Supporting Information Appendix for:

Identifying personal microbiomes using metagenomic codes

Eric A. Franzosa[1,2], Katherine Huang[2], James F. Meadow[3], Dirk Gevers[2], Katherine P. Lemon[4,5], Brendan J. M. Bohannan[3], Curtis Huttenhower[1,2*]

*Corresponding author:

chuttenh@hsph.harvard.edu

(617) 432-4912

[1] Biostatistics Department, Harvard School of Public Health,
Boston, MA 02115, USA

[2] The Broad Institute,
Cambridge, MA 02142, USA

[3] Institute of Ecology and Evolution, University of Oregon,
Eugene, OR 97403, USA

[4] Department of Microbiology, The Forsyth Institute,
Cambridge, MA 02142, USA

[5] Division of Infectious Diseases, Boston Children's Hospital, Harvard Medical School,
Boston, MA 02115, USA

SUPPORTING METHODS

Evaluating code uniqueness in an independent set of stool metagenomes

To test the generality of our findings outside of the HMP cohort (1), we separately analyzed 85 stool metagenomes collected from healthy Danish subjects enrolled in the MetaHIT project (2). This dataset was selected due to (i) its large size, (ii) the importance of stool metagenome samples in human microbiome research, and (iii) the strong performance of stool samples in the analyses of the main text. The MetaHIT dataset contains only one sample (time point) per subject. Hence, while we could not apply these data to independently evaluate metagenomic code stability, we were able to apply them to independently evaluate code uniqueness (within the MetaHIT cohort and in comparison with HMP subjects).

Following the procedures used to analyze all HMP metagenomes, the 85 MetaHIT stool metagenomes were profiled with MetaPhlAn (3) to produce profiles of species and marker gene abundance (main text, Methods). These profiles were then applied to construct species- and marker-level metagenomic codes for the MetaHIT population using the algorithm described in the main text. Although the MetaHIT stool samples outnumbered the paired HMP stool samples, they were more likely to have a unique species-level code (75% unique versus 62% unique). This may reflect differences in diversity between Danish and American gut microbiomes, or potentially differences in experimental procedures used by the MetaHIT and HMP studies. Conversely, there were two MetaHIT individuals that lacked unique marker-level codes, while this did not occur within the 50 paired HMP stool metagenomes.

We further applied the MetaHIT samples and metagenomic codes to investigate the likelihood of spurious matches to previously unseen subjects. Comparing the first-visit HMP

stool samples to MetaHIT marker-level codes revealed 6 spurious matches in 4,150 comparisons (50 HMP samples × 83 unique MetaHIT codes). Repeating the validation analysis applied in the main text in the context of single-visit HMP individuals, we predicted that MetaHIT marker-level codes would be unique among 692±282 (S.E.) individuals. The comparison of HMP marker-level codes to MetaHIT samples produced a similar result: expected uniqueness among 708±289 individuals. While we cannot speak to the stability of the MetaHIT codes over time, they provide additional, strong evidence that metagenomic codes based on strain-level variation tend to be unique among 100s of individuals.

FIGURE S1: Building metagenomic codes based on minimal hitting sets. Greedily select the best (rarest) remaining feature in the population until all other individuals have been knocked out; this procedure prioritizes small code size, but not code stability over time. Filled (black) dots represent detected features, while empty (white) dots represent absent features. Knocked-out subjects are colored gray.

