Sequence tag–based analysis of microbial population dynamics

Sören Abel¹,², Pia Abel zur Wiesch³,⁴, Hsiao-Han Chang⁵, Brigid M Davis¹,², Marc Lipsitch⁵ & Matthew K Waldor¹,²,⁶

We describe sequence tag–based analysis of microbial populations (STAMP) for characterization of pathogen population dynamics during infection. STAMP analyzes the frequency changes of genetically ‘barcoded’ organisms to quantify population bottlenecks and infer the founding population size. Analyses of intraintestinal Vibrio cholerae revealed infection-stage and region-specific host barriers to infection and showed unexpected V. cholerae migration counter to intestinal flow. STAMP provides a robust, widely applicable analytical framework for high-confidence characterization of in vivo microbial dissemination.

A pathogen’s population dynamics within a host organism reflect a plethora of factors, including the availability of hospitable niches for colonization, the extent of host barriers to infection, and the pathogen’s capacity for replication¹⁻³. However, it can be difficult to parse the relative impacts of these factors using traditional approaches, such as enumeration of colony-forming units (CFU) at different times and sites of infection, and such analyses typically require use of many experimental animals. Mathematical frameworks have been developed to characterize events that shape population structures over time based on natural variation in the genetic composition of populations, but these have generally been applied in studies of eukaryotic evolutionary biology in which numerous distinguishable alleles are present⁴⁻⁷. The inocula of infectious microbes used in laboratory analyses usually lack a sufficient number of distinguishable alleles for high-resolution analysis of pathogen population dynamics. Furthermore, the effects of natural polymorphisms are not necessarily neutral, so it can be difficult to discriminate between genetic drift and selection. Artificial tags, most recently sequence ‘barcodes’ in a method termed WITS (wild-type isogenic tagged strains), have been used to create distinguishable pathogens that are more easily analyzed and have equivalent fitness⁸⁻¹⁴. However, these studies have so far been limited by the use of small numbers of tags, which restrict the resolving power; by the need for specialized mathematical models; and by the lack of a systematic approach for analysis of tag frequencies in different populations. These limitations are not critical when the size of the founding populations (i.e., bacteria that survive host defenses and subsequently replicate) is very small, for example, when only one or a few organisms overcome the host defenses and colonize specific tissues or organs. However, they severely constrain the information that can be obtained from more complex founding populations.

In our work, we have combined population analysis frameworks with high-throughput DNA sequencing technology and large libraries of tagged pathogens to generate a new approach for dissection of microbial population dynamics during infection (STAMP) that is applicable to analyses of all populations, regardless of their complexity. In this method, we first determine the relative abundances (rather than simply presence or absence) of hundreds of individually tagged but otherwise isogenic strains in the infection inoculum. By comparing these baseline frequencies to those of samples taken at various times and sites during infection, we can estimate the number of bacteria from the inoculum whose descendants are represented in a population at the time and site of sampling. This number, which we term the founding or bottleneck population size (N₀), reflects the stringency of host barriers encountered during infection. STAMP allows the magnitude of such restrictions to be assessed retrospectively, without knowledge of their timing or location, and permits high-resolution determination of N₀ over a large dynamic range. We demonstrate STAMP’s utility through analysis of the population dynamics of the cholera pathogen, V. cholerae, in the infant rabbit model of infection¹⁵.

We first confirmed that N₀ could be estimated using approaches for determination of effective population size (Nₑ; ref. 16)—a key parameter for modeling population dynamics—in silico, using simulations in which we varied bottleneck size and the number of tags (Supplementary Fig. 1, Supplementary Data and Online Methods). These simulations revealed that 500 tags is sufficient for high-confidence determination of N₀ (Supplementary Fig. 1), whereas 50 tags is predicted to yield less robust results, particularly for high values of N₀. To validate our approach experimentally, we generated a library of V. cholerae that were individually barcoded with 1 of ~500 distinct, short sequence tags inserted into a neutral locus on the chromosome (Supplementary Figs. 2a and 3). We sampled defined numbers of bacteria (10¹⁻¹⁰ CFU) to simulate

