nature methods
ARTICLES
https://doi.org/10.1038/s41592-022-01520-4

Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data

Kristen D. Curry[1]✉, Qi Wang[2], Michael G. Nute[1], Alona Tyshaieva[3], Elizabeth Reeves[1], Sirena Soriano[4], Qinglong Wu[5,6], Enid Graeber[3], Patrick Finzer[3], Werner Mendling[7], Tor Savidge[5,6], Sonia Villapol[4], Alexander Dilthey[3,8]✉ and Todd J. Treangen[1,8]✉

16S ribosomal RNA-based analysis is the established standard for elucidating the composition of microbial communities. While short-read 16S rRNA analyses are largely confined to genus-level resolution at best, given that only a portion of the gene is sequenced, full-length 16S rRNA gene amplicon sequences have the potential to provide species-level accuracy. However, existing taxonomic identification algorithms are not optimized for the increased read length and error rate often observed in long-read data. Here we present Emu, an approach that uses an expectation-maximization algorithm to generate taxonomic abundance profiles from full-length 16S rRNA reads. Results produced from simulated datasets and mock communities show that Emu is capable of accurate microbial community profiling while obtaining fewer false positives and false negatives than alternative methods. Additionally, we illustrate a real-world application of Emu by comparing clinical sample composition estimates generated by an established whole-genome shotgun sequencing workflow with those returned by full-length 16S rRNA gene sequences processed with Emu.

Sequencing of the 16S subunit of the ribosomal RNA (rRNA) gene has been a reliable way to characterize diversity in a community of microbes since Carl Woese used this technique to identify Archaea in 1977 (ref. [1]). Today, high-throughput sequencing machines used for this analysis are dominantly Illumina devices. Although they are cost-effective and accurate, Illumina sequences are limited to approximately 500 nucleotides per joined paired-end read. Given that the 16S rRNA gene is approximately 1,550 bp, 16S rRNA-targeted amplification sequencing is limited to only a portion of the gene and is completed by targeting a selected subset of the nine hypervariable regions. This constraint ultimately prevents distinction between highly similar species and therefore short-read data can reliably generate taxonomic profiles measured down to only the genus level in most cases[2]. One workaround for this limitation is to assemble short reads through a synthetic long-read method[3]; another is sample specific barcoding[4]. Although these approaches have high accuracy, they require significantly more sequences or sample-specific library preparation, which introduces additional financial costs.

Recent developments in third-generation sequencing, from providers such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), permit amplification of sequences spanning the entire 16S rRNA gene. However, these long reads come with one notable drawback: high rates of sequencing error[5]. Errors can be corrected by deducing a consensus sequence from multiple passes on each strand of genetic material as seen in PacBio HiFi[6], or from multiple reads tagged with matching unique molecular identifiers[7]. Although these methods have produced near-perfect accuracy[8] they again come with a significant increase in cost due to increased sequencing depth. To reduce these expenses and achieve species-level resolution from single-pass 16S rRNA reads, the appropriate software to account for high error profiles is needed.

The canonical pipeline for 16S rRNA analysis operates in two main steps. First, the set of raw sequences is de-noised to identify a smaller set of core sequences, in which each set is believed to represent a distinct taxonomic unit in the community. Various algorithms are available for this process[9], however, the majority are calibrated to the low level of error associated with Illumina reads. Second, the representative sequences are compared to a database and assigned a taxonomic label. Given that reads are already corrected for error at this point, a database-lookup tool such as BLAST[10] is effective here. These pipelines are operative because the input reads are accurate, and unfortunately they produce inconsistent results when challenged with error-prone reads.

Given that the ONT sequencers are comparatively recent to the marketplace, 16S rRNA method development for these devices has only just begun[11]. In the absence of dedicated tools, some studies have chosen to use more general read mapping software such as BWA-MEM[12] or the LAST aligner[13] to align reads directly to raw 16S rRNA sequences from one of the major databases[14–16], while other studies have chosen to incorporate metagenomic classification methods designed for whole-genome shotgun sequences. Centrifuge[17] proved capable of ONT shotgun sequencing analysis

[1]Department of Computer Science, Rice University, Houston, TX, USA. [2]Department of Systems, Synthetic and Physical Biology Science, Rice University, Houston, TX, USA. [3]Institute of Medical Microbiology and Hospital Hygiene, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. [4]Houston Methodist Research Institute, Center for Neuroregeneration, Houston, TX, USA. [5]Department of Pathology and Immunology, Baylor College of Medicine, Houston, TX, USA. [6]Texas Children's Microbiome Center, Department of Pathology, Texas Children's Hospital, Houston, Texas, USA. [7]German Centre for Infections in Gynaecology and Obstetrics at Helios University Clinic Wuppertal, Wuppertal, Germany. [8]These authors contributed equally: Alexander Dilthey, Todd J. Treangen. ✉e-mail: kristen.d.curry@rice.edu; alexander.dilthey@med.uni-duesseldorf.de; treangen@rice.edu

and is now included as a step in ONT's WIMP[18] (What's in my Pot?), a long-read workflow provided on its EPI2ME analysis platform. Kraken[19] has also fared well in long-read 16S rRNA benchmarking[20] and has performed favorably compared with QIIME 2 (ref. [21]) for 16S rRNA short reads when the results are re-estimated with its Bayesian cousin Bracken[22,23].

NanoClust[24] was the first published purpose-built method for taxonomic abundance profiling using full-length 16S rRNA amplicon sequencing from ONT devices. Here, the two-stage cluster and database-lookup procedure is implemented in Nextflow[25] with external tools for demultiplexing, quality filtering, clustering, polishing and taxon assignment. Although the use of clustered consensus sequences increases computational efficiency, this approach is susceptible to overlooking the identification of species that are truly present in the error-prone dataset.

One method that has successfully overcome high error rates in long-read data is MetaMaps[26], a method designed for taxonomic binning of long, high-error shotgun sequencing reads. MetaMaps uses an approximate read mapping algorithm to identify multiple candidate species and locations for each read, then applies an expectation–maximization (EM) algorithm to adjust the relative confidence in each mapping based on the mapping density of other reads in the sample. This has the effect of smoothing out some of the noise that is inherently created by the ONT error profile. Although the approximation methods built into MetaMaps make it incompatible for the analysis of highly similar 16S rRNA genes, it and other EM algorithms that have successfully disambiguated ambiguous read mappings[27,28] provoke interest in an EM method for error correction of long 16S rRNA reads.

Here, we present Emu, a microbial community profiling software tool tailored for full-length 16S rRNA data with high error rates. Emu benefits from the increased precision potential provided by the full gene while accounting for high error rates produced by single-pass third-generation sequencing. Emu's algorithm involves a two-stage process. First, proper alignments are generated between the reads and the supplied reference database. In the second stage an EM-based error-correction step is performed to iteratively refine species-level relative abundances based on total read mapping counts. This results in microbial community profile estimations from full-length 16S reads that are more accurate than existing methods at both the genus and species level.

Results

To generate an accurate microbial community composition estimate from noisy full-length reads sequenced from the16S rRNA gene, an EM algorithm with a composition-dependent prior is developed in Emu. The algorithm is shown in Fig. 1 and a more detailed version of the algorithm (including the equations used) is shown in Extended Data Fig. 1.

To demonstrate the performance of Emu, four studies were completed. The first study was a quantitative comparison of two distinct sets of simulated ONT sequences (Supplementary Tables 1 and 2). The second study was a quantitative comparison of two distinct communities sequenced with both ONT and Illumina devices (Supplementary Tables 3 and 4), for which a de facto ground truth could be used to evaluate accuracy and compare methods. The third study was a series of analyses highlighting the various facets of Emu via a breakdown of profile estimations throughout the EM algorithm, a database comparison, a novel species simulation, a read mapper comparison and a naive application of the Emu default minimum abundance threshold. The final study was a demonstration of Emu's applicability to understanding dynamics in actual microbial communities. In this real-world model, human vaginal microbiome clinical samples were processed with two separate pipelines: 16S rRNA long reads analyzed with Emu and whole-genome shotgun sequences processed with Bracken.

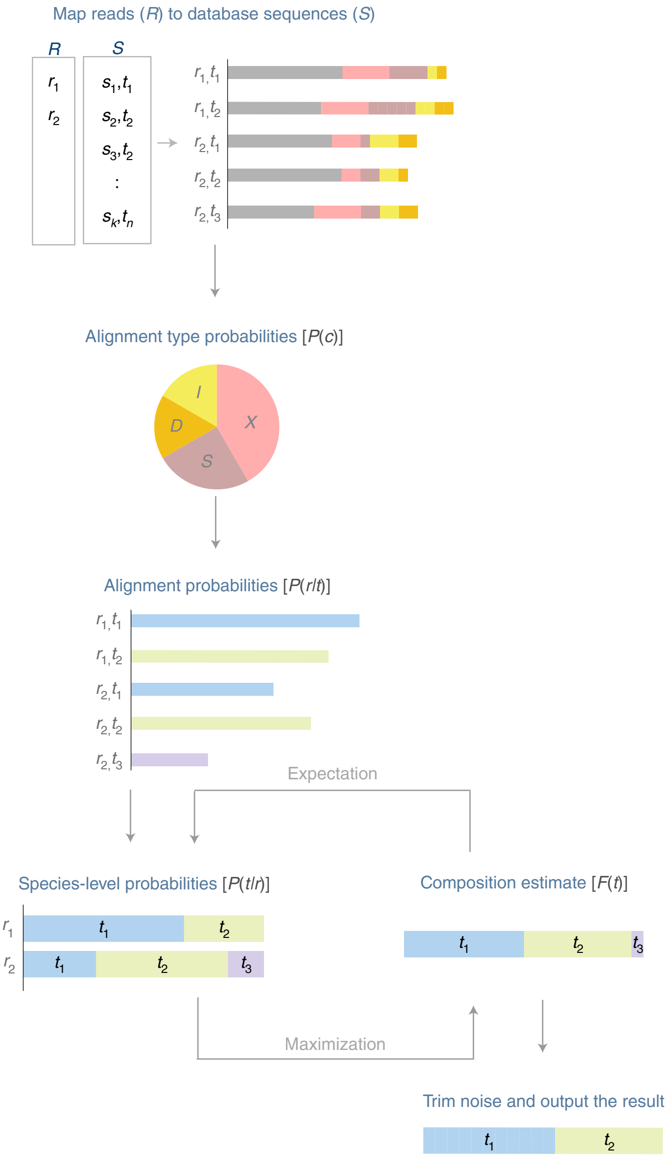


Fig. 1 | The Emu algorithm. The Emu algorithm begins by generating alignments between input reads (*R*) and database sequences (*S*). The probability of each non-matching character alignment type (mismatch (*X*), insertion (*I*), deletion (*D*), softclip (*S*)) is calculated based on the number of occurrences of each character alignment type in all of the primary alignments from the read mapping. The probability of each alignment in the read mapping is then generated as *P*(*r*|*t*) from the counts of each character alignment type and their corresponding established probabilities. The EM phase is then entered, in which each read is broken down into the likelihood that it is derived from each possible species in the database *P*(*t*|*r*), and its overall composition estimate *F*(*t*) (which is deduced). This cycle repeats as the composition estimate influences read-taxonomy probabilities to give more weight to taxa with higher abundances, then the composition estimate is updated accordingly. Once minimal changes are detected between cycle iterations, the EM loop is exited. The composition estimate is then trimmed based on the specified minimum abundance probability threshold to complete one final EM iteration and the final composition estimate is produced.

