ORIGINAL ARTICLE 10.1111/j.1469-0691.2012.03855.x

The adult intestinal core microbiota is determined by analysis depth and health status

A. Salonen[1,2], J. Salojärvi[1], L. Lahti[1,3] and W. M. de Vos[1,2,3]

1) *Department of Veterinary Biosciences, University of Helsinki, Helsinki, Finland,* 2) *Department of Bacteriology and Immunology, University of Helsinki, Helsinki, Finland and* 3) *Laboratory of Microbiology, Wageningen University, Wageningen, The Netherlands*

Abstract

High-throughput molecular methods are currently exploited to characterize the complex and highly individual intestinal microbiota in health and disease. Definition of the human intestinal core microbiota, i.e. the number and the identity of bacteria that are shared among different individuals, is currently one of the main research questions. Here we apply a high-throughput phylogenetic microarray, for a comprehensive and high-resolution microbiota analysis, and a novel computational approach in a quantitative study of the core microbiota in over 100 individuals. In the approach presented we study how the criteria for the phylotype abundance or prevalence influence the resulting core in parallel with biological variables, such as the number and health status of the study subjects. We observed that the core size is highly conditional, mostly depending on the depth of the analysis and the required prevalence of the core taxa. Moreover, the core size is also affected by biological variables, of which the health status had a larger impact than the number of studied subjects. We also introduce a computational method that estimates the expected size of the core, given the varying prevalence and abundance criteria. The approach is directly applicable to sequencing data derived from intestinal and other host-associated microbial communities, and can be modified to include more informative definitions of core microbiota. Hence, we anticipate its utilization will facilitate the conceptual definition of the core microbiota and its consequent characterization so that future studies yield conclusive views on the intestinal core microbiota, eliminating the current controversy.

Keywords: 16S rRNA, core microbiota, health, intestinal microbiota, phylogenetic microarray
Original Submission: 9 April 2011; **Revised Submission:** 10 October 2011; **Accepted:** 11 October 2011
Editors: A. Moya, Rafael Cantón, and D. Raoult
Clin Microbiol Infect 2012; **18** (Suppl. 4): 16–20

Corresponding author: W. M. de Vos, Department of Veterinary Biosciences, P.O. Box 66, FI-00014, University of Helsinki, Finland
E-mail: willem.devos@wur.nl

Introduction

From birth the gastrointestinal (GI) microbiota constitute the largest microbial ecosystem of the human body. Recent studies with culture-independent molecular methods have revealed that, while the exact GI microbiota composition is highly individual specific [1], a typical gut ecosystem harbours thousands of phylotypes from less than ten bacterial phyla dominated by the Firmicutes, Bacteroidetes, Actinobacteria, Proteobacteria and Verrucomicrobia. The collective genome of the intestinal microbes vastly surpasses the coding capacity of the human genome with more than 3 million genes [2]. Hence, we are composite organisms co-programmed by the inherited human genome and the environmentally acquired microbiome. The health relevance of the GI microbiome lies in its capacity to provide the host with vital and irreplaceable functions ranging from the energy and vitamin metabolism to epithelial barrier integrity and immune modulation [3].

The vast majority of the GI tract microbes have not yet been cultured and are only recognized with molecular methods based on 16S rDNA sequences. A panoply of high-throughput approaches have been developed to describe the GI microbiota, including deep new generation sequence analysis and phylogenetic microarrays [1]. Using these approaches, the loss of homeostasis in the host–microbe symbiosis has recently been associated with a wide variety of intestinal and systemic diseases (reviewed in [4]). Disease-associated compositional and functional alterations of the GI microbiota are

actively being investigated. However, the immense complexity and large inter-individual variability of the microbiota hamper the current ability to resolve differences between the bacterial communities of patients and controls and call for robust and sufficiently powered studies to get an insight into the microbial aetiology of specific diseases.

Despite the accumulating data provided by modern molecular techniques, current knowledge does not yet offer a definition for a normal or optimal GI microbiota composition. In parallel with mining the entire diversity of host-associated microbial communities, recently significant effort has been devoted to a more focused approach that aims to define a core microbiota that is potentially shared across adult individuals [2,5–10]. The specific interest towards universally shared bacteria arises from the fact that, in contrast to transient gut inhabitants that fluctuate depending on the diet and other environmental factors, the common core bacteria are conserved during the mutual coevolution of man and his intestinal microbes. Consequently, the core microbiota is anticipated to represent a selected set of health-associated symbionts. Once catalogued, the targeted characterization of the core bacteria would provide a scientifically sound and economically relevant strategy to access the GI microbes that are the most relevant for human health and may hold diagnostic or therapeutic potential. Although the basic definition of the core microbiota is intuitive, there are currently several unaddressed questions relating to its biological and analytical parameters. How many individuals need to be studied? Should their overall or even intestinal health status be defined and, if yes, based on which parameters [11]? Do we qualify only bacteria that are detected in 100% of the individuals or is a lower prevalence threshold justifiable to provide robustness against technical variation? Finally, are we interested only in the dominant bacteria or should we use analytical methods that also allow mining of the rare intestinal biosphere?