¹Department of Microbiology and Immunobiology, Harvard Medical School, Boston, Massachusetts, USA. ²Division of Infectious Diseases, Brigham and Women’s Hospital, Boston, Massachusetts, USA. ³Department of Global Health and Social Medicine, Harvard Medical School, Boston, Massachusetts, USA. ⁴Division of Global Health Equity, Brigham and Women’s Hospital, Boston, Massachusetts, USA. ⁵Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA. ⁶Howard Hughes Medical Institute, Boston, Massachusetts, USA. Correspondence should be addressed to M.K.W. (mwaldor@research.bwh.harvard.edu).

RECEIVED 29 SEPTEMBER 2014; ACCEPTED 10 DECEMBER 2014; PUBLISHED ONLINE 19 JANUARY 2015; DOI:10.1038/NMETH.3253

© 2015 Nature America, Inc. All rights reserved.
bottleneck events and assessed whether changes in the frequency of individual tags in these samples relative to the frequency in the initial library could be used to estimate $N_b$ after the sample was expanded on agar plates (Fig. 1a and Supplementary Fig. 2a). Among several approaches we tried for sequencing-based estimation of $N_b$, optimal results were achieved using equations from Krimbas and Tsakas$^5$ (Fig. 1a and Supplementary Fig. 4). There was a very high correlation ($R^2 = 0.99$) between the estimated $N_b$ and the associated bacterial load over a range of approximately five orders of magnitude (Fig. 1a), suggesting that this approach could enable accurate assessment of population bottlenecks over this range in vivo. For smaller populations, our calculations yielded an estimated $N_b$ slightly lower than the CFU-based value. Consequently, these in vitro data were used as a calibration curve for our subsequent in vivo experiments, and the corrected in vivo values are denoted $N_b'$. Like infected humans, infant rabbits oro-gastrically infected with V. cholerae develop severe and potentially fatal diarrhea because of the pathogen’s colonization of the small intestine (SI) and subsequent secretion of cholera toxin$^1$. We collected bacteria in vitro calibration curve. Correlation between experimentally determined bottleneck population size (bacterial load) and estimated bottleneck size ($N_b$). The red, blue and yellow symbols represent biologically independent samples. The solid line indicates the median; the dotted black lines indicate the 95% confidence interval. The dashed gray lines mark the resolution limit for $N_b$ estimation. (b) Representative example of nine biologically independent experiments of bottleneck populations corrected with the calibration curve ($N_b'$, black dots) and bacterial load (CFU, red squares) at 20 h post infection throughout the GI tract of a single animal after inoculation (Inoc) with $10^6$ CFU of barcoded V. cholerae. An additional example is shown in Supplementary Figure 5. (c) Sampling sites used in this study are indicated in light red or blue. S, stomach content; P1–P10, proximal SI sections used in Figure 1; I1, proximal SI section used in Figure 2; I2, middle SI; I3, distal SI; Ce, cecum tissue; Cf, cecal fluid; Co, colon.

