Camarinha-Silva, A., Jáuregui, R., Chaves-Moreno, D., Oxley, A.P.A., Schaumburg, F., Becker, K., Wos-Oxley, M.L., Pieper, D.H.

Comparing the anterior nare bacterial community of two discrete human populations using Illumina amplicon sequencing

(2014) Environmental Microbiology, 16 (9), pp. 2939-2952.

Comparing the anterior nare bacterial community of two discrete human populations using Illumina amplicon sequencing

Amélia Camarinha-Silva[1], Ruy Jáuregui[1], Diego Chaves-Moreno[1], Andrew P.A. Oxley[1,2], Frieder Schaumburg[3], Karsten Becker[3], Melissa L. Wos-Oxley[1], Dietmar H. Pieper[1]*

[1]Microbial Interactions and Processes Research Group, Helmholtz Centre for Infection Research, Braunschweig, Germany;

[2]Infection Immunology Research Group, Helmholtz Centre for Infection Research, Braunschweig, Germany;

[3]Institute of Medical Microbiology, University Hospital Münster, Münster, Germany

*For correspondence: E-mail dpi@helmholtz-hzi.de; Tel. (+49) 531 6181 4200; Fax (+49) 531 6181 4499

Running title: Bacterial community structure of anterior nares

Summary

The anterior nares are an important reservoir for opportunistic pathogens and commensal microorganisms. A barcoded Illumina paired-end sequencing method targeting the 16S rRNA V1-2 hypervariable region was developed to compare the bacterial diversity of the anterior nares across distinct human populations (volunteers from Germany vs a Babongo Pygmy tribe, Africa). Of the 251 phylotypes detected, 231 could be classified to the genus level and 109 to the species level, including the unambiguous identification of the ubiquitous *Staphylococcus aureus* and *Moraxella catarrhalis*. The global bacterial community of both adult populations revealed that they shared 85% of the phylotypes, suggesting that our global bacterial communities have likely been with us for thousands of years. Of the 34 phylotypes unique to the non-westernized population, most were related to members within the suborder *Micrococcineae.* There was an even more overwelming distinction between children and adults of the same population, suggesting a progression of a childhood community of high diversity comprising species of *Moraxellaceae* and *Streptococcaceae* to an adult community of lower diversity comprising species of *Propionibacteriaceae*, *Clostridiales* Incertae Sedis XI, *Corynebacteriaceae* and *Staphylococcaceae.* Thus, age was a stronger factor for accounting for differing bacterial assemblages than the origin of the human population sampled.

Introduction

The human anterior nares are known to harbor bacterial species representing predominantly the phyla *Actinobacteria*, *Bacteroidetes*, *Firmicutes* and *Proteobacteria* (Costello, 2009, Grice*, et al.*, 2009, Lemon*, et al.*, 2010, Wos-Oxley*, et al.*, 2010, Camarinha-Silva*, et al.*, 2012, Oh*, et al.*, 2012). They are the principal

habitat for one of the major human pathogens, *Staphylococcus aureus* (van Belkum*, et al.*, 2009), and for other commensal microorganisms and opportunistic pathogens such as *Staphylococcus epidermidis* and other coagulase-negative staphylococci, *Corynebacterium* spp., *Propionibacterium* spp., *Dolosigranulum pigrum*, *Finegoldia magna*, *Peptoniphilus* sp., *Moraxella* spp., *Anaerococcus* spp., uncultured *Actinomycetes*, and uncultured *Neisseriale*s, all of which constitute the normal and asymptomatic core bacterial community (Wos-Oxley*, et al.*, 2010). However, our current paradigms on the anterior nare bacterial community are based on the characterization of individuals of westernized populations of Europe and the U.S.A. (Rasmussen*, et al.*, 2000, Costello, 2009, Grice*, et al.*, 2009, Frank*, et al.*, 2010, Lemon*, et al.*, 2010, Wos-Oxley*, et al.*, 2010, Laufer*, et al.*, 2011, Camarinha-Silva*, et al.*, 2012, Oh*, et al.*, 2012). There is currently no work describing the anterior nare bacterial communities of humans from non-westernized populations, even though it has been suggested that bacterial clones have adapted to specific human populations reflecting coevolution between bacteria and their hosts, as described for *Mycobacterium tuberculosis* (Gutierrez*, et al.*, 2005, Ruimy*, et al.*, 2010), *Helicobacter pylori* (Linz*, et al.*, 2007) and *S. aureus* (Ruimy*, et al.*, 2008, Fan*, et al.*, 2009, Ruimy*, et al.*, 2010).

Next Generation Sequencing (NGS) platforms offer high-throughput culture-independent analyses where previous studies have characterized human anterior nare communities using a range of different culture-independent approaches, including 454-pyrosequencing (Costello, 2009, Grice*, et al.*, 2009, Frank*, et al.*, 2010, Lemon*, et al.*, 2010, Wos-Oxley*, et al.*, 2010, Camarinha-Silva*, et al.*, 2012, Oh*, et al.*, 2012). A handful of publications have now assessed the applicability of using Illumina for amplicon deep-sequencing and reported its advantages and shortcomings over other technologies, (Caporaso*, et al.*, 2012, Degnan & Ochman,

2012, Soergel*, et al.*, 2012, Werner*, et al.*, 2012, Bokulich*, et al.*, 2013). However, despite the recent spurt in work reporting the potential use of Illumina-based amplicon deep-seqencing, its use to assess and compare hundreds of samples is still in its infancy.

In the current work, a cross-sectional study assessing the anterior nare bacterial diversity among and between the indigenous Pygmy population of Gabon, known to be genetically divergent from nonpygmy populations for between 54,000-90,000 years (Verdu*, et al.*, 2009), and a population within Germany was performed using a barcoded Illumina paired-end sequencing method targeting the V1-2 hypervariable region of the 16S rRNA gene. This region provides sufficient resolution to specifically identify *S. aureus* among other *Staphylococcus* species, and *M. catarrhalis* among other *Moraxella* spp.

Results

Gauging sequencing error rates and classifying phylotypes

The classification accuracy of short sequence reads not only depends on the resolution level of the region analyzed but crucially on the introduced error rate. Samples containing sequences belonging to *S. aureus* (DSM 3463), *C. accolens* (GU074966), *S. epidermidis (*JF927883*)* and *M. nonliquefaciens* (JF927886), served as controls for assessing sequencing errors along the full length paired-end read. Each nucleotide of each of the 146 nt sequence reads, both forward and reverse, for each control sample was compared to its reference sequence to explore for the variation along the sequence and at what point the sequence quality changes. Plotting the error rate of each nt over the full paired-end sequence read (245 nt in length after removal of primers and barcodes) showed an increase in the error rate

after 80 nt on both of the paired-end reads in all 4 species (Fig. 1). Thus, a total of 160 nt were used for classification of sequence reads.

