LETTER

doi:10.1038/nature13178

Dynamics and associations of microbial community types across the human body

Tao Ding[1] & Patrick D. Schloss[1]

A primary goal of the Human Microbiome Project (HMP) was to provide a reference collection of 16S ribosomal RNA gene sequences collected from sites across the human body that would allow microbiologists to better associate changes in the microbiome with changes in health[1]. The HMP Consortium has reported the structure and function of the human microbiome in 300 healthy adults at 18 body sites from a single time point[2,3]. Using additional data collected over the course of 12–18 months, we used Dirichlet multinomial mixture models[4] to partition the data into community types for each body site and made three important observations. First, there were strong associations between whether individuals had been breastfed as an infant, their gender, and their level of education with their community types at several body sites. Second, although the specific taxonomic compositions of the oral and gut microbiomes were different, the community types observed at these sites were predictive of each other. Finally, over the course of the sampling period, the community types from sites within the oral cavity were the least stable, whereas those in the vagina and gut were the most stable. Our results demonstrate that even with the considerable intra- and interpersonal variation in the human microbiome, this variation can be partitioned into community types that are predictive of each other and are probably the result of life-history characteristics. Understanding the diversity of community types and the mechanisms that result in an individual having a particular type or changing types, will allow us to use their community types to assess disease risk and to personalize therapies.

Building on previous analysis of a healthy cohort of 300 individuals, we analysed a 16S rRNA gene sequence data set from the HMP Consortium[2,3]. The final data release for this cohort provided 16S rRNA gene sequence data and clinical metadata (Extended Data Table 1) from two time points for each of 300 healthy individuals and from a third time point for 100 of the individuals at 15 body sites for men and 18 for women[5]; the interval between samplings varied between 30 and 451 days (median = 224 days). A significant difficulty in analysing microbiome data has been the considerable intra- and interpersonal variation in the composition of the human microbiome[3,6,7]. A recently proposed approach for overcoming this difficulty within the gastrointestinal tract has been the concept of enterotypes, or more generically, stool community types[4,8,9]. In this approach samples are clustered into bins based on their taxonomic similarity. Specific enterotypes have been associated with the amount of protein, fat and carbohydrates in one's diet, obesity, inflammatory bowel disease, and Crohn's disease[4,9–11]. Others have found associations between specific vaginal community types and the sexually transmitted *Trichomonas vaginalis*, pH, and ethnicity[12–14] and associations between skin community types and psoriasis[15]. Using bacterial community structures collected from 18 body sites and up to three time points, we applied community typing analysis to understand better the factors that affect the structure of the microbiome and contribute to human health.

Concern has been expressed regarding whether community types reflect partitioning of an abundance gradient or the presence of clusters of relative abundance profiles[8,16]. Two general approaches have been developed to assign samples to community types: partitioning around the medoid (PAM) and Dirichlet multinomial mixture (DMM) models[4,8]. To compare these methods we first generated simulated communities where there were one or four community types. Analysis of the simulated communities indicated that the negative log model evidence metric used by the DMM-based approach was superior to the metrics used to assess clusters within the PAM-based approach (Supplementary Information). Next, we assigned the samples for each body site to community types using both methods. Calculation of the negative log model evidence demonstrated that the community types identified using DMM were superior to those identified using the PAM-based approach (Extended Data Table 2 and Extended Data Fig. 1). Thus, our analysis of simulated data and the HMP data suggests that the community types represent clusters of relative abundance profiles.

Using the DMM-based approach, we identified between two (anterior nares) and seven (tongue dorsum) community types per body site (see Source Data associated with Fig. 1 for community data and DMM fits). As an example, bacteria from stool samples fell into four distinct community types (Fig. 1a). We observed that 63 genera were needed to account for 90% of the difference between a model with a single community type and four community types (see Source Data associated with Fig. 1). Thus, it was not merely the most abundant bacterial population that differentiated the types as has been previously reported (for example, *Bacteroides*, *Prevotella*, or *Ruminococcus*)[8–10,17]; rather, community types were identified based on complex configurations of numerous taxa. In fact, this supports the findings of the original study; that is, the taxa that typify each enterotype represent networks of co-occurring bacterial populations[9]. Inspection of the five most important genera, which accounted for 54% of the difference in fit between four community types and one, indicated that each community type represented a cluster of relative abundance profiles (Fig. 1b). Community type A had the highest levels of *Bacteroides* but lacked *Prevotella* and Ruminococcaceae. Similar to community type A, community type C also lacked *Prevotella*, but had a lower relative abundance of *Bacteroides* and had higher levels of *Alistipes*, *Faecalibacterium* and Ruminococcaceae. Community type D had fewer *Bacteroides* than community types A and C, but had higher levels of *Prevotella*. Community type B had the fewest *Bacteroides* and was dominated by a variety of populations affiliated within the Firmicutes. Furthermore, the diversity of the samples assigned to each of the community types indicated that type A had a significantly lower diversity than the other three types ($P < 0.001$). Community types A, C and D resembled the previously identified *Bacteroides*, *Ruminococcus* and *Prevotella* enterotypes, respectively[9,10,17]. Analysis of the other body sites yielded analogous patterns.