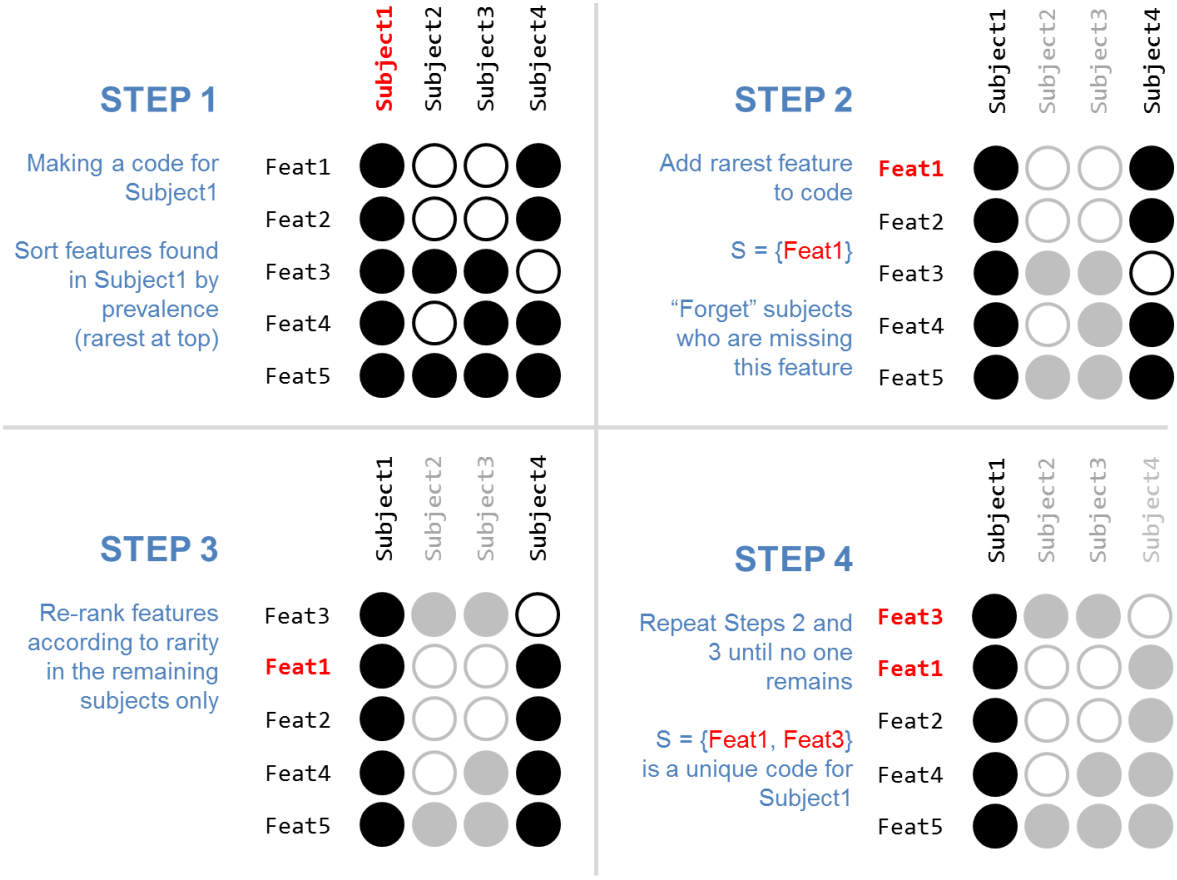

FIGURE S2: Quantifying technical variation. (A) For 325 replicate 16S metagenomes and 26 replicate shotgun metagenomes, we computed the probability of a feature being detected in one sample given that it was confidently detected in the paired sample. Replicate pairs were binned based on this measurement. Agreement was very strong for the majority of replicate pairs. **(B)** For technical replicates A and B corresponding to (body site X, subject Y, and time 1), we separately constructed codes for all site-X samples including only replicate A and again including only replicate B. We then compared the code derived for replicate A and the code derived for replicate B to all time-2 samples. If the results were precisely the same (e.g. both the A code and the B code only matched subject Y at time 2), then we scored the replicates as "consistent"; if there was any deviation then we scored the replicates as "inconsistent." An analogous procedure was repeated for technical replicates of time-2 samples. Coding results based on shotgun sequencing were very robust to technical variation.

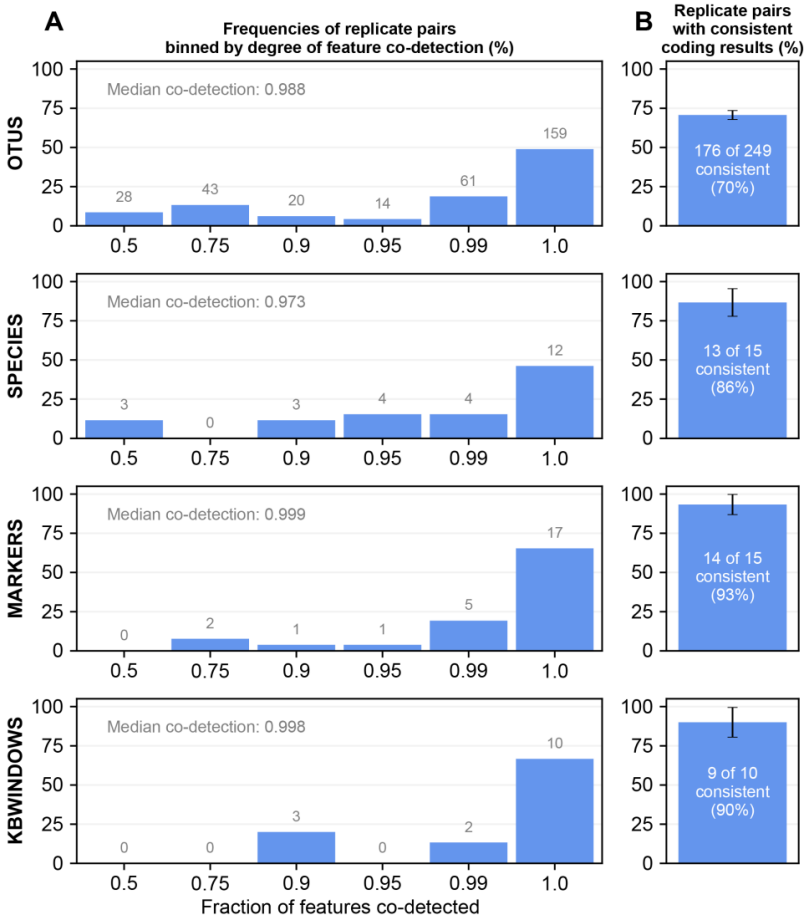

FIGURE S3: Example of a true positive marker gene-based code. For multi-visit subject 158499257, we were able to construct a unique marker gene-based code from stool composed of seven genes from four species. This marker-based code was stable over time (a true positive). Note that only a few marker genes from *Clostrium nexile* were detected, of which one was included in the code. These "orphaned" markers are most likely present in another genomic background (**Table S1**).