Quantitative comparison. To quantify the output of Emu in relation to several existing methods, four communities were used. The first two are single datasets of simulated ONT reads that follow the distribution of a published mock community. The other two are synthetic mock communities, each of which was sequenced with both

Table 1 | Performance summary of 16S rRNA gene relative abundance estimates on ONT and Illumina sequences for all four communities

		ONT								Illumina				
		Emu	minimap2	Kraken 2	Bracken	NanoCLUST	Centrifuge	MetaMaps	QIIME 2	Emu	minimap2	Kraken 2	Bracken	QIIME 2
MBARC-26														
Genus	L1-norm	2E−5	3E−4	0.03	0.72	1.77	0.01	0.42	0.16					
	TP (24)	24	24	24	19	4	24	24	22					
	FP	0	18	339	15	0	484	229	8					
Species	L1-norm	3E−05	0.01	0.14	0.80	1.77	0.11	0.51	0.96					
	TP (26)	26	26	25	19	5	26	26	16					
	FP	0	73	626	43	0	860	415	28					
CAMI2														
Genus	L1-norm	0.01	0.02	0.13	1.12	-	0.10	0.05	-					
	TP (179)	171	179	176	105	-	180	177	-					
	FP	69	1482	2134	542	-	2296	1366	-					
Species	L1-norm	0.03	0.13	0.43	1.45	-	0.24	0.11	-					
	TP (345)	330	343	338	162	-	343	339	-					
	FP	250	4665	5879	1175	-	6780	3271	-					
ZymoBIOMICS mock community														
Genus	L1-norm	1E−3	0.01	0.25	0.24	0.02	0.45	0.16	0.75	0.04	0.07	0.39	0.30	0.31
	TP (8,8)	8	8	8	8	8	8	8	7	8	8	8	8	7
	FP	0	5	77	61	0	245	49	5	9	27	65	64	1
Species	L1-norm	0.03	0.18	0.65	0.66	0.24	1.11	0.61	1.14	0.34	0.39	1.18	0.89	0.93
	TP (8,8)	8	8	8	8	8	8	8	1	8	8	8	8	1
	FP	6	45	219	191	1	480	146	15	81	120	262	195	2
Synthetic gut microbiome mock community														
Genus	L1-norm	0.03	0.05	0.41	0.79	0.34	0.30	0.53	0.67	0.34	0.35	0.90	0.77	0.45
	TP (18,19)	18	18	18	14	14	18	17	14	17	18	17	16	14
	FP	15	93	505	80	2	1007	381	31	14	557	291	59	3
Species	L1-norm	0.43	0.44	0.76	0.83	0.50	0.66	0.74	1.17	0.51	0.54	1.26	1.20	1.13
	TP (20,21)	18	19	19	12	14	19	17	9	16	18	15	12	6
	FP	40	252	1156	166	4	2372	836	55	71	1230	539	110	4

FP, number of false positives; L1-norm, linear error; TP, number of true positives ((*n*), estimated number of actual true positives for each dataset; (*n*,*n*), expected TP for the ONT and Illumina datasets, respectively).

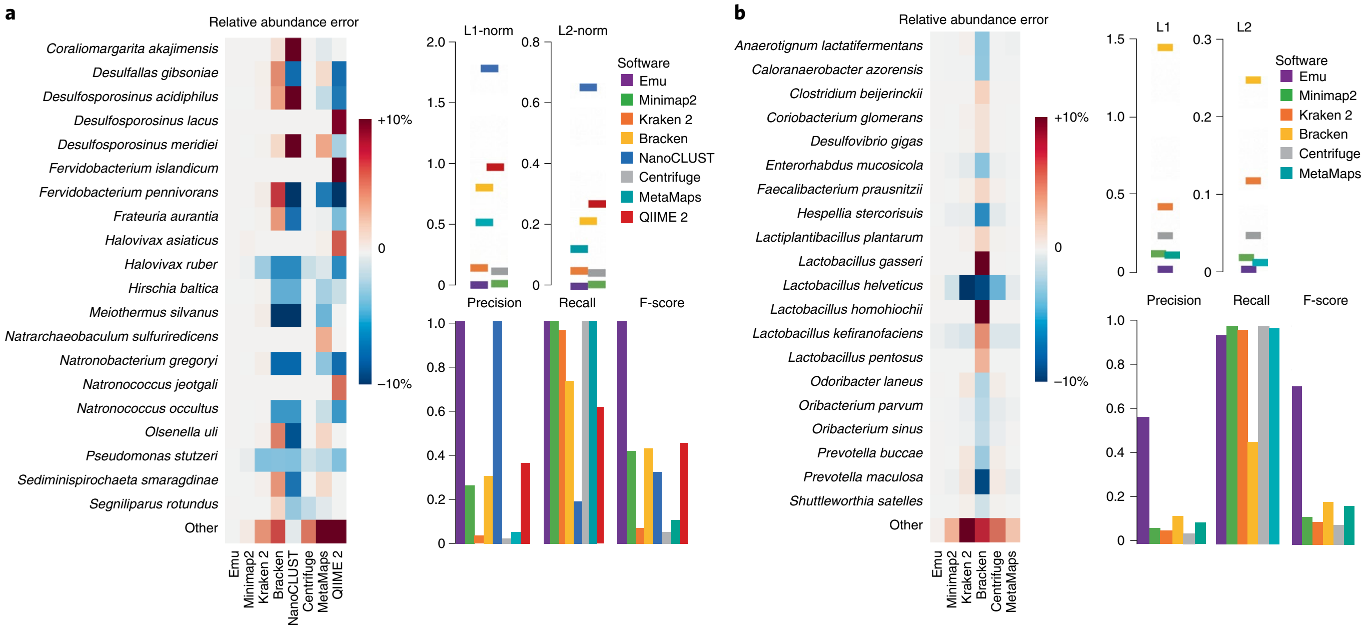


Fig. 2 | Performance on simulated ONT reads. a, Quantitative result statistics for our MBARC-26 simulated dataset. A heatmap is shown of species-level error between expected and inferred relative abundances, where darker blue denotes an underestimate by the software, darker red denotes an overestimate, and white represents no error. The color scheme is capped at ±10, meaning that errors greater than ±10% will be shown in the maximum error colors. Displayed are the 20 species claiming the largest abundance in any of the included results. 'Other' represents the sum of all species not shown in the figure for the respective column. Species-level L1-norm, L2-norm, precision, recall and F-score are also plotted for the methods evaluated. **b**, The same statistics shown in **a** for our CAMI2 simulated dataset.

ONT and Illumina devices. The performance of each method was evaluated at both the genus and species level using three metrics: the L1-norm (that is, linear error) of the taxonomic abundance profile, the count of true-positive taxa, and the count of false-positive taxa. Computational resources required by each method were measured by recording the run time and memory usage for each software.

The set of methods used for comparison includes those discussed above: Kraken 2 (ref. [29]), Bracken, NanoClust, Centrifuge and MetaMaps. We also include QIIME 2 and the primary alignment generated by minimap2 (ref. [30]). Although minimap2 is not a composition estimator or read-level classifier in itself, it is included because it is instrumental in the Emu algorithm: minimap2 is the read mapping software that Emu uses to compute likelihood scores and iteratively estimate relative abundance. Inclusion of minimap2 in the comparison separates the effect of the EM implementation in Emu from the read mapping output it uses as a starting point. Identical reference databases were built for each software to ensure consistent comparison for all of the methods (see Methods for details).

Ground truth relative abundance values for synthetic communities are based on sample-specific imputed values. This was done to correct for fluctuations from the theoretical abundances that may occur during handling, storage or library preparation (including potential primer bias during 16S rRNA gene amplification). The two ZymoBIOMICS community profiles are reasonably similar to their abundance claims (Extended Data Fig. 2 and Supplementary Table 5), but the synthetic gut community is subject to greater variation by nature of the microbes included and the skewed distribution. Details on this process are given in the Establishing Ground Truth section in the Methods, and for these two communities the term 'ground truth' here refers to this imputed value.

Simulated mock community datasets. The ONT reads were simulated following the composition of the published mock communities MBARC-26 (ref. [31]) and the mouse gut profile from the Critical Assessment of Metagenome Interpretation II (CAMI2) Challenge[32]. MBARC-26 represents a simple community with 23 bacterial and 3 archaeal strains, while our subsampled version of the CAMI2 profile increases microbial richness and contains 345 unique species, each of which is present in the Emu default database. Detailed information on the reference sequences and distribution of simulated reads is contained in Supplementary Tables 1 and 2.

ZymoBIOMICS mock community standard dataset. A previous study compared 16S rRNA sample composition accuracy across a series of hypervariable regions as well as the full-length gene using the ZymoBIOMICS community standard catalog D3605 (ref. [33]). We retrieved the ONT full-length dataset and one of the Illumina datasets for our analysis. We selected the Illumina dataset with targeted regions V4–V6 to represent short-read data, given that this dataset has been shown in the previous study[33] to produce classification results that were among the most accurate for this specific community.

Synthetic mock gut microbiome dataset. To challenge our software, a synthetic community mimicking the human gut microbiome was created and sequenced with both ONT and Illumina devices, as described in the Creation of Gut Microbiome Mock Community in the Methods. To represent a real-world scenario with unknown species, *Romboutsia hominis* is included in the sample, even though this new species is not present in our database. The derived relative abundances of 21 species present in the sample are listed in Supplementary Table 4. One notable difference between the two datasets for this community is that *Bifidobacterium dentium* is not considered to be a true positive in the ONT sequences. This is a result of a recently noted issue with the standard ONT forward primer, which contains three mismatching bases to the family *Bifidobacteriaceae* and thus fails to amplify microbes of this taxa[15]. Consequently, the ONT dataset does not contain reads from this

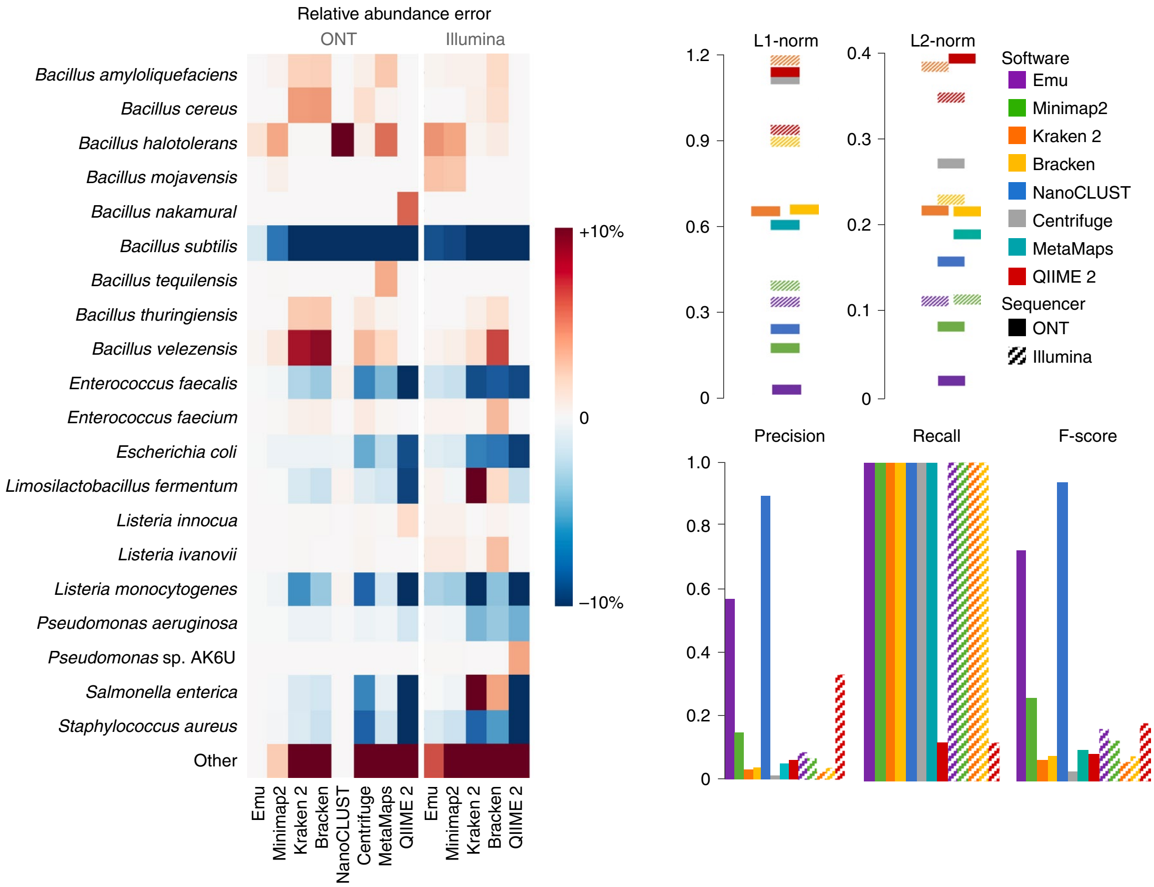


Fig. 3 | Performance on our ZymoBIOMICS community standard dataset. Heatmap of species-level error between the calculated ground truth and estimated relative abundances, where darker blue denotes an underestimate by the software, darker red denotes an overestimate, and white represents no error. All ONT errors are measured in relation to the ground truth of the ONT dataset, while Illumina errors are measured in relation to the ground truth for the Illumina dataset. The color scheme is capped at ±10, meaning that errors greater than ±10% will be shown in the maximum error colors. Displayed are the 20 species claiming the largest abundance in any of the ONT or Illumina results. 'Other' represents the sum of all species not shown in the figure for the respective column. Species-level L1-norm, L2-norm, precision, recall and F-score are also plotted for the methods evaluated. The true- and false-positive counts used to calculate precision, recall and F-score are restricted to species with relative abundance ≥0.01% to align with guidance from ZymoBIOMICS on maximum levels of contamination.

microbe. This demonstrates the importance of an imputed ground for the mock communities and additionally highlights the need for further research to identify reliably universal primers for this region.