Using the responses that subjects gave to an extensive survey (summarized in Extended Data Table 1), we identified demographic and life-history characteristics that could be correlated with different community types at each body site. Of the numerous characteristics tested, we observed significant associations between community types and whether the subject was ever breastfed, their gender, and their education level (see Source Data associated with Fig. 1). Whether an individual was ever breastfed was strongly associated with their stool community type ($P = 1 \times 10^{-4}$; Fig. 1c). Individuals who had been breastfed at some point as infants

[1]Department of Microbiology and Immunology, 1500 W. Medical Center, University of Michigan, Ann Arbor, Michigan 48109, USA.

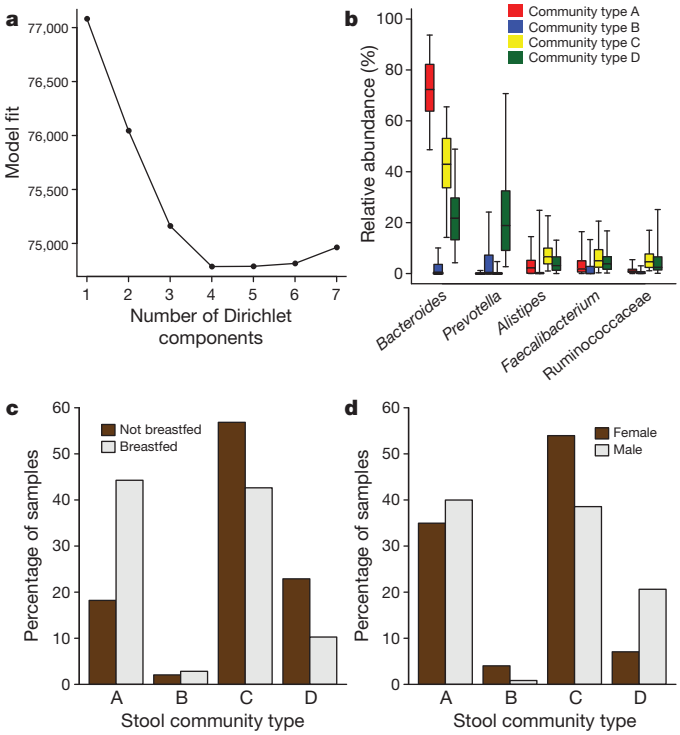


Figure 1 | Analysis of stool samples reveals four community types. **a**, Fitting the genera-level relative abundance data from 597 stool samples to Dirichlet multinomial mixture models provided support for four types when using the Laplace approximation to the negative log model evidence. **b**, The relative abundance of the most abundant genera in the samples assigned to each of the types (the boxes represent the interquartile range and the error bars represent the 95% confidence interval; *n* (community type A) = 221; *n* (community type B) = 15; *n* (community type C) = 80; *n* (community type D) = 281). **c**, **d**, There were significant associations between stool community types ($n = 287$ unique individuals) and whether the subject was breastfed as an infant (**c**; median $P = 1 \times 10^{-4}$) and their gender (**d**; median $P = 4 \times 10^{-4}$).

were 2.4-times more likely to belong to community type A, and those who were not breastfed were 2.2-times more likely to belong to community type D. Gender was associated with community types identified in the stool ($P = 4 \times 10^{-4}$; Fig. 1d), tongue ($P = 2 \times 10^{-3}$; Extended Data Fig. 2a), right retroauricular crease ($P = 9 \times 10^{-5}$; Extended Data Fig. 2b), and right antecubital fossa ($P = 3 \times 10^{-5}$; Extended Data Fig. 2c). For example, men were 3.0-times more likely than women to harbour stool community type D (Fig. 1b). Whether a woman had a baccalaureate degree had a strong association with the community types observed within the vaginal introitus ($P = 2 \times 10^{-3}$; Extended Data Fig. 3a), mid vagina ($P = 8 \times 10^{-4}$; Extended Data Fig. 3B), and posterior fornix ($P = 4 \times 10^{-4}$; Extended Data Fig. 3C). At each of these sites, women with a baccalaureate degree were more likely to be dominated by *Lactobacillus* (type E) and those without a baccalaureate degree were likely to have very low levels of *Lactobacillus* and moderate abundances of *Atopobium*, *Prevotella*, *Bifidobacterium* and unclassified members of the Firmicutes (type D). Together, our analysis indicates that an individual's life-history characteristics can be associated with their microbiome composition.

The second important observation that we identified was that the community type at one body site was predictive of the community type at another body site. Previously, cross-body site comparisons were made by calculating the ecological distance between samples collected at different body sites based on the taxonomic composition of those communities[3]. Our approach allowed us to identify similar associations within a body region (for example, oral, skin, vagina), but also allowed us to detect associations between communities that had very different taxonomic compositions. Community type membership was correlated among sites within the oral cavity, in the vagina, and between the left and right antecubital fossa and the left and right retroauricular crease (Fig. 2). Surprisingly, stool samples showed a significant association with samples from within the oral cavity; the strongest association was with the community types observed in saliva ($P = 10^{-3}$; Extended Data Table 3). Saliva was dominated by members of the *Prevotella*, *Streptococcus*, Pasteurellaceae, *Veillonella* and *Fusobacterium*; among these taxa, only *Prevotella* were abundant in the stool communities. Individuals with stool community type D, which had the highest level of *Prevotella*, were 2.1-times more likely to harbour saliva community types A and C, which were also high in *Prevotella* relative to saliva community types B and D. Stool community types A and C, which had low levels of *Prevotella*, were less likely to co-occur with saliva community types A and C (Extended Data Table 3). These results are intriguing because they suggest that although the oral and stool communities share little taxonomic resemblance, oral bacterial populations seed the gut, and those populations experience the ecological environment of the gut to give rise to consistent community types by the time they reach the stool.