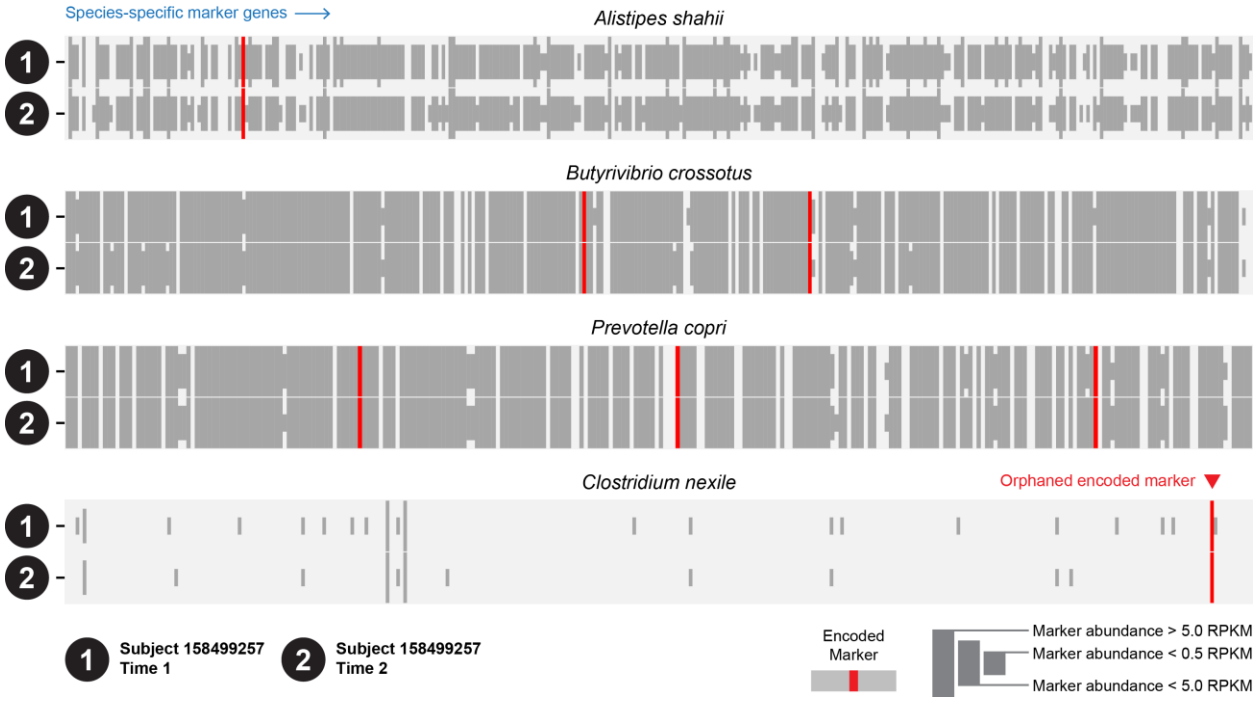

FIGURE S4: Example of a false negative marker gene-based code suggesting strain perturbation. For multi-visit subject 158499257, we were able to construct a marker gene-based code at the cheek (buccal mucosa) body site that was unique among all multi-visit subjects. This code (which was composed of seven genes from four species) was not robust to temporal variation between time 1 and time 2, resulting in a false negative. In this case, the false negative resulted from strain-level perturbation: encoded markers from *Neisseria sicca* and *Haemophilus parainfluenzae* were lost between time 1 and time 2, although the two species remained confidently detected.