To gain further insight into the kinetics and directionality of V. cholerae spread within the intestinal tract, we estimated $N_b'$ for a variety of intestinal sites (Fig. 1b) at three distinct phases of infection (Fig. 2). During the early phase of infection (~2 h PI), $N_b'$ and bacterial load were relatively high (~$10^5$ and ~$10^7$, respectively) in all sections, indicating that the bacteria quickly disseminated from the site of inoculation in the stomach to the distal intestine. In the middle phase of infection (~7 h PI), both $N_b'$ and CFU were reduced (relative to values in the early phase) for most regions of the intestine, suggesting that most of the inoculum was cleared relatively early after inoculation. However, during this middle phase, $N_b'$ for the distal SI (I3) remained at ~$10^5$, suggesting that I3 contained abundant niches that were permissive for V. cholerae growth from the onset of infection onwards. Given the absence of viable bacteria in the stomach at ~7 h PI and the low $N_b'$ (~$10^5$) and bacterial load of V. cholerae in I1 and I2 in this phase, our observation that $N_b'$ for the proximal SI (I1) increased to ~$10^4$ at the late phase of infection (~20 h PI) likely reflects reseeding of I1 with bacteria from I3. Notably, consumption of contaminated food or stool was not required for reseeding of I1 (Supplementary Fig. 6), strongly suggesting that the increased diversity of founders in I1 late in the infection was caused (at least in part) by migration of bacteria from the distal to the proximal region of the GI tract, i.e., movement counter to the usual direction of intestinal flow. Such migration may be aided by the onset of cholera toxin-induced fluid secretion, which typically occurred around this time (Supplementary Fig. 7). Although the biological significance of the backward migration of V. cholerae is not clear, these observations provide new insight into V. cholerae–host dynamics: the difference in I1 $N_b'$ indicates either that the number of permissive niches for V. cholerae growth in this region changes during the course of infection or that bacteria become better adapted to replicate in I1 after initial growth in I3. Intriguingly, the backward migration of V. cholerae from I3 to I1 did not substantially alter the number of founders in I2, which remained at ~$10^2$ during the late phase of infection. Despite the difference between $N_b'$ of I1 and I2, these segments contained
similar quantities of recoverable \textit{V. cholerae} (~$10^8$ CFU) in the late phase of infection, suggesting that both sites allow for replication once bacteria establish a foothold.

Large numbers of tags also enable estimation of the 'genetic distance' separating pathogen populations by comparative analyses of barcode frequencies\textsuperscript{17}. The results from these analyses were congruent with our conclusions regarding \textit{V. cholerae}'s intra-intestinal dynamics (Fig. 2b). All populations in the SI were closely related in the early phase of infection, probably reflecting relatively even spread of the inoculum, but the genetic distance between them increased by the middle phase. However, by the late phase, the populations in I1 and I3 were again closely related, a result supporting our hypothesis that the former is derived from the latter. Thus, our STAMP-based spatial and temporal analyses of \textit{V. cholerae} population dynamics in vivo revealed unexpected complexity in pathogen migration patterns and in the host landscape that largely could not be deduced from traditional approaches to investigation of colonization\textsuperscript{18}.

---

**Figure 2** Spatial and temporal dynamics of founding \textit{V. cholerae} populations along the intestine. (a) Corrected bottleneck population size ($N_b'$, black dots) and bacterial load (CFU, red squares) at different loci during early, middle and late phases (~2 h, ~7 h and ~20 h post infection, respectively) of infection from 19 animals from 12 independent litters. Open circles represent $N_b'$ values above the resolution limit or no detected colonies; dashed lines indicate the resolution limit for $N_b'$ estimation. Sample medians are represented by horizontal lines. Corresponding $N_b'$ and bacterial load from the same animal are aligned vertically and in the same sequential order. Significance was tested with one-sided Wilcoxon signed-rank tests; *$P < 0.05$, **$P < 0.01$. (b) Genetic distance of populations during different phases of the disease. The genetic distance between I1 and I3 in the middle phase of the disease is significantly different from the early ($P = 0.026$) and the late phase ($P = 0.015$; two-sided Wilcoxon rank-sum test; sample size as in a). (c) Model of the spatiotemporal dynamics of \textit{V. cholerae} infection in the infant rabbit host.
In contrast to earlier studies, \(^8\text{--}^{14,18\text{--}20}\), STAMP enables systematic and robust analysis of populations with a large number of barcodes, which is critical for high-confidence definition of population parameters, and it can resolve bottlenecks over a dynamic range orders of magnitude larger than those seen in previous analyses (Fig. 1a and Supplementary Fig. 1). STAMP's power allows us to identify bottleneck sizes with very small technical error on the basis of a single animal as opposed to animal population averages (Supplementary Fig. 1), thereby permitting us to approximate biological variance between hosts. STAMP should be universally applicable to investigation of the in vivo population dynamics of diverse bacterial pathogens as well as of host, microbiota and pathogen factors that govern these dynamics. Such analyses may be particularly interesting for pathogens that disseminate through uncharacterized bottlenecks to secondary sites of infection or for quantification of pathogen transmission between hosts. Additionally, the analytical approach underlying STAMP is equally valid for in vivo studies of viruses, parasites or other organisms that can be barcoded or equivalently tagged. Finally, if coupled with high-throughput approaches for tagging eukaryotic cells, STAMP's analytical framework could be used to dissect eukaryotic cell population dynamics, for example, in models of stem cell dissemination, immune cell maturation or cancer metastasis.