Aside from life-history characteristics and inoculation from other body sites, the structure of the human microbiome is probably shaped by an individual's recent interactions with their environment, diet, medications, and overall health. We quantified the stability of each community type at every body site by estimating the probability that the type would change between sampling visits (Fig. 3a). The most stable body sites were in the stool and vagina and the least stable site was the supragingival plaque. Among the four stool community types, type D was the most stable followed by types A, C and B (Fig. 3b). Unfortunately, the metadata describing changes in health or lifestyle are unable to provide us with an explanation for why community types change.

The human microbiome is a complex ecosystem that varies considerably across the body and between individuals. This study demonstrates that given the myriad permutations of genetics, life histories, behaviours, environments and exposures, an individual's microbiome is an emergent property whereby a potentially limitless number of microbial community structures can be distilled into a finite number of types. Knowledge of

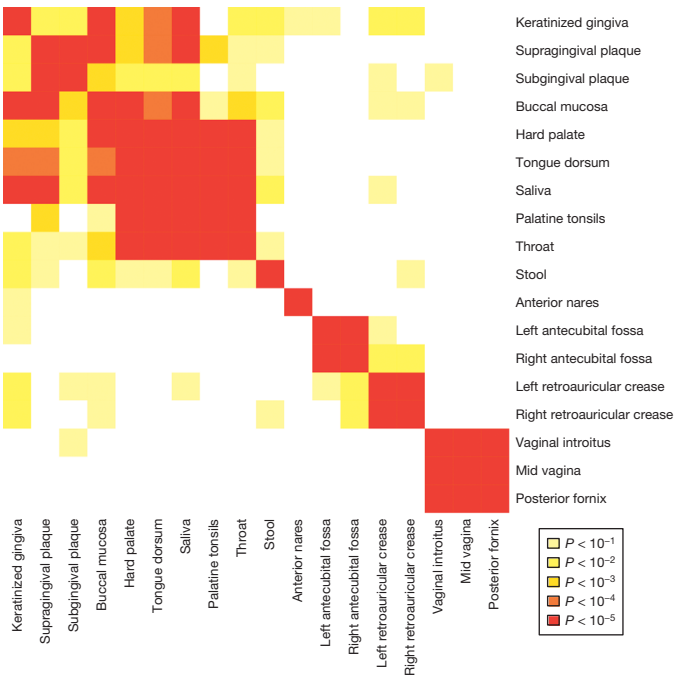


Figure 2 | Community-type associations are strongest within a body region, but also exist between stool and the oral cavity. Heat-map colours represent the magnitude of the median *P* value for the comparison of community type membership using Fisher's exact test. Median *P* values are found in the Source Data.

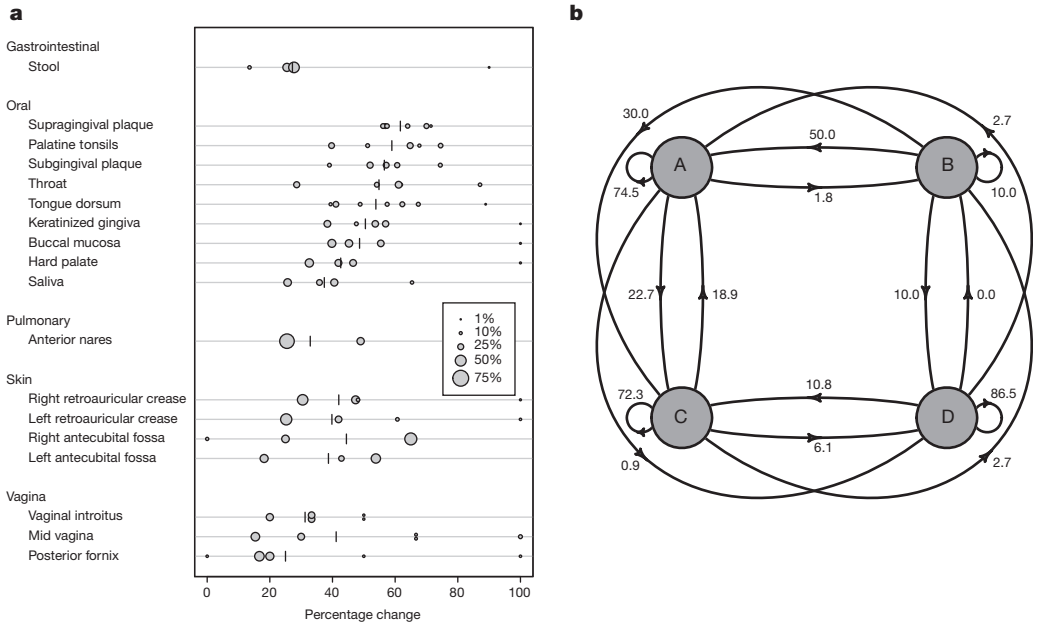


Figure 3 | Dynamics of community types at various body sites indicates that community type stability is correlated with the diversity of the community type. **a**, The community types at each body site differ in the fraction of samples that change their community type membership between visits. (Size of circles represents percentage of samples that affiliated with each community type and the vertical line represents the weighted average.) **b**, Rate of change between stool community types (n (community type A) = 221; n (community type B) = 15; n (community type C) = 80; n (community type D) = 281). The numbers on directed edges indicate the percentage of samples that changed community types.