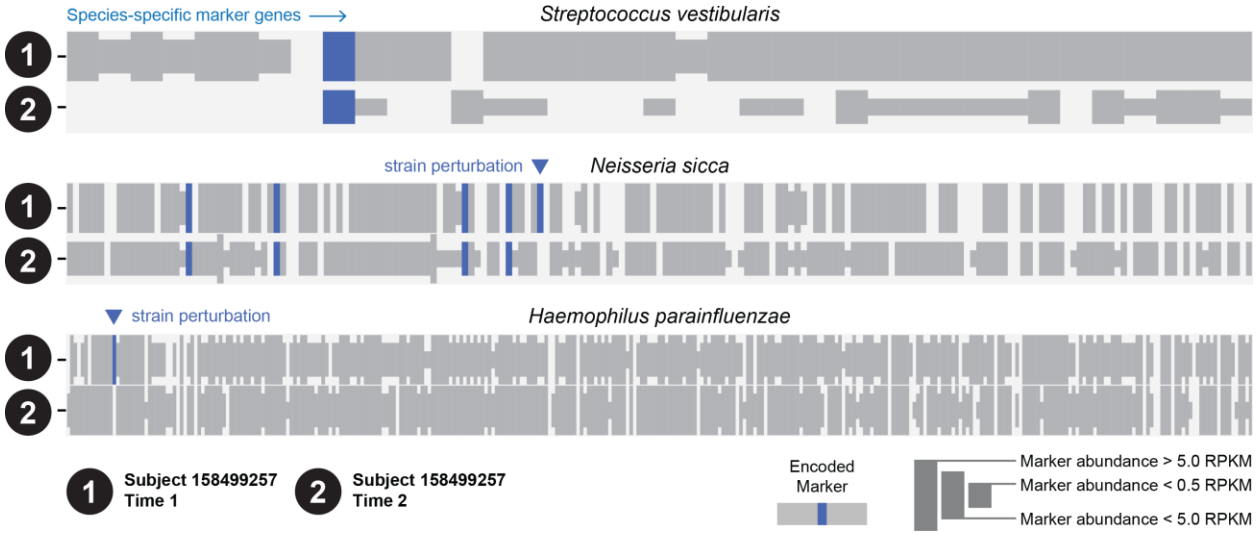

FIGURE S5: Example of a false positive marker-based code. For multi-visit subject 159268001, we were able to construct a marker gene-based code at the cheek (buccal mucosa) body site that was unique among all multi-visit subjects. This code included 10 marker genes from three species. This marker-based code yielded a false positive in our validation against single-visit subjects, specifically individual 160400887.

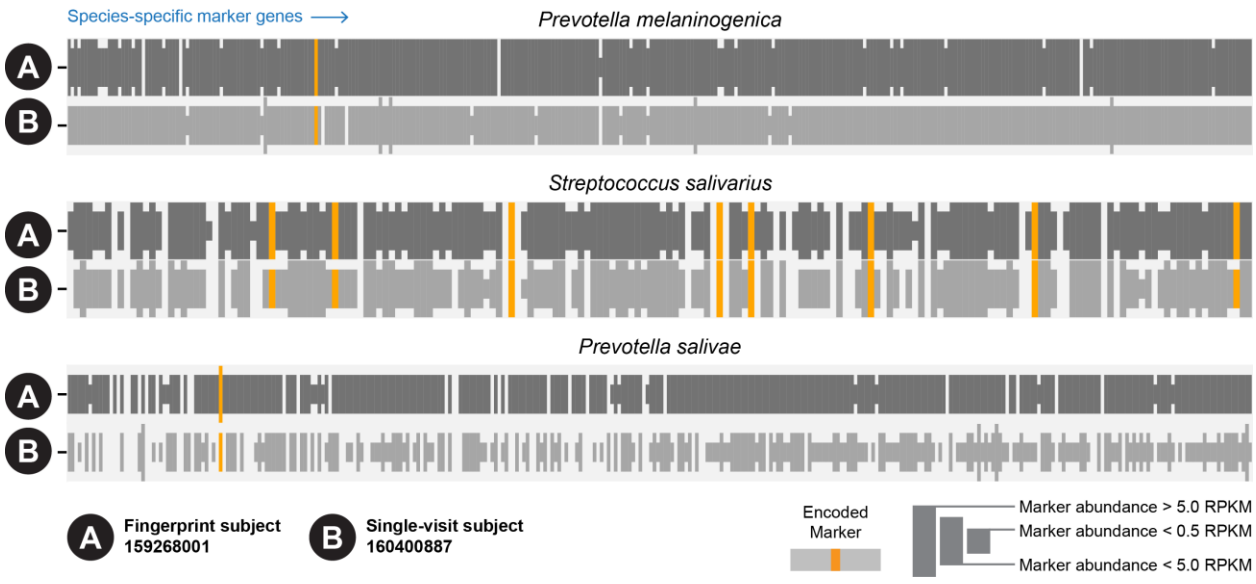

FIGURE S6: Effects of feature detection thresholds on metagenomic code performance. Panel (**A**) reproduces **Fig. 3A** from the main text for comparative purposes. In that case, stringent feature detection thresholds were used during code definition (>0.1% relative abundance for taxa and >5 RPKM for gene-level features) and relaxed detection thresholds were used during code evaluation (i.e. re-detection at later time points; >0.01% relative abundance for taxa and >0.5 RPKM for gene-level features). (**B**) Using relaxed detection thresholds in both code definition and evaluation produced more unique taxon-level codes, most of which produced false negatives in comparisons with later time points. Gene-level codes were minimally affected. (**C**) Using stringent detection thresholds in both code definition and evaluation systematically decreased the frequency of false positives, but at the expense of decreased sensitivity (fewer true positives).

