Bayesian community-wide culture-independent microbial source tracking

Dan Knights[1], Justin Kuczynski[2], Emily S Charlson[3,4], Jesse Zaneveld[2], Michael C Mozer[1], Ronald G Collman[3], Frederic D Bushman[3], Rob Knight[5,6] & Scott T Kelley[7]

Contamination is a critical issue in high-throughput metagenomic studies, yet progress toward a comprehensive solution has been limited. We present SourceTracker, a Bayesian approach to estimate the proportion of contaminants in a given community that come from possible source environments. We applied SourceTracker to microbial surveys from neonatal intensive care units (NICUs), offices and molecular biology laboratories, and provide a database of known contaminants for future testing.

Advances in sequencing technology and informatics, including the minimum information about metadata standards (minimum information about a marker gene sequence, a metagenome sequence and a genome sequence), are resulting in an exponential increase in the acquisition and sharing of microbial data. These advances are revolutionizing our understanding of the roles microbes have, for example, in health and disease or in biogeochemical cycling. Considerable attention has been paid to reducing error from PCR[1] and sequencing[2], but the problem of sample contamination has been relatively unstudied. Preparing contaminant-free DNA is challenging, and the sensitivity of PCR and whole-genome amplification methods means that even trace contamination can become a serious issue[3]. Ideally, computational methods could identify both the source and quantity of contamination, and this knowledge could help prevent future instances of contamination. Furthermore, accurately estimating the proportion of contamination from a given source environment would have far-reaching applications in source tracking, for example, in forensics, pollution and public health.

We developed SourceTracker, a Bayesian approach to identifying sources and proportions of contamination in marker-gene and functional metagenomics studies. Our approach models contamination as a mixture of entire source communities into a sink community, where the mixing proportions are unknown. Previous approaches to microbial source tracking have been focused on detection of fecal contamination in water[4–6], limited to detection of predetermined indicator species and custom-tailored biomarkers from source communities. One notable exception[7] uses community structure to measure similarity between sink samples and potential source environments. Other prior work uses data-driven identification of indicator species but lacks a probabilistic framework[8]. SourceTracker's distinguishing features are its direct estimation of source proportions and its Bayesian modeling of uncertainty about known and unknown source environments.

We collected barcoded pyrosequencing datasets of bacterial 16S ribosomal RNA gene sequences representing surface contamination in office buildings, hospitals and research laboratories, and reagents used for metagenomics studies (**Supplementary Table 1** and Online Methods). Using SourceTracker, we compared these data to published datasets from environments likely to be sources of indoor contaminants, namely human skin, oral cavities, feces[9] and temperate soils[10] (**Supplementary Table 2**). We treated these natural environments as sources contributing organisms to the indoor sink environments through natural migration (as with office samples) or inadvertent contamination (as with no-template PCR controls) (**Supplementary Fig. 1**).

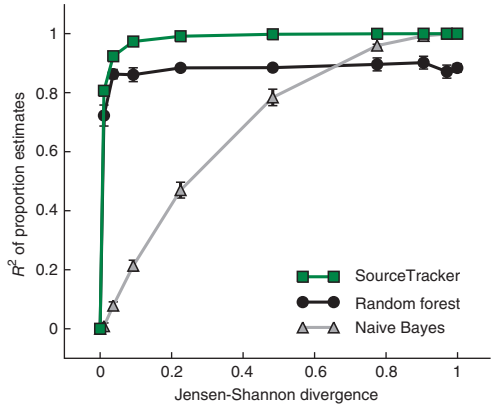

Figure 1 | Comparison of SourceTracker and other models. Indicated models estimate the proportions of two source environments in simulated samples, as the degree of overlap between the environments was varied from a Jensen-Shannon divergence of 0 (completely identical and thus impossible to disambiguate) to 1 (completely non-overlapping and thus trivial to disambiguate). The coefficients of determination (R^2) of the estimated proportions are plotted. Each point represents the mean R^2 for three trials of 100 samples each; error bars show s.e.m. (n = 3).