the factors that affect one's community type profile will be critical as they continue to be associated with predisposition to diseases. Furthermore, understanding why community types change will be useful in developing therapies that can alter one's community type using pre- and probiotics, faecal transplants, or antibiotics. Given the varying levels of flux between community types at different body sites, it is remarkable that we were still able to detect life-long legacy effects on the microbiome, such as whether the subject was ever breastfed as an infant. This result could represent a true long-term impact of breastfeeding on the microbiome or it could represent the effect of the individual's childhood environment or care. The result raises the possibility that there may be other legacy effects on the microbiome, such as duration of breastfeeding, mode of birth, level of early antibiotic exposure, and childhood disease[18–20]. The four gender-based associations are intriguing and support previous studies showing that men and women have different skin communities[21] and that autoimmune diseases may be mediated via the microbiome and hormonal differences[22]. The association between one's level of education and their vaginal microbiome type is less clear; it is most likely that a baccalaureate degree represents a composite variable of numerous factors known to affect the vaginal microbiome, including race/ethnicity, sexual behaviour and socioeconomic class. Regardless, that such considerable variation was observed among a population of healthy women supports the observation that there is no single normal vaginal microbiome[23]; this is probably true for every body site. Looking forward, prospective studies that include individuals with varied levels of health and varied backgrounds (study groups that are more representative of society) are needed to achieve a better understanding of the mechanisms of change in community types as well as to provide more details about correlations between community type and life-history factors such as genetics, age, diet, health status, and environment (that is, rural or urban). Furthermore, future prospective studies with a longitudinal component need to control for the time between samplings and perhaps synchronize sampling with host physiology (for example, menses). Perhaps most exciting is the prospect that community types may be associated with complex diseases such as bacterial vaginosis, periodontitis, cancer, and diabetes where it has not been possible to establish a causative relationship between a specific bacterium and the disease.

METHODS SUMMARY

The Human Microbiome Project carried out three phases of sequencing the 16S rRNA gene and we obtained the unprocessed data for the V35 region from the NCBI Short Read Archive (SRA): the Clinical Pilot Project (accession SRP002012), Phase I (accession SRP002395) and Phase II (accession SRP002860). The Clinical Pilot Project and Phase I data sets have been described previously[2,3]. The metadata and clinical data associated with the samples from the subjects were obtained from dbGap (accession phs000228.v3.p1). The 16S rRNA gene sequence curation pipeline was implemented using the mothur software package (http://nbviewer.ipython.org/gist/pschloss/9815766/notebook.ipynb)[24,25].

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 September 2013; accepted 20 February 2014.
Published online 16 April 2014.

1. Peterson, J. *et al.* The NIH Human Microbiome Project. *Genome Res.* **19,** 2317–2323 (2009).
2. The Human Microbiome Consortium. A framework for human microbiome research. *Nature* **486,** 215–221 (2012).
3. The Human Microbiome Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214 (2012).
4. Holmes, I., Harris, K. & Quince, C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS ONE* **7,** e30126 (2012).
5. Aagaard, K. *et al.* The Human Microbiome Project strategy for comprehensive sampling of the human microbiome and why it matters. *FASEB J.* **27,** 1012–1022 (2013).
6. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–484 (2009).
7. Costello, E. K. *et al.* Bacterial community variation in human body habitats across space and time. *Science* **326,** 1694–1697 (2009).
8. Koren, O. *et al.* A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLOS Comput. Biol.* **9,** e1002863 (2013).
9. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473,** 174–180 (2011).
10. Wu, G. D. *et al.* Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334,** 105–108 (2011).
11. Quince, C. *et al.* The impact of Crohn's disease genes on healthy human gut microbiota: a pilot study. *Gut* **62,** 952–954 (2013).
12. Brotman, R. M. *et al.* Association between *Trichomonas vaginalis* and vaginal bacterial community composition among reproductive-age women. *Sex. Transm. Dis.* **39,** 807–812 (2012).
13. Gajer, P. *et al.* Temporal dynamics of the human vaginal microbiota. *Sci. Transl. Med.* **4,** 132ra152 (2012).

14. Ravel, J. *et al.* Vaginal microbiome of reproductive-age women. *Proc. Natl Acad. Sci. USA* **108** (suppl. 1), 4680–4687 (2011).
15. Statnikov, A. *et al.* Microbiomic signatures of psoriasis: feasibility and methodology comparison. *Sci. Rep.* **3,** 2620 (2013).
16. Arumugam, M. *et al.* Addendum: Enterotypes of the human gut microbiome. *Nature* **506,** 516 (2014).
17. Moeller, A. H. *et al.* Chimpanzees and humans harbour compositionally similar gut enterotypes. *Nature Commun.* **3,** 1179 (2012).
18. Palmer, C., Bik, E. M., Digiulio, D. B., Relman, D. A. & Brown, P. O. Development of the human infant intestinal microbiota. *PLoS Biol.* **5,** e177 (2007).
19. Koenig, J. E. *et al.* Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl Acad. Sci. USA* **108** (suppl. 1), 4578–4585 (2011).
20. Pantoja-Feliciano, I. G. *et al.* Biphasic assembly of the murine intestinal microbiota during early development. *ISME J.* **7,** 1112–1115 (2013).
21. Fierer, N., Hamady, M., Lauber, C. L. & Knight, R. The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proc. Natl Acad. Sci. USA* **105,** 17994–17999 (2008).
22. Markle, J. G. *et al.* Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity. *Science* **339,** 1084–1088 (2013).
23. Ma, B., Forney, L. J. & Ravel, J. Vaginal microbiome: rethinking health and disease. *Annu. Rev. Microbiol.* **66,** 371–389 (2012).
24. Schloss, P. D. *et al.* Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75,** 7537–7541 (2009).
25. Schloss, P. D., Gevers, D. & Westcott, S. L. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE* **6,** e27310 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Crabtree of the HMP Data Analysis and Coordination Center for his assistance in obtaining the sequencing and metadata files. The analysis described in this study was supported by grants from the National Institutes of Health (R01HG005975, R01GM099514 and P30DK034933).