METHODS

Methods and any associated references are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors thank D. Munera for help with animal experiments, M. Chao and T. Lieberman for discussions, and S. Lory, L. Comstock and members of the Waldor lab for comments on the manuscript. This work was supported by Swiss Foundation for Grants in Biology and Medicine (http://www.samw.ch) grant P3N3_142724 /1 (S.A.), Swiss National Science Foundation (http://www.snf.ch) grant PBEZP3_140163 and German Academic Exchange Service (http://www.daad.org) grant D/11/45747 (P.A.z.W.), the National Institute of General Medical Sciences of the US National Institutes of Health (NIH) award number U54GM088558 (H.-H.C. and M.L.), NIH grant R37 AI–042347 and Howard Hughes Medical Institute (M.K.W.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript, and the content is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

AUTHOR CONTRIBUTIONS

S.A., P.A.z.W. and M.K.W. designed experiments. S.A. performed experiments. S.A., P.A.z.W. and H.-H.C. analyzed the data. S.A., P.A.z.W., B.M.D., M.L. and M.K.W. wrote the manuscript. All authors discussed the results and commented on the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

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

1. Levin, B.R., Lipsitch, M. & Bonhoeffer, S. Science 283, 806–809 (1999).
2. Gutiérrez, S., Michalakis, Y. & Blanc, S. Curr. Opin. Virol. 2, 546–555 (2012).
3. Watson, K.G. & Holden, D.W. Cell. Microbiol. 12, 1389–1397 (2010).
4. Wright, S. Nature 166, 247–249 (1950).
5. Kimbas, C.B. & Tsakas, S. Evolution 25, 454–460 (1971).
6. Nel, M. & Tajima, F. Genetics 98, 625–640 (1981).
7. Pollak, E. Genetics 104, 531–548 (1983).
8. Moxon, E.R. & Murphy, P.A. Proc. Natl. Acad. Sci. USA 75, 1534–1536 (1978).
9. Margolis, E. & Levin, B.R. J. Infect. Dis. 196, 1068–1075 (2007).
10. Barnes, P.D., Bergman, M.A., Mecsas, J. & Isberg, R.R. J. Exp. Med. 203, 1591–1601 (2006).
11. Li, Y., Thompson, C.M., Trzcinski, K. & Lipsitch, M. Infect. Immun. 81, 4534–4543 (2013).
12. Grant, A.J. et al. PLoS Biol. 6, e74 (2008).
13. Kaiser, P., Slack, E., Grant, A.J., Hardt, W.-D. & Regoes, R.R. PLoS Pathog. 9, e1003532 (2013).
14. Lim, C.H. et al. PLoS Pathog. 10, e1004270 (2014).
15. Ritchie, J.M., Rui, H., Bronson, R.T. & Waldor, M.K. mBio 1, e00047–10 (2010).
16. Charlesworth, B. Nat. Rev. Genet. 10, 195–205 (2009).
17. Cavalli-Sforza, L.L. & Edwards, A.W. Am. J. Hum. Genet. 19, 233–257 (1967).
18. Angelichio, M.J., Spector, J., Waldor, M.K. & Camilli, A. Infect. Immun. 67, 3733–3739 (1999).
19. Fu, Y., Waldor, M.K. & Mekalanos, J.J. Cell Host Microbe 14, 652–663 (2013).
20. Chiang, S.L. & Mekalanos, J.J. Mol. Microbiol. 27, 797–805 (1998).
ONLINE METHODS