After sampling the global bacterial communities of 190 human anterior nares from westernized and non-westernized individuals and generating 1.5 million usable sequence reads, 251 phylotypes could be discretely separated and classified into Order (98% of phylotypes), Family (96% of phylotypes), Genus (92% of phylotypes) and Species (43% of phylotypes) levels (Table S1, S2). In total, 241 phylotypes belong to the phyla *Actinobacteria, Firmicutes, Proteobacteria* and *Bacteroidetes*, previously described as the most abundant anterior nare bacterial phyla (Costello, 2009, Grice*, et al.*, 2009, Frank*, et al.*, 2010, Lemon*, et al.*, 2010, Wos-Oxley*, et al.*, 2010, Camarinha-Silva*, et al.*, 2012, Oh*, et al.*, 2012). Of the remaining 10 phylotypes, 6 were classified as belonging to the phylum *Fusobacteria* while 2 indicate the presence of chloroplasts. Chloroplasts have been shown to be ubiquitous in house dust (Pakarinen*, et al.*, 2008, Taubel*, et al.*, 2009) and it is thus reasonable to assume that the chloroplast sequences here are indicative of those organisms that only pass through the anterior nares. Only 2 phylotypes could not be classified into any phyla.

To assess the influence of using phylotype-level (species-level), genus-, family-, order-, class-, phylum-level abundance matrices to discern for patterns across samples, that is whether the pattern across all 190 samples is maintained when using either of these data matrices, a Mantel-like test (RELATE) was performed. Correlating the pattern generated using the original phylotype-level data matrix with the higher-level data matrices resulted in high correlation co-efficients (Fig. S1) for the genus-level and family-level matrices, determining that using species-, genus- and family-level abundance provides the same level of distinction across samples.

The global community profiles generated from 63 samples using this approach was correlated with the global community profiles previously obtained using the T-RFLP fingerprinting technique (Camarinha-Silva*, et al.*, 2012). The very strong correlation between both methods (Rho 0.876, p=0.001) implies that they produced highly similar profiles for each volunteer and conserved the differences observed between them. However, in contrast to T-RFLP, direct sequence information results in a higher resolution at resolving species.

Gauging sampling effort

To estimate whether the sampling depth per sample was sufficient, rarefaction curves (Fig. S2) were plotted. They all show a plateau indicating that sequencing depth per sample was ample for resolving total phylotype numbers and thus for capturing the bacterial diversity for each anterior nare sample. The mean number of sequences per sample was 7,682. Each anterior nare sample contained between 14 and 164 phylotypes, with on average 88±29 phylotypes per person for the non-westernized adults, 59±21 phylotypes per person for the westernized adults and 65±36 phylotypes per person for the non-westernized children.

To estimate whether the number of samples per population was sufficient, species accumulation curves using S_{obs} (species observed) and the Chao 2 species richness estimator were used. Estimates of cumulative species richness against sampling effort (the number of samples collected per population) reached asymptotic values (Fig. S3). The total number of phylotypes observed across all westernized and non-westernized samples was 217 and 246, respectively, while the Chao 2 estimate of species richness gave 219±2 and 254±12, respectively. Likewise, estimates of cumulative species richness against sampling effort of the non-westernized adults and children independently, showed total numbers of observed phylotypes of 243

and 233, respectively, while the Chao 2 estimate of species richness gave 247±4 and 242±6, respectively. Since both the S_{obs} and Chao 2 values are in good agreement and the plotted species accumulation curves of S_{obs} and Chao 2 meet and level off (after 78 collected samples), it can be assumed that sampling effort was great enough to accurately characterize most of the phylotypes likely to be present in the nares of both populations.

The bacterial community across the anterior nares

Of the 251 phylotypes, 246 were detected in the non-westernized population, while 217 phylotypes were detected in the westernized population (Fig. 2, Fig. S4). Eighty-five percent of all phylotypes were shared between both human populations. However, of the 34 phylotypes unique to the non-westernized population, half were related to members within the phylum *Actinobacteria*, and most within the suborder *Micrococcineae* (Fig. 2, Fig. S4 and Table S2) such as *Brachybacterium* sp. (PT90-2), *Brevibacterium linens/iodinum* (PT184-1), *Janibacter* sp. (PT199-1) and *Janibacter anopheles* (PT166-1). Also 3 phylotypes indicating the presence of *Kocuria* spp. (PT40-1, PT27-1 and PT152-1) were evidently more prevalent in the non-westernized compared to the westernized population (Fig. S5). A further 5 phylotypes, belonging to the families *Staphylococcaceae* (PT38-2 and PT509-1), Incertae Sedis XI of the *Clostridiales* (PT63-1), *Prevotellaceae* (PT160-1) and an unclassified bacterium (PT171-1) were specific to the westernized population (Fig. 2, Fig. S4).

Exploring the global bacterial community structure of the 190 anterior nares using ordination revealed a distinction between westernized adults and the non-westernized population, but an even greater distinction between children (0-14 years-of-age) and adults (18-85 years-of-age) of the non-westernized population (Fig. 3).

Calculating the degree of distinction using the relative abundance of phylotypes purports that the communities of westernized and non-westernized adults were significantly different (R=0.117, p=0.001) (Fig. S6), but the ANOSIM R-value suggests that these two communities still share some community members. The difference between non-westernized adults and children was greater (R=0.535, p=0.001) (Fig. 3, Fig. S6), indicating that these 2 groups comprise very different community members. There is also much less intra-variation within the non-westernized adult group (having a low dispersion index of 0.686) compared to the westernized adults (1.04) or non-westernized children (1.333), as they clustered more tightly together, suggesting that they have a more consistent community membership from person to person. Correlations of the global bacterial profiles of the non-westernized population with other demographic data (except for age) such as weight, gender and dwelling status (residence of a particular village and hut), and also travel habits and daily activities (hunting/gathering) were explored for and not found.

Comparing the bacterial community of westernized and non-westernized adults

Most of the phylotypes of the adult populations belong to a few families (Fig. 4). *Corynebacteriaceae* members were observed in relative higher abundance in the non-westernized adults (representing 57% of the total reads), contrasting with the westernized adults (38%). In contrary, members of the *Propionibacteriaceae* and *Staphylococcaceae* were in higher abundance in the westernized adults comprising 21% and 13% in westernized versus 8% and 6% in non-westernized adults, respectively. Phylotypes belonging to the families *Carnobacteriaceae, Moraxellaceae*, *Streptococcaceae* and Incerteae Sedis XI were detected in similar abundances in both adult populations. Members of the *Neisseriaceae* and

Pasteurellaceae were represented by less than 2% of the total reads in both adult populations (Fig. 4).