Performance. The results of all of the methods on the simulated dataset and two synthetic mock communities are listed in Table 1. The computational resources required by each method are listed in Supplementary Table 6. Complete abundance profile output from all methods on all datasets is provided in Supplementary Tables 7–12. All of the generated results use the default Emu database.

MBARC-26 simulation. For the MBARC-26 simulated data, Emu outperforms every method (Fig. 1). Not only does Emu produce the lowest L1 and L2 distances, but it is also the only method to correctly identify all 26 species without producing any false positives. The difference in both the false-positive counts and the relative abundance error measurements between Emu and minimap2 is substantial, and reflects the accuracy gains produced by the EM algorithm compared with a simple similarity-based taxon assignment approach. It is evident from the memory and run time data between these two methods (Supplementary Table 6) that the majority of computational resources used by Emu are in fact due to its use of minimap2 for alignment generation. NanoCLUST results differ from the other methods shown in that it has no false positives but it fails to identify several of the present taxa; in other words, it is generally conservative in its identifications.

CAMI2 simulation. Results for the more complex simulated data shown in the CAMI2 dataset align with those reported from our first simulated dataset. Again, Emu reports the lowest relative abundance error through both the L1 and L2 distances. In addition, Emu reports the best precision and F-score due to its ability to find a balance between true-positive detection and false-positive reports. Bracken's re-estimated Kraken 2 results cause both the true and false positives to drop off significantly. However, Emu's re-estimation of minimap2 findings reduces the false-positive counts by an order of magnitude at the cost of only 11 true positives. NanoCLUST and QIIME 2 analyses were not completed here because the CAMI2 dataset lacked quality score information and was not compatible with either software. However, we expect neither software to be a top performer in species detection due to their inability to do so on the simpler simulated data shown above.

ZymoBIOMICS. For the ONT reads Emu produces the lowest measured error distances across the methods tested at both the genus and species levels. Although almost all of the methods accurately

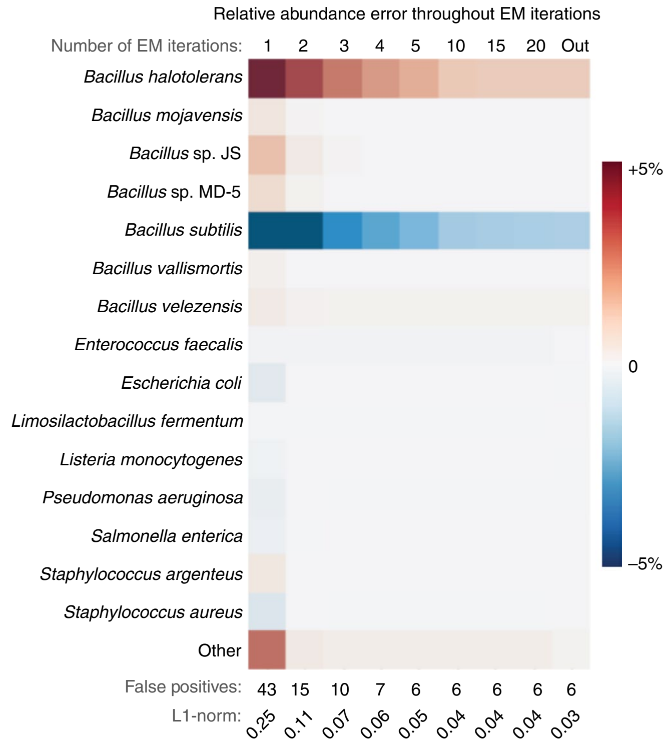


Fig. 4 | Relative error after consecutive EM iterations within Emu on ZymoBIOMICS ONT reads. The relative error of the Emu algorithm after 1, 2, 3, 4, 5, 10, 15 and 20 EM iterations is shown, as well as the final Emu output (out) on our ZymoBIOMICS sample sequenced by an ONT device. The 15 most abundant species in the computational estimate are displayed. The final Emu output includes threshold trimming and final re-estimation after 22 EM iterations. Darker blue represents an underestimate by the method, while darker red represents an overestimate. The color scheme is capped at ±5, meaning that errors greater than ±5% will be shown in the maximum error colors. The false-positive count and L1-norm are reported for each iteration with the ZymoBIOMICS guaranteed minimum abundance threshold of 0.01% applied.

detect the eight species in the sample, the number of false positives reported varies. Of the methods with perfect recall, NanoCLUST returns the fewest false positives and Emu returns the second fewest. It is also important to note that the abundance accuracy and sensitivity measured in the ONT dataset prove superior to those of the Illumina dataset, especially at the species level. When restricted to only the Illumina results, Emu again has the lowest L1 distance. Although Emu is not primed for Illumina 16S reads, this shows that Emu is a sensible approach regardless of the read-error profile. Figure 2 provides a graphical representation of the accuracy measures displayed in Table 1.

Synthetic gut microbiome. For the ONT reads Emu once again has the best or near-best performance for these metrics on the synthetic gut microbiome community. This is an intentionally challenging community because it contains several microbes that have a relative abundance of below 0.01%, even based on putative input abundance. This is a particular form of stress test for Emu because the EM algorithm specifically down-weights low-abundance taxa that are closely related to those in higher abundance (reflecting the likelihood of sequencing error accounting for the match). Nonetheless, Emu reports the best L1 distance at the species level. Centrifuge reports the best L2 distance, although this statistic is affected by the species that is not present in the database: *Romboutsia hominis*. In standard classification methods these reads are classified under an assortment of *Romboutsia* species, however, in methods involving statistical re-estimation these reads are labeled as a close relative, which ultimately increases the squared error. In terms of presence–absence calls, Emu is only one species short of having the highest true-positive count but it has far fewer false positives than every method aside from NanoCLUST. Although NanoCLUST does report the lowest false-positive counts, it also detects fewer true positives than others. The results are shown in Extended Data Fig. 3.

Comparison of EM iterations within Emu. To get a sense of the value of the error-correction step, Fig. 4 shows the relative abundance error calculated after subsequent EM iterations in Emu on the ZymoBIOMICS mock community ONT dataset. The error reduction with each iteration is especially clear for the *Bacillus* genus. The species that is present as per the manufacturer information is *B. subtilis*, but *B. halotolerans* differs from it by fewer than 15 bases over the length of the entire 16S rRNA gene. As a result of this similarity and the high error in ONT reads, we would expect a large fraction of reads to map to *B. halotolerans*. Our minimap2 primary alignment results do just that by classifying the *Bacillus* species reads as approximately 67% *B. subtilis*, 18% *B. halotolerans* and 15% distributed among 21 other species. In Emu's final estimate, however, more than 92% of the *Bacillus* reads are dedicated to *B. subtilis*, while only five additional *Bacillus* species are falsely identified to account for the remaining 8%.

Database comparison. To evaluate the default Emu database compared with a larger, well-reputed[34] 16S rRNA gene database, results were generated with the Ribosomal Database Project (RDP)[35] using Emu, minimap2, Kraken 2 and Bracken for our four ONT test datasets. The performance of each method was evaluated at both the genus and species level using three metrics: the L1-norm of the taxonomic abundance profile, the count of true-positive taxa and the count of false-positive taxa (Supplementary Table 13). Computational resources required by each method were measured by recording the run time and memory usage for each software (Supplementary Table 14). Complete abundance profile outputs for all RDP database results are listed in Supplementary Tables 15–18. These results show that Emu with the Emu default database has the lowest L1-norm for all four test datasets at both the species and genus level compared with all other software tool and database combinations evaluated. Although we would expect this from our simulated datasets (MBARC-26 and CAMI2) because these were simulated from sequences in the Emu database, these findings were mirrored in our mock communities as well.

Novel species simulation. To simulate the real-world scenario of communities containing novel species, or species that are not yet in our database, we used our CAMI2 dataset and removed sequences from our Emu and RDP databases. First, a list of 35 of the 345 species in our simulated CAMI2 dataset was randomly selected, in which reads from these species comprised 9.5% of the total CAMI2 simulated reads. All database sequences classified under these 35 species were removed from both the Emu and RDP databases. Results were then generated for our complete CAMI2 dataset with both the incomplete Emu and RDP databases by Emu, minimap2, Kraken 2 and Bracken. Performance statistics L1-norm, TP (true positive), FP (false positive) and unclassified percent are given in Supplementary Table 19, while complete abundance profile outputs are given in Supplementary Tables 20 and 21. These results show that Emu with the Emu database still produces the lowest L1-norm across taxonomic ranks evaluated when presented with novel species. A heatmap of the relative abundance error for families of the removed species with both the incomplete Emu and RDP databases is shown in Extended Data Fig. 4. This highlights the ability of Emu

Table 2 | Relative abundance of dominant and marker taxa assigned by Emu from 16S rRNA gene ONT data and by Bracken from WGS ONT data

		Control group						Vaginosis group					
Sample		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Dominant genera													
Lactobacillus	Emu	1.00	1.00	1.00	0.99	1.00	1.00	0.40	0.96	0.16	0.06	0.00	0.68
	Bracken	0.88	0.95	0.95	0.60	0.98	0.97	0.11	0.64	0.05	0.57	0.02	0.34
Gardnerella	Emu	–	–	–	–	–	–	–	–	–	0.00	–	–
	Bracken	0.03	0.02	0.03	0.37	0.01	0.02	0.57	0.03	0.44	0.31	0.07	0.48
Prevotella	Emu	–	–	–	–	–	–	0.03	0.00	0.04	0.01	0.04	0.02
	Bracken	0.01	0.01	0.01	0.01	0.00	0.01	0.09	0.09	0.13	0.12	0.53	0.05
Megasphaera	Emu	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.32	0.50	0.00	0.25
	Bracken	–	–	–	–	–	–	0.00	0.00	0.06	–	0.00	0.05
Aerococcus	Emu	–	–	–	–	–	–	0.29	–	0.01	0.15	0.00	0.02
	Bracken	–	0.00	–	–	–	–	0.11	0.00	0.01	–	0.00	0.01
***Lactobacillus* CST marker species**													
L. crispatus	Emu	0.00	0.50	0.99	0.00	1.00	0.99	0.00	0.96	0.00	0.00	0.00	0.00
	Bracken	0.06	0.63	0.92	0.06	0.89	0.96	0.03	0.64	0.02	0.61	0.01	0.04
L. gasseri	Emu	0.00	0.00	0.00	0.98	0.00	0.00	0.00	–	–	0.00	–	0.00
	Bracken	–	–	–	0.46	–	–	0.00	–	–	–	–	0.00
L. iners	Emu	0.99	0.47	0.00	0.00	0.00	0.00	0.34	0.00	0.16	0.06	0.00	0.68
	Bracken	0.86	0.28	0.01	0.01	0.00	–	0.06	0.01	0.03	–	0.00	0.29
L. jensenii	Emu	–	0.02	0.01	–	–	0.01	0.05	–	–	–	–	–
	Bracken	–	0.02	0.02	0.05	–	0.01	0.01	–	–	–	–	–
Inferred CST	**Emu**	**3**	**1**	**1**	**2**	**1**	**1**	**4**	**1**	**4**	**4**	**4**	**4**
	Bracken	**3**	**1**	**1**	**2/4**	**1**	**1**	**4**	**1**	**4**	**1/4**	**4**	**4**

CST, community state type. Dominant genera are defined as those with more than 10% abundance in at least one sample. The CST marker species of *Lactobacillus* have been defined previously[40,41]. The data are rounded and '–' denotes true zero values.

with the Emu database to accurately classify reads from novel species at the lowest taxonomic rank that is in the database.

