firmicutes phyla. Controls had a higher prevalence of *Streptococcus salivarious* (66 sequences in controls versus 10 in cases. *P*<0.005)*, Streptococcus mitis group 7* (controls 100 versus cases 20. *P*<0.005)*, Veillonella sp.* (controls 96 versus cases 48. *P*<0.005) and *Veillonella dispar* (controls 51 versus cases 2. *P*<0.005).

Actinobacteria were found only in controls, in contrast with Synergistetes that were only present in cases although in both instances, these diferences were not statistically different (Figure 3.4).

Fusobacteria, represented by *Leptotrichia* sp*., Fusobacterium* sp*.* and *Streptobacillus* sp., were more common in cases but again this was not statistically significant (Figure 3.4). No significant differences between cases and controls for any of the OTUs observed, Flavovactereacea, *Leptotrichia* sp*., Porphyromonas* sp*.* and *Prevotella* sp. for the Bacteroides phylum.

The most important differences between cases and controls were found for the phylum Proteobacteria. *Neisseria* was the most prevalent genus although discrimination at the species level (*N. perflava* or *subflaba, N. meningitidis* or *subflava, N. cinerea* or *meningitidis)* or genera level (Pasturellacea, Costridiales and Burkholderiales) was not feasible in all cases. It was however possible to categorize

at species level several OTUs and species included *Haemophilus parainfluenzae.* For every OTU of the Proteobacteria phylum, statistical analysis showed significant differences between cases and controls (P<0.001) (Figure 3.4).

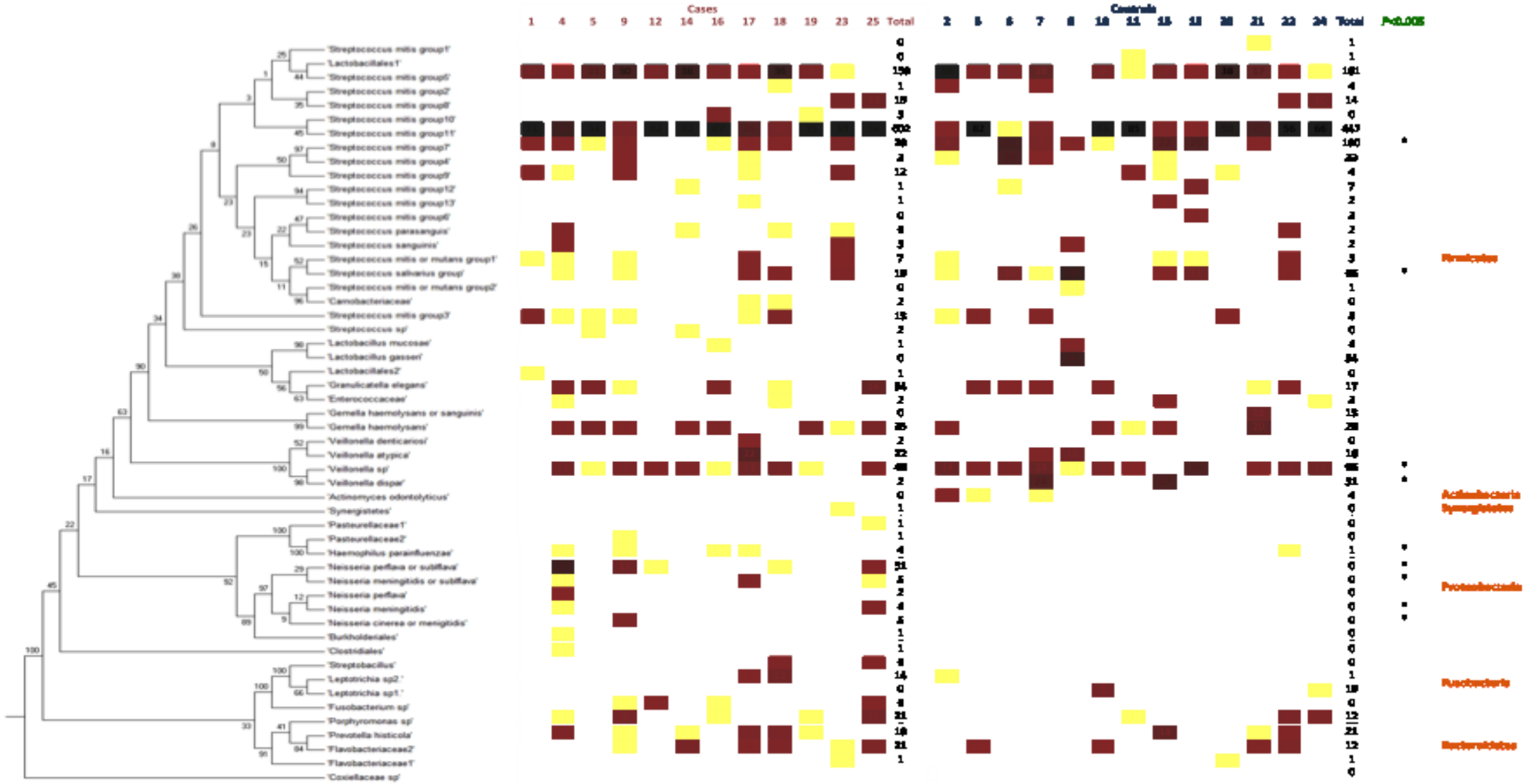


Figure 3.4. Evolutionary Relationships of Bacterial Taxa. The evolutionary history was inferred using the Neighbour-Joining method. The optimal tree with the sum of branch length = 4.78 is shown. (*Coxielacea* sp. was used as the 'out group' control). Phylogenetic analyses were conducted in MEGA6. The Heat map scale shows the abundance of sequences from each OTU found per sample (yellow = 1<5%, orange = 5–15%, red = 15–30% and black >30%). On the right there are shown with * the significant *P* values when compare cases versus controls.

Using the UniFrac metric (Lozupone *et al.*, 2006), which weights the branches based on abundance (number of sequences per branch) (Chapter 2: Section 2.6.6), a whole community comparison (beta diversity analysis) revealed the bacterial microbiome communities from wheezing children and healthy controls to be significantly different (P<0.001).

In order to determine diversity differences within each individual, alpha diversity measures were performed (Chapter 1: Section 2.6.6). There were 43 OTUs in the cases group and 35 in the healthy control group, and both groups shared 27 OTUs (51.9%) (Figure 3.5). Alpha diversity analyses in cases and controls were estimated by the Shannon index (Marcon *et al.*, 2014, Shannon, 1948):

Shanon Cases=0.42(CI 95% 0.33-0.51)

Shanon Controls=0.19 (CI 95% 0.13-0.25).

These results show that cases are more diverse than controls and contrasting the number of OTUs per group this was statistically significant (P<0.001).

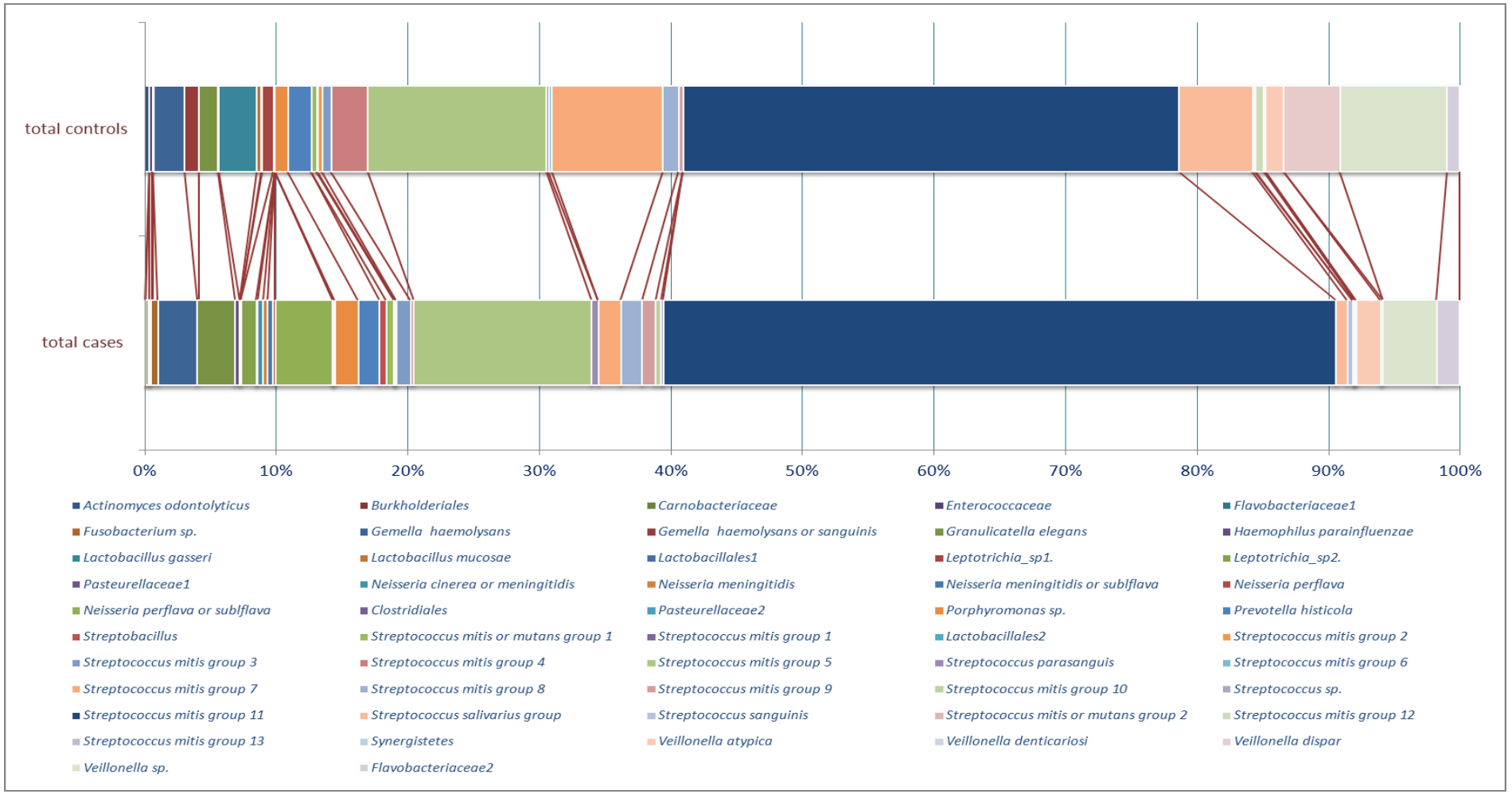


Figure 3.5. Comparative Diversity and OTUs distribution per Group. Diversity patterns per group (cases and controls) are shown. Each OTU has a different colour. A greater number of OTUs were present in the cases compared to controls (43 versus 35).

Comparison by per sample diversity showed a wide variability across all samples. A standard OTUs pattern for cases or controls was not evident (Figure 3.6).

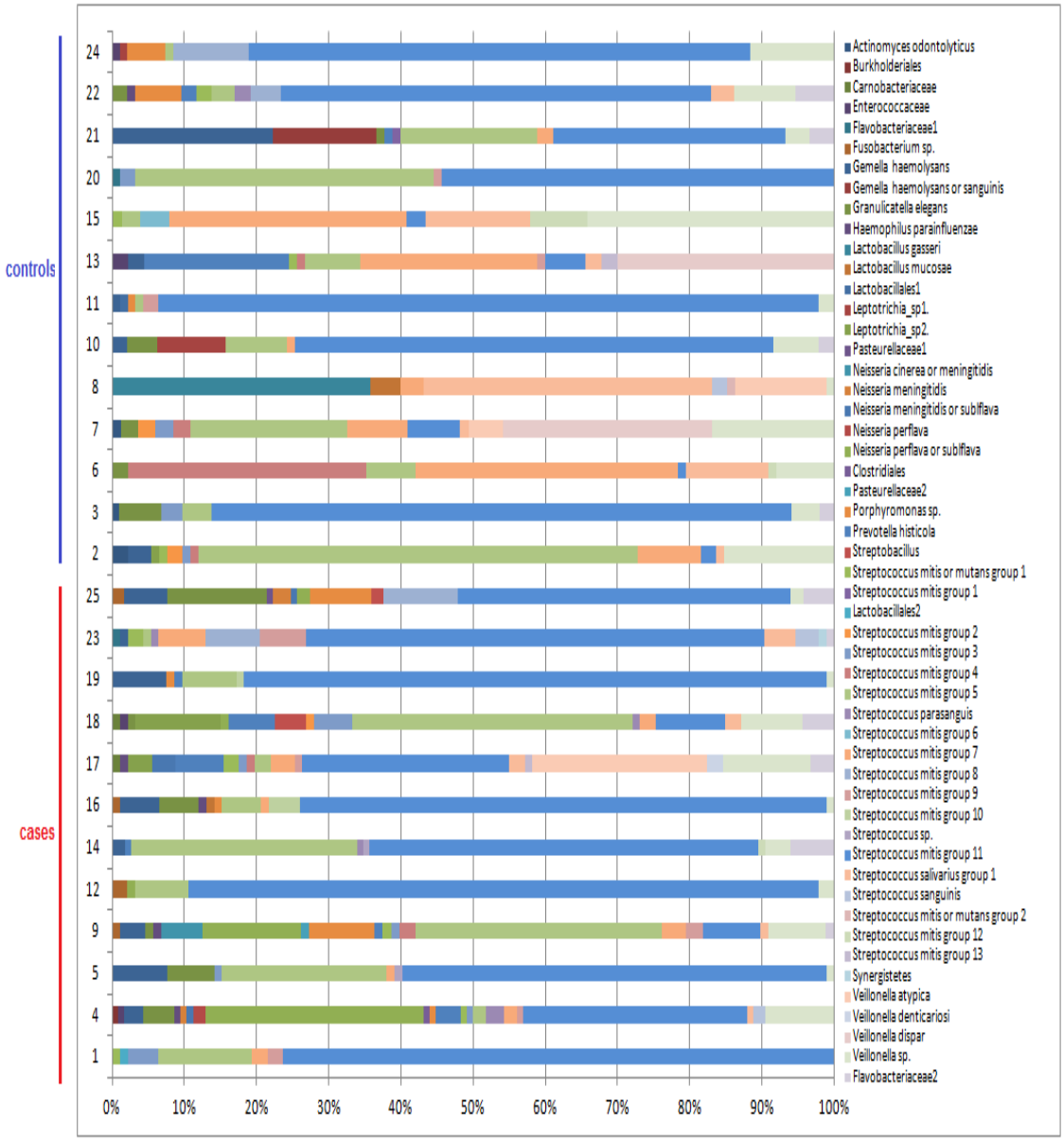


Figure 3.6. Distribution of OTUs per Sample. Diversity patterns per sample are shown. Each bar represents a single individual. Each OTU has a colour assigned. An evident OTU pattern of diversity was not evident for cases or control groups.

Alpha diversity richness statistics was performed using Chao1 index at 97% (Chapter 2: Section 2.6) (Chao et al., 2012), which suggested the presence of 147.66 expected OTUs. The current analysis was able to identify only the 35.21% (52 OTUs) of the estimated. Since diversity between cases and controls was found to be significantly different Chao1 Richness estimates were performed for each group (Chapter 2: Section 2.6.6).

For cases Chao1 estimated 312 OTUs of which 18% were identified experimentally whilst for controls the estimate was ~34 with 66.9% identified experimentally.

3.3. Stage 1a using Next Generation Sequencing

Following the initial Stage 1a pilot study using Sanger sequencing (at this point pyrosequencing technology was only able to sequence amplicons with a maximum length of 200 bp) Roche Diagnostics released the 454 technologies, which had the capability of sequencing longer amplicons (600-700 bp). The 454-pyrosequencing approach then became the gold standard technique to characterize in depth the diversity and abundance of mixed microbial communities (Lewis *et al.*, 2013, Zemanick *et al.*, 2011, Petrosino *et al.*, 2009, Parahitiyawa *et al.*, 2010). In comparison to conventional sequencing it delivers a much higher number of sequences per run, which allows one to obtain many more sequences per sample. As a consequence, the portrayal of the microbiome is more accurate, and the bacteria targeted range from the most commonly present to the minor ones within a community. It is important to remark however, that the technique's efficiency is directly correlated with the

quality of the samples, the DNA extraction protocols and the PCR products (Kim and Yu, 2014).

As this technology has the capability of sequence amplicons of a maximum of 700bp, in this part of the study new primers of region V3 to V5 were constructed for an accurate taxonomic determination. And additionally to the 25 samples used on the Stage 1a pilot study, there further 23 samples were included (cases and controls) giving a total of 48 samples.

3.3.1. Materials and Methods

After DNA extraction, the 16S rRNA amplification using barcoded primers of the V3-V5 regions was successfully performed for 48 samples part of stage 1a including all the samples previously analysed using DGGE and Sanger sequencing. Twenty-four infants with non-infectious wheezing and 24 controls were used (both groups had an average age of 10.2 months of life), and the amplicons obtained were pooled through use of the MIDs into a single 454 Junior sequencing run.

Data analysis was performed using the third party platform QIIME 1.7 (Caporaso *et al.*, 2010b), and statistical analysis was performed using Metastats V 2.0 (metagenomeSeq) (White *et al.*, 2009, Paulson *et al.*, 2013) and the R package phyloseq (McMurdie and Holmes, 2013) (Chapter 2: Section 2.7.5).

Representative sequences from significantly different OTUs of further interest were investigated using more intensive phylogenetic approaches in order to maximize the level of the identification. The representative sequences

were aligned using the online SINA aligner (http://www.arb-silva.de/aligner/version 1.29 (Pruesse *et al.*, 2012)) and then imported into the ARB phylogenetic software (version 5.1, http://www.arb-home.de/ (Ludwig *et al.*, 2004a)) running on Biolinux 6.0 (http://nebc.nerc.ac.uk/tools/bio-linux/b io-linux-6.0 (Field *et al.*, 2006a)). The aligned SILVA reference database SSU_REF108 of 618,442 high quality 16S rRNA gene sequences was downloaded and merged with the aligned test sequences. All *Haemophilus* spp. and *Streptococcus* spp. sequences within the database were selected and the SINA alignment individually checked for each test sequence in the ARB alignment editor. The length of the alignments used depended on the length of the available reference reads for each OTU. Thus, for the *Haemophilus* spp. alignment, the 522 bp region corresponding to the region between positions 384 and 908 of the *Escherichia coli* reference was selected, and for the *Streptococcus* spp. alignment 470 bp between positions 470 and 908 were selected. Columns of the alignment containing uninformative positions (gaps) were masked from the phylogenetic analysis. Three trees were constructed for each of the genera, an ARB neighbour joining (NJ) tree with 1000 bootstraps, a Maximum Parsimony (MP) tree with 500 bootstraps and a RAxML Maximum Likelihood (ML) tree (version 7.0.3 (Stamatakis *et al.*, 2005b)) with GTR substitution model in rapid hill-climbing mode. Trees were rooted with sequences from near neighbours outside the genus of interest. Tree topology was compared between the three methods and bootstrap values for the NJ and MP trees were used to determine stability of the phylogeny. Accession numbers for the reference sequences used are recorded in the tree labels (Figure 3.12)

and whilst those for the ‘out group’ are DQ358146, AJ301681, AF008582, AJ301682, AF008581, AM040491, and AM040495.

3.3.2. Pyrosequencing Results

An initial 108,042 raw sequences were obtained from the pyrosequencing run. After denoising, singleton exclusion, chimera checking and removal of OTUs present in only one sample a total of 76,627 sequences remained (37,235 in cases and 39,392 in controls). Between 969 and 6,269 sequences were obtained per sample, and consequently sequences were rarefied to remove the sample heterogeneity using the minimum number of sequences (969) obtained in a sample. After exclusion of singletons, chimeras and OTUs present in only one-sample 182 operational taxonomic units (OTUs) at 97% sequence identity level were identified.

Cladistic classification of all 76,627 sequences (not split by case and control status) showed a high prevalence (72%) of the phylum Firmicutes followed by Proteobacteria (12%), Actinobacteria (8%) Bacteroidetes (7%) and Fusobacteria (1%). The Firmicutes was the most diverse phylum containing 93 distinct OTUs (51% of the total OTUs), followed by Actinobacteria with 30 OTUs (16%), Proteobacteria with 20 OTUs (11%), then Fusobacteria and Bacteroidetes with 18 and 19 OTUs respectively (10%) and finally Cyanobacteria with 2 OTUs (1%). There were no statistical differences when phyla present in cases versus controls were compared).

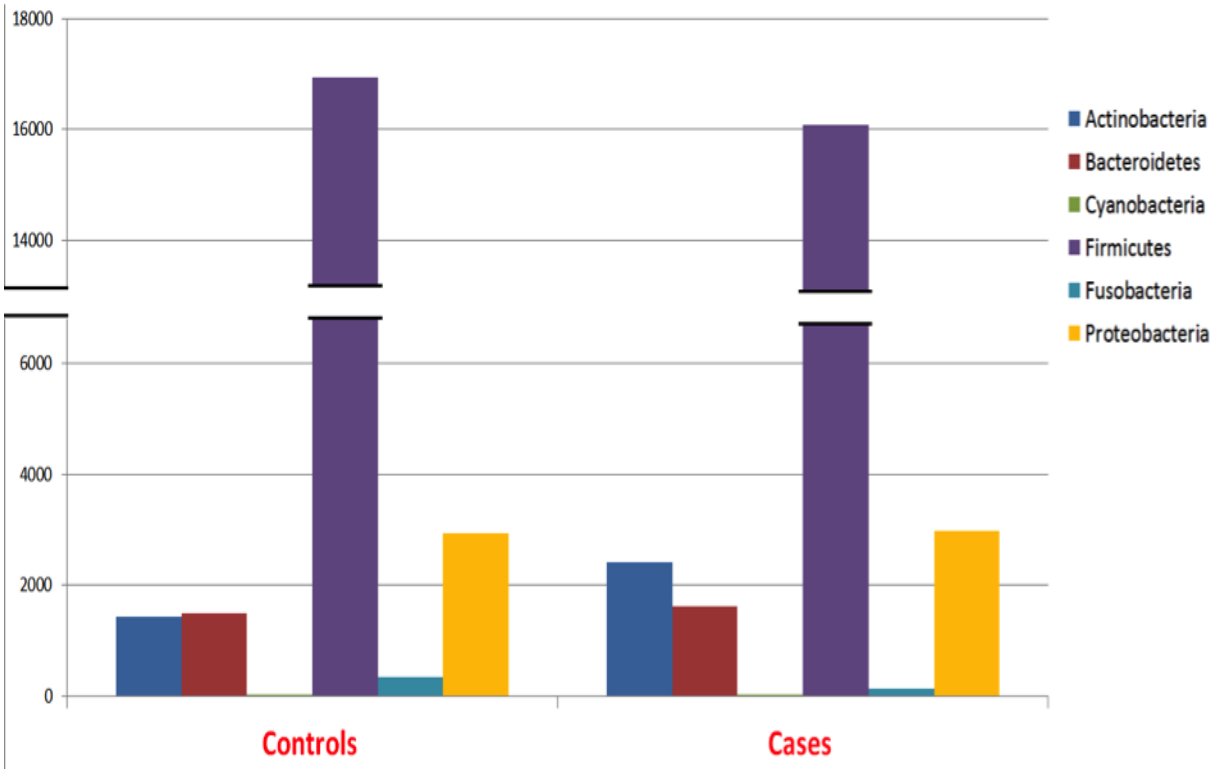


Figure 3.7 Phyla Distribution Cases versus Controls. Y axis shows the number of sequences rarefied. There were no statistical differences.

For the 182 OTUs identified at 97% sequence identity level, a representing sequence of each of the OTUs was used to assemble a phylogenetic tree using FASTREE (Price *et al.*, 2010b) and heatmaps representing the number of sequences per OTU were constructed using ITOL (Letunic and Bork, 2011). Figure 3.8 shows a heatmap with every sample included whilst Figure 3.9 shows the total of sequences pooled in cases and controls.

The OTUs assigned to *Streptococcus* were the most prevalent (32 OTUs), only 6 of the 32 however showed statistical differences (calculated by metastats with 1000 iterations) between cases and controls decreasing to three if only those with a sequence frequency of 100 were considered. Of these three OTUs, one assigned to *Streptococcus* was more prevalent in cases (*P*=0.0003,

OR 1.20), whilst the other 2 OTUs (also assigned to *Streptococcus)* were more prevalent in controls (*P*=0.0001, OR 0.70 and *P*=0.0220, OR 0.23 respectively).

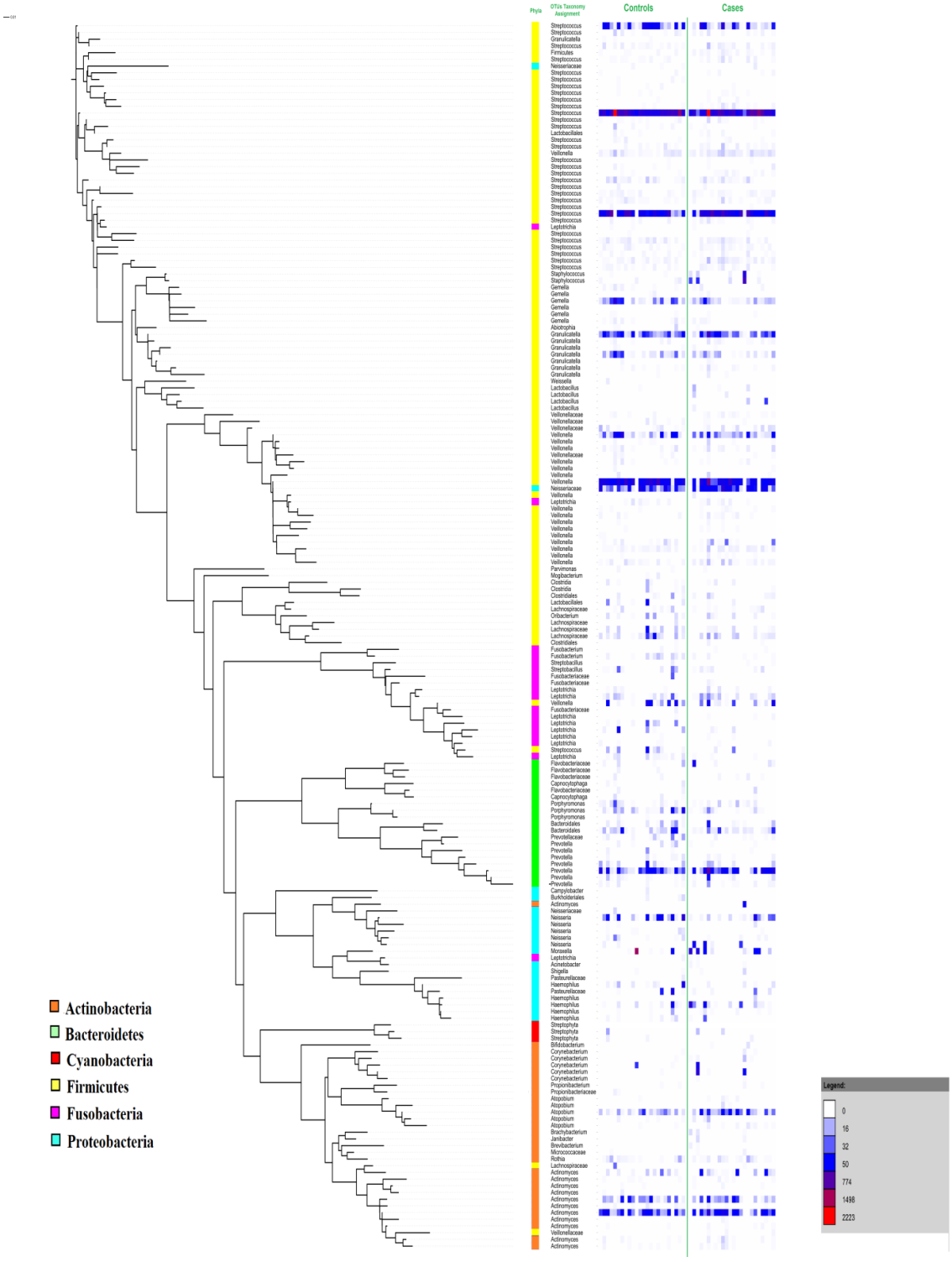


Figure 3.8. Neighbour Joining Phylogenetic Tree of Bacterial 16S rRNA using the OTUs found using Pyrosequencing. Boxes in the right show the Heatmap of sequences frequencies per OTU in Controls and Cases, and the taxonomy assignments at Phyla level (inner column – colour coding for the Phyla left panel). The Legend in the right bottom describes the number of sequences assigned to each colour.

In Firmicutes the OTUs assigned to *Gemella* (*P*=0.0039, OR 0.39), *Lachnospiracea* (*P*=0.013, OR 0.44) and two OTUs assigned to *Veilonella* (*P*=0.0209, OR 0.52; *P*=0.0397, OR 0.54 and *P*=4.75×10^{-72}, OR 0.58) were more common in controls compared to cases. The OTUs assigned to *Staphylococcus* were different when compared between numbers of sequences between cases and controls (on pooled samples) although further statistical analysis using MetagenomeSeq (which takes account the overall distribution on groups) it was apparent that the differences were due to the presence of a higher number of sequences in just a few individuals (Figures 3.8 and 3.9).

In Actinobacteria the most abundant OTUs were significantly more prevalent in cases than in controls e.g. *Actinomyces* (*P*=0.0017, OR 2.85), *Atopobium* (*P*=0.0014, OR 2.17) and *Corynebacterium* (*P*=0.0004, OR 7.95).

Within the Phylum Bacteroidetes, an OTU assigned to *Porphyromonas* (*P*=0.0039, OR 0.11) was more prevalent in controls whilst the OTU assigned to Flavobacteriacea (*P*=0.0270, OR 12.16) was more prevalent in cases.

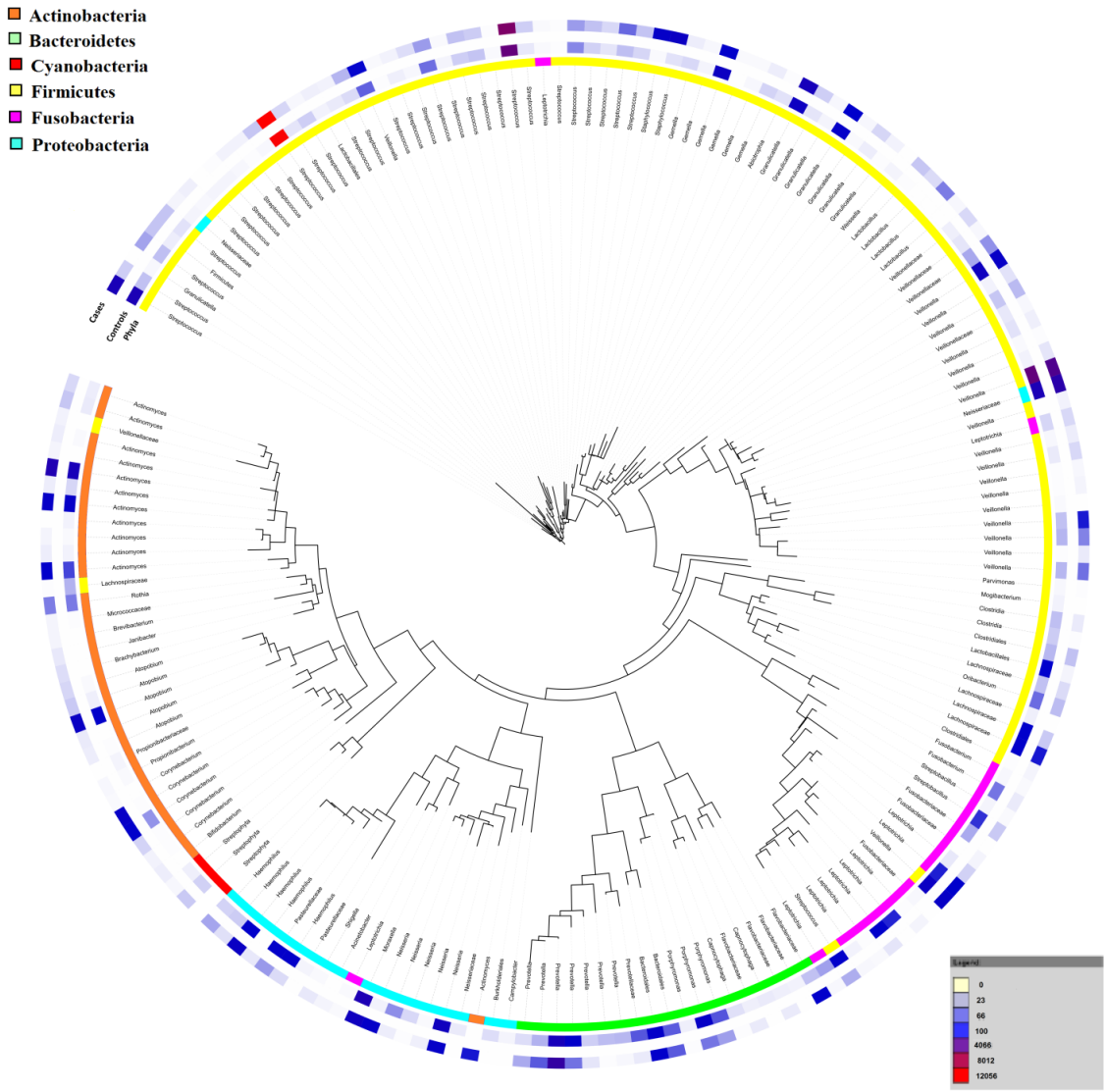


Figure 3.9. Neighbour Joining Phylogenetic Tree of Bacterial 16S rRNA grouped by OTUs and pooled by Cases and Controls. Boxes in the border rings depict the frequency of sequences for each operational taxonomic unit (OTU) in Controls (middle ring) and Cases (outside ring). The taxonomy assignments at the Phyla level are shown in the inner ring. The legend in the right describes the abundance of OTUs (number of sequences/per group) whilst top left the colour coding for the different Phyla is given.

For the phylum Proteobacteria, one of the OTUs assigned to *Haemophilus* (*P*=0.0040, OR 5.06) and one assigned to Neisseriaceae

(*P*=0.0031, OR 1.20) were more prevalent in cases. In contrast an OTU assigned to *Neisseria* (*P*=0.001, OR 0.40) and another OTU assigned to *Haemophilus* (*P*=0.012, OR 0.45) were more abundant in controls.

The most important differences overall at the OTU level are summarised in Table 3.1.

OTUs	No. Sequences		*P** value	Odds Ratio	Confidence Interval 95%
	Controls	Cases			
Actinobacteria/Actinomyces	50	142	0.0017	2.85	2.06 to 3.94
Actinobacteria/Atopobium	169	364	0.0014	2.17	1.80 to 2.61
Actinobacteria/Corynebacterium	26	205	0.0004	7.95	5.28 to 11.96
Bacteroidetes/Flavobacteriaceae	13	157	0.0270	12.16	6.90 to 21.42
Bacteroidetes/Porphyromonas	226	26	0.0039	0.11	0.07 to 0.17
Firmicutes/Gemella	336	133	0.0039	0.39	0.32 to 0.48
Firmicutes/Lachnospiraceae	129	57	0.0013	0.44	0.32 to 0.60
Firmicutes/Streptococcus 1	3221	3756	0.0003	1.20	1.13 to 1.26
Firmicutes/Streptococcus 2	948	671	0.0001	0.70	0.63 to 0.77
Firmicutes/Streptococcus 3	87	20	0.0220	0.23	0.14 to 0.37
Firmicutes/Veillonella 1	494	257	0.0209	0.52	0.44 to 0.59
Firmicutes/Veillonella 2	3154	1948	0.0397	0.58	0.54 to 0.61
Proteobacteria/Haemophilus (OTU 162)	97	482	0.0040	5.06	4.06 to 6.2
Proteobacteria/Haemophilus (OTU 190)	86	43	0.0112	0.45	0.38 to 0.52

Proteobacteria/Neisseria	525	213	0.0001	0.40	0.34 to 0.47
Proteobacteria/Neisseriaceae	1025	1217	0.0031	1.20	1.10 to 1.30

Table 3.1. Differences at OTU level in Bacterial 16S rRNA sequences from Throat Swabs of Infants with Wheeze versus Healthy Controls. *Only groups with more than 100 sequences and statistically significant differences are shown.

Alpha diversity analysis (Chapter 2: Section 2.7.5) using the Chao1 was estimated in 289.84 OTUs (95% CI 218.39-361.30). Statistical analysis revealed no differences in richness (the maximum of expected OTUs that could be found if the sequencing continues until the infinite) when comparing cases versus controls (*P*=0.852) and in overall it (Figure 3.10 A). Using the Shannon Index and observed OTUs to measure the diversity in taxa, no significant difference between groups was found (*P*=0.921) (Figure 3.10 B). Evenness was measured using the Equitability index and again no differences were seen (*P*=0.702) (Figure 3.10 C).

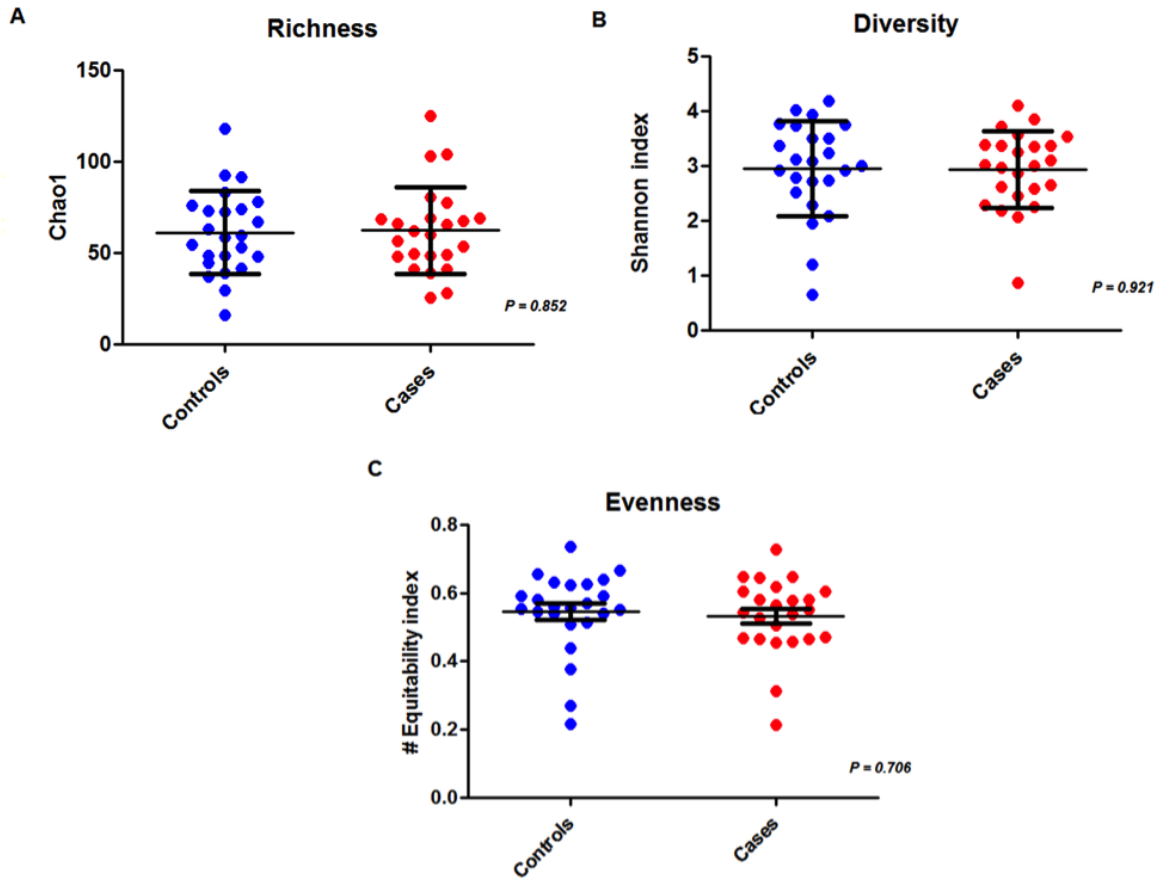


Figure 3.10. Alpha Diversity Statistics. A) Scatter dot plot comparing cases versus controls values of Chao1 richness index. B) Scatter dot plot comparing values of Shannon diversity index. C) Scatter dot plot comparing equitability evenness index.

When compared the results obtained in the same samples using conventional sequencing (Stage 1a Pilot Study Section 3.2.1.2) and pyrosequencing, a higher alpha diversity was encountered. Due to the greater depth of sequencing that is cost and time effective by the pyrosequencing approach one consequently will detect less common OTUs hence the higher alpha diversity. When comparisons at the whole microbial community level or at the major OTUs level were performed, there were no statistically significant differences observed.

Multiple rarefaction curves using the Shannon index (Figure 3.11) showed that a plateau of diversity was achieved at around 360 sequences per sample. This value would be considered as the minimum sampling depth to capture diversity. In the Stage 1a pilot 120 sequences per sample were generated (by Sanger sequencing of randomly picked clones) considerably below the plateau of diversity. Using pyrosequencing, the rarefaction of the complete Stage 1a sample set was performed at a minimum of 969 sequences per sample, which will therefore have captured a realistic panorama of each sample's diversity.

Beta diversity statistics (Chapter 1: Section 1.12) using the Principal Coordinate Analysis (PCoA) of the UniFrac distance matrix (weighted and unweight UniFrac, distance-based ANOVA with permutations) revealed no significant differences in the microbial communities cluster patterns between cases and controls (Figure 3.11).

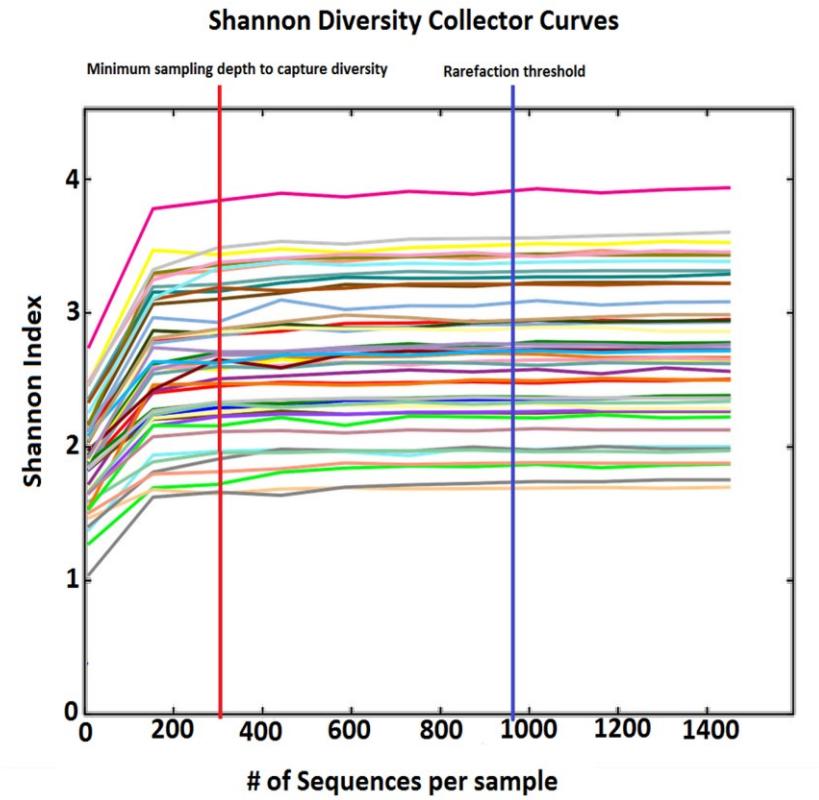


Figure 3.11. Shannon Diversity Collector Curves. Multiple rarefaction curves were collated from each sample's Shannon diversity index. The graphic shows the estimated diversity plotted against the number of sequences per sample. Each line represents one sample. The plateau in each estimated diversity curve indicates the minimum number of sequences to capture diversity. For all samples the plateau was achieved at approximately 320 sequences (red vertical line), well below the chosen rarefaction threshold of 969 sequences (blue line).

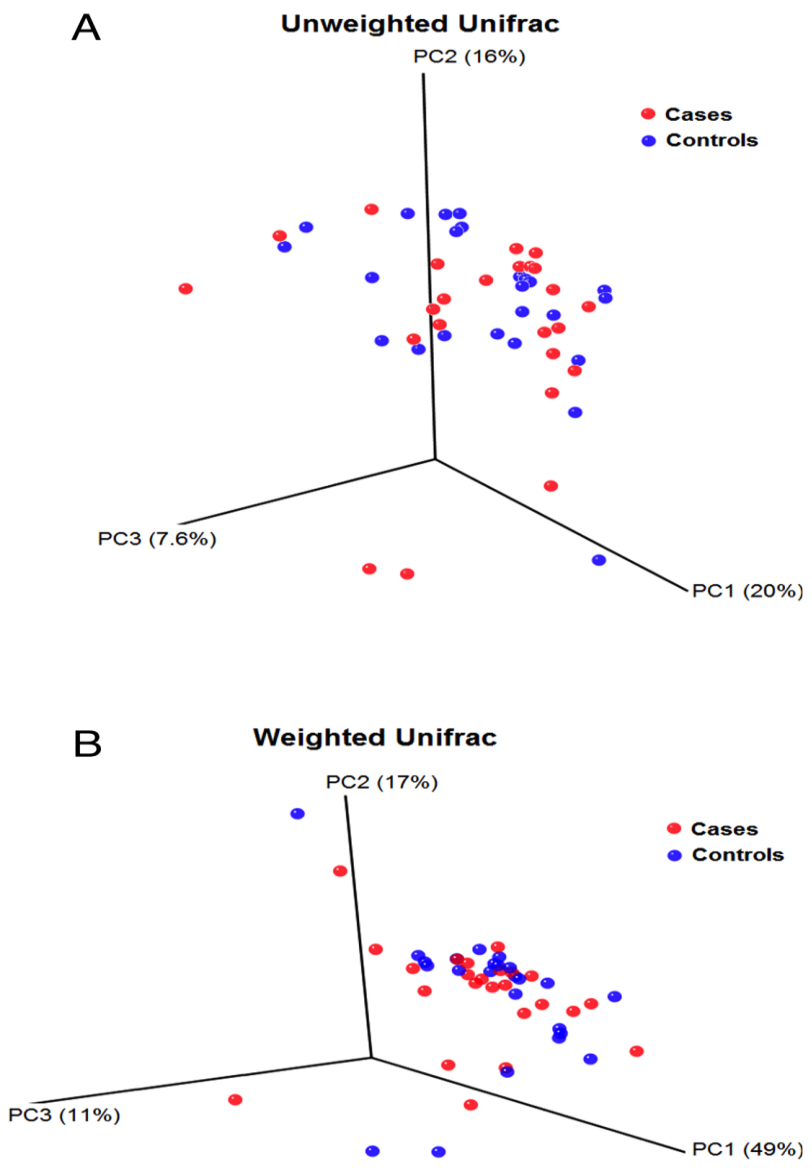


Figure 3.12. Beta Diversity Statistics using the Principal Coordinate Analysis (PCoA). A) Unweighted UNIFRAC Principal Coordinate Analysis (PCoA) plot comparing presence/absence metrics. B) Weighted UNIFRAC Principal Coordinate Analysis (PCoA) plot comparing presence/absence metrics and abundance using the UniFrac distance matrix (weighted unweight UniFrac). The results show no cluster patterns different between cases (red dots) and controls (blue dots).

3.3.3. Phylogenetic Identification using Sequence Alignments

Using three independent phylogenetic treeing methods it was not possible to increase the specificity of the identification of the *Streptococcus* spp. OTUs beyond that of the basic Ribosomal Database Project (RDP) classifier (Cole *et al.*, 2009). Tree topology between the three methods of treeing was not conserved and significance of assignment to major nodes was low.

By contrast, the tree topology for *Haemophilus* spp. was conserved in all three methods of phylogenetic inference (Figure 3.12), with robust significance for assignment for each *Haemophilus* spp. OTU. This enabled confident assignment of OTU 162 to *Haemophilus influenzae*, OTU 32 and 38 to *Haemophilus haemolyticus* (both with less than 100 sequences and therefore not detailed in Table 3.1), and OTU 190 to *Haemophilus parainfluenzae*. OTU 162 which was assigned to the known pathogen *Haemophilus influenza*e was significantly more common in cases (P = 0.004, OR 5.06) compared with controls, whilst OTU 190 (assigned to *Haemophilus parainfluenza*e) was significantly more prevalent in controls (P = 0.0112, OR 0.45).

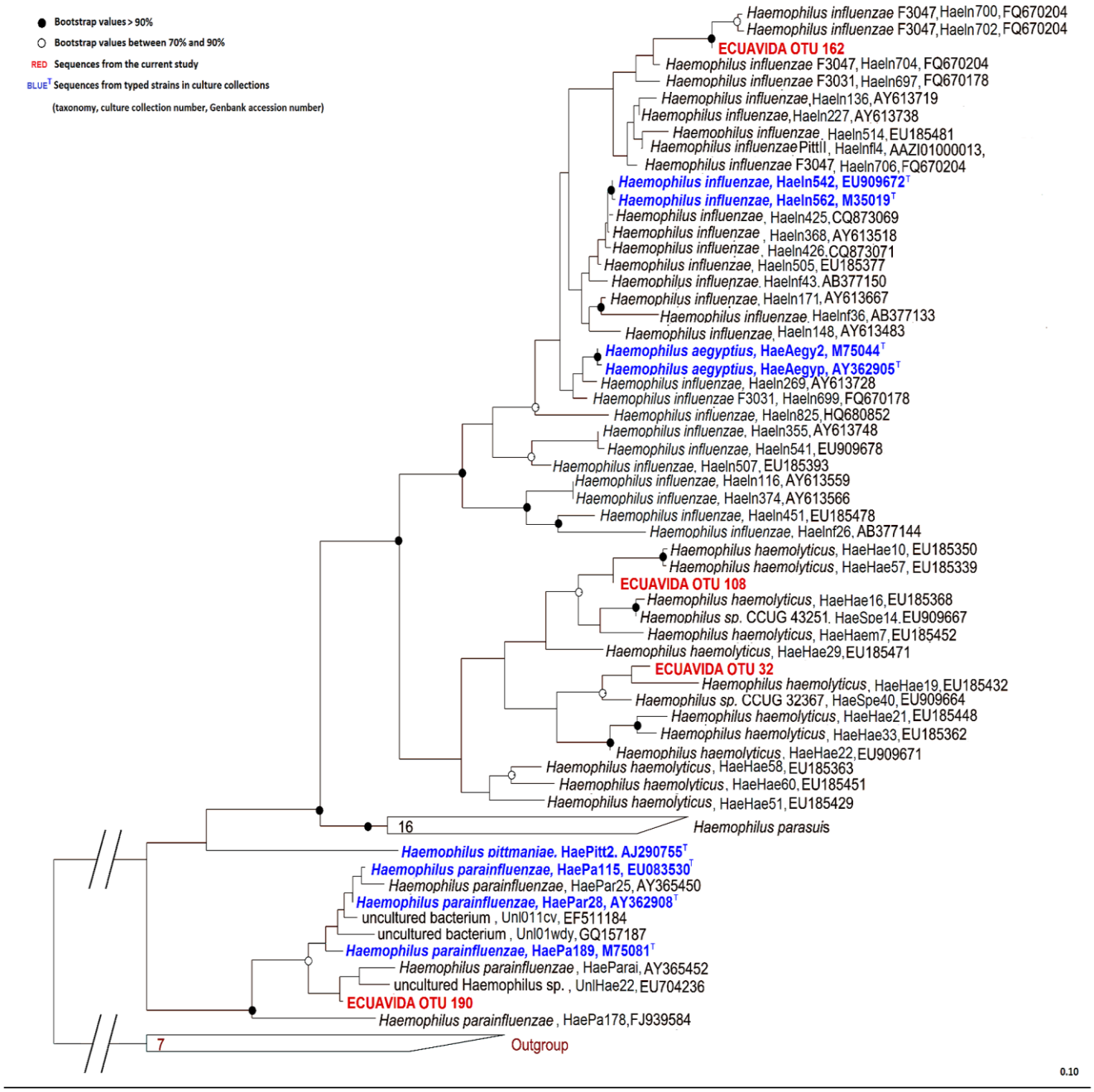


Figure 3.12. Phylogenetic Identification of *Haemophilus* OTUs. Phylogenetic analysis of the 16S rRNA sequences of the OTUs assigned taxonomically to *Haemophilus* genus (OTUs 32, 108, 162 and 190, shown in Red) together with reference *Haemophilus* sequences from the SILVA database, using the ARB alignment editor. The scale bar indicates 10% sequence divergence, and NCBI accession numbers are included. The tree was rooted with a near neighbour outgroup constructed with sequences from *Morganella morganii*, *Proteus mirabilis* and *Providencia stuartii* (Accession numbers Section 3.2.2.1).