The majority of the volunteers of both adult populations (>85%) were colonized by phylotypes previously identified as core nare community members in German adults (Wos-Oxley*, et al.*, 2010), such as *Corynebacterium accolens* (PT1-1)*, Propionibacterium acnes* (PT3-1), *Propionibacterium granulosum* (PT61-1) and *S. epidermidis/Staphylococcus capitis/Staphylococcus caprae* (PT38-1) (Fig. S7). Phylotypes indicating the presence of *Finegoldia magna* (PT16-1), *Peptoniphilus sp.* (PT44-1), *Streptococcus cristatus/Streptococcus infantis* (PT22-1), *Anaerococcus sp.* (PT6-1 and PT9-1) and uncultured *Actinomycetes* (PT60-1) were identified in 40-85% of the volunteers in both populations. *S. aureus* was detected in 12 of the westernized volunteers in an abundance higher than 10%. This is in contrast to the non-westernized population where its relative abundance never exceeded 2.2% of the total community (Fig. S8).

Moraxella lacunata/Moraxella nonliquefaciens (PT76-1 and PT 233-1) and *Moraxella catarrhalis* (PT123-1) were detected in more than 75% of the non-westernized adult population, colonizing both genders in similar relative abundances (Fig. S8). This is in contrast to the westernized adults, where *M. lacunata/M. nonliquefaciens* was more prevalent in women (n=32) than in men (n=16) and specifically more abundant in women (Fig. S8), a finding that has been previously described by Camarinha-Silva *et al.* (2012). Colonization by *M. catarrhalis* was detected only in a minority of the westernized population (female n=14, male n=9) at abundances similar in men and women (Fig. S8).

An overall of five different phylotypes were assigned to *Dolosigranulum sp.* (PT72-1, PT72-2, PT182-1, PT226-1 and PT66-1), where PT66-1 was more prevalent and abundant in the westernized adult population and PT72-1 was

observed in all non-westernized adults in relatively high abundances up to 47% (Fig. S9). PT182-1 and PT226-1 were observed in relatively low abundances in both adult populations, though were more prevalent in non-westernized adults. This implies that although both adult populations were colonized by bacteria related to *Dolosigranulum* species, there is a population specific trend.

Comparing the microbial community of non-westernized adults and children

Phylotypes belonging to the *Moraxellaceae* and *Streptococcaceae* families were more abundant in children (27% and 17%, respectively), contrasting with the lower values observed in both westernized (6% and 4%, respectively) and non-westernized (7% and 3%, respectively) adults. Also, *Neisseriaceae,* which were detected in the non-westernized adults in abundances lower than 1%, were observed in greater abundances in the non-westernized children (2-3%). However, *Corynebacteriaceae* were only represented by 21% of the total reads in the children, contrasting with the high value observed in the non-westernized adults (57%). Specifically, PT59-1, indicating the presence of bacteria related to *Streptococcus pneumoniae/ Streptococcus mitis,* was more than 7-fold more abundant in children (average relative abundance of 14.3%) compared to adults (2%) (Fig. S9). Also, *Moraxella lincolnii* (PT98-1) and *M. catarrhalis* (PT123-1) were detected in higher abundances in children (Fig. S10). In contrary, phylotypes representing the *Propionibacteriaceae* and *Clostridiales* Incerteae Sedis XI families, such as *P. acnes* (PT3-1)*, F. magna* (PT5-1 and PT 16-1) and *Peptoniphilus sp*. (PT44-1) were poorly represented in children (Fig. S7).

Of the previously described core community members of the anterior nares, which were also shown to be core community members in both the westernized and non-westernized adults in this work, only *Dolosigranulum sp., C. accolens* and *C.*

propinquum/C. tuberculostearicum were observed in the majority of children (>90%). In contrast, *M. catarrhalis* (PT123-1) and *S. pneumoniae/ S. mitis* (PT59-1) (Fig. S9, S9), were prevalent in almost all children (85% and 98%, respectively), thus collectively comprising the core community of children.

The diversity of the anterior nare communities

Phylotype diversity, richness and evenness were explored using both conventional and contemporary measures. The conventional indices, total phylotypes (S), Shannon diversity (H'), Pielou's evenness (J') and Simpsons index (1-lambda) (Fig. S11) are based on species richness/abundance data from each sample, while the contemporary indices, average taxonomic distinctness (delta+) and variation in taxonomic distinctness (lambda+) address the taxonomic relatedness of species within each sample, where taxa that are distantly related to each other contribute more to taxonomic diversity than those that are closely related.

When delta+ was plotted against the number of phylotypes (Fig. 5), most of the samples derived from adults had values that were below the 'expected range', indicating that adults of both populations have lower taxonomic breadth and diversity per sample than the children. In contrast, most of the children returned delta+ values that fell within the expected range, and since having larger values for delta+ indicate that their nare bacterial communities were more diverse. Lambda+, a measure of consistency between levels of taxonomic classification and thus taxonomic evenness within a sample was also plotted against the number of phylotypes (Fig. 5), where most of the adults (from both populations) had a greater disparate range of values for delta+ which led to higher lambda+ values outside of the expected range, and so greater taxonomic unevenness. Again, most children returned lambda+ values that

fell within the expected range, and since having lower values for lambda+ indicate a more consistent taxonomic distance between species and thus taxonomic evenness.

Besides being less taxonomic diverse and even, the westernized adults comprised the lowest species richness (with an average of 59±2 phylotypes per sample), the lowest species diversity (using both Shannon and Simpson indices) and the lowest species evenness (Fig. S11). In contrary, while the non-westernized adults comprised the highest species richness with an average of 88±4 phylotypes per sample, they had the lowest taxonomic diversity. The non-westernized children having an average 65±6 phylotypes per sample comprised the highest species diversity, species evenness, taxonomic diversity and taxomonic evenness (Fig. S11). Noteworthy, while the conventional diversity indices rely on species richness and abundance within each sample and not on the taxonomic relatedness of the species themselves, they are also greatly dependent on sample size/effort and lack a statistical framework for testing the departure from expectation (Warwick and Clarke, 1995).

Discussion

Next generation sequencing like the Illumina's ultra-high-throughput technology now allows deep sampling of bacterial niches at a fraction of the cost of other technologies (Degnan & Ochman, 2012). We deem that the paired-end reads of 80 nt (160 nt together) used here were enough to distinguish closely related phylotypes and irrespective of their classification at species, genus, family or higher-order, the patterns across the samples were maintained. Targeting the 16S rRNA V1-2 hypervariable region provides sufficient information for the unambiguous identification of the ubiquitous species *S. aureus* and *M. catarrhalis,* while also allowing 231 of the 251 observed phylotypes to be classified to the genus level and

109 phylotypes to be classified to the species level. With this newly optimized deep-sequencing approach, we could phylogenetically characterize the anterior nare bacterial communities of hundreds of samples, particularly from populations that have yet to be investigated, and could provide an even more in-depth analysis of anterior nare community structure, taxonomic diversity and distinctness than has previously been performed.