Medium and growth conditions. All *V. cholerae* strains used here were generated in a streptomycin-resistant mutant of *V. cholerae* El Tor O1 Inaba strain C6706 (ref. 21). Bacteria were grown in LB-medium (Difco) supplemented with antibiotic when necessary at 37 °C. Carbenicillin (Sigma-Aldrich) was used at a final concentration of 50 μg/ml (LB-Carb), streptomycin (Sigma-Aldrich) at 200 μg/ml (LB-Strep). Growth-curve analyses were conducted in a Bioscreen C growth plate reader and 100-well honeycomb plates (Oy Growth Curves Ab Ltd.), measuring the absorption at 600 nm in 10-min intervals.

Construction of the barcoded *V. cholerae* library. All cloning procedures were conducted using isothermal assembly22. Supplementary Table 1 contains the sequences of all primers used in this study. The plasmids pSoA160 and pSoA158.mix, used for generating the tagged *V. cholerae* library, were created as follows.

1. pSoA160: a 719-bp fragment from pMK2010 (ref. 23) containing the ccbD toxin was amplified with primers P78 and P79 and inserted into pGP704, a suicide plasmid for *V. cholerae* carrying a beta lactamase gene, at the SacI and XbaI sites, yielding pSoA160. The correct plasmid sequence was confirmed by sequencing.

2. pSoA158.mix: an ~1,055-bp fragment of VC0610 that included 93 bp of the intergenic region between VC0610 and VC0611 was amplified using primer P110, which contained a 30-bp stretch of random sequence, and P80 and inserted into pSoA160 at the SacI and XbaI sites, yielding pSoA158.mix. The correct plasmid sequence of 24 individual colonies was confirmed by sequencing.

pSoA158.mix was transferred to *V. cholerae* by conjugation with SM10 λpir24. Transconjugants that successfully integrated the plasmid into the genome by homologous recombination were selected with streptomycin and carbenicillin. After 93 of 93 tested colonies were found to have the correct insertion of pSoA158 by PCR using primers P9 and P10, the remaining colonies were washed off the LB-Carb plates and pooled. After addition of 1 ml sterile phosphate-buffered saline (PBS) with a mini-beadbeater-16 and 3.2-mm stainless steel beads (BioSpec Products). A total of 750 μl of each sample, including stomach content, cecal fluid and three replicate samples of the inoculum, were spread on three separate LB-Carb plates and grown for ~18 h to be harvested for N6 estimation. Bacterial load (CFU) was enumerated by plating serial dilutions. A graphical overview of the experimental setup is depicted in Supplementary Figure 2a.

DNA sample preparation and sequencing. To sequence the barcodes, we washed bacterial colonies off the LB-Carb plates with cold PBS, and triplicate samples were pooled. Genomic DNA was extracted from samples containing ~3 × 1010 cells using the Wizard Genomic DNA Purification Kit (Promega). A 313-bp fragment containing the tag region was amplified with primer P47 and either P48 or one of P51–P73 using ~2 g of genomic DNA as template. P47 contains complementary sequence to Illumina’s P5 grafting primer, and P48 and P51–P73 contain complementary sequence to Illumina’s P7 grafting primer and additionally one unique TruSeq index barcode each. PCRs were performed in triplicates, and the PCR products were pooled before purification with a MinElute PCR Purification Kit (Qiagen).

The DNA was quantified by a Qubit 2.0 fluorometer and Qubit dsDNA HS Assay kit (Life Technologies) and by quantitative PCR with primer P74 and P75 using a Step One Plus Real-Time PCR machine and Fast Sybr Green Master Mix (Applied Biosystems).

