LETTER

doi:10.1038/nature10571

Ecology drives a global network of gene exchange connecting the human microbiome

Chris S. Smillie[1]*, Mark B. Smith[2]*, Jonathan Friedman[1], Otto X. Cordero[3], Lawrence A. David[4] & Eric J. Alm[3,5,6]

Horizontal gene transfer (HGT), the acquisition of genetic material from non-parental lineages, is known to be important in bacterial evolution[1,2]. In particular, HGT provides rapid access to genetic innovations, allowing traits such as virulence[3], antibiotic resistance[4] and xenobiotic metabolism[5] to spread through the human microbiome. Recent anecdotal studies providing snapshots of active gene flow on the human body have highlighted the need to determine the frequency of such recent transfers and the forces that govern these events[4,5]. Here we report the discovery and characterization of a vast, human-associated network of gene exchange, large enough to directly compare the principal forces shaping HGT. We show that this network of 10,770 unique, recently transferred (more than 99% nucleotide identity) genes found in 2,235 full bacterial genomes, is shaped principally by ecology rather than geography or phylogeny, with most gene exchange occurring between isolates from ecologically similar, but geographically separated, environments. For example, we observe 25-fold more HGT between human-associated bacteria than among ecologically diverse non-human isolates ($P = 3.0 \times 10^{-270}$). We show that within the human microbiome this ecological architecture continues across multiple spatial scales, functional classes and ecological niches with transfer further enriched among bacteria that inhabit the same body site, have the same oxygen tolerance or have the same ability to cause disease. This structure offers a window into the molecular traits that define ecological niches, insight that we use to uncover sources of antibiotic resistance and identify genes associated with the pathology of meningitis and other diseases.

The human body is a complex biological network comprising ten microbes for each human cell and 100 microbial genes for each unique human gene[6]. Because this hidden microbial majority is known to have profound impacts on many aspects of human health including immunity[7], inflammatory disease[8] and obesity[9], considerable efforts are underway to document the genetic diversity of the human microbiome. The role of HGT in the generation and distribution of this biochemical repertoire is unclear, although anecdotal findings suggest that it may be significant[4,5,10]. In addition to informing our understanding of microbial evolution, predictive models of gene transfer are needed for the effective engineering of the human microbiome because HGT facilitates rapid adaptation to drugs and other perturbations[4,5]. Until now, however, a dearth of available genome sequences and appropriate analytical techniques have left an incomplete view of the forces that govern HGT[11].

Many previous efforts to explore these forces have highlighted the relationship between phylogeny and HGT[11–14]. Phylogeny is expected to influence HGT strongly because shared evolutionary history is associated with overlap in the host range of mobile elements[14], establishing a mechanistic basis for the phylogenetic control of gene exchange. Meanwhile, upon transfer, selection favours the persistence of genes acquired from close relatives, because these genes have greater compatibility with native molecular machinery[15,16].

Geography might provide an alternative structure to HGT by restricting dispersal, as suggested by the geographically organized distribution of *Vibrio cholera* integrons[17] and NDM-1 antibiotic resistance genes[18].

A third possibility is that ecological similarity shapes networks of gene exchange by selecting for the transfer and proliferation of adaptive traits or by increasing physical interactions between community members. Reports of enriched levels of HGT between hyperthermophiles[19] and spatially segregated exchange among *Shewanella* isolates[20] offer suggestive glimpses of such an ecological structure. However, it has been difficult to determine whether ecology has a broader function in HGT because of the limited availability of genomes from similar environments and because most previous work has ignored the distinction between recent transfers and ancient events. The inclusion of transfers from millions or billions of years in the past can obscure ecological structure, because historical niches may not reflect modern environmental associations.