3.4 Discussion

In this study the characterization of the upper airways bacterial microbiome was conducted to enable comparisons between early onset wheezing children suffering a current wheezing episode (cases) and healthy controls from the tropics of Ecuador. The results of the study reveal that significant differences in diversity and abundance of bacteria exist between cases and controls when a wheezing episode is occurring. As has been reported in previous studies (Hilty *et al.*, 2010, Bisgaard *et al.*, 2007), and independently of the technique used (DGGE, Sanger sequencing and pyrosequencing) it was possible to identify in cases a higher prevalence of potentially pathogenic bacteria (*Neisseria* sp. and *Haemophilus* sp.). This finding is consistent with studies that have used standard culture techniques for bacterial identification and classification (Bisgaard *et al.*, 2007). Additionally in concordance with previous studies of the airways microbiome (Hilty *et al.*, 2010), Firmicutes was the most prevalent phylum found in the infants´ airways microbiome with the majority of the Firmicutes belonging to the *Streptococcus sp.* genus.

At the beginning, a pilot study using throat swab samples from 25 infants of Stage 1a and PCR-DGGE was conducted to estimate the diversity present on each sample as well as to provide an indication of how many clones should be Sanger sequenced. PCR-DGGE however encountered several issues when determining the microbiome taxonomy and diversity, for example the number of bands observed do not always reflect the diversity present, a single band can represent more than one organism, it is necessary to use a ‘marker’ on the gel to extrapolate each band bacteria taxonomy, and usually there are bands that

do not correspond to any assignation by the marker therefore their identification remains unknown.

PCR-DGGE failed in the quantification of bacterial diversity (by the identification of a real number of species), for instance several samples with a single band on the PCR-DGGE had a several number of different bacterial species when analysed by sequencing. In regard of taxonomical identification, PCR-DGGE was able to detect the high prevalence of *Streptococcus* spp. later determined using DNA sequencing but the technique was not accurate enough to differentiate them from *Prevotella* spp. The second problem of bacterial identification using PCR-DGGE is the difficulty that was encountered in assigning the gel bands of the samples to bands of the markers used as this was completely subjective and done by eye. As a result for the majority of the gel bands it was impossible to assign a bacterial genus. In conclusion, this technique did not allow comparison quantitatively of the bacterial microbiome in cases versus controls.

The PCR-DGGE experiment however did provide an indication of the number of clones that should be sequenced per sample in order to capture the real diversity present in each sample. Based on this tool and with the previous work within the Molecular Genetics Group (Hilty *et al.*, 2010) one hundred and twenty clones were sequenced per sample in the pilot study. Using Sanger sequencing, the cases group showed a more complex diversity in the microbiome compared with the control group, and most of the diversity was due to the presence of a higher frequency of *Proteobacteria* and a lower frequency of *Streptococcus* spp. Infants with early onset wheezing and suffering a current wheezing episode therefore had more bacterial species compared with controls.

As anticipated from the initial PCR-DGGE screening, Firmicutes was the most common (90.9%) and most diverse phylum seen (32 of 52 OTUs) using Sanger sequencing. Most of the OTUs found in this group belong to the *Streptococcus* spp. group, which was expected based on prior observations in children (Hilty *et al.*, 2010). In the pilot study all the OTUs in Proteobacteria were significantly different between cases and controls, as their prevalence in healthy individuals was almost zero. Strikingly the diversity in the cases was greater than that seen for the controls, and most of this increased diversity was attributable to a higher amount of Proteobacteria spp. being present. Cases therefore had more bacterial species than controls with the majority of the species being potentially pathogenic bacteria (*Neisseria sp.* and *Haemophylus sp.*). This finding is consistent with the prior study of Bisgaard *et al.* (2007) that used culture-based approaches to look at microbial diversity.

Although there was a reduction in Bacteroidetes present in cases in the pilot study, this was not statistically significant. A decrease in the presence of Bacteroidetes has however been seen previously in adults and children with disease (Hilty *et al.*, 2010) and species of this phylum may therefore be considered possible 'protective' bacteria. Several studies have shown that *Prevotella sp.* (member of the Bacteriodetes phylum) inhibits the grown of other bacteria in the oropharynx (Murray and Rosenblatt, 1976).

When the whole microbial community of the cases versus the controls was compared using UNIFRAC, statistical differences between both communities were found in the pilot study. This means that in the upper airways, diseased infants with current wheezing and healthy children do appear to have distinct microbiomes.

Using conventional cloning and sequencing methods, it was possible to identify 52 OTUs. Chao 1 richness statistics using rarefaction curves have however calculated a maximum of 147 OTUs would exist in the airways microbiome. That means that with the current sequencing data generated only 35% of the possible oropharynx bacterial microbiome in these samples has been identified. If cases and controls are considered separately as they have shown different diversities in the pilot study, the Chao1 index indicated that for cases 18% of the expected OTUs have been identified in contrast to 66.9% for controls.

On the pyrosequencing analysis the profundity of sequencing was increased to 969 sequences per sample. This allowed a more complex analysis of the upper airways bacterial microbiome with an increased range of exploration of the less common bacterial species, and as consequence a better panorama of the real diversity. Likewise using Sanger sequencing, pyrosequencing analysis found similar relations Firmicutes/Proteobacteria in cases and controls, but the diversity measured by alpha and beta diversity indexes did not show dissimilarities. These results also could be explained by the improved sensitivity of the next generation sequencing techniques in detecting less common bacterial species, specifically by not falling in bias caused by the randomization at the time of clones picking (Petrosino *et al.*, 2009). Using pyrosequencing, Firmicutes likewise turned out to be the most prevalent phyla found (72% of the total sequences on the 48 samples), and the most diverse with 93 OTUs found (51% of the total).

Likewise as by conventional sequencing, pyrosequencing encountered that the most common genus consistently identified was *Streptococcus* spp.,

and overall, when sequences are sumarized by genera level, this genus is more abundant in controls than in cases. Studies have hypothesised that in the upper airways, *S. viridans* could act as protective bacteria replacing possible pathogenic bacteria (Tano *et al.*, 2000). The results from Stage 1a harmonize with these findings, as not only do they demonstrate that there is a higher prevalence of *Streptococcus* spp. in healthy children but also that there is a lower frequency of posible pathogenic bacteria especially those belonging to the Proteobacteria phylum. However when the analysis is performed at OTU level, some OTUs assigned to this genus were more common in cases whilst others in controls, and the ambiguity of 16S rRNA did not allow differentiation between them at species level.

Other common Firmicutes that were found in the study are: *Veillonella* sp. (Gram negative strict anaerobic cocci) and *Gemella* sp. (Gram positive facultatic aerobic cocci), which are commensal bacteria that similar to *S. viridans* can potentially cause of disease (bacteraemia and endocarditis) after invasive orthodontic procedures or in immunocompromised patients (Shenep, 2000, Bochud *et al.*, 1994), and could also be considered protective against asthma in the airways microbiome.

Proteobacteria was the second most diverse phylum in both the pilot study and the pyrosequencing study. In the pyrosequencing study the majority of OTUs belonging to this phylum were more common in cases, however if only the OTUs with more than 100 sequences are taken account then the OTUs found are equally distributed between the two groups. Strikingly, most of the OTUs found in this phylum were important etiological agents of disease in children such as: *Neisseria* sp. and *Haemophilus* sp. These bacteria are the

most prevalent etiological agents of diseases in the airway related structures for example including lungs, meninges, paranasal sinus, etc (Rouphael and Stephens, 2012, Sabra and Benger, 2011, Watt *et al.*, 2009, Gkentzi *et al.*, 2012).

In Bacteroidetes using pyrosequencing, *Porphyromonas* was encountered more commonly in controls whilst Flavovacteriacea in cases. Moreover, *Prevotella* was more common in controls (in concordance with the pilot study) but these results were driven by a few samples and therefore were not found to be significant when using microbiome specialized statistics (UNIFRAC and metastats).

Actinobacteria was consistently more commonly found in cases (also using pyrosequencing), and within this phylum are potentially pathogenic genera such as *Actinomyces, Atopobium* and *Corynebacterium* (Kutluhan *et al.*, 2011, Francavilla *et al.*, 2014). These genera are important agents of diseases in infants, and their range of diseases go from mild cold symptoms to important opportunistic infections and life threatening illnesses as diphtheria (Byard, 2013). It is interesting however than *Atopobium* spp. have been related with other chronic inflammatory diseases as celiac disease (Francavilla *et al.*, 2014, Collado *et al.*, 2007).

Due to the limitation of cloning and sequencing (sequencing a wide number of clones selected randomly), it is usually not possible to identify more than 70% of the microbiome present in a sample (Hilty *et al.*, 2010). Consequently, it is estimated that more than 90% of the microbiome in a sample could be captured by pyrosequencing (Turnbaugh *et al.*, 2007). Using alpha

diversity collector curves, it was possible to determine that the sampling depth performed on Stage 1a reached the plateau of diversity identification and therefore the perspective obtained of the microbiome was very accurate.

The biological interpretation of the differences in the frequencies of individual OTUs is limited by the imprecision of OTU assignments in identifying individual species. In particular, this problem was obvious in our data with OTUs assigned to *Streptococcus* and *Haemophilus spp.*, which were likely to contain a mixture of pathogenic and non-pathogenic strains. An attempt to improve the classification of these OTUs was made by including them in phylogenetic tress constructed from reference sequences. The incorporation of reference strain sequences into our phylogenetic trees allowed the discrimination between *Haemophilus* spp. OTUs at species level, and to show that the pathogen *H. influenzae* was more prevalent in wheezing infants, in concordance with previous studies (Hilty *et al.*, 2010, Bisgaard *et al.*, 2007). *H. parainfluenzae* was more abundant in healthy children which might be associated with wheezing protection. The 16S rRNA gene has previously been suggested for use as a marker in MLST of *H. influenzae* (Sacchi *et al.*, 2005) and the results presented in this chapter confirm that its phylogeny is well matched to species and strain identification in airway samples.

In conclusion, the upper airways microbiome of early onset wheezing infants from the tropics at the time of a wheezing episode have important differences in the number and types of bacteria when compared to healthy infants from the same environment. It is noticeable that in controls there is an elevated presence of possible ‘protective’ bacteria, whilst in cases there is a lack of them in addition to a higher level of potentially pathogenic bacteria.

Although the 16S rRNA sequences through more sophisticated phylogenetic tree analysis allowed identification of *Haemophilus* spp., this was not feasible for *Streptococcus* spp. due to the genetic uniformity of the 16S rRNA gene in this genus. This is clearly an important group and discrimination would increment our understanding further of the respiratory microbiome (it is not possible to discriminate other important groups of bacteria as *S. pneumoniae* from their non-pathogenic pairs the *S. viridans*). Consequently an approach involving alternative genes in combination with pyrosequencing to allow further differentiation of the *Streptococcus* spp. was developed (Chapter 5).

Chapter 4: Mid and Long Term Changes in the Upper Airways Microbiota and their Relationship with Wheezing - ECUAVIDA Stages 1b and 2

As described earlier in Chapter 2, Section 2.3, the ECUAVIDA microbiota study was separated into two Stages (1 and 2) according to age. Stage 1 was focused on infants less than 12 months of age with Stage 2 being focused on infants' 2 years ± 2 months old. Additionally Stage 1 was further subdivided into two parts. Part 1a: samples taken during a current wheezing episode and Part 1b: samples taken during a past wheezing episode. For Stage 2, sampling was as per part 1b (wheezing history and no current wheezing at the time of sampling) but when children were 24 months old ± 2 month.

This chapter deals with the data generation and subsequent analysis and results for Stages 1b and 2. As the two stages only differed by age (all children had a history of past wheezing episodes), it affords the opportunity to investigate the mid and long term changes of the upper airways microbiota that occur as a consequence of multiple "non-infectious" wheezing episodes. Additionally the current chapter also tries to determine the differences produced in the upper airways bacterial microbiota between a current wheezing episode and the variations that occur after the episode by comparing the results from Chapter 3 (children Stage 1a) and this chapter (children Stage 1b and 2). A summary of the different stages compared in this chapter is shown in Figure 4.1.

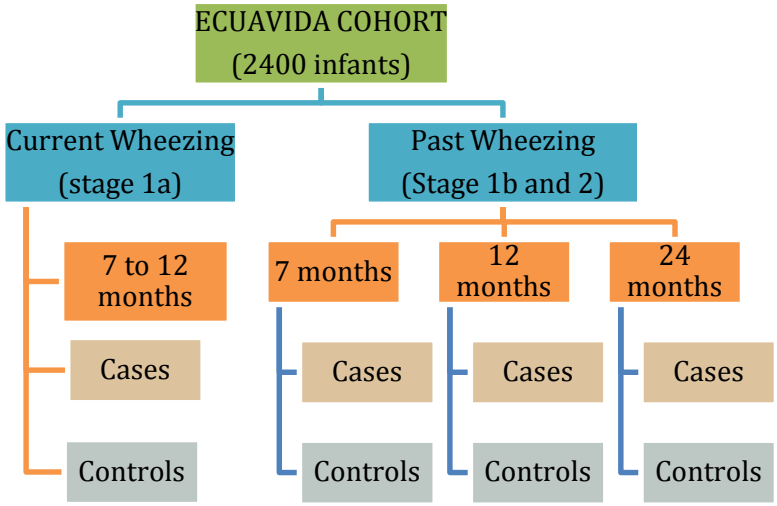


Figure 4.1. Summary of the Project Stages.

4.1 Materials and Methods

As part of the general inclusion criteria of the study, cases and controls have no history of frequent antibiotic use and no recurrent respiratory infections (Chapter 2: Section 2.3). Additionally it was ensured that, at least during the last two weeks prior to sampling, the subjects were not using antibiotics for any reason and did not have any evidence of a current airway infection (absence of cold symptoms and fever). All the subjects belonged to the same mixed ethnical background, live in the same town and had access to the same basic services (electricity, water, sanitation, etc.). All of them received the vaccines recommended by the Ecuadorian Public Health Ministry (BCG, mumps, measles and rubella, tetanus, diphtheria, hepatitis B and *Haemophilus influenza* B; MSP, 2005) but none of the subjects received anti Streptococcal vaccination (not compulsory for the MSP) (Chapter 2: Section 2.3).

After collecting the samples between October 2010 and April 2011 in the rural District of Quininde in the tropics of Ecuador, they were shipped on dry ice to the United Kingdom. DNA extraction was performed using a modified kit protocol standardized by the Molecular Genetics and Genomics Group (Chapter 2: Sections 2.4 and 2.5) PCR of the 16S rRNA gene V3-V5 segments was performed also as previously described (Chapter 2: Section 2.7).

For the comparison between current and past wheezing section, 3 pyrosequencing runs were merged. For this part of the study the DNA sequences obtained from the Stage 1a of the project were compared with the results from the Stage 1b. Both stages contain samples from infants that range on age between 7 and 12 months and belong to the same community of the Quininde District on the rural tropics of Ecuador. For the mid and long-term changes after wheezing section, six pyrosequencing runs were performed on the Roche 454 Junior sequencer using the samples from Stages 1b and 2 of the study.

The resulting data were merged together after sequencing error correction had been performed using Ampliconnoise (Chapter 2: Section 2.7.5). Results were analysed with QIIME 1.7 and the phyloseq script on R. Statistical analysis was performed using Metastats adnon the stat4 script on R. Graphics were developed on QIIME, GraphPad 6.0 and R (Chapter 2: Sections 2.7.5).

4.2 Results

4.2.1 Current versus Past Wheezing (Stages 1a versus 1b)

4.2.1.1 OTUs Comparisons

For Stages 1a versus 1b a total of 370,000 sequences were analysed from 246 samples.

At phyla level the patterns on the upper airways microbiota were similar between acute wheezing sufferers and past wheezing sufferers. In both groups Firmicutes remained the most common phylum, whilst overall the rest of phyla maintained their abundance and diversity (between Stages 1a and 1b).

At OTUs level there were encountered several differences when comparing the upper airways microbiota in infants with a current wheezing episode and infants with a past wheezing episode. However the most abundant OTUs in the current wheezing group were the same as for the past wheezing group, *Streptococcus* being the most common bacteria followed by *Veillonella* and *Prevotella* (Figure 4.2).

When statistically comparing only cases in both groups (current and past wheezing infants), it was possible to determine differences in several OTUs. In the phylum Actinobacteria there were encountered differences in three OTUs taxonomically assigned to *Actinomyces* spp. that were more commonly found in the current wheezing group (*P* values = 0.0051, 0.00095 and 0.000091 respectively). Also in this phylum one OTU assigned to *Corynebacterium* was more common in the current wheezing group (*P*=0.010).

In Bacteroidetes, an OTU appointed as *Prevotella* sp. was more abundant in the past wheezing group (*P*=0.017) but another OTU assigned to *Prevotella melaninogenica* was more commonly found in the current wheezing group (*P*=0.00009). Two OTUs assigned to *Porphyromonas* spp. were more commonly found in the current wheezing group (*P* values = 0.001 and 0.006).

In Firmicutes, *Lactobacillus* was more abundant in the past wheezing group than in the current wheezing one (*P*=0.001). In contrast the genus *Sthaphylococcus* was more common in the current wheezing group (*P*=0.012). Between the OTUs assigned to *Streptococcus* spp., several of them (OTUs 1041 and 1120) were more common in current wheezers (*P*=0.017 and 0.016), but most of them (OTUs 427, 179, 524 and 1142) were more common in past wheezers (*P values* = 0.0001, 0.0002, 0.03 and 0.024). In this phylum *Veillonella* spp., Gemellacea and *Leptotrichia* spp. were more common also in past wheezers (*P*=0.002, 0.0009 and 0.0044).

In Proteobacteria, the OTUs assigned to *Neisseria, Haemophilus* and *Moraxella* were all more abundant in the current wheezing infants (*P values* = 0.0009, 0.0009 and 0.022) compared with the past wheezing infants.

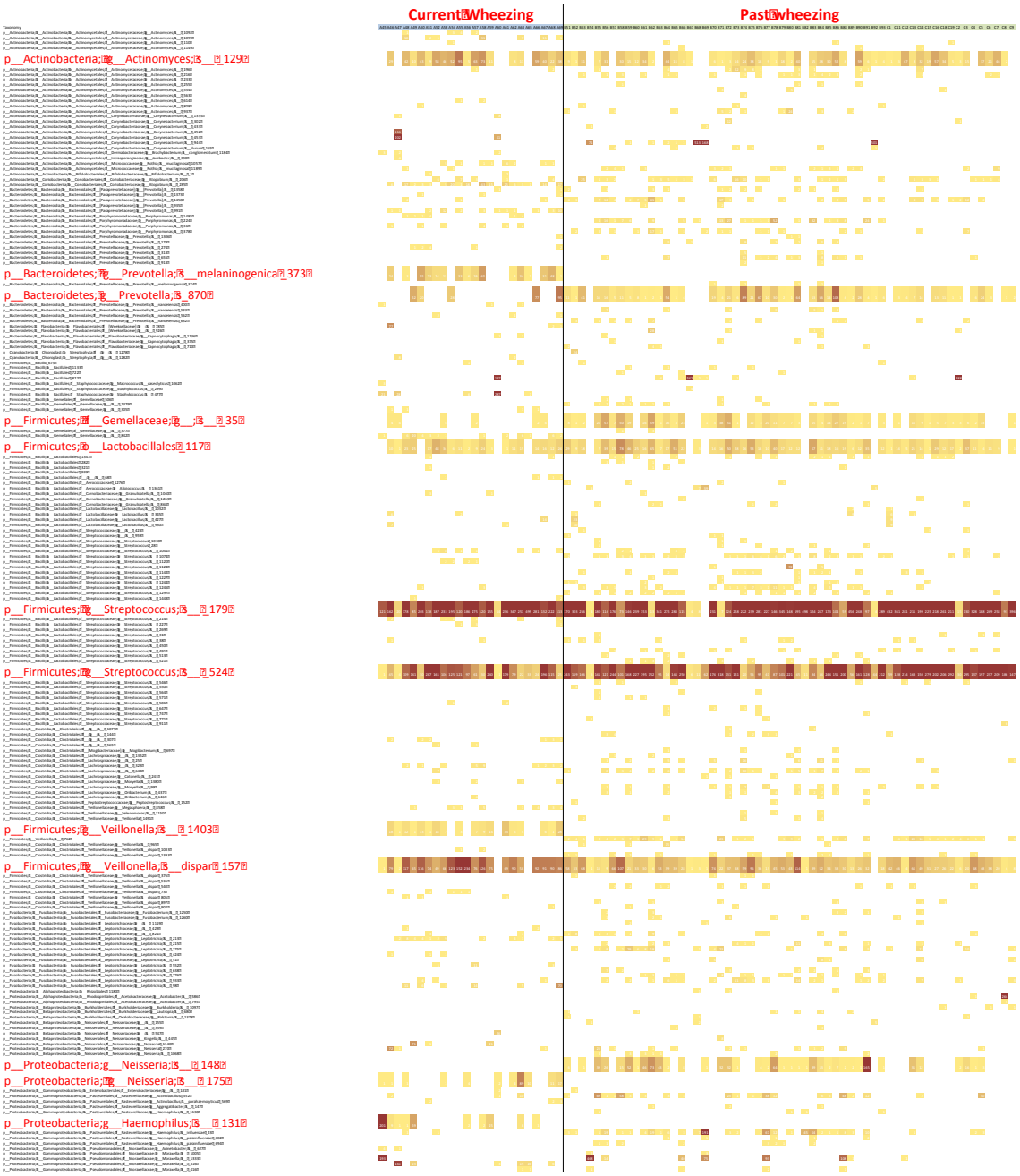


Figure 4.2. Current versus Past Wheezing Heatmap. that compares the sequences abundance per OTU in cases between current wheezing and past wheezing infants. The taxonomy assignation is shown in the left and the boxes in right show the number of sequences. Dark red represents a high number of reads per OTU, yellow low number and white represents absence.

4.2.1.2 Beta Diversity

For the beta diversity analysis of Stages 1a and b, analysis was almost identically done to the pilot study analysis with the exception of the addition of the UNIFRAC statistics there have been added to the analysis and the Bray-Curtis and Canberra statistics in order to determine differences without the use of phylogenetic information. Inclusion of these approaches has the potential to detect more sensitively dissimilarities in larger data sets. When comparing between groups at OTUs level, there were found statistical differences using Adonis Canberra (P<0.001), Bray-Curtis (P<0.001), UNIFRAC (P<0.001), and Weighted UNIFRAC (P<0.001). However current and past wheezing status explains a different proportion of the variance each time: Canberra = 58%, Bray-Curtis 14%, UNIFRAC = 15%, weighted UNIFRAC = 7%. One can conclude that the abundance of the different OTUs is not important (UNIFRAC and Canberra are presence/absence and show higher values), and that phylogenetic information is not relevant to show differences (UNIFRAC versus Canberra).

On the NMDS plots, the stress (disagreement between 2-D configuration and predicted values from regression) is 0.10 in Bray-Curtis, 0.10 in Canberra, 0.19 in UNIFRAC and 0.11 Weighted-UNIFRAC showing a very good representation in a bi-dimensional plot. The NMDS plots in Canberra show a clear clustering between current and past wheezers (Figure 4.3). The NMDS and PCoA plots using UNIFRAC and weighted UNIFRAC did not show a clear clustering between groups.

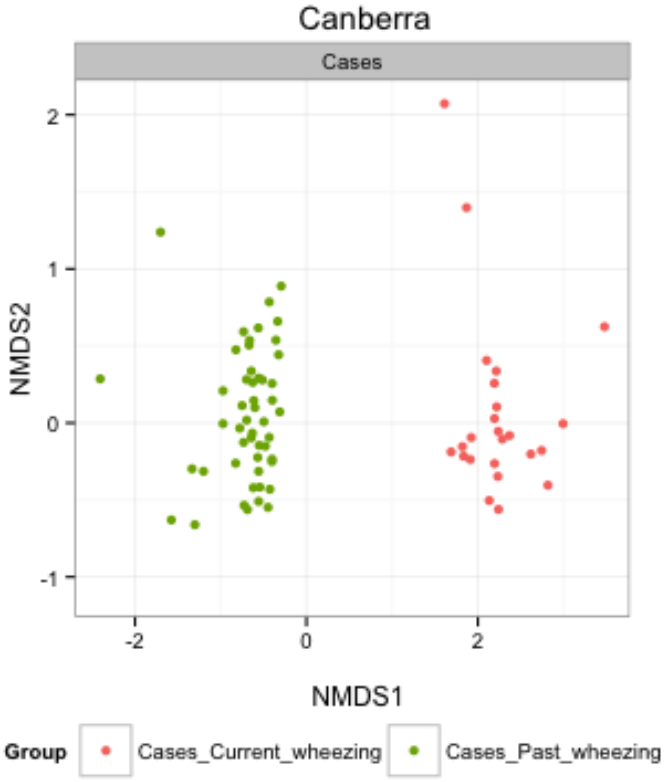


Figure 4.3. NMDS Plot using Canberra Statistics (presence/absence of OTUs). Each dot represents a sample and current wheezers are coloured on red whilst past wheezers are in green.

4.2.1.3 Alpha Diversity

For this analysis the inverse Simpson diversity was also included as it increases the sensitivity of detecting differences in non-normal distributed data. This index is complementary to Shannon but instead of measuring entropy associated with the prediction of similar abundances between groups (Shannon, 1948), the inverse Simpson measures the grade of concentration when individuals are grouped (Simpson, 1949). Alpha diversity grouped on quantiles on multiple rarefaction curves using Shapiro-Wilkinson normality tests, showed

a non-normal distribution on Richness (P<0.0001), Evenness (P<0.0001), Shannon diversity (P=0.025) and Inverse Simpson (P<0.001) (Figure 4.4). For this reason, non-parametric tests were used to analyse between the current and past wheezing groups.

Alpha diversity analysis using the Shannon index per group was performed using a rarefied OTU table at 565 sequences per sample to see if any group shows a more diverse microbiota. There were no statistical differences found (P>0.05). Kruskal-Wallis non-parametric analysis showed no statistical differences in richness, Shannon and inverse Simpson indexes when compared between current and past wheezing groups. Evenness however, was found to be statistically different between the two groups (P=0.019).

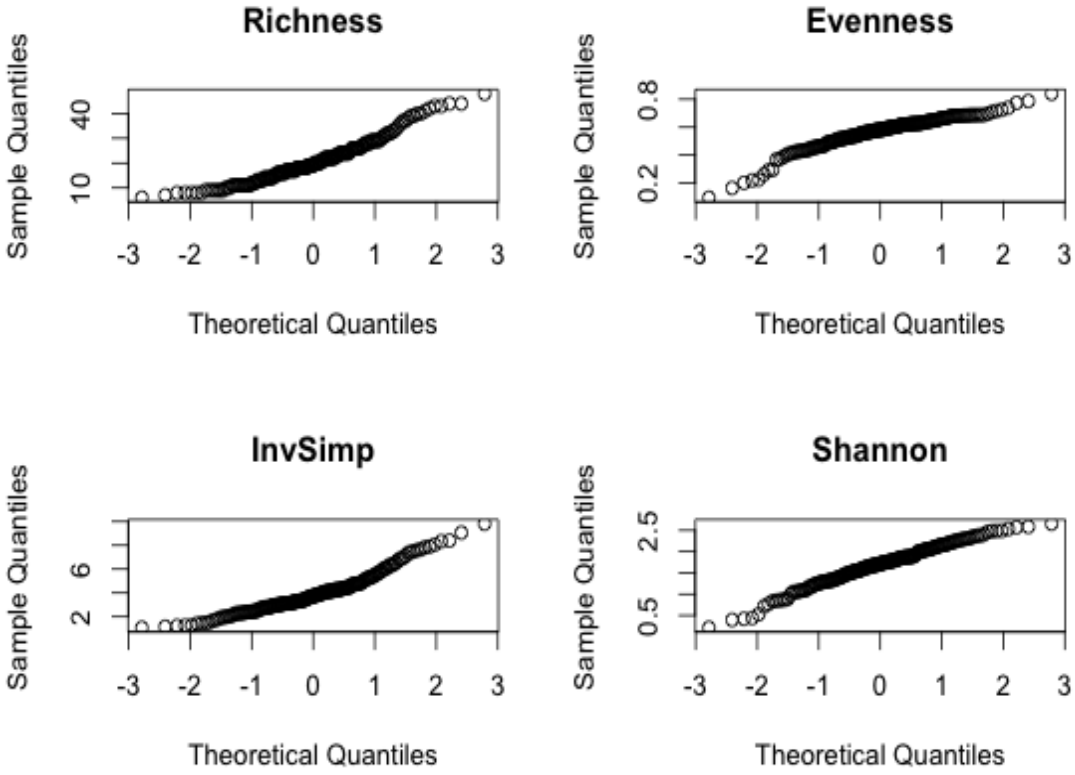


Figure 4.4. Alpha Diversity normality plots using Shapiro-Wilkinson test. Richness, Evenness, and Diversity with Inverse Simpson and Shannon indexes were calculated. Plots show the comparison of the theoretical versus sample quantiles to determine normality. None of the indexes showed normal distribution.

4.2.2 Mid and Long-term Changes after Recurrent Wheezing Episodes (Stages 1b and 2)

For Stages 1b and 2, 260 throat swabs were obtained and underwent DNA extraction. The 16S rRNA amplification was successful for 244 samples. Using Multiplex Identifier (MID) adaptors and six sequencing runs an initial 792,000 raw sequences were obtained for the 244 samples. After denoising, singleton exclusion, chimera checking and removal of OTUs present in only one sample a total of 478,988 sequences remained assigned to 455 OTUs (at 97% sequence identity level). The number of sequences obtained varied between

samples from 575 to 4,395. Consequently to remove the sample heterogeneity sequences were rarefied to 575 sequences per sample.

4.2.2.1 OTUs Comparisons

Cladistic classification of all the sequences (cases and controls combined) showed a high prevalence of the phylum Firmicutes (76% of the total number of sequences obtained) followed by Bacteroidetes (9%), Proteobacteria (7%), Actinobacteria (6%) and Fusobacteria (2%) respectively. Firmicutes was the most diverse phylum containing 113 distinct OTUs (51% of the total OTUs) followed by Proteobacteria with 37 OTUs (17%), Bacteroidetes with 36 OTUs (16%), Actinobacteria with 23 OTUs (10%), and Fusobacteria with 14 OTUs (6%).

At the OTU level, *Streptococcus* spp. were found to be the most common with more than half of the sequences identified assigned to this genus (55.3% of the total). *Veillonella* spp. were the second most common with 8.3% of sequences followed by *Prevotella* spp. with 5.5%, *Granulicatella* spp. with 3.8%, *Actinomyces* spp. with 3.6%, *Neisseria* spp. with 3.3%, and *Gemella* spp. with 2.5% of sequences.

At Phyla level when comparing all cases versus all controls, Actinobacteria and Proteobacteria were slightly more prevalent in cases but these differences were not significant. In contrast Firmicutes, Fusobacteria and Bacteroidetes were more prevalent in controls although only for Bacteroidetes was this statistically significant (*P*=0.017, OR 0.73) (Figure 4.5).

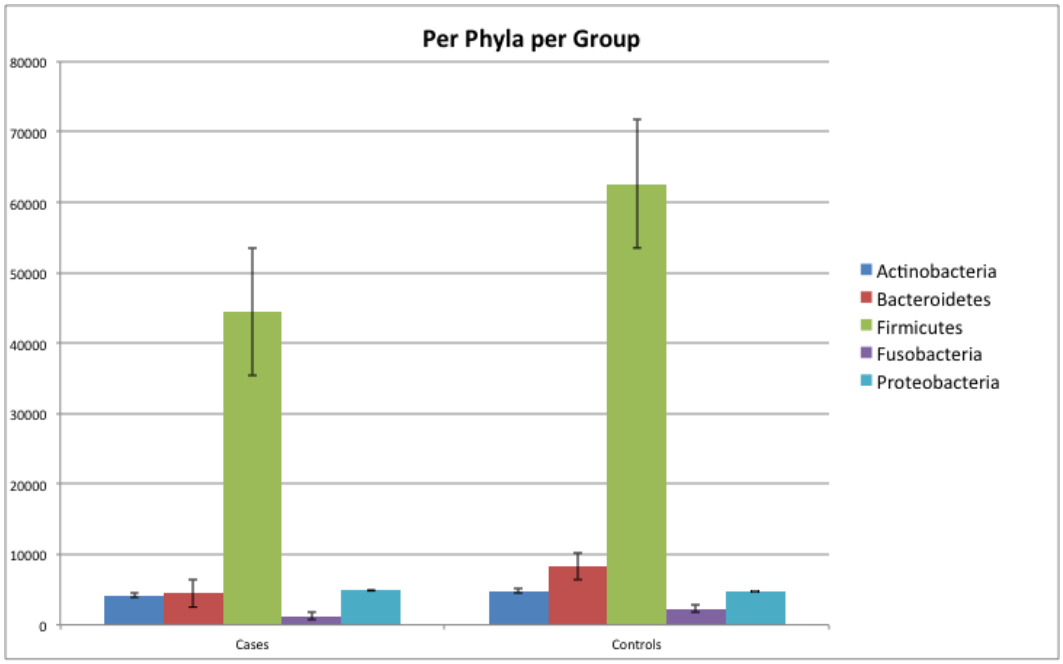


Figure 4.5. Comparisons at Phyla level between Cases versus Controls. Each bar represents the OTUs taxonomy assignation summarized per phylum. Error bars are also shown.

Taking age and case control status into consideration, analysis at the phyla level revealed statistically significant differences using two-way ANOVA (*P*<0.0001) and when multiple comparisons were performed (Tucke´s multiple comparison test). Statistical differences were seen between cases versus controls at 7 months of age for Firmicutes (*P*<0.0001), as well as for cases at 7 months compared with controls at 12 and 24 months (*P*<0.001), with more Firmicutes present in wheezing infants at 7 months compared with the other

groups. In addition at 7 months, an increase in Bacteroidetes in controls was observed when compared to cases (P<0.001) (Figure 4.6).

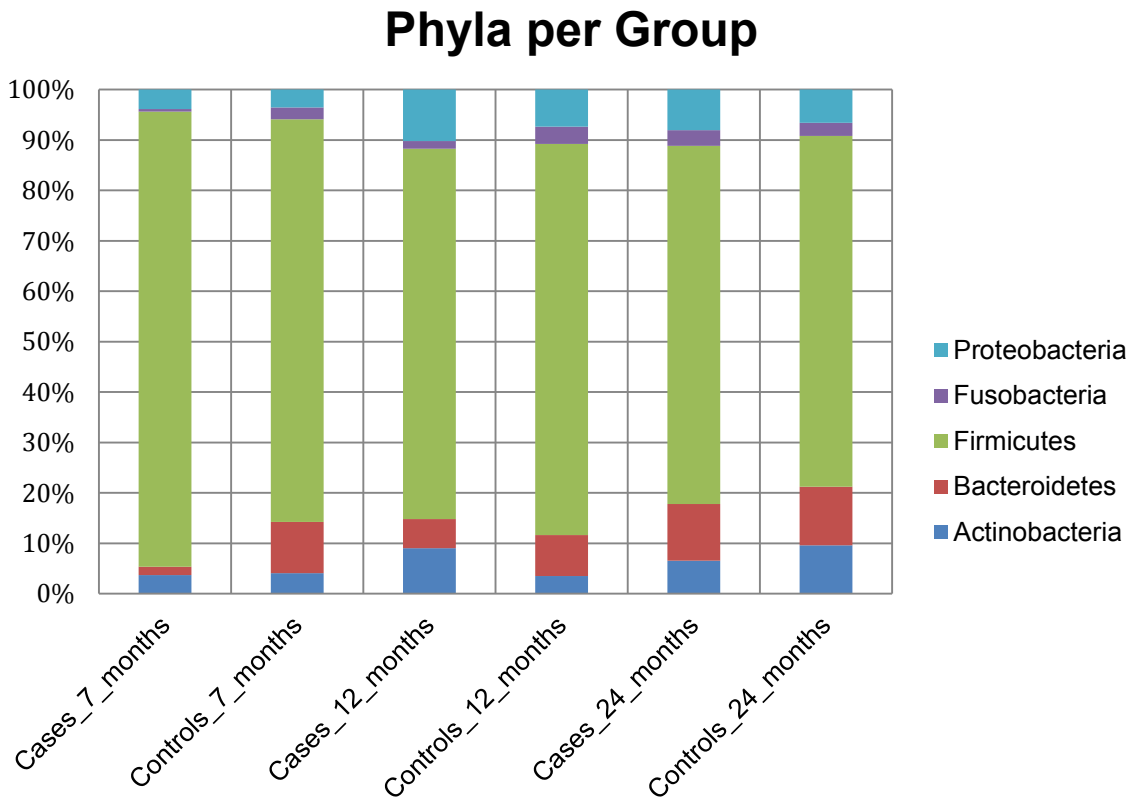


Figure 4.6. Comparisons at Phyla level per Group. Each bar represents the OTUs taxonomy assignation summarized per phylum percentage.

At the OTU level, there were several differences found when comparing between the pooled samples i.e. all cases versus all controls irrespective of age, and additionally when comparisons per status as well as age were examined (Table 4.1 – note only the OTUs with more than 100 sequences in total are shown).

OTUs	7 months	12 months	24 months	Total		
	P value	*P* value	*P* value	*P* value	OR	CI (95%)
Actinobacteria_Actinomyces	NS	0.0029	0.0010	0.0330	0.9021	0.8556 to 0.9511
Actinobacteria_Corynebacterium	NS	0.0006	0.0000	0.0049	3.5149	3.1852 to 3.8787
Actinobacteria_Propionibacterium	NS	NS	0.0021	0.0216	0.8907	0.819 to 0.9747
Bacteroidetes_Bacteroidales	0.0064	NS	0.0005	<0.0001	1.8294	1.4235 to 2.351
Bacteroidetes_Prevotellaceae	0.0090	0.0233	0.0002	<0.0001	0.7479	0.6048 to 0.9249
Bacteroidetes_Prevotella	0.0010	0.0021	0.0131	0.0001	0.6659	0.6367 to 0.6964
Bacteroidetes_Capnocytophaga	NS	<0.0001	0.0400	<0.0001	1.9237	1.5432 to 2.3981
Firmicutes_Lactobacillus	<0.0001	0.0013	<0.0001	<0.0001	0.7051	0.5665 to 0.8775
Firmicutes_Streptococcus	0.0010	NS	NS	0.0020	0.9719	0.9514 to 0.9928
Firmicutes_Streptococcus	0.0449	0.0450	0.0005	<0.0001	0.8919	0.8715 to 0.9128
Firmicutes_Gemella	NS	NS	0.0090	0.0012	1.6980	1.5916 to 1.8115
Firmicutes_"Lachnospiraceae"	0.0277	0.0000	NS	0.0270	1.3966	1.2579 to 1.5505
Firmicutes_Veillonella	0.0100	0.0240	0.0179	0.0040	0.6808	0.6547 to 0.708
Fusobacteria_Fusobacterium	NS	0.0000	0.0456	0.0056	0.7615	0.6252 to 0.9275
Fusobacteria_Leptotrichia	0.0010	0.0006	NS	0.0003	0.7005	0.6457 to 0.7599
Proteobacteria_Burkholderia	NS	0.0088	0.0131	<0.0001	2.1818	1.5815 to 3.0099
Proteobacteria_Neisseriaceae	0.0175	0.0119	<0.0001	0.0166	0.8317	0.6088 to 1.1362
Proteobacteria_Pasteurellacea	NS	0.0011	NS	0.0002	0.9986	0.9434 to 1.0571
Proteobacteria_Haemophilus	0.0100	0.0300	0.0430	0.0400	2.1295	1.896 to 2.3918
Proteobacteria_Moraxella	0.0010	0.0440	NS	0.0487	3.3673	2.9784 to 3.8069

Table 4.1. Summary of Statistical Analysis at OTU level. Only OTUs with greater than 100 sequences are shown. *P values* obtained on Metastats with 1,000 iterations were calculated comparing cases versus controls. The comparisons were performed per age group and in total. Additionally Odds Ratio and Confidence Intervals at 95% were calculated for the total analysis. • indicates the results when the OR was less than 1 (protective), and • when the OR is higher than 1 (increased risk).

For Actinobacteria, one single OTU (*Corynebacterium*) was more common in cases whilst two OTUs (*Actinomyces* and *Propionibacterium*) were more common in controls. The OTU assigned to *Actinomyces* sp. was most common in controls for the overall comparison (*P*=0.033, OR 0.902 95% CI: 0.86 - 0.95) and these differences were also observed at 12 months (*P*=0.0028) and at 24 months (*P*=0.0009). In the same phylum *Propionibacterium* was more prevalent in controls (*P*=0.0216, OR 0.890, CI=0.819 to 0.974), with the differences when age was considered only being significant at 24 months (*P*=0.0205). Both *Actinomyces* and *Propionibacterium* are commensal bacteria of the upper airways and rarely cause opportunistic infections. In contrast, *Corynebacterium* sp. (which has been associated with disease more commonly

but also could act as a commensal bacteria) was more prevalent in cases than in controls (*P*=0.0049, OR 3.514, CI=3.185 to 3.878). This finding was also seen at 12 months (*P*=0.0006) and at 24 months (*P*=5.2E-11).

For the phylum Bacteroidetes, two OTUs were more common in controls (Prevotellacea and *Prevotella*) and two in cases (Bacteroidales and *Capnocytophaga*). On overall, Prevotellacea and *Prevotella* sp. were more commonly found in controls (*P*=3.02E-05, OR 0.747 CI=0.60 to 0.92 and *P*=0.0001, OR 0.66, CI=0.63 to 0.69 respectively). Considering age, both OTUs were also significantly different for controls versus cases at 7 months (*P*=0.0089 and *P*=0.0009), 12 months (*P*=0.023 and *P*=0.002), and 24 months (*P*=0.0002 and *P*=0.013). *Prevotella* has been usually considered a protective commensal in the oral and upper airways. Within this phylum, Bacteroidales was more prevalent in cases (*P*=1.66E-05, OR=1.829, CI=1.423 to 2.351), particularly at 7 months (*P*=0.0066) and 24 months (*P*=0.0004). *Capnocytophaga* sp. was more common in cases (*P*=19.59E-10, OR=1.923, CI=1.543 to 2.398), and at 12 months (*P*=2.21E-06) and 24 months (*P*=0.039). Both Bacteroidales and *Capnocytophaga* are also considered commensals of the mouth but can produce in rare cases opportunistic infections.

For Firmicutes, two OTUs assigned to *Streptococcus* spp. were more common in controls compared to cases (*P*=0.0020, OR=0.971, CI=0.951 to 0.992 and *P*<0.0001), OR=0.891, CI=0.871 to 0.912). The first OTU was only statistically different at 7 months (*P*=0.0009), whilst the other was different for all ages: 7 months (*P*=0.0448), 12 months (*P*=0.0449) and 24 months (*P*=0.0005). In the same phylum, *Lactobacillus* sp. was also more common in controls (*P*=1.50E-11, OR=0.705, CI=0.566 to 0.877), similarly at 7 months

(*P*=1.93E-05), 12 months (*P*=0.0012) and 24 months (*P*=1.16E-11). *Veillonella* sp. was more prevalent in controls (*P*=0.004, OR=0.680, CI=0.654 to 0.708), and the differences were seen at 7 months (*P*=0.009), 12 months (*P*=0.023) and 24 months (*P*=0.017). *Gemella* sp. (*P*=0.0012, OR=1.698, CI=1.5916 to 1.811) and Lachnospiraceae (*P*=0.027, OR=1.396, CI=1.257 to 1.550) were more common in cases, and whilst Lachnospiraceae was significant at 7 months (*P*=0.027) and 12 months (*P*=4.2E-05), *Gemella* sp. was only statistically different at 24 months (*P*=0.008). All these bacterial generas are considered commensals of the upper airways usually associated with protective functions with the exception of *S.pneumoniae* that acts as a virulent pathogen in the airways of children.

For Fusobacteria, both OTUs (Table 4.1) were more prevalent in controls. *Fusobacterium* sp. (*P*=0.0056, OR=0.761, CI=0.625 to 0.927), and at 12 (*P*=1.94E-06) and 24 months (*P*=0.0456). *Leptotrichia* sp. (*P*=0.0003, OR=0.7005, CI=0.645 to 0.759) with differences at 7 months (*P*=0.0009) and 12 months (*P*=0.0005). Both *Fusobacterium* and *Leptotrichia* are commensals of the oral and airways microbiotas.

An OTU assigned to Neisseriaceaea, within the phylum Proteobacteria, was more common in controls (*P*=0.016, OR=0.83, CI=0.608 to 1.13). This OTU was also significantly different when compared at 7 months (*P*=0.017), 12 months (*P*=0.011) and 24 months (*P*=2.63E-08). Pasteurellaceae was more prevalent in controls (*P*=0.0002, OR=0.998, CI=0.608 to 1.13), this OTU presented differences at 7 months (*P*=0.017), 12 months (*P*=0.011) and 24 months (*P*=2.63E-08). Within this phylum *Burkholderia* sp., *Haemophilus* sp. and *Moraxella* sp. were more abundant in cases (*Burkholderia P*=1.78E-05,

OR=2.18, CI=1.58 to 3.00; *Haemophilus P*=0.040, OR=2.129, CI=1.896 to 2.39; and *Moraxella P*=0.048, OR=3.367, CI=2.97 to 3.80). For *Burkholderia* sp. the differences were also found at 12 months (*P*=0.0088) and 24 months (*P*=0.013), in *Haemophilus* sp. at 7 months (*P*=0.0099), 12 months (*P*=0.029) and 24 months (*P*=0.042), and *Moraxella* sp. at 7 months (*P*=0.0009) and 12 months (*P*=0.043). *Haemophiuls, Neisseria* and *Burkholderia* are considered as important pathogens of the airways, however in many circumstances are encountered being part of the ´commensal´ ecosystem in the upper airways.

The OTUs distribution of samples grouped by age and also case/control status is graphically represented (heatmap) in Figure 4.7. This highlights as previously commented that the most abundant OTUs present are *Streptococcus* spp. followed by *Prevotella* spp., *Veillonella* spp., *Neisseria* spp. and *Moraxella* spp. Although from the figure it is a little difficult to appreciate differences between cases and controls for each group (7, 12 and 24 months respectively), it is nonetheless possible to see how the diversity (by OTUs abundance) increases with age.

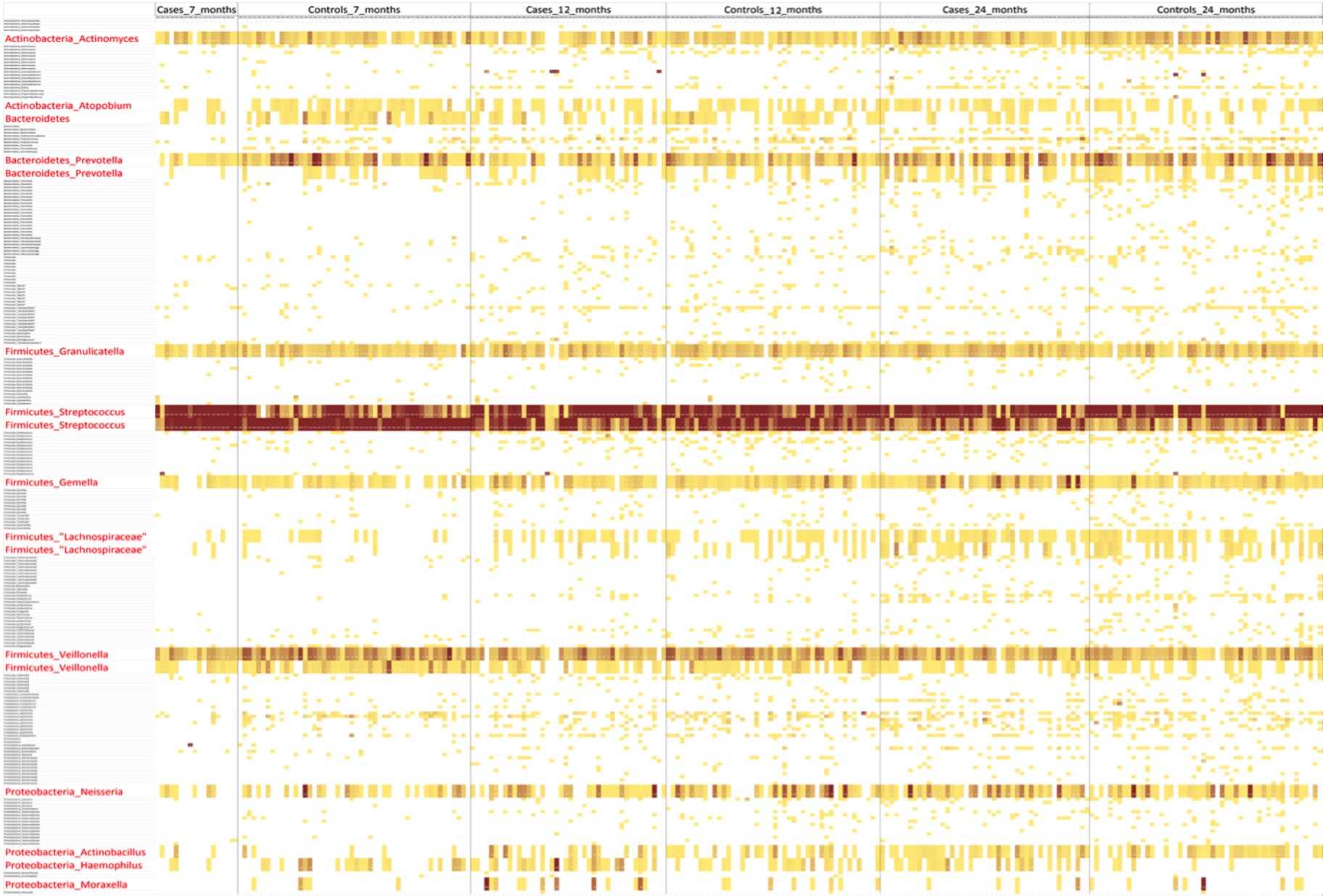


Figure 4.7. Heatmap that represents the OTUs abundance per sample. Each box ranges from white being the complete absence of sequences in that OTU, to yellow with a few and dark red with a high number of sequences. In red font are the most common OTUs found.