Read mapping software comparison. To evaluate the impact of the read mapping software in Emu, results were generated with a version of Emu with BWA-MEM[12] as the read mapper (Supplementary Table 22). In both datasets tested, the error measured (L1-norm and L2-norm) and the number of false positives decreased after the EM algorithm is applied to the read mapping results.

Minimum abundance threshold comparison. To evaluate the error-correction step in Emu compared with the use of a minimum abundance threshold, we have applied the Emu default minimum abundance threshold of 10 reads to the results listed in Table 1. These results show that Emu still reports the fewest false positives across the tested datasets (Supplementary Table 23).

Research application: human vaginal microbiome. Variation in vaginal microbiota is associated with several urogenital diseases (for example, bacterial vaginosis)[36,37], a variety of sexually transmitted infections (for example, HIV)[38], and uncategorized phenotypes such as reproductive success[36,39]. Vaginal microbial communities can be classified into six so-called 'community state types', that is, I, II, III, IV-A, IV-B and V, which are defined by the relative abundance of several *Lactobacillus* species and the presence of anaerobic bacteria[40,41]. We generated community composition from 12 vaginal samples, six with diagnosed bacterial vaginosis and six controls, using Emu and an established whole-genome shotgun (WGS) metagenomic approach. We compared the characterization of community state types between the two pipelines to test Emu's ability to reproduce accepted community clusters.

Experimental design. Twelve vaginal swabs were obtained from the German Centre for Infections in Gynecology and Obstetrics at Helios Hospital Wuppertal and prepared at the Institute of Medical Microbiology, Virology and Hospital Hygiene at the University of Duesseldorf. Samples 1–6 originate from control group patients and samples 7–12 from patients with diagnosed bacterial vaginosis. Each sample was sequenced using whole-genome and 16S rRNA ONT workflows. The whole-genome reads were processed into species- and genus-level abundance profiles using Kraken 2 and Bracken, while the 16S rRNA reads were processed with Emu.

16S rRNA gene and whole-genome shotgun data. Comparison of 16S rRNA and WGS sequencing data is not trivial, even when sequencing libraries are prepared from the same nucleic acid preparation; bias introduced during amplification and sequencing in marker gene sequencing may differ from the bias produced in WGS sequencing, which ultimately influences the bioinformatic analysis in each approach[42]. Still, this comparison is useful to present the benefits and limitations of 16S rRNA gene amplicon sequencing. Given that swabs contained a significant portion of host DNA (98–99% of reads classified as human by Kraken 2 and Bracken), the number of bacterial reads was lower in WGS than in 16S rRNA sequencing. To reduce the bias due to imbalance in sensitivity between the two methods, a species detection threshold of 0.01% was set for Emu.

Table 2 lists the most abundant bacterial genera and the four *Lactobacillus* species that are used as markers for the inference of vaginal community state type. Healthy vaginal microbial communities are reported to be dominated by *Lactobacillus* species, while vaginal dysbiosis has been associated with high abundance of the genera *Gardnerella*, *Prevotella*, *Megasphaera* and *Aerococcus*[43]. Both pipelines, 16S rRNA and WGS, show relative abundance results that agree with that previous research.

The most notable discrepancy between the WGS and 16S rRNA methods for these genera is the relative abundance of *G. vaginalis*, in that the WGS method shows this species as being significantly abundant while the 16S rRNA method misses it almost entirely. This is a result of the same primer mismatch problem noted earlier for the family *Bifidobacteriaceae*, the parent family of *G. vaginalis*. Even with this bias, the inferred community state type is consistent between the Emu and Bracken workflows across the samples: both pipelines express the same marker species of the dominant community state type in 11 of the 12 samples. Sample 10 is the only sample for which the assignment is different between the methods, and this may be explained by the low sequencing depth acquired in the WGS approach for this sample. Despite the variation in the community profiles generated for these two pipelines (Extended Data Fig. 5), the clearly inferred community state types were identical between the pipelines (except in samples 4 and 10), reflecting the congruency in the clinical outcome of these two approaches.

Discussion

Emu is a homology-aware alignment likelihood approach in which read classification probabilities are adaptively updated based on read alignments to multiple reference sequences and the current community profile estimate. This iterative approach goes beyond the simple classification of individual reads and instead uses information gathered from the entire community to enable accurate and robust community profiling despite high error rates in the input sequences. Demonstrated error reduction (Fig. 4) as well as the superior results reported when comparing Emu's output to both read mappers tested alone (Supplementary Table 22) are indicative of the performance and flexibility of the adaptive likelihood model used in Emu.

Emu is impactful for two main reasons: the reduced number of false positives and the ability to differentiate between genomically similar species. To demonstrate the false-positive count reduction accomplished by the EM portion of Emu, we can compare results between Emu and minimap2. In each of the four ONT test sets, the false positives drop significantly, namely from 73 to 0, from 4,665 to 250, from 45 to 6 and from 252 to 40.

To observe Emu's ability to distinguish between genomically similar organisms, we can zoom in on two pairs of similar species in our MBARC-26 dataset. The first pair includes *Salmonella bongori* and *Salmonella enterica*, which have true relative abundances 0.04% and 0.17%, respectively. The reference sequences for these species have an average nucleotide identity of 97%, but Emu is able to accurately determine the appropriate relative abundance for each of these species within 0.001%. A second similar pair includes *Desulfosporosinus acidiphilus* and *Desulfosporosinus meridiei* with relative abundances of 7% and 2.6%, respectively, and an average nucleotide identity between the two reference species of 94%. Emu accurately estimated each of their relative abundances within 0.0005% of the expected value.

While both MetaMaps and Emu were developed for long-read data and incorporate a form of the EM algorithm, their difference is best understood in terms of 'horizontal' alignment (MetaMaps) and 'vertical' alignment (Emu). MetaMaps was developed for the analysis of shotgun metagenomic data and can make use of homology information from entire microbial genomes to correctly place individual reads (that is, horizontal alignment); this enables MetaMaps to skip (computationally expensive, in particular at the scale of the reference databases used for whole-genome metagenomic analysis) base-level alignment between reads and reference sequences and limit itself to more efficient approximate alignment. Emu, by contrast, is designed for the accurate analysis of an individual locus across a very large number of reference records (that is, vertical alignment); given that the 16S rRNA gene sequences of different species may differ by as little as a few bases, Emu operates under conditions in which individual alignment matches and mismatches need to be carefully examined while taking into account the increased error rate of ONT sequencing. Base-level alignment between individual reads and the 16S rRNA gene reference database (that is, vertical alignment) thus becomes necessary, along with an EM approach that uses an alignment likelihood model tailored to the requirements of 16S rRNA gene sequence analysis.

Due to the nature of probabilistic models, Emu creates a long tail of low-abundance species. To avoid this long inaccurate list in the results, the built-in threshold for Emu is the equivalent abundance of 1 read for samples with less than 1,000 reads or 10 reads for anything larger. This means that Emu will not be able to detect microbes with abundance lower than this threshold. This occurs in our synthetic gut mock community, which contains only five *Clostridium leptum* reads, and thus Emu does not report this species. Therefore, in cases in which the detection of ultra-low-abundance organisms is imperative, Emu would probably not be the best tool.

The optimal abundance threshold cut-off to distinguish between true assignments and noise resulting from the EM algorithm is an open question for Emu. A future model could use the statistical information from the sample to establish a minimum abundance threshold instead of the current somewhat arbitrary cut-off explained above. A second parameter setting that is open for future development is the number of secondary alignments kept from the mimimap2 output. The current default of up to 50 alignments was selected with a parameter sweep to evaluate the trade-off between accuracy and additional computing cost, and can be modified at the command line. A future version of Emu could incorporate an algorithm to determine an optimal value given the input sample.

Given that Emu is a full-length alignment-based approach, more computational resources (that is, memory and time) are required than alternative methods. This may prevent certain users from incorporating Emu into their pipeline depending on the availability of appropriate computing resources. Future work in this area could reduce these requirements.

Emu is a closed-reference approach, which ultimately restricts output to only those bacteria and archaea that are present in the database. As seen in our novel species simulation experiment, we would expect Emu to classify sequences from novel species as a near neighbor that is present in the database and therefore to accurately classify these sequences at the lowest taxonomic rank that is present in the database. Further work in this area could label reads from novel species as 'unclassified' instead. However, as shown in Supplementary Table 19, this is a complex task given that even with the use of the larger database RDP and the k-mer-based technologies Kraken 2 and Bracken, the novel species still fail to be labeled as 'unclassified' when it comes to high-error full-length 16S rRNA reads.

In addition to comparing 16S rRNA gene microbial community profiling methods, the present results also highlight the differences between the ONT and Illumina 16S rRNA gene sequencing technologies. In both mock communities, ONT reads are able to deliver lower L1 and L2 distances than Illumina reads. An argument can be made that selection of a different hypervariable region could have produced better results for the Illumina reads, however, this resembles the actual decision-making process for Illumina users and the potential bias it may produce. It is also important to note that Emu and NanoCLUST are the only tools evaluated here that were designed for 16S rRNA, thus we expect full-length reads with these two methods to produce the most accurate results.

The potential for long, single-molecule reads to deliver higher resolution pictures of microbial communities from single-pass sequencing remains enticing, but the high rate of sequencing error is a formidable obstacle. Specifically, although short reads are constrained in sensitivity below the genus level, long reads are not; instead, their difficulty is with specificity. In the case of long reads applied to 16S rRNA amplicon sequencing, Emu represents an important improvement in minimizing this trade-off and it has the potential to show the communities of well-studied environments in a new light. Given that the error profiles are dynamically learned from the input data, Emu has the flexibility to adjust as sequencing technologies develop and would continue to be appropriate for use in novel third-generation sequencing technologies in future applications.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-022-01520-4.