To explore the effects of phylogeny, geography and ecology on HGT we use an evolutionary-rate heuristic to identify recent transfers between thousands of microbial genomes. Our heuristic finds blocks of nearly identical DNA (more than 500 nucleotides, more than 99% identity) in distantly related genomes (less than 97% 16S rRNA similarity). HGT is the best explanation for these observations because the highly conserved 16S gene evolves about 25-fold more slowly than protein-coding synonymous sites[21]. As a result, vertically inherited orthologues in such divergent genomes are nearly saturated with mutations at synonymous sites[22], in contrast to the almost perfect identity that we require. To avoid overcounting transfers, we cluster similar genomes and normalize against the number of possible comparisons.

We have confirmed that at least 98% of all HGT events identified with our approach include a predicted protein-coding gene, indicating that potentially problematic non-coding elements do not significantly affect our results. To validate our HGT detection method further, we use two phylogenetic inference methods to evaluate the evolutionary origins of putatively transferred sequences. Quartet mapping and a gene loss analysis each support 99% of identified HGTs (Supplementary Fig. 1).

As expected, a large fraction of observed transfers (27%) include at least one predicted mobile element, underscoring the importance of these genes in facilitating exchange. However, when we account for redundancies we find that mobile elements such as plasmids (2%), phages (1%) and transposons (9%) reflect only a promiscuous minority of the 10,770 total unique proteins that we observe, whereas the majority of unique genes (87%) provide other functions.

Direct exchange between any two bacteria in our data set is unlikely, both because we limit our analysis to distantly related bacteria and because strains were isolated from different human subjects or

[1]Computational and Systems Biology Initiative, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [2]Microbiology Graduate Program, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [3]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [4]Society of Fellows, Harvard University, Cambridge, Massachusetts 02138, USA. [5]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [6]Broad Institute, Cambridge, Massachusetts 02139, USA.
*These authors contributed equally to this work.

environments, often on different continents. An average pairwise distance of 7,000 km separates bacteria engaging in HGT. Therefore, each observed HGT probably reflects two independent acquisitions from a shared pool of mobile DNA, followed by proliferation.

To quantitatively explore the connectivity of bacteria in the human microbiome relative to other environments, we compare gene transfer between the 1,183 human-associated bacteria and 1,052 non-human-associated isolates from a broad range of aquatic, terrestrial and host-associated environments across the world. Even after correcting for biased sampling of human-associated clades (see Methods), pairs of bacteria isolated from the human body are 25-fold more likely to share transferred DNA than pairs from other environments ($P = 3.0 \times 10^{-270}$; combined Mann–Whitney *U*-test).

This enrichment in human-associated transfer may be caused by the prevalence of overlapping selective pressures in the tightly regulated, endothermic human host in comparison with diverse, non-human environments that experience significant temporal and spatial variation in selective pressures. Consistent with this hypothesis, when the environment is specified more precisely by focusing on human isolates from the same body site, we observe twofold higher rates of transfer ($P = 9.9 \times 10^{-108}$; combined Mann–Whitney *U*-test). Among the most closely related isolates from the same body site, this corresponds to recent HGT in more than 40% of comparisons. This elevated transfer between bacteria isolated from similar environments extends beyond the human body, with threefold more HGT between bacteria isolated from the same non-human environment relative to isolates from different non-human environments ($P = 1.3 \times 10^{-31}$; combined Mann–Whitney *U*-test).

However, an alternative explanation for these observations is that closely related bacteria colonize similar environments, creating an apparent ecological effect that is actually driven by shared evolutionary history. To control for such a phylogenetic effect, we plotted observed HGT over a range of phylogenetic divergences and found that the strong enrichment for exchange within similar environments (same host, same body site, same non-human environment) persisted across all distances (Fig. 1).

For a direct comparison of the relative contributions of phylogeny and ecology to the enrichment in human-associated transfer, we computed recent HGT between bacteria isolated from the human body (same ecology) and between these human-associated bacteria and all non-human-associated isolates (different ecology) over a range of phylogenetic distances. As shown by the dashed line in Fig. 2a, even the most deeply divergent bacteria that are separated by billions of years of evolution but share the same ecology engaged in more HGT than the mostly closely related isolates with different ecology. Thus, this recent gene exchange is structured by ecology more than by phylogeny.