Owing to lack of the above-mentioned definitions, the number and health status of the study subjects as well as the required prevalence for core species has varied considerably among current studies describing the human intestinal core microbiota. Moreover, the effect of analysis depth has so far been largely ignored. Current estimates of the taxonomic overlap between individuals range from 0–2% [5,6] to over 30% [2,7] and thus lack consensus. To tackle the current controversy, we carried out phylogenetic microarray analysis of the GI microbiota derived from more than 100 individuals. The data were used to examine the impact of analytical resolution (depth) and coverage (width) as well as of biological variables (subject number and health status) as determinants of the common core microbiota.

Materials and Methods

The data set used consisted of faecal microbiota profiles obtained by the Human Intestinal Tract Chip (HITChip), a phylogenetic microarray covering over 1000 different intestinal phylotypes [9]. Single time point HITChip profiles from 127 unrelated European individuals that were derived from ongoing clinical or observational trials were extracted from an in-house data collection of over 1000 microarray experiments [12]. The main data set was composed of 115 healthy subjects, devoid of GI or other diseases. The mean age of the subjects was 40, the mean body mass index was 24 kg/m^2 and two-thirds of the subjects were female. To benchmark the healthy microbiota, we used as a reference HITChip data measured from faecal samples collected from 12 ulcerative colitis (UC) patients. The faecal samples were collected and stored according to established procedures and the faecal DNA was extracted as previously described [12].

All HITChip microarray analyses and computational preprocessing including signal thresholding were performed as previously described [9,10]. The analysis was carried out using phylotype (species-like) level signals that were estimated from the hybridizing HITChip probes by using the robust probabilistic averaging algorithm [13]. This method provides a robust estimate of the average signal of the HITChip probes targeting the same phylotype by giving less weight to probes showing sensitivity to noise attributable to unintended cross-hybridization with non-target sequences. It is anticipated that this method reduces the number of false positives in the common core analysis.

In rarefaction analysis, for each sample size a set of 10 000 bootstrap replicates were sampled from the full data set of healthy or UC patients. For each set, the detection threshold was chosen randomly from a uniform distribution between the minimum detection threshold and the observed maximum intensity, and the number of detected species was counted. The common core microbiota was addressed by thresholding the HITChip phylotype (species-like) level data in a grid of logarithmic signal intensity (range 1.93–4.98) and prevalence (number of carriers, range 1–115) as previously described [12]. The resulting surface was visualized with a perspective plot [14].

The number of common microbes in an arbitrary set of samples was estimated by bootstrapping, where in each bootstrap set the intensity threshold was selected randomly and the number of common microbes was computed. The set of phylotypes belonging to the core microbiota of UC or healthy patients was estimated from 10 000 bootstrap samples where, in addition to the intensity threshold also the

prevalence was selected randomly. Additionally, to balance the difference in sizes between the two data sets (UC vs healthy), 12 samples from healthy patients were chosen for each bootstrap round. The results were collected into a frequency table reporting the number of times each phylotype fulfilled the criteria. The core sizes estimated above were then invoked to select the most frequently occurring phylotypes in the core microbiota of the UC or healthy patients. All data analyses were performed in R version 2.12.1 [15].

Results

We addressed the common core of the human GI microbiota by using high-resolution and highly reproducible microarray data sets derived from over 100 individuals that were mined with a flexible computational approach. To generate an overview of the data, we carried out principal component analysis and hierarchical clustering that ensured sufficient homogeneity of the data to carry out the meta-level analysis of the common core (data not shown).

To estimate the representativeness of the data set, and to assess how many individuals are actually needed to reliably determine the core size, we performed a rarefaction analysis for the detected phylotypes (Fig. 1). We observed that already in a set of only a few dozen subjects the vast majority of the total richness was captured, indicated by the levelling of the line towards the horizontal. However, there was no plateau in the number of detected phylotypes, signifying a constant increase in the detected richness even after 100 individuals (Fig. 1). Altogether, the rarefaction analysis indicated that the number of samples provided a sufficient and representative data set to address the core microbiota.

