Mobile genes in the human microbiome are structured from global to individual scales

I. L. Brito1,2, S. Yilmaz3,8, K. Huang28, L. Xu21, S. D. Jupiter4, A. P. Jenkins5, W. Naisalisili4, M. Tamminen6,7, C. S. Smillie1, J. R. Wortman2, B. W. Birren2, R. J. Xavier2,8,9, P. C. Blainey3, A. K. Singh3, D. Gevers2 & E. J. Alm1,2,9

Recent work has underscored the importance of the microbiome in human health, and has largely attributed differences in phenotype to differences in the species present among individuals1–5. However, mobile genes can confer profoundly different phenotypes on different strains of the same species. Little is known about the function and distribution of mobile genes in the human microbiome, and in particular whether the gene pool is globally homogenous or constrained by human population structure. Here, we investigate this question by comparing the mobile genes found in the microbiomes of 81 metropolitan North Americans with those of 172 agrarian Fiji islanders using a combination of single-cell genomics and metagenomics. We find large differences in mobile gene content between the Fijian and North American microbiomes, with functional variation that mirrors known dietary differences such as the excess of plant-based starch degradation genes found in Fijian individuals. Notably, we also observed differences between the mobile gene pools of neighbouring Fijian villages, even though microbiome composition across villages is similar. Finally, we observe high rates of recombination leading to individual-specific mobile elements, suggesting that the abundance of some genes may reflect environmental selection rather than dispersal limitation. Together, these data support the hypothesis that human activities and behaviours provide selective pressures that shape mobile gene pools, and that acquisition of mobile genes is important for colonizing specific human populations.

Bacteria rapidly evolve and adapt to changing environments by acquiring new genes from other bacteria in their environments. Comparison of reference genomes from the microbiomes of individuals living in different countries around the world has shown that geography does not present a substantial barrier to gene flow6. Moreover, there are specific examples of genes that have swept through global populations, such as the antibiotic resistance gene New Delhi metallo-beta-lactamase 1 (ref. 7). These observations raise the possibility that the mobile gene pools of neighbouring Fijian villages, even though microbiome composition across villages is similar. Finally, we observe high rates of recombination leading to individual-specific mobile elements, suggesting that the abundance of some genes may reflect environmental selection rather than dispersal limitation. Together, these data support the hypothesis that human activities and behaviours provide selective pressures that shape mobile gene pools, and that acquisition of mobile genes is important for colonizing specific human populations.

Here, we combine single-cell genomics with metagenomics to survey the mobile gene pool and investigate whether mobile genes are mainly globally distributed or population-specific. We compared the mobile gene pools of 81 participants in the US-based Human Microbiome Project (HMP)16 (Supplementary Table 1) with those of 172 Fijian islanders participating in the Fiji Community Microbiome Project (FijiCOMP) (Supplementary Table 2). FijiCOMP represents the first terabase-scale metagenomic view of the microbiome in the developing world.

Cataloguing the mobile gene pool in a large cohort using short-read metagenomic sequences is difficult, so previous analyses have been constrained to individual species17 or specific mobile elements such as plasmids18–20 and phages21,22. One reliable method for cataloguing mobile genes, including integrated transposons and prophages, depends on assembled genomes and is based on identifying identical or nearly identical genes present in distantly related bacterial hosts6. We used this method to identify 15,585 mobile genes from the 387 HMP gut microbiome reference and draft genomes23 (Supplementary Tables 3, 5). Similar to complete genomes, even draft single-cell genomes provide enough context to link genes to hosts and to identify mobile genes. We used 180 single-cell genomes, derived from seven FijiCOMP participants, to identify an additional 22,268 mobile genes (Supplementary Tables 4, 5 and Extended Data Fig. 1). We then investigated which metagenomic libraries contained reads that mapped to this set of mobile genes to survey their abundances across individuals and populations (Extended Data Fig. 2).

Most mobile genes (62.4%) were present to some extent in both study populations, consistent with the previous finding that horizontal gene transfer (HGT) is not strongly structured by geography. Surprisingly, however, the abundance of those genes across populations was noticeably distinct (Extended Data Fig. 3a; P < 10−6, permutational multivariate analysis of variance (PERMANOVA) based on Bray–Curtis dissimilarity,106 permutations), an observation that holds true even when considering a subset of the most stringently defined mobile genes (Extended Data Fig. 3b; P < 10−6, PERMANOVA based on Bray–Curtis dissimilarity,106 permutations). Surveys of organisms endogenous to each population are required to fully observe both mobile gene pools (Extended Data Fig. 3c). We reasoned that diet might be a strong factor in differentiating the mobile gene pool across populations because many of the most highly consumed Fijian food items (taro, cassava breadfruit, coconut, and certain seafood) are not widely consumed in the US (see Methods).

We specifically examined genes encoding glycoside hydrolases as these include substrate-specific dietary enzymes. In the Fijian microbiomes, we found a high abundance of mobile genes from the glycoside hydrolase GH13 family (Fig. 1a, Supplementary Table 6 and Extended Data Fig. 4), which encompasses most of the dietary starch.
across populations, reflecting the high usage of beta-lactams and difficult to link directly to environmental factors such as diet. Specific genes rather than species-level effects that might be more advantageous of looking at mobile genes is that it narrows our focus to we were unable to determine what fraction of these are mobile. The FijiCOMP cohort versus the HMP cohort (see Methods), although there are likely to be additional in previous cross-cultural comparisons, may be due in part to differencies at a global scale, we hypothesized that gene distributions varied abundance of plant matter (read alignments to rRNA from the kingdom Viridiplantae) in the metagenomes of stool samples from the FijiCOMP (red) and HMP (blue) populations. c, Prevalence and abundance of annotated mobile antibiotic resistance genes across the FijiCOMP (red) and HMP (blue) metagenomic stool samples. P value based on Mann–Whitney test. d, Prevalence and abundance of eight village-specific mobile genes (of 31 total village-specific genes) across four Fijian villages. q values (a, b, d) of prevalence comparisons based on FDR-corrected Fisher’s exact tests; q values of abundance comparisons based on FDR-corrected Mann–Whitney tests.