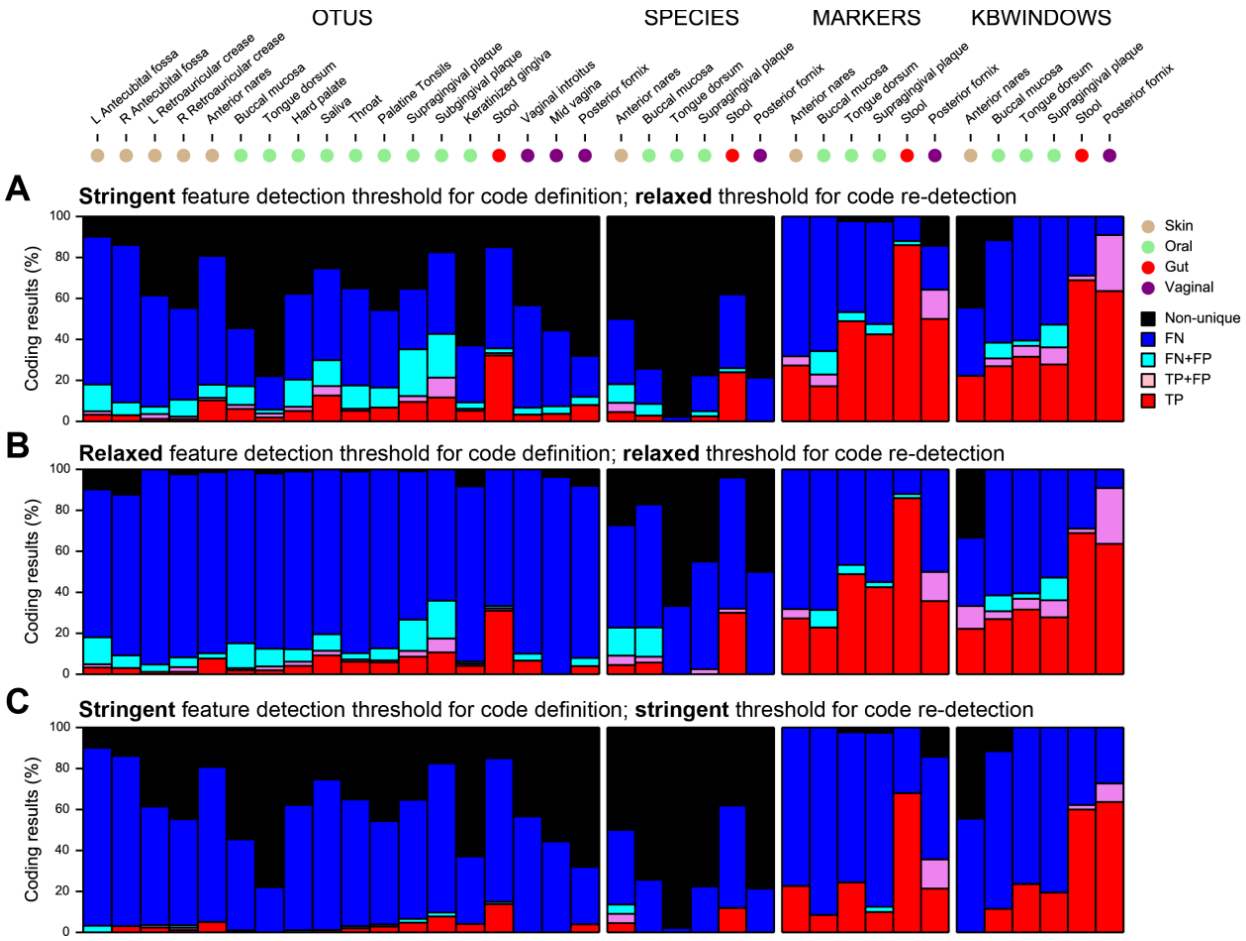

FIGURE S7. Evaluation of ecological distance-based identifiability. For each metagenomic feature type, we compared each second-visit sample to the collection of first-visit samples from the same body site using two ecological similarity measures: (**A**) Bray-Curtis distance and (**B**) Canberra distance. An individual was scored as a true positive (TP) if the closest match to his/her second-visit sample was his/her first-visit sample. If the individual's second-visit sample matched another individual's first-visit sample, then we scored this as a false negative (FN) and a false positive (FP): the former because the sample failed to match its owner, and the latter because it spuriously matched someone else. For both distance measures, false positive rates were considerably worse than those achieved using metagenomic codes (main text, **Fig. 3**).