4.2.2.2 Comparisons of Cases versus Controls

4.2.2.2.1 Alpha Diversity

The alpha diversity analysis was performed calculating multiple rarefaction curves to determine Richness, Evenness and Diversity (using the Shannon index and the Inverse Simpson index) (Chapter 2: Section 2.7.5.2). To determine the normal distribution of the results, the Shapiro-Wilkinson normality test was performed. It showed that Richness (P<0.0001), Evenness (P=0.0012), and Inverse Simpson (P<0.0001) indexes were not normally distributed and therefore non-parametric statistics were performed. In contrast the Shannon index of diversity showed normal distribution (P=0.5718) and therefore parametric statistics were used (Figure 4.8).

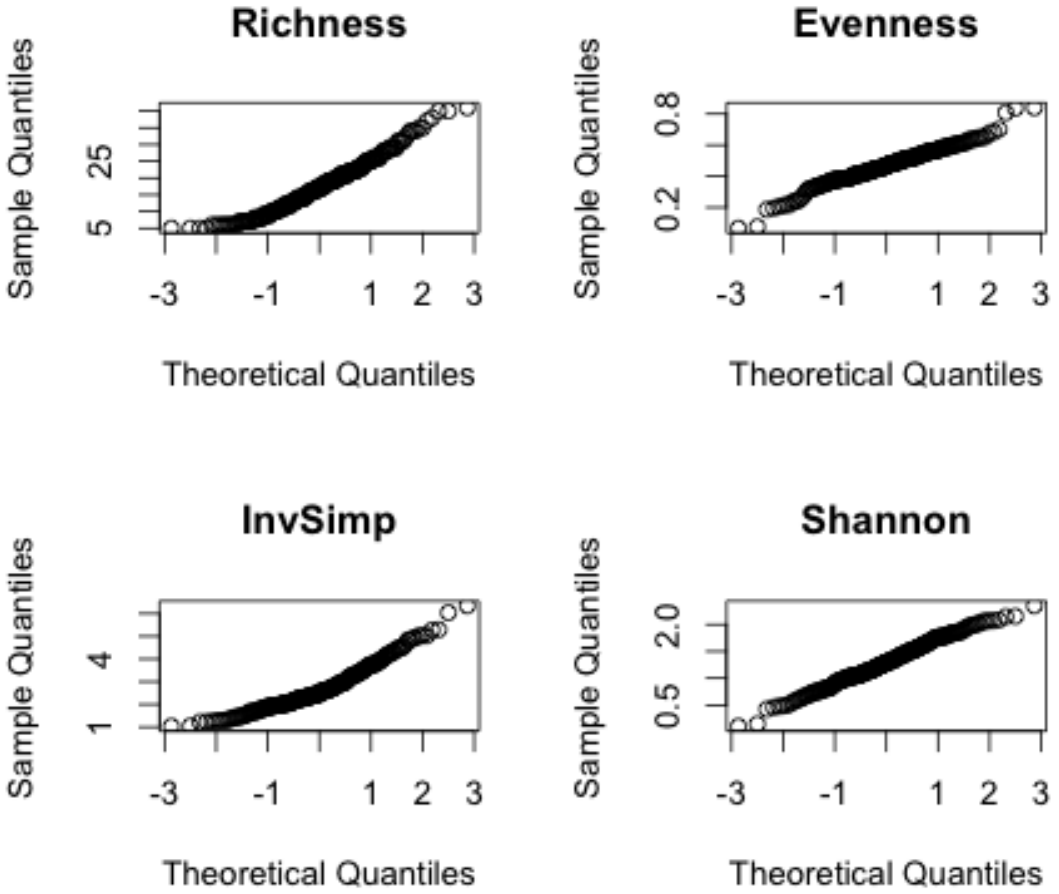


Figure 4.8. Alpha Diversity normality plots using Shapiro-Wilkinson test. Richness, Evenness, and Diversity with Inverse Simpson and Shannon indexes were calculated. Plots show the comparison of the theoretical versus sample quantiles to determine normality. Only Shannon index showed normal distribution.

There were encountered statistical differences on alpha diversity when case/control status was compared. Richness (Kruskal-Wallis *P*=<0.001), Inverse Simpson (Kruskal-Wallis *P*=0.04), and Shannon diversity index (ANOVA *P*<0.001) all confirmed there were significant differences. On the other hand Evenness was not statistically significant (Kruskal-Wallis *P*=0.11).

Comparison of alpha diversity plots using the Shannon diversity index revealed significant differences only at 7 months of age when comparisons between pairs of age groups were conducted (t-test *P*<0.001) (Figure 4.9). A trend in the same direction -although not significant- was seen for the 12 months age group comparison (cases versus controls).

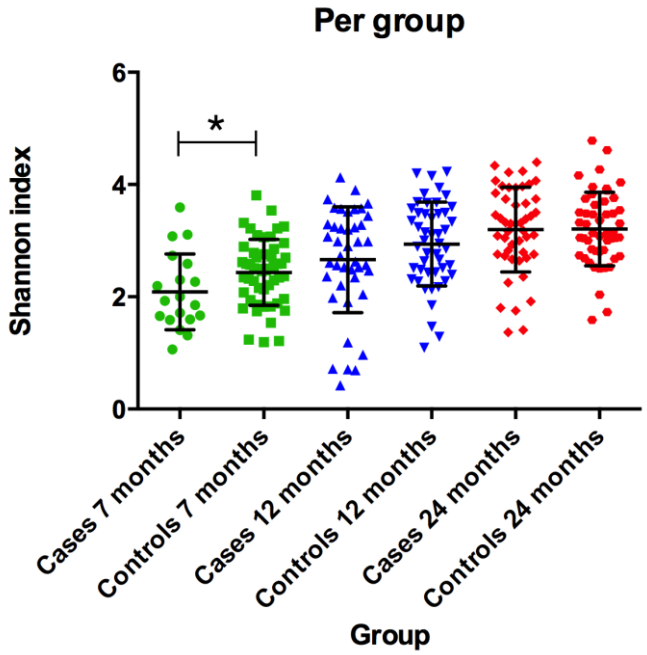


Figure 4.9 Per group Shannon Index Plot Analysis. * represents significant differences (*P*<0.001). Standard Deviation bars are shown.

When samples at different ages were pooled into cases versus controls, no significant differences for Richness (Kruskal-Wallis *P*=0.55), Evenness (Kruskal-Wallis *P*=0.3907), diversity in Inverse Simpson (Kruskal-Wallis *P*=0.1148) or diversity using Shannon index (ANOVA *P*=0.356 [Figure 4.10]) were seen.

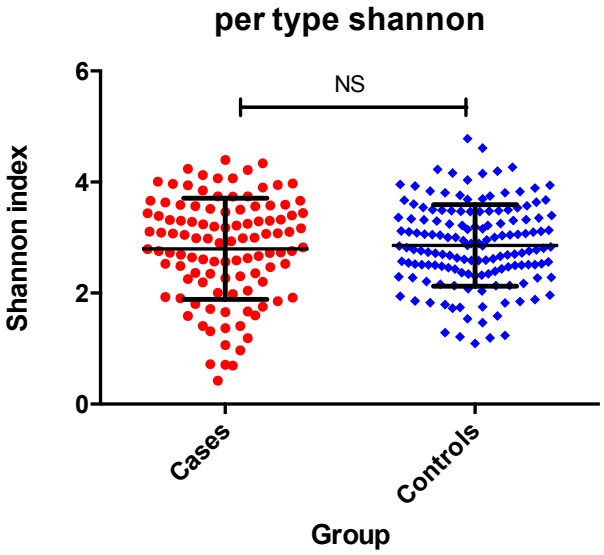


Figure 4.10. Alpha Diversity Shannon Index Comparison per Type of Sample: Cases versus controls. NS=non-significant differences. Standard Deviation (SD) bars are shown.

4.2.2.2.2 Beta Diversity

Comparing between groups at OTU level, statistical differences were found using Canberra (*P*<0.0001), Bray-Curtis (*P*<0.0001), UNIFRAC (*P*<0.0001), and Weighted UNIFRAC (*P*<0.0001). The type of group (cases versus controls) however was found to explain a different proportion of the variance each time, which means how belonging to a group (cases or controls) explains the microbiota differences: Canberra = 17%, Bray-Curtis 7%, UNIFRAC = 12%, weighted UNIFRAC = 8%. It can be concluded that the abundance of the different OTUs is not particularly important (UNIFRAC and Canberra are presence/absence and show higher values), and that phylogenetic information is not relevant to show differences (UNIFRAC versus Canberra).

On the NMDS plots, the stress (disagreement between 2-D configuration and predicted values from regression) is 0.17 in Bray-Curtis, 0.22 in Canberra,

0.23 in UNIFRAC and 0.10 Weighted-UNIFRAC, which denotes a very good representation in condensed dimensions, which means that 2D plots are a good representation of what happens during clustering. The NMDS bi-dimensional plots on the whole do not show very clear clustering when each group is compared separately. The exception is at 7 months when clustering is clearer in Canberra and UNIFRAC (Figure 4.11).

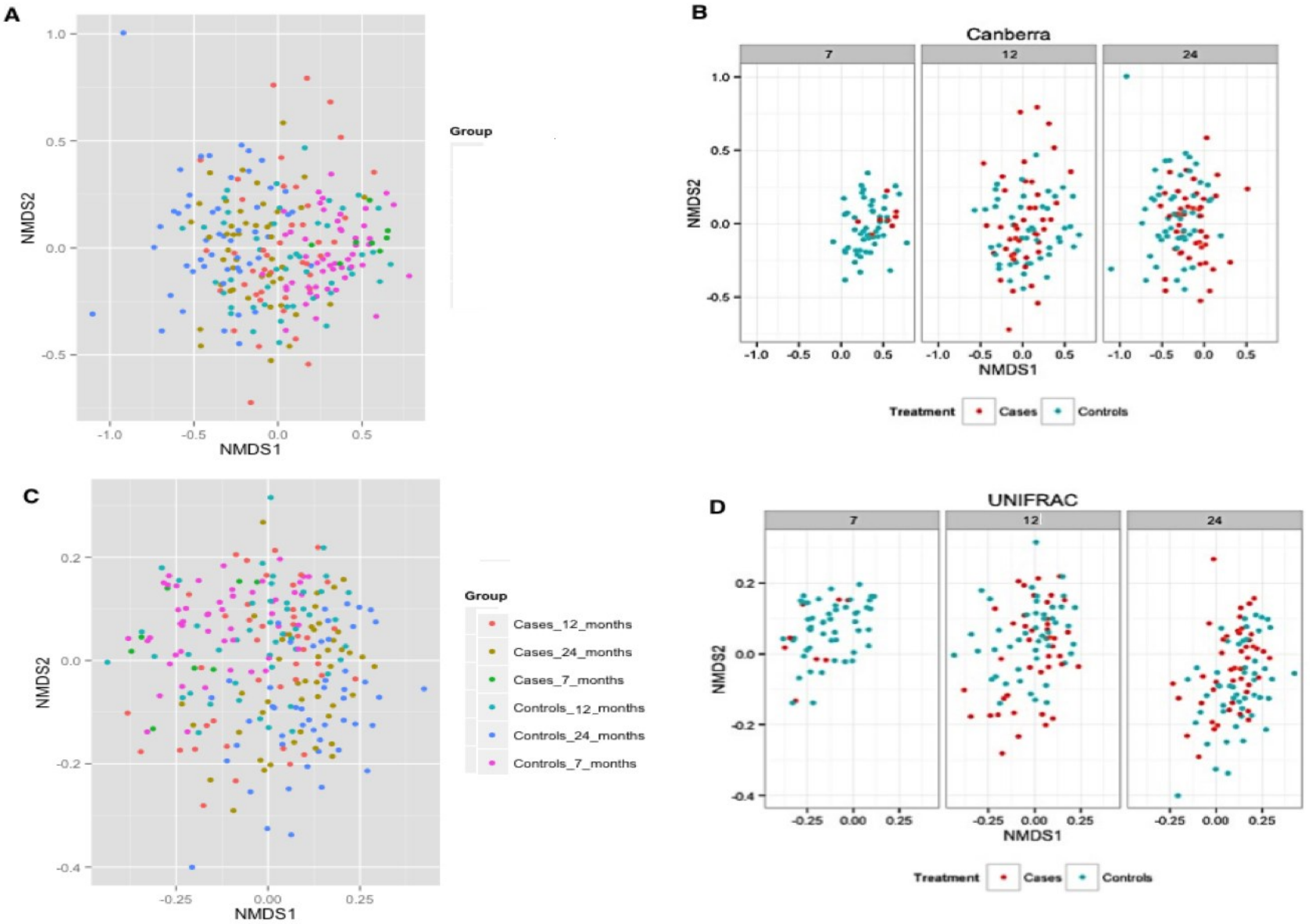


Figure 4.11. Beta Diversity NMDS2 Bi-dimensional Plots. Red dots represent cases and light blue dots represent controls. A) Canberra total, B) Canberra split by age C) UNIFRAC total D) UNIFRAC split by age.

By principal coordinate analysis (PCoA 3D plots), good clustering was observed when comparing cases versus controls at 7 months (Figure 4.12 A). Using Bi-plots at 7 months of age, demonstrated significant differences in clustering driven by a higher presence of *Streptococcus* spp. and *Prevotella* spp. in the controls group and a high prevalence of *Staphylococcus* spp. and *Haemophilus* spp. in cases were revealed (Figure 4.12 A).

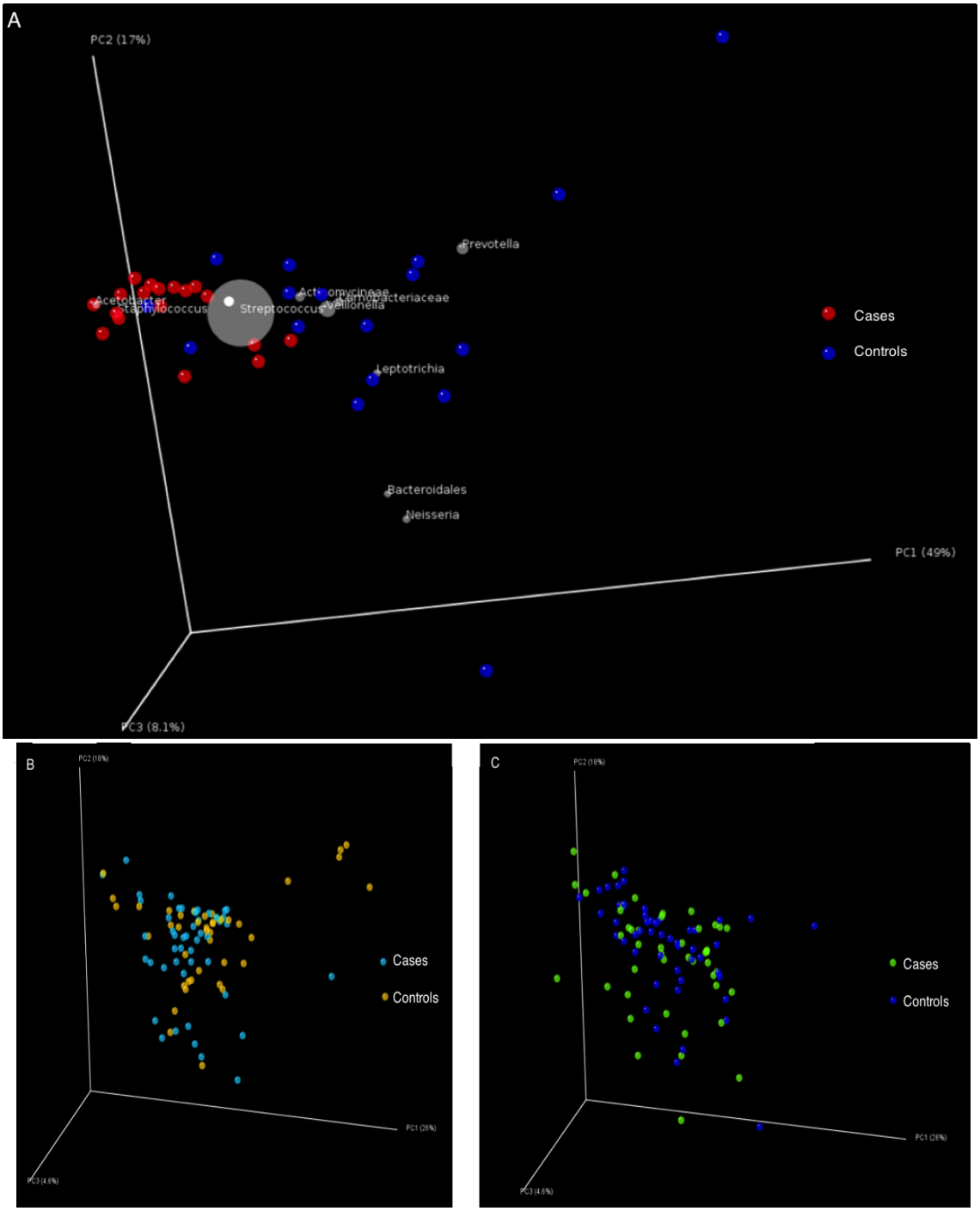


Figure 4.12. Beta Diversity: UNIFRAC PCoA analysis for each Age group Cases and Controls. Each dot represents a sample. A) 7 months PCoA with additional taxonomic information of the main genus that drives clustering B) 12 months PCoA C) 24 months PCoA.

Beta diversity analysis, with samples pooled into cases and control groups, showed no significant differences using Bray-Curtis (*P*=0.062) and Weighted UNIFRAC (*P*=0.076), whilst significant results were observed by Canberra (*P*=0.001) and UNIFRAC (*P*=0.024). The type of overall group (cases versus controls) however only explains on the variance in 7% on Canberra and 8% on UNIFRAC. The NMDS plots do not show any evidence of clustering on the bi-dimensional plots and in the PCoA tri-dimensional plots (Figure 4.13).

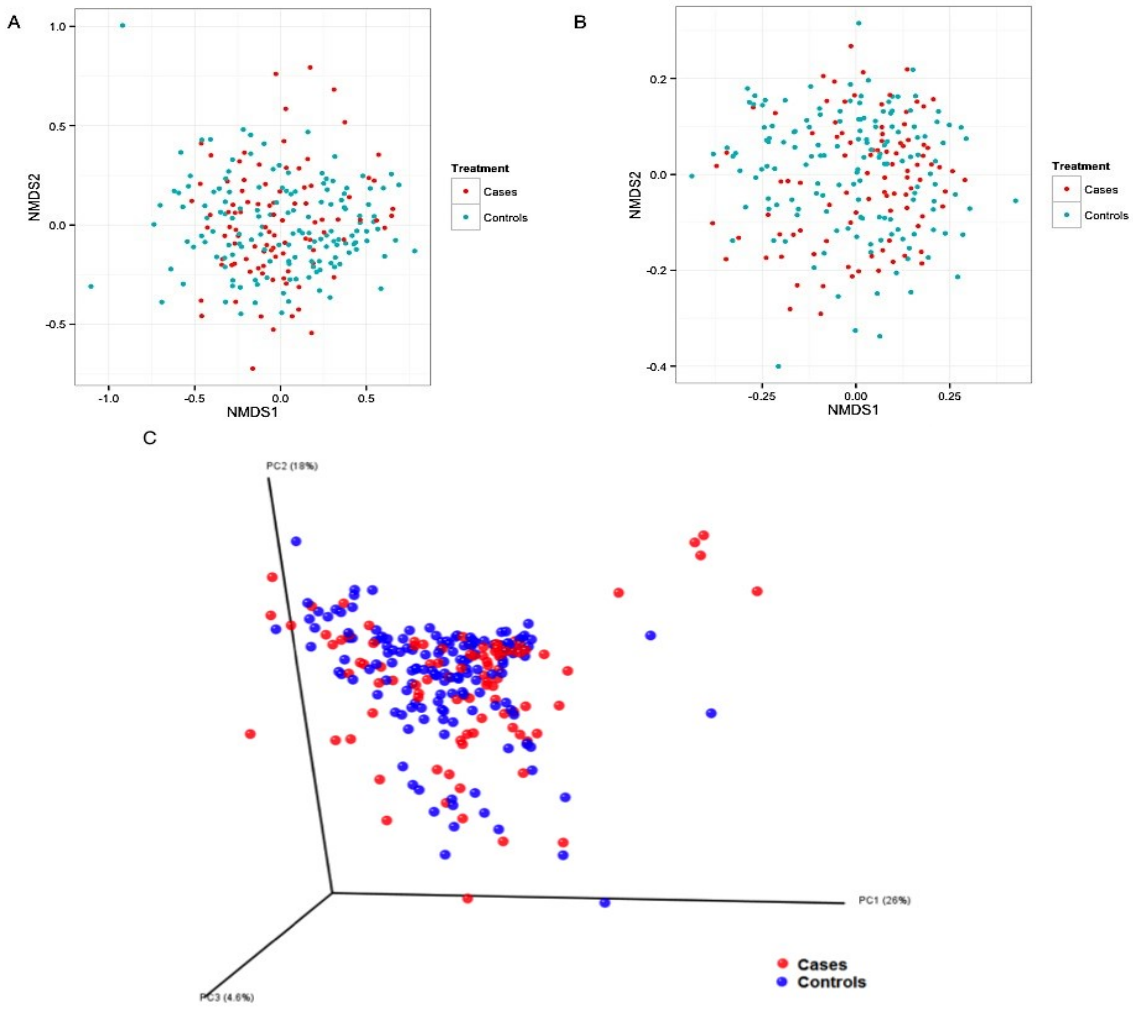


Figure 4.13. Beta Diversity Analysis between Cases and Controls. Each dot represents a sample with red being cases and blue controls. A) Canberra NMDS bi-dimensional plot, B) UNIFRAC NMDS bi-dimensional plot, C) PCoA UNIFRAC tri-dimensional plot.

4.2.2.3 Age Group Comparisons

In contrast with the previous analysis in Section 4.2.2.1, the following statistical comparisons were not performed using cases or controls grouping but only differences per age groups. This analysis was performed to see how the upper airways microbiota develops over time.

4.2.2.3.1 Alpha Diversity

Alpha diversity analysis for the per age group comparisons used the same values of normal distribution, because the richness, evenness and diversity indexes Shapiro-Wilkinson tests were calculated before grouping. Richness (Kruskal-Wallis *P*=2.2E-16), diversity using Inverse Simpson (Kruskal-Wallis *P*=0.023), and diversity using the Shannon index (ANOVA *P*<0.0001) were statistically different, whilst Evenness was not (Kruskal-Wallis *P*=0.549). Pairwise comparisons using *t tests* showed that between groups of age the Shannon diversity index and the raw observed species were statistically different with an increase in the number of species with age (Figure 4.14).

Figure 4.14. Alpha Diversity Plots for Age Group Comparison. A) Shannon index comparisons, B) Observed species comparisons. * Statistically significant on pairwise analysis. SD bars are shown.

4.2.2.3.2 Beta Diversity

Beta diversity analysis for the per age group comparisons revealed significant statistical differences using all four methods: Adonis Canberra (*P*<0.0001), Bray-Curtis (*P*<0.0001), UNIFRAC (*P*<0.0001), and Weighted UNIFRAC (*P*<0.0001). Still the type of age group (7 months, 12 months and 24 months) explains a different proportion of the variance each time: Canberra = 11%, Bray-Curtis 4%, UNIFRAC = 8%, weighted UNIFRAC = 5%. Consequently the variety of OTUs found is more important than abundance (UNIFRAC and Canberra are presence/absence and show higher values), and the phylogenetic information is not relevant to show such differences (UNIFRAC has lower values than Canberra).

On the NMDS plots, the stress found was similar than previously with good representation of clustering in them. The NMDS bi-dimensional plots show evidence of some clustering related with increase of age. When case control status is considered the clustering is more evident for the control group (Figure 4.15).

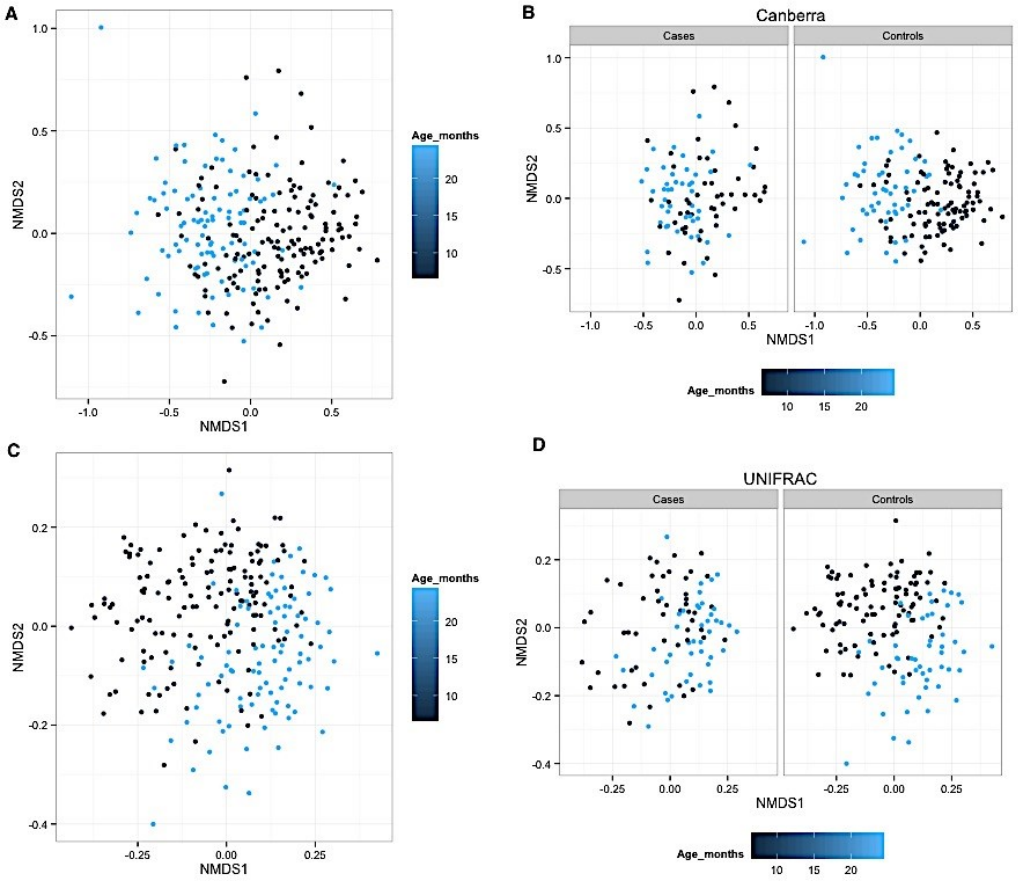


Figure 4.15. Beta Diversity: NMDS Bi-dimensional Plots comparing Groups by Age in months. Each dot represents a sample and are coloured on a scale according to age (lighter colour represents oldest) A) Canberra analysis on the total NMDS plot, B) Canberra analysis on plots separated by case control status, C) UNIFRAC analysis on the total NMDS plot, D) UNIFRAC analysis on plots split by case control status.

Principal coordinate analysis using UNIFRAC did not show a clear clustering when samples were grouped per age (Figure 4.16).

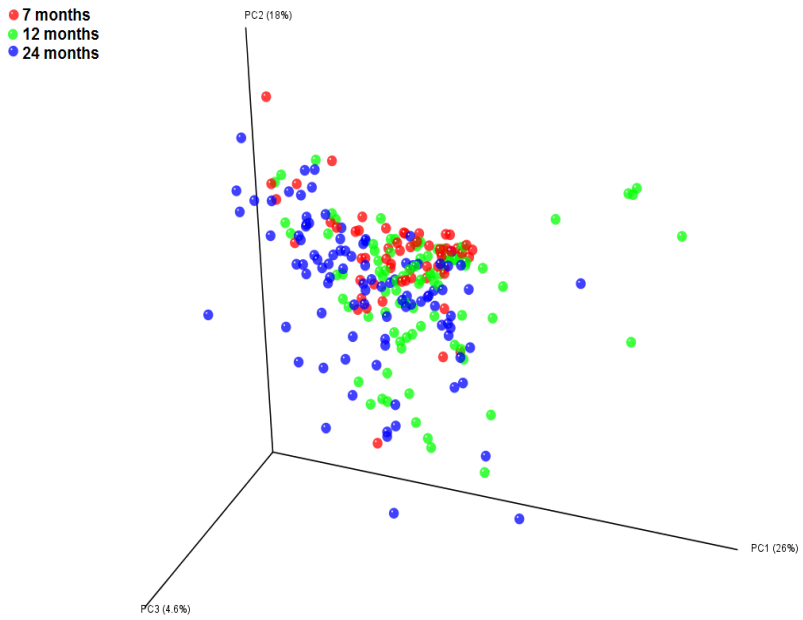


Figure 4.16. Beta Diversity PCoA using UNIFRAC. Each dot represents a sample and each colour represents an age group as indicated by the key.

4.2.2.4 Investigation of Sequencing Batch Effect

With all next generation sequencing platforms it is only possible to sequence a set number of samples per run and this number varies depending as well on depth of sequencing required. For the Roche 454 Junior platform it is only possible to sequence a maximum of 170,000 bacteria at the same time, consequently to achieve the desired depth in sequencing for around 400 samples six independent sequencing runs had to be conducted and the data then merged. For that reason it was important

to verify that the effect of merging multiple runs had not had any impact or underpinned the results in relation to OTU differences and diversity.

Metastats was therefore used to examine whether at OTU level significant differences in abundance where seen when sequencing runs were compared. No differences were found. Similarly no differences were seen when alpha and beta diversity were examined using Shannon index and in NMDS and PCoA plots (Figure 4.17).

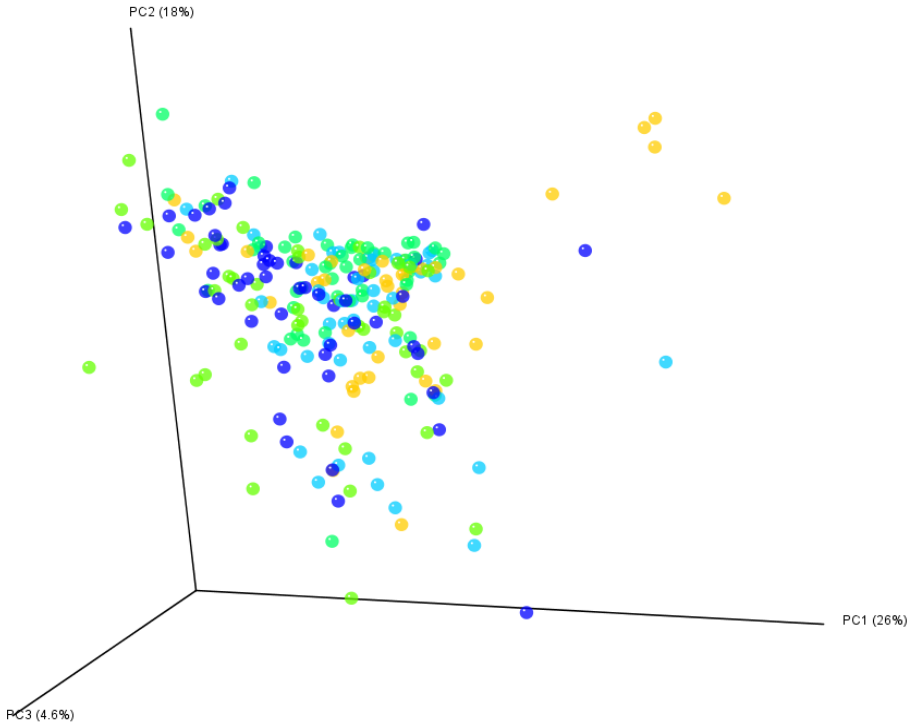


Figure 4.17. Beta Diversity PCoA between runs using UNIFRAC. Each dot represents a sample and each colour represents a sequencing run – six runs in total.

4.2.2.5 Epidemiological Analysis

To determine if the epidemiological variables were different between wheezing infants and healthy controls, several parameters were compared

using the questionnaires taken during each sampling occasion. Gender was not different between groups with 54% in the wheezing infants and 46% in the controls. The crowdedness was measured by the number of individuals per room of the house, and was not different between groups with 3.15 individuals/room in cases and 3.06 in controls. Both groups also presented a similar parental income with 222 USD in cases and a slightly higher 229 in controls (Table 4.2).

	Cases	**Controls**	***P*<0.05**
Sex (% male)	54	46	
Number of individuals per rooms of the house	3.15	3.06	
Average parental income (USD per month)	222	229	
Average birth weight (grams)	3390	3341	
Average birth height (cm)	49.6	49.1	
Maternal Education			
Illiterate	2%	2%	
Primary Incomplete	17%	13%	
Primary Complete	23%	22%	
Secondary Incomplete	15%	16%	
Secondary Complete	32%	44%	*
University Incomplete	10%	3%	*
University Complete	1%	0%	
Respiratory Tract Infections (episodes per person)	0.02	0.04	
Antibiotics (numbers of times used per person)	0.06	0.07	
Malnutrition (% reported)	2%	2%	
Anaemia (% reported)	1%	3%	

Table 4.2. Epidemiological Characteristics of Cases and Controls. * *P*<0.05

Beta diversity analysis was performed to determine whether gender, crowdedness and parental income, however none of these variables account for the differences observed in the microbiota between cases and controls (Figure 4.18). Canberra statistics (stress 0.10) were the best representation of the beta diversity in bi-dimensional plots.

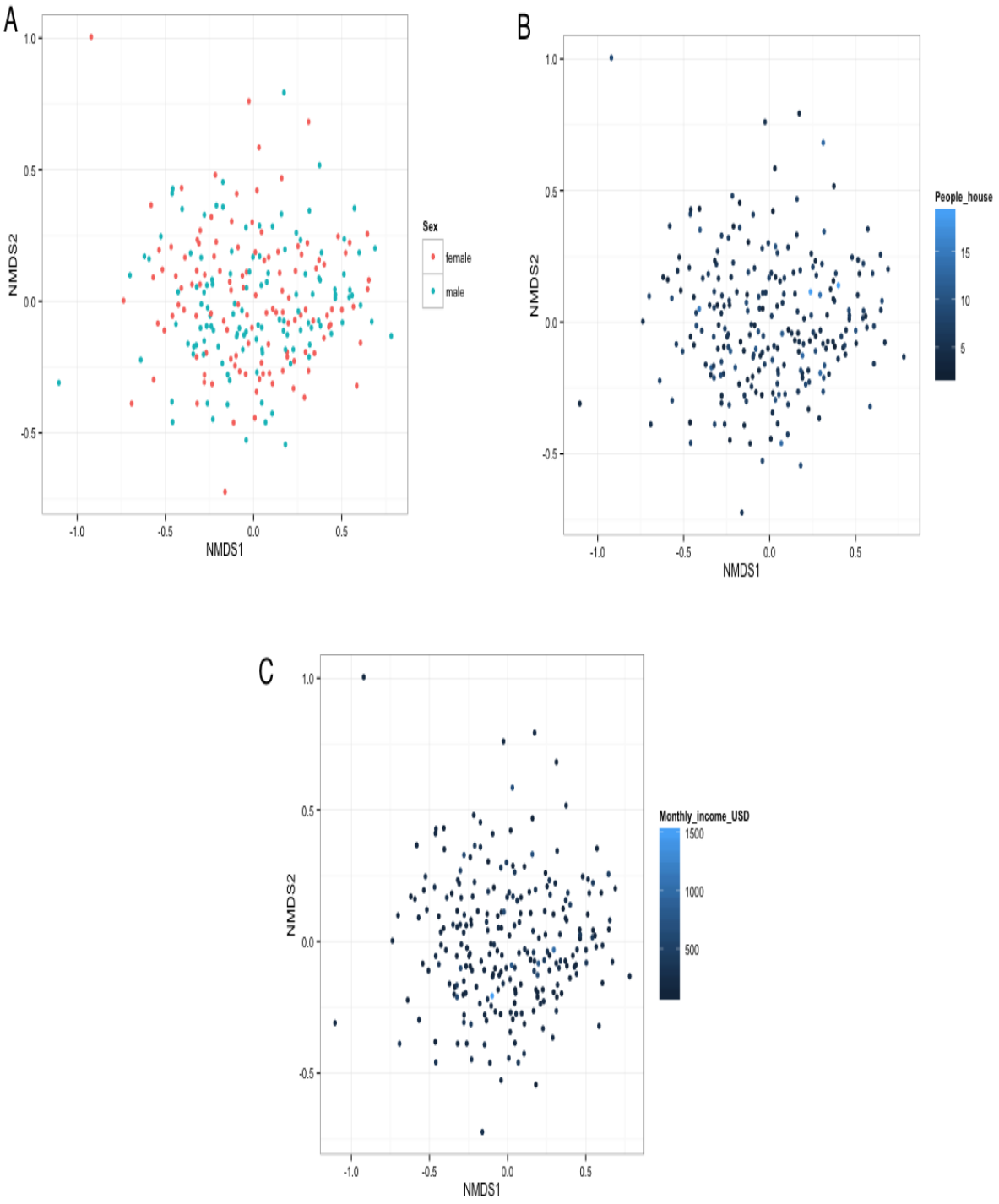


Figure 4.18. NMDS Beta Diversity Plots considering Epidemiological Variables: parental income, gender, people per room. Canberra statistics were used to determine the influence of epidemiological variables on the microbiota changes. Each dot represents a sample. A) Gender, B) Individuals per room per house, C) parental income in USD.

Biometrical measures at birth such as weight and height were also not different between groups. The average birth weight in grams for the

cases was 3390 whilst in the controls 3341, and the average height in centimetres for the cases was 49.6 whilst for controls it was 49.1. These variables also were not able to explain the microbiota patterns present at 7, 12 and 24 months of age in cases or controls (Figure 4.19).

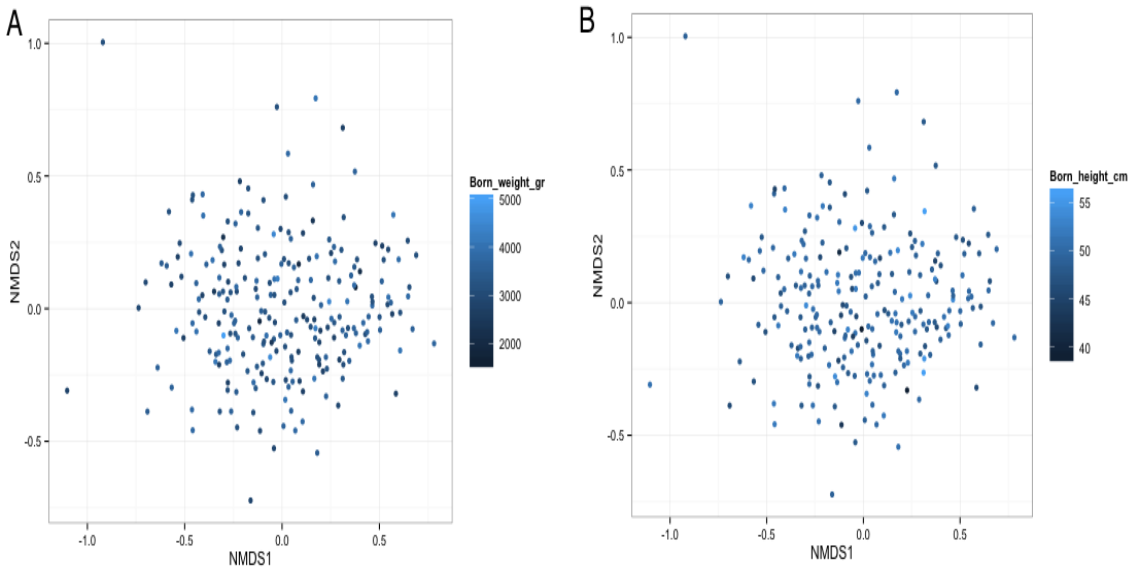


Figure 4.19. NMDS Beta Diversity Plots considering Epidemiological Variables: birth weight and height. Canberra statistics was used to determine the influence of epidemiological variables on the microbiota changes. Each dot represents a sample. A) birth weight B) birth height.

Maternal education was not different in the illiterate, primary incomplete, primary complete and university complete groups between cases and controls. However there were more mothers in the controls group that have the secondary studies complete compared with cases (44% versus 32%), and more mothers with university incomplete in the cases group compared with controls. Overall maternal education did not explain the microbiota patterns present in the samples (Figure 4.20).

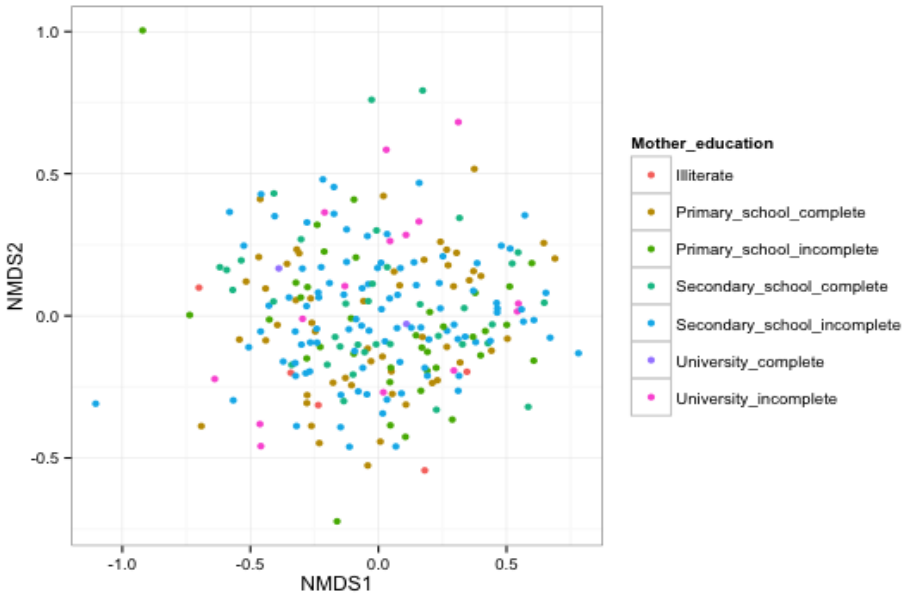


Figure 4.20. NMDS Beta Diversity Plots considering Epidemiological Variable: maternal education. Canberra statistics were used to determine the influence of maternal education on the microbiota patterns.

The infants included on the study did not have a high number of respiratory tract infections (pharyngitis, pneumonia, bronchitis, etc.), or a history of high antibiotic usage. The respiratory tract episodes in the 2 years follow up per person were only 0.02 for the cases group whilst for the controls it was 0.04, the use of antibiotics per any reason in the clinical records per infant was 0.06 times use in cases and 0.07 in controls. All the infants in the study had no history of oral or inhaled corticosteroid use.

Malnutrition and anaemia were present in the clinical records of a few infants in both groups, but were not present at the time of sampling. However these two conditions were not significantly different between cases and controls. Malnutrition was found in 2% of infants in both groups and anaemia 1% in cases and 2% in controls. Neither variable explained the microbiota patterns found on beta diversity analysis using Canberra (Figure 4.21).

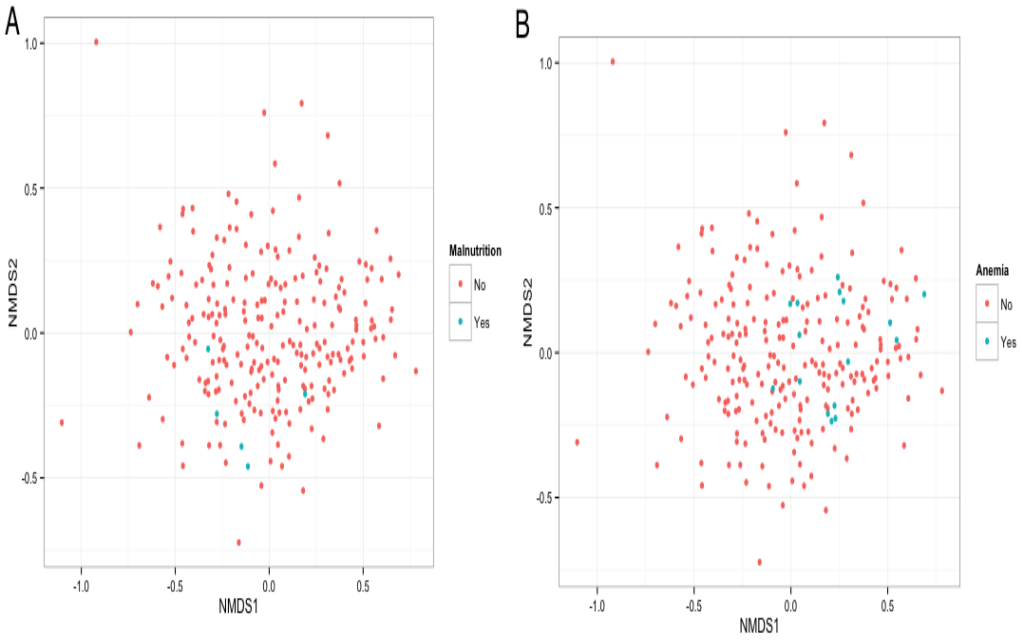


Figure 4.21. NMDS Beta Diversity Plots considering Epidemiological Variables (Malnutrition and anaemia). Canberra statistics was used to determine the influence of epidemiological variables on the microbiota patterns. A) malnutrition, B) anaemia.

4.3 Discussion

There were encountered several differences in the upper airways microbiota of infants with a current wheezing episode (Stage 1a) versus infants at the same age with history of wheezing but that at the sampling time did not have an episode (Stage 1b). When comparing the overall microbiota patterns between current and past wheezers the main features of bacterial diversity and abundance were conserved. Firmicutes was the most common phylum in both groups and streptococci was the most abundant genus as reported in previous chapters and in other studies (Hilty *et al.*, 2010, Bisgaard *et al.*, 2007).

The other commonly found bacteria i.e. *Prevotella* and *Veillonella* were also present in both groups equivalently when compare at genus level, however some OTUs of them were more common in the infants with history

of past wheezing. For instance *Prevotella* spp. (not possible to assign taxonomy at species level) were more common in past wheezers but *Prevotella melanogenica* was more common in current wheezers. Interestingly *Prevotella* spp. (non-melaninogenica) have been encountered as possible protective bacteria against the development of asthma in adults and children (Hilty *et al.*, 2010), and were consistently more abundant in healthy infants as shown in the previous chapter. On the other hand, *P. melaninogenica*, that in this part of the study is present very abundantly in current wheezers but very rarely in past wheezers, has been associated with the development of otitis media, oral plaque, and tonsillitis (Jensen *et al.*, 2013, Saini *et al.*, 2005, Tanaka *et al.*, 2008). *Veillonella* spp. the other reported bacteria present mostly in healthy infants in previous chapters and in conventional culture-dependant studies (Bisgaard *et al.*, 2007), is present more commonly in past wheezers as well. Its abundance however is not comparable with the higher proportions of these bacteria in healthy infants.

In the Actinobacteria phylum the genera *Actinomyces* and *Corynebacterium* was more common in current wheezers, and moreover these bacteria have been commonly reported as causes of upper respiratory tract infections (Celebi *et al.*, 2009, Byard, 2013).

Other usually reported 'protective' bacteria like *Lactobacillus* (Messaoudi *et al.*, 2013), were encountered more in past wheezers compared with current wheezers. Its abundance however did not reach the higher values found in healthy controls with no history of wheezing.

Strikingly *Staphyloccoccus*, that is a well-known pathogen (Vissing *et al.*, 2013, Stokken *et al.*, 2014, Wolter *et al.*, 2013), was also more abundant

in the current wheezing infants compared with the past wheezers, and almost absent in healthy infants. Other Firmicutes present regularly in the upper airways and in the oral normal microbiota such as Gemmellacea and *Leptotrichia* (Hilty *et al.*, 2010, Wolff *et al.*, 2013), were more common in past wheezers.

The phylum Proteobacteria has been reported to be usually more frequent in asthmatics (Hilty *et al.*, 2010, Bisgaard *et al.*, 2007), and several potentially pathogenic bacteria that belong to this phylum were found more commonly in current wheezers. *Neisseria, Haemophilus* and *Moraxella* were significantly more frequent in current wheezers and this higher prevalence compared with controls was maintained over time (Chapter 3).

Beta diversity analysis revealed statistical differences when the current and past wheezing infants were compared, with the presence/absence of particular OTUs determining the differences. The sequences abundance did not explained the main differences. It therefore would appear that purely the presence of new bacteria is related with a wheezing episode as opposed to the overgrowth of the bacteria already present in a healthy individual. For that reason, Canberra statistics were the best way to represent the clustering on NMDS plots.

In terms of alpha diversity no differences were observed between current and past wheezers in richness and diversity, however evenness was statistically different. This means that the number of taxa was not different between the groups but the distribution of taxa encountered was dissimilar. These results show that particular OTUs appear temporarily during a wheezing episode but the core microbiota remains unaltered.

Stages 1b and 2 of the ECUAVIDA were analysed in order to determine the mid and long term changes in microbiota after a history of wheezing. The current part of the study (Section 4.2.2) has resulted in the identification of significant differences in the upper airways microbiota at different ages of development between wheezing and healthy infants.

As in previous studies the upper airways microbiota was found to be represented principally by Firmicutes, with the most common genus being *Streptococcus* spp. (Hilty *et al.*, 2010). In addition, common to prior publications, Firmicutes was found to be in this study the most diverse phylum of the airways with the highest number of OTUs (Hilty *et al.*, 2010, Marsland, 2013, Molyneaux *et al.*, 2013, Marsland *et al.*, 2013). Proteobacteria, Bacteroidetes and Actinobacteria were also prevalent in the oropharynx microbiota in both non-wheezing and wheezing infants again agreeing with early observations (Hilty *et al.*, 2010, Zemanick *et al.*, 2011).