A majority of the anterior nare bacterial community (85%) was shared by both westernized and non-westernized adult populations, with those same members known to constitute the core nare community of previously described westernized populations (Wos-Oxley*, et al.*, 2010). Given that these populations are both geographically and evolutionally separated, this shared core community with only some species-specific patterns suggests that the global bacterial community of the human anterior nares is likely to have persisted with its host throughout recent evolution, and that there is a definite 'fingerprint' for all human anterior nares.

However, the prevalence and abundance of particular phylotypes differed between both adult populations. For example, while *Propionibacteriaceae* and *Staphylococcaceae* families were prevalent across all adults, they were more abundant in the westernized adults, whereas *Corynebacteriaceae* were more abundant in the non-westernized adults. Also, there was evidence of species-population specific trends, where for example phylotypes related to *D. pigrum* showed population specific lineages. Considering that the pygmies split from other humans thousands of years ago and that they are still a semi-nomadic population (Verdu*, et al.*, 2009), it is not unexpected that the non-westernized adults have some species-specific patterns. In fact, other studies have shown that bacterial clones have adapted to specific human populations reflecting coevolution between bacteria and their hosts, such as *Mycobacterium tuberculosis* (Gutierrez*, et al.*, 2005, Ruimy*, et*

al., 2010), *Helicobacter pylori* (Linz*, et al.*, 2007) and *S. aureus* (Ruimy*, et al.*, 2008, Fan*, et al.*, 2009, Ruimy*, et al.*, 2010).

Both host and environmental factors have been reported to shape host bacterial communities (Benson*, et al.*, 2010, Li & Hotamisligil, 2010). Given that a low degree of variability among the microbiomes of non-westernized adults compared to those of the westernized adults was observed here, it could be hypothesized that this is correlated to the assumed low human genetic variability among the non-westernized population. However, this low intra-variation among microbiomes may as well be due to the non-westernized adults residing in close proximity and being exposed to more consistent environmental conditions. In fact, the anterior nare microbiome might likely be influenced by direct environmental factors due to inhalation, as interestingly, 3 phylotypes indicating a high prevalence of *Kocuria* spp. in the anterior nares of non-westernized people, were also found to be the second most abundant group of species from African desert dust (Favet*, et al.*, 2013). This supports the idea that bacterial community composition can be driven by the exposure to different environmental conditions such as different sources of dust.

Even though a significant difference between the non-westernized and westernized adults was observed, the most striking difference was between the non-westernized adults and children. Two recent studies have characterized the global bacterial structure of the anterior nares of children (Laufer*, et al.*, 2011, Oh*, et al.*, 2012). Both papers report that the nasal communities of American children were dominated by *Moraxellaceae* and *Streptococcaceae*, as was observed here for the Pygmy children of Gabon. However, members of *Corynebacteriaceae* were 3 times more abundant in Pygmy children analyzed here compared to the American children analyzed by Laufer *et al*. (2011) and Oh *et al*. (2012). Contrasting results were also found between Laufer *et al*. (2011) and Oh *et al.* (2012), where Laufer *et al*. (2011)

reported of higher abundances of *Propionibacteriaceae* in children, also in contrary to the non-westernized Pygmy children in this study. It was recently shown that propionibacterial densities increase even more sharply in the nares of children who develop acne, suggesting that the nasal epithelium responds to the rise in androgen levels or to sebum production (Mourelatos*, et al.*, 2007). If the reported lack of incidence of acne in two non-westernized populations of the Kitavan Islanders of Papua New Guinea and the Aché hunter-gatherers of Paraguay (Cordain*, et al.*, 2002) is in fact related to low *Propionibacteriaceae* abundances in non-westernized populations remains to be elucidated.

Among the most prominent differences reported to date between westernized adults and children was the higher abundance of taxa related to *M. catarrhalis* and *S. pneumoniae/mitis*. *S. pneumonia* and *M. catarrhalis* have been reported to asymptomatically colonize the nasopharynx of 71% and 88% of children, respectively (Bogaert*, et al.*, 2011), suggesting that both respiratory pathogens are constituents of the core upper respiratory tract community of children.

In this work, the non-westernized adults comprised the highest species richness and contained medium to high levels of species diversity compared to westernized adults and non-westernized children. However, taking into account the taxonomic relatedness of species within each sample, they comprised the lowest taxonomic diversity. Interestingly, since their introduction by Warwick and Clarke (1995), these contemporary indices have gained much attention in the field of macro-ecology but little attention in microbial ecology. To our knowledge, this is the first report on the use of these contemporary measures of diversity using data generated via next-generation sequencing technologies. Since phylotype classification (using Linnaean classification) can be made from the short sequences generated from NGS technologies, it only seems appropriate to go one step further and couple phylotype

classification with taxonomic distinctness and diversity assessment, giving more detailed and accurate knowledge on microbial community diversity.

Experimental Procedures

Study population

Anterior nare swabs were collected from 190 human volunteers using a standard swabbing procedure with dry sterile cotton swabs (Schaumburg*, et al.*, 2011). From the Lower Saxony and North Rhine-Westphalia regions of Germany (herein referred to as the westernized population), 92 healthy volunteers provided swabs (Table S3). These swabs were stored at -20°C until further analysis. From a semi-nomadic Babongo Pygmy tribe from the Ikobé region in the Gabonese Republic (herein referred to as the non-westernized population), 98 healthy volunteers provided swabs (Table S4), which equates to one third of this population living in close proximity (within 15 km). These swabs were stored in cool boxes and transferred to -20°C within four days after sampling until further analysis. Informed consent was obtained from all volunteers. With regards to the volunteers from Gabon, ethical clearance and consent was obtained as previously described (Schaumburg*, et al.*, 2011). In brief, ethical clearance was obtained from the institutional review board (IRB, “Comité d'Éthique Régional Indépendant de Lambaréné”, Lambaréné, Gabon, protocol number: CERIL 15–09). As the majority of the Babongo Pygmies are illiterate and mainly speak the tribal language, a local interpreter provided detailed information about the study and obtained documented oral informed consent. A short written summary was prepared in French that described the information presented to the Pygmies. This document was signed or finger-printed by the participant, the researcher and a witness who spoke French and Babongo. The IRB approved the

use of documented oral informed consent. Exclusion criteria were (i) infections of nostrils and (ii) a purulent rhinitis. Demographic data (self reported age, height, weight, gender, and dwelling status) were recorded for each subject, as well as travel habits since birth and daily activities such as hunting/gathering. Global positioning data of each village were taken using a GPS-device (Garmin76 csx).