We used a similar approach to explore the influence of geography relative to phylogeny and found that exchange between continents was slightly lower than exchange within the same continent (Fig. 2b; $P = 0.02$; combined Mann–Whitney *U*-test). However, this geographic effect was much weaker than that of phylogeny, which was itself less informative than ecology. Taken together, these analyses indicate that recent HGT frequently crosses continents and the Tree of Life to connect the human microbiome globally in an ecologically structured network.

This ecological architecture might reflect only the especially pronounced ecological differences between human-associated and non-human-associated bacteria. To determine whether ecology has a broad influence on recent gene exchange we searched for enriched HGT in narrower spatial, functional, and niche resolutions within the human host. Across all of these dimensions ecology strongly predicted gene exchange.

In addition to the previously discussed finding that transfer was enriched among bacteria from the same body site (Fig. 1), we found that further specifying the subsite of isolation (for example, separating vaginal isolates from other urogenital isolates) revealed even higher

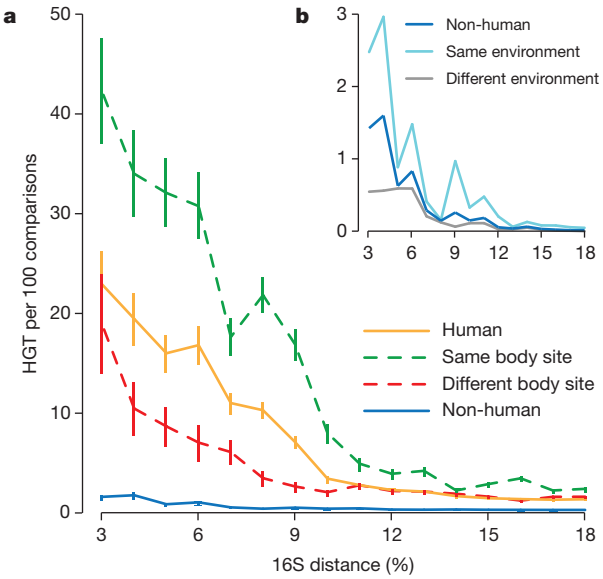

Figure 1 | Recent HGT is enriched in the human microbiome across all phylogenetic distances. HGT frequency is plotted as a function of the phylogenetic divergence between species for human-associated bacteria (**a**) and non-human-associated bacteria (**b**). We define species as clusters of genomes separated by less than 2% 16S rRNA divergence. HGT frequency is calculated in bins of 1% 16S rRNA divergence. Error bars indicate s.d. (see Supplementary Methods), with sample sizes described in Supplementary Table 8. These trends are also observed after controlling for the potential effects of sequencing centre contamination (Supplementary Fig. 4) and cosmopolitan strains (Supplementary Fig. 6).

levels of transfer across all three annotated body subsites (vagina, gingiva and nasopharynx) (Fig. 3a and Supplementary Figs 2 and 3; $P = 1.7 \times 10^{-9}$; combined Mann–Whitney *U*-test). When all human and non-human environments were considered, with scales ranging from tissues to ecosystems, we found that exchange at a narrow spatial scale, within an environment, always exceeded exchange at a broader spatial scale, with all other environments (Fig. 3b; $P = 1.3 \times 10^{-273}$; combined χ^2).

Up to this stage, our analysis relied on isolation environment as a proxy for ecological similarity, ignoring heterogeneities within these sites. Next we explored these differences, by evaluating the effects on HGT of oxygen tolerance and pathogenicity—the only other sufficiently annotated ecological features. Even after controlling for the effects of