Received: 30 April 2021; Accepted: 10 May 2022;
Published online: 30 June 2022

References

1. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl Acad. Sci. USA* **74**, 5088–5090 (1977).
2. Martínez-Porchas, M., Villalpando-Canchola, E. & Vargas-Albores, F. Significant loss of sensitivity and specificity in the taxonomic classification occurs when short 16S rRNA gene sequences are used. *Heliyon* **2**, e00170 (2016).
3. Callahan, B. J., Grinevich, D., Thakur, S., Balamotis, M. A. & Yehezkel, T. B. Ultra-accurate microbial amplicon sequencing with synthetic long reads. *Microbiome* **9**, 130 (2021).
4. Miller, C. S. et al. Short-read assembly of full-length 16S amplicons reveals bacterial diversity in subsurface sediments. *PLoS ONE* **8**, e56018 (2013).
5. Workman, R. E. et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* **16**, 1297–1305 (2019).
6. Callahan, B. J. et al. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Res.* **47**, e103 (2019).
7. Karst, S. M. et al. High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nat. Methods* **18**, 165–169 (2021).
8. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
9. Nearing, J. T., Douglas, G. M., Comeau, A. M. & Langille, M. G. I. Denoising the denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ* **6**, e5364 (2018).
10. 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).
11. Santos, A., van Aerle, R., Barrientos, L. & Martinez-Urtaza, J. Computational methods for 16S metabarcoding studies using Nanopore sequencing data. *Comput. Struct. Biotechnol. J.* **18**, 296–305 (2020).
12. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://doi.org/10.48550/arxiv.1303.3997 (2013).
13. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493 (2011).
14. Benítez-Páez, A., Portune, K. J. & Sanz, Y. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. *GigaScience* **5**, 4 (2016).
15. Fujiyoshi, S., Muto-Fujita, A. & Maruyama, F. Evaluation of PCR conditions for characterizing bacterial communities with full-length 16S rRNA genes using a portable nanopore sequencer. *Sci. Rep.* **10**, 12580 (2020).
16. Shin, J. et al. Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing. *Sci. Rep.* **6**, 29681 (2016).
17. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**, 1721–1729 (2016).
18. Juul, S. et al. What's in my pot? Real-time species identification on the MinION™. Preprint at *bioRxiv* https://doi.org/10.1101/030742 (2015).
19. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
20. Valenzuela-González, F., Martínez-Porchas, M., Villalpando-Canchola, E. & Vargas-Albores, F. Studying long 16S rDNA sequences with ultrafast-metagenomic sequence classification using exact alignments (Kraken). *J. Microbiol. Methods* **122**, 38–42 (2016).
21. Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).
22. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **2017**, e104 (2017).
23. Lu, J. & Salzberg, S. L. Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2. *Microbiome* **8**, 124 (2020).
24. Rodríguez-Pérez, H., Ciuffreda, L. & Flores, C. NanoCLUST: a species-level analysis of 16S rRNA nanopore sequencing data. *Bioinformatics* **37**, 1600–1601 (2021).
25. Di Tommaso, P. et al. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
26. Dilthey, A. T., Jain, C., Koren, S. & Phillippy, A. M. Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. *Nat. Commun.* **10**, 3066 (2019).
27. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
28. Roberts, A. & Pachter, L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods* **10**, 71–73 (2013).
29. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
30. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
31. Singer, E. et al. Next generation sequencing data of a defined microbial mock community. *Sci. Data* **3**, 160081 (2016).
32. Meyer, F. et al. Critical Assessment of Metagenome Interpretation: the second round of challenges. *Nat. Methods* **19**, 429–440 (2022).
33. Winand, R. et al. Targeting the 16S rRNA gene for bacterial identification in complex mixed samples: comparative evaluation of second (Illumina) and third (Oxford Nanopore Technologies) generation sequencing technologies. *Int. J. Mol. Sci.* **21**, 298 (2020).
34. Edgar, R. Taxonomy annotation and guide tree errors in 16S rRNA databases. *PeerJ* **6**, e5030 (2018).
35. Cole, J. R. et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* **42**, D633–D642 (2014).
36. Smith, S. B. & Ravel, J. The vaginal microbiota, host defence and reproductive physiology. *J. Physiol.* **595**, 451–463 (2017).
37. Pybus, V. & Onderdonk, A. B. Microbial interactions in the vaginal ecosystem, with emphasis on the pathogenesis of bacterial vaginosis. *Microbes Infect.* **1**, 285–292 (1999).
38. Petrova, M. I., van den Broek, M., Balzarini, J., Vanderleyden, J. & Lebeer, S. Vaginal microbiota and its role in HIV transmission and infection. *FEMS Microbiol. Rev.* **37**, 762–792 (2013).
39. Mendling, W. Vaginal microbiota. *Adv. Exp. Med. Biol.* **902**, 83–93 (2016).
40. Gajer, P. et al. Temporal dynamics of the human vaginal microbiota. *Sci. Transl. Med.* **4**, 132ra52 (2012).
41. Ravel, J. et al. Vaginal microbiome of reproductive-age women. *Proc. Natl Acad. Sci. USA* **108**, 4680–4687 (2011).
42. Brooks, J. P. et al. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiol.* **15**, 66 (2015).
43. Onderdonk, A. B., Delaney, M. L. & Fichorova, R. N. The human microbiome during bacterial vaginosis. *Clin. Microbiol. Rev.* **29**, 223–238 (2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Emu algorithm. Figure 1 shows a high-level schematic diagram of the algorithm, while Extended Data Fig. 1 provides a more detailed picture of the equations in the algorithm. The pipeline begins by generating alignments between input reads and database sequences with minimap2 (ref. [30]). With these mappings, the following steps are completed: establishment of the initial probabilities, redistribution of the sample composition with an EM algorithm, and then the trimming of noise for a final estimation.

Initial probabilities. To apply the EM framework we need, first, an initial sample composition estimate for vector *F*, and second, alignment likelihoods $P\left(r|s\right)$ between each sample read *r* and database reference sequence $s \in S$. Given that we do not have any pre-existing knowledge about the sample composition, *F* starts as an evenly distributed vector $F\left(t\right)_{t\in T} = \frac{1}{|T|}$, where *T* is the set of all taxonomy identifications in *S*. To identify the alignment likelihoods $P\left(r|s\right)$ we start by generating pairwise sequence alignments between $r \in R$ and $s \in S$ with minimap2, where *R* represents all reads in the sample. We determine the likelihood of the nucleotide alignment types mismatch (X), insertion (I), deletion (D) and softclip (S) by counting the number of occurrences of each nucleotide alignment type in the primary alignments. We define these probabilities with a simple proportion, $P\left(c\right) = \frac{n_c}{\sum_{c\in C} n_c}$, where C = [X,I,D,S] and n_c is the sum of occurrences of that type among all of the primary alignments.

Using the likelihood for each type of nucleotide alignment, the likelihood for each pairwise sequence alignment $r \in R$, $s \in S$ is calculated as $P\left(r|s\right) = \prod_{c\in C} P\left(c\right)^{\hat{n}_{c(r,s)}}$, where $\hat{n}_{c(r,s)}$ is the normalized number of occurrences of alignment type *c* observed in the alignment between *r* and *s*. The count for each alignment type is normalized by dividing the length of the longest alignment for read *r* by the length of the alignment: $\hat{n}_{c(r,s)} = n_{c(r,s)} * \frac{\max\limits_{s'\in S}\left(\mathrm{len}\left(r,s'\right)\right)}{\mathrm{len}\left(r,s\right)}$. This normalization accounts for variation in alignment lengths for a given read, which is caused by deletions. In the event that no alignment is generated between *r* and *s*, $P\left(r|s\right) = 0$. Given that we are interested in the most likely taxonomy of *r* rather than the most similar sequence *s*, we keep only the highest $P\left(r|s\right)$ for any *s* with species-level taxonomy identification (ID) *t*. Thus, the alignment probability between each read *r* and species-level taxonomy *t* is calculated with $P\left(r|t\right) = \max\limits_{s\in t}\left(\prod_{c\in C} P\left(c\right)^{\hat{n}_{c(r,s)}}\right)$, where $s \in t$ represents all *s* with the taxonomy ID *t*. With the initial probabilities set, we can now improve our sample composition estimation with an EM probabilistic model.

Redistribution of sample composition. The likelihood of *r* emanating from species *t* is constructed for each $P\left(r|t\right)$ using Bayes' theorem, $P\left(t|r\right) = \frac{P(r|t)*F(t)}{\sum_{t\in T} P(r|t)*F(t)}$. With these probabilities, *F* is redistributed as $F\left(t\right)_{t\in T} = \frac{\sum_{r\in R} P(t|r)}{|R|}$. The accuracy of this estimate is evaluated using the total log likelihood, $L\left(R\right) = \sum_{r\in R} \log\left[\sum_{s\in S} P\left(r|t\right) * F\left(t\right)\right]$, which increases with each iteration. If this *L*(*R*) improvement from the previous iteration is substantial (>0.01), then this re-estimation step is repeated with the updated *F*. Otherwise, redistribution is complete and we move to the final phase of the algorithm.

Noise trimming for final estimation. Due to the nature of the probabilistic structure in an EM model, vector *F* is likely to contain a long tail of species claiming low abundance. To avoid this long list of false positives in the output, any abundance below the set threshold will be modified to 0 at this stage. The default threshold for Emu is an abundance equivalent to 1 read for samples with under 1,000 reads and 10 reads for larger samples; however, the user can modify this parameter. Once *F* is trimmed, Emu enters one final round of abundance redistribution. The resulting *F* is given as the final sample composition estimation.

Simulated read generation. Two simulated datasets were generated to mimic the ONT full-length 16S rRNA reads. First, 958,655 ONT reads were simulated using DeepSimulator v1.5 (ref. [44]) with default settings on a synthetic metagenomic community structure following the composition of the published mock community MBARC-26 (ref. [31]). Reference 16S rRNA gene sequences were obtained from 16S RefSeq (Reference Sequence) nucleotide sequence records[45]. For strains not present in the RefSeq 16S rRNA sequence database, all strains under the same species as the desired strain were used instead.

Given that CAMISIM does not currently have the functionality to simulate 16S rRNA data, the simulator in its pipeline, NanoSim[46], was used in isolation following the CAMI2 (ref. [32]) mouse gut profile (https://www.microbiome-cosi.org). 16S rRNA sequences were selected from 16S RefSeq[45] based on taxonomy IDs in the described CAMI2 mouse gut profile. For unfound organisms, 16S rRNA sequences were selected from the Ribosomal RNA Operon Copy Number Database (rrnDB) v5.6 (ref. [47]) by name instead. The number of reads simulated for each microbe was determined by multiplying the relative abundance by 10^7 to ensure that each species contained at least one simulated read. Given that the generated dataset contained more than 400 million reads, this dataset was then subsampled down to 1% to reduce computational load, resulting in 4,310,093 reads.

Creation of the gut microbiome mock community. Each gut bacterium was activated and propagated individually in brain heart infusion (BHI) medium supplemented with hemin (5 mg l^{-1}) and yeast extract (10 g l^{-1}). The plate counting method was used to determine viable cells in the cultures after 4 h of anaerobic cultivation at 37 °C; all bacterial strains were combined with an equal volume of 100 µl. Cultures were then centrifuged at 12,000 *g* for 10 min before extra bacterial lysis with lysozyme followed by DNA extraction using the MasterPure Complete DNA and RNA Purification Kit. DNA was quantified using the Qubit kit.

Sequencing mock communities. *ZymoBIOMICS.* A detailed description of the steps taken to sequence the ZymoBIOMICS sample can be found in the Methods section of the study that produced these sequences[33].

Synthetic gut microbiome. Library construction and sequencing of the V4 region of the 16S rRNA gene were performed using the NEXTflex 16S V4 Amplicon-Seq Kit 2.0 (Bio Scientific) with 20 ng input DNA, and sequences were generated on the Illumina MiSeq platform (Illumina).

Library construction and sequencing of the full-length 16S rRNA gene were performed using the MinION nanopore sequencer (ONT) and 16S Barcoding Kit 1-24 (ONT, cat. no. SQK-16S024). The polymerase chain reaction amplification and barcoding was completed with 15 ng template DNA added to the LongAmp Hot Start Taq 2X Master Mix (New England Biolabs). Initial denaturation at 95 °C was followed by 35 cycles of 20 s at 95 °C, 30 s at 55 °C, 2 min at 65 °C, and a final extension step of 5 min at 65 °C. Purification of the barcoded amplicons was performed using the AMPure XP Beads (Beckman Coulter) as per ONT's instructions. Samples were then quantified using Qubit fluorometer (Life Technologies) and pooled in an equimolar ratio to a total of 50–100 ng in 10 µl. The pooled library was then loaded into an R9.4.1 flow cell and run as per the manufacturer's instructions. MINKNOW v19.12.5 was used for data acquisition.

Emu 16S database. The default database of Emu is a combination of rrnDB v5.6 (ref. [47]) and NCBI (National Center for Biotechnology Information) 16S RefSeq downloaded on 17 September 2020 (ref. [45]). Duplicate species-level entries, defined as entries with identical sequences and species-level identification, were removed. The resulting database contains 49,301 sequences from 17,555 unique microbial species. Database taxonomy was also retrieved from NCBI on the same date as the RefSeq download. This database can be reproduced by using the build custom database option in Emu on both the rrnDB and RefSeq sequences separately, then concatenating the results.

Emu was first tested with the NCBI 16S RefSeq database on our ONT ZymoBIOMICS sample. This yielded subpar accuracy (Supplementary Table 24), which we attribute to the large number of reference sequences containing ambiguous bases. To increase the number of complete sequences in our database, rrnDB v5.6 was added because it contains species-level taxonomy and few ambiguous bases.

Three popular 16S rRNA gene databases are Greengenes[48], RDP[35] and SILVA[49]. Although each of the three contains far more sequences than our curated Emu database, species-level annotation in Greengenes is relatively low, SILVA does not map completely to the NCBI taxonomy, and RDP did not perform as well in our experiments. Given Emu's reliance on mapping each read to several database sequences, we have found that a smaller, well-curated database performs better in Emu. We have therefore created the default database of Emu as explained above, but have also pre-built an RDP database for Emu that is publicly available.

Emu RDP database. An Emu-compatible RDP[35] v11.5 database was generated with Emu's build-database function for the database comparison quantitative results listed in Supplementary Table 13 and the computational resource results listed in Supplementary Table 14. To construct this database, bacterial and archaeal 16S rRNA gene unaligned fasta sequences were downloaded from the RDP website, and the NCBI taxonomy database was downloaded in January 2022 (ref. [50]). Mappings between the RDP fasta sequence IDs and the NCBI taxonomy IDs were generated using the NCBI accession2taxid database[50]. Sequences that mapped to a taxonomy ID that was no longer in the NCBI taxonomy were removed. In addition, sequences that mapped to 'uncultured organism' (taxonomy ID: 155900), bacteria sequences that mapped to 'uncultured bacterium' (taxonomy ID: 77133) and archaea sequences that mapped to 'uncultured archaeon' (taxonomy ID: 115547) were removed. The resulting Emu database contains 1,089,863 sequences and can be downloaded via GitLab (https://gitlab.com/treangenlab/emu). The input files to create the Emu RDP database were then used to create a Kraken 2 database and generate both Kraken 2 and Bracken results. The Emu-compatible RDP database fasta file was used to generate minimap2 RDP results.