DNA extraction

DNA was extracted from the swabs using the FastDNA Spin Kit for Soil (MP Biomedicals, Solon, OH, USA) following the manufacturer's instructions. In brief, the swabs were placed into Lysing Matrix E tubes containing MT buffer and sodium phosphate buffer and cells were lyzed in a Fast Prep®-24 instrument for 30 sec at an intensity setting of 6.0. DNA was eluted in 50 µL of DES and quantified using a NanoDrop 2000 spectrophotometer (Thermo Scientific, Waltham, USA).

Amplicon library preparation

The V1-2 region of the 16S rRNA gene was amplified using primers based on the previously described 27F and 338R primers (Lane, 1991, Etchebehere & Tiedje, 2005) (Table S5). The forward primer contains a 6 nucleotide (nt) barcode (Meyer & Kircher, 2010) and a 2 nt CA linker (Hamady*, et al.*, 2008). Both primers comprised sequences complementary to the Illumina specific adaptors to the 5'-ends. Amplification was performed in a total volume of 50 µL with 5x PrimeSTAR TM buffer, containing each deoxynucleoside triphosphate at a concentration of 2.5 mM, each primer at a concentration of 0.2 µM, 1 µL of template DNA and 0.5 µL PrimeSTAR TM HS DNA polymerase (2.5U). An initial denaturation step of 95°C for 3 min was followed by 15 cycles of denaturation at 98°C for 10 sec, annealing at

55°C for 10 sec and extension at 72°C for 45 sec. One µL of this reaction mixture served as template in a second PCR performed under the same conditions as described above, but for 20 cycles using PCR primers designed to integrate the sequence of the specific Illumina multiplexing sequencing primers and index primers (Table S5). Non-template controls (using water as template) were performed and were free of any amplification products after both rounds of PCR. PCR amplicons were verified by agarose gel electrophoresis, purified using Macherey-Nagel 96-well plate purification kits (Macherey-Nagel, Düren, Germany) following the manufacturer's instructions and quantified with the Quant-iT PicoGreen dsDNA reagent and kit (Invitrogen). Libraries were prepared by pooling equimolar ratios of amplicons (200 ng of each sample) derived from approximately 40 samples, all having been tagged with a unique barcode. In total, 5 libraries were prepared comprising all 190 samples. To remove any contaminants or PCR artefacts, each library (626–1169 µl) was precipitated on ice for 30 min after addition of 20 µl of NaCl (3M) and 3 volumes of ice-cold 100% ethanol. The precipitated DNA was centrifuged at 13,000 x g for 30 min at 4°C. The supernatant was removed, the pellet air dried, resuspended in 30 µL of double-distilled water and separated on a 2% agarose gel. PCR products of the correct size were extracted and recovered using the QIAquick gel extraction kit (Qiagen). In order to assess the quality of sequence data and for possible sequencing errors, positive control samples containing known 16S rRNA gene sequences were used, where control #1 comprised equimolar concentrations of 16S rRNA gene amplicons derived from *Staphylococcus aureus* (DSM3463) genomic DNA and from a pGEM-T easy vector containing the 16S rRNA gene from *Corynebacterium accolens* (GU074966) and control #2 comprised equimolar concentrations of 16S rRNA gene amplicons derived from pGEM-T easy vectors containing the 16S rRNA genes from *Moraxella nonliquefaciens* (JF927886) and

Staphylococcus epidermidis (JF927883). These additional samples were treated together with the other anterior nare samples and pooled in the final library. Libraries were sent for 150 nt paired-end sequencing on a GAIIx Genome Analyzer (Illumina, Inc., California, USA). Image analysis and base calling were accomplished using the Illumina Pipeline (version 1.7).

Bioinformatic analysis

A total of 3 million sequence reads were obtained. A quality filter that runs a sliding window of 10% of the read length at a time, and calculates the local average score based on the Phred quality score of the fastq file, was used to trim the 3'-ends of the reads that fall below a quality score of 10. All reads that had an N character in their sequence, any mismatches within the primers and barcodes or more than 10 homopolymer stretches were discarded. Sequences were sorted into their respective samples based on their barcodes, where primers and barcodes were then trimmed from each read. All reads were further trimmed to account for additional sequencing errors found to increase with read length. For this, the error rate of sequence reads obtained for the known 16S rRNA gene fragments within the internal controls were assessed by applying the Needleman-Wunsch global alignment algorithm (Needleman & Wunsch, 1970) embedded in EMBOSS (Rice*, et al.*, 2000). This algorithm was used to determine the optimum alignment of each read present in the control samples with a sequence of reference along their entire length at 98% identity. As the sequence quality decreased at the 3'-end of both paired-end reads, all reads were trimmed conservatively to 80 nt. The paired-ends were subsequently matched to give 160 nt for downstream analysis, where such conservative shortening of paired-end and single direction Illumina amplicon reads has been performed and validated by others (Soergel*, et al.*, 2012, Werner*, et al.*, 2012).

To determine how many phylotypes were present across all samples, forward reads were first collapsed and clustered together allowing for 1 mismatch (Uclust algorithm on USEARCH (Edgar, 2010)). Then, reverse reads were paired with their respective forward read and clustered again allowing for 1 mismatch. Due to diversity in some reverse reads, phylotypes could be further resolved once the reverse sequences were considered. The data-set was then filtered to consider only those phylotypes that: a) were present in at least one sample at a relative abundance >1% of the total sequences of that sample or b) were present in at least 2% of samples at a relative abundance >0.1% for a given sample, or c) were present in at least 5% of samples at any abundance level. A total of 251 phylotypes could be resolved for further analysis.

All samples comprised >816 sequence reads, where the mean number of sequences per sample was 7,682 ± 484, totalling 1,459,615 usable paired-end sequence reads. All phylotypes were assigned a taxonomic affiliation based on naïve Bayesian classification (RDP classifier) (Wang*, et al.*, 2007). Phylotypes were then manually analyzed against the RDP database using the Seqmatch function as well as against the NCBI database to define the discriminatory power of each sequence read. A species name was assigned to a phylotype when only 16S rRNA gene fragments of previously described isolates of that species showed ≤2 mismatches with the respective representative sequence read. Similarly, a genus name was assigned to a phylotype when only 16S rRNA gene fragments of previously described isolates belonging to that genus and of 16S rRNA gene fragments originating from uncultured representatives of that genus showed ≤2 mismatches (Table S2). The short nucleotide sequences for each of the determined phylotypes obtained in this work have been made available in Table S1.

Sequence analysis

16S rRNA gene sequences from the closest taxonomic relatives assigned to each of the 251 phylotypes using RDP/NCBI were obtained as a pre-aligned set of manually curated sequences from the SILVA database (Pruesse*, et al.*, 2007) and a maximum likelihood tree constructed using MEGA5 (Tamura*, et al.*, 2011). Evolutionary distances were determined across all sites using the Jukes-Cantor correction model (Jukes & Cantor, 1969) with branch support values calculated from 1000 bootstrap re-samplings (values $\geq$ 85% given at the nodes).