The most common pathogens that have been found in the airways are Gram-negatives belonging to the phyla Actinobacteria and Proteobacteria (Watt *et al.*, 2009, Sabra and Benger, 2011, Gkentzi *et al.*, 2012, Celebi *et al.*, 2009). Interestingly in this current study these pathogens were found to be more prevalent in infants with a history of wheezing. Additionally, the previously reported protective bacteria Firmicutes and Bacteroidetes (Hilty *et al.*, 2010, van der Gast *et al.*, 2014, Huang *et al.*, 2010) were found to be more common in the control infants.

When the samples were separated by case/control status, significant differences were seen at the phyla level. At this taxonomic level the most remarkable dissimilarities were encountered between Firmicutes for the 7

months group of cases compared with the controls for all three age groups (7 months group but also 12 and 24 months groups). For instance, cases at 7 months show an unusual high abundance of Firmicutes with lower abundances seen for the other phyla. These differences were not attributable to particular Firmicutes genera but as will be discussed further are due to a decreased diversity of other phyla in wheezing infants at 7 months of age.

Resembling the results seen in Stage 1a (Chapter 3: Section 3.2.2.2), when compared at genus level in infants with wheezing history there were more potential pathogenic bacteria (e.g. Corynebacterium, *Haemophilus, Moraxella* and *Burkholderia*) present and this was statistically significant. Furthermore, a lack of some of the reported 'protective' bacteria species (e.g. *Veillonella*, *Prevotella,* and *Lactobacillus*) in wheezing infants compared with controls was observed (Chapter 3: Section 3.2.2.2 and Section 4.2.2 above). These concordances reveal that changes during an episode of wheezing remained stable both medium and long term during infancy, and the alterations in the normal microbiota may in turn be of importance in relation to host resistance to pathogen colonization in the airways (Crowe *et al.*, 1973).

At the OTU level, *Corynebacterium* spp. were more common in infants with history of wheezing, and this finding was evident at 12 and 24 months of age. This bacterial genus includes many bacteria that are part of the normal oral (Jiang *et al.*, 2014) and nasopharynx microbiota (Biesbroek *et al.*, 2014). The genus also includes some extremely pathogenic bacteria that cause high rates of morbidity e.g. *C. diphteriae* (Byard, 2013). The

Actinomyces and *Propionibacterium* genera showed an increased abundance in controls. Although they can cause opportunistic infections (Celebi *et al.*, 2009, Kutluhan *et al.*, 2011) they are common constituents of the healthy oral and airways microbiota (Francavilla *et al.*, 2014, Hilty *et al.*, 2010).

Bacteroidales is a common phylum of the upper airways, oral and nasopharyngeal microbiota. Several of their genera have been implicated for causing opportunistic infections in immunocompetent and immunodepressed patients (Piau *et al.*, 2013, Papadaki *et al.*, 2008, Mahenthiralingam, 2014, Sherrard *et al.*, 2014, Rath *et al.*, 2014). One of the genera within Bacteroidales is *Prevotella* which has been considered protective against the development of asthma in adults (Hilty *et al.*, 2010). Reinforcing this premise, the present study found that *Prevotella* and other Prevotellacea were significantly more prevalent in controls (at all ages and in total) compared with infants with a history of wheezing.

Firmicutes, being the most diverse phylum in the airways microbiota, presented several genera that were more abundant in infants with history of wheezing. The most common genus found in the infants' upper airways microbiota was *Streptococcus,* with two OTUs seen more frequently in controls than in infants with history of wheezing. One of them presented significant differences at all ages whilst the other only at 7 months. *Streptococcus* spp. can be considered as commensal bacteria part of the "normal" microbiota (Hilty *et al.*, 2010, Francavilla *et al.*, 2014) but members of the genus, e.g. pneumococcus, do cause life-threatening infections in childhood (Harmes *et al.*, 2013, Gil *et al.*, 2013). Due to the uniformity of the

16S rRNA gene within *Streptococcus* spp. it was not possible to differentiate at the species level. Consequently although the OTUs found are related with protection against wheezing further discrimination at species level is clearly desirable (see Chapter 5).

Strikingly an OTU assigned to the well-recognized “protector” of the microbiota *Lactobacillus* was more prevalent in controls in the total analysis as well as at all ages. *Lactobacillus* spp. are associated with the protection of the microbiota against the development of pathogenic bacteria due to their fermentation properties which acidify the microbial environment and inhibit the growth of non-acidophilic bacteria (Messaoudi *et al.*, 2013, Nardis *et al.*, 2013, Walker, 2013). *Veillonella* spp. are Gram-positive anaerobic bacteria with interesting metabolic pathways that can cause them to appear to be Gram-negatives (Vesth *et al.*, 2013). They are commonly present in saliva of healthy people (Belstrom *et al.*, 2014), although they are usually considered opportunistic pathogens in patients with chronic pulmonary disease (Rogers *et al.*, 2014, Abel Nour *et al.*, 2013). In this present study they were found to be more abundant in controls with differences seen for all three time points - 7, 12 and 24 months.

Gemella and Lachnospiraceae are common bacteria of the oral and gut microbiota but have also been associated with opportunistic infections (Iwai *et al.*, 2014, Garcia-Lechuz *et al.*, 2002). They have been shown to be important in the early colonization in infants’ mucosal barriers depending on the delivery method (Hansen *et al.*, 2014). Anaerobic bacteria of the genus *Clostridium,* which are important pathogens in the human (Gabriel and Beriot-Mathiot, 2014, Li *et al.*, 2013), are members of the family

Lachnospiraceae. The genus *Clostridium* are potent triggers of the immune system notably chemotaxis of macrophages (Nakanishi *et al.*, 2014). In the present study *Gemella* and Lachnospiraceae were found to be at a higher abundance in infants with history of wheezing although not at all ages. For *Gemella* differences were seen at 24 months only whilst for Lachnospiraceae the differences were at 7 and 12 months.

Within the phylum Fusobacteria there are bacteria commonly found in the oral microbiota which form biofilms like *Fusobacterium* and *Leptotrichia* (Wolff *et al.*, 2013) and that are only associated with infections in immune-compromised patients (Eribe and Olsen, 2008). Interestingly *Fusobacterium* has been implicated to have a role in chronic inflammation of the gut and inflammatory bowel disease (IBD) and colorectal carcinoma (Allen-Vercoe and Jobin, 2014). Stages 1a and 2 of the ECUAVIDA microbiota study revealed the phylum to be present more commonly in healthy infants compared with infants with a history of wheezing although the significance of this was somewhat variable with age (*Fusobacterium* at 12 and 24 months, *Leptotrichia* at 7 and 12 months).

Members of the phylum Proteobacteria include a large variety of Gram-negative bacteria that previously have been shown to be related with the development of disease in children. Members include: *Haemophilus influenza, Neisseria meningitidis*, *Moraxella catarrhalis* and *Burkholderia* spp. In this current study *Burkholderia, Haemophilus* and *Moraxella* were found (Section 4.2.2) to be more abundant in wheezing infants; particularly *Haemophilus*, which at all ages, was related with disease. This finding is in agreement with the microbiota studies in adult asthmatics (Hilty *et al.*, 2010)

and with culture-dependent studies in children (Bisgaard *et al.*, 2007). Pasteurellacea and Neisseriaceae were found to be more common in controls (Pasteurellacea at 12 months and Neissereaceae all ages). Even though it was not possible to identify them at species level nonetheless from the 16S rRNA sequences it was possible to establish that the OTU assigned to Neisseriaceae did not belong to the genus *Neisseria*. Consequently it therefore might be another member of that family Neisseriaceae like *Eikenella* or *Kingella* (part of the oral normal microbiota) (Belstrom *et al.*, 2014).

When comparisons between cases and controls were performed using beta diversity inter-groups comparisons, significant differences were seen. The differences found were more important when statistics that take account the presence/absence of OTUs were implemented whilst the abundance levels of OTUs was of less importance. Additionally the phylogenetic information was also not relevant, since the UNIFRAC statistics were no better at discriminating differences than the conventional statistics.

When comparing between groups statistically significant differences were seen overall as well as between infants with history of wheezing and controls at 7 months. These differences were also revealed through the cluster analysis (NMDS and PCoA plots). Of particular interest was that for the 7 months age group the PCoA bi-plot revealed that *Streptococcus* spp. was distributed evenly across groups (Figure 4.12 A), whilst *Acetobacter* and *Staphylococcus* underpinned the clustering of the cases, with *Prevotella,* Bacteroidales and *Neisseria* being key in the clustering of the

controls. When the samples were plotted all together however, or at 12 and 24 months, clear clusters could not be discriminated (Figure 4.12 B and C).

The alpha diversity analysis revealed that diversity was different when comparisons between all groups were conducted using the raw number of observed species. Richness was statistically different but not Evenness (Li *et al.*, 2012).

Considering healthy infants versus infants with a history of wheezing at 7 months of age, it is interesting to note that for the healthy infants the microbiota is more diverse. This suggests that a lack of diversity results in the infant having a higher risk of developing wheezing and asthma later in life.

One of the key aims of this part of the ECUAVIDA microbiota study was to establish the composition of the airways microbiota and its development during early life. What has been revealed is that a more diverse microbial community is acquired with age (7 months up to 24 months). These differences are evident from the alpha diversity (analysis by Shannon and Inverse Simpson diversity indexes)/ between age groups, and was mainly caused by an increase in Richness (number of unique taxa on the multiple rarefaction curves). In contrast, the distribution of taxa was not found to be different between ages (analysis of Evenness). Paired comparisons between all ages revealed that the differences are maintained, and there was clear evidence of an increase of diversity with age (Figure 4.7 heat map and Figure 4.9 alpha diversity plots).

The beta diversity analysis between age groups also revealed significant differences specifically presence/absence of new taxa with

increasing age but with no abundance or phylogenetic differences (Section 4.2.2.3.2). The NMDS plots show a clear clustering of samples dependent on their age particularly for the control group.

The epidemiological characteristics of the infants used on the study show firstly that there is a very low frequency of antibiotics usage reported in the clinical records for cases and controls. Secondly there is no history of corticosteroids use for these infants. These results reassure that the microbiome patterns encountered were not skewed by the use of antibiotics or corticosteroids.

Several epidemiological parameters were tested in order to characterize the population studied in both groups. The gender ratio was almost a half-half relation between male and female for each group, and the microbiome patterns seen were not attributable to sex. Parental income was on average 226 USD with no differences between cases and controls, strikingly this value is much lower than the minimum salary necessary established by the Ecuadorian government (Diario, 2014), and shows that the population living in the tropics of Ecuador have high rates of poverty.

A number of epidemiological variables were variables were examined for cases and controls to determine whether any of them accounted for the differences seen in microbiota patterns between the two groups. Between the variables studied there were included: relative ratio of crowdedness (the number of people that live with each child in the house per the number of rooms in the house), biometric measures at birth, maternal education level, and comorbidities like respiratory tract infections, anaemia and malnutrition.

None of them were able to explain the microbiota differences found between these two groups.

In conclusion the microbiome patterns are different between past and current wheezing infants, and unlike healthy controls in whom diversity of the airways microbiota increases over time, children with repetitive past wheezing episodes during their first year of age maintained mid and long-term differences. The wheezing episode is characterized by the presence of potentially pathogenic bacteria that decrease in abundance afterwards but remain present at mid and long-term.

Additionally, another key finding of this part of the study is that the development of the airways microbiota in infants consists principally of the acquirement over time of a more complex and diverse flora. The differences in bacterial diversity are also important in infants that reported wheezing episodes until the 7 months of age, and based on the current results a more diverse microbiota at this time is protective against the development of wheezing.

Chapter 5: Streptococcus Identification at Species Level using a New Culture Independent Technique based on Pyrosequencing

Analysis of the 16S rRNA gene to characterize bacterial communities is a very useful tool to obtain a wide panorama of the whole bacterial taxa present. The 16S rRNA gene however has several limitations in the taxonomy assignation at species level for a number of genera. Consequently pending on initial findings from 16S rRNA gene analysis it can be important to perform a further characterization of particular genera of interest.

In the ECUAVIDA microbiota study of oropharynx samples, the characterization of possible pathogenic bacteria was of particular importance in order to fully understand the relationships between the bacteria present and the wheezing syndrome phenotype. The 16S rRNA gene analysis conducted in the prior chapters (Chapters 3 and 4) revealed that more than 70% of the airways microbiota belonged to the phylum Firmicutes, which was also the most diverse phylum (more than 50% of OTUs found). The most common and diverse genus within Firmicutes was consistently *Streptococcus.* Due to the genetic uniformity of the 16S rRNA gene however, it was not possible to differentiate between some members of the *Streptococcus* genus.

Consequently the pathogen *S. pneumoniae* (also known as Pneumococcus) which is a major cause of morbidity and mortality in infants (Galli *et al.*, 2013, Harmes *et al.*, 2013, Nau *et al.*, 2013), ends up

categorized as part of the *S. mitis* ´group´ that encompasses mostly commensal, harmless bacteria (Mitchell, 2011, Kilian *et al.*, 2014).

A good strategy to achieve differentiation between members of a genus is to perform further genotyping of other housekeeping genes. This allows the identification of *Streptococcus* at species level and potentially even serotype categorization (El Aila *et al.*, 2010). Methods in order to conduct this genotyping still necessitate an element of culture (multi locus sequence typing [MLST] and multi locus sequence assignation [MLSA]). To date there are no culture-independent methods for this genotyping and cutting edge next generation sequencing has not been applied.

The aim of this part of the study was therefore to develop and optimise a culture-independent robust technique to identify streptococci at species level using next generation sequencing. The technique was then applied to the ECUAVIDIA oropharyngeal swab DNA samples to enable species differentiation of the *Streptococcus* genus, determination of the relative abundance of the streptococci identified and examination of their relationship with the wheezing syndrome phenotype.

5.1 Development of Streptococci Species Assay

To develop a new robust next generation sequencing based technique capable of identifying streptococci at species level and that could be used starting from mixed microbial DNA samples, a strategic plan was created as shown in Figure 5.1.

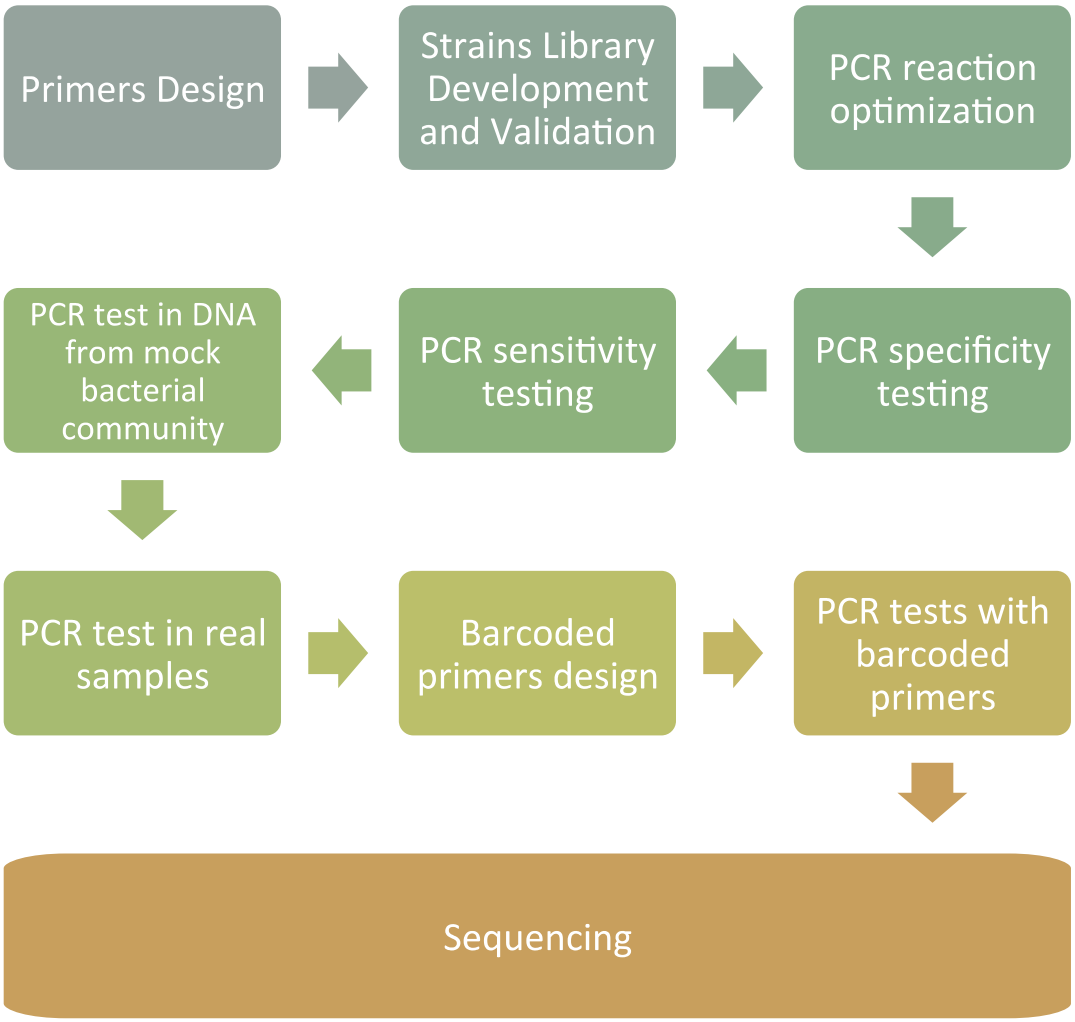


Figure 5.1 Strategic Plan to Create a new Molecular Tool to Identify Streptococci at the Species Level.

5.1.1 Candidate Genes Selection

Multi locus sequence typing (MLST) and multi locus sequence assignation (MLSA) have been and are currently the gold standard methods to identify streptococci taxonomically at species level (Bishop *et al.*, 2009, Picard *et al.*, 2004). Such techniques however require the prior culturing and isolation of the *Streptococcus* strain to be identified.

Based on the housekeeping genes commonly used for MLSA and MLST, candidate genes to develop a culture-independent identification technique using next generation sequencing (pyrosequencing) were identified. The gene selection was additionally informed by the currently

available MLSA and MLST online databanks (www.emlsa.org and http://www.mlst.net). For each gene selected, primers were chosen to ensure an amplicon length between 200 and 700 bp would be generated, as this is optimal for successful pyrosequencing.

Based on the prior literature nine-primer pairs were identified for 8 selected genes (Table 5.1).

Name	Gene Product	Primers sequences (5' – 3')	Amplicon size (bp)
guaA	GMP synthase	guaA_spratt_fwd: ATYCARTTYCACCCMGAAGT	567
		guaA_spratt_rev: CWGGNCCWGGRAATGGTTG	
map	Methionine aminopeptidase	map_spratt_fwd: GCWGACTCWTGTTGGGCWTATGC	348
		map_spratt_rev: TTARTAAGTTCYTTCTTCDCCTTG	
pfl	Pyruvate formate lyase	pfl_spratt_fwd: AACGTTGCTTACTCTAAACAAACTGG	351
		pfl_spratt_rev: ACTTCRTGGAAGACACGTTGWGTC	
ppaC	Inorganic pyrophosphatase	ppaC_spratt_fwd: GACCAYAATGAATTYCARCAATC	552
		ppaC_spratt_rev: TGAGGNACMACTTGTTTSTTACG	
pyk	Pyruvate kinase	pyk_spratt_fwd: GCGGTWGAAWTCCGTGGTG	492
		pyk_spratt_rev: GCAAGWGCTGGGAAAGGAAT	

rpoB	RNA polymerase beta subunit	rpoB_spratt_fwd: AARYTIGGMCCTGAAGAAAT	516
		rpoB_spratt_rev: TGIARTTTRTCATCAACCATGTG	
sodA	Superoxide dismutase	sodA_spratt_fwd: TRCAYCATGAYAARCACCAT	378
		sodA_spratt_rev: ARRTARTAMGCRTGYTCCCARACRTC	
tuf	Elongation factor Tu	tuf_spratt_fwd: GTTGAAATGGAAATCCGTGACC	426
		tuf_spratt_rev: GTTGAAGAATGGAGTGTGACG	
		tuf_picard_fwd: GTACAGTTGCTTCAGGACGTATC	865
		tuf_picard_rev: ACGTTCGATTTCATCACGTTG	

Table 5.1. Summary of the Genes and Primers used to Develop the Molecular Tool for Streptococci Identification. Based on online databases (Bishop *et al.*, 2009, Picard *et al.*, 2004).

As can be seen in Table 5.1, one set of primers (tuf_picard_fwd and tuf_picard_rev combination) for the candidate gene *tuf* unavoidably result in a larger than ideal amplicon. They were nonetheless still considered as the key polymorphisms used to determine species identification were located at the beginning and middle of the amplicon [Picard *et al.* (2004)]. For each primer a blast search was conducted to ensure primer specificity for streptococci versus other bacteria and human DNA. Using the online databases for molecular typing based on MLSA (http://www.emlsa.org), MLST (www.mlst.net) and the *Streptococcus* spp. sequences found using Entrez search on the Genebank, multiple alignments were performed using the ClustalW algorithm for each of

the eight candidate genes in order to double check the amplicon sequences and the localization of the key polymorphisms within them (Figure 5.2).

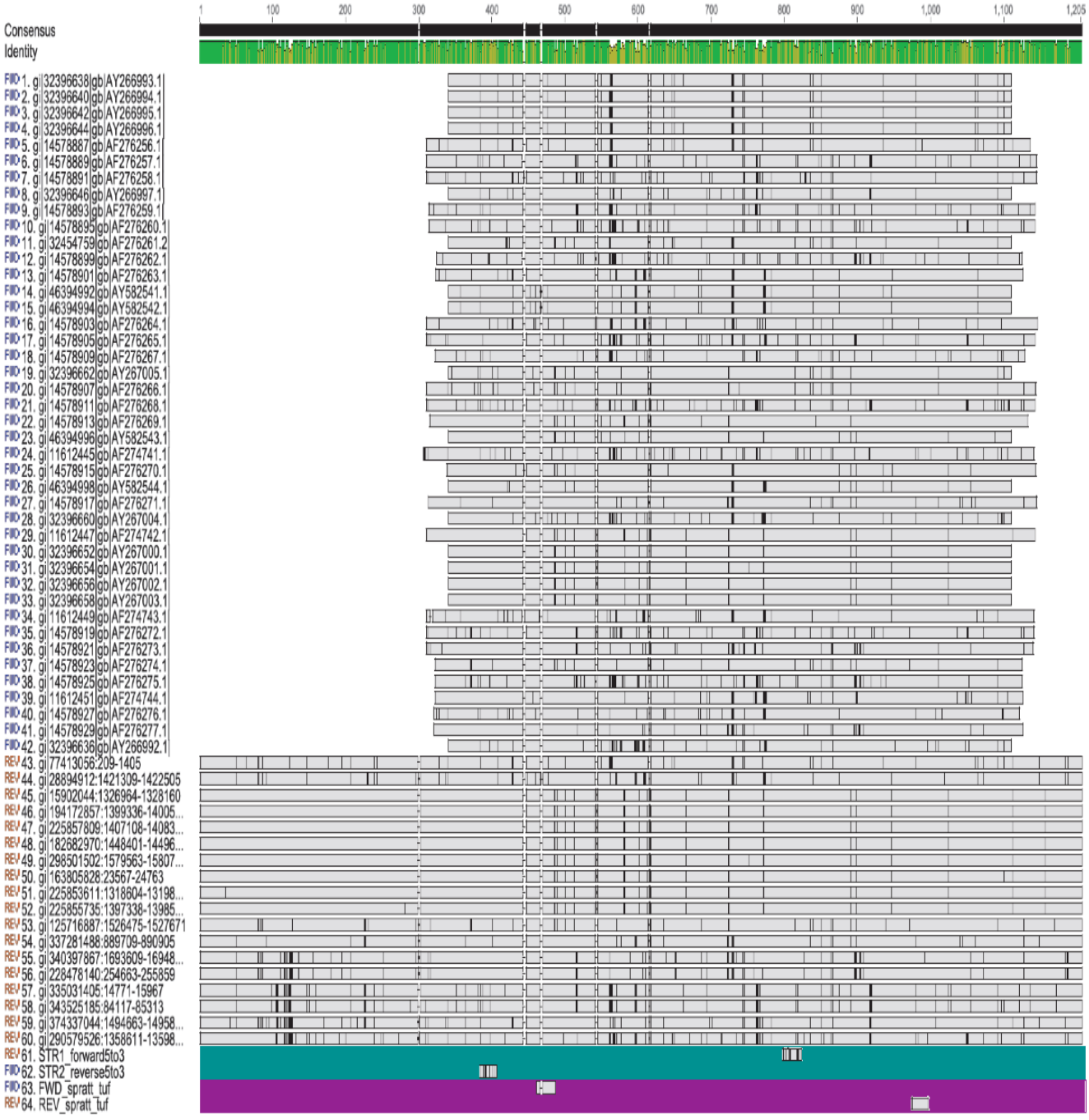


Figure 5.2. Example (for *tuf*) of the ClustalW Multiple Alignments performed on Geneious version 7.1.6. Sequences were obtained from www.emlsa.org and Pubmed nucleotide for the *tuf* gene downloaded and aligned to determine the principal polymorphisms hotspots and how each set of primers captures them. On light blue are shown the *tuf* primers based on Piccard *et al*. 2004, and in purple the primers pair based on Bishop *et al.* 2009. There are included in the alignment the forward (FWD) and complementary (REV) chains obtained in the databases.

Additionally, phylogenetic trees of all the 9 primer set candidates were created on MEGA 6 based on the emlsa database sequences to determine the molecular distances of each gene. This was performed to validate their importance on taxonomic identification. Examples are shown for *map* and *tuf* (Figure 5.3).

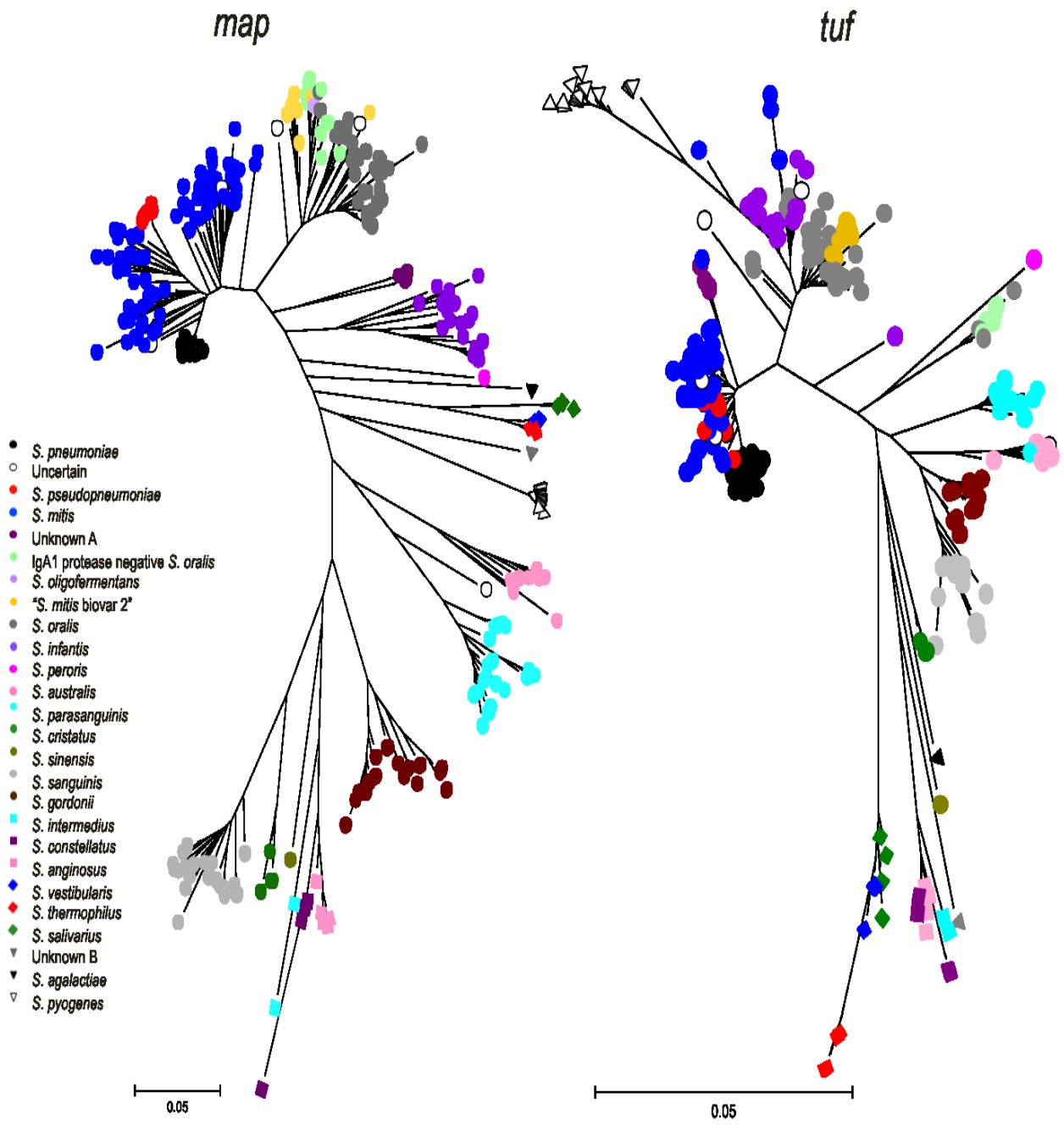


Figure 5.3. Neighbour Joining Non-rooted Phylogenetic trees of *map* and *tuf* to Determine the Taxonomy Usefulness for Streptococci. *Map* shows a good taxonomy differentiation between *S. mitis* and *S. pneumoniae,* whilst *tuf* is better to discriminate *S. pyogenes* from the others. Images based from the emlsa.org database and the study of Bishop *et al.* (2009) under permission (see appendix).

5.1.2 Construction of a Bacterial Library

In order to test primer specificity, a bacterial library was constructed using conventional culturing. A total of 22 strains of bacteria (obtained from Professor Simon Kroll strains library, St. Mary´s Campus, Imperial College London), previously cultured from the human microbiome in health and disease, as well as reference strains of commonly found bacteria, were selected including different species in order to determine specificity in taxonomy assignation. Each strain was cultured using different types of media (solids or broths) according to individual bacterial requirements (Table 5.2).

Strain	Media
Escherichia coli DH5α	BHI Agar
Haemophilus ducrey	
Haemophilus haemolyticus ATCC 33390	
Haemophilus parasuis	
Micrococcus luteus ATCC 9341	
Moraxella catarrhalis MIDG 635 B59455	
Moraxella haemolytica	
Neisseria flava	
Neisseria flavescens	
Neisseria lactamica	
Neisseria meningitidis	
Pasteurella sp.	
Pseudomonas multocida	
Streptococcus pyogenes ATCC 19615	

Sallmonella dublin	
Shigella flexneri	
Staphylococcus aureus ATCC	
Bifidobacterium dentium	Chocolate Agar
Streptococcus agalactiae ATCC 13813	
Streptococcus mitis ATCC 49456	
Streptococcus pneumoniae ATCC 27336	
Haemophilus influenzae Rd	BHI broth + Factors X and V

Table 5.2. List of Bacterial strains. Strains used to create the bacterial library and the media used for their culture.

After culturing for 24 hours at 37ºC a single colony was isolated from each plate. Each colony was then re-cultured in BHI broth and then cultured to exponential growth phase confirmed by spectrophotometry (Chapter 2: Section 2.8). DNA was extracted using the MPBio kit for soil and quantified using the Nanodrop (DM-1000 Thermo Scientific, USA) (Chapter 2: Section 2.8).

Each bacterial DNA obtained was verified in terms of taxonomy assignation using conventional Sanger sequencing of the entire 16S rRNA gene. For verification the universally used 16S rRNA gene primers at positions 27F, 341F and 1942R (according to *E. coli* genome) (Frank *et al.*, 2008, Marchesi *et al.*, 1998, Lane, 1991) were utilised. Sanger sequencing was performed and the sequences results were aligned with the Ribosomal

Database Project (Cole *et al.*, 2003) sequences bank. Each strain obtained by culture was confirmed 100% by DNA alignment.

5.1.3 PCR Validation

5.1.3.1 PCR Conditions Optimization

PCRs were optimized by testing a number of cycling conditions including different annealing temperatures, primers, BSA and Betaine concentrations. The final optimized PCR reaction set up is detailed in Table 5.3.

	µl 1X
DNA (10 ng)	1
Fwd primer 1µM	1
Rev primer 1µM	1
dNTP mx (10nM each)	0.2
Buffer 2 10X	1
BSA (20mg/ml)	0.4
Betaine (1.3M final)	2.6
Taq 2.5 U/µl	0.1
H2O	2.7
TOTAL	9 µl

Table 5.3. Optimized PCR reaction.

The PCR temperature and cycling conditions were also optimized in order to produce successful amplification (Table 5.4).

Step	**temp**	**time**	**cycles**	***gua, map, pfl, tuf***	***ppaC, pyk, rpoB,***
hot start	94°C	2 min			
denaturation	94°C	30 s			
annealing		30S	30	55°C	50°C
extension	72°C	30s			

final extension	72°C	5 min

Table 5.4. PCR Temperature and Cycling Conditions.

5.1.3.2 PCR Specificity Verification

Using DNA extracted from the 22 bacterial strains, the nine sets of primers were tested. This test was performed to determine which primers resulted in specific amplification of the streptococci species only. If false bands or amplification of non-streptococcal strains were observed this would highlight further optimisation was required and/or potential primer redesign (Table 5.5).

Strain	*guaA*	*map*	*pfl*	*ppaC*	*pyk*	*rpoB*	*sodA*	*tuf_S*	*tuf_P*
Bifidobacterium dentium	-	-	-	-	-	-	-	-	-
E. coli DH5α	-	-	-	-	-	-	-	-	-
Haemophilus ducrey	-	-	-	-	-	-	-	-	-
Haemophilus haemolyticus ATCC 33390	-	-	-	-	-	-	-	-	-
Haemophilus influenzae Rd	-	-	-	-	-	-	-	-	-
Haemophilus parasuis	-	-	-	-	-	-	-	-	-
Micrococcus luteus ATCC 9341	-	-	-	-	-	-	-	-	-
Moraxella catarrhalis MIDG 635 B59455	-	-	-	-	-	-	-	-	-
Moraxella haemolytica	-	-	-	-	-	-	-	-	-
Neisseria flava	-	-	-	-	-	-	-	-	+

Neisseria flavescens	-	-	-	-	-	-	-	-	+
Neisseria lactamica	-	-	-	-	-	-	-	-	-
Neisseria meningitidis	-	-	-	-	-	-	-	-	+
Pasteurella sp.	-	-	-	-	-	-	-	-	-
Pseudomonas multocida	-	-	-	-	-	-	-	-	-
S. agalactiae ATCC 13813	+	+	+	+	+	+	+	+	+
S. mitis ATCC 49456	+	+	+	+	+	+	+	+	+
S. pneumoniae ATCC 27336	+	+	+	+	+	+	+	+	+
S. pyogenes ATCC 19615	+	+	+	+	+	+	+	+	+
Sallmonella dublin	-	-	-	-	-	-	-	-	-
Shigella flexneri	-	-	-	-	-	-	-	-	-
Staphylococcus aureus ATCC 25923	-	-	-	-	-	-	-	-	-

Table 5.5. Specific PCR Tests in the Nine Primer Sets. There were used the 22 bacterial strains and positive signs show when there was produced amplification.

All the primer sets were designed to specifically target *Streptococcus* and 8 of the nine primer sets did only give products when the *Streptococcus* strains were used as PCR templates. The exception was the PCR using one of the primer sets targeting the *tuf* gene – primers by Picard *et al.* These were found to amplify the *Neisseria* spp. strains but as seen later (Section 5.1.2.5) nonetheless the sequencing still enabled to differentiate between *Streptococcus* spp. and *Neisseria* spp.

5.1.3.3 PCR Sensitivity Verification

Since the amount of template of DNA of each bacterium in a mixed microbial sample, like that of the oropharyngeal swab, is likely to be variable it was necessary to determine the range of concentrations required to obtain amplification. For that reason, a DNA concentration series test was performed using *Streptococcus* pyogenes DNA (used as a standard *Streptococcus* strain) at decreasing concentrations 100, 10, 1, 0.1, 0.01, 0.001 ng/µl (Table 5.6).

PRIMERS	***S. pyogenes* (ng/µl)**						
	100	**10**	**1**	**0.1**	**0.01**	**0.001**	**0**
guaA	+	+	+	-	-	-	-
map	+	+	+	+	+	+	-
pfl	+	+	+	+	+	+	-
ppaC	+	+	+	+	-	-	-
pyk	+	+	+	+	+	-	-
rpoB	+	+	+	+	+	-	-
sodA	+	+	+	+	+	-	-
tuf_P	+	+	+	+	+	+	-
tuf_S	+	+	+	+	+	+	-

Table 5.6. Sensitivity PCR tests using the Nine Primer Sets and *Streptococcus pyogenes* DNA. Decreasing concentrations ranging from 100 ng/µl to 0.001 ng/µl of DNA were used. + indicates successful amplification, and - indicates no amplification.

No PCR amplification was seen at with a concentration of 0.1 ng/µl template for the *guaA* set of primers, at 0.01 ng/µl for *ppa*C, and at 0.001 ng/µl for *pyk, rpoB* and *sodA*. *Map, pfl, tuf* (P and S) were found to be the most sensitive sets of primers resulting in amplicons for all concentrations of template DNA.

5.1.3.4 PCR Sensitivity using a Mixed Microbial DNA Template

To determine the sensitivity in primers detection when dealing with a mixed microbial DNA sample/template, decreasing concentrations of *Streptococcus. pyogenes* DNA and increasing concentrations of *Escherichia coli* were used (*S.pyogenes/E.coli* in ng/µl: 10/0, 10/10, 1/10, 0.1/10, 0.01/10 and 0/10) (Figure 5.4).

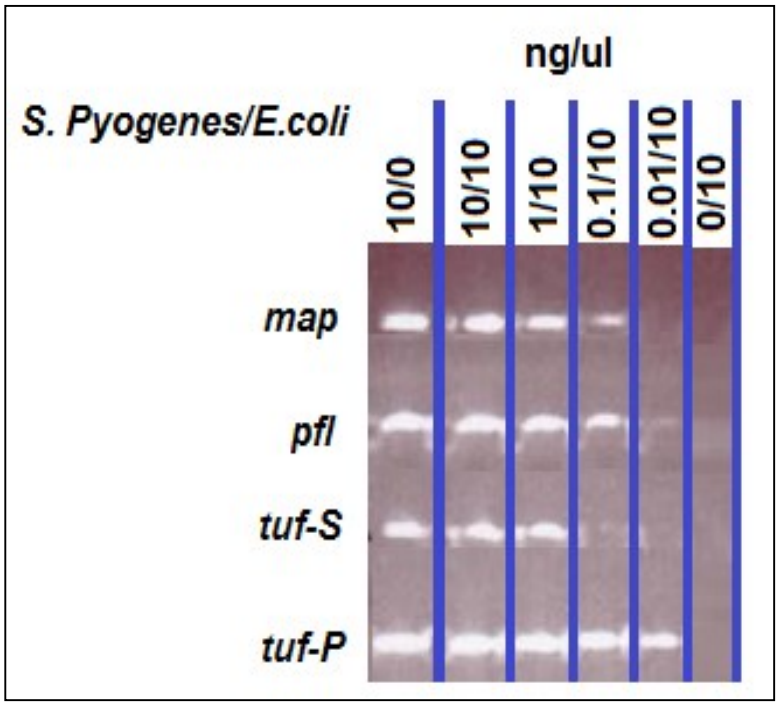


Figure 5.4. PCR Amplification Tests using a Mixed DNA Template. There were used varying concentrations of *Streptococcus pyogenes* and *Escherichia coli* DNA.

The mixed combination of *S. pyogenes* and *E. coli* highlighted that for two of the candidate genes: *map* and *pfl* no amplification inhibition by the presence of non-streptococcal DNA was observed for one set of primers for each gene. No amplification inhibition was seen for either of the primer sets for the candidate gene *tuf*. These set of primers were therefore focussed upon for the next optimization experiment.

5.1.3.5 PCR Validation using Oropharyngeal DNA Samples

Four DNA samples (extracted from throat swabs) which had undergone previous 16S rRNA gene sequencing were selected. Two of the samples had

shown the presence of *Streptococcus*, the other two samples being *Streptococcus* negative. PCR was performed using the two sets of primers for *tuf* and the single set of primers for *map* and *pfl* (Figure 5.5).

Amplification products using the four sets of primers was only seen for those samples that had through 16S rRNA gene sequencing been shown to be positive for the *Streptococcus* genus (Samples 3 and 4 labelled Strep Pos figure 5.5).

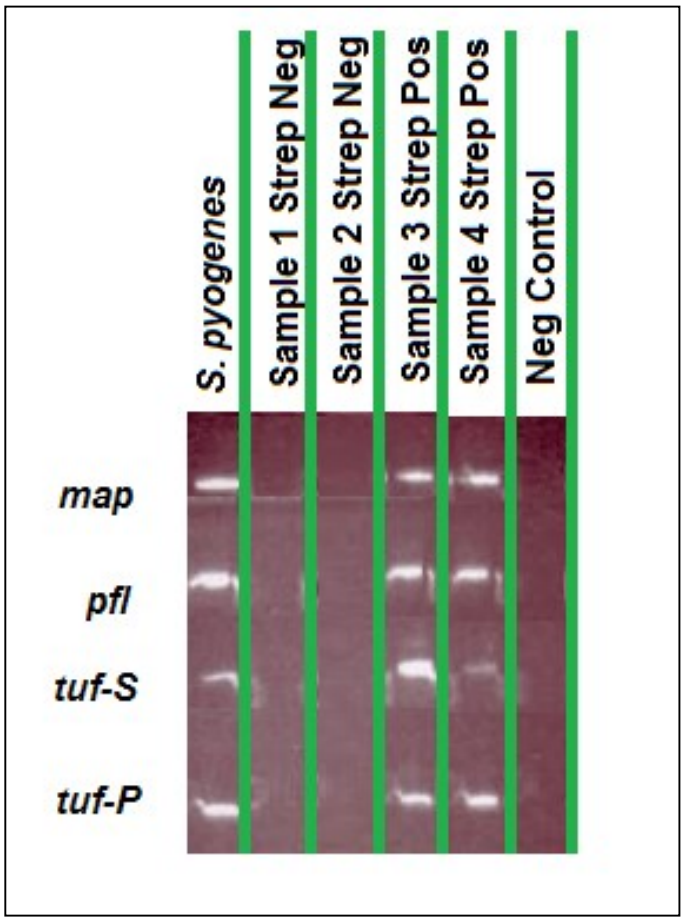


Figure 5.5. PCR Validation Tests using DNA Extracted from Real Orophayngeal Swabs.

5.1.3.6 Pyrosequencing: Primer Design

In light of the results of the PCR primer tests with the bacterial DNAs library and real sample templates (Sections 5.1.1.2 to 5.1.1.5), modifications

in the design of the primers for *map, pfl* and *tuf*_S were made to allow inclusion of barcoded sequences and adaptors both being required for pyrosequencing.

For each forward primer a sequence adaptor was added: 5' CCATCTCATCCCTGCGTGTCTCCGACTCAG 3'. The sequence adaptor enables bead ligation (Chapter 2: Section 2.7.1). On each reverse primer the same adaptor sequence was added but in addition a barcode sequence was included (between 15 and 20 nucleotides). In total 30 different barcode sequences were utilized to allow amplicons from up to 30 different samples to be multiplexed and combined into one pyrosequencing run. The barcodes were selected from the list of barcodes previously tested and released to the scientific community by the Human Microbiome Project Consortium (Jumpstart Consortium Human Microbiome Project Data Generation Working, 2012).

As for the standard primers, the modified primers were tested for sensitivity, specificity and inhibition of PCRs using the same approach and set of experiments as outline previously (Sections 5.1.1.2 to 5.1.1.5). The results of these experiments for three sets of primers were identical to those obtained with the standard primers confirming that the addition of adaptor and barcode sequences did not have any undesirable impact on amplification of the target genes *map* and *pfl*, or *tuf* using the *_Bishop* primers. The primers corresponding to *tuf* based on Picard *et al.* (2004) encountered several errors during PCR that included the barcoded sequences and these primers were not pursued any further. Consequently

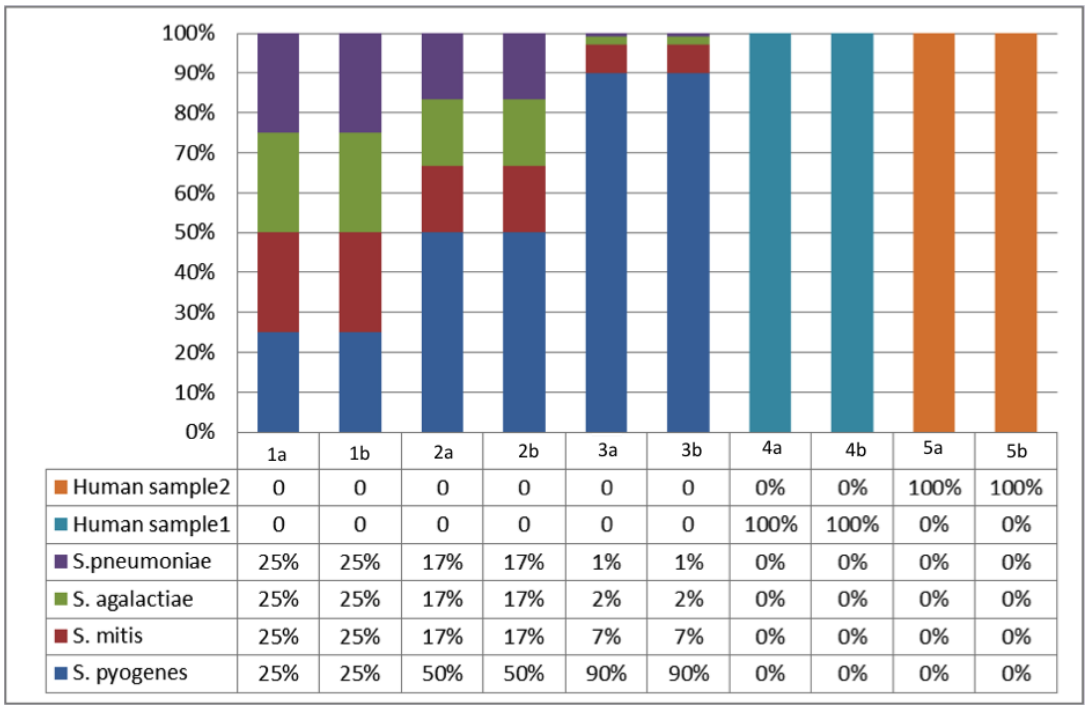
10 primer sets for each of the three genes were finally designed and used in the downstream sequencing.

5.1.4 Sequencing: Test Experiment

5.1.4.1 Test Set Construction

A sequencing test set of three samples was created using DNA extracted from four different streptococci strains: *S. pneumoniae, S. mitis, S. pyogenes* and *S.agalactiae.* The three samples were created with varying concentrations of each strain in order to determine the sensitivity of the pyrosequencing assay in detecting the varying abundances of the streptococci (Figure 5.6). The first constructed sample contained the same quantity of each streptococcal DNA strain (25% *S. pneumoniae,* 25% *S. mitis,* 25% *S. pyogenes* and 25% *S.agalactiae*) whilst the second sample was constructed with 50% *S. pyogenes* with the other three strains (*S. pneumoniae, S. mitis* and *S.agalactiae*) each at 17%. The third sample was constructed with *S. pyogenes* DNA almost completely dominant (90%) and the other remaining strains at variable percentages - 7% *S. mitis,* 2% *S.agalactiae* and 1% *S. pneumoniae*. The design of the third sample, with only 1% of *S.pneumoniae*, was specifically to determine the sensitivity of the assay in detecting such an important pathogen when present at very low levels.

Additionally included in the test set of samples were two of the ECUAVIDA DNA oropharyngeal samples identified by 16S rRNA sequencing to have different streptococcal patterns. All five of the test set samples were sequenced twice to determine reproducibility of the assay.



	1a	1b	2a	2b	3a	3b	4a	4b	5a	5b
Human sample2	0	0	0	0	0	0	0%	0%	100%	100%
Human sample1	0	0	0	0	0	0	100%	100%	0%	0%
S.pneumoniae	25%	25%	17%	17%	1%	1%	0%	0%	0%	0%
S. agalactiae	25%	25%	17%	17%	2%	2%	0%	0%	0%	0%
S. mitis	25%	25%	17%	17%	7%	7%	0%	0%	0%	0%
S. pyogenes	25%	25%	50%	50%	90%	90%	0%	0%	0%	0%

Figure 5.6. Samples used for the Pyrosequencing Test of the Four Candidate Primer Sets. Sample 1a and 1b: duplicates containing 25% *S. pneumoniae,* 25% *S. mitis,* 25% *S. pyogenes* and 25% *S.agalactiae*. Samples 2a and 2b: duplicates containing 50% *S. pyogenes,* 17% *S. pneumoniae,* 17% *S. mitis* and 17% *S.agalactiae*. Samples 3a and 3b: duplicates containing 90% *S. pyogenes,* 7% *S. mitis,* 2% *S.agalactiae* and 1% *S. pneumoniae.* Samples 4a and 4b: duplicates of oropharyngeal human sample one, with 5a and 5b: duplicates for oropharyngeal human sample two.

5.1.4.2 Sequencing Results for Test Set Samples

The 10 samples (Figure 5.5) underwent PCR and were multiplexed into a single sequencing run on the Roche 454 Junior titanium machine. The run resulted in a total of 148,000 sequences and after denoising 75,238 valid sequences remained.

5.1.4.2.1 *map* gene Sequencing Results

A range (70% to 99%) of percentage similarity between sequences was tested for OTUs assignation. The most accurate results, when compared with the OTU assignment by 16S rRNA sequencing, were obtained at the level of 80% similarity. At this level in the streptococci constructed test samples *map* sequencing resulted in eight different OTUs although four 4 OTUs accounted for more than 95% of the sequences.

For OTUs taxonomic assignation a manual alignment using the streptococci MLSA database found in www.emlsa.org was performed. Taxonomic assignment with the RDP databases part of QIIME was not feasible because it was analysed a gene different from the 16SrRNA. A total of 421 different *Streptococcus* sequences were included in the phylogenetic alignments. The phylogenetic tree was constructed on Mega 6 (Tamura *et al.*, 2013b) using the neighbour-joining algorithm and bootstrap with 1000 iterations included to measure phylogenetic distances.