**Figure 1** Enrichment of functional mobile genes is locale-specific. a, Prevalence and abundance of all of the annotated mobile glycoside hydrolase families present in the FijiCOMP (n = 172, red) and HMP (n = 81, blue) metagenomic stool samples. Abundances were measured in fragments per kilobase of protein coding sequence per million mapped reads (FPKM) for each of the horizontally transferred genes, aggregated according to glycoside hydrolase family and plotted as a function of the density across samples. For each glycoside hydrolase family, the number of unique horizontally transferred genes present across the two cohorts is plotted (in grey), as are the sources of their substrates. b, Prevalence and degradation enzymes. By contrast, glycoside hydrolases that degrade animal- and fungus-derived glycans exhibited bimodality between the American and Fijian cohorts, which may represent different dependencies on farm animals versus seafood that predominate in each of these cultures’ diets. We confirmed the dietary differences between the two cohorts directly by metagenomic sequencing of stool samples, as the metagenomes of Fijians harboured significantly higher levels of plant matter (P < 10^{-15}, Mann–Whitney test; Fig. 1b). Thus, differences in diet-related genes in the microbiome, which have also been observed in previous cross-cultural comparisons, may be due in part to differential abundance of mobile genes, rather than to the presence of genes that can be attributed to specific taxa. There are likely to be additional non-mobile glycoside hydrolase genes that differ across populations. GH13 family genes were indeed enriched in the overall microbiomes of the FijiCOMP cohort versus the HMP cohort (see Methods), although we were unable to determine what fraction of these are mobile. The advantage of looking at mobile genes is that it narrows our focus to specific genes rather than species-level effects that might be more difficult to link directly to environmental factors such as diet.

We next hypothesized that antibiotic resistance genes might differ across populations, reflecting the high usage of beta-lactams and abundance of quinolones in Fiji (see Methods). We found that quinolone resistance was more pervasive in the Fijian cohort than the American cohort (Fig. 1c and Supplementary Table 7). Resistance to cephalosporins, which have been introduced relatively recently, was primarily an American phenomenon. Despite limited access to the full range antibiotics in Fiji, resistance genes for most classes of antibiotics were found in the Fijian population, consistent with recent findings in other developing world communities. These results highlight our limited understanding of the forces that drive antibiotic resistance, and could reflect other uses of antibiotics, such as their use in agriculture, or even point to new classes of antibiotic resistance that have not yet been characterized.

As the distribution of mobile genes in the microbiome shows differences at a global scale, we hypothesized that gene distributions varied at even finer scales of resolution. The FijiCOMP cohort provides a unique opportunity to look at fine-scale structure because it includes individuals living in three villages in the remote Bua province and a fourth village in the agricultural Macuata province. Therefore, we checked whether any genes were specifically associated with one or more villages. Notably, we identified many village-associated mobile genes (31 genes varied significantly according to abundance, based on FDR-corrected Mann–Whitney test). Values based on FDR-corrected Fisher’s exact test: q values of abundance comparisons based on FDR-corrected Mann–Whitney tests.© 2016 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
or prevalence, based on a Fisher’s exact test of village association using FDR with $q < 0.05$ (Fig. 1d and Supplementary Table 8). Village-associated mobile genes include those encoding HGT machinery, such as proteins involved in plasmid segregation, as well as proteins that may directly provide a fitness advantage to their host, such as ferredoxins.

The identification of village-associated mobile genes was unexpected because individuals’ microbiomes are more similar among villages ($P > 0.05$, PERMANOVA based on Jensen-Shannon divergence, $10^5$ permutations) than countries ($P < 10^{-6}$, PERMANOVA based on Jensen-Shannon divergence, $10^6$ permutations) (Fig. 2, Extended Data Fig. 5 and Supplementary Table 9). Even though the mobile genes in our data set are defined by their occurrence in multiple phylogenetic backgrounds, gene exchange is known to be more frequent among closely related species. We therefore investigated whether the village-associated genes were associated with species that were also partitioned unevenly across villages. For six of the 31 village-associated genes, we found a correlation with one of the ten village-associated species, although five of these six were also correlated with other non-village-associated species (see Methods). Overall, our findings suggest that these mobile genes are not restrained by their host genomes.

This finding prompted us to investigate the extent to which species composition determines mobile gene pool composition. First, we tested whether mobile genes originally identified in Bacteroides genomes or Prevotella genomes were more common in the metagenomes of individuals whose microbiomes were dominated by the same species. However, we did not find a significant association between the dominant taxon in an individual and their likelihood of carrying mobile genes of similar origins. Next, we investigated the taxonomy of bacteria hosting specific mobile elements by examining paired reads

Figure 3 | Personal mobile genetic element architecture displays high variation due to recombination. a–c, Examples showing comparisons of assembled mobile genetic elements between the microbiomes of individuals from different continents (a, b) or different villages (c). Gene linkages between mobile genes are coloured according to the broad COG category. d, For each mobile gene end, the median and quartile proportions of neighbours (as determined by the proportion of metagenomic read pairs) is plotted according to whether the adjacent gene is in broad functional concordance (detected by COG category) and whether they are situated on the same DNA strand, denoting whether they are likely to comprise the same operon. e, The average number of gene families connected to mobile genes of each type of functional category, as determined by paired reads
that span junctions between mobile genes integrated next to evolutionarily conserved tRNAs. Out of 838 mobile genes found adjacent to tRNAs, 65 were found near tRNAs from more than one genera, with the most promiscuous occurring next to tRNAs from 25 different genera (Supplementary Table 10; examples shown in Extended Data Fig. 6). Even within a single individual, these genes were not tied to the dominant bacteria present within that individual’s microbiome, and often spanned multiple genera, including those in both Prevotellaceae and Bacteroidaceae families. Thus, horizontal gene transfer is likely to be driven by forces independent of species community composition.