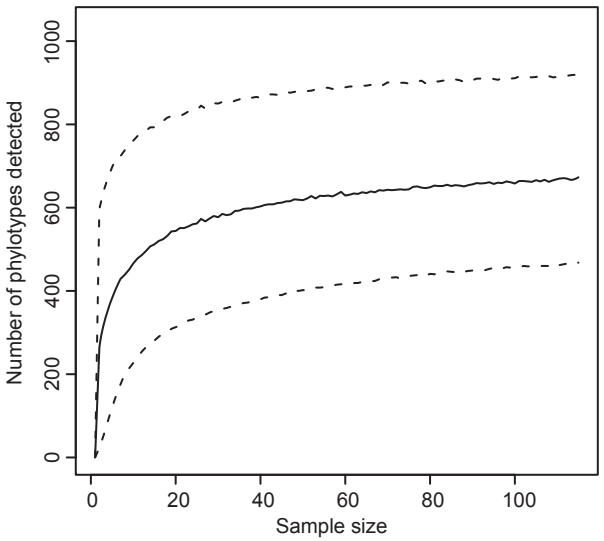

FIG. 1. Overview of the total phylotype richness in 115 healthy subjects. Rarefaction curve showing the number of phylotypes (*y*-axis) that are detected after analysing the number of subjects shown on the *x*-axis. Dashed lines indicate the 95% confidence intervals of the mean.

In the absence of consensus criteria for the analytical common core parameters, we included the complete range of abundance and prevalence values, from minimum to maximum. The HITChip signal intensities varied over 1000-fold ($\log_{10}$ ratio of 3.1), signifying high variability in the abundance of the phylotypes. All possible prevalence values were accommodated, denoting the presence of a given phylotype in 1–115 subjects. As a result, the common core size appears as a continuum from zero to several hundreds of phylotypes, depending on the selected abundance and prevalence values as visualized with a perspective plot (Fig. 2a). Plotting the number of core phylotypes on different abundance thresholds visualizes the strong dependence between these factors (Fig. 2b). Consequently, there was no common core if a phylotype was required to be present in high abundance (>2.5% of the total signal) in all subjects, but when we included also the low abundance bacteria, as many as 30% (290 phylotypes) of the microbiota were shared by all 115 individuals. The true core size should lie somewhere between these two extremes.

To estimate the true core size, we computed the mean core size that could be detected in a random sample of healthy individuals. We required absolute (100%) prevalence in 115 individuals and applied bootstrap analysis to average over different abundance thresholds between the minimum and maximum (Fig. 3). The expected number of shared phylotypes in the given cohort was around 100 phylotypes. Notably, the size of the core levelled off fast, suggesting that the mean core size can be estimated already from a few healthy individuals.

The effect of health status on the core size was analysed by including HITChip data from 12 UC patients in the data set and comparing them with the microbiota of 12 healthy subjects derived with the bootstrap procedure. Separate analysis of the UC and healthy cores indicated a significantly smaller core size in the UC patients (Fig. 4a). This finding was confirmed by pooling the UC and healthy data sets, which yielded a core size intermediate between the health-status-specific ones. To study the compositional overlap between the healthy and UC cores, a comparative analysis was carried out (Fig. 4b). Altogether 58% of the core phylotypes were common and thus independent of health status, while 25% and 17% were specific to healthy or UC subjects, respectively.

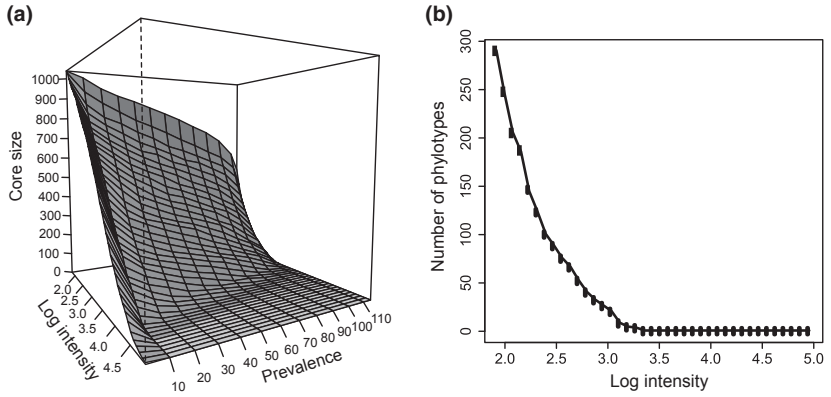

FIG. 2. Definition of the common core microbiota. (a) Perspective plot visualizing the number of core phylotypes as a function of the prevalence and abundance (indicated as logarithmic values of the signal intensity). (b) The common core size in 115 healthy subjects. The *y*-axis represents the number of shared phylotypes at different abundance (logarithmic values of the signal intensity, *x*-axis).