16S rRNA quantitative comparison. Barcodes were removed from each mock community dataset using the trim_barcodes function in Guppy Basecalling Software v4.4.2 (ref. [51]) for our ONT datasets and Trimmomatic v0.39 for the Illumina data. An equivalent of the default database of Emu was built for each software. 'Unclassified' reads, or those that failed to match a reference sequence, were removed prior to the calculation of relative abundance for each method. Supplementary Note 1 contains a detailed list of all commands used.

Minimap2. Minimap2 v2.24 classification was generated by selecting the top database hit for each input read. The preset option for ONT was used for our long-read data and the genomic short-read mapping preset was used for our Illumina data. The relative abundance of each species was calculated as the number of species classifications divided by the total number of aligned reads.

Kraken 2. Kraken 2 v2.1.1 was used to generate a custom database matching our Emu default, and then to produce classification results. To calculate relative abundance from the Kraken 2 classification, the 'clade counts' column from the Kraken 2 report (kreport) was used. For species-level results, only rows with 'rank:S' were kept. Relative abundance for each species was then defined as the clade counts for that species divided by the total number of clade counts in the reduced kreport. This process is then repeated at the genus level by restricting the kreport to only those rows with 'rank:G'.

Bracken. Bracken v2.5.0 was used to gather microbial abundance estimates from our Kraken 2 results. For full-length ONT reads, our custom Kraken 2 database was converted to a Bracken database with a read length of 1,500. The same process was applied to Illumina data, with read lengths of 250 and 300 used for the ZymoBIOMICS and the synthetic gut microbiome mock communities, respectively. Bracken abundance estimations were then generated for each dataset at the genus and species level.

NanoCLUST. Given that NanoCLUST uses a BLAST database, a custom BLAST database was created to match our Emu default database. NanoCLUST v1.0 was then run on each of our long-read samples with the docker profile option. And given that NanoCLUST generates relative abundance estimates at each taxonomic rank by default, no further processing was necessary.

Centrifuge. Centrifuge v1.0.4 was used to build a custom database and generate taxonomic classification of the four ONT datasets. The kreport generation functionality in Centrifuge was then incorporated to create Kraken-style reports for each Centrifuge classification result. Genus- and species-level relative abundance results were calculated from these kreports in the same manner as the Kraken 2 results described above.

MetaMaps. MetaMaps v.0.1 was used to build a custom database from the Emu default 16S rRNA database. The datasets were analyzed with the following alterations to the default settings: the estimated alignment identity target parameter was set to 90 (--pi 90) for all datasets for improved performance, and the minimal read length was set to 500 (-m 500) for CAMI2 data given that the dataset had shorter reads. Genus- and species-level relative abundance results were calculated from the output file ending in EM.WIMP by selecting all rows with the appropriate 'AnalysisLevel' (genus or species) then using the values in the 'EMFrequency' column directly.

QIIME 2. QIIME 2 results were produced with the classify-sklearn Naive Bayes classifier workflow of QIIME 2 2020.11.1. First, a QIIME 2 artifact representation of the default Emu database was generated with the appropriate QIIME 2 import command. Then, reference sequences were extracted appropriately based on the primer used for each sample and fitted to the reference taxonomy to produce a QIIME 2 classifier. The already demultiplexed sample reads were de-noised (Illumina) or de-replicated (ONT), and then classified with the appropriate pre-fit classifier. The taxonomic classifications were then collapsed to genus and species levels, and relative abundances were calculated separately for the two taxonomic rank results.

Establishing ground truth. A 'Zymo-exclusive' database containing only the provided 16S rRNA gene assembled reference genomes for the eight bacterial species in the sample was created. The ZymoBIOMICS samples were then mapped (BWA-MEM v0.7.17 for short-read data, minimap2 v2.17 for long-read data) to this Zymo-exclusive database for accurate classification of each read. Reads were classified as the top hit, and the ground truth relative abundances were derived from these results.

A restricted database for the 21 species known to be in our synthetic gut microbiome community was created by retrieving the NCBI 16S RefSeq entries for those species. This resulted in 45 sequences from 20 of the 21 species. Given that *Romboutsia hominis* is present in the sample but not in RefSeq, a *Romboutsia hominis* sequence was selected from GenBank[52] and included in the restricted database. Mapping, classification and sample composition calculation follow the workflow for the ZymoBIOMICS community described above. This community, however, is subject to other undocumented contamination that may introduce bias.

Accuracy evaluation metrics. L1-norm is essentially the linear error and is calculated using the equation $\sum_{s\in S}|E_s - I_s|$, where set S consists of the union between all of the species in the database and the ground truth, and E_s and I_s are the expected and inferred relative abundances for species s, respectively. A perfect L1 distance is 0, while an entirely inaccurate sample composition estimate would return an L1 distance of 2 given that $\sum_{s\in S} E_s = 1$ and $\sum_{s\in S} I_s = 1$. L2-norm is the sum of the squared error, which magnifies the cost of larger differences and is calculated using the equation $\sqrt{\sum_{s\in S}(E_s - I_s)^2}$. Precision, recall and F-score are used to evaluate the accuracy of microbe presence. For this explanation, TP represents true positives, FP represents false positives, and FN represents false negatives. Precision is the proportion of claimed true positives that are truly present in the sample: $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$. Recall is the percentage of expected positives that were detected by the software: $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$. The F-score is simply the harmonic mean between the two values: $\frac{2\cdot \mathrm{precision}\cdot \mathrm{recall}}{\mathrm{precision}+\mathrm{recall}}$. Given that the ZymoBIOMICS sample is guaranteed to contain <0.01% foreign microbial DNA, all ZymoBIOMICS results are trimmed to include only taxa with abundance ≥0.01%, prior to the calculation of performance metrics.

Computational resources. All software analysis was completed on a Ubuntu 18.04.4 LTS system, with the exception of the MetaMaps runs, which were completed on CentOS Linux release 7.9.2009. The /usr/bin/time command was used to gather time and memory statistics. Reported CPU (central processing unit) time is calculated by summing the user and the system time, and the RAM (random access memory) requirements are determined using the maximum resident set size. The only exception is NanoCLUST, for which computational requirements were instead extracted from the Nextflow execution report and timeline. Here, run time was gathered from the 'CPU-Hours' output in the execution report, and the maximum resident set size was determined by the step with the largest memory usage (RAM) in the execution timeline. The computational requirements recorded for Bracken are an accumulation of both the Bracken and Kraken 2 commands, given that both are required to produce the Bracken abundance estimation. Computational requirements for the QIIME 2 workflow are left out of this analysis because QIIME 2 involves several commands.

Clinical vaginal samples. *Data generation.* Total DNA and RNA was extracted using the ZymoBIOMICS DNA/RNA Miniprep Kit (cat. no. R2002). The 16S Nanopore sequencing library was prepared from 10 ng total DNA using the 16S Barcoding Kit (ONT, cat. no. SQK-RAB204). The whole-genome Nanopore library was prepared from the remaining total DNA using a Native Barcoding Expansion 1-12 (PCR-free) Kit (ONT, cat. no. EXP-NBD104) and Ligation Sequencing Kit (ONT, cat. no. SQK-LSK109). Data were sequenced on a MinION flow cell type R9 (ONT, cat. no. FLO-MIN106D) in two runs (a 16S rRNA gene run and a whole-genome run). Data were acquired with MINKNOW core v.4.0.5. Basecalling and demultiplexing were done using Guppy v.4.0.15.

Data analysis and databases. Computational analysis of vaginal samples was performed on a machine with CentOS Linux release 7.9.2009. Whole-genome sequencing data were analyzed with Kraken v.2.1.1 and Bracken v.2.5.

The Kraken 2 database was built from a custom metagenomic database, which includes all latest complete and reference genomes derived from the RefSeq database in the divisions bacteria, fungi, protozoa and viral of RefSeq (state 26.12.2019). The host portion of the metagenomic database is represented by a 1000 Genomes Project reference sequence and two well-characterized human assemblies (GCA_001524155.4 and GCA_002009925.1).

Retrieved Bracken abundances at both the genus and species levels were recalculated considering only bacteria to align with 16S rRNA gene results. Therefore, total Bracken results belonging to the superkingdom 'Bacteria' were assumed as having 100% abundance for each sample.

Emu was run on 16S rRNA sequencing data with a species detection threshold of 0.01%. Species- and genus-level abundances were retrieved from the Emu output. Community state types were inferred from the abundance profile by considering the dominance of four marker *Lactobacillus* species.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All sequenced samples used in this study are publicly available on Sequence Read Achieve (SRA). Both ZymoBIOMICS datasets are under BioProject ID PRJNA587452 with SRA accessions SRR10391201 for ONT and SRR10391187 for Illumina[33]. Our gut mock community is under BioProject ID PRJNA725207. The 12 vaginal samples used for our real-world application demonstration are uploaded under BioProject ID PRJNA723982. Our simulated sequences are publicly available on OSF under project 56UF7. Databases used in this paper include 16S RefSeq nucleotide sequence records (https://www.ncbi.nlm.nih.gov/refseq/targetedloci/16S_process/), Ribosomal Database Project (RDP) v11.5 (https://rdp.cme.msu.edu/) and rrnDB v5.7 (https://rrndb.umms.med.umich.edu/). Study of vaginal microbiomes was approved by the ethics committee of the Medical Faculty of Heinrich Heine University. All patient samples were collected with informed consent from individuals in the context of an exploratory clinical microbiome study approved by the Ethics Committee of the Medical Faculty of Heinrich Heine University Düsseldorf (institutional review board study identification '2019–600-andere Forschung erstvotierend').

Code availability

Emu and all associate code are available on GitLab (https://gitlab.com/treangenlab/emu). Emu can be installed via Bioconda (https://anaconda.org/bioconda/emu). A Code Ocean capsule of the package is provided (https://doi.org/10.24433/CO.7761675.v1). All scripts and data used to compile quantitative comparison results can be found on GitLab (https://gitlab.com/treangenlab/emu-benchmark).

References

44. Li, Y. et al. DeepSimulator: a deep simulator for Nanopore sequencing. *Bioinformatics* **34**, 2899–2908 (2018).
45. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
46. Yang, C., Chu, J., Warren, R. L. & Birol, I. NanoSim: nanopore sequence read simulator based on statistical characterization. *GigaScience* **6**, 1–6 (2017).
47. Stoddard, S. F., Smith, B. J., Hein, R., Roller, B. R. & Schmidt, T. M. rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Res.* **43**, D593–D598 (2015).
48. 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).
49. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
50. Schoch, C. L. et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* **2020**, baaa062 (2020).
51. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**, 129 (2019).
52. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **44**, D67–D72 (2016).

Acknowledgements

We thank two additional members of the Treangen Laboratory, B. Kille for technical support and N. Sapoval for algorithm development. Computational support and infrastructure were provided by the Centre for Information and Media Technology (ZIM) at the University of Düsseldorf (Germany). This work has been supported by Jürgen Manchot Foundation and Deutsche Forschungsgemeinschaft (DFG) award 428994620 (A.D., A.T., W.M., P.F. and E.G.). This work has also been supported by NIH grants from NIDDK P30-DK56338, NIAID R01-AI10091401, U01-AI24290 and P01-AI152999, and NINR R01-NR013497 (T.S. and Q. Wu). Q. Wang and S.V. were supported in part by NIH grant R21NS106640 from the National Institute for Neurological Disorders and Stroke (NINDS). K.D.C. was supported in part by a Ken Kennedy Institute Computational Science and Engineering Graduate Recruiting Fellowship. K.D.C., M.G.N. and T.J.T. were supported in part by NIH grant P01-AI152999 from the National Institute of Allergy and Infectious Diseases (NIAID). K.D.C. and T.J.T. were supported in part by NSF EF-2126387. M.G.N. was funded by a fellowship from the National Library of Medicine Training Program in Biomedical Informatics and Data Science (T15LM007093, PI: Kavraki).