[1]Department of Computer Science, University of Colorado, Boulder, Colorado, USA. [2]Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, Colorado, USA. [3]Department of Medicine, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, USA. [4]Department of Microbiology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, USA. [5]Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado, USA. [6]Howard Hughes Medical Institute, Chevy Chase, Maryland, USA. [7]Department of Biology, San Diego State University, San Diego, California, USA. Correspondence should be addressed to S.K. (skelley@sciences.sdsu.edu).

RECEIVED 10 FEBRUARY; ACCEPTED 2 JUNE; PUBLISHED ONLINE 17 JULY 2011; DOI:10.1038/NMETH.1650

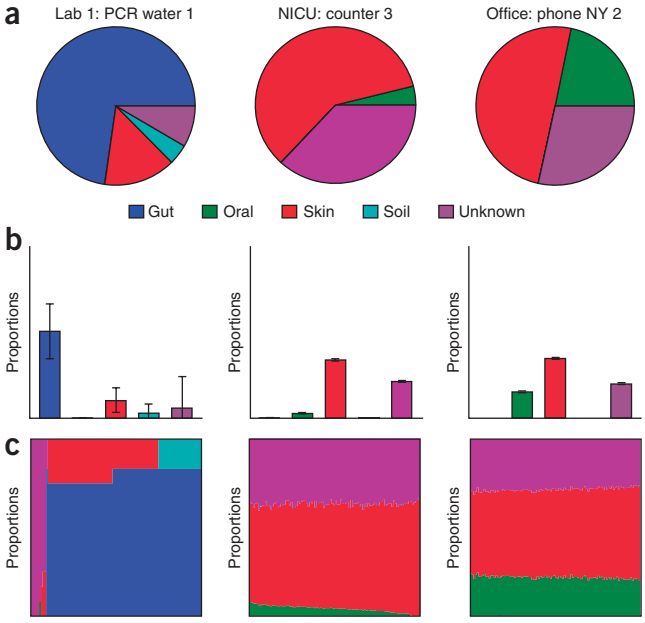

Figure 2 | SourceTracker proportion estimates for a subset of sink samples. (**a**–**c**) Source environment proportions for three sink samples estimated using SourceTracker and 45 training samples from each source environment: mean proportions for 100 draws from Gibbs sampling (**a**), data for the same samples, including s.d. of the proportion estimates (**b**), and visualization of the 100 Gibbs draws; each column shows the mixture from one draw, with columns ordered to keep similar mixtures together (**c**).

Although qualitative assessment of source and sink similarities can be performed by visualizing UniFrac distances[11] (**Supplementary Fig. 2**) or taxon relative abundance (**Supplementary Fig. 3**), these methods cannot tell us the proportion of each sink sample (such as a cotton swab) comprising taxa from a known source environment (such as soil). The problem would be trivial if source and sink environments had no taxa in common, but usually some taxa are common to both. Source-tracking methods must therefore leverage potentially useful information contained in the abundance of species with low or moderate source environment endemicity.

Previous work has used probabilistic indicator species for naive Bayes estimation[6]. Although naive Bayes actually estimates the probability that each source generated the entire sink sample, these probabilities can sometimes act as proxies for the proportions of the sink contributed by each source. We compared the accuracies of naive Bayes modeling and SourceTracker analyses as we varied the distributions of taxa in two simulated source environments from perfectly identical to perfectly non-overlapping (**Fig. 1**). Naive Bayes modeling was accurate when disambiguation was easy but inaccurate elsewhere. SourceTracker performed well even when disambiguation was difficult ($R^2 \geq 0.8$, Jensen-Shannon divergence ≥ 0.05; **Fig. 1**). We also evaluated the accuracy of the random forests classifier used in previous source-tracking work[7]. Like naive Bayes modeling, the random forests classifier estimates the probability that the entire test sample came from a single source, but these probabilities are often reasonable estimates of the mixing proportions for source tracking. Random forests classification generally performed better than naive Bayes analysis but worse than SourceTracker. SourceTracker outperformed these methods because it allows uncertainty in the source and sink distributions, and because it explicitly models a sink sample as a mixture of sources.