The high rates of recombination observed among mobile genes support the idea that mobile genes are even less dispersal-limited than their host genomes. Following the logic of the Baas-Becking hypothesis that “everything is everywhere, but the environment selects,” we might infer that environmental selection rather than dispersal drives differences in gene abundance among populations. Extending this logic from species to genes, however, is not straightforward because genes can change in abundance as a result of genetic ‘hitchhiking’ on mobile elements under selection at a different locus. We reasoned that if recombination rates are sufficiently high, then genes should be present in many contexts. On the other hand, if gene abundances are driven by selection on larger mobile elements, then genes should appear in a limited number of contexts.

To investigate the genomic contexts of our mobile genes, we used the alignment and orientation of paired-end metagenomic reads to assemble short contigs encompassing each mobile gene (Fig. 3a–c). As expected, we found that recombination was limited, although not absent, within operons (as defined by adjacent genes in the same orientation and of the same functional category; Fig. 3d), and that horizontal transfer machinery, such as phage-, plasmid- or transposon-specific genes, had higher levels of recombination (Fig. 3e). However, despite the high prevalence of genes and broader gene functions, we found that 34.9% of the gene contexts, defined as the set of unique combinations between adjacent genes that were observed, were specific to individuals; very few of these gene contexts were conserved across populations (Fig. 4 and Extended Data Fig. 7).

Together, high recombination rates and population-specificity support the notion that environmental selection on individual genes, rather than dispersal limitation alone, plays a key role in driving gene abundances. The ‘everything is everywhere’ concept is reinforced for mobile genes by the observation that the majority of genes were found in both the HMP and FijiCOMP populations (Fig. 4). Even among the village-associated genes, all but five were found in all of the villages. Dispersal may play a role in shaping the distribution of a small, but abundant, subset of genes that are restricted to one population (that is, gene families (GH67, GH28 and GH110, Fig. 1a) or subsets of villages (Fig. 1e)). Nevertheless, environmental selection is underscored by the impact of diet on gene abundance. In fact, only one of the universal genes, defined as those present in more than 75% of both populations, was annotated as carbohydrate-metabolizing, despite carbohydrate-metabolizing genes being significantly enriched among population-specific genes, defined as those present in more than 50% of one population and less than 10% of the other ($P < 10^{-5}$, Pearson’s $\chi^2$ test).

How selection and dispersal affect gene exchange within physically proximal environments, that is, within a single individual’s body sites, is still an open question. On one hand, differences in composition and ecology may result in distinct mobile gene pools at various body sites. On the other hand, the direct route between oral and gut communities should facilitate transmission of mobile elements, and systemic selective pressures, such as those imposed by orally administered antibiotics, may homogenize personal gene pools. Our analysis of saliva samples derived from FijiCOMP participants shows that there is little overlap between gut and oral mobile gene pools, as only 0.94% of the genes represented in the gut were detectable in any of the saliva samples. For any particular gene, there was no correlation between its presence or abundance between stool and saliva samples from the same individual. These data support the hypothesis that shared selective pressures and common ecologies structure horizontal transfer, although they do not rule out the possibility that physical proximity plays a role.

As well as presenting a data set describing the microbiome of a population in the developing world, we have also described an approach that can be used to identify environmentally relevant genes. Previous shotgun metagenome approaches have focused on the abundance of a gene as a proxy for its importance. This approach will identify important genes but also has the potential to identify many spurious genes because a single highly abundant species can carry many genes that are not specifically relevant to the environment it occupies. Instead, we look for abundant genes that are present in multiple species, using horizontal gene transfer as an additional filter for gene importance. Our approach is subject to several important caveats. First, its sensitivity for detecting mobile genes is low, and there are likely to be many more genes transferred across species than we can detect. Second, even though each of the mobile genes within our data set exists in more than one species, some of those genes may be primarily associated with a single taxon. This is especially true if a single taxon is much more abundant than the other species that carry the gene.

Despite these caveats, we found that the human-associated mobile gene pool differed between populations and carried gene functions that are likely to be associated with cultural practices; these findings provide functional insights not possible based on surveys of phylogenetic markers, such as the 16S rRNA gene, alone. These insights have the potential to improve public health at multiple levels. For example, a better understanding of the distribution of antibiotic-resistance genes among different populations’ microbiomes could inform antibiotic stewardship at a regional level, by avoiding the use of specific antibiotics where resistance to them is highest. We also showed that the mobile genetic elements are highly diverse among individuals, raising the possibility that they may vary within bacterial lineages in an individual over short time spans. If this is true, then diet and drugs could modify the functions of the microbiome, even if the long-term species composition is stable. To assess whether changes to cultural practices will influence human health via mobile genetic elements, future studies should test the speed at which selective pressures alter mobile gene frequencies within single individuals and larger populations.

Online Content Methods, along with any additional 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 16 February 2015; accepted 15 June 2016. Published online 13 July 2016.
1. Turnbaugh, P. J. et al. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444, 1027–1031 (2006).
2. Giongo, A. et al. Toward defining the autoimmune microbiome for type 1 diabetes. ISME J. 5, 82–91 (2011).
3. Qin, J. et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490, 55–60 (2012).
4. Scher, J. U. et al. Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. MBio 5, e01305–e01314 (2014).