The amplicon libraries were combined in an equimolar fashion and sequenced on an Illumina MiSeq sequencer using a 50-cycle V2 MiSeq reagent kit (Illumina) with custom sequencing primer P49. The libraries were clustered to a density of ~106 per mm2. Image analysis, base-calling, data quality assessment and demultiplexing were performed on the MiSeq instrument.

Using reaper-12-340 (ref. 25), we discarded all sequences that contained undefined base calls (N) or that did not contain the 14 bp of the constant region directly following the random sequence tag (on average 15.3% (±9.6 s.d.) of the sequenced reads). In the remaining sequences, the constant region, as well as all of the following sequence, was trimmed. The sequences were converted to FASTA format with convert_fastaqual_fastaq.py from QIIME.
1.6.0 (ref. 26) and clustered and enumerated with pick_otus.py using uclust27 and a sequence similarity threshold. The effect of different thresholds for clustering on the \( N_b \) estimation was tested; performance was best with a threshold of 0.9, and so this value was used throughout the study (Supplementary Fig. 4). For each cluster, the most abundant sequence was picked as representative with pick_rep_set.py. In order to remove remaining nonspecific tags, all clusters that were not represented in the INOC54 reference set were discarded (see below). The reproducibility of sequencing results was confirmed by comparing the sequences of the same inoculum sample or different independent inocula, resequenced on the same or separate sequencing runs (Supplementary Fig. 8). A graphical overview of the analysis setup is depicted in Supplementary Figure 2b.

**INOC54 reference set.** Sequences from 54 independently sequenced inocula samples were analyzed as described above. However, after the trimming with reaper, all sequences that contained base calls with a quality score below Q30 were discarded. Clusters that were present in 53 out of 54 inocula were included in the INOC54 reference set. A list of all tags is given in Supplementary Table 2.

**Bottleneck population size estimation.** In an idealized experiment, to answer the question "How many \( V. \) cholerae cells from the inoculum have descendants in the population of an intestinal segment sampled at time \( t \)" we would inoculate rabbits with a \( V. \) cholerae population in which each cell carried a unique tag, harvest the entire population in a segment of intestine at time \( t \) and count the number of tags therein. This was not possible for technical and logistical reasons, so we employed mathematical techniques developed in population genetics, specifically the estimation of the effective population size \( (N_e; \text{ ref. } 16) \) based on temporal allele frequency data, to estimate this quantity that we call the founding population size or bottleneck size \( (N_b) \). We assume that the changes in allele frequencies are introduced by genetic drift, i.e., by random survival of pathogens that pass through a population bottleneck. However, other sources of changes in allele frequency (for example, niche-specific differences in growth rates) can potentially confound the analysis. We hypothesized that applying an estimation formula for \( N_e \) would provide a very good estimate of \( N_b \).

Although we cannot observe the tag diversity of the \( V. \) cholerae population at each point in space and time during an infection, we hypothesized that \( N_b \) could be estimated by applying a formula to estimate \( N_e \) under the simplifying assumption that a single-step bottleneck had occurred, reducing the diversity in the inoculum down to that observed at the sampling time. Methods for estimating \( N_e \) also permit correction for the fact that not every member of the population at time \( t \) is sampled.5,16 In reality, the loss of diversity from the inoculum probably occurred in every one of several bacterial generations before sampling, though it was likely concentrated in one or a few generations because we observed severe bottlenecks followed by a robust expansion of the pathogen population in the host (Fig. 2). However, because our goal is to determine the number of bacteria that have descendants in the population, \( N_b \), at a given time and location rather than to map each step of the population constriction, we set the generation \( g \) in equation (1) equal to 1 to summarize the loss of diversity as having all occurred in a single step. The simulation results shown in Supplementary Figure 1 confirm that the estimates obtained for \( N_b \) using our approach accurately reflect the true bottleneck population size. This approach is further supported by the excellent fit of the in vitro calibration curve (Fig. 1) and the finding that the \( N_b \) estimates remain constant throughout the infection in several sections of the GI tract (Fig. 2). Several population genetic methods were used to estimate \( N_b \) (Supplementary Fig. 4 and data not shown)5,7. The best correlation between experimental CFU and estimated \( N_b \) was achieved by using equations from Krimbas and Tsakas, which was then employed throughout this study.