The Bayesian approach requires consideration of all possible assignments of the test sample sequences to the different source environments, but direct exploration is intractable. Fortunately, we can explore this joint distribution using Gibbs sampling, a technique widely used in the exploration of complex posterior distributions in applications such as topic modeling[12]. Community-wide source tracking is analogous to inferring the mixing proportions of conversation topics in a test document, except that the source environment distributions over taxa (topic distributions over words) are known from the training data, and each test sample may contain taxa from an unknown, uncharacterized source. The application of Gibbs sampling to topic modeling has been discussed previously[13].

SourceTracker considers each sink sample $\mathbf{x}$ as a set of n sequences mapped to taxa, in which each sequence can be assigned to any one of the source environments $v \in 1 \ldots V\}$, including an unknown source. These assignments are treated as hidden variables, denoted $\mathbf{z}_{i=1\ldots n} \in 1 \ldots V\}$. To perform Gibbs sampling, we initialized $\mathbf{z}$ with random source environment assignments and then iteratively reassigned each sequence based on the conditional distribution:

$$P(\mathbf{z}_i = v \mid \mathbf{z}^{\neg i}, \mathbf{x}) \propto P(\mathbf{x}_i \mid v) \times P(v \mid \mathbf{x}^{\neg i}) = \left(\frac{m_{x_i v} + \alpha}{m_{\cdot v} + \alpha m_{\cdot v}} \right) \times \left(\frac{n_v^{\neg i} + \beta}{n - 1 + \beta V} \right),$$

in which m_{tv} is the number of training sequences from taxon t in environment v, n_v is the number of test sequences currently assigned to environment v, and $\neg i$ excludes the i^{th} sequence. The first fraction gives the posterior distribution over taxa in the source environment; the second gives the posterior distribution over source environments in the test sample. Both are Dirichlet distributions, and Gibbs sampling allows us to integrate over their uncertainty. The Dirichlet parameters, α and β, act as imaginary prior counts that smooth the distributions for low-coverage source and sink samples, respectively. They also allow unknown source assignments to accumulate when part of a sink sample is unlike any of the known sources. By inferring source proportions for

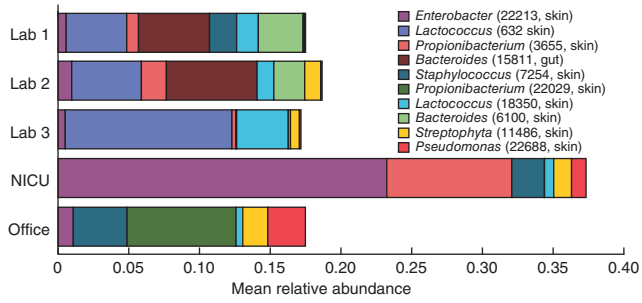

Figure 3 | Relative abundance of common contaminating operational taxonomic units (OTUs). SourceTracker may assign a different source environment to each observation (sequence) of an OTU in the sink samples. These ten OTU-source pairs had the highest average relative abundance across sink environments, excluding the unknown source. The legend gives the genus-level taxonomic classification[14] of the OTU, the OTU identifier and the source environment assigned to these observations of the OTU. Note that the OTU classified as *Enterobacter*, a lineage commonly seen in the gut, was more prevalent in the skin training samples than the gut training samples.

multiple sink samples simultaneously, we can allow them to have an unknown source in common. We could also include several unknown sources. Full details and an overview of Gibbs sampling are available in Online Methods.