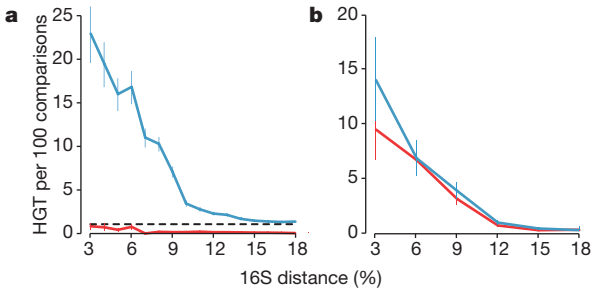

Figure 2 | Ecology is the dominant force shaping recent HGT in the human microbiome. **a**, The frequency of HGT between human-associated isolates (same ecology; blue) and between human-associated and non-human-associated isolates (different ecology; red). **b**, The frequency of HGT between bacteria isolated from the same continent (blue) and different continents (red). As a result of a reduced sample size in **b**, we pooled comparisons into larger phylogenetic distance bins of 3%. Error bars were calculated as in Fig. 1. The role of ecology in **a** is recovered when we control for sequencing centre contamination (see Supplementary Fig. 5).

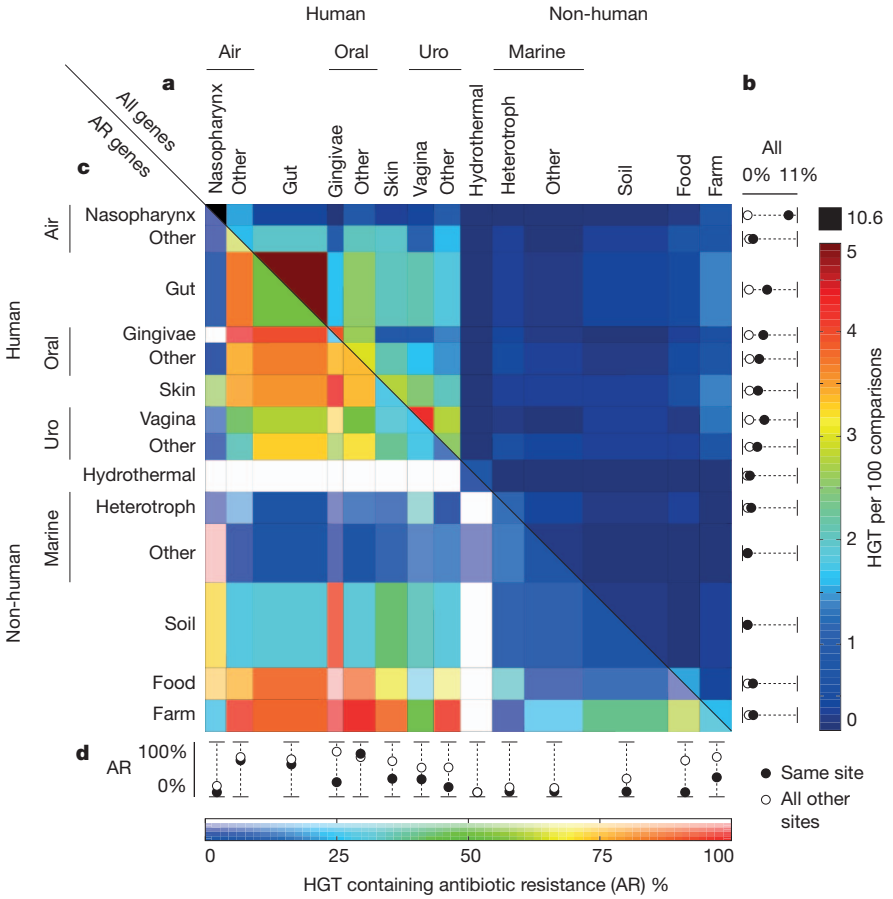

Figure 3 | HGT is ecologically structured by functional class and at multiple spatial scales. The frequency of transfer between different environments is shown for all functional groups (**a**, **b**) and for antibiotic resistance (AR) genes only (**c**, **d**). Box widths indicate the number of genomes from each environment. **a**, When all genes are considered (upper half), human isolates form a block of enrichment (upper left). **b**, For every environment examined we observe more transfer within the same environment (black dots) than between environments (white dots). **c**, The fraction of gene transfers that includes at least one AR gene for each environment. Statistical uncertainty in the proportion of AR transfer is indicated by decreased colour saturation (see Methods). **d**, AR genes comprise a significantly higher fraction of observed HGT between different environments (white dots) relative to that within the same environment (black dots) in contrast to **b**. Uro, urogenital.