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

Acknowledgements We thank our field collaborators in the Fiji Islands: the Wildlife Conservation Society, Fiji; Wetlands International-Oceania, K. Jenkins, S. Korovou, N. L tidamau, and K. Kishore. We thank T. Poons for sample sequencing, and data coordination, and A. Materna (QIAGEN) for technical assistance. This work was supported by grants from the National Human Genome Research Institute (U54HG003067) to the Broad Institute, the Center for Environmental Health Sciences at MIT, the Center for Microbiome Informatics and Therapeutics at MIT, and the Fijian Ministry of Health. Additional support was provided by a Columbia University Earth Institute Fellowship (I.L.B.); a Broad Institute Lawrence Summers Fellowship (L.X.); a Burroughs Wellcome Fund Career Award at the Scientific Interface (P.C.B.); and an R01 DE020891 funded by the NICDR and ENIGMA and a Lawrence Berkeley National Laboratory Scientific Focus Area Program supported by the US Department of Energy, Office of Science, Office of Biological and Environmental Research (S.Y. and A.K.S.). Sandia is a multi-program laboratory operated by Sandia Corp., a Lockheed Martin Co., for the United States Department of Energy under Contract DE-AC04-94AL85000.

Author Contributions I.L.B. and E.J.A. designed the study. I.L.B., S.D.J., A.P.J. and W.N. oversaw and performed the field collection of FijicOMP data and samples. I.L.B., L.X., S.Y., and M.T. performed all experimental work. D.G., B.W.B., J.R.W., P.C.B., R.I.J.X. and A.K.S. oversaw the DNA sequencing production. I.L.B. and K.H. processed the shotgun data and performed alignments. I.L.B., K.H., and D.G. provided new analytical tools. I.L.B. and E.J.A. wrote the manuscript.

Reviewer Information Nature thanks P. Bork, K. Forslund, P. Hugenholtz and C. Rinke for their contribution to the peer review of this work.
METHODS

Overview of the Fiji Community Microbiome Project. The Fiji Community Microbiome Project was developed to characterize the role of human-associated bacteria in health and disease from a developing world population recruited in the Fiji Islands. The study sought to understand the transmission of microbiome components among individuals and their environmental surroundings. The goal was to be as comprehensive as possible in the study villages. The study included 300 individuals, regardless of health status, each of whom chose to provide any or all of stool, saliva, or skin swab samples. To date, this is one of the largest cohorts on which metagenomic sequencing has been performed. Surveys were performed by visiting all households within each community. Individuals under the age of 5 years were excluded, as were individuals who were deemed mentally incapable of providing informed consent. Informed consent was received from all participants and parental consent was additionally required for all minors. IRB approval was received from Institutional Review Boards at Columbia University, the Massachusetts Institute of Technology and the Broad Institute and ethics approvals were received from the Research Ethics Review Committees at the Fiji National University and the Ministry of Health in the Fiji Islands. Whole genome metagenomic shotgun sequencing was performed on individuals’ stool and saliva samples, in addition to environmental samples from individuals’ proximal environments. More information and links to the data set can be found at http://www.fijicomp.org.

Samples used in this study. Saliva samples were collected in 20% glycerol and frozen within 30 min of collection. Stool samples to be used for metagenomic sequencing were collected within 30 min of voiding, stored in RNALater (QIAGEN) and frozen at 80°C. Stool samples used for retrieval of intact cells for single-cell analysis were collected within 30 min of voiding, stored in 20% glycerol and PBS, and frozen at −80°C. Seven individuals’ samples were selected from the FijiCOMP cohort for isolation of single-cell genomes (Supplementary Table 2). Samples from the single-cell methods were prepared using two different single-cell amplification approaches: one based on sorting individual cells and a second based on capturing individual cells within a hydrogel.

Flow-sorted single-cell amplification. For the sorting method, thawed cells were resuspended in PBS glycerol (20%) to a concentration of 10⁶ cells per ml. Samples were serially filtered through 30-μl and 11-μl membranes then briefly sonicated (20 s) and diluted 50–100-fold. They were sorted into individual wells of a 384-well plate containing 0.5 μl PBST (0.1%). Samples were sonicated for 30 μl hydrogel and 10 pg DNA was captured in each cluster. This was dissolved and denatured in 1 μl 400 mM KOH with 0.1 mM EDTA and 0.1 M DTT at 72 °C for 10 min before neutralization. The neutralized product underwent a second MDA reaction (for 10 h) within a hydrogel. The gel was then denatured and neutralized.

MDA products were purified and quantified using the Quant-iT HS assay (Thermo Fisher Scientific) and normalized to 5 ng/pL. Tagmentation reactions (Nextera) were carried out on 10 ng of purified DNA, and were followed by SPRI cleanup. Unique PCR library barcoding using Index primers N7 and S5 (Illumina) was performed, followed with SPRI twice using equal volumes of beads to DNA. Libraries of were sequenced to a depth of ∼1.5M 125-bp paired-end reads on Illumina’s HiSeq 2500. The Broad Institute’s Internal Genomics Platform’s custom designed paired-end library barcodes were used. Final library size was 200–300 bp.