For each of our indoor sink environments, we used SourceTracker to estimate the proportion of bacteria from 'gut', 'oral', 'skin', 'soil' and 'unknown' sources (that is, one or more sources absent from the training data) (**Fig. 2** and **Supplementary Figs. 4** and **5**). In general, wet-lab surface communities tended to be composed mainly of bacteria from 'skin' and 'unknown', with the exception of PCR water, which was generally more similar to 'gut'. Neonatal intensive care units (NICU) and office communities were dominated by skin bacteria, except for two samples from Arizona, USA, which were dominated by soil bacteria, and several telephone samples, which were dominated by oral bacteria. SourceTracker also reported its confidence in the estimated mixtures. For example, sample lab 1 PCR water 1 had several possible mixtures (all 'unknown', 'gut and skin', and 'gut and soil' sources), and NICU counter 3 had mostly 'skin' and 'unknown' components with an unstable gut component; we can visualize the posterior distribution over mixtures directly (**Fig. 2c**). From these results we can also determine the most common contaminating taxa (**Fig. 3**).

For low-coverage sink samples or when source environments lacked a 'core' set of taxa, SourceTracker will report high variability in the proportion estimates (**Fig. 2**). In some datasets, variation in each source environment (the 'non-core' taxa) might be accounted for by using phylogenetic information, by automatically identifying distinct niches in the broader source environment, by modeling postmixture population dynamics or by modeling potential biases inherent in the DNA extraction procedures used; these are important directions for future work. SourceTracker also assumes that an environment cannot be both a source and a sink, and we recommend research into bidirectional models.

SourceTracker can also be used to detect low-level contamination, with sensitivity adjusted by the prior parameter β. For simulations with 1% and 5% contamination, SourceTracker achieved nearly perfect specificity for a wide range of sensitivities, demonstrating that it is not restricted to low-biomass sink environments in which contamination rates are likely to be higher (area under the receiver operating characteristic curve = 0.971 for 1% and 0.989 for 5%; **Supplementary Fig. 6**).

Based on our results, simple analytical steps can be suggested for tracking sources and assessing contamination in newly acquired datasets. Although source-tracking estimates are limited by the comprehensiveness of the source environments used for training, large-scale projects such as the Earth Microbiome Project will dramatically expand the availability of such resources. SourceTracker is applicable not only to source tracking and forensic analysis in a wide variety of microbial community surveys (where did this biofilm come from?), but also to shotgun metagenomics and other population-genetics data. We made our implementation of SourceTracker available as an R package (http://sourcetracker.sf.net/), and we advocate automated tests of deposited data to screen samples that may be contaminated before deposition.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We acknowledge funding from US National Institutes of Health (R01HG4872, R01HG4866, U01HL098957 and P01DK78669), the Crohn's and Colitis Foundation of America and the Howard Hughes Medical Institute, and B. Prithiviraj for helpful insight into previous related work.