Author Contributions T.D. and P.D.S. designed and executed the analysis and prepared the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.D.S. (pschloss@umich.edu).

METHODS

Sequence analysis pipeline. The Human Microbiome Project carried out three phases of sequencing the 16S rRNA gene, which were performed using the 454 Titanium sequencing platform. We obtained the unprocessed sff files for the V35 region from the NCBI Short Read Archive (SRA) for each of these phases: the Clinical Pilot Project (accession SRP002012), Phase I (accession SRP002395) and Phase II (accession SRP002860). The Clinical Pilot Project and Phase I data sets have been described previously[2,3]. The sequencing was performed by sequencing from the 3′ to the 5′ end of the 16S rRNA gene[26]. Although the V13 and V69 regions were also sequenced by the HMP sequencing centres, the number of data sets generated for those regions was considerably smaller than was obtained for the V35 region. The 16S rRNA gene sequence curation pipeline was implemented using the mothur software package[24,25]. This approach has been shown to result in a sequencing error rate of 0.02%[25]. Briefly, flowgrams were extracted from the sff files and any that had more than one mismatch to the barcode, more than two mismatches to the primer, had fewer than 450 flows, contained homopolymers longer than 8 nucleotides, or contained an ambiguous base call were culled. The flows for each sequencing run were trimmed to 450 flows and de-noised separately using the PyroNoise algorithm as implemented within mothur[27]. The de-noised sequences were then aligned against a customized reference alignment based on the SILVA database using the NAST algorithm implemented within mothur[28]. The customized database included small subunit rRNA sequences from bacteria, archaea, eukarya, chloroplasts and mitochondria. Sequences that did not align to the predicted V35 region were culled from further analysis and the alignments were trimmed so that the sequences fully overlapped the same alignment coordinates[29,30]. These sequences were then subjected to a pre-clustering step that first sorted the sequences by their abundance within each sample and then clustered sequence abundances together if a sequence was within 2 nucleotides of a more abundant sequence[25]. Treating each sample separately, we interrogated each sequence for the presence of chimaeras using the *de novo* UChime chimaera detection algorithm[31]. Once chimaeric sequences were culled from the data sets, the sequences were classified using the naive Bayesian Classifier trained against a customized version of the RDP training set (version 9) as implemented within mothur[32]. The training set was customized by supplementing sequences derived from chloroplasts, mitochondria and members of the Eukarya. The reference sequences were trimmed to only include the V35 region of the 16S rRNA gene. We required a minimum classification confidence score of 80% and used 1,000 pseudo-bootstrap iterations. Because the PCR target was bacterial 16S rRNA gene sequences, we culled any sequences that classified as being derived from archaea, eukarya, mitochondria, chloroplasts or sequences that could not be classified to a kingdom with at least 80% confidence. The taxonomy of the remaining sequences was used to assign the sequences to genus-level phylotypes. Those sequences without a genus-level classification were assigned to a phylotype represented by the lowest level taxonomy with a confidence score of at least 80%. This allowed us to create a table of counts for the number of times each genus-level phylotype was observed in each sample. As some samples were sequenced multiple times to obtain additional sequence data, we pooled replicate sequencing runs to create a single sample. Samples with fewer than 1,000 reads were removed from further analysis and all samples were either sub-sampled or rarefied ($n = 1{,}000$ iterations) to 1,000 reads to perform subsequent analyses. Sub-sampling and rarefaction were necessary to limit the effects of differential sampling that are known to affect alpha and beta diversity metrics and differentially increase the representation of PCR and sequencing artefacts in data sets.