The taxonomy assignations on *map* were accurate identifying the four strains used for the three constructed samples *S. pneumoniae, S. mitis, S. pyogenes* and *S.agalactiae* (Figure 5.7). For the two oropharyngeal DNA samples, sample 4 despite good PCR amplification failed to produce any sequences. This will have been due to interference of the nucleotide barcode in the sequencing process. Errors caused by barcodes (including barcodes that the HMP Consortium has released) are common and non-predictable until sequencing is performed. For the other oropharyngeal DNA sample, sample 5, most of the OTUs were assigned to *S. salivarius, S. termophilus* and *S.parasanguis.*

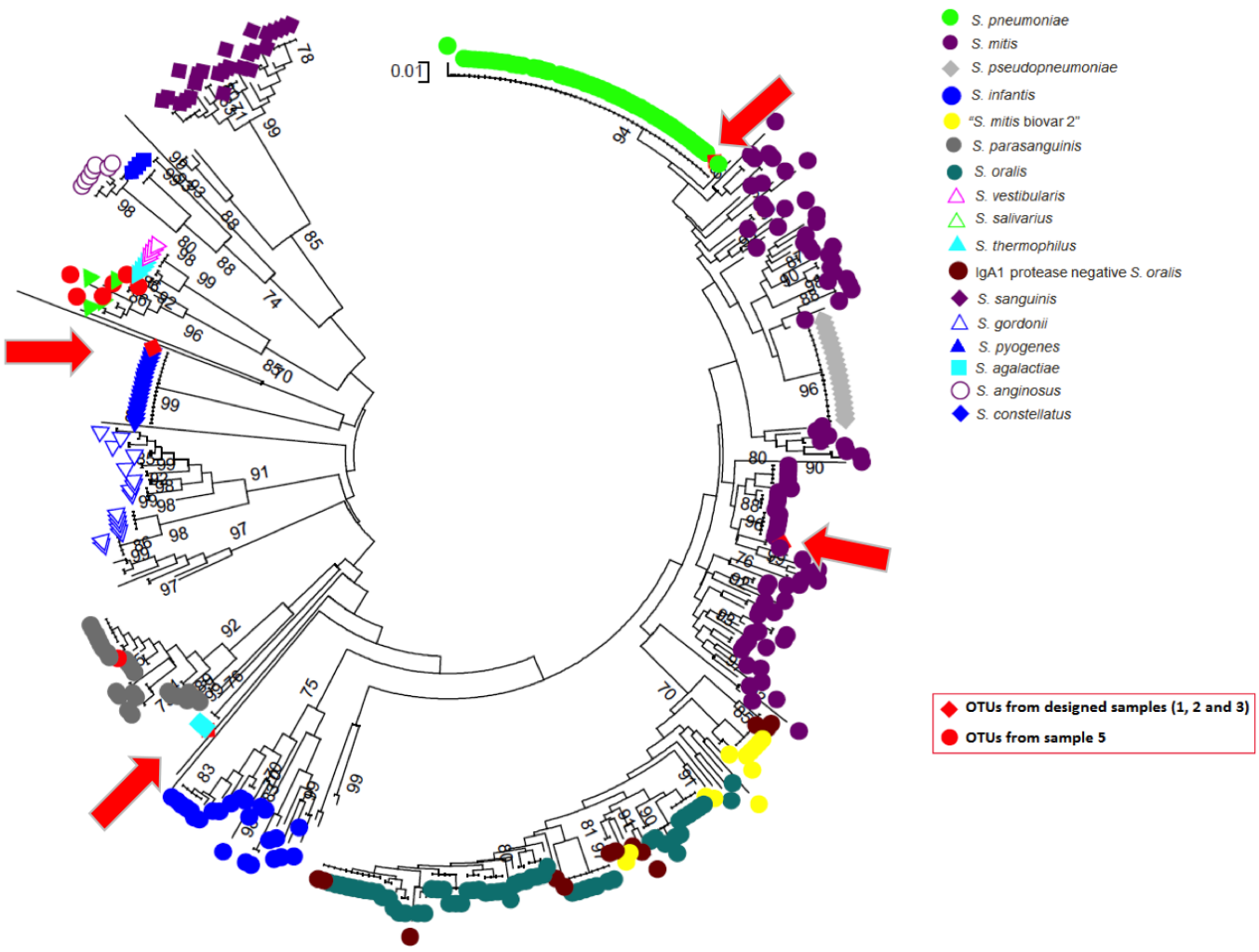
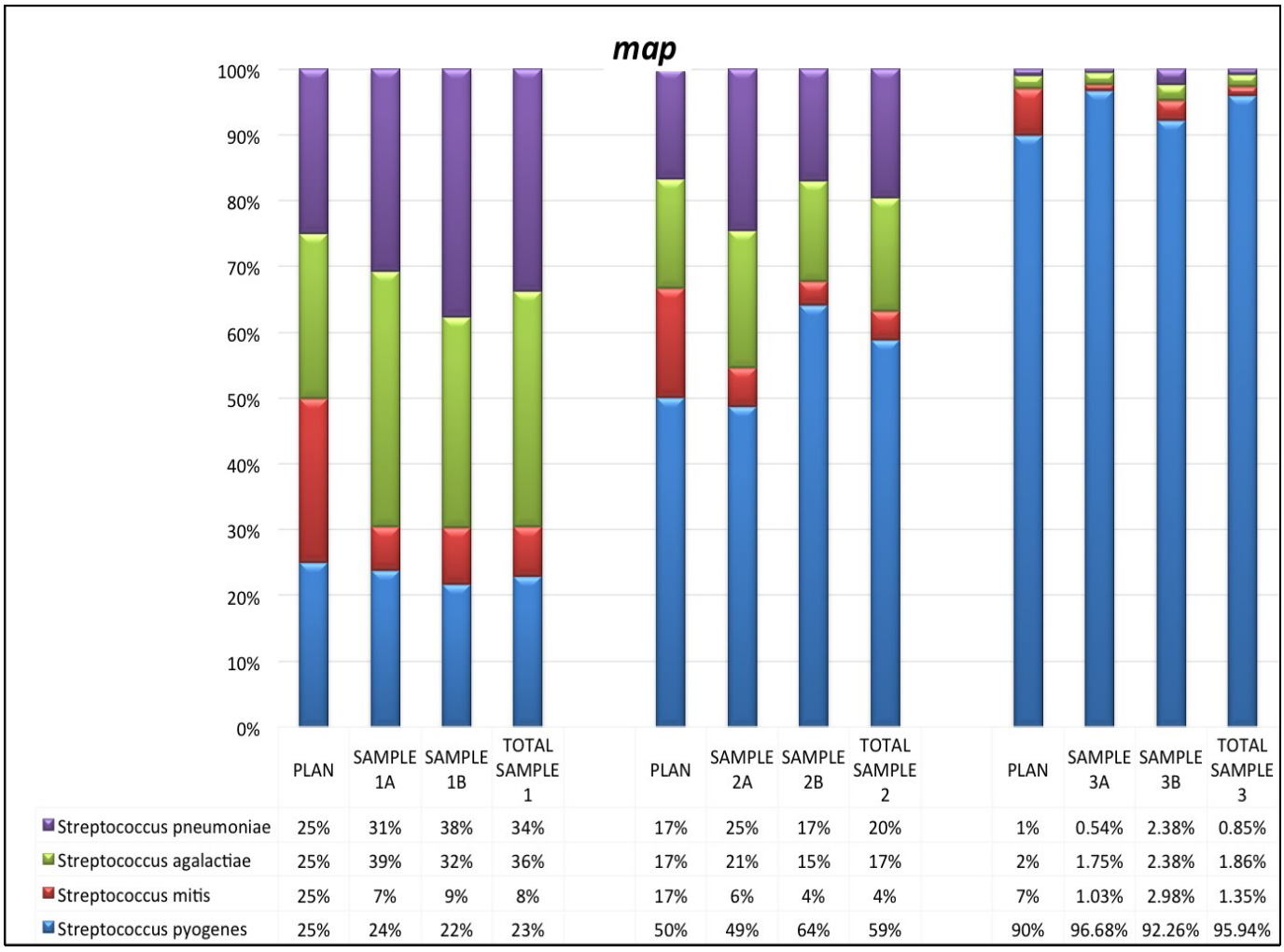


Figure 5.7. Neighbour-joining Phylogenetic Tree Showing the Taxonomic Assignations of the OTUs Identified in the Constructed Samples and in the Oropharyngeal Swab DNAs. The arrows show how the main four OTUs encountered in the constructed samples are accurately aligned with *S. pneumoniae, S. mitis, S. pyogenes* and *S.agalactiae* respectively. Scale bar is shown and the branches with more than 70% of bootstrap level are labelled.

Looking at the relative abundance (Figure 5.8), the results revealed that *map* did reasonably well in determining the percentages and ratios to one another for three of the four streptococci (main OTUs) present in the constructed test samples (P>0.05). The abundance of *S. mitis* was however

somewhat underestimated (P<0.001). Comparing the relative abundance proportions seen for the duplicate samples revealed that the assay was robust. The presence of *S.pneumoniae* was accurately identified even at the minimal percentage of 1% (Samples 3a and 3b).



	PLAN	SAMPLE 1A	SAMPLE 1B	TOTAL SAMPLE 1	PLAN	SAMPLE 2A	SAMPLE 2B	TOTAL SAMPLE 2	PLAN	SAMPLE 3A	SAMPLE 3B	TOTAL SAMPLE 3
Streptococcus pneumoniae	25%	31%	38%	34%	17%	25%	17%	20%	1%	0.54%	2.38%	0.85%
Streptococcus agalactiae	25%	39%	32%	36%	17%	21%	15%	17%	2%	1.75%	2.38%	1.86%
Streptococcus mitis	25%	7%	9%	8%	17%	6%	4%	4%	7%	1.03%	2.98%	1.35%
Streptococcus pyogenes	25%	24%	22%	23%	50%	49%	64%	59%	90%	96.68%	92.26%	95.94%

Figure 5.8. Relative Abundance Results for the Constructed Test Samples. Each sample duplicate is shown as A and B, with the average abundance obtained from the 2 duplicates labelled as TOTAL SAMPLE. In all samples there is an underestimation of *S.mitis* abundance but the relative proportions of the other three streptococci are maintained.

The oropharyngeal sample (Sample 5) produced interesting results in terms of relative abundance. It distinguished four different OTUs of *S. salivarius* and one of *S. parasanguis.* The proportions identified were comparable with the proportions of OTUs assigned to S*treptococcus* spp. using 16S rRNA sequencing although the latter data could not identify the actual species (i.e. OTUs 1328 and 1057 were assigned as *S. salivarius* using *map* but 16SrRNA could not differentiate them from other 'mitis' streptococci) (Figure 5.9). Again agreement between the duplicates (Samples 5a and 5b) was high.

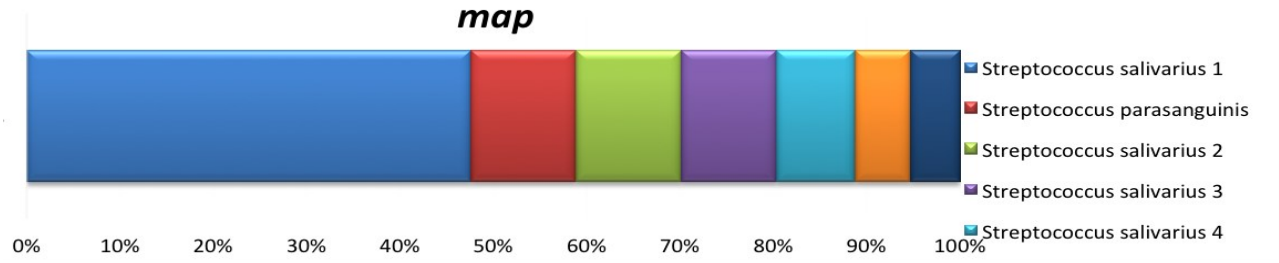


Figure 5.9. Relative Abundance of Streptococcal OTUs using *map* Gene Sequencing. *S.salivaius* OTU1 was the most common OTU and then in decreasing abundance *S. parasanguis* and *S. salivarius OTUs 2, 3, and 4.*

In conclusion *map* is a very good and reliable candidate gene in this pyrosequencing assay as it has enabled streptococci taxonomic assignation and has provided an accurate insight of the relative abundance of streptococci within each sample.

5.1.4.2.2 *pfl* gene Sequencing Results

As for *map,* a range (70% to 99%) of percentage similarity between sequences was tested for OTUs assignation and 80% was again found to be

the most accurate value. The taxonomy assignation was also performed manually by the construction of a phylogenetic tree on Mega 6 (Tamura *et al.*, 2013b) using the neighbour-joining algorithm and bootstrap with 1000 iterations that included the OTUs representative sequences obtained and the *pfl* streptococci database found in www.emlsa.org. A total of 421 different *Streptococcus* sequences were included in the phylogenetic alignments.

Disappointingly the barcodes for samples 1A and 2B failed to produce any sequences despite good amplification (Section 5.1.2.3) so could not be analysed any further. The taxonomic assignation for the other samples showed an accurate assignation on *S. pyogenes, S. agalactiae* and *S. pneumoniae* (Figure 5.10). *S. mitis*, correctly identified by *map,* was however identified as *S. infantis*. Interestingly *S. infantis* belongs to the same group ('mitis') as *S. mitis.*

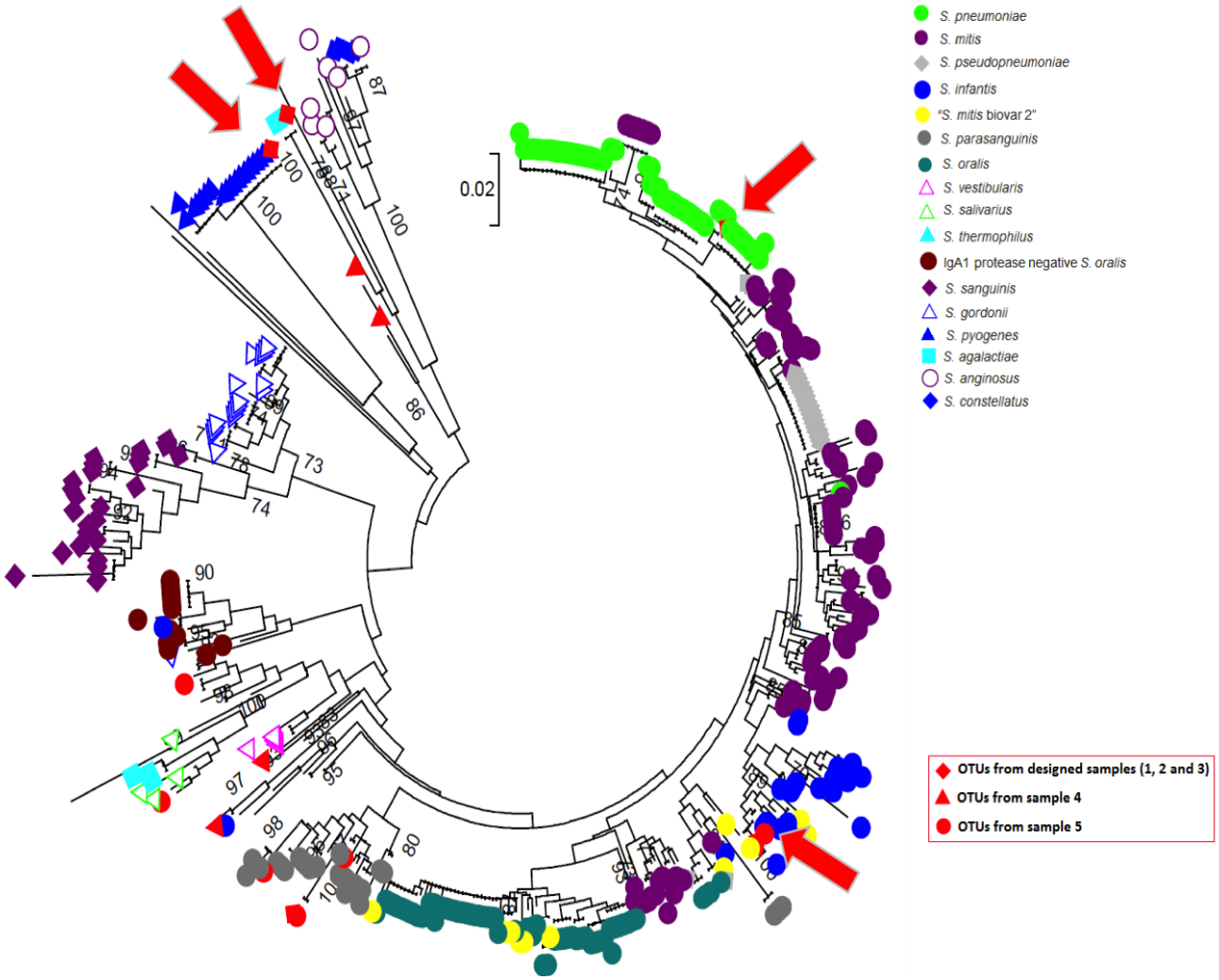
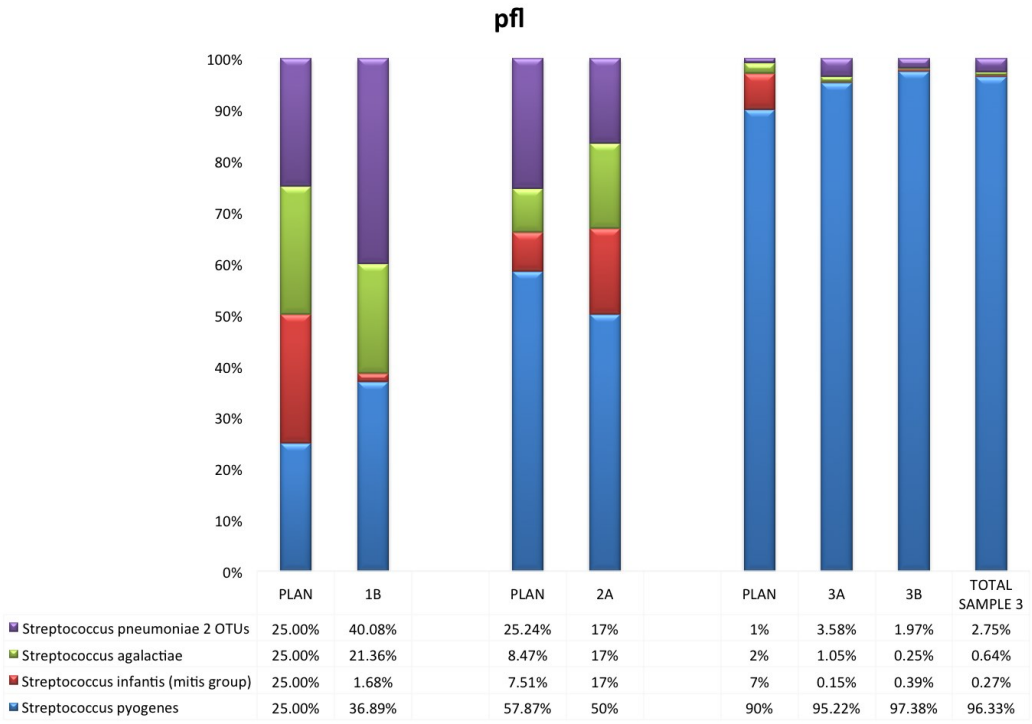


Figure 5.10. Neighbour-joining Phylogenetic Tree Showing the Taxonomic Assignations of the OTUs found by *pfl* in the Constructed Samples and in the Oropharyngeal Swab DNAs. The arrows show how the three OTUs encountered in the constructed samples are accurately aligned with *S. pneumoniae, S. pyogenes* and *S.agalactiae* respectively. Disappointedly *S. mitis* was assigned to *S. infantis.* Scale bar is shown and the branches with more than 70% of bootstrap level are labelled.

The relative abundance results showed a comparable proportionality of OTUs between the sequencing results and the expected concentrations on the constructed samples. In the results it is shown a high variability of *S.infantis* (wrongly assigned to *S. infantis* as it should be *S. mitis)*

abundance compared with the expected, there is an underestimation on samples 1 and 3 but an overestimation on sample 2 (*P*<0.001). The rest of strains were fairly precise with the expected abundance, and in all samples it was possible to identify *S. pneumoniae* (Figure 5.11).



	PLAN	1B		PLAN	2A		PLAN	3A	3B	TOTAL SAMPLE 3
Streptococcus pneumoniae 2 OTUs	25.00%	40.08%		25.24%	17%		1%	3.58%	1.97%	2.75%
Streptococcus agalactiae	25.00%	21.36%		8.47%	17%		2%	1.05%	0.25%	0.64%
Streptococcus infantis (mitis group)	25.00%	1.68%		7.51%	17%		7%	0.15%	0.39%	0.27%
Streptococcus pyogenes	25.00%	36.89%		57.87%	50%		90%	95.22%	97.38%	96.33%

Figure 5.11. Relative Abundance Results for *pfl* the Constructed Samples and the Comparison with the Expected Results (PLAN). Sample 3 has duplicates A and B which abundances were averaged on TOTAL SAMPLE 3.

For the oropharyngeal samples DNA, *pfl* was able to identify several streptococcal OTUs but in contrast to *map* and reflecting the results obtained for the constructed test samples, there were issues in relation to *S. mitis* assignation. For Sample 4, OTUs were predominantly belonged to the 'mitis' group, being to species of this group including *S. vestibularis, S. infantis* and *S. constellatus.* Sample 5 also had several OTUs assigned to the 'mitis' group but again assigned to species like *S. salivarius, S. infantis, S. parasanguis* and *S.*

australis (Figure 5.12). For both the samples the taxonomy assignation diverged with the approximation of species identification by 16S rRNA sequencing. The latter and the results of the constructed test samples highlight that *pfl* performs poorly for streptococci identification in this assay.

5.1.4.2.3 *tuf* gene Sequencing Results

For the *tuf* sequencing results, as for the prior two candidate genes, a range (70% to 99%) of percentage similarity between sequences was tested for

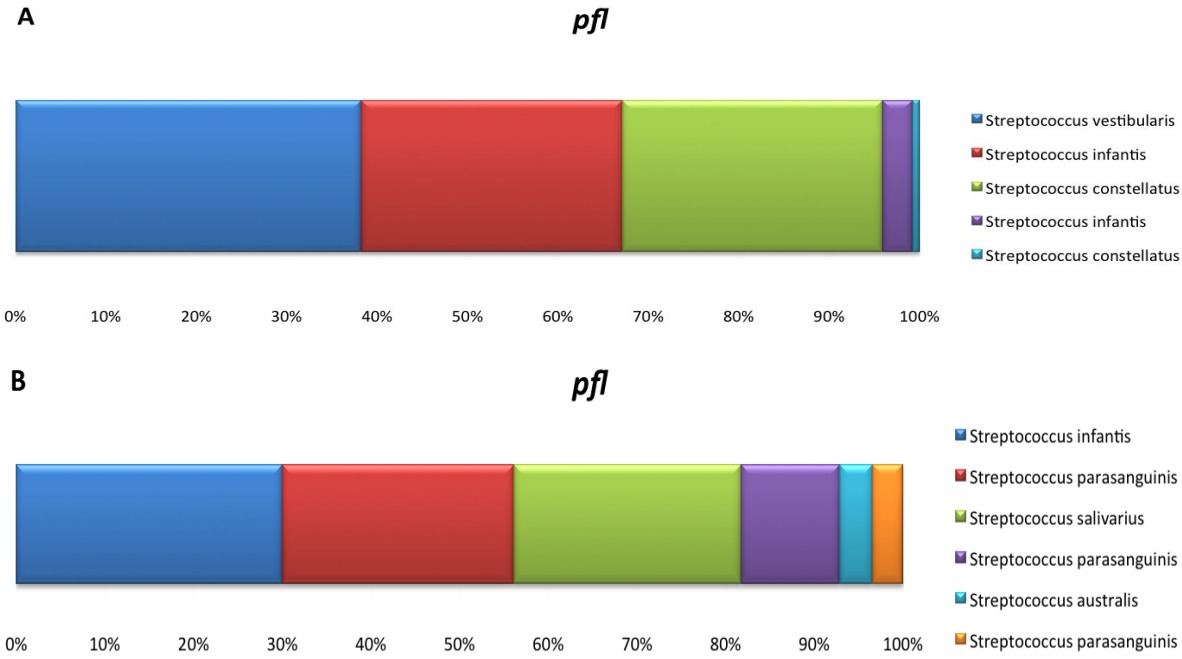


Figure 5.12. Relative Abundance for Two Oropharyngeal Samples previously Analysed using 16SrRNA Pyrosequencing. A) sample 4 and B) sample 5.

OTUs assignation and 80% was again found to be the most accurate value. In terms of taxonomic identification accuracy when phylogenetic trees were performed using Mega 6 (Tamura *et al.*, 2013b), *tuf* was able to identify all four species in the constructed test samples: *S. pneumoniae, S. mitis, S. pyogenes* and *S. agalactiae* (Figure 5.13).

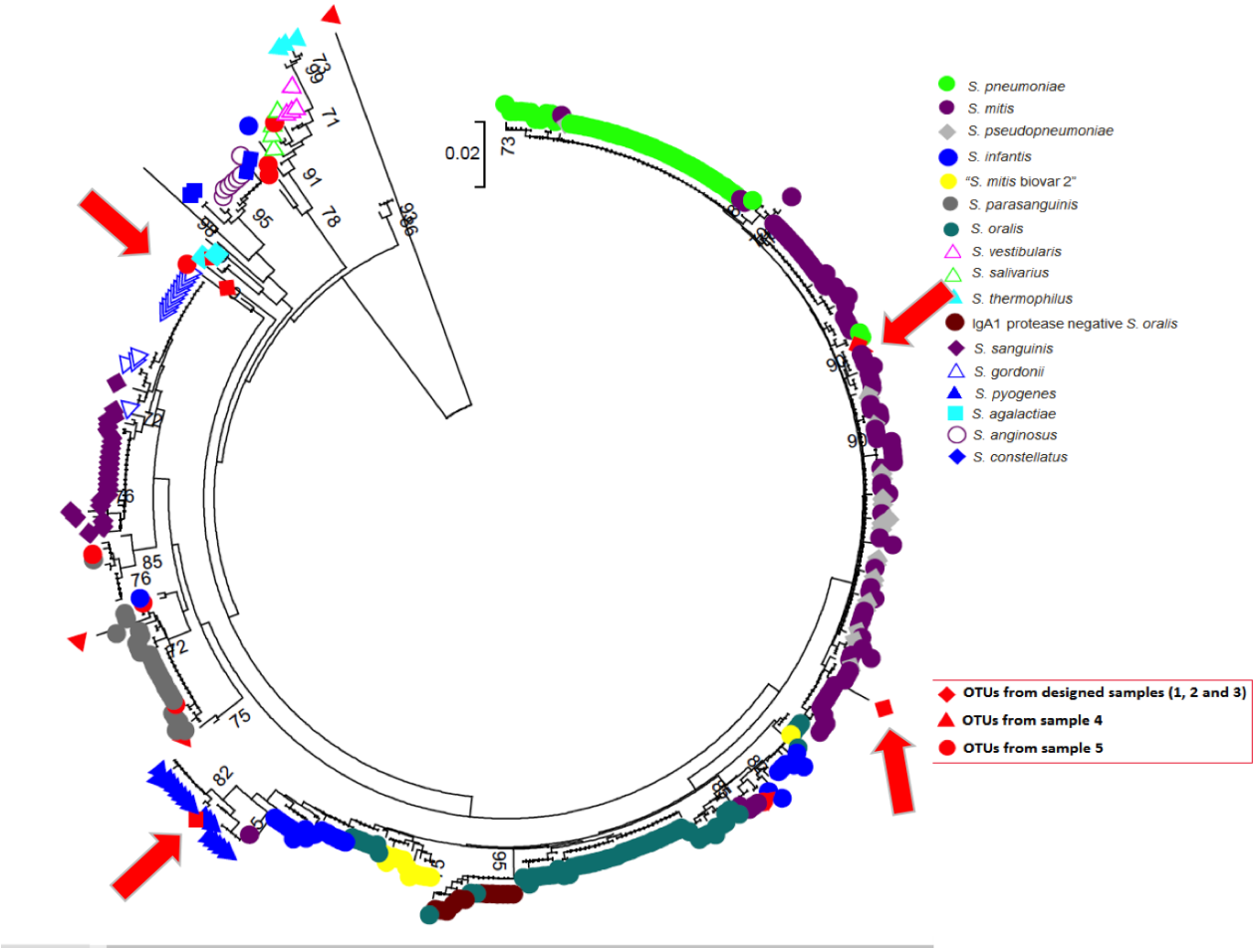
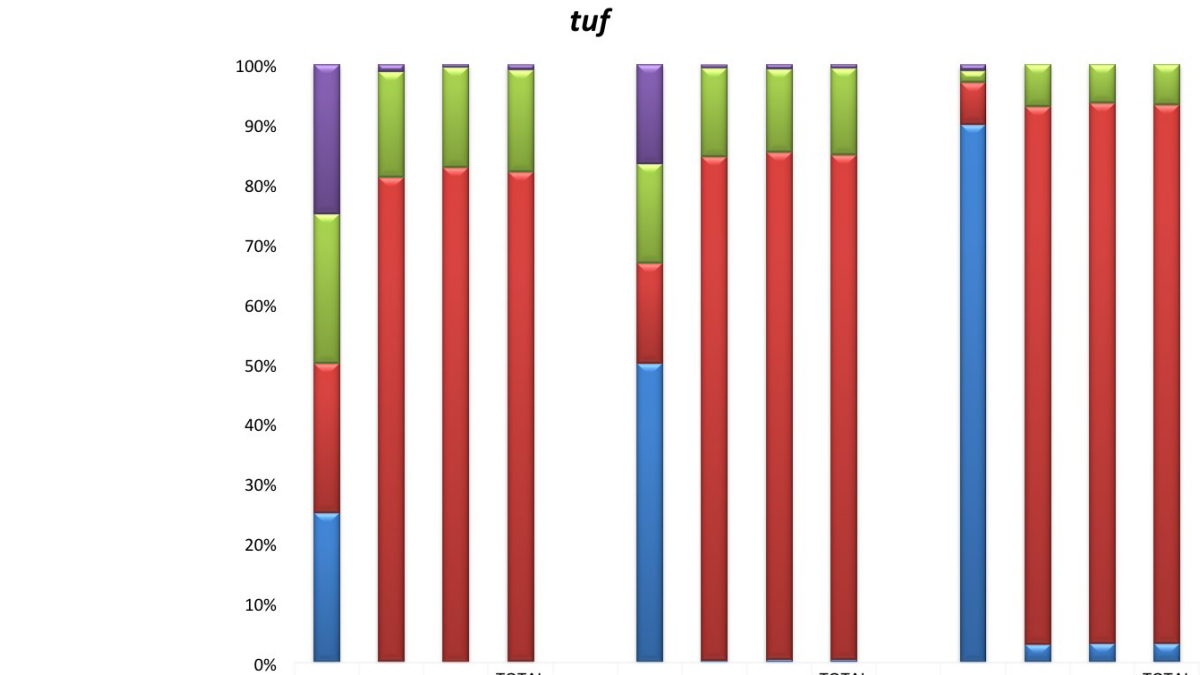


Figure 5.13. Neighbour-joining Phylogenetic Tree Showing the Taxonomic Assignations of the OTUs found by *tuf* in the Constructed Test samples and in the Oropharyngeal Swab DNAs. The arrows show how the main four OTUs encountered in the constructed samples are accurately aligned with *S. pneumoniae, S. mitis, S. pyogenes* and *S.agalactiae* respectively. Scale bar is shown and the branches with more than 70% of bootstrap level are labelled.

With regard to the relative abundance of OTUs, *tuf* encountered 2 OTUs of *S. pneumoniae* and 3 OTUs of *S.agalactiae* that were pooled to compare with the expected proportions of streptocci strains. This gene compared with the prior *map* and *pfl* could not assign accurately single OTUs from a strain but encountered inexistent differences and therefore more OTUs than the existent.

Also *tuf* overrepresented the abundance of *S. mitis* and underestimated the abundance of the rest of species (*P*<0.001), particularly *S.pyogenes* that to be identified by *tuf* had to be present in at least 90% of the sample (sample 3) (Figure 5.14).



	PLAN	1A	1B	TOTAL 1	PLAN	2A	2B	TOTAL 2	PLAN	3A	3B	TOTAL 3
Streptococcus pneumoniae 2 OTUs	25.00%	1.25%	0.53%	0.88%	17%	0.29%	0.78%	0.67%	1%	0.00%	0.00%	0.00%
Streptococcus agalactiae 3 OTUs	25.00%	17.57%	16.71%	17.12%	17%	7.61%	13.96%	14.43%	2%	7.05%	6.40%	6.78%
Streptococcus mitis	25.00%	80.93%	82.71%	81.85%	17%	43.07%	84.68%	84.43%	7%	89.62%	90.25%	89.88%
Streptococcus pyogenes	25.00%	0.25%	0.06%	0.15%	50%	0.19%	0.58%	0.48%	90%	3.06%	3.28%	3.15%

Figure 5.14. Relative Abundance Results of *tuf* in the constructed samples and the Comparison with the Expected Results (PLAN). Samples 1, 2 and 3 have duplicates A and B, which abundances were averaged on TOTAL SAMPLE.

The sequencing analysis of the oropharyngeal DNA samples (Samples 4 and 5) identified streptococci species similar to what was anticipated (although not possible to definitely determine) by 16S rRNA sequencing for example *S. vestibularis, S. infantis, S.peroris* and *S. salivarius.* For both samples however, the relative abundance of sequences could not be

contrasted with the obtained by16S rRNA sequencing because OTUs abundances could not be extrapolated (Figure 5.15).

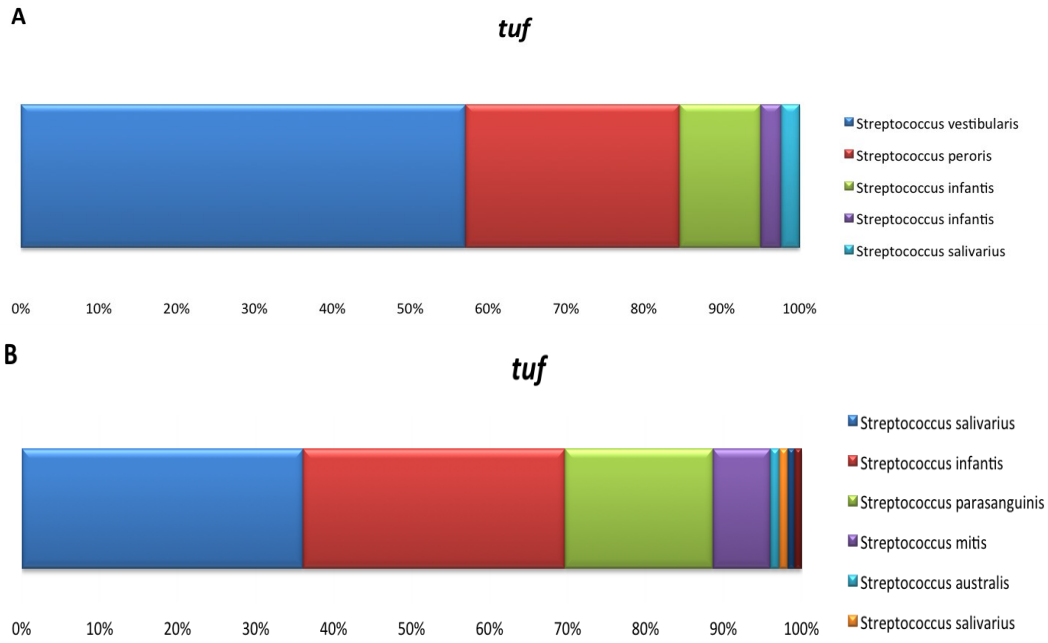


Figure 5.15. Relative Abundance Determined using the *tuf* Gene for the Two Oropharyngeal DNA samples Previously Analysed using 16S rRNA Pyrosequencing. A) Sample 4 and B) Sample 5.

5.1.5 Selection of a Primer Set

Due to the high accuracy and reliability of the *map* gene for both the taxonomic identification of the streptococci as well as the relative abundance measurements, the gene was selected to take forward to perform the streptococci species identification for the oropharyngeal samples of the ECUAVIDA infants' cohort. The other genes *pfl* and *tuf* were not pursued further in light of neither being as robust as the *map* gene particularly when it came to taxonomic identification and to some extent relative abundance.

5.2 ECUAVIDA Microbiota Study: *Streptococcus* species Analysis using *map* Pyrosequencing

5.2.1 Materials and Methods

From the cases/controls study of wheezing in infants from the rural tropics of Ecuador, a total of 76 samples were selected from Stage 1b (from 172 samples) for the streptococcal species characterization by *map* (following the protocols optimized in Section 5.1). The rest of samples from the study (Stages 1a and 2) did not have enough DNA to perform the streptococcus analysis (122 samples) or did not produce amplification (136 samples) (mainly because some samples were frozen at least 4 years prior the experiments). In the streptococci species typing using *map* infants included had a history of wheezing that had been sampled between 7 and 12 months of age. A total of 37 cases and 39 controls were studied. Amplification and sequencing was performed as previously described using the Roche 454 Junior Platform, and a total of 2 runs were performed (Section 5.1.2 and Chapter 2: Sections 2.7.1 and 2.7.4).

The sequences were filtered with ampliconnoise and analysed with QIIME 1.7 (Caporaso *et al.*, 2010b) (Chapter 2: Sections 2.7.5). The OTUs were picked at 80% similarity range and OTU taxonomic assignation was performed manually using the *map* streptococci database from www.emlsa.org (Bishop *et al.*, 2009) as well as the *map* sequences uploaded on the NCBI Genebank. The alignments were performed on Mega 6 using the ClustalW algorithm (Tamura *et al.*, 2013b). The alpha and beta diversity calculations were performed on the R package phyloseq (Dixon, 2003, Pages *et al.,* McMurdie and Holmes, 2013). Statistical analysis of beta diversity was

performed using UNIFRAC statistics (Lozupone and Knight, 2005a) and OTUs comparisons on Metastats (White *et al.*, 2009).

5.2.1 Results

From a total of 138,000 sequences after annoising and singletons removal 136,000 sequences remained. Although a total of 51 OTUs were five OTUs accounted for 95% of the 136,000 sequences with the remaining 46 OTUs being less common. The taxonomy assignation using the Neighbour-joining phylogeny determined that the most common OTU corresponded to *S. mitis* (OTU 5), the second most abundant (OTU 32) was assigned to *S.salivarius,* with the third and fourth (OTUs 45 and 18) assigned as *S.parasanguinis* and the fifth more abundant (OTU 22) was *S. peroris* (Figure 5.16). For the less common OTUs *S.pneumoniae* was assigned once but only seen in 2 samples and with less than 5 sequences each. For all the rest of less common OTUs, the taxonomy assignation as species level was equally straightforward. Since the rarity of these OTUs, they were dropped from subsequent analysis, which only was focused on the five main OTUs IDs: 5, 18, 22, 32 and 45.

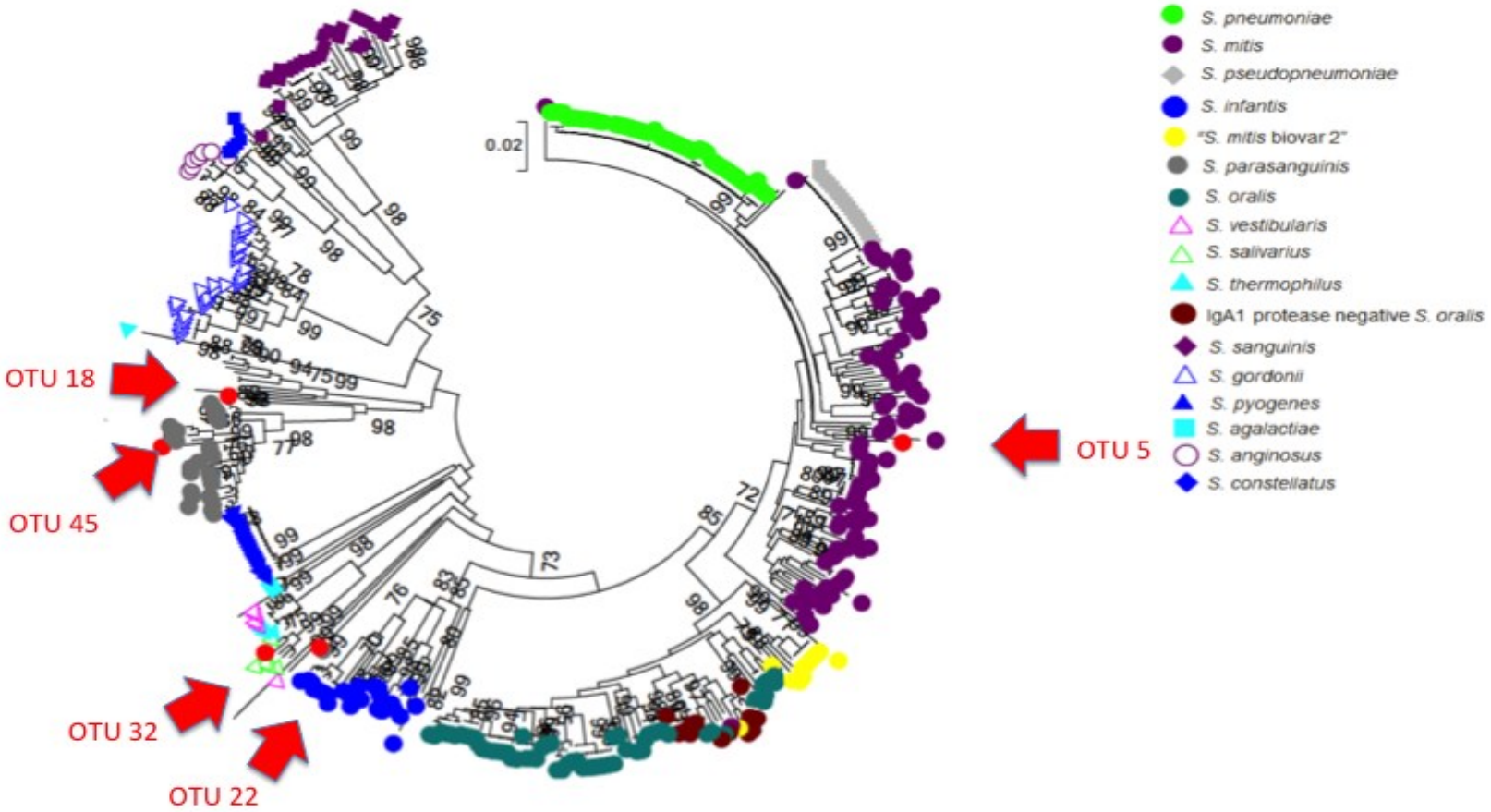


Figure 5.16. Neighbour-joining Phylogenetic Tree Showing the Taxonomic assignations by the *map* Gene for the Stage 1b Oropharyngeal DNA Samples. The arrows show how the main five OTUs encountered in the sample set are aligned allowing taxonomic identification to be determined. Scale bar is shown and the branches with more than 70% of bootstrap level are labelled.

5.2.1.1 OTUs Analysis

Previously the 16S rRNA sequencing had found that one OTU (OTU1328) assigned to *Streptococcus* spp. (with similar distribution of sequences per sample) was consistently more common in controls (P<0.0001) (Chapter 3: Section 4.2.2). Thanks to this new technique it was possible to identify at species level the OTUs assigned to streptococci, as consequence OTU1328 in 16SrRNA could be inferred as OTU5 in *map* analysis and therefore assigned t*o S.mitis.* When comparing at OTUs level between cases and controls, there were encountered statistical differences in this particular OTU *(P*=0.045), which was more commonly found in controls (Figure 5.17).

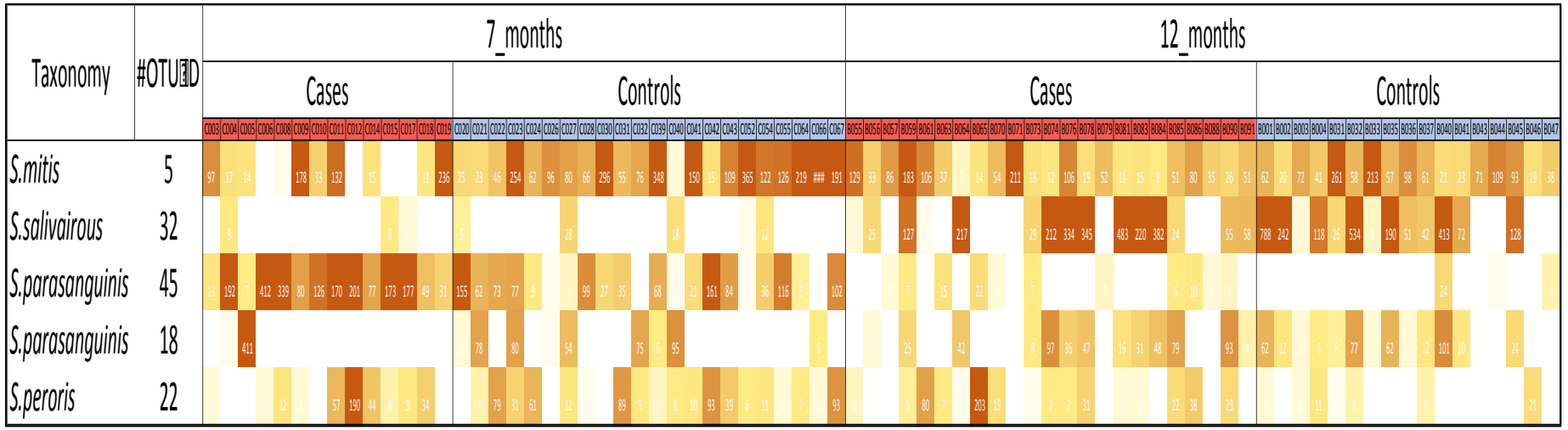


Figure 5.17. Heatmap Showing OTU Abundance per Sample for the Five Main OTUs Identified by *map* Gene Pyrosequencing. Each box ranges from white complete absence of sequences in that OTU, to yellow with few and dark red with high number of sequences.

Using metastats doing comparisons between age groups, at 7 months *S. parasinguinis* OTU 45 was more common (*P*<0.0001) as was *S. peroris* OTU 22 (*P*<0.001), whilst at 12 months there was a higher abundance of *S.mitis* OTU 5 (*P*=0.042), *S. salivarius* OTU 32 (*P*<0.0001), and *S. parasanguinis* OTU 18 (*P*=0.0006).

5.2.1.2 Alpha Diversity

Alpha diversity analysis was performed calculating richness, evenness and diversity (using the Shannon Index and the Inverse Simpson Index). To determine the normal distribution of the results, the Shapiro-Wilkinson normality test was performed. It showed that richness (*P*=0.002) and inverse Simpson (*P*=0.0006) indexes were not normally distributed and therefore non-parametric statistics were performed. On contrary Shannon index of diversity (*P*=0.154) and evenness (*P*=0.054) showed normal distribution and therefore parametric statistics were used.

Statistical analysis of the streptococci data revealed no differences in alpha diversity indexes between cases and controls, irrespective of whether age, 7 and 12 months, was taken into consideration.

5.2.1.3 Beta Diversity

When compare between cases versus controls at OTUs level, there were not found statistical differences using Canberra, Bray-Curtis, UNIFRAC or Weighted UNIFRAC. NMDS plots do not show a clear clustering when comparing cases versus controls. However when compare between ages (7 versus 12 months), there were statistical differences in Canberra (*P*<0.0001), Bray-Curtis (*P*<0.0001), UNIFRAC (*P*<0.0001), and Weighted UNIFRAC

(*P*<0.0001). However the age explains a different proportion of the variance each time: Canberra = 8%, Bray-Curtis 16%, UNIFRAC = 13%, weighted UNIFRAC = 18%. It can be inferred that the abundance of the different OTUs is important in determining differences as weighted UNIFRAC and Bray-Curtis show higher values), and that the phylogenetic information is also relevant to show differences (weighted UNIFRAC versus Bray-Curtis). Though Bray-Curtis has a reasonable level of stress and is a good representation of clustering in 2 dimensions. NMDS plots show a clear clustering of samples when grouped by age (Figure 5.18).

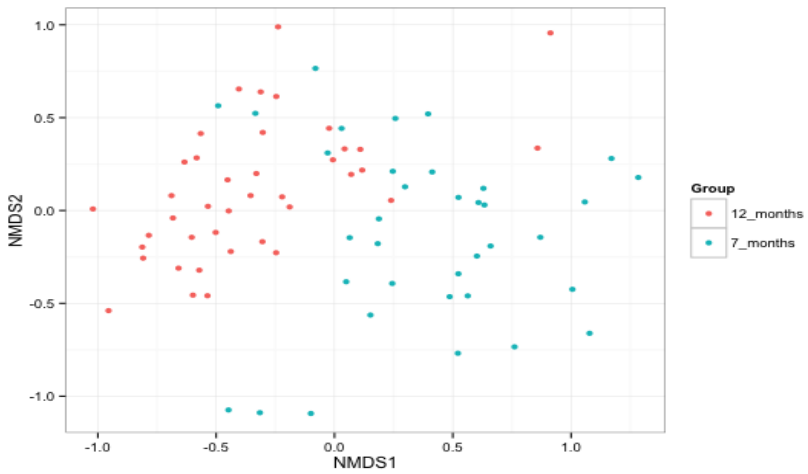


Figure 5.18. Beta Diversity NMDS Plots: Clustering between Samples. Bray-Curtis analysis between ages at 7 months versus 12 months.

5.3 Discussion

In order to complement the ECUAVIDA analysis of the human microbiota using 16S rRNA pyrosequencing, a new tool was designed for characterization of sub-populations of bacteria (streptococci) at species level. The tool designed was also capable of calculating relative abundance measures comparable to that obtained with the 16S rRNA sequencing. It was necessary to investigate several candidate genes commonly utilized by MLSA and MLST methodologies (Bishop *et al.*, 2009, Picard *et al.*, 2004) in order to find a reliable candidate gene for the pyrosequencing assay. After PCR standardization, validation and sequencing tests, the methionine aminopeptidase gene (*map*) was the best candidate differentiating streptococci species and providing a good insight of the sequences abundance.

The candidate genes selection and the primers design were based on the availability of broad databases www.emlsa.org, www.mlst.net and the NCBI Genebank that included all known *Streptococcus* species including *S.mitis, S.pyogenes, S.pneumoniae, S.agalactiae,* etc. (Section 5.1.1.1). Each set of primers was aligned with the streptococci sequences present within the databases and phylogenetically tested to determine how accurate they were for taxonomy assignation. PCR optimization of parameters and the validation process implemented (Section 5.1.1.2 to 5.1.1.5) determined three candidate genes to be suitable for further investigation *map, pfl* and *tuf.* The sequencing pilot study using constructed samples (with known concentrations of streptococci strains) and with real oropharyngeal bacterial DNA determined that *map* was the best candidate to study the streptococcal microbiota on the airways. *Map* provides accurate taxonomic assignation and relative abundance results comparable with that observed for 16S rRNA sequencing.

The analysis of the *Streptococcus* microbiota of the ECUAVIDA Stage 1b samples determined that infants have many different streptococcal species. Fifty-one OTUs in total were observed with 95% of sequences being assigned to 5 OTUs. The most common species were *S. mitis (*subtype mitis*), S.salivarious, S.parasanguinis* and *S. peroris*, which are members of the *Streptococcus* 'mitis' group. The ECUAVIDA infants had surprisingly a very low prevalence of *S.pneumoniae,* an important pathogen, which was seen only in 2 samples.