AUTHOR CONTRIBUTIONS

D.K. designed the algorithm and software, and performed computational experiments; D.K., R.K. and S.T.K. wrote the manuscript; J.K., E.S.C., J.Z., M.C.M., R.G.C. and F.D.B. contributed to writing the manuscript; J.K. and M.C.M. contributed to algorithm design; J.K. processed the data after sequencing; E.S.C. collected the data; R.G.C. and F.D.B. organized and supervised the data collection; R.G.C., F.D.B., R.K. and S.T.K. supervised the project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Acinas, S.G., Sarma-Rupavtarm, R., Klepac-Ceraj, V. & Polz, M.F. *Appl. Environ. Microbiol.* **71**, 8966–8969 (2005).
2. Quince, C. *et al. Nat. Methods* **6**, 639–641 (2009).
3. Tanner, M.A., Goebel, B.M., Dojka, M.A. & Pace, N.R. *Appl. Environ. Microbiol.* **64**, 3110–3113 (1998).
4. Simpson, J.M., Santo Domingo, J.W. & Reasoner, D.J. *Environ. Sci. Technol.* **36**, 5279–5288 (2002).
5. Wu, C.H. *et al. PLoS ONE* **5**, e11285 (2010).
6. Greenberg, J., Price, B. & Ware, A. *Water Res.* **44**, 2629–2637 (2010).
7. Smith, A., Sterba-Boatwright, B. & Mott, J. *Water Res.* **44**, 4067–4076 (2010).
8. Dufrêne, M. & Legendre, P. *Ecol. Monogr.* **67**, 345–366 (1997).
9. Costello, E.K. *et al. Science* **326**, 1694–1697 (2009).
10. Lauber, C.L., Hamady, M., Knight, R. & Fierer, N. *Appl. Environ. Microbiol.* **75**, 5111–5120 (2009).
11. Lozupone, C. & Knight, R. *Appl. Environ. Microbiol.* **71**, 8228–8235 (2005).
12. Blei, D.M., Ng, A.Y. & Jordan, M.I. *J. Mach. Learn.* **3**, 993–1022 (2003).
13. Griffiths, T.L. & Steyvers, M. *Proc. Natl. Acad. Sci. USA* **101**, 5228–5235 (2004).
14. Wang, Q., Garrity, G.M., Tiedje, J.M. & Cole, J.R. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007).

ONLINE METHODS

Data collection. We collected the 'office' samples from surfaces in 54 offices in three office buildings (18 per building) located in New York City, San Francisco and Tucson, Arizona, USA (Hewitt, K.M., Gerba, C.P., Maxwell, S.L. & S.T.K., unpublished data). In each office, we sampled the same two surfaces, phone and chair, by swabbing ~13 cm^2 with dual-tip sterile cotton specimen collection and transport swabs (BD Diagnostics). Phone and chairs had already been determined by culture-based methods to be the most contaminated surfaces in these offices (unpublished data). We also collected samples from surfaces in two different large level-three NICUs in San Diego using the same methods. After sampling, we stored swabs in sterile labeled tubes, placed them on ice and shipped them overnight or drove them directly to the lab for DNA extraction.

For the lab 1 and lab 2 datasets, we cut sterile nylon-flocked swabs (Copan) and swabs of sterile scissors into MoBio 0.7 mm garnet bead tubes (Mo Bio Laboratories) using autoclaved and flamed scissors in a biosafety cabinet, placed them at −80 °C within 1 h and stored them for <1 week before DNA extraction.

For the lab 3 dataset, we used sterile nylon-flocked swabs to sample indoor surfaces including desktops, lab benches, windowsills, a keyboard and a door handle over a three-month period from January–March 2010 in Philadelphia. We cut swabs into MoBio 0.7 mm garnet bead tubes using autoclaved and flamed scissors in a biosafety cabinet, placed them at −80 °C within 1 h and stored them for <1 week before DNA extraction.

DNA extraction, PCR and pyrosequencing. For the 'office' and 'NICU' samples, we removed the cotton from the swab using a flame-sterilized razor blade and deposited the cotton threads into a lysozyme reaction mixture. The reaction mixture had a total volume of 200 µl and the following components (final concentration given): 20 M Tris, 2 mM EDTA (pH 8.0), 1.2% P40 detergent, 20 mg ml^{-1} lysozyme and sterile water (filtered through 0.2-µm filter) (Sigma Chemical). We incubated the samples in a 37 °C water bath for 30 min. Next, we added proteinase K (DNeasy Tissue Kit, Qiagen) and AL Buffer (DNeasy Tissue Kit) to the tubes and gently mixed them. We incubated the samples in a 70 °C water bath for 10 min. We purified all samples using the DNeasy Tissue kit. After extraction, we quantified the DNA using a NanoDrop ND-1000 Spectrophotometer (NanoDropTechnologies). PCR-barcoded primers and conditions have been previously described[15]. PCR purification, dilutions and pyrosequencing (FLX instrument) were all conducted by the core facility at the University of South Carolina (Environmental Genomics Core Facility).