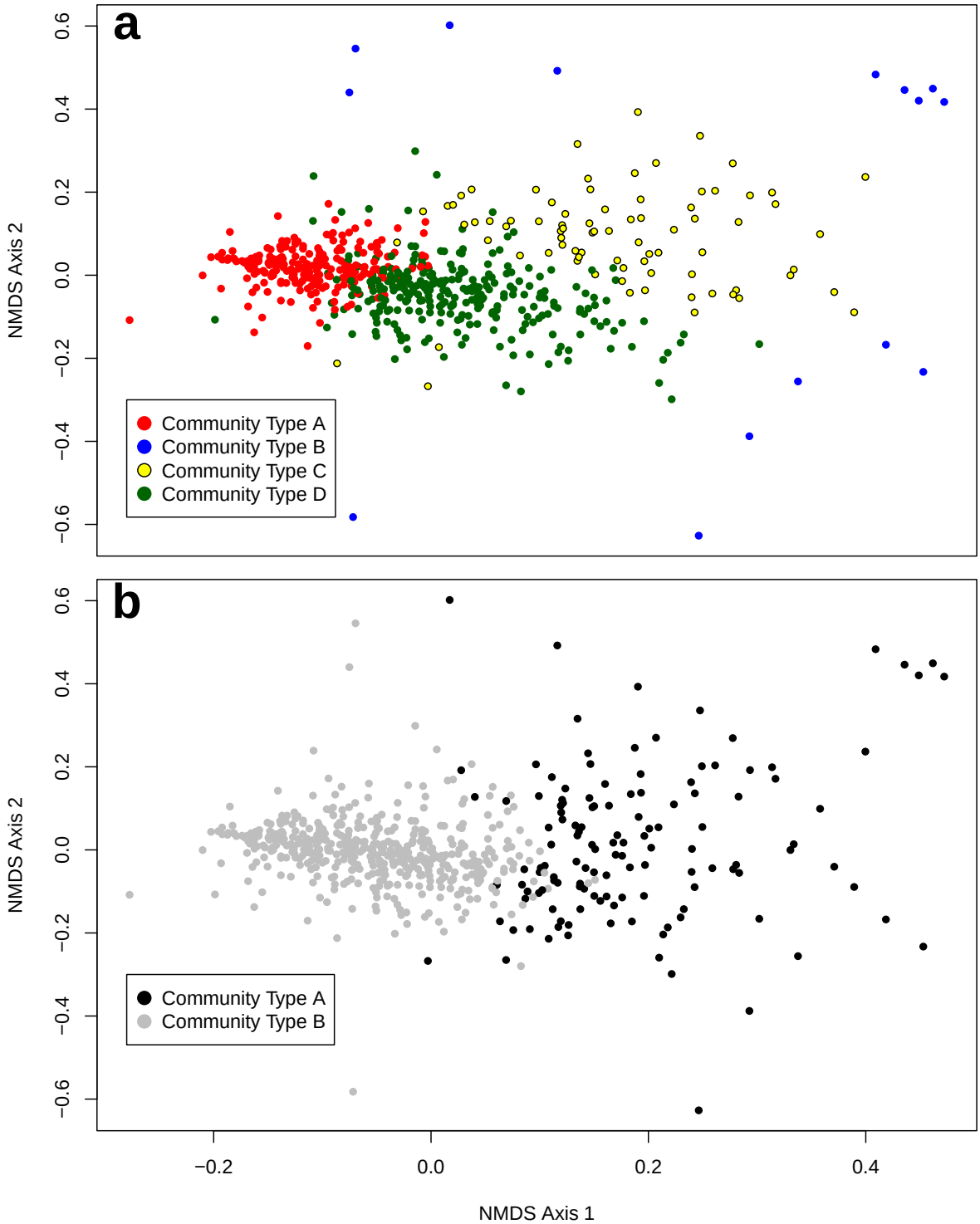
Assignment to community types. The table of counts was partitioned according to the 18 body sites. Because the communities were similar and we wanted to use the maximum number of samples per body site when assigning samples to community types, we pooled the three vaginal body sites (that is, vaginal introitus, mid-vagina, and posterior fornix) into one vaginal data set, the two antecubital fossa sites (that is, left and right) into one antecubital fossa data set, and the two retroauricular crease sites (that is, left and right) into one retroauricular crease data set. The resulting 14 tables were used as input to partition the samples according to community types at each body site using the Dirichlet multinomial mixture model[4]. We selected the number of community types at each body site by selecting the number of components that gave the minimum Laplace approximation to the negative log model evidence. Samples were assigned to their community type based on the maximum posterior probability. For all body sites, between 89.2% and 99.7% of the samples had a posterior probability of at least 0.90. The mean abundance and 95% confidence interval predicted by the model are provided for each body site.

Selection of metadata. A large amount of metadata and clinical data were collected for each of the samples and subjects[5]. We obtained the most recent version of these data from dbGap (accession phs000228.v3.p1). Because of the uniformity and healthy nature of the cohort, a number of the clinical data fields could not be included in our analysis. Furthermore, there was evidence that several variables were collected from subjects in one city but not the other (see Supplementary Information for a discussion of the difficulties in analysing the city of origin data). We interrogated, a priori, the categorical metadata to identify those variables where we were able to identify at least 10 instances of the condition and that was represented in the subjects from both cities. In addition, to increase the number of variables under consideration, we pooled responses. For example, there were 13 categories for country of birth with only one of those having more than 10 respondents (US/Canada; $n = 260$). In this case we pooled the other responses to create a non-US/Canada group ($n = 40$). We used a similar pooling strategy for parents' country of origin, meat eaters/vegetarians, number of children the subject had given birth to, occupation, and level of education. The data available through dbGap partitioned medications into broad categories and indicated whether the subject was using the medication at the time of the visit. This created three classes of subjects. The first never used the medication during the study, the second class used the medication for one or two of the visits, but not all of their samples, and the third class used the medication for all of their visits. For the purpose of correlating medication usage with community type, we used the data for the first and third classes of subjects and ignored the second. The number of subjects in the second class was below 10 for each type of medication. For example, there were only 8 subjects that had more than one visit and used antibiotics within 30 days of any of their visits. Because of the general paucity of subjects in this category of medication users, it was not possible to associate medication usage with changes in community type. Finally, we converted the subjects' body mass index (BMI) into categories of normal (18.5–25), overweight (25–30) and obese (>30); there were no underweight subjects (<18.5). The resulting list of categorical clinical data that were considered is provided in Extended Data Table 1. In addition to these categorical data, we also had access to continuous clinical metadata for each of the subjects. These included their age, BMI, pulse and blood pressure. A summary of these data is provided in Extended Data Table 1.