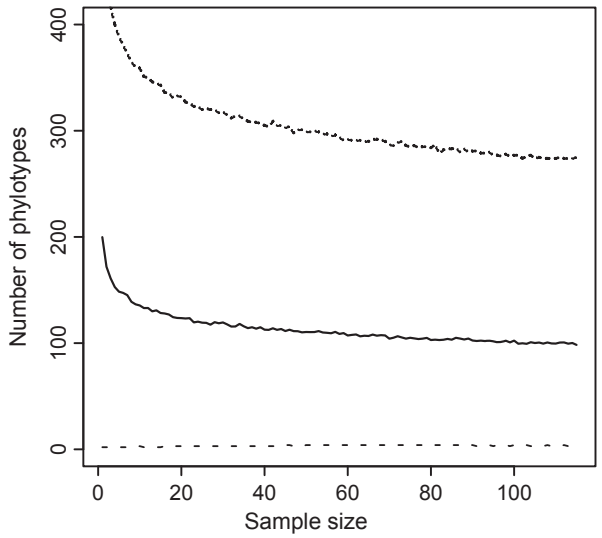

FIG. 3. Averaged, abundance-independent core size in 115 healthy subjects. In order to refrain from using predefined abundance thresholds in the definition of the core, an average core was calculated by bootstrapping the signal intensities and supposing 100% prevalence. Dashed lines indicate the 95% confidence intervals of the mean.

Discussion

In this work we utilized one of the largest human gut microbiota data sets available, consisting of 16S rRNA gene-based high-resolution microarray profiles. Collective data from over 100 subjects were analysed using a novel computational approach to characterize the common core microbiota in healthy subjects, and to test the impact of three key variables in the common core analysis: (i) analysis depth, by using different abundance thresholds for the phylotypes; (ii) number of subjects, by carrying out rarefaction and bootstrap analyses; and (iii) the health status of the subjects, by including microbiota of UC patients.

Our results show that the core size is highly conditional, depending on both technical and biological variables, i.e. the depth of the analysis, the prevalence of the taxa as well as the number and health status of the study subjects. The deterministic impact of the coverage of analysis has been indicated also in a previous study where doubling of the sequencing depth increased the amount of shared phylotypes by 25% [2]. So far, most studies on core microbiota have targeted the phylotypes that are predominant in all individuals and thus have excluded a substantial part as the abundance of phylotypes may vary over 2000-fold across individuals [2].

By using a criterion of 100% prevalence and including also the low abundant phylotypes, we found that one-third of these were shared among the 115 healthy subjects. The proportion of shared phylotypes reported here is considerably larger than those reported previously in sequencing-derived estimates, which have ranged from 0–2% [5,6] to about 30% [2,7] using notably lower stringency for the prevalence (≥50%). It should be noted that while sequencing discovers novel sequences, microarrays are limited to previously detected phylotypes for which they provide a rapid and powerful profiling. The HITChip thus provides a closed system, which covers also phylotypes with low relative abundance (below 0.02%). These are not accessible with conventional sequencing depth [2,16] and therefore have been missed in previous core analyses. However, overestimation of our core size due to cross-hybridization of non-target phylotypes cannot be excluded, and thus further studies are needed to verify the true dimensions of the common core microbiota.

In this study, the health status introduced much more variation to the core than the sole number of studied subjects.