Assembly of single-cell genomes. Single genome amplicons were quality filtered (Phred score ≥3), and filtered for reads that were less than 45 bp and for those that aligned with the human genome, the P. aeruginosa PA01 genome (a laboratory contaminant) and the E. coli BL21_D3E genome (from which the Φ29 polymerase used in the MDA reaction was expressed and purified) using BMDtagger. A small number of adaptor sequences were found in the raw data due to small inserts or primer-dimers. These adaptor sequences (30–61 bp in length) were easily identified using BLASTn and were removed before analysis. Amplicons were then assembled into genomes using either CLC Assembly Cell (v. 4.2), for the flow-sorted cells, or SPADES [26] (with the–careful flag) (v.3.6.0) for the hydrogel-captured cells. We retained assembled contigs that were at least 500 bp and resultant genomes where at least 100 kb could be assembled.

Filtering single-cell assemblies. To further vet the quality and purity of our assemblies, we used BLASTp to assign taxonomies to a set of 31 predetermined core genes that are both phylogenetically conserved and single copy in almost all genomes [22]. Although we could not identify the full set of 31 core genes in any of the assemblies, we removed several cells in which the core genes reflected mixed taxonomies. Additional validations of the single cell assemblies included quantifying the levels of contamination using CheckM [31]. We retained cells with less than 10% putative contamination. We used RNAMMER [20] to identify 16S sequences present in the assembled genome. We discarded a small number of cells that had multiple 16S sequences or those in which the RNAMMER-identified 16S rRNA sequence conflicted with the Sanger-sequenced 16S rRNA V68 region. Our final data set included 196 single cell assemblies (Supplementary Table 4).

Identification of horizontally transferred genes between divergent genomes. To identify horizontally transferred regions, we used a previously benchmarked method [6]. All assembled and reference genomes were compared in a pairwise manner using default in BLAST + (v. 2.2.24). Recent DNA transfers are defined as the presence of near-identical DNA fragments (99% per cent identity or greater) with >100 mutations per 500 bp in two distantly related genomes. Note that the identification of identical DNA does not imply direct transfer between two cells. We found that approximately two-thirds of the identified transferred regions contained at least one SNP, suggesting that they could not have been the result of contamination. We restricted our analysis to comparisons between genomes whose full length 16S distance would be at least 3% dissimilar. This cut-off allowed us to distinguish between signatures of vertical inheritance and horizontal transfer, as 97% divergence at the 16S corresponds with roughly 75 million years of evolution, during which time well signatures of vertical inheritance and horizontal transfer, as 97% divergence at the 16S corresponds with roughly 75 million years of evolution, during which time well
Cells that have unusual or multiple divergent 16S sequences can cause highly conserved (e.g., ribosomal) genes to appear horizontally acquired. Although some ribosomal genes may be transferred, especially across related species, because they confer antibiotic resistance, to minimize the contribution of closely related strains, we excluded all HGT events inferred between a pair of cells for which any ribo-

mal gene was inferred to be transferred. HGT was observed to occur less frequently between cell pairs as the phylogenetic distance between them increased (Extended Data Fig. 8b), as previously reported.56

To illustrate the diversity of genes used in our study and their sources, a maximum likelihood-based phylogenetic tree was constructed with FastTree2 (v.2.1.3)45 (GTR nucleotide substitution model) using a multiple sequence align-

ment generated by RDP of the full length 16S sequences where available and the 16S V68 region those were full length sequences could not be identified. 70 of the hydrogel-captured single-cell genomes lacked 16S rRNA sequences and were therefore not included in this tree. The tree was rooted using Archeal taxa as the outgroup, although placement of the root within Bacteria is unsupported.

Creating a non-redundant mobile gene set. In order to align reads to genes to attain relative gene abundances, we built a non-redundant gene data set (Extended Data Fig. 2). After pair-wise BLASTs between genomes (Step1), we clustered hori-

zontally transferred regions using single-linkage clustering (Step2), identifying all with partial or full overlapping regions. Next, we identified ORFs within each contig using Prodigal (v.2.5.3)39 (Step3). We included ORFs that overlapped at a minimum of 50% with a horizontally transferred region. This cut-off avoids false positive genes that overlap minimally with horizontally transferred regions, but allows for the inclusion of genes that may have been truncated due to highly frag-

mented draft genomes. We then clustered ORFs from each group of overlapping transferred regions into non-redundant sets using UCLUST37 (v.1.5.579), with a 90% identity cut-off (Step4). The vast majority of genes were within 99–100% identity with the gene closest to read alignment (that is, the centroid) (Extended Data Fig. 9). No gene with more than 10% ambiguous base pairs was chosen for read alignment. Since non-overlapping horizontally transferred regions may con-

tain identical genes, we performed an additional final BLASTn search between genes to further reduce redundancy, although this step resulted in the removal of a relatively small number of genes (Step5).

Functional annotation of horizontally transferred genes. We relied on several methods for functional annotation. Details on the functional annotation of each gene are provided in Supplementary Table 5. We queried sequences using BLASTp against the Kyoto Encyclopedia of Genes and Genomes (KEGG)44, using BLASTp to attain their KO classifications (e value 10−5, per cent identity 30%); against the Clusters of Orthologous Groups (COG) database using rpsBLAST (e value 10−5, per cent identity = 30%); and the TIGRFAM (v12.0) and PFAM (v.26) databases using HMMER (v.3.0) (e value 10−4, score 22). KO numbers were mapped directly back to COG; and TIGRFAM roles that could not be assigned COGs retained the TIGRFAM designation. For each gene cluster, functional annotations were aggregated by retaining annotations to any gene within the cluster first according to COG classification, followed by KEGG, TIGRFAM and finally PFAM. We also assigned metabolic functions using the Automated Carbohydrate-Active Enzyme Annotation (http://cbcb.umd.edu/bcACAN/49 (v.4.0), which employs an HMM-

based search protocol. These were then assigned substrate categories.44 To annotate antibiotic resistance genes, we used the ResFam40 core genes database, which uses HMMS. Still, 32.8% of our HGT gene set could not be annotated by any of these means. To annotate the type of mobile genetic element specifically, we performed an additional final BLASTn search between genes to further reduce redundancy, although this step resulted in the removal of a relatively small number of genes (Step5).