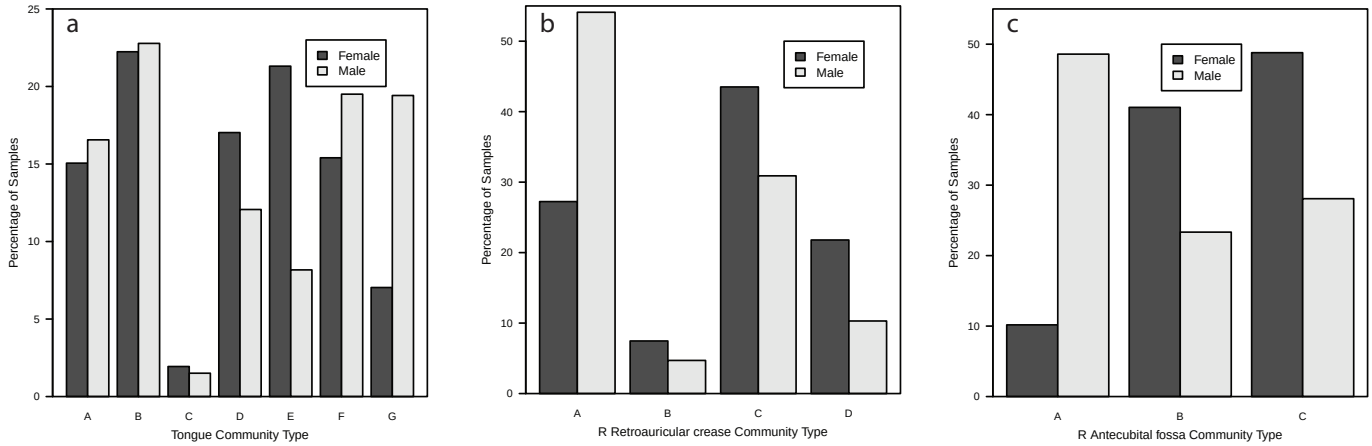
Tests of association. Because individuals provided up to three samples it would have been arbitrary to select one visit from each subject (for example, the first visit) on which to base our analyses. For the categorical metadata associations, we performed an iterative procedure where we selected a single visit for each individual's body site and tested the association between the community type at the body site and the metadata using Fisher's exact test. For the continuous metadata we performed a similar procedure except we tested the association between the community type at the body site and the metadata using analysis of variance. Finally, to test associations across the body we performed a randomization procedure where each iteration consisted of selecting one visit for each individual and then testing for inter-body site associations using the Fisher exact test. We performed 1,000 iterations and we calculated the percentage of iterations that each variable was significant according to the Benjamini–Hochberg step-up procedure that we used to limit the false discovery rate to 5%. We report the median P value and the percentage of iterations that were significant.

A complete description of the analysis pipeline including scripts, mothur commands, and intermediate files are available at http://nbviewer.ipython.org/gist/pschloss/9815766/notebook.ipynb.

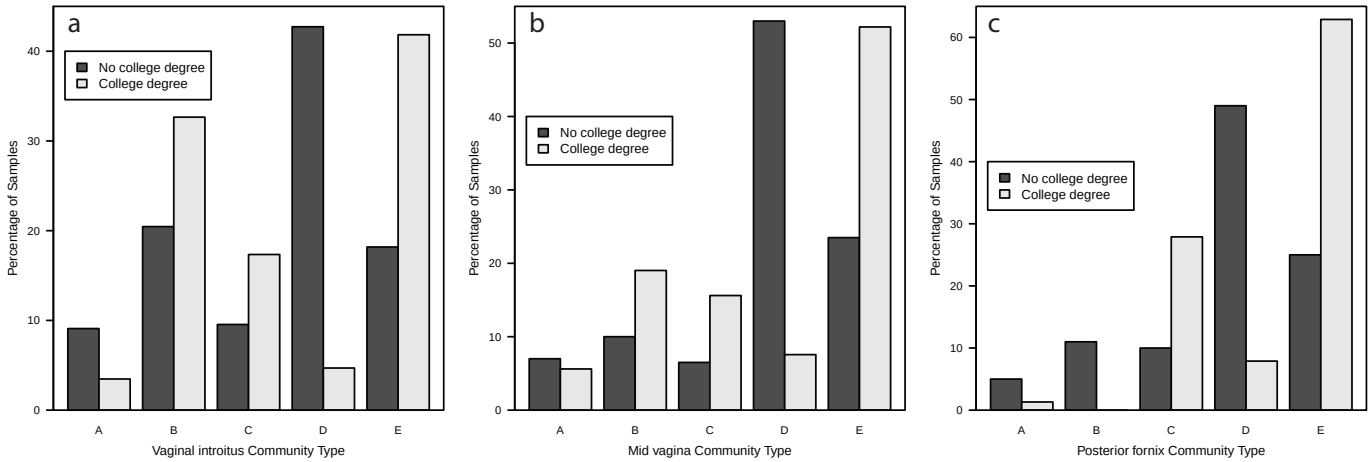
26. Jumpstart Consortium Human Microbiome Project Data Generation Working Group. Evaluation of 16S rDNA-based community profiling for human microbiome research. *PLoS ONE* **7,** e39315 (2012).
27. Quince, C., Lanzen, A., Davenport, R. J. & Turnbaugh, P. J. Removing noise from pyrosequenced amplicons. *BMC Bioinform.* **12,** 38 (2011).
28. Schloss, P. D. A high-throughput DNA sequence aligner for microbial ecology studies. *PLoS ONE* **4,** e8230 (2009).
29. Schloss, P. D. Secondary structure improves OTU assignments of 16S rRNA gene sequences. *ISME J.* **7,** 457–460 (2013).
30. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35,** 7188–7196 (2007).
31. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27,** 2194–2200 (2011).
32. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73,** 5261–5267 (2007).