Author contributions

A.D. and T.J.T. derived the Emu concept and supervised the project. K.D.C., Q. Wang and M.G.N. developed the software. K.D.C., Q. Wang, A.T. and E.R produced results for benchmarking. P.F., E.G., W.M., S.S, Q. Wu, T.S. and S.V. generated sequencing data for analysis and contributed to the interpretation of results. K.D.C., Q. Wang, M.G.N., A.T., Q. Wu, E.R., A.D. and T.J.T. contributed to writing the original draft of the manuscript. All authors read, revised and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

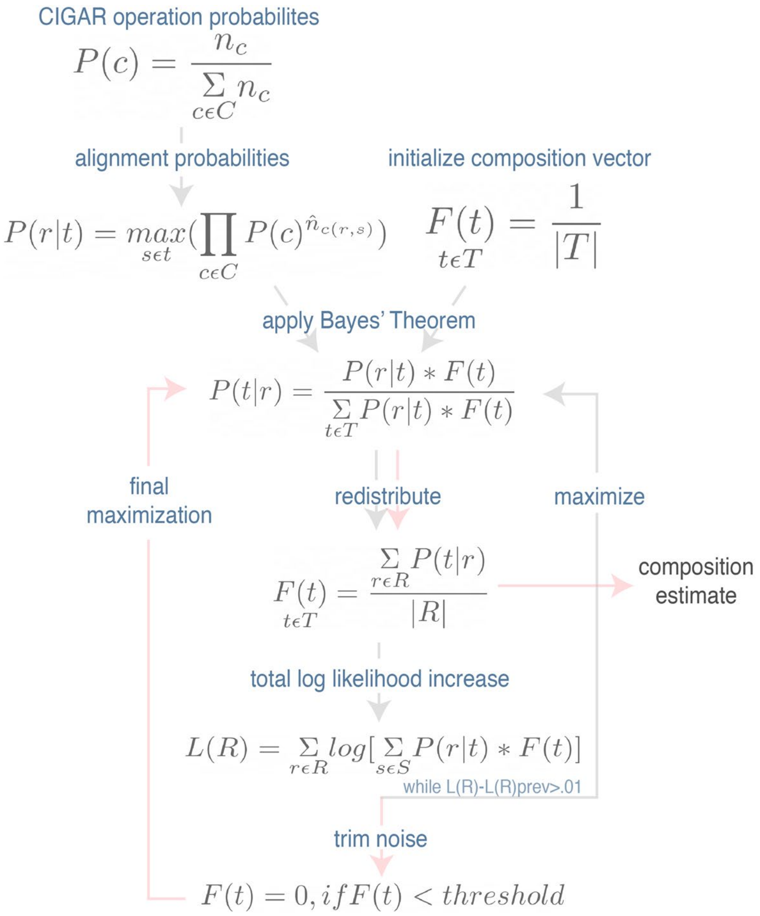
Extended data are available for this paper at https://doi.org/10.1038/s41592-022-01520-4.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-022-01520-4.

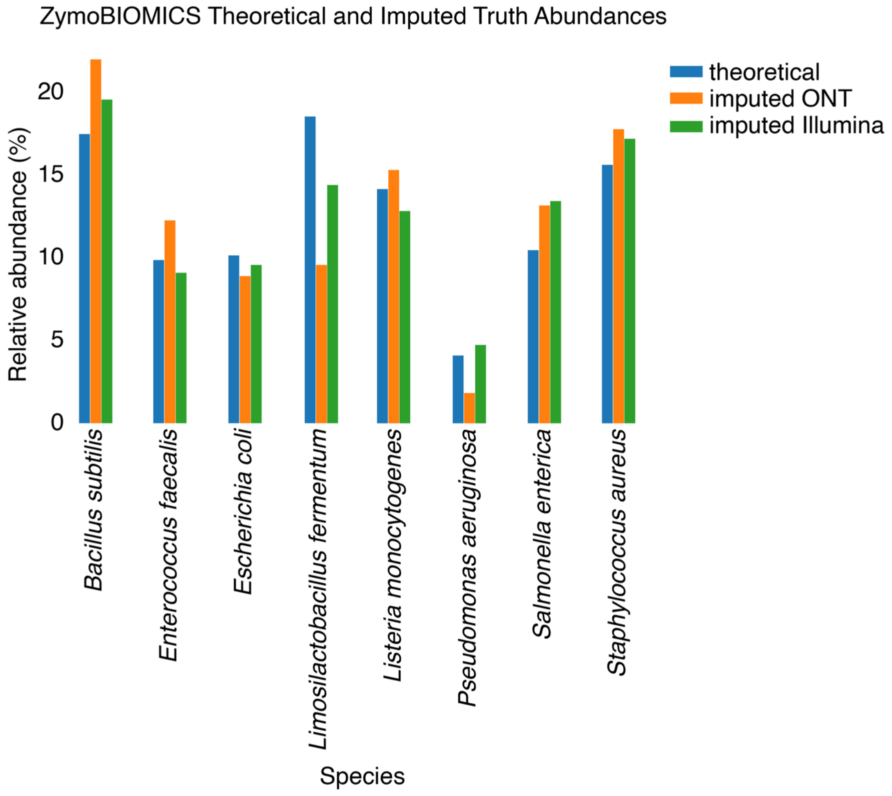
Correspondence and requests for materials should be addressed to Kristen D. Curry, Alexander Dilthey or Todd J. Treangen.

Peer review information *Nature Methods* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Lin Tang, in collaboration with the *Nature Methods* team.

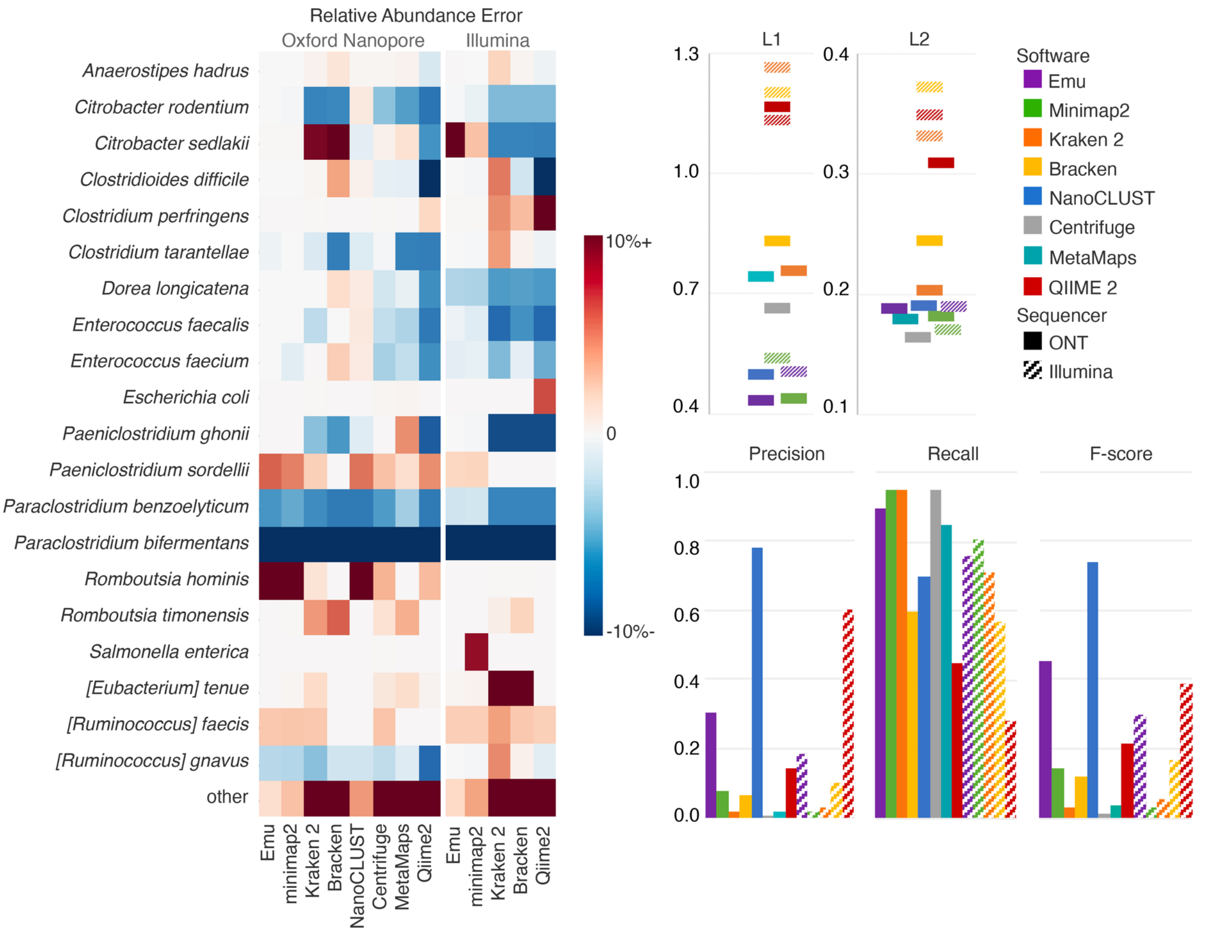
Reprints and permissions information is available at www.nature.com/reprints.


CIGAR operation probabilites
$P(c) = \frac{n_c}{\sum_{c\epsilon C} n_c}$
alignment probabilities
initialize composition vector
$P(r|t) = \max_{s\epsilon t}(\prod_{c\epsilon C} P(c)^{\hat{n}_{c(r,s)}})$
$F(t)_{t\epsilon T} = \frac{1}{|T|}$
apply Bayes' Theorem
$P(t|r) = \frac{P(r|t) * F(t)}{\sum_{t\epsilon T} P(r|t) * F(t)}$
final maximization
redistribute
maximize
$F(t)_{t\epsilon T} = \frac{\sum_{r\epsilon R} P(t|r)}{|R|}$
composition estimate
total log likelihood increase
$L(R) = \sum_{r\epsilon R} log[\sum_{s\epsilon S} P(r|t) * F(t)]$
while L(R)-L(R)prev>.01
trim noise
$F(t) = 0, if F(t) < threshold$

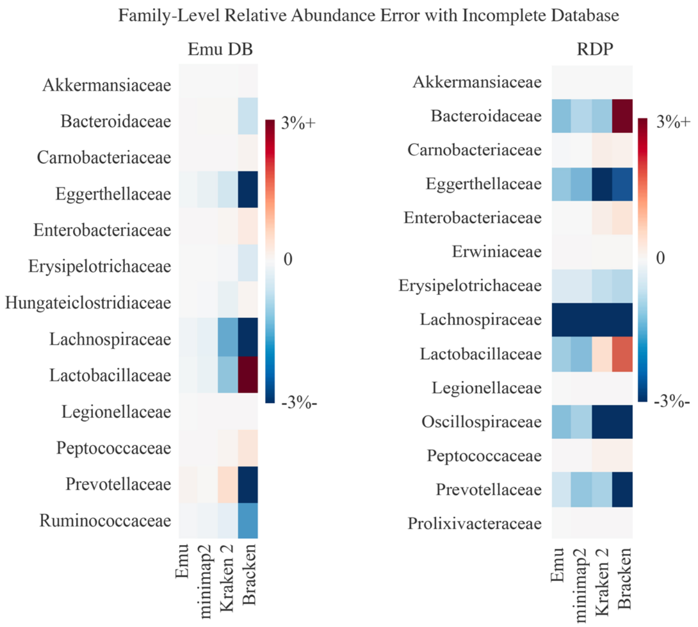
Extended Data Fig. 1 | Pictorial representation of the complete Emu algorithm. Follow the gray-arrowed path until expectation–maximization (EM) iterations are complete, then pink arrows are followed to the final composition estimate. The method starts by establishing probabilities for each alignment type *C* = *[mismatch (X), insertion (I), deletion (D), softclip (S)]* through occurrence counts in the primary alignments. Next, alignment probability *P(r|t)* is calculated for each read, taxonomy pair *(r,t)* by assuming the maximum alignment probability between *r* and *t*. Meanwhile, an evenly distributed composition vector *F* is initialized. The EM phase is entered by determining *P(t|r)*, the probability that *r* emanated from *t*, for all *P(r|t)*. *F* is updated accordingly, and the total log likelihood of the estimate is calculated. If the total log likelihood is a significant increase over the previous iteration (>.01), then EM iterations continue. Otherwise, the loop is exited, and *F* is trimmed to remove all entries less than the set threshold. Now following the pink arrows, one final round of estimation is completed with the trimmed *F* to produce the final sample composition estimate.