Comparing cases and controls at 7 and at 12 months, a higher abundance of *S. mitis* at both ages was seen for controls in controls. These bacteria have been related with protection against the development of pathogens due to their ability to secrete hydrogen peroxide which inhibits the growth of Gram negative bacteria (Tano *et al.*, 2000). No differences were seen for alpha and beta diversity measures when wheezing infants were compared to healthy controls irrespective of age.

When comparisons between age groups, 7 and 12 months, were done marked differences were observed. Some species like *S.peroris* were more common at 7 months whilst at 12 months addidtional species were acquired including *S.mitis, salivarious* and *S.parasanguinis.*

In conclusion a new culture-independent pyrosequencing based tool that is robust in assessing the streptococcal communities of the upper airways has been successfully designed and optimised. This is an important advancement as it now allows characterization of streptococci at species level which is not feasible and a limitation of current 16S rRNA sequencing. The application of this new tool to the ECUAVIDA microbiota study has revealed that *S. mitis* is potentially an important protective factor against the development of airways chronic inflammation and that the microbiota of the

upper airways becomes more complex with age involving a higher abundance of streptococci species.

Chapter 6: General Discussion

Asthma is a chronic disease of the airways characterized by a dysfunctional mucosa, intermittent airway inflammation and symptoms of wheezing, dyspnoea and cough. The syndrome results from a complex interplay between genetic and environmental factors (Cookson, 2004, Cookson, 2002, Cookson *et al.*, 2011) (Chapter 1). Several microbiological studies have determined the important effects of bacteria in asthma recurrences (Kraft, 2000) as well as in the development of asthma. A rich microbial environment in early life has been shown to be protective against the development of asthma (Eder *et al.*, 2006, Ege *et al.*, 2011). The airways of the lung have been shown to contain a characteristic flora that in asthma is disturbed by the presence of pathogens (Hilty *et al.*, 2010) whilst birth cohort studies, using standard culture methods and throat swabs, have found the presence of some pathogens to be predictive of the subsequent development of asthma (Bisgaard *et al.*, 2007).

Laboratory microbiological culture techniques do not however always result in successful bacterial identification and around 90% of bacteria cannot be grown by standard routine culture (Turnbaugh *et al.*, 2007). Consequently, the most powerful approach to study the microbial diversity in a community is to implement molecular culture-independent techniques (Turnbaugh *et al.*, 2007, Cox *et al.*, 2013a) employing key bacterial genes to enable bacterial identification (Chapter 1: Section 12). In the work presented in this thesis, this approach was used focusing on the 16S rRNA gene in the first instance and subsequently performing subtyping of the Streptococci seen using the *map* gene (Chapter 5). The use of state of the art bioinformatic tools in the data analysis was also essential in determining microbiota

patterns and to look for statistical differences between groups (Morgan and Huttenhower, 2012).

The focus of this thesis has been the examination of the respiratory microbiome in a set of infants selected from the ECUAVIDA case/control retrospective study of infants from the rural tropics of Ecuador. In this region of Ecuador general asthma prevalence is low (Mallol *et al.*, 2010) but the wheezing syndrome occurrence is comparable with that seen in the urban populations (Cooper *et al.*, 2014) (Chapter 1: Section 1.4). The ECUAVIDA population characteristically has very low history of antibiotic use (a key inclusion criterion of the study) and the selected infants had never taken inhaled corticosteroids. Both antibiotics and corticosteroids could be important biases in determining if the microbiota acts as an aetiological factor of airways inflammation or is a consequence of the medication use.

Furthermore, there have been to date no studies that have characterized the development of the airways core microbiome in early infancy. Additionally no studies of the airway microbiome using culture independent techniques have been done in Latin American countries where environmental exposures and disease incidence is very different to that seen in Europe and North America. According to the hygiene hypothesis, factors in the rural tropics of Ecuador including the rural environment, low antibiotic usage and the high rate of parasitic infections, would be protective against the development of asthma and atopic disease (Strachan, 1989, Cooper *et al.*, 2014).

The study conducted in this thesis was divided into several stages according to age (7, 12 and 24 months) and wheezing status (current wheezing, history of

wheezing, or healthy controls with no wheezing history). During sampling strict inclusion criteria were implemented in order to avoid skewed results in the microbiota patterns caused by for example current infectious processes, recent use of antibiotics or incorrect oropharyngeal swabbing techniques.

Considering all infants, irrespective of case/control status, overall the most common bacteria seen in the upper airways was *Streptococcus* spp. and their phylum Firmicutes was the most diverse accounting for 70% of sequences obtained when compared with the other phyla (Chapters 3 and 4). These results were expected based on findings reported by previous microbiome culture-independent studies (Hilty *et al.*, 2010).

When the bacterial microbiota was compared between infants with current wheezing and healthy controls (Stage 1a - ages between 7 and 12 months of age) a higher prevalence of potential pathogenic bacteria, principally *Neisseria* spp. and *Haemophilus influenzae,* was found in cases (Chapter 4: Section 4.2.1). Both *Neisseria* spp. and *Haemophilus influenzae* have been associated with serious infections in childhood (Gkentzi *et al.*, 2012, Watt *et al.*, 2009, Sabra and Benger, 2011). They have also been found to be related with asthmatics airways in adults (Hilty *et al.*, 2010) and in children, the latter using conventional culturing techniques (Bisgaard *et al.*, 2007). *Neisseria meningitidis*, also known as Meningococcus, is a main cause of morbidity and mortality during infancy in developed and developing countries (Khatami and Pollard, 2010). It is present in the airways in adult healthy carriers in a prevalence of 10 to 35% (Caugant and Maiden, 2009), and causes meningitis, otitis, pneumonia, bacteraemia, cerebral abscesses. Its mortality rate can reach up to 10%, and the reason why sometimes it acts as a harmless commensal and at other times as a highly virulent bacterium is not as yet understood (Khatami

and Pollard, 2010). *Haemophilus parainfluenzae* is a Gram negative rod shaped bacteria found in the oral cavity and the pharynx of humans. Usually it has a saprophytic behaviour but occasionally can cause disease especially bacteraemia and endocarditis, and less commonly meningitis, otitis, pneumonia (Nwaohiri *et al.*, 2009). Interestingly other potentially pathogenic bacteria have also been found in current wheezers like Clostrodiales, Burkholderiales, Pasteurellacea, which are enviromental bacteria that act as oportunistic pathogens (Murphy *et al.*, 2009, Corti *et al.*, 2009).

In this present study, when compared with controls current wheezers from Stage 1a had a lower abundance of the so-called 'protective' bacteria, that have previously been reported, including *Veillonella, Prevotella* and *Gemella.* These genera have been observed to be more common in healthy subjects (Hilty *et al.*, 2010), and interestingly it has been shown that *Prevotella* has the ability to inhibit the growth of other bacteria in the oropharynx (Murray and Rosenblatt, 1976).

To determine the mid and long term changes in the airways microbiome as a consequence of recurrent wheezing, infants with a history of wheezing were compared with healthy controls at 7, 12 and 24 months. This paired analysis identified distinct different microbial patterns with notable increases in potential pathogenic bacteria, *Haemophilus, Moraxella, Corynebacterium* and *Burkholderia*, in infants with a history of wheezing (Chapter 4: Section 4.2.2). As for current wheezers, past wheezers were also found to have a lower abundance of 'protective' bacteria including *Prevotella, Veillonella* and *Lactobacillus.* Additionally an OTU assigned to *Streptococcus* was consistently more common in controls however due to16S rRNA uniformity in this genus it was not possible to identify the species from this experiment. The differences found between the cases and controls were

maintained with increase in age and still present at 24 months (2 years) of age. Infants with a history of early wheezing (commencing at less than 7 months of age) have very apparent differences in both diversity and abundance differences compared with corresponding controls, as result a more diverse microbiome at this time appears to be protective against the development of wheezing.

Due to the structure of the ECUAVIDA study, it was possible to examine for the first time how the upper airways microbiota changes during infancy. The key observation was that over time a more complex and diverse flora is acquired (Chapter 4: Section 4.2.2.3). The main OTUs remain constant from 7 months up to and including 24 months (2 years), however other minor OTUs begin to appear through the time period. Consequently the number of species and the other alpha diversity parameters were found to increase with age (Chapter 4: Section 4.2.2.3.1).

Differences at the OTU level were also seen when comparing current versus past wheezers (Stage 1a versus Stage 1b). The wheezing episode (current) was characterized by a higher presence of potentially pathogenic bacteria compared with controls the differences remain present at mid and long-term (Chapter 4: Section 4.2.1). The same potentially pathogenic bacteria present in current wheezers i.e. *Haemophilus, Neisseria, Moraxella, Corynebacterium* and *Burkholderia*, remained present in past wheezers. Correspondingly the lack of protective bacteria was maintained over time. There was a higher abundance of pathogens and a lower abundance of protective bacteria in current wheezers compared with infants with a history of wheezing (Chapter 4: Section 4.2.1). Additionally, analysis and comparison of the epidemiological data for cases and controls at all stages was conducted and was found not to underpin the microbiota patterns observed.

As discussed earlier (Chapter 1) 16S rRNA gene sequencing is unable to discriminate at species level within the streptococci genus. Consequently it is incapable of distinguishing *S. pneumoniae* from the *Streptococcus mitis* ´group´ bacteria. *S. mitis* ´group´ are commensal bacteria of the airways and the upper intestinal tract. The *S.* mitis ´group´ encompasses different species of the *Streptococcus viridians* group *such* as: *S. mitis* (subtype *mitis*), *S. gordonii, S. pneumoniae, S. oralis, S. sanguis* and *S. parasanguis* (Kawamura *et al.*, 1995). Since the results from Chapters 3 and 4 had highlighted the high levels of streptococci present in the upper airways of the ECUAVIDA infants, to complement the 16S rRNA analysis a new tool to allow subtyping the streptococcal populations at species level was designed and established (Chapter 5).

Using pyrosequencing and variability in the bacterial *map* gene, the tool enables one to classify taxonomically, as well as distinguish, several species of streptococci present in a mixed microbial sample. Additionally relative abundance results obtained for each streptococci present can be correlated with 16S rRNA pyrosequencing results. The latter is particularly important as it allows one to establish how important each S*treptococcus* species is in the overall microbial community.

The results of the *map* sequencing for the seventy-six ECUAVIDA infant oropharyngeal samples revealed that S*treptococcus mitis* was consistently more commonly found in healthy subjects compared with current wheezers or infants with history of wheezing (Chapter 5: Section 5.2). S*. mitis* are alpha hemolytic Gram-positive cocci, facultative anaerobes and do not form spores. The *S. mitis* ´group´ bacteria are not pathogenic in normal circumstances (Coykendall, 1989) although they are an important cause of bacterial endocarditis and bacteraemia especially

after orthodontic treatments (Matsui and Jin, 2014). There are several *in vivo* and *in vitro* studies where *Streptococcus* spp. has been catalogued as: “inhabitants of the oropharynx that suppress the growth of Gram-positive and Gram-negative microorganisms” (Sachs *et al.*, 1993). Studies have hypothesised that in the upper airways, *S. viridans* could act as protective bacteria replacing possible pathogenic bacteria (Tano *et al.*, 2000). There are several mechanisms postulated as the molecular reason behind this inhibition, the most important of which is the production of hydrogen peroxide by *S. viridians,* which would restrict the growth of pathogenic bacteria (Tano *et al.*, 2003).

In the ECUAVIDA infants, both cases and controls, the prevalence of *Streptococcus pneumoniae* was found to be notably low (Chapter 5: Section 5.2). This is of particular interest because *S. pneumoniae* (also known as Pneumococcus) is categorized as part of the ‘mitis’ group using the 16S rRNA gene sequencing approach (Johnson and Tunkel, 2000). Pneumococcus is also part of the human normal upper respiratory tract microbiome but compared with *S. viridans* is more prevalent in causing disease. *S. pneumoniae* commonly causes pneumonia, sinusitis, otitis media and meningitis (secondary to one of the previous infections). Less common pathologies, secondary to *S. pneumoniae* infection, are cellulitis, peritonitis, osteomyelitis, septic arthritis and endocarditis especially in immunocompromised patients (Krzysciak *et al.*, 2013).

In the national regulations of the Ecuadorian Health System, pneumococcus vaccination is not considered compulsory and therefore is only provided to selected populations at risk. In this current study none of the infants had received the pneumococcus vaccination. The new streptococci typing tool also confirmed the prior 16S rRNA results (Chapter 4: Section 4.2.2) that the abundance and diversity of the

streptococci group of bacteria increases over time and they contribute to the formation of a more complex respiratory microbiome.

In conclusion the upper airways bacterial microbiota is different between infants with wheezing versus healthy age matched controls. There is a marked increase of potentially pathogenic bacteria in the wheezing infants and a lower abundance of the previously reported ‘protective’ bacteria. Such differences are of relevance to the pathophysiology of asthma and its development later in life, and also because could potentially provide a new therapeutic angle based in the bacterial microbiota modification. It is important to mention that the statistical software used has been refined over time to reflect the data structure, however the *p* values quoted in this thesis will need correction before publication in final form.

3.1. Future Work

The current study identifies the microbiota patterns related with the development of early onset wheezing on infants from the rural Tropics of Ecuador. In light of the findings it would be interesting to investigate further the microbiome structure in an extended part of the ECUAVIDA cohort (that in total has more than 2,500 children). As oropharyngeal samples have been continued to be collected since initiation of this thesis project as infants have been recruited into the birth-cohort, the project could retrospectively analyse such samples (collected during current wheezing and in past wheezers). Additionally as the project plans to continue prospectively until children will reach the 8 years old, there is an opportunity to continue to study the development of the upper airways microbiota over time and as children develop asthma (not only wheezing).

Another complementary study to be performed would be the relationships of the gut microbiota and the airways microbiota at different time points in development. As the mouth is the entrance for both the digestive and respiratory systems it is likely that a close relationship exists, and that dysbiotic changes in one body-system microbiome may influence the other. In the ECUAVIDA cohort where antibiotic intake is low overall hence limited skewing of results, the relationship between both systems in health and disease could be analysed as well as their influence in the development of asthma and allergies.

Another goal of the ECUAVIDA cohort is to determine the environmental influences in the development of asthma and allergies, and an important part of the study is the 'protective' effect conferred by geohelminth infection. The effect of these infections in the gut microbiome has been reported in a pilot study (Cooper *et al.*,

2013) with a reduced number of samples, however it would be interesting to analyse the gut and airways microbiome patterns related with wheezing and asthma in an extended study and the effect of anti-parasitic therapies on this.

Technologically the study could be further continued using the new generation sequencing technologies like the Illumina HiSeq and MiSeq (released in the last months of 2013), which would allow a deeper analysis of the microbiota with increased number of sequences per sample, and the possibility of longer amplicon sequencing (which could provide a better taxonomical identification of bacterial OTUs). The development of metagenomics experiments for a better microbiota analysis would also be important allowing not only more even more accurate characterization of the bacteria present but also their activity and the activity of the host. Consequently consideration of the use of metabolomics and proteomic platforms could also be key in enabling the molecular relationship between microbiome and the epithelial and immune cells in the airways to be determined.

References

ABEL NOUR, S. M., MEHTA, J. B., ROY, T. M. & BYRD, R. P., JR. 2013. Veillonella as a cause for septic pulmonary emboli. *Tenn Med,* 106**,** 39-40.

AFSHAR, R., MEDOFF, B. D. & LUSTER, A. D. 2008. Allergic asthma: a tale of many T cells. *Clin Exp Allergy,* 38**,** 1847-57.

AHMED, S., MACFARLANE, G. T., FITE, A., MCBAIN, A. J., GILBERT, P. & MACFARLANE, S. 2007. Mucosa-associated bacterial diversity in relation to human terminal ileum and colonic biopsy samples. *Appl Environ Microbiol,* 73**,** 7435-42.

ALLAN, K. & DEVEREUX, G. 2011. Diet and Asthma: Nutrition Implications from Prevention to Treatment. *Journal of the American Dietetic Association,* 111**,** 258-268.

ALLEN-VERCOE, E. & JOBIN, C. 2014. Fusobacterium and Enterobacteriaceae: Important players for CRC? *Immunol Lett*.

AOSHIBA, K. & NAGAI, A. 2004. Differences in airway remodeling between asthma and chronic obstructive pulmonary disease. *Clin Rev Allergy Immunol,* 27**,** 35-43.

BATEMAN, E. D. & JITHOO, A. 2007. Asthma and allergy - a global perspective. *Allergy,* 62**,** 213-5.

BELSTROM, D., HOLMSTRUP, P., NIELSEN, C. H., KIRKBY, N., TWETMAN, S., HEITMANN, B. L., KLEPAC-CERAJ, V., PASTER, B. J. & FIEHN, N. E. 2014. Bacterial profiles of saliva in relation to diet, lifestyle factors, and socioeconomic status. *J Oral Microbiol,* 6.

BENICIO, M. H., FERREIRA, M. U., CARDOSO, M. R., KONNO, S. C. & MONTEIRO, C. A. 2004. Wheezing conditions in early childhood: prevalence and risk factors in the city of Sao Paulo, Brazil. *Bull World Health Organ,* 82**,** 516-22.

BIESBROEK, G., BOSCH, A. A., WANG, X., KEIJSER, B. J., VEENHOVEN, R. H., SANDERS, E. A. & BOGAERT, D. 2014. The Impact of Breastfeeding on Nasopharyngeal Microbial Communities in Infants. *Am J Respir Crit Care Med*.

BINIA, A., KHORASANI, N., BHAVSAR, P. K., ADCOCK, I., BRIGHTLING, C. E., CHUNG, K. F., COOKSON, W. O. & MOFFATT, M. F. 2011. Chromosome 17q21 SNP and severe asthma. *J Hum Genet,* 56**,** 97-8.

BISGAARD, H., HERMANSEN, M. N., BUCHVALD, F., LOLAND, L., HALKJAER, L. B., BONNELYKKE, K., BRASHOLT, M., HELTBERG, A., VISSING, N. H., THORSEN, S. V., STAGE, M. & PIPPER, C. B. 2007. Childhood asthma after bacterial colonization of the airway in neonates. *N Engl J Med,* 357**,** 1487-95.

BISHOP, C. J., AANENSEN, D. M., JORDAN, G. E., KILIAN, M., HANAGE, W. P. & SPRATT, B. G. 2009. Assigning strains to bacterial species via the internet. *BMC Biol,* 7**,** 3.

BOCHUD, P. Y., CALANDRA, T. & FRANCIOLI, P. 1994. Bacteremia due to viridans streptococci in neutropenic patients: a review. *Am J Med,* 97**,** 256-64.

BOGAERT, D., KEIJSER, B., HUSE, S., ROSSEN, J., VEENHOVEN, R., VAN GILS, E., BRUIN, J., MONTIJN, R., BONTEN, M. & SANDERS, E. 2011. Variability and Diversity of Nasopharyngeal Microbiota in Children: A Metagenomic Analysis. *PLoS ONE,* 6**,** e17035.

BOUSQUET, J., CLARK, T. J., HURD, S., KHALTAEV, N., LENFANT, C., O'BYRNE, P. & SHEFFER, A. 2007. GINA guidelines on asthma and beyond. *Allergy,* 62**,** 102-12.

BOWYER, P., FRACZEK, M. & DENNING, D. W. 2006. Comparative genomics of fungal allergens and epitopes shows widespread distribution of closely related allergen and epitope orthologues. *BMC Genomics,* 7**,** 251.

BRAR, T., NAGARAJ, S. & MOHAPATRA, S. 2012. Microbes and asthma: the missing cellular and molecular links. *Curr Opin Pulm Med,* 18**,** 14-22.

BROCKMAN, W., ALVAREZ, P., YOUNG, S., GARBER, M., GIANNOUKOS, G., LEE, W. L., RUSS, C., LANDER, E. S., NUSBAUM, C. & JAFFE, D. B. 2008. Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Research,* 18**,** 763-770.

BURNS, J. L. & ROLAIN, J. M. 2014. Culture-based diagnostic microbiology in cystic fibrosis: can we simplify the complexity? *J Cyst Fibros,* 13**,** 1-9.

BUSSE, W. W. & LEMANSKE, R. F., JR. 2001. Asthma. *N Engl J Med,* 344**,** 350-62.

BYARD, R. W. 2013. Diphtheria - 'The strangling angel' of children. *J Forensic Leg Med,* 20**,** 65-8.

CAMARGO, C. A., JR. 2013. Human rhinovirus, wheezing illness, and the primary prevention of childhood asthma. *Am J Respir Crit Care Med,* 188**,** 1281-2.

CAPORASO, J. G., BITTINGER, K., BUSHMAN, F. D., DESANTIS, T. Z., ANDERSEN, G. L. & KNIGHT, R. 2010a. PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics,* 26**,** 266-7.

CAPORASO, J. G., KUCZYNSKI, J., STOMBAUGH, J., BITTINGER, K., BUSHMAN, F. D., COSTELLO, E. K., FIERER, N., PENA, A. G., GOODRICH, J. K., GORDON, J. I., HUTTLEY, G. A., KELLEY, S. T., KNIGHTS, D., KOENIG, J. E., LEY, R. E., LOZUPONE, C. A., MCDONALD, D., MUEGGE, B. D., PIRRUNG, M., REEDER, J., SEVINSKY, J. R., TURNBAUGH, P. J., WALTERS, W. A., WIDMANN, J., YATSUNENKO, T., ZANEVELD, J. & KNIGHT, R. 2010b. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods,* 7**,** 335-6.

CARDENAS, P. A., COOPER, P. J., COX, M. J., CHICO, M., ARIAS, C., MOFFATT, M. F. & COOKSON, W. O. 2012. Upper Airways Microbiota in Antibiotic-Naïve Wheezing and Healthy Infants from the Tropics of Rural Ecuador. *PLoS ONE,* 7**,** e46803.

CASTRO, M., SCHWEIGER, T., YIN-DECLUE, H., RAMKUMAR, T. P., CHRISTIE, C., ZHENG, J., COHEN, R., SCHECHTMAN, K. B., STRUNK, R. & BACHARIER, L. B. 2008. Cytokine response after severe respiratory syncytial virus bronchiolitis in early life. *Journal of Allergy and Clinical Immunology,* 122**,** 726-733.e3.

CAUGANT, D. A. & MAIDEN, M. C. 2009. Meningococcal carriage and disease--population biology and evolution. *Vaccine,* 27 Suppl 2**,** B64-70.

CELEBI, S., SEVINIR, B., SARAYDAROGLU, O., GURPINAR, A. & HACIMUSTAFAOGLU, M. 2009. Pulmonary actinomycosis. *Indian J Pediatr,* 76**,** 236-8.

CHANG, Q., LUAN, Y. & SUN, F. 2011. Variance adjusted weighted UniFrac: a powerful beta diversity measure for comparing communities based on phylogeny. *BMC Bioinformatics,* 12**,** 118.

CHAO, A., CHIU, C. H. & HSIEH, T. C. 2012. Proposing a resolution to debates on diversity partitioning. *Ecology,* 93**,** 2037-51.

CHARLSON, E. S., BITTINGER, K., HAAS, A. R., FITZGERALD, A. S., FRANK, I., YADAV, A., BUSHMAN, F. D. & COLLMAN, R. G. 2011. Topographical Continuity of Bacterial Populations in the Healthy Human Respiratory Tract. *American Journal of Respiratory and Critical Care Medicine,* 184**,** 957-963.

CHARLSON, E. S., CHEN, J., CUSTERS-ALLEN, R., BITTINGER, K., LI, H., SINHA, R., HWANG, J., BUSHMAN, F. D. & COLLMAN, R. G. 2010. Disordered Microbial Communities in the Upper Respiratory Tract of Cigarette Smokers. *PLoS ONE,* 5**,** e15216.

CHIN, K. J., HAHN, D., HENGSTMANN, U., LIESACK, W. & JANSSEN, P. H. 1999. Characterization and identification of numerically abundant culturable bacteria from the anoxic bulk soil of rice paddy microcosms. *Appl Environ Microbiol,* 65**,** 5042-9.

COLE, J. R., CHAI, B., MARSH, T. L., FARRIS, R. J., WANG, Q., KULAM, S. A., CHANDRA, S., MCGARRELL, D. M., SCHMIDT, T. M., GARRITY, G. M. & TIEDJE, J. M. 2003. The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res,* 31**,** 442-3.

COLE, J. R., WANG, Q., CARDENAS, E., FISH, J., CHAI, B., FARRIS, R. J., KULAM-SYED-MOHIDEEN, A. S., MCGARRELL, D. M., MARSH, T., GARRITY, G. M. & TIEDJE, J. M. 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res,* 37**,** D141-5.

COLLADO, M. C., CALABUIG, M. & SANZ, Y. 2007. Differences between the fecal microbiota of coeliac infants and healthy controls. *Curr Issues Intest Microbiol,* 8**,** 9-14.

COOKSON, W. 2002. Genetics and genomics of asthma and allergic diseases. *Immunol Rev,* 190**,** 195-206.

COOKSON, W. 2004. The immunogenetics of asthma and eczema: a new focus on the epithelium. *Nat Rev Immunol,* 4**,** 978-88.

COOKSON, W., MOFFATT, M. & STRACHAN, D. P. 2011. Genetic risks and childhood-onset asthma. *J Allergy Clin Immunol,* 128**,** 266-70; quiz 271-2.

COOKSON, W. O. & MOFFATT, M. F. 2011. Genetics of complex airway disease. *Proc Am Thorac Soc,* 8**,** 149-53.

COOPER, P., WALKER, A. W., REYES, J., CHICO, M., SALTER, S. J., VACA, M. & PARKHILL, J. 2013. Patent human infections with the whipworm, Trichuris trichiura, are not associated with alterations in the faecal microbiota. *PLoS One,* 8**,** e76573.

COOPER, P. J., CHICO, M. E., BLAND, M., GRIFFIN, G. E. & NUTMAN, T. B. 2003. Allergic symptoms, atopy, and geohelminth infections in a rural area of Ecuador. *Am J Respir Crit Care Med,* 168**,** 313-7.

COOPER, P. J., CHICO, M. E., GUADALUPE, I., SANDOVAL, C. A., MITRE, E., PLATTS-MILLS, T. A., BARRETO, M. L., RODRIGUES, L. C., STRACHAN, D. P. & GRIFFIN, G. E. 2011. Impact of early life exposures to geohelminth infections on the development of vaccine immunity, allergic sensitization, and allergic inflammatory diseases in children living in tropical Ecuador: the ECUAVIDA birth cohort study. *BMC Infect Dis,* 11**,** 184.

COOPER, P. J., CHICO, M. E., VACA, M. G., RODRIGUEZ, A., ALCANTARA-NEVES, N. M., GENSER, B., DE CARVALHO, L. P., STEIN, R. T., CRUZ, A. A., RODRIGUES, L. C. & BARRETO, M. L. 2006. Risk factors for asthma and allergy associated with urban migration: background and methodology of a

cross-sectional study in Afro-Ecuadorian school children in Northeastern Ecuador (Esmeraldas-SCAALA Study). *BMC Pulm Med,* 6**,** 24.
COOPER, P. J., RODRIGUES, L. C. & BARRETO, M. L. 2012. Influence of poverty and infection on asthma in Latin America. *Curr Opin Allergy Clin Immunol,* 12**,** 171-8.
COOPER, P. J., RODRIGUES, L. C., CRUZ, A. A. & BARRETO, M. L. 2009. Asthma in Latin America: a public heath challenge and research opportunity. *Allergy,* 64**,** 5-17.
COOPER, P. J., VACA, M., RODRIGUEZ, A., CHICO, M. E., SANTOS, D. N., RODRIGUES, L. C. & BARRETO, M. L. 2014. Hygiene, atopy and wheeze-eczema-rhinitis symptoms in schoolchildren from urban and rural Ecuador. *Thorax,* 69**,** 232-9.
CORTI, M., PALMERO, D. & EIGUCHI, K. 2009. Respiratory infections in immunocompromised patients. *Curr Opin Pulm Med,* 15**,** 209-17.
COSTELLO, E. K., LAUBER, C. L., HAMADY, M., FIERER, N., GORDON, J. I. & KNIGHT, R. 2009. Bacterial Community Variation in Human Body Habitats Across Space and Time. *Science,* 326**,** 1694-1697.
COX, M. J., COOKSON, W. O. & MOFFATT, M. F. 2013a. Sequencing the human microbiome in health and disease. *Hum Mol Genet,* 22**,** R88-94.
COX, M. J., COOKSON, W. O. C. M. & MOFFATT, M. F. 2013b. Sequencing the human microbiome in health and disease. *Human Molecular Genetics,* 22**,** R88-R94.
COYKENDALL, A. L. 1989. Classification and identification of the viridans streptococci. *Clin Microbiol Rev,* 2**,** 315-28.
CROWE, C. C., SANDERS, W. E., JR. & LONGLEY, S. 1973. Bacterial interference. II. Role of the normal throat flora in prevention of colonization by group A Streptococcus. *J Infect Dis,* 128**,** 527-32.
CULLEY, F. J., POLLOTT, J. & OPENSHAW, P. J. 2002. Age at first viral infection determines the pattern of T cell-mediated disease during reinfection in adulthood. *J Exp Med,* 196**,** 1381-6.
DAVIS, C. 2014. Enumeration of probiotic strains: Review of culture-dependent and alternative techniques to quantify viable bacteria. *J Microbiol Methods,* 103**,** 9-17.
DESANTIS, T. Z., HUGENHOLTZ, P., LARSEN, N., ROJAS, M., BRODIE, E. L., KELLER, K., HUBER, T., DALEVI, D., HU, P. & ANDERSEN, G. L. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol,* 72**,** 5069-72.
DIARIO, E. 2014. Minimum Salary in 2014 has increased. *El Diario*.
DIXON, P. 2003. VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science,* 14**,** 927-930.
DRAY, S., DUFOUR, A. & CHESSEL, D. 2007. The ade4 Package--{II}: Two-table and ${K}$-table Methods. *R News,* 7**,** 47-52.
DROSTE, J. H., WIERINGA, M. H., WEYLER, J. J., NELEN, V. J., VERMEIRE, P. A. & VAN BEVER, H. P. 2000. Does the use of antibiotics in early childhood increase the risk of asthma and allergic disease? *Clin Exp Allergy,* 30**,** 1547-53.
DURRANT, C., ZONDERVAN, K. T., CARDON, L. R., HUNT, S., DELOUKAS, P. & MORRIS, A. P. 2004. Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *Am J Hum Genet,* 75**,** 35-43.

ECKBURG, P. B., BIK, E. M., BERNSTEIN, C. N., PURDOM, E., DETHLEFSEN, L., SARGENT, M., GILL, S. R., NELSON, K. E. & RELMAN, D. A. 2005. Diversity of the human intestinal microbial flora. *Science,* 308**,** 1635-8.

EDER, W., EGE, M. J. & VON MUTIUS, E. 2006. The asthma epidemic. *N Engl J Med,* 355**,** 2226-35.

EDGAR, R. C. 2010a. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics,* 26**,** 2460-1.

EDGAR, R. C. 2010b. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics,* 26**,** 2460-2461.

EDWARDS, K. M., ZHU, Y., GRIFFIN, M. R., WEINBERG, G. A., HALL, C. B., SZILAGYI, P. G., STAAT, M. A., IWANE, M., PRILL, M. M. & WILLIAMS, J. V. 2013. Burden of human metapneumovirus infection in young children. *N Engl J Med,* 368**,** 633-43.

EDWARDS, M. R., BARTLETT, N. W., HUSSELL, T., OPENSHAW, P. & JOHNSTON, S. L. 2012. The microbiology of asthma. *Nat Rev Micro,* 10**,** 459-471.

EGE, M. J., MAYER, M., NORMAND, A. C., GENUNEIT, J., COOKSON, W. O., BRAUN-FAHRLANDER, C., HEEDERIK, D., PIARROUX, R. & VON MUTIUS, E. 2011. Exposure to environmental microorganisms and childhood asthma. *N Engl J Med,* 364**,** 701-9.

EL AILA, N. A., EMLER, S., KAIJALAINEN, T., DE BAERE, T., SAERENS, B., ALKAN, E., DESCHAGHT, P., VERHELST, R. & VANEECHOUTTE, M. 2010. The development of a 16S rRNA gene based PCR for the identification of Streptococcus pneumoniae and comparison with four other species specific PCR assays. *BMC Infect Dis,* 10**,** 104.

ENGELMANN, I., MORDACQ, C., GOSSET, P., TILLIE-LEBLOND, I., DEWILDE, A., THUMERELLE, C., POUESSEL, G. & DESCHILDRE, A. 2013. Rhinovirus and Asthma: Reinfection, Not Persistence. *American Journal of Respiratory and Critical Care Medicine,* 188**,** 1165-1167.

ERB-DOWNWARD, J. R., THOMPSON, D. L., HAN, M. K., FREEMAN, C. M., MCCLOSKEY, L., SCHMIDT, L. A., YOUNG, V. B., TOEWS, G. B., CURTIS, J. L., SUNDARAM, B., MARTINEZ, F. J. & HUFFNAGLE, G. B. 2011. Analysis of the Lung Microbiome in the “Healthy” Smoker and in COPD. *PLoS ONE,* 6**,** e16384.

ERIBE, E. R. & OLSEN, I. 2008. Leptotrichia species in human infections. *Anaerobe,* 14**,** 131-7.

FIELD, D., TIWARI, B., BOOTH, T., HOUTEN, S., SWAN, D., BERTRAND, N. & THURSTON, M. 2006a. Open software for biologists: from famine to feast. *Nat Biotechnol,* 24**,** 801-3.

FIELD, D., TIWARI, B., BOOTH, T., HOUTEN, S., SWAN, D., BERTRAND, N. & THURSTON, M. 2006b. Open software for biologists: from famine to feast. *Nat Biotech,* 24**,** 801-803.

FITZGERALD, M., BATEMAN, E. D., HURD, S. S., BARNES, P. J., BOUSQUET, J., DRAZEN, J. M., GIBSON, P., OHTA, K., O'BYRNE, P., PEDERSEN, S. E., PIZZICHINI, E., SULLIVAN, S. D., WENZEL, S. E. & ZAR, H. J. 2013. Global strategy for asthma management and prevention: GINA report. *Global Initiative for asthma*.

FRANCAVILLA, R., ERCOLINI, D., PICCOLO, M., VANNINI, L., SIRAGUSA, S., DE FILIPPIS, F., DE PASQUALE, I., DI CAGNO, R., DI TOMA, M., GOZZI, G., SERRAZANETTI, D. I., DE ANGELIS, M. & GOBBETTI, M. 2014. Salivary

microbiota and metabolome associated with celiac disease. *Appl Environ Microbiol,* 80**,** 3416-25.
FRANK, D. N., FEAZEL, L. M., BESSESEN, M. T., PRICE, C. S., JANOFF, E. N. & PACE, N. R. 2010. The Human Nasal Microbiota and <italic>Staphylococcus aureus</italic> Carriage. *PLoS ONE,* 5**,** e10598.
FRANK, D. N., ST AMAND, A. L., FELDMAN, R. A., BOEDEKER, E. C., HARPAZ, N. & PACE, N. R. 2007a. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci U S A,* 104**,** 13780-5.
FRANK, D. N., ST. AMAND, A. L., FELDMAN, R. A., BOEDEKER, E. C., HARPAZ, N. & PACE, N. R. 2007b. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proceedings of the National Academy of Sciences,* 104**,** 13780-13785.
FRANK, J. A., REICH, C. I., SHARMA, S., WEISBAUM, J. S., WILSON, B. A. & OLSEN, G. J. 2008. Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes. *Appl Environ Microbiol,* 74**,** 2461-70.
FRIEDLANDER, S. L., JACKSON, D. J., GANGNON, R. E., EVANS, M. D., LI, Z., ROBERG, K. A., ANDERSON, E. L., CARLSON-DAKES, K. T., ADLER, K. J., GILBERTSON-WHITE, S., PAPPAS, T. E., DASILVA, D. F., TISLER, C. J., PLEISS, L. E., MIKUS, L. D., ROSENTHAL, L. A., SHULT, P. A., KIRK, C. J., REISDORF, E., HOFFJAN, S., GERN, J. E. & LEMANSKE, R. F., JR. 2005. Viral infections, cytokine dysregulation and the origins of childhood asthma and allergic diseases. *Pediatr Infect Dis J,* 24**,** S170-6, discussion S174-5.
GABRIEL, L. & BERIOT-MATHIOT, A. 2014. Hospitalization stay and costs attributable to Clostridium difficile infection: a critical review. *J Hosp Infect*.
GALLI, L., MONTAGNANI, C., CHIAPPINI, E. & DE MARTINO, M. 2013. Treating paediatric community-acquired pneumonia in the era of antimicrobial resistance. *Acta Paediatr Suppl,* 102**,** 25-33.
GARCIA-LECHUZ, J. M., CUEVAS-LOBATO, O., HERNANGOMEZ, S., HERMIDA, A., GUINEA, J., MARIN, M., PELAEZ, T. & BOUZA, E. 2002. Extra-abdominal infections due to Gemella species. *Int J Infect Dis,* 6**,** 78-82.
GERN, J. E., ROSENTHAL, L. A., SORKNESS, R. L. & LEMANSKE JR, R. F. 2005. Effects of viral respiratory infections on lung development and childhood asthma. *Journal of Allergy and Clinical Immunology,* 115**,** 668-674.
GIL, E., NOURSADEGHI, M. & BROWN, J. S. 2013. The clinical and ecological impact of childhood pneumococcal vaccination. *Br J Hosp Med (Lond),* 74**,** 212-6.
GKENTZI, D., SLACK, M. P. & LADHANI, S. N. 2012. The burden of nonencapsulated Haemophilus influenzae in children and potential for prevention. *Curr Opin Infect Dis,* 25**,** 266-72.
GOLIN, V., MIMICA, I. M. & MIMICA, L. M. 1998. Oropharynx microbiota among alcoholics and non-alcoholics. *Sao Paulo Med J,* 116**,** 1727-33.
GREEN, S. J., MICHEL, F. C., JR., HADAR, Y. & MINZ, D. 2004. Similarity of bacterial communities in sawdust- and straw-amended cow manure composts. *FEMS Microbiol Lett,* 233**,** 115-23.
GRICE, E. A., KONG, H. H., RENAUD, G., YOUNG, A. C., BOUFFARD, G. G., BLAKESLEY, R. W., WOLFSBERG, T. G., TURNER, M. L. & SEGRE, J. A. 2008. A diversity profile of the human skin microbiota. *Genome Res,* 18**,** 1043-50.

GUADALUPE, I., MITRE, E., BENITEZ, S., CHICO, M. E., NUTMAN, T. B. & COOPER, P. J. 2009. Evidence for in utero sensitization to Ascaris lumbricoides in newborns of mothers with ascariasis. *J Infect Dis,* 199**,** 1846-50.

GUPTA, K., MARTINELLO, R. A., YOUNG, M., STRYMISH, J., CHO, K. & LAWLER, E. 2013. MRSA Nasal Carriage Patterns and the Subsequent Risk of Conversion between Patterns, Infection, and Death. *PLoS ONE,* 8**,** e53674.

HALEY, K. J. & DRAZEN, J. M. 1998. Inflammation and airway function in asthma: what you see is not necessarily what you get. *Am J Respir Crit Care Med,* 157**,** 1-3.

HALL, B. G. 2013. Building Phylogenetic Trees from Molecular Data with MEGA. *Molecular Biology and Evolution,* 30**,** 1229-1235.

HANSEN, C. H., ANDERSEN, L. S., KRYCH, L., METZDORFF, S. B., HASSELBY, J. P., SKOV, S., NIELSEN, D. S., BUSCHARD, K., HANSEN, L. H. & HANSEN, A. K. 2014. Mode of Delivery Shapes Gut Colonization Pattern and Modulates Regulatory Immunity in Mice. *J Immunol*.

HAPFELMEIER, S., LAWSON, M. A. E., SLACK, E., KIRUNDI, J. K., STOEL, M., HEIKENWALDER, M., CAHENZLI, J., VELYKOREDKO, Y., BALMER, M. L., ENDT, K., GEUKING, M. B., CURTISS, R., MCCOY, K. D. & MACPHERSON, A. J. 2010. Reversible Microbial Colonization of Germ-Free Mice Reveals the Dynamics of IgA Immune Responses. *Science,* 328**,** 1705-1709.

HARMES, K. M., BLACKWOOD, R. A., BURROWS, H. L., COOKE, J. M., HARRISON, R. V. & PASSAMANI, P. P. 2013. Otitis media: diagnosis and treatment. *Am Fam Physician,* 88**,** 435-40.

HENGSTMANN, U., CHIN, K. J., JANSSEN, P. H. & LIESACK, W. 1999. Comparative phylogenetic assignment of environmental sequences of genes encoding 16S rRNA and numerically abundant culturable bacteria from an anoxic rice paddy soil. *Appl Environ Microbiol,* 65**,** 5050-8.

HENRICKSON, K. J. 2003. Parainfluenza viruses. *Clin Microbiol Rev,* 16**,** 242-64.

HENRICKSON, K. J. & SAVATSKI, L. L. 1997. Antigenic structure, function, and evolution of the hemagglutinin-neuraminidase protein of human parainfluenza virus type 1. *J Infect Dis,* 176**,** 867-75.

HERBST, T., SICHELSTIEL, A., SCHAR, C., YADAVA, K., BURKI, K., CAHENZLI, J., MCCOY, K., MARSLAND, B. J. & HARRIS, N. L. 2011. Dysregulation of allergic airway inflammation in the absence of microbial colonization. *Am J Respir Crit Care Med,* 184**,** 198-205.

HILTY, M., BURKE, C., PEDRO, H., CARDENAS, P., BUSH, A., BOSSLEY, C., DAVIES, J., ERVINE, A., POULTER, L., PACHTER, L., MOFFATT, M. F. & COOKSON, W. O. C. 2010. Disordered Microbial Communities in Asthmatic Airways. *PLoS ONE,* 5**,** e8578.

HOLGATE, S. T. 2013. Mechanisms of Asthma and Implications for Its Prevention and Treatment: A Personal Journey. *Allergy Asthma Immunol Res,* 5**,** 343-347.

HOLT, P. G., UPHAM, J. W. & SLY, P. D. 2005. Contemporaneous maturation of immunologic and respiratory functions during early childhood: Implications for development of asthma prevention strategies. *Journal of Allergy and Clinical Immunology,* 116**,** 16-24.

HUANG, Y. J. 2013. Asthma microbiome studies and the potential for new therapeutic strategies. *Curr Allergy Asthma Rep,* 13**,** 453-61.

HUANG, Y. J., KIM, E., COX, M. J., BRODIE, E. L., BROWN, R., WIENER-KRONISH, J. P. & LYNCH, S. V. 2010. A persistent and diverse airway microbiota present during chronic obstructive pulmonary disease exacerbations. *Omics,* 14**,** 9-59.

HUANG, Y. J., NELSON, C. E., BRODIE, E. L., DESANTIS, T. Z., BAEK, M. S., LIU, J., WOYKE, T., ALLGAIER, M., BRISTOW, J., WIENER-KRONISH, J. P., SUTHERLAND, E. R., KING, T. S., ICITOVIC, N., MARTIN, R. J., CALHOUN, W. J., CASTRO, M., DENLINGER, L. C., DIMANGO, E., KRAFT, M., PETERS, S. P., WASSERMAN, S. I., WECHSLER, M. E., BOUSHEY, H. A. & LYNCH, S. V. 2011. Airway microbiota and bronchial hyperresponsiveness in patients with suboptimally controlled asthma. *Journal of Allergy and Clinical Immunology,* 127**,** 372-381.e3.

HUCKABEE, M. M. & PEEBLES, R. S., JR. 2009. Novel concepts in virally induced asthma. *Clin Mol Allergy,* 7**,** 2.

HUSE, S. M., YE, Y., ZHOU, Y. & FODOR, A. A. 2012. A Core Human Microbiome as Viewed through 16S rRNA Sequence Clusters. *PLoS ONE,* 7**,** e34242.

IWAI, S., HUANG, D., FONG, S., JARLSBERG, L. G., WORODRIA, W., YOO, S., CATTAMANCHI, A., DAVIS, J. L., KASWABULI, S., SEGAL, M., HUANG, L. & LYNCH, S. V. 2014. The lung microbiome of Ugandan HIV-infected pneumonia patients is compositionally and functionally distinct from that of San Franciscan patients. *PLoS One,* 9**,** e95726.

JACKSON, D. J., EVANS, M. D., GANGNON, R. E., TISLER, C. J., PAPPAS, T. E., LEE, W. M., GERN, J. E. & LEMANSKE, R. F., JR. 2012. Evidence for a causal relationship between allergic sensitization and rhinovirus wheezing in early life. *Am J Respir Crit Care Med,* 185**,** 281-5.

JACKSON, D. J., SYKES, A., MALLIA, P. & JOHNSTON, S. L. 2011. Asthma exacerbations: origin, effect, and prevention. *J Allergy Clin Immunol,* 128**,** 1165-74.

JAMES, D. G. 2009. International Medical Congress, London 1881. *Journal of Medical Biography,* 17**,** 216-216.

JANSSEN, R., BONT, L., SIEZEN, C. L. E., HODEMAEKERS, H. M., ERMERS, M. J., DOORNBOS, G., SLOT, R. V. T., WIJMENGA, C., GOEMAN, J. J., KIMPEN, J. L. L., VAN HOUWELINGEN, H. C., KIMMAN, T. G. & HOEBEE, B. 2007. Genetic Susceptibility to Respiratory Syncytial Virus Bronchiolitis Is Predominantly Associated with Innate Immune Genes. *Journal of Infectious Diseases,* 196**,** 826-834.

JENSEN, A., FAGO-OLSEN, H., SORENSEN, C. H. & KILIAN, M. 2013. Molecular mapping to species level of the tonsillar crypt microbiota associated with health and recurrent tonsillitis. *PLoS One,* 8**,** e56418.

JIANG, W., LING, Z., LIN, X., CHEN, Y., ZHANG, J., YU, J., XIANG, C. & CHEN, H. 2014. Pyrosequencing analysis of oral microbiota shifting in various caries states in childhood. *Microb Ecol,* 67**,** 962-9.

JOHNSON, C. & TUNKEL, A. 2000. Viridans streptococci and groups C and G streptococci. *In:* G, M., J, B. & R, D. (eds.) *Principles and Practice of Infectious Diseases* 5th ed. New York: Churchill Livingstone.

JOINT, I., MUHLING, M. & QUERELLOU, J. 2010. Culturing marine bacteria - an essential prerequisite for biodiscovery. *Microb Biotechnol,* 3**,** 564-75.

JUMPSTART CONSORTIUM HUMAN MICROBIOME PROJECT DATA GENERATION WORKING, G. 2012. Evaluation of 16S rDNA-Based

Community Profiling for Human Microbiome Research. *PLoS ONE,* 7**,** e39315.
JUST, J., BELFAR, S., WANIN, S., PRIBIL, C., GRIMFELD, A. & DURU, G. 2010. Impact of innate and environmental factors on wheezing persistence during childhood. *J Asthma,* 47**,** 412-6.
KABAT, A. M., SRINIVASAN, N. & MALOY, K. J. 2014. Modulation of immune development and function by intestinal microbiota. *Trends Immunol*.
KALINA, U., KOYAMA, N., HOSODA, T., NUERNBERGER, H., SATO, K., HOELZER, D., HERWECK, F., MANIGOLD, T., SINGER, M. V., ROSSOL, S. & BOCKER, U. 2002. Enhanced production of IL-18 in butyrate-treated intestinal epithelium by stimulation of the proximal promoter region. *Eur J Immunol,* 32**,** 2635-43.
KAPLAN, J. L., SHI, H. N. & WALKER, W. A. 2011. The Role of Microbes in Developmental Immunologic Programming. *Pediatr Res,* 69**,** 465-472.
KAWAMURA, Y., HOU, X. G., SULTANA, F., MIURA, H. & EZAKI, T. 1995. Determination of 16S rRNA sequences of Streptococcus mitis and Streptococcus gordonii and phylogenetic relationships among members of the genus Streptococcus. *Int J Syst Bacteriol,* 45**,** 406-8.
KAWAMURA, Y., WHILEY, R. A., SHU, S. E., EZAKI, T. & HARDIE, J. M. 1999. Genetic approaches to the identification of the mitis group within the genus Streptococcus. *Microbiology,* 145 (Pt 9)**,** 2605-13.
KHATAMI, A. & POLLARD, A. J. 2010. The epidemiology of meningococcal disease and the impact of vaccines. *Expert Rev Vaccines,* 9**,** 285-98.
KILIAN, M., RILEY, D. R., JENSEN, A., BRUGGEMANN, H. & TETTELIN, H. 2014. Parallel Evolution of Streptococcus pneumoniae and Streptococcus mitis to Pathogenic and Mutualistic Lifestyles. *MBio,* 5.
KIM, M. & YU, Z. 2014. Variations in 16S rRNA-based microbiome profiling between pyrosequencing runs and between pyrosequencing facilities. *J Microbiol,* 52**,** 355-65.
KING, P. 2011. Pathogenesis of bronchiectasis. *Paediatr Respir Rev,* 12**,** 104-10.
KLEPAC-CERAJ, V., LEMON, K. P., MARTIN, T. R., ALLGAIER, M., KEMBEL, S. W., KNAPP, A. A., LORY, S., BRODIE, E. L., LYNCH, S. V., BOHANNAN, B. J., GREEN, J. L., MAURER, B. A. & KOLTER, R. 2010. Relationship between cystic fibrosis respiratory tract bacterial communities and age, genotype, antibiotics and Pseudomonas aeruginosa. *Environ Microbiol,* 12**,** 1293-303.
KOHL, J. & WILLS-KARP, M. 2007. A dual role for complement in allergic asthma. *Curr Opin Pharmacol,* 7**,** 283-9.
KOTANIEMI-SYRJANEN, A., VAINIONPAA, R., REIJONEN, T. M., WARIS, M., KORHONEN, K. & KORPPI, M. 2003. Rhinovirus-induced wheezing in infancy--the first sign of childhood asthma? *J Allergy Clin Immunol,* 111**,** 66-71.
KRAFT, M. 2000. The role of bacterial infections in asthma. *Clin Chest Med,* 21**,** 301-13.
KROEGEL, C. 2009. Global Initiative for Asthma (GINA) guidelines: 15 years of application. *Expert Rev Clin Immunol,* 5**,** 239-49.
KRZYSCIAK, W., PLUSKWA, K. K., JURCZAK, A. & KOSCIELNIAK, D. 2013. The pathogenicity of the Streptococcus genus. *Eur J Clin Microbiol Infect Dis,* 32**,** 1361-76.

KUTLUHAN, A., SALVIZ, M., YALCINER, G., KANDEMIR, O. & YESIL, C. 2011. The role of the actinomyces in obstructive tonsillar hypertrophy and recurrent tonsillitis in pediatric population. *Int J Pediatr Otorhinolaryngol,* 75**,** 391-4.