\[
\hat{N}_b = N_e \frac{g}{\hat{F} - \frac{1}{S_0} - \frac{1}{S_k}}
\]

where \( k \) is the total number of distinct alleles (i.e., number of unique tags); \( f_{i,0} \) the frequency of allele \( i \) at time 0, \( f_{i,s} \) the frequency of allele \( i \) at sampling, \( g \) the number of generations during competitive growth, and \( S_0 \) and \( S_k \) the sample size used to determine the population composition (i.e., the number of sequence reads) at time 0 or at sampling, respectively. The R code to estimate \( N_b \) is available upon request.

**In vitro calibration curve.** The inoculum was prepared as described above in triplicates. For each sample, independent 1:10 dilution series in PBS were prepared, spread on LB-Carb plates and grown for ~18 h. The bacterial load was determined and the colonies were harvested for \( N_b \) estimation as described above. The data were spline interpolated using the spline function in R (http://www.r-project.org/) with a step width of 0.01 log units, and the median as well as 95% confidence interval was determined. None of the tested \( N_b \) estimation methods was perfect; therefore, all estimated \( N_b \) were corrected using the calibration curve and the resulting values were denoted \( N_b' \). Additionally, when the bottleneck size is larger than the number of obtained sequences, the sampling error becomes larger than the level of genetic drift, which is used for estimating \( N_b \), and \( N_b \) can no longer be estimated accurately. Therefore, a resolution limit is given in all figures. It depicts the \( N_b' \) estimate for which it was still possible to calculate the median and both upper as well as lower confidence boundaries from the in vitro calibration data. For given values above the detection limit, the median and lower confidence boundary, but not the upper confidence boundary could be determined.

**Statistics.** The majority of data was not normally distributed; therefore, nonparametric tests were used. Wilcoxon signed-rank tests (paired data) or rank-sum tests (unpaired data) were used to compare two groups, and the Kruskal-Wallis rank-sum test was used for more than two groups. R was used for all statistical analysis.
Genetic distance. Genetic distances were calculated by the Cavalli-Sforza chord distance method\textsuperscript{17} using tag frequency distributions in populations according to

\[ D_{\text{ch}} = \frac{2\sqrt{2}}{\pi} \sqrt{1 - \cos \theta} \]  

(3)

and

\[ \cos \theta = \sum_{i=1}^{k} \sqrt{f_{P1,i} f_{P2,i}} \]  

(4)

where \( D_{\text{ch}} \) is the chord distance, \( k \) is the total number of distinct alleles (number of unique tags), and \( f_{P1,i} \) and \( f_{P2,i} \) are the frequencies of allele \( i \) in population 1 and population 2, respectively. We assume that populations that are in exchange with each other or have only been separated recently are composed of similar relative amounts of organisms carrying individual barcode tags.

21. Thelin, K.H. & Taylor, R.K. Infect. Immun. 64, 2853–2856 (1996).
22. Gibson, D.G. et al. Nat. Methods 6, 343–345 (2009).
23. House, B.L., Mortimer, M.W. & Kahn, M.L. Appl. Environ. Microbiol. 70, 2806–2815 (2004).
24. Simon, R., Priefer, U. & Pühler, A. Biotechnology (NY) 1, 784–791 (1983).
25. Davis, M.P.A., van Dongen, S., Abreu-Goodger, C., Bartonicek, N. & Enright, A.J. Methods 63, 41–49 (2013).
26. Caporaso, J.G. et al. Nat. Methods 7, 335–336 (2010).
27. Edgar, R.C. Bioinformatics 26, 2460–2461 (2010).