Statistical analysis

A multivariate data-set comprising the relative percent abundance of each phylotype across each of the 190 samples was analyzed using a suite of univariate and non-parametric multivariate routines in PRIMER (v.6.1.6, PRIMER-E, Plymouth Marine Laboratory, UK, (Clarke & Warwick, 2001). To estimate whether the sampling depth per sample was sufficient, rarefaction analyses were performed (Sanders, 1968, Hurlbert, 1971). To estimate whether the number of samples per group (human population) was sufficient, species accumulation curves using Sobs (species observed) and the Chao 2 species richness estimator (= Sobs + (L2/2M), where L is the number of species that occurred in only one sample and M is the number of species that occurred in exactly two samples were constructed using 999 permutations (Chao, 1984, Chao, 1987, Clarke & Warwick, 2001).

To make comparisons between the samples themselves, a sample-similarity matrix was generated using the Bray-Curtis coefficient (Bray & Curtis, 1957) and the bacterial community structures were explored by ordination using non-metric multidimensional scaling (nMDS) (50 random restarts). Significant differences between *a priori* predefined groups of samples (non-westernized adults, non-

westernized children and westernized adults) were evaluated using Analysis of Similarity (ANOSIM) (999 permutations). Groups of samples (anterior nare communities) were considered significantly different if the *p*-value falls <0.05. The accompanying *R* statistic measures the degree of separation between groups and ranges from −1 to 1, in which the higher its value (closer to 1) the more distinct the groups (Clarke & Warwick, 2001). Multivariate dispersion analysis was used to calculate the degree of intra-variation among the microbiomes originating from the same population, where a low dispersion index indicates low within-group heterogeneity.

To compare whether using the phylotype list and their relative abundance yielded similar patterns if only genus-level, family- etc data could be reached, the current species-level data matrix was aggregated to genera-, family-, order- class- and phylum-level by its taxonomic assignment. With each new data matrix, a new resemblance matrix was produced. Then, each matrix was compared to the original species-similarity matrix by a Mantel-like test (RELATE in PRIMER) using the Spearman Rank correlation method and 999 permutations. This same test was also used to quantify the pattern match between the bacterial community profiles of 63 samples generated using this Illumina technique compared to the more traditional T-RFLP fingerprinting technique (Camarinha-Silva*, et al.*, 2012).

Phylotype diversity, richness and evenness were explored using both conventional and contemporary measures. For the conventional measures, total phylotypes (S), Shannon diversity (H'), Pielou's evenness (J') and Simpsons index (1-lambda) were calculated (Clarke & Warwick, 2001). For contemporary measures, average taxonomic distinctness, (delta+) and variation in taxonomic distinctness, (lambda+) within each sample were calculated (Warwick & Clarke, 1995). Specifically, average taxonomic distinctness (delta+) is a measure of the average

taxonomic distance between all pairs of species in a sample and thus is a measure of taxonomic breadth within a sample, while the variation in taxonomic distinctness (lambda+) reports how consistent each level of organization within the Linnaean classification is represented. In brief, a species masterlist comprising all 251 phylotypes was classified according to Linnaean classification. The expected delta+ value was calculated by sampling different numbers of phylotypes (from 10 to 100 in increments of 10) from the masterlist, repeating 999 times (Pienkowski*, et al.*, 1998). Values for delta+ and lambda+ are plotted against the number of phylotypes within each sample and compared to those 'expected values' derived from a species master list ascertained across all samples, where the funnel indicates the limits within 95% of the simulated taxonomic distinctness (TD) values and the middle line represents the mean expected TD, thus providing a statistical framework to test whether these measures depart from expectation (Warwick & Clarke, 1995, Pienkowski*, et al.*, 1998).

Lastly, any report of data dispersion in this work is given as the Standard Error of the Mean (S.E.M), unless otherwise stated.

Acknowledgments

The authors would like to thank Iris Plumeier and Silke Kahl for technical support. This work was funded by the BMBF project "Medical Infection Genomics" to DHP (0315832B) and KB (0315832A) and in part by the BMBF project "SkIn Staph" to KB and DHP (01KI1009A).

References

Benson AK, Kelly SA, Legge R*, et al.* (2010) Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc Natl Acad Sci U S A* **107**: 18933-18938.

Bogaert D, Keijser B, Huse S*, et al.* (2011) Variability and diversity of nasopharyngeal microbiota in children: a metagenomic analysis. *PLoS One* **6**: e17035.

Bokulich NA, Subramanian S, Faith JJ*, et al.* (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Methods* **10**: 57-59.

Bray JR & Curtis JT (1957) An ordination of the upland forest communities of southern Wisconsin. *Ecologic Monog* **27**: 325-349.

Camarinha-Silva A, Wos-Oxley ML, Jauregui R, Becker K & Pieper DH (2012) Validating T-RFLP as a sensitive and high-throughput approach to assess bacterial diversity patterns in human anterior nares. *FEMS Microbiol Ecol* **79**: 98-108.

Caporaso JG, Lauber CL, Walters WA*, et al.* (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISMEJ* **6**: 1621-1624.

Chao A (1984) Nonparametric estimation of the number of classes in a population. *Scand J Statist* **11**: 265-270.

Chao A (1987) Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43**: 783-791.

Clarke KR & Warwick RM (2001) Change in marine communities: an approach to statistical analysis and interpretation, 2nd edition. PRIMER-E. *Plymouth*.

Cordain L, Lindeberg S, Hurtado M, Hill K, Eaton SB & Brand-Miller J (2002) Acne Vulgaris: a disease of western civilization. *Arch Dermatol* **138**: 1584-1590.

Costello EK (2009) Bacterial community variation in human body habitats across space and time. *Science* **326**: 1694-1697.

Degnan PH & Ochman H (2012) Illumina-based analysis of microbial community diversity. *ISMEJ* **6**: 183-194.

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*.

Etchebehere C & Tiedje J (2005) Presence of two different active *nirS* nitrite reductase genes in a denitrifying *Thauera* sp. from a high-nitrate-removal-rate reactor. *Appl Environ Microb* **71**: 5642-5645.

Fan J, Shu M, Zhang G*, et al.* (2009) Biogeography and virulence of *Staphylococcus aureus*. *PLoS One* **4**: e6216.

Favet J, Lapanje A, Giongo A*, et al.* (2013) Microbial hitchhikers on intercontinental dust: catching a lift in Chad. *ISMEJ* **7**: 850-867.

Frank DN, Feazel LM, Bessesen MT, Price CS, Janoff EN & Pace NR (2010) The human nasal microbiota and *Staphylococcus aureus* carriage. *PLoS One* **5**: e10598.