LANE, D. 1991. 16S/23S rRNA sequencing. *In:* STACKEBRANDT, E. & GOODFELLOW, M. (eds.) *Nucleic acid techniques in bacterial systematics.* Chichester, United Kingdom: John Wiley and Sons.

LAWSON, A. J. 2004. Discovering new pathogens: culture-resistant bacteria. *Methods Mol Biol,* 266**,** 305-22.

LEBOUDER, E., REY-NORES, J. E., RABY, A. C., AFFOLTER, M., VIDAL, K., THORNTON, C. A. & LABÉTA, M. O. 2006. Modulation of neonatal microbial recognition: TLR-mediated innate immune responses are specifically and differentially modulated by human milk. *Journal of Immunology,* 176**,** 3742-3752.

LEDERBERG, J. 2000. Infectious history. *Science,* 288**,** 287-93.

LEMANSKE, R. F., JR., JACKSON, D. J., GANGNON, R. E., EVANS, M. D., LI, Z., SHULT, P. A., KIRK, C. J., REISDORF, E., ROBERG, K. A., ANDERSON, E. L., CARLSON-DAKES, K. T., ADLER, K. J., GILBERTSON-WHITE, S., PAPPAS, T. E., DASILVA, D. F., TISLER, C. J. & GERN, J. E. 2005. Rhinovirus illnesses during infancy predict subsequent childhood wheezing. *J Allergy Clin Immunol,* 116**,** 571-7.

LEMON, K. P., KLEPAC-CERAJ, V., SCHIFFER, H. K., BRODIE, E. L., LYNCH, S. V. & KOLTER, R. 2010. Comparative analyses of the bacterial microbiota of the human nostril and oropharynx. *MBio,* 1.

LETUNIC, I. & BORK, P. 2011. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res,* 39**,** W475-8.

LEWIS, D. A., BROWN, R., WILLIAMS, J., WHITE, P., JACOBSON, S. K., MARCHESI, J. R. & DRAKE, M. J. 2013. The human urinary microbiome; bacterial DNA in voided urine of asymptomatic adults. *Front Cell Infect Microbiol,* 3**,** 41.

LI, J., ADAMS, V., BANNAM, T. L., MIYAMOTO, K., GARCIA, J. P., UZAL, F. A., ROOD, J. I. & MCCLANE, B. A. 2013. Toxin plasmids of Clostridium perfringens. *Microbiol Mol Biol Rev,* 77**,** 208-33.

LI, K., BIHAN, M., YOOSEPH, S. & METHÉ, B. A. 2012. Analyses of the Microbial Diversity across the Human Microbiome. *PLoS ONE,* 7**,** e32118.

LICCIARDI, P. V., WONG, S. S., TANG, M. L. & KARAGIANNIS, T. C. 2010. Epigenome targeting by probiotic metabolites. *Gut Pathog,* 2**,** 24.

LIU, A. H. 2007. Hygiene theory and allergy and asthma prevention. *Paediatr Perinat Epidemiol,* 21 Suppl 3**,** 2-7.

LOTZ, M. T., MOORE, M. L. & PEEBLES, R. S., JR. 2013. Respiratory syncytial virus and reactive airway disease. *Curr Top Microbiol Immunol,* 372**,** 105-18.

LOZUPONE, C., HAMADY, M. & KNIGHT, R. 2006. UniFrac--an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics,* 7**,** 371.

LOZUPONE, C. & KNIGHT, R. 2005a. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol,* 71**,** 8228-35.

LOZUPONE, C. & KNIGHT, R. 2005b. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology,* 71**,** 8228-8235.

LU, Q. B., WO, Y., WANG, H. Y., WEI, M. T., ZHANG, L., YANG, H., LIU, E. M., LI, T. Y., ZHAO, Z. T., LIU, W. & CAO, W. C. 2013. Enterovirus 68 is one of the

commonest types of Enterovirus found in patients with Acute Respiratory Tract Infection in China. *J Med Microbiol*.

LUDWIG, W., STRUNK, O., WESTRAM, R., RICHTER, L., MEIER, H., YADHUKUMAR, BUCHNER, A., LAI, T., STEPPI, S., JOBB, G., FORSTER, W., BRETTSKE, I., GERBER, S., GINHART, A. W., GROSS, O., GRUMANN, S., HERMANN, S., JOST, R., KONIG, A., LISS, T., LUSSMANN, R., MAY, M., NONHOFF, B., REICHEL, B., STREHLOW, R., STAMATAKIS, A., STUCKMANN, N., VILBIG, A., LENKE, M., LUDWIG, T., BODE, A. & SCHLEIFER, K. H. 2004a. ARB: a software environment for sequence data. *Nucleic Acids Res,* 32**,** 1363-71.

LUDWIG, W., STRUNK, O., WESTRAM, R., RICHTER, L., MEIER, H., YADHUKUMAR, BUCHNER, A., LAI, T., STEPPI, S., JOBB, G., FÖRSTER, W., BRETTSKE, I., GERBER, S., GINHART, A. W., GROSS, O., GRUMANN, S., HERMANN, S., JOST, R., KÖNIG, A., LISS, T., LÜSSMAN, R., MAY, M., NONHOFF, B., REICHEL, B., STREHLOW, R., STAMATAKIS, A., STUCKMANN, N., VILBIG, A., LENKE, M., LUDWIG, T., BODE, A. & SCHLEIFER, K. H. 2004b. ARB: a software environment for sequence data. *Nucleic Acids Res,* 32**,** 1363-1371.

MAHENTHIRALINGAM, E. 2014. Emerging cystic fibrosis pathogens and the microbiome. *Paediatr Respir Rev,* 15 Suppl 1**,** 13-5.

MALLOL, J., SOLE, D., BAEZA-BACAB, M., AGUIRRE-CAMPOSANO, V., SOTO-QUIROS, M. & BAENA-CAGNANI, C. 2010. Regional variation in asthma symptom prevalence in Latin American children. *J Asthma,* 47**,** 644-50.

MAMESSIER, E. & MAGNAN, A. 2006. Cytokines in atopic diseases: revisiting the Th2 dogma. *Eur J Dermatol,* 16**,** 103-13.

MARCHESI, J. R., SATO, T., WEIGHTMAN, A. J., MARTIN, T. A., FRY, J. C., HIOM, S. J., DYMOCK, D. & WADE, W. G. 1998. Design and evaluation of useful bacterium-specific PCR primers that amplify genes coding for bacterial 16S rRNA. *Appl Environ Microbiol,* 64**,** 795-9.

MARCON, E., SCOTTI, I., HERAULT, B., ROSSI, V. & LANG, G. 2014. Generalization of the partitioning of shannon diversity. *PLoS One,* 9**,** e90289.

MARRI, P. R., STERN, D. A., WRIGHT, A. L., BILLHEIMER, D. & MARTINEZ, F. D. 2013. Asthma-associated differences in microbial composition of induced sputum. *J Allergy Clin Immunol,* 131**,** 346-352 e3.

MARSLAND, B. J. 2013. Influences of the microbiome on the early origins of allergic asthma. *Ann Am Thorac Soc,* 10 Suppl**,** S165-9.

MARSLAND, B. J., YADAVA, K. & NICOD, L. P. 2013. The airway microbiome and disease. *Chest,* 144**,** 632-7.

MARTINEZ, F. D. & VERCELLI, D. Asthma. *The Lancet,* 382**,** 1360-1372.

MASOLI, M., FABIAN, D., HOLT, S., BEASLEY, R. & GLOBAL INITIATIVE FOR ASTHMA, P. 2004. The global burden of asthma: executive summary of the GINA Dissemination Committee report. *Allergy,* 59**,** 469-78.

MATSUI, A. & JIN, J. O. 2014. Pathogenic bacterial species associated with endodontic infection evade innate immune control by disabling neutrophils.

MCAULIFFE, L., ELLIS, R. J., LAWES, J. R., AYLING, R. D. & NICHOLAS, R. A. 2005. 16S rDNA PCR and denaturing gradient gel electrophoresis; a single generic test for detecting and differentiating Mycoplasma species. *J Med Microbiol,* 54**,** 731-9.

MCDONALD, D., PRICE, M. N., GOODRICH, J., NAWROCKI, E. P., DESANTIS, T. Z., PROBST, A., ANDERSEN, G. L., KNIGHT, R. & HUGENHOLTZ, P. 2012.

An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *Isme j,* 6**,** 610-8.

MCKEEVER, T. M., LEWIS, S. A., SMITH, C., COLLINS, J., HEATLIE, H., FRISCHER, M. & HUBBARD, R. 2002. Early exposure to infections and antibiotics and the incidence of allergic disease: A birth cohort study with the West Midlands General Practice Research Database. *Journal of Allergy and Clinical Immunology,* 109**,** 43-50.

MCMURDIE, P. J. & HOLMES, S. 2013. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One,* 8**,** e61217.

MESSAOUDI, S., MANAI, M., KERGOURLAY, G., PREVOST, H., CONNIL, N., CHOBERT, J. M. & DOUSSET, X. 2013. Lactobacillus salivarius: bacteriocin and probiotic activity. *Food Microbiol,* 36**,** 296-304.

MITCHELL, J. 2011. Streptococcus mitis: walking the line between commensalism and pathogenesis. *Mol Oral Microbiol,* 26**,** 89-98.

MOFFATT, M. F., GUT, I. G., DEMENAIS, F., STRACHAN, D. P., BOUZIGON, E., HEATH, S., VON MUTIUS, E., FARRALL, M., LATHROP, M. & COOKSON, W. O. C. M. 2010. A Large-Scale, Consortium-Based Genomewide Association Study of Asthma. *New England Journal of Medicine,* 363**,** 1211-1221.

MOFFATT, M. F., KABESCH, M., LIANG, L., DIXON, A. L., STRACHAN, D., HEATH, S., DEPNER, M., VON BERG, A., BUFE, A., RIETSCHEL, E., HEINZMANN, A., SIMMA, B., FRISCHER, T., WILLIS-OWEN, S. A., WONG, K. C., ILLIG, T., VOGELBERG, C., WEILAND, S. K., VON MUTIUS, E., ABECASIS, G. R., FARRALL, M., GUT, I. G., LATHROP, G. M. & COOKSON, W. O. 2007. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature,* 448**,** 470-3.

MOLYNEAUX, P. L., MALLIA, P., COX, M. J., FOOTITT, J., WILLIS-OWEN, S. A., HOMOLA, D., TRUJILLO-TORRALBO, M. B., ELKIN, S., KON, O. M., COOKSON, W. O., MOFFATT, M. F. & JOHNSTON, S. L. 2013. Outgrowth of the bacterial airway microbiome after rhinovirus exacerbation of chronic obstructive pulmonary disease. *Am J Respir Crit Care Med,* 188**,** 1224-31.

MONCAYO, A. L., VACA, M., OVIEDO, G., ERAZO, S., QUINZO, I., FIACCONE, R. L., CHICO, M. E., BARRETO, M. L. & COOPER, P. J. 2010. Risk factors for atopic and non-atopic asthma in a rural area of Ecuador. *Thorax,* 65**,** 409-16.

MORGAN, X. C. & HUTTENHOWER, C. 2012. Chapter 12: Human microbiome analysis. *PLoS Comput Biol,* 8**,** e1002808.

MOSER, S., PERONI, D. G., COMBERIATI, P. & PIACENTINI, G. L. 2014. Asthma and viruses: is there a relationship? *Front Biosci (Elite Ed),* 6**,** 46-54.

MUHLING, M., WOOLVEN-ALLEN, J., MURRELL, J. C. & JOINT, I. 2008. Improved group-specific PCR primers for denaturing gradient gel electrophoresis analysis of the genetic diversity of complex microbial communities. *Isme j,* 2**,** 379-92.

MURPHY, T. F., BAKALETZ, L. O. & SMEESTERS, P. R. 2009. Microbial interactions in the respiratory tract. *Pediatr Infect Dis J,* 28**,** S121-6.

MURRAY, C. S., TANNOCK, G. W., SIMON, M. A., HARMSEN, H. J. M., WELLING, G. W., CUSTOVIC, A. & WOODCOCK, A. 2005. Fecal microbiota in sensitized wheezy and non-sensitized non-wheezy children: a nested case–control study. *Clinical & Experimental Allergy,* 35**,** 741-745.

MURRAY, P. R. & ROSENBLATT, J. E. 1976. Bacterial interference by oropharynegeal and clinical isolates of anaerobic bacteria. *J Infect Dis,* 134**,** 281-5.

MUYZER, G. & SMALLA, K. 1998. Application of denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) in microbial ecology. *Antonie Van Leeuwenhoek,* 73**,** 127-41.

NAKANISHI, Y., SATO, T. & OHTEKI, T. 2014. Commensal Gram-positive bacteria initiates colitis by inducing monocyte/macrophage mobilization. *Mucosal Immunol*.

NARDIS, C., MOSCA, L. & MASTROMARINO, P. 2013. Vaginal microbiota and viral sexually transmitted diseases. *Ann Ig,* 25**,** 443-56.

NAU, R., DJUKIC, M., SPREER, A. & EIFFERT, H. 2013. Bacterial meningitis: new therapeutic approaches. *Expert Rev Anti Infect Ther,* 11**,** 1079-95.

NELSON, K. A. & ZORC, J. J. 2013. Asthma update. *Pediatr Clin North Am,* 60**,** 1035-48.

NELSON, K. E., WEINSTOCK, G. M., HIGHLANDER, S. K., WORLEY, K. C., CREASY, H. H., WORTMAN, J. R., RUSCH, D. B., MITREVA, M., SODERGREN, E., CHINWALLA, A. T., FELDGARDEN, M., GEVERS, D., HAAS, B. J., MADUPU, R., WARD, D. V., BIRREN, B. W., GIBBS, R. A., METHE, B., PETROSINO, J. F., STRAUSBERG, R. L., SUTTON, G. G., WHITE, O. R., WILSON, R. K., DURKIN, S., GIGLIO, M. G., GUJJA, S., HOWARTH, C., KODIRA, C. D., KYRPIDES, N., MEHTA, T., MUZNY, D. M., PEARSON, M., PEPIN, K., PATI, A., QIN, X., YANDAVA, C., ZENG, Q., ZHANG, L., BERLIN, A. M., CHEN, L., HEPBURN, T. A., JOHNSON, J., MCCORRISON, J., MILLER, J., MINX, P., NUSBAUM, C., RUSS, C., SYKES, S. M., TOMLINSON, C. M., YOUNG, S., WARREN, W. C., BADGER, J., CRABTREE, J., MARKOWITZ, V. M., ORVIS, J., CREE, A., FERRIERA, S., FULTON, L. L., FULTON, R. S., GILLIS, M., HEMPHILL, L. D., JOSHI, V., KOVAR, C., TORRALBA, M., WETTERSTRAND, K. A., ABOUELLLEIL, A., WOLLAM, A. M., BUHAY, C. J., DING, Y., DUGAN, S., FITZGERALD, M. G., HOLDER, M., HOSTETLER, J., CLIFTON, S. W., ALLEN-VERCOE, E., EARL, A. M., FARMER, C. N., LIOLIOS, K., SURETTE, M. G., XU, Q., POHL, C., WILCZEK-BONEY, K. & ZHU, D. 2010. A catalog of reference genomes from the human microbiome. *Science,* 328**,** 994-9.

NICHOLSON, K. G., KENT, J. & IRELAND, D. C. 1993. Respiratory viruses and exacerbations of asthma in adults. *BMJ,* 307**,** 982-6.

NOVERR, M. C. & HUFFNAGLE, G. B. 2005. The 'microflora hypothesis' of allergic diseases. *Clin Exp Allergy,* 35**,** 1511-20.

NWAOHIRI, N., URBAN, C., GLUCK, J., AHLUWALIA, M. & WEHBEH, W. 2009. Tricuspid valve endocarditis caused by Haemophilus parainfluenzae: a case report and review of the literature. *Diagn Microbiol Infect Dis,* 64**,** 216-9.

OKAYAMA, Y. 2013. Cellular and humoral immunity of virus-induced asthma. *Frontiers in Microbiology,* 4.

OPENSHAW, P. J., DEAN, G. S. & CULLEY, F. J. 2003. Links between respiratory syncytial virus bronchiolitis and childhood asthma: clinical and research approaches. *Pediatr Infect Dis J,* 22**,** S58-64; discussion S64-5.

OTERO, C., PAZ, R. D., GALASSI, N., BEZRODNIK, L., FINIASZ, M. R. & FINK, S. 2013. Immune response to Streptococcus pneumoniae in asthma patients: comparison between stable situation and exacerbation. *Clin Exp Immunol,* 173**,** 92-101.

PAGES, H., ABOYOUN, P., GENTLEMAN, R. & DEBROY, S. Biostrings: String objects representing biological sequences, and matching algorithms.
PAPADAKI, T. G., EL MOUSSAOUI, R., VAN KETEL, R. J., VERBRAAK, F. D. & TAN, H. S. 2008. Capnocytophaga canimorsus endogenous endophthalmitis in an immunocompetent host. *Br J Ophthalmol,* 92**,** 1566-7.
PARADIS, E., CLAUDE, J. & STRIMMER, K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics,* 20**,** 289-290.
PARAHITIYAWA, N. B., SCULLY, C., LEUNG, W. K., YAM, W. C., JIN, L. J. & SAMARANAYAKE, L. P. 2010. Exploring the oral bacterial flora: current status and future directions. *Oral Dis,* 16**,** 136-45.
PAULSON, J. N., STINE, O. C., BRAVO, H. C. & POP, M. 2013. Differential abundance analysis for microbial marker-gene surveys. *Nat Meth,* 10**,** 1200-1202.
PETROSINO, J. F., HIGHLANDER, S., LUNA, R. A., GIBBS, R. A. & VERSALOVIC, J. 2009. Metagenomic pyrosequencing and microbial identification. *Clin Chem,* 55**,** 856-66.
PIAU, C., ARVIEUX, C., BONNAURE-MALLET, M. & JOLIVET-GOUGEON, A. 2013. Capnocytophaga spp. involvement in bone infections: a review. *Int J Antimicrob Agents,* 41**,** 509-15.
PICARD, F. J., KE, D., BOUDREAU, D. K., BOISSINOT, M., HULETSKY, A., RICHARD, D., OUELLETTE, M., ROY, P. H. & BERGERON, M. G. 2004. Use of tuf sequences for genus-specific PCR detection and phylogenetic analysis of 28 streptococcal species. *J Clin Microbiol,* 42**,** 3686-95.
PIEDIMONTE, G. 2013. Respiratory syncytial virus and asthma: speed-dating or long-term relationship? *Curr Opin Pediatr,* 25**,** 344-9.
POMMERVILLE, J. C. 2011. *Alcamo´s Laboratory Fundamentals of Microbiology,* United Kingdom, Jones & Bartlet.
PRESCOTT, S. L., MACAUBAS, C., HOLT, B. J., SMALLACOMBE, T. B., LOH, R., SLY, P. D. & HOLT, P. G. 1998. Transplacental priming of the human immune system to environmental allergens: Universal skewing of initial T cell responses toward the Th2 cytokine profile. *Journal of Immunology,* 160**,** 4730-4737.
PRICE, M. N., DEHAL, P. S. & ARKIN, A. P. 2010a. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One,* 5**,** e9490.
PRICE, M. N., DEHAL, P. S. & ARKIN, A. P. 2010b. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One,* 5**,** e9490.
PRUESSE, E., PEPLIES, J. & GLOCKNER, F. O. 2012. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics,* 28**,** 1823-9.
PULIMOOD, T. B., CORDEN, J. M., BRYDEN, C., SHARPLES, L. & NASSER, S. M. 2007. Epidemic asthma and the role of the fungal mold Alternaria alternata. *J Allergy Clin Immunol,* 120**,** 610-7.
RAKOFF-NAHOUM, S., PAGLINO, J., ESLAMI-VARZANEH, F., EDBERG, S. & MEDZHITOV, R. 2004. Recognition of commensal microflora by toll-like receptors is required for intestinal homeostasis. *Cell,* 118**,** 229-41.
RATH, S. K., MUKHERJEE, M., KAUSHIK, R., SEN, S. & KUMAR, M. 2014. Periodontal pathogens in atheromatous plaque. *Indian J Pathol Microbiol,* 57**,** 259-64.
REEDER, J. & KNIGHT, R. 2010. Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. *Nat Methods,* 7**,** 668-9.

RELMAN, D. A. 1999. The search for unrecognized pathogens. *Science,* 284**,** 1308-10.
RELMAN, D. A. 2002. New technologies, human-microbe interactions, and the search for previously unrecognized pathogens. *J Infect Dis,* 186 Suppl 2**,** S254-8.
ROGERS, G. B., ZAIN, N. M., BRUCE, K. D., BURR, L. D., CHEN, A. C., RIVETT, D. W., MCGUCKIN, M. A. & SERISIER, D. J. 2014. A novel microbiota stratification system predicts future exacerbations in bronchiectasis. *Ann Am Thorac Soc,* 11**,** 496-503.
ROUPHAEL, N. G. & STEPHENS, D. S. 2012. Neisseria meningitidis: biology, microbiology, and epidemiology. *Methods Mol Biol,* 799**,** 1-20.
RUSSELL, S. L. & FINLAY, B. B. 2012. The impact of gut microbes in allergic diseases. *Curr Opin Gastroenterol,* 28**,** 563-9.
SABRA, A. & BENGER, J. 2011. Meningococcal disease in children: a clinical review. *Turk J Pediatr,* 53**,** 477-88.
SACCHI, C. T., ALBER, D., DULL, P., MOTHERSHED, E. A., WHITNEY, A. M., BARNETT, G. A., POPOVIC, T. & MAYER, L. W. 2005. High level of sequence diversity in the 16S rRNA genes of Haemophilus influenzae isolates is useful for molecular subtyping. *J Clin Microbiol,* 43**,** 3734-42.
SACHS, A. P., VAN DER WAAIJ, D., GROENIER, K. H., KOETER, G. H. & SCHIPHUIS, J. 1993. Oropharyngeal flora in asthma and in chronic obstructive pulmonary disease. Indigenous oropharyngeal microorganisms in outpatients with asthma or chronic obstructive pulmonary disease. *Am Rev Respir Dis,* 148**,** 1302-7.
SAINI, S., GUPTA, N., APARNA, SEEMA & SACHDEVA, O. P. 2005. Bacteriological study of paediatric and adult chronic suppurative otitis media. *Indian J Pathol Microbiol,* 48**,** 413-6.
SAITOU, N. & NEI, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol,* 4**,** 406-25.
SAVAGE, D. C. 1977. Microbial ecology of the gastrointestinal tract. *Annu Rev Microbiol,* 31**,** 107-33.
SCHAUER, U., HOFFJAN, S., BITTSCHEIDT, J., KOCHLING, A., HEMMIS, S., BONGARTZ, S. & STEPHAN, V. 2002. RSV bronchiolitis and risk of wheeze and allergic sensitisation in the first year of life. *Eur Respir J,* 20**,** 1277-83.
SCHLOSS, P. D. & HANDELSMAN, J. 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol,* 71**,** 1501-6.
SCHUSTER, J. E. & WILLIAMS, J. V. 2013. Human metapneumovirus. *Pediatr Rev,* 34**,** 558-65.
SEBASTIAN, L. J., PHILIP, K. P., GWENDOLYN, S., SANDRA, S., FIONA, L., LYNN, J., PENNY, S., SUSAN, O., TOOLE, STEVEN, H. M., DAVID, A. J. T. & STEPHEN, T. H. 1995. Community study of role of viral infections in exacerbations of asthma in 9-11 year old children. *BMJ,* 310.
SHANNON, C. E. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal,* 27.
SHAPIRO, S. S. & WILK, M. B. 1965. An analysis of variance test for normality (complete samples). *Biometrika,* 52**,** 591-611.
SHENEP, J. L. 2000. Viridans-group streptococcal infections in immunocompromised hosts. *Int J Antimicrob Agents,* 14**,** 129-35.

SHERRARD, L. J., SCHAIBLE, B., GRAHAM, K. A., MCGRATH, S. J., MCILREAVEY, L., HATCH, J., WOLFGANG, M. C., MUHLEBACH, M. S., GILPIN, D. F., SCHNEIDERS, T., ELBORN, J. S. & TUNNEY, M. M. 2014. Mechanisms of reduced susceptibility and genotypic prediction of antibiotic resistance in Prevotella isolated from cystic fibrosis (CF) and non-CF patients. *J Antimicrob Chemother*.

SIGURS, N., BJARNASON, R., SIGURBERGSSON, F. & KJELLMAN, B. 2000. Respiratory syncytial virus bronchiolitis in infancy is an important risk factor for asthma and allergy at age 7. *Am J Respir Crit Care Med,* 161**,** 1501-7.

SIMPSON, E. 1949. Measurement of diversity. *Nature,* 163**,** 688.

STALEY, J. T. & KONOPKA, A. 1985. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu Rev Microbiol,* 39**,** 321-46.

STAMATAKIS, A., LUDWIG, T. & MEIER, H. 2005a. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics,* 21**,** 456-463.

STAMATAKIS, A., LUDWIG, T. & MEIER, H. 2005b. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics,* 21**,** 456-63.

STERN, D. A., MORGAN, W. J., HALONEN, M., WRIGHT, A. L. & MARTINEZ, F. D. 2008. Wheezing and bronchial hyper-responsiveness in early childhood as predictors of newly diagnosed asthma in early adulthood: a longitudinal birth-cohort study. *Lancet,* 372**,** 1058-64.

STOKKEN, J., GUPTA, A., KRAKOVITZ, P. & ANNE, S. 2014. Rhinosinusitis in children: A comparison of patients requiring surgery for acute complications versus chronic disease. *Am J Otolaryngol*.

STRACHAN, D. P. 1989. Hay fever, hygiene, and household size. *BMJ,* 299**,** 1259-60.

STRESSMANN, F. A., ROGERS, G. B., KLEM, E. R., LILLEY, A. K., DONALDSON, S. H., DANIELS, T. W., CARROLL, M. P., PATEL, N., FORBES, B., BOUCHER, R. C., WOLFGANG, M. C. & BRUCE, K. D. 2011. Analysis of the Bacterial Communities Present in Lungs of Patients with Cystic Fibrosis from American and British Centers. *Journal of Clinical Microbiology,* 49**,** 281-291.

SWENSON, N. G., ENQUIST, B. J., PITHER, J., THOMPSON, J. & ZIMMERMAN, J. K. 2006. The problem and promise of scale dependency in community phylogenetics. *Ecology,* 87**,** 2418-24.

SZEFLER, S. J., CHMIEL, J. F., FITZPATRICK, A. M., GIACOIA, G., GREEN, T. P., JACKSON, D. J., NIELSEN, H. C., PHIPATANAKUL, W. & RAISSY, H. H. 2014. Asthma across the ages: Knowledge gaps in childhood asthma. *Journal of Allergy and Clinical Immunology,* 133**,** 3-13.

TAMURA, K., STECHER, G., PETERSON, D., FILIPSKI, A. & KUMAR, S. 2013a. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol,* 30**,** 2725-9.

TAMURA, K., STECHER, G., PETERSON, D., FILIPSKI, A. & KUMAR, S. 2013b. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution*.

TANAKA, S., YOSHIDA, M., MURAKAMI, Y., OGIWARA, T., SHOJI, M., KOBAYASHI, S., WATANABE, S., MACHINO, M. & FUJISAWA, S. 2008. The relationship of Prevotella intermedia, Prevotella nigrescens and Prevotella

melaninogenica in the supragingival plaque of children, caries and oral malodor. *J Clin Pediatr Dent,* 32**,** 195-200.

TANO, K., GRAHN HAKANSSON, E., WALLBRANDT, P., RONNQVIST, D., HOLM, S. E. & HELLSTROM, S. 2003. Is hydrogen peroxide responsible for the inhibitory activity of alpha-haemolytic streptococci sampled from the nasopharynx? *Acta Otolaryngol,* 123**,** 724-9.

TANO, K., GRAHN-HAKANSSON, E., HOLM, S. E. & HELLSTROM, S. 2000. Inhibition of OM pathogens by alpha-hemolytic streptococci from healthy children, children with SOM and children with rAOM. *Int J Pediatr Otorhinolaryngol,* 56**,** 185-90.

TEAM, R. C. 2014. *A language and environment for statistical computing.* [Online]. Viena, Austria: R Foundation for Statistical Computing. Available: http://www.r-project.org./.

THOMPSON, J. D., GIBSON, T. J. & HIGGINS, D. G. 2002. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics,* Chapter 2**,** Unit 2.3.

THORNHILL, D. J., LAJEUNESSE, T. C. & SANTOS, S. R. 2007. Measuring rDNA diversity in eukaryotic microbial systems: how intragenomic variation, pseudogenes, and PCR artifacts confound biodiversity estimates. *Mol Ecol,* 16**,** 5326-40.

THORPE, J. E., BAUGHMAN, R. P., FRAME, P. T., WESSELER, T. A. & STANECK, J. L. 1987. Bronchoalveolar Lavage for Diagnosing Acute Bacterial Pneumonia. *Journal of Infectious Diseases,* 155**,** 855-861.

TLASKALOVA-HOGENOVA, H., STEPANKOVA, R., HUDCOVIC, T., TUCKOVA, L., CUKROWSKA, B., LODINOVA-ZADNIKOVA, R., KOZAKOVA, H., ROSSMANN, P., BARTOVA, J., SOKOL, D., FUNDA, D. P., BOROVSKA, D., REHAKOVA, Z., SINKORA, J., HOFMAN, J., DRASTICH, P. & KOKESOVA, A. 2004. Commensal bacteria (normal microflora), mucosal immunity and chronic inflammatory and autoimmune diseases. *Immunol Lett,* 93**,** 97-108.

TORRAZZA, R. M. & NEU, J. 2011. The developing intestinal microbiome and its relationship to health and disease in the neonate. *J Perinatol,* 31**,** S29-S34.

TSUKAGOSHI, H., ISHIOKA, T., MASAHIRO, N., KOZAWA, K. & KIMURA, H. 2013. Molecular epidemiology of respiratory viruses in virus-induced asthma. *Frontiers in Microbiology,* 4.

TURNBAUGH, P. J., LEY, R. E., HAMADY, M., FRASER-LIGGETT, C. M., KNIGHT, R. & GORDON, J. I. 2007. The human microbiome project. *Nature,* 449**,** 804-10.

TZENEVA, V. A., HEILIG, H. G., VAN VLIET, W. A., AKKERMANS, A. D., DE VOS, W. M. & SMIDT, H. 2008. 16S rRNA targeted DGGE fingerprinting of microbial communities. *Methods Mol Biol,* 410**,** 335-49.

VAEL, C. & DESAGER, K. 2009. The importance of the development of the intestinal microbiota in infancy. *Current Opinion in Pediatrics,* 21**,** 794-800 10.1097/MOP.0b013e328332351b.

VAN DER GAST, C. J., CUTHBERTSON, L., ROGERS, G. B., POPE, C., MARSH, R. L., REDDING, G. J., BRUCE, K. D., CHANG, A. B. & HOFFMAN, L. R. 2014. Three Clinically Distinct Chronic Pediatric Airway Infections Share a Common Core Microbiota. *Ann Am Thorac Soc*.

VAZQUEZ-BAEZA, Y., PIRRUNG, M., GONZALEZ, A. & KNIGHT, R. 2013. EMPeror: a tool for visualizing high-throughput microbial community data. *GigaScience,* 2**,** 16.

VESTH, T., OZEN, A., ANDERSEN, S. C., KAAS, R. S., LUKJANCENKO, O., BOHLIN, J., NOOKAEW, I., WASSENAAR, T. M. & USSERY, D. W. 2013. Veillonella, Firmicutes: Microbes disguised as Gram negatives. *Stand Genomic Sci,* 9**,** 431-48.

VISSING, N. H., CHAWES, B. L. & BISGAARD, H. 2013. Increased risk of pneumonia and bronchiolitis after bacterial colonization of the airways as neonates. *Am J Respir Crit Care Med,* 188**,** 1246-52.

WALKER, W. A. 2013. Initial intestinal colonization in the human infant and immune homeostasis. *Ann Nutr Metab,* 63 Suppl 2**,** 8-15.

WANG, M., AHRNE, S., JEPPSSON, B. & MOLIN, G. 2005. Comparison of bacterial diversity along the human intestinal tract by direct cloning and sequencing of 16S rRNA genes. *FEMS Microbiol Ecol,* 54**,** 219-31.

WARD, J., WARD, J. & WIENER, C. 2006. *The respiratory system at a glance*, Wiley-Blackwell.

WATT, J. P., WOLFSON, L. J., O'BRIEN, K. L., HENKLE, E., DELORIA-KNOLL, M., MCCALL, N., LEE, E., LEVINE, O. S., HAJJEH, R., MULHOLLAND, K. & CHERIAN, T. 2009. Burden of disease caused by Haemophilus influenzae type b in children younger than 5 years: global estimates. *Lancet,* 374**,** 903-11.

WECHSLER, M. E. 2009. Managing asthma in primary care: putting new guideline recommendations into context. *Mayo Clin Proc,* 84**,** 707-17.

WEINBERG, E. G. 2000. Urbanization and childhood asthma: an African perspective. *J Allergy Clin Immunol,* 105**,** 224-31.

WEINMAYR, G., WEILAND, S. K., BJORKSTEN, B., BRUNEKREEF, B., BUCHELE, G., COOKSON, W. O., GARCIA-MARCOS, L., GOTUA, M., GRATZIOU, C., VAN HAGE, M., VON MUTIUS, E., RIIKJARV, M. A., RZEHAK, P., STEIN, R. T., STRACHAN, D. P., TSANAKAS, J., WICKENS, K. & WONG, G. W. 2007. Atopic sensitization and the international variation of asthma symptom prevalence in children. *Am J Respir Crit Care Med,* 176**,** 565-74.

WENZEL, S. E. 2012. Asthma phenotypes: the evolution from clinical to molecular approaches. *Nat Med,* 18**,** 716-725.

WHITE, J. R., NAGARAJAN, N. & POP, M. 2009. Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. *PLoS Comput Biol,* 5**,** e1000352.

WOESE, C. R. & FOX, G. E. 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences,* 74**,** 5088-5090.

WOLFF, D., FRESE, C., MAIER-KRAUS, T., KRUEGER, T. & WOLFF, B. 2013. Bacterial biofilm composition in caries and caries-free subjects. *Caries Res,* 47**,** 69-77.

WOLTER, D. J., EMERSON, J. C., MCNAMARA, S., BUCCAT, A. M., QIN, X., COCHRANE, E., HOUSTON, L. S., ROGERS, G. B., MARSH, P., PREHAR, K., POPE, C. E., BLACKLEDGE, M., DEZIEL, E., BRUCE, K. D., RAMSEY, B. W., GIBSON, R. L., BURNS, J. L. & HOFFMAN, L. R. 2013. Staphylococcus aureus small-colony variants are independently associated with worse lung disease in children with cystic fibrosis. *Clin Infect Dis,* 57**,** 384-91.

WOOD, L. G., SIMPSON, J. L., HANSBRO, P. M. & GIBSON, P. G. 2010. Potentially pathogenic bacteria cultured from the sputum of stable asthmatics are

associated with increased 8-isoprostane and airway neutrophilia. *Free Radic Res,* 44**,** 146-54.
YAMANAKA, T., HELGELAND, L., FARSTAD, I. N., FUKUSHIMA, H., MIDTVEDT, T. & BRANDTZAEG, P. 2003. Microbial colonization drives lymphocyte accumulation and differentiation in the follicle-associated epithelium of Peyer's patches. *J Immunol,* 170**,** 816-22.
YONG, C. H., LIU, G., CHUA, H. N. & WONG, L. 2012. Supervised maximum-likelihood weighting of composite protein networks for complex prediction. *BMC Syst Biol,* 6 Suppl 2**,** S13.
ZEMANICK, E. T., SAGEL, S. D. & HARRIS, J. K. 2011. The airway microbiome in cystic fibrosis and implications for treatment. *Curr Opin Pediatr,* 23**,** 319-24.
ZHANG, Y., MOFFATT, M. F. & COOKSON, W. O. 2012. Genetic and genomic approaches to asthma: new insights for the origins. *Curr Opin Pulm Med,* 18**,** 6-13.
ZHAO, Y. F., LUO, Y. M., XIONG, W. & WU, X. L. 2014. Genetic variation in ORMDL3 gene may contribute to the risk of asthma: A meta-analysis. *Hum Immunol*.

Appendices

1. Published paper based on this project: Cardenas PA, Cooper PJ, Cox MJ, Chico M, Arias C, et al. (2012) Upper Airways Microbiota in Antibiotic-Naïve Wheezing and Healthy Infants from the Tropics of Rural Ecuador. PLoS ONE 7(10): e46803. doi: 10.1371/journal.pone.0046803.
2. Reprint permissions for figures 1.1, 1.2 and 1.3, 1.4, 2.2 (The rest of figures used from published work are under the rights of the open access journals and therefore do not require formal reprint permission. They however have been properly cited).

OPEN ACCESS Freely available online

PLOS ONE

Upper Airways Microbiota in Antibiotic-Naïve Wheezing and Healthy Infants from the Tropics of Rural Ecuador

Paul Andres Cardenas[1,2,3], Philip J. Cooper[2,4,5], Michael J. Cox[1], Martha Chico[2], Carlos Arias[2], Miriam F. Moffatt[1☯], William Osmond Cookson[1*☯]

1 National Heart & Lung Institute, Imperial College London, London, United Kingdom, **2** Fundación Ecuatoriana para la Investigación en Salud FEPIS, Quinindé, Esmeraldas Province, Ecuador, **3** Institute of Microbiology, Universidad San Francisco de Quito, Quito, Ecuador, **4** Colegio de Ciencias de la Salud, Universidad San Francisco de Quito, Quito, Ecuador, **5** Liverpool School of Tropical Medicine, Liverpool, United Kingdom

Abstract

Background: Observations that the airway microbiome is disturbed in asthma may be confounded by the widespread use of antibiotics and inhaled steroids. We have therefore examined the oropharyngeal microbiome in early onset wheezing infants from a rural area of tropical Ecuador where antibiotic usage is minimal and glucocorticoid usage is absent.

Materials and Methods: We performed pyrosequencing of amplicons of the polymorphic bacterial 16S rRNA gene from oropharyngeal samples from 24 infants with non-infectious early onset wheezing and 24 healthy controls (average age 10.2 months). We analyzed microbial community structure and differences between cases and controls by QIIME software.

Results: We obtained 76,627 high quality sequences classified into 182 operational taxonomic units (OTUs). Firmicutes was the most common and diverse phylum (71.22% of sequences) with *Streptococcus* being the most common genus (49.72%). Known pathogens were found significantly more often in cases of infantile wheeze compared to controls, exemplified by *Haemophilus* spp. (OR = 2.12, 95% Confidence Interval (CI) 1.82–2.47; $P = 5.46 \times 10^{-23}$) and *Staphylococcus* spp. (OR = 124.1, 95%CI 59.0–261.2; $P = 1.87 \times 10^{-241}$). Other OTUs were less common in cases than controls, notably *Veillonella* spp. (OR = 0.59, 95%CI = 0.56–0.62; $P = 8.06 \times 10^{-86}$).

Discussion: The airway microbiota appeared to contain many more Streptococci than found in Western Europe and the USA. Comparisons between healthy and wheezing infants revealed a significant difference in several bacterial phylotypes that were not confounded by antibiotics or use of inhaled steroids. The increased prevalence of pathogens such as *Haemophilus* and *Staphylococcus* spp. in cases may contribute to wheezing illnesses in this age group.

Citation: Cardenas PA, Cooper PJ, Cox MJ, Chico M, Arias C, et al. (2012) Upper Airways Microbiota in Antibiotic-Naïve Wheezing and Healthy Infants from the Tropics of Rural Ecuador. PLoS ONE 7(10): e46803. doi:10.1371/journal.pone.0046803

Editor: Susanne Krauss-Etschmann, Ludwig-Maximilians-University Munich, Germany

Received April 6, 2012; **Accepted** September 10, 2012; **Published** October 5, 2012

Funding: The cohort and collection of samples in Ecuador was supported by the Wellcome Trust (grant number, 088862/Z/09/Z). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: w.cookson@imperial.ac.uk

☯ These authors contributed equally to this work.

Introduction

Asthma is a chronic disease of the airways that is characterized by an abnormal mucosa, intermittent airway inflammation and symptoms of wheezing, dyspnea and cough. The syndrome results from a complex interplay between genetic and environmental factors [1].

A worthwhile understanding of the causes of asthma needs to reconcile consistent epidemiological indications of the importance of the microbiome (also known as microbiota) to the disease [1]. These include the protection afforded by a rich microbial environment in early life [2,3], observations that the bronchial tree contains a characteristic flora that is disturbed by the presence of pathogens such as *Haemophilus infuenzae* in asthma [4,5], birth cohort studies showing that the presence of the same pathogens in throat swabs predicts the later development of asthma [6] and recognition that these bacteria have consistently been associated with exacerbations of asthma [7].

Of potential importance is the finding that organisms commonly found in healthy airways and mucosal surfaces are significantly reduced in asthmatic airways [4]. Investigations of inflammatory bowel disease have shown that a normal bacterial flora is essential in maintaining a healthy mucosa [8] and similar mechanisms are likely to be important in the airways [9–11]. Murine studies have shown that sterility of the airways and intestinal tract results in enhanced inflammatory responses to a variety of stimuli [10,12].

Ninety percent of the cells in the human body are microorganisms including bacteria, parasites and archaea [13]. These microorganisms are commensal on body surfaces exposed to the external environment including the gut, respiratory tract and skin. Allergy, and other immune diseases are associated with differences in microbial communities, but it is unclear if these differences are causes or consequences of disease [4,5,12]. Although most bacteria

Table 1. Epidemiologic characteristics of the children investigated in the study.

	Cases	Controls
Mean age (months)	9.9	10.5
Sex (% male)	50%	42%
Number of Individuals per room of the house	3.57	3.11
Average parental income (USD per month)	186	238
Average Birth weight (grams)	3306	3154
Maternal Education Level (percentage)		
Illiterate	0%	4%
Primary School Incomplete	21%	8%
Primary School Complete	17%	33%
High school Incomplete	42%	46%
High school Complete	21%	4%
University Incomplete	0%	4%
University Complete	0%	0%
Respiratory Tract Infections (average occurrences per month of age)	0.21	0.15
Average Number of Antibiotic Courses	0.166	0.083
Antibiotics (number of times used)		
Trimethoprim/sulfamethoxazole	1	
Cephalexine		1
Amoxicilin	3	1
Antibiotic Use for Respiratory Tract Infection		
Upper RTIs	1	1
Lower RTIs	1	1

doi:10.1371/journal.pone.0046803.t001

are not cultivable with standard methods [14], the membership of complex microbial communities can be quantified and classified by DNA sequencing of the conserved bacterial 16S rRNA gene [15,16]. Bacteria are classified by these sequences into Operational Taxonomic Units (OTUs). OTUs approximate closely but not completely to taxonomy derived from classical techniques, and sequences of other regions may be necessary for precise discrimination at the species level.

Previous studies of the airway microbiome in healthy and in diseased subjects have been carried out in Westernized societies [4–6] where antibiotic use and the prescription of inhaled corticosteroids is almost ubiquitous and confounds understanding of the microbiome. We have therefore carried out a sequence-based study of the upper airway microbiome in children from the Esmeraldas province in rural Ecuador who have had minimal exposure to antibiotic medications and no exposure to inhaled steroids.

Ecuador has strong regional differences in the prevalence of wheeze in rural compared to urban areas. The prevalence of wheezing in children has been estimated to be 16.6% in urban areas [17] while in rural areas of the Pichincha province the rate of current wheezing has been estimated between 0.8% and 2.2% [18,19]. Factors that may be protective against asthma in this region include exposures associated with living in a rural environment, low antibiotic usage and a high rate of geohelminth parasitic infections. We therefore sought to compare and contrast the airway microbiome in infants with non-infective wheeze and healthy controls, and to relate our findings to surveys of the airway microbiome in European children.

Materials and Methods

Subjects

A case-control study was designed to investigate the upper airway microbiota profiles of early onset non-infectious wheezing infants (cases) and healthy infants (controls). The project used DNA extracted from oropharyngeal swabs samples collected from the hypopharynx of 48 infants (average age 10.2 months) that were recruited as part of a birth cohort (the ECUAVIDA cohort) in the Esmeraldas Province in Ecuador. The aim of the ECUAVIDA cohort study is to investigate the effects of early infant infections on the development of immunity, allergic sensitization and allergic disease and the methodology has been previously described in detail [18,19]. The study is an unselected population-based birth cohort that has recruited 2,403 newborns in the rural District of Quininde in the Esmeraldas Province, Ecuador. Detailed data has been collected from the mothers at the time of the first antenatal visit using questionnaires and environmental sampling. The protocol for the ECUAVIDA cohort was approved by the Ethical Committees of the Hospital Pedro Vicente Maldonado and Universidad San Francisco de Quito, Quito, Ecuador.

Twenty-four infants were selected with early onset multiple-trigger non-infectious wheezing according to the GINA guidelines (http://www.ginasthma.org/). Wheezing illness was diagnosed by a physician. Twenty-four healthy controls (no history of wheezing, current respiratory disease, chronic disease or current infections) were selected and paired by age range to cases. The samples were collected when the cases and controls did not have any evidence of a current airway infection (cold symptoms and fever). The infants in both groups had a minimal history of antibiotic use, with 87% never being exposed to antibiotics, and the whole group receiving

on average 0.12 courses of antibiotics per child (Table 1). None of the infants had received antibiotics for any reason for at least two weeks prior to sampling. None of the subjects had ever received corticosteroids. All subjects were of the same mixed ethnic background, lived in the same town and had access to the same basic electricity, water, and sanitation services. All subjects had received vaccines recommended by the Ecuadorian Ministry of Public Health. None had received anti-Streptococcal vaccination.

Sample Collection and Storage

Throat swabs were collected by a physician using sterile cotton swabs and were placed in collection tubes (Qiagen, UK). Sampling was performed carefully without touching any surface other than the oropharynx and using a tongue depressor. Each swab was rubbed approximately five times around the oropharynx. An even pressure was applied and the swab was rotated without interruption. Post sampling, the swab was immediately placed back into the collection tube and stored at $-20^{\circ}C$ and subsequently within the next 24 hours at $-80^{\circ}C$. Samples were shipped to Imperial College, London on dry ice.

Bacterial DNA Extraction from Throat Swabs

Bacterial DNA was extracted from the throat swabs using a modified protocol of the commercial QIAmp DNA Mini Kit (Qiagen). Additional steps at the beginning of the protocol were included to improve the lysis of Gram positive bacteria [4].

Each swab head was transferred into a 2 ml microcentrifuge tube, and 432 µl TE (Tris EDTA pH 8.0)+18 µl 4X lysozyme solution was added (4X lysozyme in a concentration of 1000 U/µl was prepared from lysozyme stock at 30,000 U/µl Ready-Lyse™ Lysozyme Solution of EPICENTRE, UK).

Samples were incubated for 1 hour at $37^{\circ}C$; to allow improved lysis of Gram +ve bacterial walls. During this hour, samples were vortexed for 20 seconds at intervals of 15 minutes. Next 30 µl of Proteinase K and 450 µl of Buffer AL were added to the tube and samples were incubated at $56^{\circ}C$ for 30 minutes. To terminate the Proteinase K step, samples were then incubated for 5 min at $95^{\circ}C$.450 µl of Ethanol (96–100%) was added to the sample. In order to obtain a homogenous solution this was mixed by vortexing. This solution was then applied to the QIAamp Spin Column as per the manufacturer's protocol. In the final step 40 µl of nuclease free water was added instead of the elution buffer supplied by the kit. If the DNA was not being used immediately after extraction samples were stored at $-20^{\circ}C$ until required.

Amplification of 16S rRNA Gene

Polymerase chain reaction (PCR) was used to amplify the variable regions 3 to 5 (V3–V5, Primers 454B_357F and 454A_926R [20]) of the gene that encodes for 16S rRNA in bacteria. Samples were multiplexed using sample specific barcodes [20] and Roche 454 adaptor sequences (Roche Diagnostics, Oakland). To minimize PCR nucleotide insertion mistakes, a high fidelity *Taq* polymerase was used, and samples were amplified in quadruplicate reactions with 20 cycles each and then pooled.

Pyrosequencing and Data Analysis

The DNA amplicons were pyrosequenced using a GS Junior Titanium 454 (Roche Diagnostics, Oakland) following manufacturer's protocols (http://www.gsjunior.com/454-gs-junior.php). Data analysis was performed using the software "Quantitative Insights into Microbial Ecology" (QIIME) [21]. Reads were removed if they were <200 and >800 nucleotides (nt) in length, if there were mismatches in the barcodes or primers, if ambiguous nucleotides were present or if the read quality score was <25. The denoiser algorithm version 1.2.1 [22] was used to avoid overestimation of diversity and chimeras were removed using ChimeraSlayer [23]. Phylogenetic classification was assigned using the Ribosomal Database Project database (RDP). [24] Sequences were clustered in Operational Taxonomic Units (OTUs) using UCLUST version 2.1 at 97% sequence identity [25]. Any sequences present once (singletons) or in only one sample were filtered out.

Remaining sequences were aligned using PyNast [26]. Sequences were rarefied (to remove the heterogeneity of the number of sequences per sample) prior to calculation of alpha and beta diversity statistics. Alpha diversity indexes were calculated in QIIME from rarefied samples using for diversity the Shannon index [27], for richness the Chao1 index [28], and evenness by the equitability index [29]. Beta diversity was calculated using weighted and unweighted UniFrac [30] and principal coordinate analysis (PCoA) performed. Neighbour joining with nearest neighbour interchange phylogenetic trees were created using the representative sequences of each OTU and FastTree version 2.1.3 [31]. The heatmap to represent the abundance of sequences was constructed in iTOL [32]. Microbial community comparisons were performed using parametric statistics in METASTATS, with the P values corrected by multiple hypothesis testing using the false discovery rate (FDR) [20]. Sequences are available in EBI (European Bioinformatics Institute) with the accession number ERP001558.

Representative sequences from significantly different OTUs of further interest were investigated using more intensive phylogenetic approaches in order to maximize the quality of the identification. These test sequences were aligned using the online SINA aligner (http://www.arb-silva.de/aligner/version 1.29 [33]) and this was imported into the ARB phylogenetic software (version 5.1, http://www.arb-home.de/[34]) running on Biolinux 6.0 (http://nebc.nerc.ac.uk/tools/bio-linux/bio-linux-6.0, ref 3 [35]). The aligned SILVA reference database SSU_REF108 of 618,442 high quality 16S rRNA gene sequences was downloaded and merged with the aligned test sequences. All *Haemophilus* spp. (or *Streptococcus* spp.) sequences within the database were selected and the SINA alignment individually checked for each test sequence in the ARB alignment editor. The length of the alignments used depended on the length of the available reference reads for each OTU. Thus, for the *Haemophilus* spp. alignment, the 522 bp region corresponding to the region between positions 384 and 908 of the *Escherichia coli* reference were selected, and for *Streptococcus* spp. 470 bp between positions 470 and 908. Columns of the alignment containing uninformative positions (gaps) were masked from the phylogenetic analysis. Three trees were constructed for each of the genera, an ARB neighbor joining (NJ) tree with 1000 bootstraps, a Maximum Parsimony (MP) tree with 500 bootstraps and a RAxML Maximum Likelihood (ML) tree (version 7.0.3, [36]) with GTR substitution model in rapid hill-climbing mode. Trees were rooted with sequences from near neighbours outside the genus of interest. Tree topology was compared between the three methods and bootstrap values for the NJ and MP trees were used to determine stability of the phylogeny. Accession numbers for the reference sequences used are recorded in the tree labels and those for the outgroup were DQ358146, AJ301681, AF008582, AJ301682, AF008581, AM040491, and AM040495.