For the lab 1 and 2 datasets, we extracted genomic DNA from swabs using the QIAamp DNA Stool Minikit (Qiagen) with the following modifications. We added 1,500 µl of the first lysis buffer of the kit and 5 mM DTT to the nylon tips of frozen swabs. We bead-beat tubes with BioSpec Products Minibeadbeater-16 for 1 min and incubated at 95 °C for 10 min. We performed the remaining steps according to the manufacturer's protocol. We PCR-amplified 16S rRNA genes using the V1V2 primers and conditions described previously[16] in duplicate. We quantified purified amplicons using Quant-iT PicoGreen kit (Invitrogen) and pooled them in equimolar ratios. We also performed PCR on molecular biology grade water (Sigma) and included it in the pool. We carried out pyrosequencing using primer A and the Titanium amplicon kit on a 454 Life Sciences Genome Sequencer FLX instrument (Roche).

For the lab 3 dataset, we extracted genomic DNA from swabs using the same extraction kit and technique as lab 1 and 2 above. We performed PCR amplification of 16S rRNA genes using the V1V2 primers and conditions described previously[16]. We quantified purified amplicons using Quant-iT PicoGreen kit (Invitrogen) and pooled them in equimolar ratios. We also performed PCR on molecular biology grade water and included it in the pool. We carried out pyrosequencing using primer A and the Titanium amplicon kit on a 454 Life Sciences Genome Sequencer FLX instrument.

DNA barcodes and primers for all samples collected are available in **Supplementary Table 1**.

Combined preprocessing of contamination datasets. We processed the DNA sequence data for all source and sink samples in combination using the quantitative insights into microbial ecology (QIIME) pipeline[17]. To avoid bias, we selected subsets of the same size (45 samples) from each of the four source environments (**Supplementary Table 2**). We sequenced samples in multiplex using error-correcting nucleotide barcodes, and we used QIIME to de-multiplex the samples and perform quality filtering. We then used flowgram clustering[18] to remove sequencing noise. We clustered similar sequences (≥97% similarity) into OTUs with uclust[19] and assigned taxonomic identity to each OTU using the Ribosomal Database Project's taxonomy assignment tool[14]. We aligned representative sequences from each OTU against the greengenes reference 'core set' of 16S rRNA gene sequences (http://greengenes.lbl.gov/cgi-bin/nph-index.cgi). We then removed likely chimeric PCR products using Chimera Slayer[20]. We used the remaining aligned sequences to construct a phylogeny relating the sequences, via FastTree[21].

Identification and removal of chimeras. We removed likely chimeric PCR products using Chimera Slayer[20]. Note that we first aligned representative sequences from each OTU to the greengenes core set. Any OTU not aligning to the greengenes core set at >75% identity to the nearest basic local alignment search tool (BLAST) hit in the core set was discarded. These discarded sequences may contain chimeras as well as other artefacts. However, once completed we also used Chimera Slayer to screen the resulting sequences for chimeras. The number of chimeras removed were: 58 sequences from lab 1 samples (4%), 105 from lab 2 samples (4%), 4,208 from lab 3 samples (5%), 422 from office samples (0.3%) and 1,365 from NICU samples (0.6%).

Principal coordinates plots. After randomly selecting 500 sequence reads per sample and dropping low-coverage samples to control for sequencing effort, we used UniFrac[11] to measure the phylogenetic dissimilarity of all samples and performed principal coordinates analysis on the matrix of unweighted UniFrac distances using QIIME[17].

Gibbs sampling overview. To begin the Gibbs sampling procedure we assigned each sequence to a random source environment. We assumed that these assignments are correct (even though they are random) and tallied the current proportions of the source environments in the test sample. We then removed one

doi:10.1038/nmeth.1650

sequence from the tallies and reselected its source environment assignment, in which the probability of selecting each source was proportional to the probability of observing that sequence's taxon in that source, multiplied by the current estimate of the probability of observing that source in the test sample. After the reassignment, we updated the tally for the selected source environment, and repeated the process on another randomly selected sequence. After we reassigned all of the sequences many times in this manner, each set of assignments we observed was a representative draw from the distribution over all possible sequence-source assignments. To estimate the variability of this distribution, we can repeat the procedure as many times as we like, and we can report summary statistics for the mixing proportions or even visualize their distributions directly (**Fig. 2c**).