To determine whether the GH13 genes not identified in this study as horizontally transferred are part of the core or flexible genome. Vector contamination. Cloning vectors may have been used in the sequencing of HMP reference and draft assemblies. (Cloning vectors were not used in the sequencing of the FijiCOMP single-cell genomes and adapters were removed before assembly.) As cloning vector sequence is sometimes retained in the assembly, these assemblies were screened for vector contamination against the UniVec data-

base (http://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/), using suggested parameters. Putative contamination was observed in all 387 of the HMP reference genomes, although only 102 cells had regions of putative contamination of at least 50%, which would have affected the method used to identify HGT events (Supplementary Table 3). The metagenomic libraries were not created using cloning vectors and therefore it may be possible to distinguish between those that are contamination (that is, those that failed to recruit reads) and those that are present. Only 102 genes linked to possible vector contamination failed to recruit reads.
tRNA analysis. Metagenomic alignments were also used to determine which mobile genes were adjacent to tRNA genes. We then examine mobile elements that had integrated next to tRNAs to determine the taxonomy of the bacterial host. tRNAs have stable phylogenies that can be used for phylogenotyping49. Metagenomic alignments were filtered for paired reads that had only one read-pair aligning to a horizontally transferred gene. tRNA genes were identified using the program ARAGORN50. Taxonomies of the tRNA genes were ascertained through BLASTn searches against NT, requiring an alignment of 85 nucleotides at least 80% identical. We identified 838 mobile genes adjacent to tRNA genes. Of these, 194 had multiple bacterial hosts, multiple isoacceptors, or multiple tRNA genes. 394 of the genes adjacent to tRNA genes had at least one bacterial host that could be taxonomically identified. Plots reflect the number of reads observed. 72.7% of genes adjacent to tRNAs that had annotatable functions were genes involved in the process of horizontal gene transfer (phage, transposon, plasmid, etc.).

Defining high confidence mobile genes. We also performed our analysis on the mobile gene pools in Fijian and American metagenomes on a subset of higher confidence mobile genes representing known HGT machinery genes and/or genes on scaffolds that have additional independent support for their phylogenetic placement. BLASTn searches were performed for each of the 1,662 FijiCOMP single-cell contigs containing a mobile gene against closely related organisms where available, resulting in 634 contigs that had additional phylogenetic support that could be compared. Additionally, phylogenetic placement was attained by examining tRNA genes on paired-reads aligning to the mobile gene set. This higher confidence set includes 6,187 unique genes (59.1% of the total mobile gene data set). Even with this smaller subset of high confidence genes, distinct functional gene pools can be observed between the HMP and FijiCOMP cohorts (Extended Data Fig. 3).

Statistical analysis of gene/function presence and abundance across samples. FPKMs associated with the same protein family were summed before analysis. To assess which genes were enriched in each population, the FPKMs of each gene were compared according to country/village using FDR-adjusted Mann–Whitney tests. Comparisons of prevalence were tested using Fisher's exact tests. All analyses were performed in R (v.3.1.0).

Identifying linkage between pairs of horizontally transferred genes. Metagenomic alignments were filtered for paired reads that matched two distinct genes in the mobile gene set. The HMP and FijiCOMP metagenomic libraries were both made with an average insert size of 180 bp. We required both genes to have 99% sequence identity or greater to their respective genes. We retained only those alignments that could be compared. Additionally, phylogenetic placement was attained by examining tRNA genes on paired-reads aligning to the mobile gene set. This higher confidence set includes 6,187 unique genes (59.1% of the total mobile gene data set). Even with this smaller subset of high confidence genes, distinct functional gene pools can be observed between the HMP and FijiCOMP cohorts (Extended Data Fig. 3).

Microbiome composition. The bacterial composition of all samples was determined using MetaPhylerSR (v.0.11.3)51. The default database was supplemented with AMPHORA genes from the genome Succinibacillus sp. CAG-777 genome. The prevalence and abundance of plant matter was determined after aligning paired reads that had information on alignment and orientation. Analysis of gene linkages was performed using the igraph package (v.0.7.1) in R. Multiple contexts, where available, resulting in 634 contigs that had additional phylogenetic support that could be compared. Additionally, phylogenetic placement was attained by examining tRNA genes on paired-reads aligning to the mobile gene set. This higher confidence set includes 6,187 unique genes (59.1% of the total mobile gene data set). Even with this smaller subset of high confidence genes, distinct functional gene pools can be observed between the HMP and FijiCOMP cohorts (Extended Data Fig. 3).

Additional fish and shellfish consumed in Fiji, although not assessed for the study region include: albacore tuna (Thunnus alalunga), yellowfin tuna (Thunnus albacares), skipjack tuna (Katsuwonus pelamis), bigeye tuna (Thunnus obsesus), Spanish mackerel (Scomberomorus commerson), striped marlin (Tetrapturus audax), blue marlin (Makaira mazara), barracuda (Sphyraena sp.), swordfish (Xiphias gladius), saifish (Istiphora platypterus), opah (Lampris regius), sunfish (Mola mola), mahi mahi ( Coryphaena sp.), black snapper (Macolor niger), goatfish (Parupeneus barberinus), parrotfish (Scarus sp.), rabbit fish (Siganus purpuratus), peacock cod (Cephalopholis argus), unicornfish (Naso unicornis), cockles (Anadara antiquata), freshwater mussels (Batisa violacea) and other unspecified reef fish52.