Results

Examination of the clinical data showed no significant differences between the 24 cases of non-infectious wheezing and

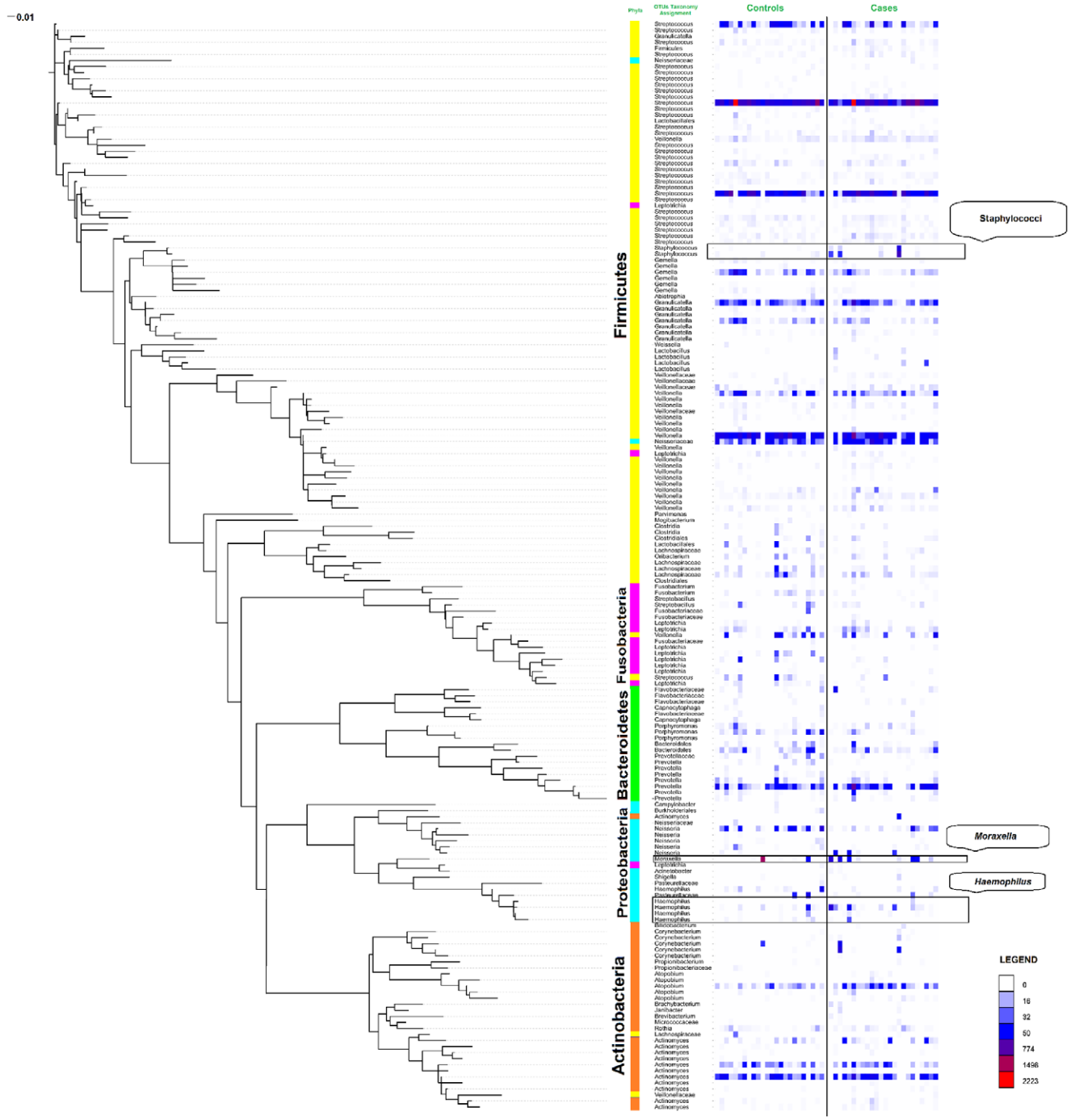


Figure 1. Phylogenetic tree and Heatmap of bacterial 16S rRNA sequences derived from throat swabs. Total sequence counts for individual operational taxonomic units (OTUs) are shown in the right column. Taxonomy assignments at the phylum level are shown in the inner column and colour coded. Intrusions of different colours within particular phyla indicate discrepancies between the phylogenetic and database classifications.
doi:10.1371/journal.pone.0046803.g001

the 24 controls in age, sex and antibiotic use (Table 1). The controls appeared to have higher parental income ($238 per month compared to $186) and slightly fewer individuals per room in their houses (3.1 compared to 3.6), although only 4% of their mothers had finished high school compared to 21% of cases.

An initial 108,042 raw sequences were obtained from all subjects. After denoising, singleton exclusion, chimera checking and removal of OTUs present in only one sample a total of 76,627 sequences remained (37,235 in cases and 39,392 in controls). Between 969 and 6269 sequences were obtained per sample. In order to control sample heterogeneity, sequences were rarified to the same minimum of 969 for all subjects. Consequently, we identified 182 operational taxonomic units (OTUs) at 97%

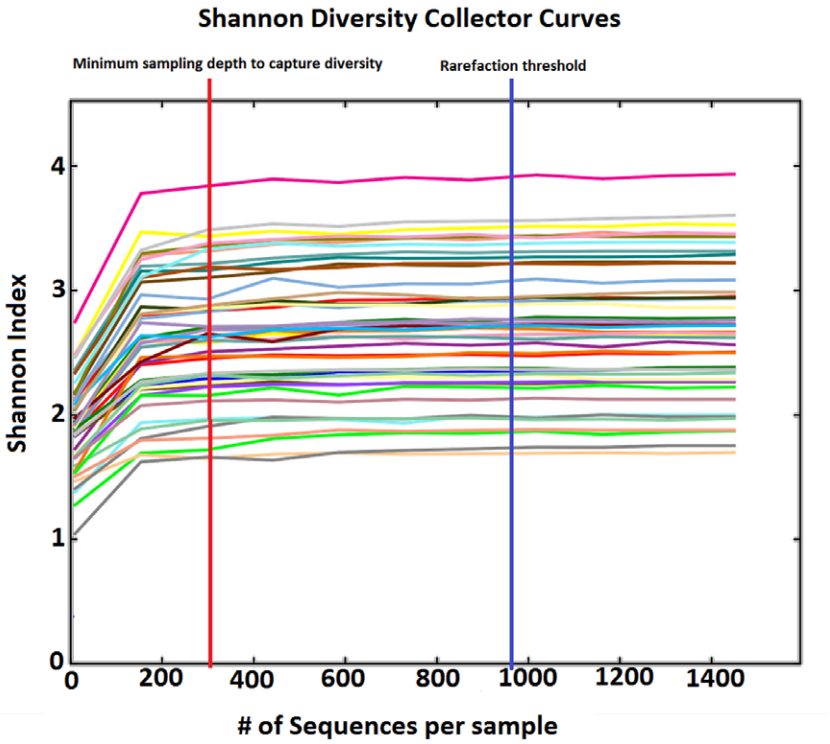


Figure 2. Shannon diversity collector curves. Multiple rarefaction curves were collated from each sample's Shannon diversity index. The graphic shows the estimated diversity plotted against the number of sequences per sample. Each line represents one sample. The plateau in each estimated diversity curve indicates the minimum number of sequences to capture diversity. For all samples the plateau was achieved at approximately 320 sequences (red vertical line), well below our chosen rarefaction threshold of 969 sequences (blue line).
doi:10.1371/journal.pone.0046803.g002

sequence identity level that were assembled into a phylogenetic tree using FastTree and iTOL (Figure 1).

By extrapolation of collectors' curves we estimated that the samples contained an average of 289.8 OTUs (95% CI 218.39–361.30). Multiple rarefaction curves using the Shannon index (Figure 2) showed that a plateau of diversity was achieved in around 360 sequences per sample. This value would be considered as the minimum sampling depth to capture diversity. Our rarefaction was performed at a minimum of 969 sequences per sample, which therefore is a realistic panorama of each sample's diversity. Phylogenetic classification of the ungrouped sequences showed a high prevalence of the phylum Firmicutes (72% of the total number of sequences obtained) followed by Proteobacteria (12%), Actinobacteria (8%), Bacteroidetes (7%) and Fusobacteria (1%) (Figure 1). Firmicutes was the most diverse phylum containing 93 distinct OTUs (51% of the total OTUs), followed by Actinobacteria with 30 OTUs (16%), Proteobacteria with 20 OTUs (11%), then Fusobacteria and Bacteroidetes with 18 and 19 OTUs respectively (10%). *Streptococcus* was the most common genus (49.72% of the total) followed by *Veillonella* (14.5%), *Atopobium* (5.37%) and *Prevotella* (4.72%).

We performed analysis of OTU taxonomy assignments summarized at genus level (when it was not possible to define at genus level, the next best determined taxonomy assignment was used), and the number of samples from where differences were detected at greater than 1% was included (Table 2). We found that members of the groups *Actinomyces* ($P = 1.89 \times 10^{-02}$, OR 1.10), *Atopobium* ($P = 8.99 \times 10^{-20}$, OR 2.27), *Corynebacterium* ($P = 1.37 \times 10^{-129}$, OR 24.99), *Flavobacteriaceae* ($P = 4.02 \times 10^{-31}$, OR 12.07), *Prevotella* ($P = 3.24 \times 10^{-13}$, OR 1.38), *Staphylococcus* ($P = 1.87 \times 10^{-241}$, OR 124.11), *Neisseriaceae* ($P = 5.84 \times 10^{-05}$, OR 1.19), and *Haemophilus* ($P = 5.46 \times 10^{-23}$, OR 2.12) occurred highly significantly more often in the cases of infantile wheeze compared to non-wheezing controls.

By contrast in controls there was a significantly higher prevalence of *Bacteroidales* ($P = 9.57 \times 10^{-08}$, OR 0.55), *Porphyromonas* ($P = 2.81 \times 10^{-32}$, OR 0.20), *Gemella* ($P = 4.29 \times 10^{-21}$, OR 0.40), *Lachnospiraceae* ($P = 7.79 \times 10^{-14}$, OR 0.39), *Veillonella* ($P = 8.06 \times 10^{-86}$, OR 0.59), *Leptotrichia* ($P = 9.37 \times 10^{-14}$, OR 0.42), *Pasteurellaceae* ($P = 1.13 \times 10^{-20}$, OR 0.20) and *Moraxella* ($P = 4.54 \times 10^{-06}$, OR 0.79) (Table 2). We repeated the statistical analysis excluding the 6 children that had a minimal use of antibiotics and obtained similar results to the full sample set.

Inspection of the data for individual subjects (Figure 1) showed that the high abundance of *Moraxella* sp. in a single sample resulted in the whole genus being significantly more prevalent in controls. If this particular sample was excluded from the analysis, *Moraxella* became more prevalent in cases.

The biological interpretation of the differences in the frequencies of individual OTUs is limited by the imprecision of OTU assignments in identifying individual species. In particular, this problem was obvious in our data with OTUs assigned to *Streptococcus* and *Haemophilus spp.*, which were likely to contain a mixture of pathogenic and non-pathogenic strains. We therefore attempted to improve the classification of these OTUs by including them in phylogenetic tress constructed from reference sequences.

Using three independent phylogenetic treeing methods it was not possible to increase the specificity of the identification of the *Streptococcus* spp. OTUs beyond that of the basic Ribosomal Database Project (RDP) classifier. Tree topology between the

Table 2. Differences in bacterial 16S rRNA sequences from throat swabs of infants in rural Ecuador with non-infectious wheeze and healthy controls.

Groups*	Number of Sequences		Number of subjects (%) in which OTU groups were detected at >1%		*P* value	Odds Ratio	95% CI
	Controls	Cases	Controls	Cases			
Actinobacteria/Actinomyces	1192	1306	23 (96%)	21 (88%)	1.89×10^{-02}	1.10	1.02 to 1.20
Actinobacteria/Atopobium	173	388	19 (79%)	20 (83%)	8.99×10^{-20}	2.27	1.89 to 2.71
Actinobacteria/Corynebacterium	27	656	6 (25%)	7 (29%)	1.37×10^{-129}	25.0	17.0 to 36.7
Bacteroidetes/Bacteroidales	227	126	18 (75%)	18 (75%)	9.57×10^{-08}	0.55	0.44 to 0.69
Bacteroidetes/Flavobacteriaceae	19	167	7 (29%)	8 (33%)	4.02×10^{-31}	12.1	7.55 to 19.3
Bacteroidetes/Porphyromonas	264	53	19 (79%)	16 (67%)	2.81×10^{-32}	0.20	0.15 to 0.27
Bacteroidetes/Prevotella	930	1263	23 (96%)	20 (83%)	3.24×10^{-13}	1.38	1.27 to 1.50
Firmicutes/Gemella	353	143	19 (79%)	22 (92%)	4.29×10^{-21}	0.40	0.33 to 0.49
Firmicutes/Lachnospiraceae	209	81	18 (75%)	15 (62%)	7.79×10^{-14}	0.39	0.30 to 0.50
Firmicutes/Staphylococcus	7	837	6 (25%)	5 (21%)	1.87×10^{-241}	124.1	59.0 to 261.2
Firmicutes/Veillonella	4117	2623	23 (96%)	23 (96%)	8.06×10^{-86}	0.59	0.56 to 0.62
Fusobacteria/Leptotrichia	236	99	17 (71%)	17 (71%)	9.37×10^{-14}	0.42	0.33 to 0.53
Proteobacteria/Haemophilus	248	520	14 (58%)	17 (71%)	5.46×10^{-23}	2.12	1.82 to 2.47
Proteobacteria/Moraxella	931	745	5 (21%)	9 (38%)	4.54×10^{-06}	0.79	0.72 to 0.88
Proteobacteria/Neisseriaceae	1032	1218	16 (67%)	18 (75%)	5.84×10^{-05}	1.19	1.09 to 1.30
Proteobacteria/Pasteurellaceae	163	32	4 (17%)	4 (17%)	1.13×10^{-20}	0.20	0.13 to 0.29

*Results are shown for the lowest level of taxonomic identification achieved. Only groups with more than 100 sequences and statistically significant differences are shown. *P* values are corrected for multiple comparisons.
doi:10.1371/journal.pone.0046803.t002

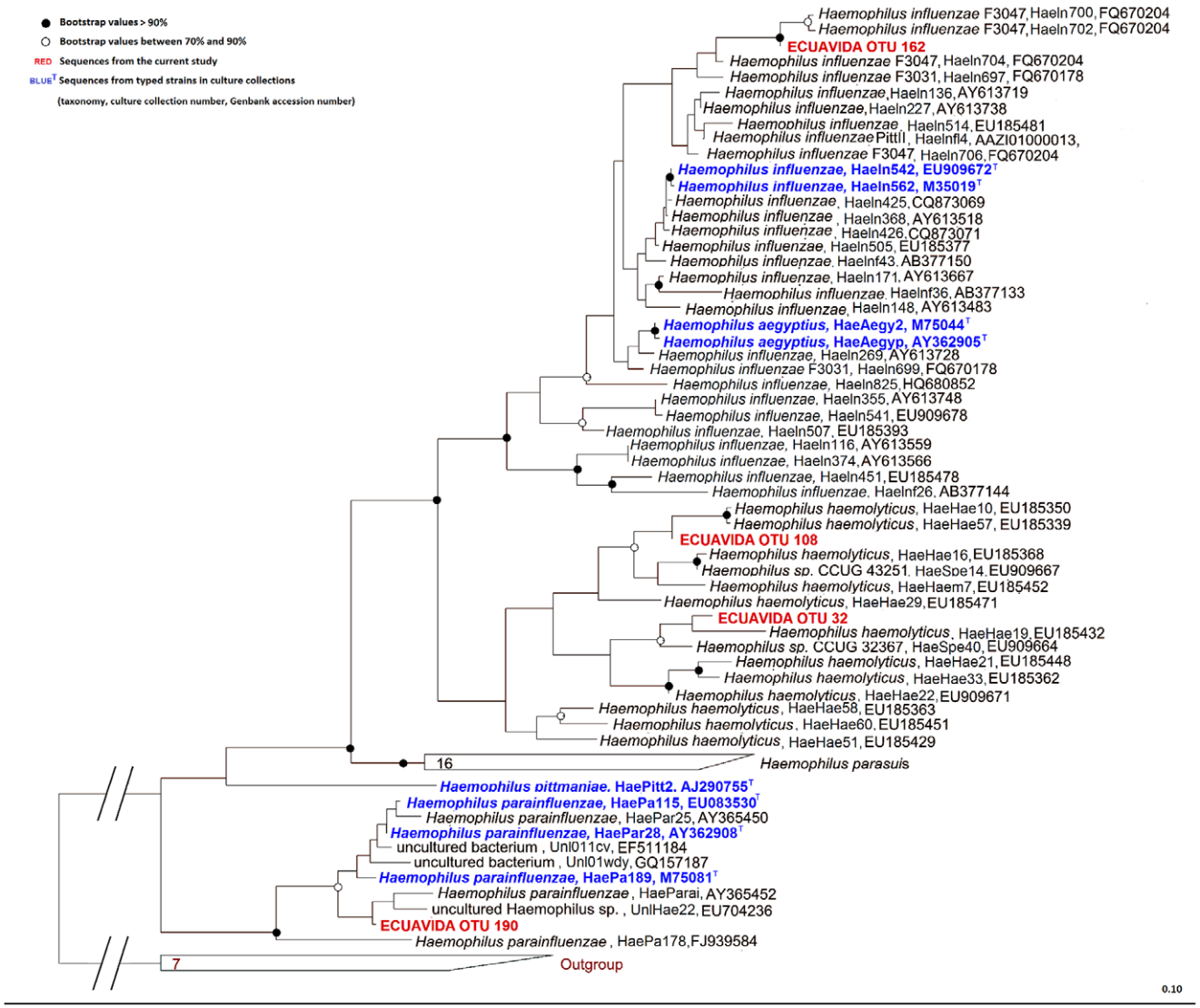


Figure 3. Phylogenetic identification of *Haemophilus* OTUs. Phylogenetic analysis of the 16S rRNA sequences of the OTUs assigned taxonomically to *Haemophilus* genus (OTUs 32, 108, 162 and 190, shown in Red) together with reference *Haemophilus* sequences from the SILVA database, using the ARB alignment editor. The scale bar indicates 10% sequence divergence, and NCBI accession numbers are included. The tree was rooted with a near neighbour outgroup constructed with sequences from *Morganella morganii, Proteus mirabilis* and *Providencia stuartii*.
doi:10.1371/journal.pone.0046803.g003

three methods of treeing was not conserved and significance of assignment to major nodes was low.

By contrast, the tree topology for *Haemophilus* spp. was conserved in all three methods of phylogenetic inference (Figure 3), with robust significance for assignment for each *Haemophilus* spp. OTU. This enabled confident assignment of OTU 162 to *Haemophilus influenzae*, OTU 32 and 38 to *Haemophilus haemolyticus*, and OTU 190 to *Haemophilus parainfluenzae*. OTU 162 which was assigned to the known pathogen *Haemophilus influenzae* was significantly more common in cases ($P = 1.66 \times 10^{-49}$, OR 3.45) compared with controls, whilst OTU 190 (assigned to *Haemophilus parainfluenzae*) was significantly more prevalent in controls ($P = 2.74 \times 10^{-48}$, OR 0.05).

We next tested if changes in the abundance of individual OTUs were associated with alterations in the overall structure of the microbial populations between cases and controls. We did not detect any differences between cases and controls in species richness, taxa abundances, or evenness (Figure 4 A, B and C). We found no large-scale differences in microbial community cluster patterns between cases and controls (beta diversity) using the principal coordinate analysis (PCoA) of the UniFrac distance matrix (weighted and un-weighted) (Figure 4, D and E).

Discussion

This study has shown significant perturbations of the airways microbiota in infants with early onset non-infectious wheeze from a rural district in the tropics of Ecuador. Previous studies conducted in a geographically adjacent area with similar climatic, population, and geographic characteristics estimated the prevalence of asthma in school age children to be just 2.2% [18] and our subjects were characterized by minimal use of antibiotics and a complete absence of use of corticosteroids and anti-Streptococcal vaccination. Our samples may therefore exemplify a naturally developing microbial community in the early months of life. Bacterial diversity in these infants (estimated mean 290 OTUs)

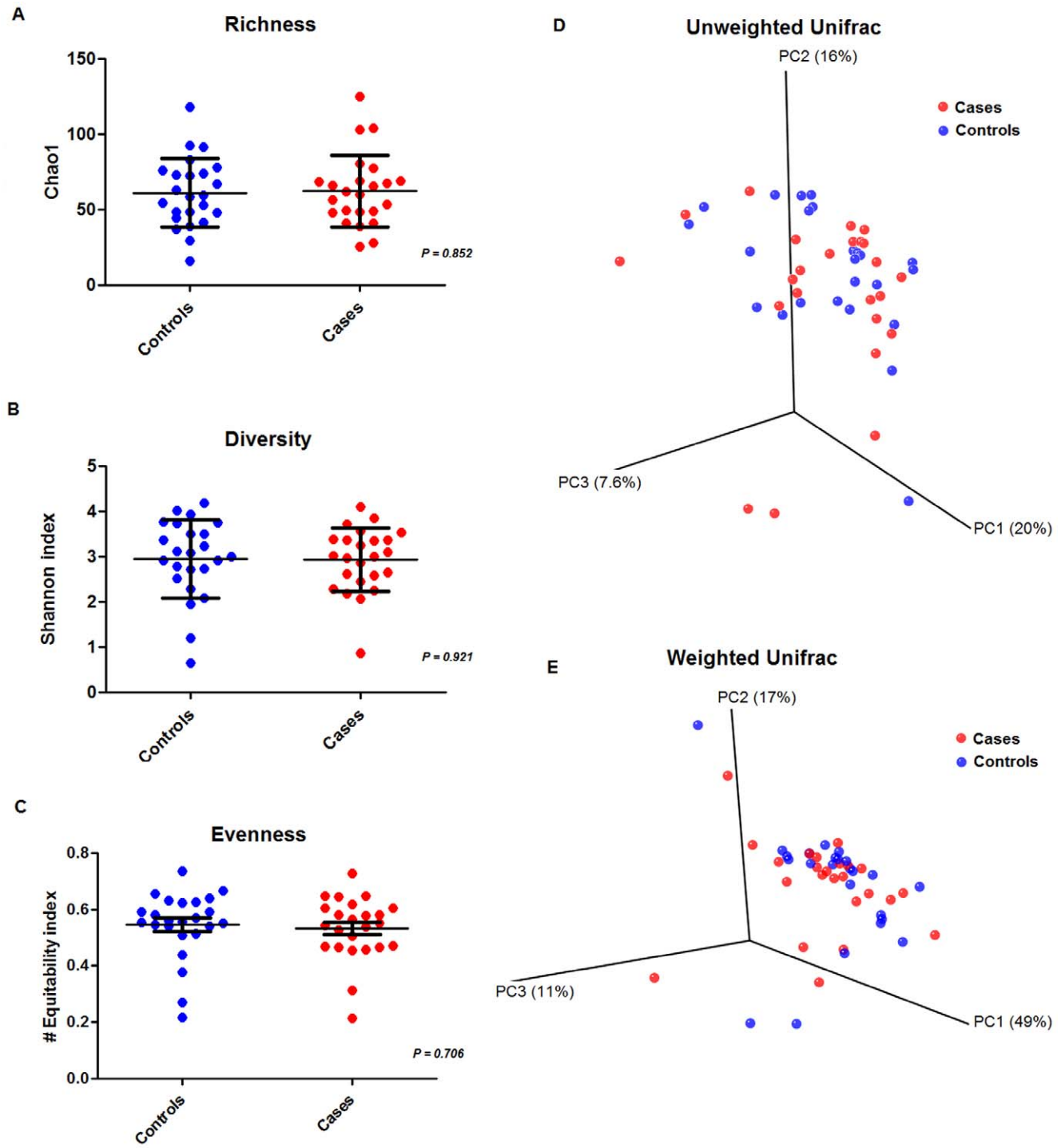


Figure 4. Alpha and Beta diversity comparisons between Cases and Controls. A) Scatter dot plot comparing cases versus controls values of chao1 richness index. B) Scatter dot plot comparing values of Shannon diversity index. C) Scatter dot plot comparing equitability evenness index. D) Unweighted UNIFRAC Principal Coordinate Analysis (PCoA) plot comparing presence/absence metrics. E) Weighted UNIFRAC Principal Coordinate Analysis (PCoA) plot comparing presence/absence metrics and abundance.
doi:10.1371/journal.pone.0046803.g004

may be higher than that in seen previously in European children (mean 85 OTUs) [4], although differences in methodology mean that this inference should be treated with caution.

It is of interest that Firmicutes was the most common and most diverse phylum with *Streptococcus* the most prevalent genus, consistent with observations of the airways microbiota in European children [4]. We were not able to differentiate well between members of the important *Streptococcus* group on the basis of the 16S sequences, and the typing of alternative chronometers (such as *pheS*, *rpoA*, and *gyrB*) and the incorporation of mutilocus sequence analysis (MLS) [37] will be important in future studies of the airway microbiome.

We defined our phenotype in infants by non-febrile episodic wheezing, according to the GINA guidelines (http://www.ginasthma.org/). These guidelines recognize that asthma diagnosis before the age of 6 years is complicated by the difficulty in performing accurate lung function tests. The recognition of recurrent wheezing not related to infections is as a consequence the most important diagnostic indicator of asthma in this age group [38]. Because not all children who experience wheezing events will develop asthma and not all children with asthma wheeze [39], we have followed the recommendation that the phenotype of recurrent non-infectious wheezing syndrome be applied to children less than 24 months old [38,39].

We used culture-independent molecular techniques to characterize microbial communities and to quantitatively investigate dissimilarities. We, and others have shown previously that the microbiome of the oropharnyx correlates with that of the bronchial tree assessed by brushings or by lavage [4,40], and therefore oropharyngeal swabs were used for this epidemiological survey. Pyrosequencing robustly determines the diversity and abundance of microbial communities in a quantitative and qualitative form. The assignment of approximately 1,600 individual sequences for each subject has conferred substantial statistical power to our study. We have reduced the possibility of bias by stringent removal of chimeric sequences, sequences present only once in the dataset and OTUs that were present in only one sample.

We were aware that a high abundance of sequences in few samples could drive the overall prevalence of bacteria in cases or controls, as occurred with *Moraxella* spp. in this study. We have therefore confined our tests for significant differences to OTUs present in three or more subjects with a total of more than 100 sequences per sample.

Our study identified a higher frequency of potential pathogens (*Neisseriaceae*, *Prevotella*, *Corynebacterium*, *Staphylococcus*, *Actinomyces* and *Haemophilus*) in wheezy infants compared to healthy controls. This finding is consistent with substantial epidemiological studies of European neonates that used standard bacterial cultures to show carriage of pathogens in neonates predicted later asthma risk [6]. The finding is also consistent with earlier studies in older children and adults with asthma that used 16S rRNA gene sequencing for bacterial characterisation [4].

The incorporation of reference strain sequences into our phylogenetic trees allowed us to discriminate between *Haemophilus* spp. OTUs at species level, and to show that the pathogen *H. influenzae* was more prevalent in wheezing infants, in concordance with previous studies [6]. *H. parainfluenzae* was more abundant in healthy children which might be associated with wheezing protection. The 16S rRNA gene has previously been suggested for use as a marker in MLST of *H. influenzae* [41] and our results confirm that its phylogeny is well matched to species and strain identification in airway samples.

In this study we found a lack of potentially 'protective' bacterial genera in non-infectious wheezing infants compared with controls, particularly *Veillonella*, *Pasteurellaceae* and *Gemella*. Alterations in the normal microbiota may alter the host resistance to pathogen colonization in the gut [42] and in the airways [43]. Possible mechanisms for this include direct inhibition of pathogen growth by commensal secreted factors [44]. Commensal bacteria may also elicit tonic signals in the gut epithelium that prevent activation of innate and adaptive immune responses [45,46]. It may be relevant that manipulating the airway microbiome in gnotobiotic mice may produce major changes in airway responsiveness and immunity [10,11].

Our results suggest that the upper airways microbiota of early onset wheezing infants from the tropics exhibit an increase in the frequency of pathogenic bacterial OTUs that is not confounded by antibiotic or steroid medications. These specific differences do not appear to have been accompanied by detectable community changes in the numbers of species and their overall relative distribution in our samples. Larger studies and direct measurements of the communities in the lower airways may change this perception. The results provide further support for a hypothesis that pathogenic bacteria may contribute to a wheezy diathesis in infants [6].

Author Contributions

Conceived and designed the experiments: PAC PJC MFM WOC. Performed the experiments: PAC. Analyzed the data: PAC MJC. Contributed reagents/materials/analysis tools: PJC MFM WOC. Wrote the paper: PAC PJC MFM WOC. Samples collection: PCA MC CA.

References

1. Cookson W (2004) The immunogenetics of asthma and eczema: a new focus on the epithelium. Nat Rev Immunol 4: 978–988.
2. Eder W, Ege MJ, von Mutius E (2006) The asthma epidemic. N Engl J Med 355: 2226–2235.
3. Ege MJ, Mayer M, Normand AC, Genuneit J, Cookson WO, et al. (2011) Exposure to environmental microorganisms and childhood asthma. N Engl J Med 364: 701–709.
4. Hilty M, Burke C, Pedro H, Cardenas P, Bush A, et al. (2010) Disordered microbial communities in asthmatic airways. PLoS One 5: e8578.
5. Huang YJ, Nelson CE, Brodie EL, Desantis TZ, Baek MS, et al. (2011) Airway microbiota and bronchial hyperresponsiveness in patients with suboptimally controlled asthma. J Allergy Clin Immunol 127: 372–381 e373.
6. Bisgaard H, Hermansen MN, Buchvald F, Loland L, Halkjaer LB, et al. (2007) Childhood asthma after bacterial colonization of the airway in neonates. N Engl J Med 357: 1487–1495.
7. Kraft M (2000) The role of bacterial infections in asthma. Clin Chest Med 21: 301–313.
8. Maslowski KM, Vieira AT, Ng A, Kranich J, Sierro F, et al. (2009) Regulation of inflammatory responses by gut microbiota and chemoattractant receptor GPR43. Nature 461: 1282–1286.
9. Noverr MC, Huffnagle GB (2005) The 'microflora hypothesis' of allergic diseases. Clin Exp Allergy 35: 1511–1520.
10. Herbst T, Sichelstiel A, Schar C, Yadava K, Burki K, et al. (2011) Dysregulation of Allergic Airway Inflammation in the Absence of Microbial Colonization. Am J Respir Crit Care Med.
11. Nembrini C, Sichelstiel A, Kisielow J, Kurrer M, Kopf M, et al. (2011) Bacterial-induced protection against allergic inflammation through a multicomponent immunoregulatory mechanism. Thorax.
12. Noverr MC, Huffnagle GB (2004) Does the microbiota regulate immune responses outside the gut? Trends Microbiol 12: 562–568.
13. Savage DC (1977) Microbial ecology of the gastrointestinal tract. Annu Rev Microbiol 31: 107–133.
14. Staley JT, Konopka A (1985) Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol 39: 321–346.
15. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, et al. (2007) The human microbiome project. Nature 449: 804–810.
16. Ahmed S, Macfarlane GT, Fite A, McBain AJ, Gilbert P, et al. (2007) Mucosa-associated bacterial diversity in relation to human terminal ileum and colonic biopsy samples. Appl Environ Microbiol 73: 7435–7442.
17. Mallol J, Garcia-Marcos L, Sole D, Brand P (2010) International prevalence of recurrent wheezing during the first year of life: variability, treatment patterns and use of health resources. Thorax 65: 1004–1009.
18. Weinmayr G, Weiland SK, Bjorksten B, Brunekreef B, Buchele G, et al. (2007) Atopic sensitization and the international variation of asthma symptom prevalence in children. Am J Respir Crit Care Med 176: 565–574.
19. Cooper PJ, Rodrigues LC, Cruz AA, Barreto ML (2009) Asthma in Latin America: a public heath challenge and research opportunity. Allergy 64: 5–17.
20. White JR, Nagarajan N, Pop M (2009) Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput Biol 5: e1000352.

21. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Meth 7: 335–336.
22. Reeder J, Knight R (2010) Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. Nat Meth 7: 668–669.
23. Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, et al. (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res 21: 494–504.
24. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, et al. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37: D141–145.
25. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics.
26. Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, et al. (2010) PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics 26: 266–267.
27. Shannon CE (1948) A mathematical theory of communication. Bell System Technical Journal 27: 379–423 and 623–656.
28. Chao A (1984) Nonparametric Estimation of the Number of Classes in a Population. Scandinavian Journal of Statistics 11: 265–270.
29. Magurran AE (1988) Ecological Diversity and its Measurement. Princeton, NJ: Princeton University Press.
30. Lozupone C, Hamady M, Knight R (2006) UniFrac–an online tool for comparing microbial community diversity in a phylogenetic context. BMC Bioinformatics 7: 371.
31. Price MN, Dehal PS, Arkin AP (2010) FastTree 2– Approximately Maximum-Likelihood Trees for Large Alignments. PLoS One 5: e9490.
32. Letunic I, Bork P (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res.
33. Pruesse E, Peplies J, Glöckner FO (2012) SINA: accurate high throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics.
34. Ludwig W, Strunk O, Westram R, Richter L, Meier H, et al. (2004) ARB: a software environment for sequence data. Nucleic Acids Res 32: 1363–1371.
35. Field D, Tiwari B, Booth T, Houten S, Swan D, et al. (2006) Open software for biologists: from famine to feast. Nat Biotech 24: 801–803.
36. Stamatakis A, Ludwig T, Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21: 456–463.
37. Bishop CJ, Aanensen DM, Jordan GE, Kilian M, Hanage WP, et al. (2009) Assigning strains to bacterial species via the internet. BMC Biol 7: 3.
38. Wechsler ME (2009) Managing asthma in primary care: putting new guideline recommendations into context. Mayo Clin Proc 84: 707–717.
39. Just J, Belfar S, Wanin S, Pribil C, Grimfeld A, et al. (2010) Impact of innate and environmental factors on wheezing persistence during childhood. J Asthma 47: 412–416.
40. Charlson ES, Bittinger K, Haas AR, Fitzgerald AS, Frank I, et al. Topographical Continuity of Bacterial Populations in the Healthy Human Respiratory Tract. Am J Respir Crit Care Med.
41. Sacchi CT, Alber D, Dull P, Mothershed EA, Whitney AM, et al. (2005) High Level of Sequence Diversity in the 16S rRNA Genes of Haemophilus influenzae Isolates Is Useful for Molecular Subtyping. Journal of Clinical Microbiology 43: 3734–3742.
42. Artis D (2008) Epithelial-cell recognition of commensal bacteria and maintenance of immune homeostasis in the gut. Nat Rev Immunol 8: 411–420.
43. Crowe CC, Sanders WE Jr, Longley S (1973) Bacterial interference. II. Role of the normal throat flora in prevention of colonization by group A Streptococcus. J Infect Dis 128: 527–532.
44. Tano K, Grahn Hakansson E, Wallbrandt P, Ronnqvist D, Holm SE, et al. (2003) Is hydrogen peroxide responsible for the inhibitory activity of alpha-haemolytic streptococci sampled from the nasopharynx? Acta Otolaryngol 123: 724–729.
45. Macpherson AJ, Harris NL (2004) Interactions between commensal intestinal bacteria and the immune system. Nat Rev Immunol 4: 478–485.
46. Mason KL, Huffnagle GB (2009) Control of mucosal polymicrobial populations by innate immunity. Cell Microbiol 11: 1297–1305.

NATURE PUBLISHING GROUP LICENSE TERMS AND CONDITIONS

Sep 25, 2014

This is a License Agreement between Paul Cardenas ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3476131238490
License date	Sep 25, 2014
Licensed content publisher	Nature Publishing Group
Licensed content publication	Nature Medicine
Licensed content title	Asthma phenotypes: the evolution from clinical to molecular approaches
Licensed content author	Sally E Wenzel
Licensed content date	May 4, 2012
Volume number	18
Issue number	5
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Figures	figure 1.1
Author of this NPG article	no
Your reference number	None
Title of your thesis / dissertation	Upper airways microbiota profiling in a case/control study between wheezing and healthy children from the tropics of Ecuador
Expected completion date	Oct 2014
Estimated size (number of pages)	300
Total	0.00 GBP
Terms and Conditions	

Terms and Conditions for Permissions

Nature Publishing Group hereby grants you a non-exclusive license to reproduce this material for this purpose, and for no other use,subject to the conditions below:

1. NPG warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to Nature Publishing Group and does not carry the copyright of another entity (as credited in the published version). If the credit line on any part of the material you have requested indicates that it was reprinted or adapted by NPG with permission from another source, then you should also seek permission from that source to reuse the material.

2. Permission granted free of charge for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to the work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.Where print permission has been granted for a fee, separate permission must be obtained for any additional, electronic re-use (unless, as in the case of a full paper, this has already been accounted for during your initial request in the calculation of a print run).NB: In all cases, web-based use of full-text articles must be authorized separately through the 'Use on a Web Site' option when requesting permission.

3. Permission granted for a first edition does not apply to second and subsequent editions and for editions in other languages (except for signatories to the STM Permissions Guidelines, or where the first edition permission was granted for free).

4. Nature Publishing Group's permission must be acknowledged next to the figure, table or abstract in print. In electronic form, this acknowledgement must be visible at the same time as the figure/table/abstract, and must be hyperlinked to the journal's homepage.

5. The credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)
   For AOP papers, the credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

   Note: For republication from the *British Journal of Cancer,* the following credit lines apply.
   Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)For AOP papers, the credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

6. Adaptations of single figures do not require NPG approval. However, the adaptation should be credited as follows:

   Adapted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)

   Note: For adaptation from the *British Journal of Cancer,* the following credit line applies.
   Adapted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)

7. Translations of 401 words up to a whole article require NPG approval. Please visit http://www.macmillanmedicalcommunications.com for more information.Translations of up

to a 400 words do not require NPG approval. The translation should be credited as follows:

Translated by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication).

Note: For translation from the *British Journal of Cancer,* the following credit line applies.
Translated by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)

We are certain that all parties will benefit from this agreement and wish you the best in the use of this material. Thank you.

Special Terms:

v1.1

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

Sep 25, 2014

This is a License Agreement between Paul Cardenas ("You") and John Wiley and Sons ("John Wiley and Sons") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by John Wiley and Sons, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3476140496791
License date	Sep 25, 2014
Licensed content publisher	John Wiley and Sons
Licensed content publication	Allergy
Licensed content title	The global burden of asthma: executive summary of the GINA Dissemination Committee Report
Licensed copyright line	Copyright © 2004, John Wiley and Sons
Licensed content author	Matthew Masoli,Denise Fabian,Shaun Holt,Richard Beasley
Licensed content date	Apr 6, 2004
Start page	469
End page	478
Type of use	Dissertation/Thesis
Requestor type	University/Academic
Format	Print and electronic
Portion	Figure/table
Number of figures/tables	2
Original Wiley figure/table number(s)	figure 1 and 4
Will you be translating?	No
Title of your thesis / dissertation	Upper airways microbiota profiling in a case/control study between wheezing and healthy children from the tropics of Ecuador
Expected completion date	Oct 2014
Expected size (number of pages)	300
Total	0.00 GBP
Terms and Conditions	

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your Rightslink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.

- You are hereby granted a personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this licence must be completed within two years of the date of the grant of this licence (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.

- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner. You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a stand-alone basis, or any of the rights granted to you hereunder to any other person.

- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc, the Wiley Companies, or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you

shall not assert any such right, license or interest with respect thereto.

- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU

- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

- This Agreement may not be assigned (including by operation of law or otherwise) by

you without WILEY's prior written consent.

- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.

- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

- WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

WILEY OPEN ACCESS TERMS AND CONDITIONS

Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles under the terms of the Creative Commons Attribution (CC BY) License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses:: Creative Commons Attribution (CC-BY) license Creative Commons Attribution Non-Commercial (CC-BY-NC) license and Creative Commons Attribution Non-Commercial-NoDerivs (CC-BY-NC-ND) License. The license type is clearly identified on the article.

Copyright in any research article in a journal published as Open Access under a Creative Commons License is retained by the author(s). Authors grant Wiley a license to publish the article and identify itself as the original publisher. Authors also grant any third party the right to use the article freely as long as its integrity is maintained and its original authors,

citation details and publisher are identified as follows: [Title of Article/Author/Journal Title and Volume/Issue. Copyright (c) [year] [copyright owner as specified in the Journal]. Links to the final article on Wiley's website are encouraged where applicable.

The Creative Commons Attribution License

The Creative Commons Attribution License (CC-BY) allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CC-BY license permits commercial and non-commercial re-use of an open access article, as long as the author is properly attributed.

The Creative Commons Attribution License does not affect the moral rights of authors, including without limitation the right not to have their work subjected to derogatory treatment. It also does not affect any other rights held by authors or third parties in the article, including without limitation the rights of privacy and publicity. Use of the article must not assert or imply, whether implicitly or explicitly, any connection with, endorsement or sponsorship of such use by the author, publisher or any other party associated with the article.

For any reuse or distribution, users must include the copyright notice and make clear to others that the article is made available under a Creative Commons Attribution license, linking to the relevant Creative Commons web page.

To the fullest extent permitted by applicable law, the article is made available as is and without representation or warranties of any kind whether express, implied, statutory or otherwise and including, without limitation, warranties of title, merchantability, fitness for a particular purpose, non-infringement, absence of defects, accuracy, or the presence or absence of errors.

Creative Commons Attribution Non-Commercial License

The Creative Commons Attribution Non-Commercial (CC-BY-NC) License permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.(see below)

Creative Commons Attribution-Non-Commercial-NoDerivs License

The Creative Commons Attribution Non-Commercial-NoDerivs License (CC-BY-NC-ND) permits use, distribution and reproduction in any medium, provided the original work is properly cited, is not used for commercial purposes and no modifications or adaptations are made. (see below)

Use by non-commercial users

For non-commercial and non-promotional purposes, individual users may access, download, copy, display and redistribute to colleagues Wiley Open Access articles, as well as adapt, translate, text- and data-mine the content subject to the following conditions:

- The authors' moral rights are not compromised. These rights include the right of "paternity" (also known as "attribution" - the right for the author to be identified as such) and "integrity" (the right for the author not to have the work altered in such a way that the author's reputation or integrity may be impugned).

- Where content in the article is identified as belonging to a third party, it is the obligation of the user to ensure that any reuse complies with the copyright policies of the owner of that content.

- If article content is copied, downloaded or otherwise reused for non-commercial research and education purposes, a link to the appropriate bibliographic citation (authors, journal, article title, volume, issue, page numbers, DOI and the link to the definitive published version on **Wiley Online Library**) should be maintained. Copyright notices and disclaimers must not be deleted.

- Any translations, for which a prior translation agreement with Wiley has not been agreed, must prominently display the statement: "This is an unofficial translation of an article that appeared in a Wiley publication. The publisher has not endorsed this translation."

Use by commercial "for-profit" organisations

Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee. Commercial purposes include:

- Copying or downloading of articles, or linking to such articles for further redistribution, sale or licensing;

- Copying, downloading or posting by a site or service that incorporates advertising with such content;

- The inclusion or incorporation of article content in other works or services (other than normal quotations with an appropriate citation) that is then available for sale or licensing, for a fee (for example, a compilation produced for marketing purposes, inclusion in a sales pack)

- Use of article content (other than normal quotations with appropriate citation) by for-profit organisations for promotional purposes

- Linking to article content in e-mails redistributed for promotional, marketing or educational purposes;

- Use for the purposes of monetary reward by means of sale, resale, licence, loan, transfer or other form of commercial exploitation such as marketing products

- Print reprints of Wiley Open Access articles can be purchased from: corporatesales@wiley.com

Further details can be found on Wiley Online Library
http://olabout.wiley.com/WileyCDA/Section/id-410895.html

Other Terms and Conditions:

v1.9

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

SPRINGER LICENSE
TERMS AND CONDITIONS

Sep 25, 2014

This is a License Agreement between Paul Cardenas ("You") and Springer ("Springer") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Springer, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3476250347368
License date	Sep 25, 2014
Licensed content publisher	Springer
Licensed content publication	Springer eBook
Licensed content title	Discovering New Pathogens
Licensed content author	Andrew J. Lawson
Licensed content date	Jan 1, 2004
Type of Use	Thesis/Dissertation
Portion	Figures
Author of this Springer article	No
Order reference number	None
Original figure numbers	figure 1
Title of your thesis / dissertation	Upper airways microbiota profiling in a case/control study between wheezing and healthy children from the tropics of Ecuador
Expected completion date	Oct 2014
Estimated size(pages)	300
Total	0.00 GBP
Terms and Conditions	

Introduction

The publisher for this copyrighted material is Springer Science + Business Media. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

Limited License

With reference to your request to reprint in your thesis material on which Springer Science

and Business Media control the copyright, permission is granted, free of charge, for the use indicated in your enquiry.

Licenses are for one-time use only with a maximum distribution equal to the number that you identified in the licensing process.

This License includes use in an electronic form, provided its password protected or on the university's intranet or repository, including UMI (according to the definition at the Sherpa website: http://www.sherpa.ac.uk/romeo/). For any other electronic use, please contact Springer at (permissions.dordrecht@springer.com or permissions.heidelberg@springer.com).

The material can only be used for the purpose of defending your thesis limited to university-use only. If the thesis is going to be published, permission needs to be re-obtained (selecting "book/textbook" as the type of use).

Although Springer holds copyright to the material and is entitled to negotiate on rights, this license is only valid, subject to a courtesy information to the author (address is given with the article/chapter) and provided it concerns original material which does not carry references to other sources (if material in question appears with credit to another source, authorization from that source is required as well).

Permission free of charge on this occasion does not prejudice any rights we might have to charge for reproduction of our copyrighted material in the future.

Altering/Modifying Material: Not Permitted
You may not alter or modify the material in any manner. Abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of the author(s) and/or Springer Science + Business Media. (Please contact Springer at (permissions.dordrecht@springer.com or permissions.heidelberg@springer.com)

Reservation of Rights
Springer Science + Business Media reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

Copyright Notice:Disclaimer
You must include the following copyright and permission notice in connection with any reproduction of the licensed material: "Springer and the original publisher /journal title, volume, year of publication, page, chapter/article title, name(s) of author(s), figure number(s), original copyright notice) is given to the publication in which the material was originally published, by adding; with kind permission from Springer Science and Business Media"

Warranties: None

Example 1: Springer Science + Business Media makes no representations or warranties with respect to the licensed material.

Example 2: Springer Science + Business Media makes no representations or warranties with respect to the licensed material and adopts on its own behalf the limitations and disclaimers established by CCC on its behalf in its Billing and Payment terms and conditions for this

licensing transaction.

Indemnity
You hereby indemnify and agree to hold harmless Springer Science + Business Media and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

No Transfer of License
This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without Springer Science + Business Media's written permission.

No Amendment Except in Writing
This license may not be amended except in a writing signed by both parties (or, in the case of Springer Science + Business Media, by CCC on Springer Science + Business Media's behalf).

Objection to Contrary Terms
Springer Science + Business Media hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and Springer Science + Business Media (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

Jurisdiction
All disputes that may arise in connection with this present License, or the breach thereof, shall be settled exclusively by arbitration, to be held in The Netherlands, in accordance with Dutch law, and to be conducted under the Rules of the 'Netherlands Arbitrage Instituut' (Netherlands Institute of Arbitration).***OR:***

All disputes that may arise in connection with this present License, or the breach thereof, shall be settled exclusively by arbitration, to be held in the Federal Republic of Germany, in accordance with German law.

Other terms and conditions:

v1.3

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

OXFORD UNIVERSITY PRESS LICENSE TERMS AND CONDITIONS

Sep 25, 2014

This is a License Agreement between Paul Cardenas ("You") and Oxford University Press ("Oxford University Press") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Oxford University Press, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3476271343529
License date	Sep 25, 2014
Licensed content publisher	Oxford University Press
Licensed content publication	Human Molecular Genetics
Licensed content title	Sequencing the human microbiome in health and disease:
Licensed content author	Michael J. Cox, William O.C.M. Cookson, Miriam F. Moffatt
Licensed content date	10/15/2013
Type of Use	Thesis/Dissertation
Institution name	None
Title of your work	Upper airways microbiota profiling in a case/control study between wheezing and healthy children from the tropics of Ecuador
Publisher of your work	n/a
Expected publication date	Oct 2014
Permissions cost	0.00 GBP
Value added tax	0.00 GBP
Total	0.00 GBP
Total	0.00 GBP
Terms and Conditions	

STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL FROM AN OXFORD UNIVERSITY PRESS JOURNAL

1. Use of the material is restricted to the type of use specified in your order details.

2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.

3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.

4. No alteration, omission or addition is made to the material without our written consent. Permission must be re-cleared with Oxford University Press if/when you decide to reprint.

5. The following credit line appears wherever the material is used: author, title, journal, year, volume, issue number, pagination, by permission of Oxford University Press or the sponsoring society if the journal is a society journal. Where a journal is being published on behalf of a learned society, the details of that society must be included in the credit line.

6. For the reproduction of a full article from an Oxford University Press journal for whatever purpose, the corresponding author of the material concerned should be informed of the proposed use. Contact details for the corresponding authors of all Oxford University Press journal contact can be found alongside either the abstract or full text of the article concerned, accessible from www.oxfordjournals.org Should there be a problem clearing these rights, please contact journals.permissions@oup.com

7. If the credit line or acknowledgement in our publication indicates that any of the figures, images or photos was reproduced, drawn or modified from an earlier source it will be necessary for you to clear this permission with the original publisher as well. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.

8. While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Oxford University Press or by Copyright Clearance Center (CCC)) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Oxford University Press reserves the right to take any and all action to protect its copyright in the materials.

9. This license is personal to you and may not be sublicensed, assigned or transferred by you to any other person without Oxford University Press's written permission.

10. Oxford University Press reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

11. You hereby indemnify and agree to hold harmless Oxford University Press and CCC, and their respective officers, directors, employs and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant

to this license.

12. Other Terms and Conditions:

v1.4

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.