body site and phylogeny, we found that HGT was also structured by oxygen tolerance (Fig. 4a; $P = 7.7 \times 10^{-13}$; χ^2) and pathogenicity (Fig. 4b; $P = 7.4 \times 10^{-11}$; χ^2). These findings demonstrate that in addition to the extensive spatial effects described above, chemical gradients and symbiotic relationships provide further ecological structure to recent HGT. Because these results persisted after controlling for explicit spatial effects, they seem to reflect selection rather than simply co-occurrence.

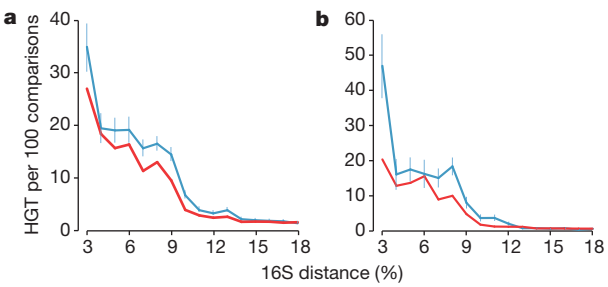

Figure 4 | Gene exchange is ecologically structured by oxygen tolerance and pathogenicity. The frequency of HGT between genomes with the same oxygen tolerance (**a**) and pathogenicity (**b**) is shown (blue) relative to their expected values (red). Expected values are based on overall frequencies of transfer between bacteria from the same distribution of body sites and phylogenetic distances. Bacteria that share the same oxygen tolerance (aerobic, anaerobic, microaerophilic or facultative aerobic) and pathogenicity (pathogenic or commensal) engage in significantly more HGT than expected under the null model, in which these traits have no influence on HGT. Error bars were calculated as in Fig. 1.

To further explore the role of selection, we probed its effects on the proliferation of different functional classes. If selection influences the rates and bounds of gene exchange, then the transfer of genes providing a non-specific selective advantage, such as antibiotic resistance, should show reduced environmental specificity relative to other, more niche-specific, functional classes. To test this prediction, for each environment, we considered the fraction of observed transfers that included at least one antibiotic resistance gene (Fig. 3c). In contrast to our earlier observation of increased transfer within sites when all functional classes were grouped together (Fig. 3a, b), here we observed that resistance comprised a higher fraction of transfers across different environments than within the same environment (Fig. 3d; $P = 6.9 \times 10^{-279}$; combined χ^2). Thus, when ecological forces transcend environmental boundaries, mobile genes do too.

We have explored networks of gene transfer to evaluate the forces that influence recent HGT, finding that ecology is profoundly important. Next we demonstrated how knowledge of this association between ecology and HGT could be used to reveal clinical insights from patterns of observed gene transfer.

Our findings, coupled with previous results[5], suggest that recently transferred genes between bacteria occupying a well-defined niche are especially likely to reflect adaptation to that niche. Consistent with this expectation, we found that many genes transferred between distantly

related meningitis isolates—such as hemolysins, adhesins and antibiotic resistance genes (Supplementary Table 1)—are known to be important in the disease[23]. We suggest that other transferred genes with unknown functions are probably cryptic virulence factors and should be prioritized for experimental annotation. Thus, in addition to recovering known virulence factors, our approach might streamline the search for novel drug targets[24], because although it is prohibitively difficult to explore all 24,095 unique meningitis genes with unknown function, it is tractable to evaluate the 13 that were recently transferred. We used this approach to identify genes associated with other diseases (for example pneumonia and endocarditis; Supplementary Tables 2 and 3) and environments (for example hot springs and soil; Supplementary Tables 4 and 5), opening a molecular window into the genetic traits that define ecological niches.