Antibiotic Use in Fiji. Specifically, we expected that orally administered antibiotics used to treat common ailments and dental infections would be the primary target of resistance, and these are limited to beta-lactams (amoxicillin, penicillin, floxacinil, tetracycline, chloramphenicol, quinolones (ciprofloxacin), and metronidazole53–55). Thus, differences may reflect antibiotic use in other societal sectors, such as in livestock, historical uses of antibiotics, or the acquisition of multiple antibiotic resistance genes transferred within single casetters. These findings highlight our relatively limited understanding of the forces that drive selection for antibiotic resistance within populations, and the reservoirs for resistance genes56. These topics are acutely important in Fiji, as Shigella infections are common and resistance to beta lactamaes, tetracyclines, chloramphenicol, cephalosporins and quinolones has already been reported57.

Data availability. FijiCOMP metagenomic samples generated for this study are all deposited on NCBI’s Short Read Archive under Project Name PRJNA217052. The specific SRS numbers corresponding to each study participant’s gut and oral microbiome can be found in Supplementary Table 2. The single-cell assemblies can be found in the Data section at http://www.fijicomp.org.
Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **32**, 11–16 (2004).

Liu, B., Gibbons, T., Ghodsi, M., Treangen, T. & Pop, M. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics* **12** (Suppl. 2), S4 (2011).

Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).

Robertson, A. F. S. (ed.) *Food and Nutrition in Fiji: Nutrition-Related Diseases and their Prevention* (Univ. South Pacific, 1991).

Rush, E., Hedges, R., Alsbersberg, B., Qionibaravi, D. & Laulu, M. Staple food intake in a rural village in Verata, Fiji. *Pac. Health Dialog* **8**, 44–46 (2001).

Food and Agricultural Organization. *Pacific Food Security Toolkit: Building Resilience to Climate Change, Root Crop and Fishery Production* (Food and Agricultural Organization, 2010).

Goetze, J. *Evidence of Artisanal Fishing Impacts and Depth Refuge in Assemblages of Reef Fish of a Fijian Island* (Thesis, Univ. Western Australia, 2009).

Jupiter, S., Saladrau, W. & Vave, R. Assessment of Sea Cucumber Fisheries Through Targeted Surveys of Lau Province, Fiji (Wildlife Conservation Society/University of the South Pacific/Fiji Department of Fisheries/Khaled bin Sultan Living Oceans Foundation, 2013).

Kumar M., Aalbersberg B. & Mosle L. *Mercury Levels in Fijian Seafoods and Potential Health Implications*. (World Health Organization, Western Pacific Region, 2004).

Murti, A. & Morse, Z. Dental antibiotic prescription in Fijian adults. *Int. Dent. J.* **57**, 65–70 (2007).

National Drugs and Therapeutics Subcommittee. *Antibiotic Guidelines 3rd Edition* (Ministry of Health, Government of Fiji, 2011).

Thompson, C. N. et al. Typhoid fever in Fiji: a reversible plague? *Trop. Med. Int. Health* **19**, 1284–1292 (2014).

Sommer, M. O. A., Fantas, G. & Church, G. M. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science* **325**, 1128–1131 (2009).

Watson, C. Death from multi-resistant shigelloses: a case study from Fiji. *Pac. Health Dialog* **13**, 111–114 (2006).
Extended Data Figure 1 | Phylogeny of assemblies used in the study span the bacterial Tree of Life. A phylogenetic tree constructed using a multiple sequence alignment of the full 16S rRNA gene or the V68 region of the 16S rRNA gene of all reference genomes and single-cell assemblies used in this analysis where available. 16S alignments were constructed using RDP. The tree was then assembled using FastTree. Support was low for all deep branches in the tree, so the archeal branch serves as the outgroup for illustrative purposes only. The outer colour bar displays taxonomic associations for archea and bacterial phyla. The inner colour bar displays the source of that operational taxonomic unit: HMP reference cells (n = 387 blue) and FijiCOMP single cell assemblies (n = 110, red). 16S rRNA gene sequences were not available for 70 FijiCOMP single-cell assemblies, which are therefore not included in this tree.
A  Identifying mobile genes and removing redundancy.

Run pair-wise BLASTs on those that with <97% similar 16S rRNA genes.

Filter BLAST hits for regions with >99% identical and >500bp

Identify ORFs using union/find algorithm.

Redundancy of the ORFs from each region is reduced using UCLUST.

Each of these is a gene cluster. A centroid is picked.

The same gene may be part of two gene clusters. For example, gene “a” and gene “m” may be the same transposon.
We used BLAST to identify identical or near identical (>99%) genes across our dataset. This step affects a small number of genes, but further helps to reduce redundancy in the final dataset.

B  Align metagenomic reads to mobile gene dataset

Only include reads that align with 99% identity over at least 50% of the read length.

Only count gene alignments with 80% coverage over their length and at least 4x coverage.

Calculate FPKM.