Extended Data Figure 1 | Comparison of community type assignments for non-metric dimensional scaling (NMDS) ordination of Jensen–Shannon divergence values between stool samples using DMM (a) and PAM-based clustering (b). The stress computed for this ordination was 0.19 and the R^2 between the input distance matrix and the distance matrix calculated between the points in the ordination was 0.90.



Extended Data Figure 2 | The frequency of community types for body sites where there was a significant association with the subject's gender. **a**–**c**, Percentage of female and male tongue communities that affiliated with each of the tongue (**a**; $n = 288$ unique individuals; median $P = 2 \times 10^{-3}$), right retroauricular crease (**b**; $n = 268$ unique individuals; median $P = 9 \times 10^{-5}$) and right antecubital fossa community types (**c**; $n = 136$ unique individuals; median $P = 3 \times 10^{-5}$).



Extended Data Figure 3 | The frequency of vaginal community types among women with and without a college degree. **a**–**c**, Percentage of women with and without a college degree whose vaginal communities affiliated with the vaginal introitus (**a**; $n = 74$ unique individuals; median $P = 2 \times 10^{-3}$), mid-vagina (**b**; $n = 64$ unique individuals; median $P = 8 \times 10^{-4}$) and posterior fornix (**c**; $n = 61$ unique individuals; median $P = 4 \times 10^{-4}$) community types.

Extended Data Table 1 | Most common characteristics of the individuals included in the HMP healthy cohort

Categorical data	Number of Individuals (Total=300)
Sampled in Houston / St. Louis	150 / 150
Female / Male	151 / 149
Born in US or Canada	260
Mother born in US or Canada	227
Father born in US or Canada	229
Hispanic, Latino, or Spanish	32
Asian	34
Black	21
White	243
Ever breastfed as infant	198 (Forgot=33, NA=23)
Eats meat at least once a week	261 (NA=23)
Occupation: Student	150
Had given birth at least once	26 (NA=156)
College educated	210 (NA=16)
Had dental insurance	265 (NA=16)
Had health insurance	217 (NA=16)
Tobacco user	19
Chronic use of antidepressants	22 (Tr=9)
Chronic use of antihistamines	9 (Tr=13)
Chronic use of hormonal contraceptives	72 (Tr=15; NA=149)
Chronic use of vitamins or supplements	34 (Tr=10)
Normal BMI	178
Overweight BMI	88
Obese BMI	34
Continuous data	**Median (Min-Max)**
Age	25 (18-40)
BMI	24 (19-34)
Pulse	71 (42-100)
Diastolic pressure	71 (50-98)
Systolic pressure	119 (91-151)
pH – Vaginal introitus	4.4 (3.3-6.5)
pH – Posterior fornix	4.0 (3.2-7.0)

NA, data was not collected; Forgot, the subject could not recall; Tr, the medication was used transiently throughout the two or three sampling events.

Extended Data Table 2 | Comparison of PAM- and DMM-based approaches to assigning samples to community types

	PAM-based using SI Index			PAM-based using CH Index			DMM-based	
Body site	**Clusters**	**SI Index**	**Laplace**	**Clusters**	**CH Index**	**Laplace**	**Clusters**	**Laplace**
Antecubital fossa	2	0.34	84858.4	2	114.3	84858.4	3	83302.1
Anterior nares	3	0.32	52136.3	2	153.5	51864.3	2	51532.0
Buccal mucosa	2	0.23	65643.3	4	166.2	64968.1	4	64588.8
Hard palate	2	0.28	72686.9	4	208.4	71573.8	4	71436.9
Keratinized gingiva	2	0.38	51392.2	2	323.8	51392.2	5	50605.3
Palatine tonsils	2	0.27	82655.3	2	237.7	82655.3	6	81446.7
Retroauricular crease	3	0.51	95797.5	3	719.9	95797.5	5	94673.5
Saliva	2	0.19	81261.3	2	120.4	81261.3	4	80656.1
Stool	2	0.40	76228.5	2	194.0	76228.5	4	74785.6
Subgingival plaque	2	0.25	90876.7	2	249.4	90876.7	5	89672.2
Supragingival plaque	2	0.24	78982.0	2	217.6	78982.0	6	78357.1
Throat	2	0.26	79238.0	2	177.8	79238.0	5	78052.8
Tongue dorsum	2	0.33	71442.3	2	293.2	71442.3	7	69923.0
Vagina	10	0.57	32407.3	2	205.7	32150.8	5	31209.5

Extended Data Table 3 | Average contingency table of stool and saliva community types

	Saliva A	Saliva B	Saliva C	Saliva D
Stool A	0.101	0.140	0.104	0.044
Stool B	0.003	0.000	0.002	0.021
Stool C	0.136	0.173	0.107	0.027
Stool D	0.048	0.024	0.052	0.017

The median *P* value from a Fisher's exact test was 1×10^{-3}.