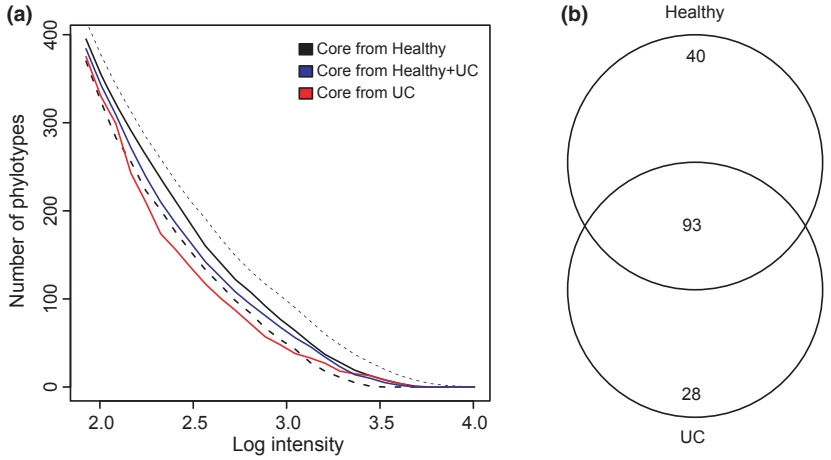

FIG. 4. Impact of health status in the common core microbiota. (a) The core size of the intestinal microbiota of 12 healthy subjects (black), 12 patients suffering from UC (red) and their pooled combination (blue) was calculated by bootstrapping. Dashed lines indicate the 95% confidence intervals for the healthy core. (b) Venn diagram showing the compositional overlap between the UC and healthy core. The phylotypes constituting the cores were compared as explained in detail in the text.

We detected a smaller and markedly different core in UC patients compared with healthy subjects (Fig. 3), in line with the previous indication [7]. The smaller core of UC patients suggests loss of certain health-specific core bacteria and potentially more heterogeneous total microbiota as a proxy of lost homeostasis. The latter would explain why it has been difficult to find consistent, disease-specific microbiota alterations.

In summary, our data indicate that when the full spectrum of the highly uneven abundance distribution of intestinal phylotypes is detected, one-third of the phylotypes are shared among all the studied individuals. These bacteria can be seen as a conserved community that does not co-vary, e.g. with the genetic or dietary variation within individuals. The remaining two-thirds of the phylotypes were shared to a variable extent, i.e. between 114 and two individuals. It can be speculated that bacteria outside the core are more strongly influenced by the genotypic and environmental variation of the subjects, and perhaps more susceptible to modulation.

Transparency Declaration

The authors declare that they have no conflicts of interest.

References

1. Zoetendal EG, Rajilić-Stojanović M, de Vos WM. High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota. *Gut* 2008; 57: 1605–1615.
2. Qin J, Li R, Raes J et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 2010; 464: 59–65.
3. Blaut M, Clavel T. Metabolic diversity of the intestinal microbiota: implications for health and disease. *J Nutr* 2007; 137: 751S–755S.
4. Gerritsen J, Smidt H, Rijkers GT, de Vos WM. Intestinal microbiota in human health and disease: the impact of probiotics. *Genes Nutr* 2011; 6: 209–240.
5. Tap J, Mondot S, Levenez F et al. Towards the human intestinal microbiota phylogenetic core. *Environ Microbiol* 2009; 11: 2574–2584.
6. Turnbaugh PJ, Hamady M, Yatsunenko T et al. A core gut microbiome in obese and lean twins. *Nature* 2009; 457: 480–484.
7. Willing B, Dicksved J, Halfvarson J et al. A pyrosequencing study in twins shows that GI microbial profiles vary with inflammatory bowel disease phenotypes. *Gastroenterology* 2010; 39: 1844–1854.
8. Sekelja M, Berget I, Naes T, Rudi K. Unveiling an abundant core microbiota in the human adult colon by a phylogroup-independent searching approach. *ISME J* 2010; 5: 519–531.
9. Rajilic-Stojanovic M, Heilig HG, Molenaar D et al. Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. *Environ Microbiol* 2009; 11: 1736–1751.
10. Jalanka-Tuovinen J, Salonen A, Nikkilä J et al. Intestinal microbiota in healthy adults: temporal analysis reveals individual and common core and relation to intestinal symptoms. *PLoS ONE* 2011; 6: e23035.
11. Bischoff SC. 'Gut health': a new objective in medicine? *BMC Med* 2011; 9: 24.
12. Nikkilä J, de Vos WM. Advanced approaches to characterize the human intestinal microbiota by computational meta-analysis. *J Clin Gastroenterol* 2010; 44: S2.
13. Lahti L, Elo LL, Aittokallio T, Kaski S. Probabilistic analysis of probe reliability in differential gene expression studies with short oligonucleotide arrays. *IEEE/ACM Trans Comput Biol Bioinform* 2011; 8: 217–225.
14. Becker RACJMWAR. *The New S Language*. London: Chapman & Hall, 1988; 702.
15. R Development CT. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing, 2010.
16. Hamady M, Knight R. Microbial community profiling for human microbiome projects: tools, techniques, and challenges. *Genome Res* 2009; 19: 1141.