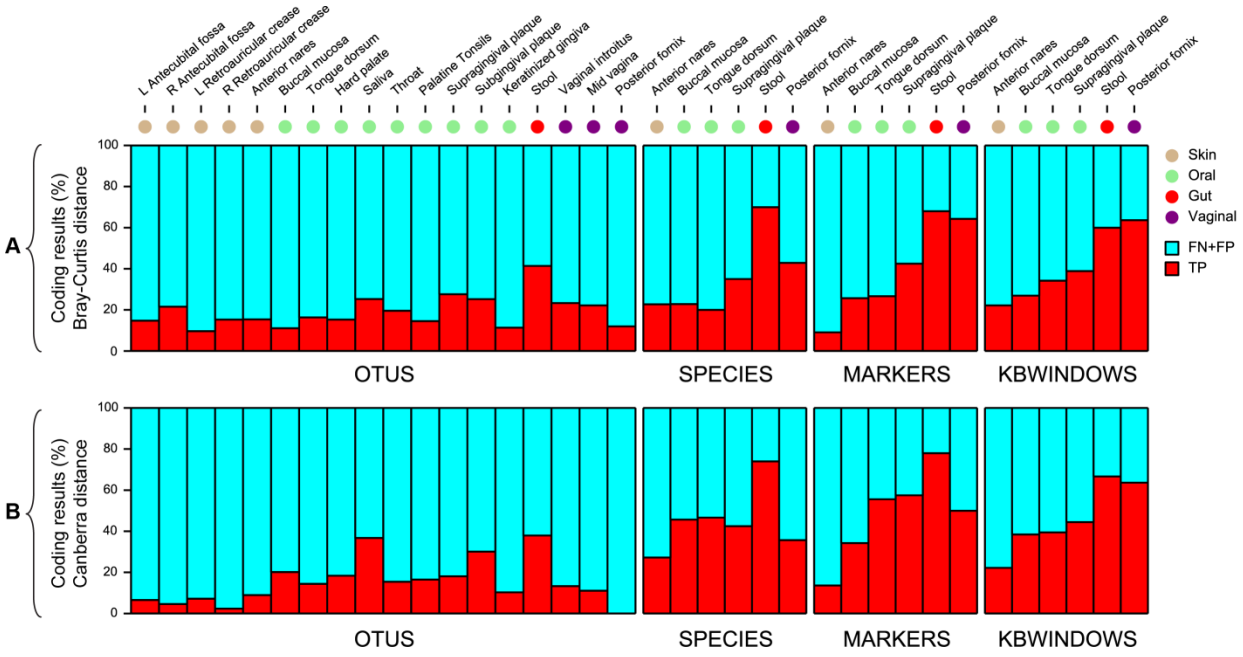

TABLE S1: Orphaned marker statistics suggestive of lateral transfer events. Orphaned markers are marker genes that were confidently detected in the absence of their parent organism. Although orphaned markers made up only 1-2% of total confidently detected markers, they made up 6-24% of markers encoded in marker-based codes, a statistically significant enrichment (Fisher's exact test, two-tailed $P<0.001$). Orphaned markers tended to be slightly less stable than other marker genes, but the difference was only significant at the anterior nares body site ($P=0.011$).

Value	Anterior nares	Buccal mucosa	Posterior fornix	Stool	Supra-gingival plaque	Tongue dorsum
# of markers detected	18,326	33,935	6,470	44,743	67,385	75,832
# of markers orphaned	738	298	153	907	160	542
Fraction of detected markers orphaned	0.040	0.009	0.024	0.020	0.002	0.007
# of markers encoded	161	258	83	356	299	321
# of markers encoded and orphaned	38	17	9	55	18	29
Fraction of encoded markers orphaned	0.236	0.066	0.108	0.154	0.060	0.090
Enrichment for orphaned markers in codes	5.9	7.5	4.6	7.6	25.4	12.6
P-value	4.26E-19	1.70E-10	0.000135	1.01E-31	4.12E-20	6.08E-23
# of encoded markers stable	105	178	80	338	248	272
Fraction of encoded markers stable	0.652	0.690	0.964	0.949	0.829	0.847
# of encoded orphaned markers stable	18	9	8	51	14	28
Fraction of encoded orphaned markers stable	0.474	0.529	0.889	0.927	0.778	0.966
Enrichment for stability in encoded orphaned markers	0.726	0.767	0.922	0.977	0.938	1.140
P-value	0.011	0.174	0.294	0.498	0.522	0.099

TABLE S2: Strain perturbation statistics. Marker gene-based codes were not always robust against changes in the strain-level composition of a microbiome. 825 total markers were encoded from species that were very confidently detected in a sample (i.e. more than half of the species' marker genes were detected with abundance >5 RPKM). Of these, 70 were lost between time 1 and time 2 (8%). For 25 of these cases (36%), the marker's parent species was still detected at time 2, indicative of perturbation at the strain level. Such perturbations could include (i) the deletion of the marker gene from a subject's time-1 strain of the species, (ii) the replacement of a subject's time-1 strain by another strain, or (iii) the loss of one of two or more time-1 strains of the same species. Of 67 marker gene-based codes that failed to match their owner at time 2, 17 (25%) involved the loss of a marker gene without the loss of its parent species (see **Fig. S4**).