Grice EA, Kong HH, Conlan S*, et al.* (2009) Topographical and temporal diversity of the human skin microbiome. *Science* **324**: 1190-1192.

Gutierrez MC, Brisse S, Brosch R*, et al.* (2005) Ancient origin and gene mosaicism of the progenitor of *Mycobacterium tuberculosis*. *PLoS Pathog* **1**: e5.

Hamady M, Walker JJ, Harris JK, Gold NJ & Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nature Meth* **5**: 235-237.

Hurlbert S (1971) The Nonconcept of Species Diversity: A Critique and Alternate Paradigms. *Ecology* **52**: 577-586.

Jukes TH & Cantor CR (1969) Evolution of Protein Molecules. *Evolution of Protein Molecules,*(Munro HN, ed.^eds.), p.^pp. 21-132. Academy Press, New York.

Lane DJ (1991) 16S/23S rRNA sequencing. *Nucleic acid techniques in bacterial systematics,*(Stackebrandt E & Goodfellow M, ed.^eds.), p.^pp. 115-175. Wiley & Sons, Chichester, United Kingdom.

Laufer AS, Metlay JP, Gent JF, Fennie KP, Kong Y & Pettigrew MM (2011) Microbial communities of the upper respiratory tract and otitis media in children. *mBio* **2**: e00245-00210.

Lemon KP, Klepac-Ceraj V, Schiffer HK, Brodie EL, Lynch SV & Kolter R (2010) Comparative analyses of the bacterial microbiota of the human nostril and oropharynx. *mBio* **1**: e00129-00110.

Li P & Hotamisligil GS (2010) Metabolism: Host and microbes in a pickle. *Nature* **464**: 1287-1288.

Linz B, Balloux F, Moodley Y*, et al.* (2007) An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* **445**: 915-918.

Meyer M & Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc* **2010**: pdb.prot5448.

Mourelatos K, Eady EA, Cunliffe WJ, Clark SM & Cove JH (2007) Temporal changes in sebum excretion and propionibacterial colonization in preadolescent children with and without acne. *Brit J Dermatol* **156**: 22-31.

Needleman SB & Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**: 443-453.

Oh J, Conlan S, Polley EC, Segre JA & Kong HH (2012) Shifts in human skin and nares microbiota of healthy children and adults. *Genome Medicine* **4**: 77.

Pakarinen J, Hyvarinen A, Salkinoja-Salonen M*, et al.* (2008) Predominance of Gram-positive bacteria in house dust in the low-allergy risk Russian Karelia. *Environ Microbiol* **10**: 3317-3325.

Pienkowski MW, Watkinson AR, Kerby G, Clarke KR & Warwick RM (1998) A taxonomic distinctness index and its statistical properties. *J Appl Ecol* **35**: 523-531.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J & Glockner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**: 7188-7196.

Rasmussen TT, Kirkeby LP, Poulsen K, Reinholdt J & Kilian M (2000) Resident aerobic microbiota of the adult human nasal cavity. *APMIS* **108**: 663-675.

Rice P, Longden I & Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276-277.

Ruimy R, Maiga A, Armand-Lefevre L*, et al.* (2008) The carriage population of *Staphylococcus aureus* from Mali is composed of a combination of pandemic clones and the divergent Panton-Valentine leukocidin-positive genotype ST152. *J Bacteriol* **190**: 3962-3968.

Ruimy R, Angebault C, Djossou F*, et al.* (2010) Are host genetics the predominant determinant of persistent nasal *Staphylococcus aureus* carriage in humans? *J Infect Dis* **202**: 924-934.

Sanders H (1968) Marine benthic diversity: A comparative study. *The American Naturalist* **102**: 243-283.

Schaumburg F, Köck R, Friedrich AW*, et al.* (2011) Population Structure of *Staphylococcus aureus* from Remote African Babongo Pygmies. *PLoS Negl Trop Dis* **5**: e1150.

Soergel DA, Dey N, Knight R & Brenner SE (2012) Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *ISMEJ* **6**: 1440-1444.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M & Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**: 2731-2739.

Taubel M, Rintala H, Pitkaranta M*, et al.* (2009) The occupant as a source of house dust bacteria. *J Allergy Clin Immunol* **124**: 834-840 e847.

van Belkum A, Verkaik NJ, de Vogel CP*, et al.* (2009) Reclassification of *Staphylococcus aureus* nasal carriage types. *J Infect Dis* **199**: 1820-1826.

Verdu P, Austerlitz F, Estoup A*, et al.* (2009) Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Curr Biol* **19**: 312-318.

Wang Q, Garrity GM, Tiedje JM & Cole JR (2007) Naïve bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**: 5261-5267.

Warwick RM & Clarke KR (1995) New "biodiversity" measures reveal a decrease in taxonomic distinctness with increasing stress. *MAR ECOL-PROG SER* **129**: 301-305.

Werner JJ, Zhou D, Caporaso JG, Knight R & Angenent LT (2012) Comparison of Illumina paired-end and single-direction sequencing for microbial 16S rRNA gene amplicon surveys. *ISMEJ* **6**: 1273-1276.

Wos-Oxley ML, Plumeier I, von Eiff C*, et al.* (2010) A poke into the diversity and associations within human anterior nare microbial communities. *ISMEJ* **4**: 839-851.

Figure legends

Fig. 1. Sequencing error rates. Error rate at each nt over the full paired-end sequence read (245 nt in length) for all 16S rRNA gene fragments derived from *S. aureus* (DSM 3463), *C. accolens* (GU074966), *S. epidermidis* (JF927883) or *M. nonliquefaciens* (JF927886) before the quality check (local average score) was applied. The error rate increased dramatically after 80 nt of each paired-end read with each of the 4 species tested.

Fig. 2. Taxonomic breadth of the 251 bacterial phylotypes. The anterior nare phylotypes of westernized and non-westernized adults, and non-westernized children, based on a maximum likelihood tree constructed from a pre-aligned set of 16S rRNA gene sequences from the closest representative taxa obtained from the SILVA database (Fig. S4). Branches denoted in grey represent phylotypes common to all three groups whereas those in bold were missing from the westernized adults (black dots), and those with broken lines from the non-westernized adults and children (white dots). Phylum level designations are indicated on the outside of the circle with corresponding suborder/family marked as: CY, *Corynebacterineae*; P, *Pseudonocardineae*; PB, *Propionibacterineae*; A, *Actinomycineae*; MC, *Micrococcineae*; L, *Leptotrichiaceae*; F, *Fusobacteriaceae*; CD, *Clostridia*; N, *Negativicutes*; BC, *Bacilli*; δ/β/γ-, proteobacteria classes; FB, *Flavobacteria*; and BD, *Bacteroidia*. Scale bar represents 5% nucleotide sequence divergence.