Extended Data Figure 2  Methodology for identifying horizontally transferred genes and assessing their distribution within the metagenomic samples. Horizontally transferred regions were first identified using pair-wise BLASTs between HMP reference genomes and FijiCOMP single cell assemblies. Open reading frames were annotated within the horizontally transferred regions. Genetic redundancy was removed in the mobile gene set to ensure accurate abundance estimates using a combination of UCLUST and BLAST. Metagenomic reads were then aligned to the data set of unique mobile genes. Alignments were filtered to retain only reads that aligned with 99% identity across over 50% of their read length. Abundances of genes in the metagenomic samples were determined for genes whose alignments had a minimum of 4 × alignment depth over 80% of the gene length.
Extended Data Figure 3 | The abundance of mobile gene families is largely determined by cohort. a, A heat map is plotted showing the abundances (FPKM) of mobile genes aggregated by functional gene family (COG assignment, KEGG, TIGRFAM or PFAM family) within each of the metagenomic samples (81 HMP samples and 172 FijiCOMP samples). Hierarchical clustering using complete linkage was performed on the Euclidean distances between profiles of functional gene families across individuals and on the distances between individuals’ mobile gene composition. Values are plotted on a logarithmic scale. b, A heat map is plotted showing the abundances (FPKM) of only those mobile gene families that were deemed of higher confidence within each of the metagenomic samples. These include mobile gene families from mobile genes that were annotated as horizontal transfer machinery or had additional support for their phylogenetic placement. The placements of gene families and individuals were maintained from a for comparative purposes. c, A heat map is plotted showing the abundances (FPKM) of only those mobile genes that were observed to be transferred between HMP reference genomes within each of the metagenomic samples. The placements of gene families and individuals were maintained from a for comparative purposes.
Extended Data Figure 4 | Distributions of GH13 genes and glycoside hydrolase families within mobile genes of higher confidence display population-specific enrichment. a, Prevalence and abundance (measured by FPKM) of mobile genes annotated as members of the GH13 family in the FijiCOMP (n = 172, red) and HMP (n = 81, blue) metagenomic stool samples. b, Prevalence and abundance of all glycoside hydrolase families within the higher confidence mobile gene subset present in the FijiCOMP (red) and HMP (blue) metagenomic stool samples. Only unique gene families from mobile genes that were annotated as horizontal transfer machinery or had additional support for their phylogenetic placement are included here. Abundances were measured by FPKM, aggregated according to glycoside hydrolase family, and plotted as a function of the density across samples. For each glycoside hydrolase family, the number of unique horizontally transferred genes observed is noted, as are the sources of their substrates.
Extended Data Figure 5 | Composition of the gut microbiomes of HMP and FijiCOMP study participants. 

a. Relative abundances of bacteria according to phylum plotted for metagenomic samples from individuals in the HMP (n = 81, blue) and FijiCOMP (n = 172, red) cohorts. Samples are sorted according to cohort and the abundance of the dominant phyla.

b. Relative abundances of families within the order Bacteroidales plotted for metagenomic samples from individuals in the HMP (blue) and FijiCOMP (red) cohorts. Samples are sorted according to cohort and the abundance of the top Bacteroidales member.
Extended Data Figure 6 | Mobile genes are observed in a wide variety of bacterial host backgrounds across the two cohorts. 

a, b. A heat map is plotted showing the number of read-pairs per person that aligned to both a tRNA gene and two specific horizontally transferred genes. Colours within the heat map reflect the read abundance according to the species associated with the specific tRNA gene. The colour bar shows which metagenomic cohort the reads are from: FijiCOMP (red) and HMP (blue).
Extended Data Figure 7 | The relative abundances of genes and contexts across populations is not sensitive to precise definitions. Percentages of gene families, as determined by COG annotations (left), identical genes (middle) and gene contexts (right) between populations for a wide range of parameters. Bars are plotted in 5% increments. Bars shaded in black are the parameters that are plotted in Fig. 4.
Extended Data Figure 8 | Horizontal transfer varies across cells at different phylogenetic distances. a, Nucleotide identity cut-offs for full length 16S rRNA and the V68 16S rRNA region were compared to avoid comparisons between closely related cells. For each pair of HMP reference genomes, nucleotide identity for their full-length 16S rRNA is plotted against that of their V68 regions. 97% identity of full-length 16S (corresponding to approximately 75 million years of evolution) was used as a cut-off, whereas 95% was used as a cut-off when only sequences in the V68 region were available. Only those genomes above 90% similar at both the full-length and V68 region are shown. b, The number of cell–cell comparisons contributing to each of the lines. c, HGT frequency plotted as a function of the phylogenetic divergence between species between all cell-cell comparisons (black), between HMP reference genomes only (blue) and between the FijiCOMP single cell assemblies (red). This plot includes only cells for which full-length 16S rRNA genes could be identified.
Extended Data Figure 9 | Representative genes chosen for the final mobile gene data set are highly similar to the genes that were filtered to reduce redundancy. For each overlapping horizontally transferred region observed in cell–cell BLASTn comparisons between the reference genomes and single-cell assemblies, genes were clustered to identify unique genes and reduce the redundancy of the gene set. This step is essential for accurate abundance measurements of these genes in the metagenomic data sets after read alignment. All open reading frames from each overlapping horizontally transferred region were grouped using UCLUST. The nucleotide identities of each of the filtered genes and the gene chosen for read alignment (that is, the centroid) are plotted.
Extended Data Figure 10 | Metagenomic reads align to mobile genes with high fidelity over their entire length. Metagenomic reads were required to align with 99% identity to a mobile gene over at least 50% of the read length. Despite the seemingly low 50% cut-off, almost all reads align with near-perfect nucleotide identity over the entire length of the gene.
Corrigendum: Mobile genes in the human microbiome are structured from global to individual scales

I. L. Brito, S. Yilmaz, K. Huang, L. Xu, S. D. Jupiter, A. P. Jenkins, W. Naisilisili, M. Tamminen, C. S. Smillie, J. R. Wortman, B. W. Birren, R. J. Xavier, P. C. Blainey, A. K. Singh, D. Gevers & E. J. Alm

Nature 535, 435–439 (2016); doi:10.1038/nature18927

In this Letter, there were many errors in the reference citations. The original ref. 6 (Parks et al., 2015) should have been cited as ref. 32 in the Methods. There were also several other errors affecting both the reference list order and the citations throughout the main text and Methods. The references have been corrected in the online version of the Letter. We thank L. M. Rodriguez-R for bringing this to our attention.