Value	Anterior nares	Buccal mucosa	Posterior fornix	Stool	Supra-gingival plaque	Tongue dorsum
# of markers encoded	161	258	83	356	299	321
# of markers encoded with parent species	82	139	72	185	224	123
# of markers encoded with parent species, then lost	12	25	2	4	20	7
# of markers encoded with parent species, then lost, parent remains	1	10	1	0	8	5
Fraction of encoded marker loss not due to parent species loss	0.083	0.400	0.500	0.000	0.400	0.714
# of marker-level codes	22	35	12	50	39	44
False negative (FN) matches	13	22	3	5	10	14
FNs involving marker loss without loss of parent species	1	7	1	0	4	4

TABLE S3: Impact of antibiotics (Abx) use on code false negatives. Abx use was very low among individuals in our focal population. We noted a slight enrichment for Abx use among individuals whose OTU-based stool codes produced false negatives at the second time point, but the enrichment was not statistically significant (Fisher's exact test, two-tailed $P>0.05$). Fold enrichment represents the ratio of Abx use among individuals with true positive codes relative to the population average. Fold enrichment <1 indicates that Abx use was associated with codes failing to match their owner at the later time point (i.e. enrichment for false negatives).

Feature Type	Body Site	TP + Abx	TP total	FN + Abx	FN Total	Fold Enrich	P-value
OTUs	Anterior nares	3	9	7	54	2.10	0.15
OTUs	Buccal mucosa	0	8	3	37	0.00	1.00
OTUs	Hard palate	1	7	6	53	1.22	1.00
OTUs	Keratinized gingiva	0	6	4	31	0.00	1.00
OTUs	L Antecubital fossa	0	3	7	52	0.00	1.00
OTUs	L Retroauricular crease	0	3	5	48	0.00	1.00
OTUs	Mid vagina	0	1	0	11	0.00	1.00
OTUs	Palatine Tonsils	0	7	3	48	0.00	1.00
OTUs	Posterior fornix	0	2	0	6	0.00	1.00
OTUs	R Antecubital fossa	0	2	8	54	0.00	1.00
OTUs	R Retroauricular crease	0	2	5	45	0.00	1.00
OTUs	Saliva	0	14	7	53	0.00	0.33
OTUs	Stool	1	31	7	45	0.31	0.13
OTUs	Subgingival plaque	1	22	5	61	0.63	1.00
OTUs	Supragingival plaque	1	13	5	55	0.87	1.00
OTUs	Throat	0	6	6	58	0.00	1.00
OTUs	Tongue dorsum	1	4	1	19	2.87	0.32
OTUs	Vaginal introitus	0	1	1	16	0.00	1.00
Species	Anterior nares	0	2	0	9	0.00	1.00
Species	Buccal mucosa	0	1	1	8	0.00	1.00
Species	Posterior fornix	0	0	0	3	0.00	1.00
Species	Stool	0	12	2	19	0.00	0.51
Species	Supragingival plaque	0	1	0	8	0.00	1.00
Species	Tongue dorsum	0	0	0	1	0.00	1.00
Markers	Anterior nares	0	7	0	15	0.00	1.00
Markers	Buccal mucosa	1	8	1	27	2.19	0.41
Markers	Posterior fornix	1	9	0	3	1.33	1.00
Markers	Stool	1	43	1	7	0.58	0.26
Markers	Supragingival plaque	1	17	0	22	2.29	0.44
Markers	Tongue dorsum	0	22	1	22	0.00	1.00
Kbwindows	Anterior nares	0	2	0	3	0.00	1.00
Kbwindows	Buccal mucosa	1	8	0	15	2.87	0.35
Kbwindows	Posterior fornix	1	10	0	1	1.10	1.00
Kbwindows	Stool	2	32	0	13	1.41	1.00
Kbwindows	Supragingival plaque	0	13	1	23	0.00	1.00
Kbwindows	Tongue dorsum	0	14	1	24	0.00	1.00

TABLE S4: Evaluation of code uniqueness in comparison with new populations (part 1). **"**New subjects" refers to individuals who were sampled during the HMP but only at one time point. As a result, these individuals were excluded from our focal population, which consisted of individuals sampled on at least two occasions.