Fig. 3. Global bacterial community structure. Non-metric multidimensional scaling (nMDS) plot comparing the global bacterial community structure of 190 human anterior nares, where westernized adults are denoted by red dots, non-westernized adults denoted by light blue dots and non-westernized children denoted by dark blue

dots. Phylotype abundance (% sequence reads) was standardized but untransformed prior to the use of the Bray-Curtis similarity algorithm. While a 2D stress value of 0.19 indicates some stress on the plot, it is deemed acceptable considering the large number of samples being ordinated.

Fig. 4. Taxonomic composition of the anterior nares' microbiota. Relative abundance of (A) bacterial phyla and (B) bacterial families within the anterior nares of each of the 3 broad groups of samples, westernized adults, non-westernized adults and non-westernized children.

Fig. 5. Phylotype diversity and richness. Funnel plots charting average Taxonomic Distinctness (delta+) and variation in Taxonomic Distinctness (lambda+) against the number of phylotypes within each sample, where westernized adults are denoted as red dots, non-westernized adults are denoted as light blue dots and non-westernized children are denoted as dark blue dots.

Supplementary Information

Fig S1. Correlation co-efficients after comparing aggregated data matrices. The correlation co-efficient (rho) derived when each of the genus-, family-, order-, class- and phylum-level data matrices were compared to the original species-level matrix, as determined by using the RELATE routine in PRIMER.

Fig. S2. Sampling effort across each sample, individual-based rarefaction curves. Rarefaction curves portraying the number of resolved phylotypes against sampling depth of each sample within the (A) non-westernized population and (B) westernized population.

Fig. S3. Sampling effort across each population, species accumulation curves. Estimators of species richness are the total number of all species (Sobs) and the Chao 2 estimator of true richness. Plotted values are the mean +/- standard deviation of 999 permutations. (A) curves representing both westernized and non-westernized populations, (B) curves representing both non-westernized adults and children.

Fig. S4. Maximum-likelihood tree depicting the taxonomic breadth of the 251 phylotypes detected from the anterior nares. The taxonomic breadth of westernized adults (red dots), non-westernized adults (light blue dots) and non-westernized children (dark blue dots), as represented by complete or near complete 16S rRNA gene sequences from their closest taxonomic relatives available from the SILVA database. GenBank accession numbers are given after each strain name. Branch support values were calculated from 1000 bootstrap re-samplings (values > 85% are given at the nodes). Scale bar represents 5% nucleotide sequence divergence.

Fig. S5. Comparing the relative abundance and prevalence of selected phylotypes in the anterior nare bacterial communities. (A) *Kocuria* sp., (B) *Kocuria marina* and (C) *Kocuria koreensis* observed in westernized and non-westernized adults.

Fig. S6. Non-metric multidimensional scaling (nMDS) plot comparing the global bacterial community structure of 190 human anterior nares. (A) Community structures across 147 adult volunteers (non-westernized (light blue dots) n=55, westernized (red dots) n=92). (B) Community structures across 92 non-westernized volunteers (non-westernized adults (light blue dots) n=55, non-westernized children (dark blue dots) n=43). For each phylotype the amount of sequence reads was standardized (%) but

untransformed prior to the use of the Bray-Curtis similarity algorithm. While 2D stress values of 0.18 and 0.19 indicate some stress on the plots, it is deemed acceptable considering that so many samples are being ordinated.

Fig. S7. Non-metric multidimensional scaling (nMDS) plot with superimposed bubbles. (A,C,E,G) Superimposed bubbles onto the ordination plot of Fig. S6A ordinating both adult populations (where westernized adults are denoted by an asterisk and non-westernized adults denoted by a hash). (B,D,F,H) Superimposed bubbles onto the ordination plot of Fig. S6B ordinating non-westernized children and adults (where non-westernized adults are denoted by a hash and non-westernized children denoted by a cross). Bubbles represent the relative abundance of (A-B) *C. accolens* (PT1-1), (C-D) *P. acnes* (PT3-1), (E-F) S*. epidermidis/ S. capitis/ S. caprae* (PT38-1) and (G-H) *Peptoniphilus* sp. (PT44-1).

Fig. S8. Comparing the relative abundance and prevalence of selected phylotypes in the anterior nare bacterial communities. (A) *S. aureus*, (B) *M. lacunata/ M. nonliquefaciens* and (C) *M. catarrhalis* observed in females and males in the non-westernized adults, (D) *M. lacunata/ M. nonliquefaciens* and (E) *M. catarrhalis* observed in females and males in the westernized adults.

Fig. S9. Non-metric multidimensional scaling (nMDS) plot with superimposed bubbles. (A,C,E) Superimposed bubbles onto the ordination plot of Fig. S6A ordinating both adult populations (where westernized adults are denoted by an asterisk and non-westernized adults denoted by a hash). (B,D,F) Superimposed bubbles onto the ordination plot of Fig. S6B ordinating non-westernized children and adults (where non-westernized adults are denoted by a hash and non-westernized

children denoted by a cross). Bubbles represent the relative abundance of (A-B) *Streptococcus pneumoniae/ S. mitis* (PT59-1), (C-D) *Dolosigranulum pigrum* (PT66-1), (E-F) *Dolosigranulum* sp. (PT72-1).

Fig. S10. Non-metric multidimensional scaling (nMDS) plot with superimposed bubbles. (A,C,E) Superimposed bubbles onto the ordination plot of Fig. S6A ordinating both adult populations (where westernized adults are denoted by an asterisk and non-westernized adults denoted by a hash). (B,D,F) Superimposed bubbles onto the ordination plot of Fig. S6B ordinating non-westernized children and adults (where non-westernized adults are denoted by a hash and non-westernized children denoted by a cross). Bubbles represent the relative abundance of (A-B) *Staphylococcus aureus* (PT26-1), (C-D) *Moraxella lincolnii* (PT98-1), (E-F) *Moraxella catarrhalis* (PT123-1).

Fig. S11. Ecological biodiversity indices of the anterior nare bacterial communities of non-westernized adults and children and westernized adults. The indices presented here are: total phylotypes (S), Shannon diversity (H'), Pielou's evenness (J'), Simpson index (1-lambda), average Taxonomic Distinctness (delta+) and variation in Taxonomic Distinctness (lambda+).

Table S1. Nucleotide sequences of all 251 phylotypes determined using Illumina-based amplicon deep-sequencing.

Table S2. Phylogenetic assignment. Description of all 251 phylotypes determined using Illumina-based amplicon deep-sequencing and the RDP database.

Table S3. Gender and age of westernized volunteers. Information on each of the 92 westernized volunteers of this study.

Table S4. Gender and age of non-westernized volunteers. Information on each of the 98 non-westernized volunteers of this study.

Table S5. Primers used in this study.