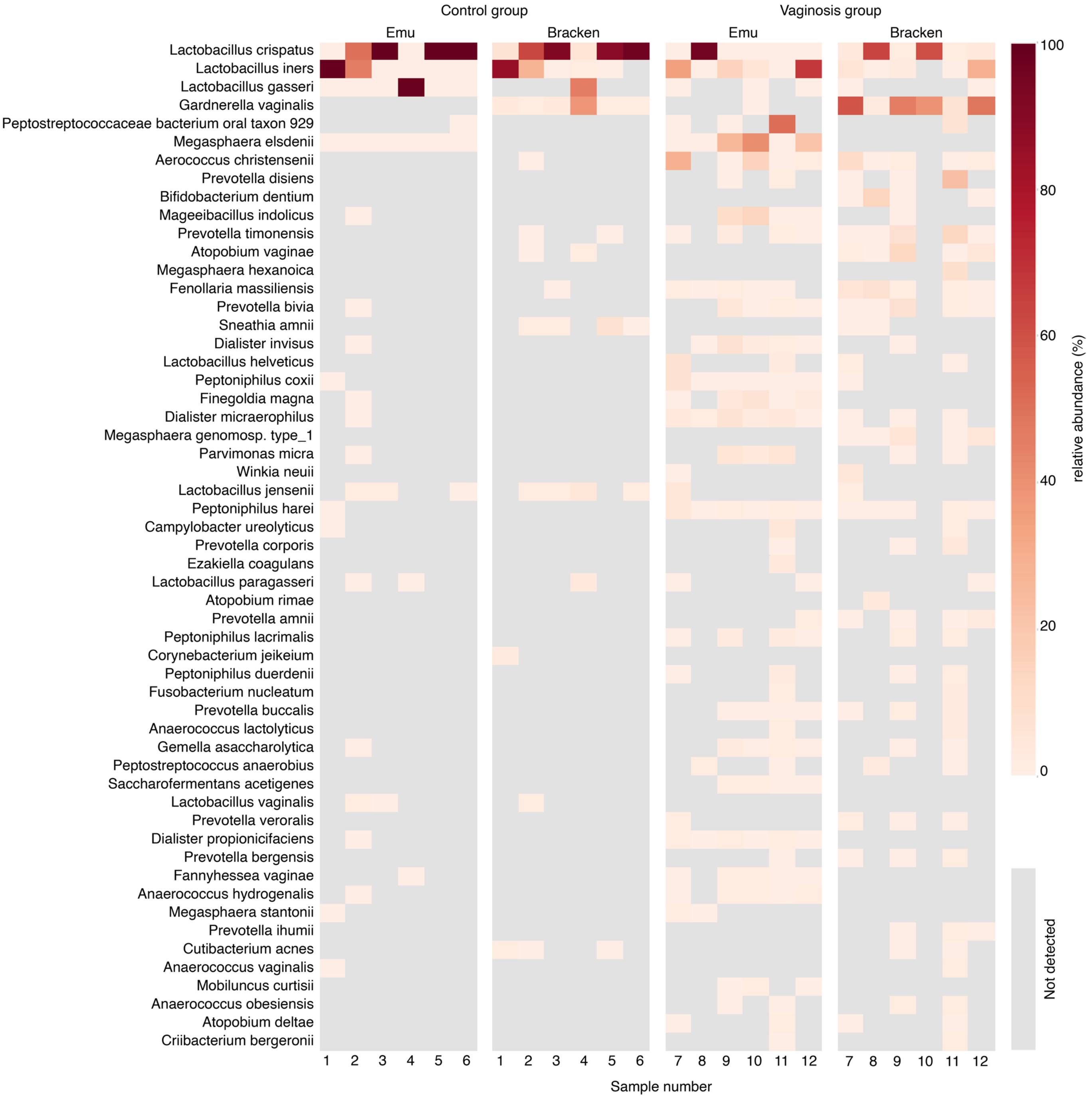
Extended Data Fig. 2 | ZymoBIOMICS theoretical and imputed ground truth community profiles. The theoretical values are taken from ZymoBIOMICS standard report of relative abundance estimates based on 16S rRNA gene copy numbers (https://files.zymoresearch.com/protocols/_d6305_d6306_zymobiomics_microbial_community_dna_standard.pdf). Truth_ONT and truth_illumina represent the ground truth relative abundances calculated for our ONT and Illumina datasets respectively, as described in the *Establishing Ground Truth* subsection under Methods.



Extended Data Fig. 3 | Performance on our synthetic gut microbiome mock community. Heatmap of species-level error between calculated ground truth and estimated relative abundances, where darker blue denotes an underestimate by the software, darker red denotes an overestimate, and white represents no error. All Oxford Nanopore Technologies (ONT) errors are measured in relation to the ground truth of the ONT dataset, while Illumina errors are measured in relation to the ground truth for the Illumina dataset. Color scheme is capped at ±10, resulting in error greater than ±10% observing the maximum error colors. Displayed are the 20 species claiming the largest abundance in any of the ONT or Illumina sample results. 'Other' represents the sum of all species not shown in figure for the respective column. Species-level L1-norm, L2-norm, precision, recall, and F-score are also plotted for the methods evaluated.



Extended Data Fig. 4 | Family-level relative abundance error heatmap of novel species simulation. Heatmap of family-level error between ground truth and estimated relative abundances for both the Emu and RDP incomplete databases (missing 35 of the 345 CAMI2 simulated species) with our CAMI2 dataset. Here, darker blue denotes an underestimate by the software, darker red denotes an overestimate, and white represents no error. Color scheme is capped at ±3, resulting in error greater than ±3% observing the maximum error colors. Displayed are the families of the 35 species that were removed from each of the databases.



Extended Data Fig. 5 | Bacterial community of 12 vaginal samples. Species with estimated abundance of over 1% in at least one sample with either Emu or Bracken are shown. Data is grouped by condition: healthy control or vaginosis.

nature research

Corresponding author(s): Todd J. Treangen, Alexander DIlthey, Kristen Curry

Last updated by author(s): April 7, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

☐ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☒ ☐ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☒ ☐ A description of all covariates tested

☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☒ ☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☒ ☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☒ ☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	DeepSimulator v1.5 and NanoSim v2.2 were used to simulate nanopore reads. Generation of sequences for our in-house mock community used the NEXTflex 16S V4 Amplicon-Seq Kit 2.0 and MiSeq platform for Illumina sequences, while 16S Barcoding Kit 1-24 (SQK-16S024) and MINKNOWN 19.12.5 were used for Oxford Nanopore Technologies sequences.
Data analysis	Guppy v4.4.2, Trimmomatic v0.39, Emuv2.0.1, Minimap2 v2.22, Kraken 2 v2.1.1, Bracken v2.5.0, NanoCLUST v1.9, Centrifuge v1.0.4, MetaMaps v0.1, QIIME 2 v2020.11, and BWA v0.7.17 were used to analyze data in this study. CAMISIM2 and MBARC-26 were used to establish microbial abundances for benchmarking simulated datasets. All scripts and data used to compile quantitative comparison results can be found on GitLab: https://gitlab.com/treangenlab/emu-benchmark.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All sequenced samples used in this study are publicly available on Sequence Read Achieve (SRA). Both ZymoBIOMICS datasets are under BioProject ID PRJNA587452 with SRA accessions SRR10391201 for ONT and SRR10391187 for Illumina.
Our gut mock community is under BioProject ID PRJNA725207. The 12 vaginal samples used for our real-world application

demonstration are uploaded under BioProject ID PRJNA723982. Additionally, our simulated sequences are publicly available on OSF under project 56UF7 (https://osf.io/56uf7/). Study of vaginal microbiomes was IRB-approved by the ethics committee of the Medical Faculty of Heinrich Heine University. Databases used in this manuscript include: 16S RefSeq nucleotide sequence records (https://www.ncbi.nlm.nih.gov/refseq/targetedloci/16S_process/), Ribosomal Database Project (RDP) version 11.5 (https://rdp.cme.msu.edu/) and rrnDB version 5.7 (https://rrndb.umms.med.umich.edu/).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences □ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

Evaluate performance of our new software Emu by comparing microbial composition accuracy produced from a variety of existing software tools on 4 different 16S rRNA datasets. Then, demonstrate application of Emu on clinical samples.

Research sample

4 communities were used for our software comparison. First, 2 simulated nanopore reads to evaluate accuracy of tools without bias generated in a real-world setting. The first simulated community (MBARC-26) was selected as an established microbial mock community (https://doi.org/10.1038/sdata.2016.81). The second (CAMI2 Mouse Gut) was selected as a more challenging stress test for Emu given that it has a larger species count (353), and specifically has similar species with vastly different abundances. Next, ZymoBIOMICS mock community standard sequenced by both Illumina and Oxford Nanopore workflows to evaluate software tools on an established community while incorporating the variation and contamination that is introduced in a real lab setting. Then, an in-house community produced to mimic the gut microbiome and challenge the software tools. This acted as a final stress test as it is subject to variation from culturing and sequencing, while including a larger number of species and abundance variation than the ZymoBIOMICS community. Finally, samples from 12 clinical vaginomes were used to demonstrate a real-world application of Emu and compare results to whole-genome sequences. The 12 samples were selected (6 from patients with vaginosis and 6 from patients without vaginosis) randomly from an ongoing exploratory clinical study comparing the vaginal microbiome between these two cohorts. The sample is to represent the population women most susceptible to bacterial vaginosis. Our samples in particular are from women ages 27-43 for the vaginosis group and 21-48 for the controls. The rationale for this study is that previous research raises belief that there may be a difference in the vaginal microbiome between women with vaginosis and those without. Participation in the exploratory clinical microbiome study was offered to patients presenting at the German Centre for Infections in Gynecology and Obstetrics at Helios Hospital Wuppertal (Germany) at the study's lead physician's discretion.

Sampling strategy

Datasets were chosen to demonstrate a range of communities with different challenges. The simulated data is used to strictly isolate the community profiling software, while the gut microbiome mock community demonstrates a real-world scenario of a sample containing a species that is not in the database.

Data collection

Collection of our gut microbiome mock community was completed by Qinglong Wu in the Tor Savidge Lab in the Department of Pathology and Immunology at Baylor College of Medicine. DNA extraction was completed with MasterPure™ Complete DNA and RNA Purification Kit with extra lysozyme lysis. Clinical vaginal samples were collected by the German Center for Infections in Gynaecology and Obstetrics at Helios University Clinic Wuppertal under Dr. Werner Mendling; Total DNA and RNA was extracted using ZymoBIOMICS DNA/RNA Miniprep Kit R2002.

Timing and spatial scale

Sequences from the ZymoBIOMICS community were generated in 2019. Sequences from our in-house gut microbiome mock community were generated in summer 2020; samples for both datasets from this community were taken at the same time. Clinical vaginal samples started collection in 2020 and terminated in 2021.

Data exclusions

No data was excluded.

Reproducibility

The main finding in this study is that Emu outperformed existing software tools for accurate community profiling from 16S reads. This is demonstrated through 3 communities in the paper, and no other dataset comparisons were performed.

Randomization

Clinical vaginal samples were divided into two groups: those diagnosed with vaginosis and those without. No randomization was required for this separation.

Blinding

Since patients were not given a treatment, blinding was not needed for the clinical samples in this study. Since the samples were not collected in the context of an interventional study, no blinding was necessary or carried out.

Did the study involve field work? □ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
□	☒ Human research participants
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics

Female; 6 patients with diagnosed bacterial vaginosis (age 27 - 43), 6 without (age 21-48).

Recruitment

12 vaginal swabs were collected at the German Centre for Infections in Gynecology and Obstetrics at Helios Hospital Wuppertal (Germany).   Participation in an exploratory clinical microbiome study was offered to patients presenting at the German Centre for Infections in Gynecology and Obstetrics at Helios Hospital Wuppertal (Germany) at the study's lead physician's discretion. No sample characteristics, apart from diagnosis of bacterial vaginosis, were taken into account for sample selection. Neither the observed microbiome community state types nor the reported comparison between Emu and Bracken are expected to be influenced by self-selection or other potential biases.

Ethics oversight

All patients samples were collected with informed consent from individuals in the context of an exploratory clinical microbiome study approved by the Ethics Committee of the Medical Faculty of Heinrich Heine University Düsseldorf (IRB study identification "2019-600-andere Forschung erstvotierend"). Ethics Committee of the Medical Faculty of Heinrich Heine University Düsseldorf

Note that full information on the approval of the study protocol must also be provided in the manuscript.