Feature type	Body site	# of Finger-prints	# Hitting	# of new subjects	# Hit	# Total hits	Hit chance (p)	Unique among N (1/p)	Standard error of N
OTUs	Anterior nares	63	7	84	11	13	0.0025	407	113
OTUs	Buccal mucosa	45	7	82	7	7	0.0019	527	199
OTUs	Hard palate	60	7	77	8	8	0.0017	578	204
OTUs	Keratinized gingiva	37	10	83	10	15	0.0049	205	53
OTUs	L Antecubital fossa	55	2	91	2	3	0.0006	1670	963
OTUs	L Retroauricular crease	51	6	97	10	10	0.0020	495	156
OTUs	Mid vagina	12	0	46	0	0	0.0018	553	553
OTUs	Palatine Tonsils	55	8	80	8	9	0.0021	489	163
OTUs	Posterior fornix	8	1	50	3	3	0.0075	133	77
OTUs	R Antecubital fossa	56	7	95	6	9	0.0017	591	197
OTUs	R Retroauricular crease	47	3	102	6	8	0.0017	599	212
OTUs	Saliva	67	6	78	6	7	0.0013	747	282
OTUs	Stool	76	9	101	8	10	0.0013	768	243
OTUs	Subgingival plaque	83	13	76	9	14	0.0022	451	120
OTUs	Supragingival plaque	68	9	76	11	13	0.0025	398	110
OTUs	Throat	64	12	81	12	16	0.0031	324	81
OTUs	Tongue dorsum	23	5	82	6	6	0.0032	314	128
OTUs	Vaginal introitus	17	3	42	7	8	0.0112	89	31
Species	Anterior nares	11	6	36	9	9	0.0227	44	15
Species	Buccal mucosa	9	0	32	0	0	0.0035	289	289
Species	Posterior fornix	3	0	16	0	0	0.0204	49	49
Species	Stool	31	5	43	5	5	0.0038	267	119
Species	Supragingival plaque	9	1	35	1	1	0.0032	315	314
Species	Tongue dorsum	1	0	37	0	0	0.0263	38	38
Markers	Anterior nares	22	6	35	5	7	0.0091	110	41
Markers	Buccal mucosa	35	6	32	9	10	0.0089	112	35
Markers	Posterior fornix	12	3	16	3	3	0.0156	64	37
Markers	Stool	50	7	42	5	8	0.0038	262	93
Markers	Supragingival plaque	39	3	35	3	3	0.0022	455	262
Markers	Tongue dorsum	44	2	37	2	2	0.0012	814	575
Kbwindows	Anterior nares	5	2	20	7	8	0.0800	13	4
Kbwindows	Buccal mucosa	23	2	37	2	2	0.0024	426	301
Kbwindows	Posterior fornix	11	7	17	7	12	0.0642	16	4
Kbwindows	Stool	45	9	42	9	10	0.0053	189	60
Kbwindows	Supragingival plaque	36	6	36	6	7	0.0054	185	70
Kbwindows	Tongue dorsum	38	2	39	2	2	0.0014	741	524

TABLE S5: Evaluation of code uniqueness in comparison with new populations (part 2). For each body site and feature type, we modeled the probability of a code spuriously matching a member of a new population (false positive rate, FPR) as a one-parameter (k) exponential function of population size, N. From this function, we estimated the size of a population in which a code would have a 50% chance of a spurious match (N_{50}). "NA" indicates that no spurious matches were observed between codes and members of the validation cohort, in which case it is not possible to produce a meaningful model fit.

Feature type	Body site	Estimated FPR growth parameter (k)	Estimated N50
OTUs	Anterior nares	0.00118	585
OTUs	Buccal mucosa	0.00137	504
OTUs	Hard palate	0.00132	526
OTUs	Keratinized gingiva	0.00208	332
OTUs	L Antecubital fossa	0.00034	2018
OTUs	L Retroauricular crease	0.00109	635
OTUs	Mid vagina	NA	NA
OTUs	Palatine Tonsils	0.00119	583
OTUs	Posterior fornix	0.00325	213
OTUs	R Antecubital fossa	0.00110	629
OTUs	R Retroauricular crease	0.00073	955
OTUs	Saliva	0.00166	418
OTUs	Stool	0.00103	673
OTUs	Subgingival plaque	0.00150	462
OTUs	Supragingival plaque	0.00155	445
OTUs	Throat	0.00202	342
OTUs	Tongue dorsum	0.00217	320
OTUs	Vaginal introitus	0.00486	142
Species	Anterior nares	0.01589	43
Species	Buccal mucosa	NA	NA
Species	Posterior fornix	NA	NA
Species	Stool	0.00269	257
Species	Supragingival plaque	0.00215	322
Species	Tongue dorsum	NA	NA
Markers	Anterior nares	0.00638	108
Markers	Buccal mucosa	0.00487	142
Markers	Posterior fornix	0.01085	63
Markers	Stool	0.00255	272
Markers	Supragingival plaque	0.00155	446
Markers	Tongue dorsum	0.00086	804
Kbwindows	Anterior nares	0.02607	26
Kbwindows	Buccal mucosa	0.00166	418
Kbwindows	Posterior fornix	0.04389	15
Kbwindows	Stool	0.00372	186
Kbwindows	Supragingival plaque	0.00363	190
Kbwindows	Tongue dorsum	0.00092	753

SUPPORTING REFERENCES

1. The Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. *Nature* 486(7402):207-214.
2. Arumugam M, *et al.* (2011) Enterotypes of the human gut microbiome. *Nature* 473(7346):174-180.
3. Segata N, *et al.* (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* 9(8):811-814.