Dirichlet prior parameters. A larger value of β causes a smoother posterior distribution over environments in the sink sample. This is valuable when we want to avoid overfitting in sink samples with few sequences. By assigning different relative values of β to each environment, we can also incorporate prior knowledge about the expected distribution of source environments in our sink samples. α represents a prior count of each taxon in each source environment. This allows taxa that are unlikely under the known source environment distributions to accumulate in an unknown environment during the sampling procedure. To simplify the choice of values for α and β, we treated them as prior counts relative to the number of sequences in the test sample, rather than absolute prior counts. For all inferences performed in this paper, we set both α and β to 0.0001. We used a separate and larger value of α (0.1) for the prior counts of each taxon in the unknown environment, to prevent that environment from overfitting each individual test sample. If we had a prior belief that some of the test samples shared the same unknown environment, we could perform inference on them jointly, and reduce this separate α value accordingly.

As is typical in Gibbs sampling, we first performed 'burn-in' passes (25 passes) through the entire set of sequences in a data sample before drawing a mixture sample from the joint posterior. We also restarted the entire sampling process with new random hidden variable values 100 times, thereby collecting a total of 100 samples from the posterior distribution for each sample. Each iteration on a sink sample with V source environments required $O(V^2n)$ operations. Before running Gibbs sampling, we rarefied all samples to an artificial sequence depth of 1,000. We kept any samples whose original sequence depth was less than 1,000 at that lower depth.

Simulations. For the comparison of SourceTracker to naive Bayes and Random Forests[22] (**Fig. 1**), we simulated two source environments with varying degrees of overlap in their distribution over taxa by defining a single uniform Dirichlet prior over 100 taxa with varying concentration levels and drawing two multinomial distributions from it. By varying the concentration parameter, we could control the extent of overlap between the two multinomials. The simulation procedure (**Supplementary Fig. 7**) was repeated three times.

For the application of SourceTracker with Gibbs sampling to the detection task, we used all of the gut and skin training samples to estimate the multinomial distribution over taxa in each environment. To generate 'contaminated' samples, we drew 100 simulated samples from each environment at sequencing depth 1,000 and mixed them together with 1% (or 5%) skin and 99% (or 95%) gut. We also generated 100 pure gut samples at depth 1,000. We then ran SourceTracker as described above to estimate the proportion of skin taxa in the simulated gut samples. We used a contamination threshold of one-half of the contamination rate and varied the Dirichlet parameter β to adjust the sensitivity of the model (higher β means higher sensitivity). For each value of β, with its corresponding level of sensitivity, we measured the specificity of the contamination predictions made by SourceTracker, and plotted the values as receiver operating characteristic curves (**Supplementary Fig. 6a**,**b**).

15. Fierer, N., Hamady, M., Lauber, C.L. & Knight, R. *Proc. Natl. Acad. Sci. USA* **105**, 17994–17999 (2008).
16. Wu, G.D. *et al. BMC Microbiol.* **10**, 206 (2010).
17. Caporaso, J.G. *et al. Nat. Methods* **7**, 335–336 (2010).
18. Reeder, J. & Knight, R. *Nat. Methods* **7**, 668–669 (2010).
19. Edgar, R.C. *Bioinformatics* **26**, 2460–2461 (2010).
20. Haas, B.J. *et al. Genome Res.* **21**, 494–504 (2011).
21. Price, M.N., Dehal, P.S. & Arkin, A.P. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
22. Breiman, L. *Mach. Learn.* **45**, 5–32 (2001).