As a second example, our analysis of recent HGT revealed potential sources of clinical antibiotic resistance. We found that bacteria from farm animals and human food were enriched in transfer of resistance with human-associated bacteria relative to other non-human-associated isolates ($P = 1.7 \times 10^{-11}$ and $P = 0.01$, respectively; Mann–Whitney U-test). In all, 42 unique antibiotic resistance genes were transferred between human and farm isolates. These transferred genes comprised nine families, all of which included both genes known to provide resistance to clinical antibiotics and genes known to confer resistance to agricultural drugs (see Supplementary Table 6). This suggests that livestock-associated bacteria can contribute to clinical resistance without directly infecting humans, because for these mobile traits, genes, not genomes, serve as the unit of evolution and proliferation. Moreover, we observed 43 unique antibiotic resistance genes crossing national borders, suggesting that because the human microbiome is globally connected, local contamination of the shared mobile gene pool can have significant transnational consequences.

We have shown that ecology governs recent HGT and used this finding to reveal the key genes and networks of exchange that facilitate colonization, and occasionally exploitation, of the human host. In the future this approach could be extended to analyse bacterial genomes from individuals or groups of individuals that differ in diet, disease or descent to search for the microbial genes that relate to these human conditions.

METHODS SUMMARY

All 16S genes were identified with the GreenGenes database[25]. A total of 115 genomes with spurious or truncated 16S sequences were excluded from our analysis. We used BLAST (version 2.2.20) with default parameters[26] to calculate an all-against-all nucleotide alignment for 2,235 genomes downloaded from IMG[27]. We inferred HGT events from blocks of nearly identical DNA (more than 99% identity, more than 500 bp) in distantly related genomes (less than 97% 16S rRNA similarity). To avoid overcounting events in ancestral lineages, we collapsed closely related genomes by using average linkage clustering into groups ('species') with a 16S dissimilarity of 2%. For each pair of these clusters, we calculated the fraction of genome comparisons between clusters that shared at least one inferred HGT event. We summed this fraction over all pairs of clusters and normalized to the total number of comparisons, to calculate the HGT per 100 comparisons. Statistical tests of HGT enrichment were performed separately for each distance bin, then combined into a single p value with Fisher's method. We modelled antibiotic resistance transfer as a binomial random variable with parameter p and calculated a 95% confidence interval around our estimate of p. The size of this confidence interval, which is the statistical uncertainty of our estimate, was used to desaturate the colour of the heat map in Fig. 3c. To explore the effects of oxygen tolerance and pathogenicity on HGT, we used a χ^2 test to compare the observed frequency of HGT with the expected value given the distribution of body sites and phylogenetic divergences. Protein-coding regions were identified and annotated with BLASTX[26] (Expect (E) value $< 10^{-50}$) and UBLAST[28] (maxtargets = 100, E value $< 10^{-50}$) searches against the NCBI nr database. Unique genes reflect unique best BLAST hits to the database. Antibiotic resistance genes were annotated with the Antibiotic Resistance Genes Database[29]. Data sets are available from http://almlab.mit.edu/data.

Received 20 May; accepted 19 September 2011.
Published online 30 October 2011.

1. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405,** 299–304 (2000).
2. Koonin, E. V., Makarova, K. S. & Aravind, L. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol.* **55,** 709–742 (2011).
3. Chen, J. & Novick, R. P. Phage-mediated intergeneric transfer of toxin genes. *Science* **323,** 139–141 (2009).
4. Lester, C. H., Frimodt-Moller, N., Sorensen, T. L., Monnet, D. L. & Hammerum, A. M. *In vivo* transfer of the *vanA* resistance gene from an *Enterococcus faecium* isolate of animal origin to an *E. faecium* isolate of human origin in the intestines of human volunteers. *Antimicrob. Agents Chemother.* **50,** 596–599 (2006).
5. Hehemann, J.-H. *et al.* Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464,** 908–912 (2010).
6. Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* **312,** 1355–1359 (2006).
7. Round, J. L. & Mazmanian, S. K. The gut microbiota shapes intestinal immune responses during health and disease. *Nature Rev. Immunol.* **9,** 313–323 (2009).
8. Xavier, R. J. & Podolsky, D. K. Unravelling the pathogenesis of inflammatory bowel disease. *Nature* **448,** 427–434 (2007).
9. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: human gut microbes associated with obesity. *Nature* **444,** 1022–1023 (2006).
10. Xu, J. *et al.* Evolution of symbiotic bacteria in the distal human intestine. *PLoS Biol.* **5,** e156 (2007).
11. Lawrence, J. G. & Hendrickson, H. Lateral gene transfer: when will adolescence end? *Mol. Microbiol.* **50,** 739–749 (2003).
12. Thomas, C. M. & Nielsen, K. M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature Rev. Microbiol.* **3,** 711–721 (2005).
13. Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* **19,** 2226–2238 (2002).
14. Mazodier, P. & Davies, J. Gene transfer between distantly related bacteria. *Annu. Rev. Genet.* **25,** 147–171 (2011).
15. Tuller, T. *et al.* Association between translation efficiency and horizontal gene transfer within microbial communities. *Nucleic Acids Res.* **39,** 1–13 (2011).
16. Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl Acad. Sci. USA* **96,** 3801–3806 (1999).
17. Boucher, Y. *et al.* Local mobile gene pools rapidly cross species boundaries to create endemicity within global *Vibrio cholerae* populations. *mBio* **2**(2), e00335–10, doi:10.1128/mBio.00335-10 (2011).
18. Kumarasamy, K. K. *et al.* Emergence of a new antibiotic resistance mechanism in India, Pakistan, and the UK: a molecular, biological, and epidemiological study. *Lancet Infect. Dis.* **10,** 597–602 (2010).
19. Aravind, L., Tatusov, R. L., Wolf, Y. I., Walker, D. R. & Koonin, E. V. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet.* **14,** 442–444 (1998).
20. Caro-Quintero, A. *et al.* Unprecedented levels of horizontal gene transfer among spatially co-occurring *Shewanella* bacteria from the Baltic Sea. *ISME J.* **5,** 131–140 (2010).
21. Ochman, H., Elwyn, S. & Moran, N. A. Calibrating bacterial evolution. *Proc. Natl Acad. Sci. USA* **96,** 12638–12643 (1999).
22. Ochman, H. & Wilson, A. C. Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. *J. Mol. Evol.* **26,** 74–86 (1987).
23. Kim, K. S. Pathogenesis of bacterial meningitis: from bacteraemia to neuronal injury. *Nature Rev. Neurosci.* **4,** 376–385 (2003).
24. Clatworthy, A. E., Pierson, E. & Hung, D. T. Targeting virulence: a new paradigm for antimicrobial therapy. *Nature Chem. Biol.* **3,** 541–548 (2007).
25. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72,** 5069–5072 (2006).
26. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
27. Markowitz, V. M. *et al.* The integrated microbial genomes (IMG) system. *Nucleic Acids Res.* **34,** D344–D348 (2006).
28. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–2461 (2010).
29. Liu, B. & Pop, M. ARDB—Antibiotic Resistance Genes Database. *Nucleic Acids Res.* **37,** D443–D447 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by National Science Foundation awards 0918333 and 0936234 to E.J.A., and by the Department of Energy's ENIGMA Scientific Focus Area. This work is part of the National Institutes of Health Human Microbiome Project.

Author Contributions C.S.S., M.B.S. and E.J.A. conceived the study. C.S.S., M.B.S., J.F. and E.J.A. analysed the data. C.S.S., M.B.S., J.F., O.X.C., L.A.D. and E.J.A. provided conceptual insight. C.S.S., M.B.S. and E.J.A. 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 this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to E.J.A. (ejalm@mit.edu).