nature biotechnology

LETTERS

https://doi.org/10.1038/s41587-020-00777-4

Check for updates

Improved metagenome binning and assembly using deep variational autoencoders

Jakob Nybo Nissen[1,2], Joachim Johansen[2], Rosa Lundbye Allesøe[2], Casper Kaae Sønderby[3], Jose Juan Almagro Armenteros[1], Christopher Heje Grønbech[3,4], Lars Juhl Jensen[2], Henrik Bjørn Nielsen[5], Thomas Nordahl Petersen[6], Ole Winther[3,4,7] and Simon Rasmussen[2]✉

Despite recent advances in metagenomic binning, reconstruction of microbial species from metagenomics data remains challenging. Here we develop variational autoencoders for metagenomic binning (VAMB), a program that uses deep variational autoencoders to encode sequence coabundance and *k*-mer distribution information before clustering. We show that a variational autoencoder is able to integrate these two distinct data types without any previous knowledge of the datasets. VAMB outperforms existing state-of-the-art binners, reconstructing 29–98% and 45% more near-complete (NC) genomes on simulated and real data, respectively. Furthermore, VAMB is able to separate closely related strains up to 99.5% average nucleotide identity (ANI), and reconstructed 255 and 91 NC *Bacteroides vulgatus* and *Bacteroides dorei* sample-specific genomes as two distinct clusters from a dataset of 1,000 human gut microbiome samples. We use 2,606 NC bins from this dataset to show that species of the human gut microbiome have different geographical distribution patterns. VAMB can be run on standard hardware and is freely available at https://github.com/RasmussenLab/vamb.

Metagenomic binning is the process of grouping metagenomic sequences by their organism of origin[1,2]. In metagenomic studies, binning allows the reconstruction of known and unknown genomes, enabling a broad description of the community and creating a starting point for further analysis of the organisms[3]. We developed a binning tool that uses deep learning in the form of variational autoencoders (VAE)[4,5] that integrates coabundance[6] and *k*-mer composition[7] data from metagenomics de novo assemblies and clusters the resulting latent representation into genome clusters and sample-specific bins. Our approach leverages multiple samples while simultaneously avoiding between-sample chimeras. It outperforms commonly used single-sample binning approaches by reconstructing 29–98% more NC genomes from simulated datasets, as well as 45% more NC genomes from a dataset of 1,000 human gut microbiome samples. Furthermore, our clustering method automatically groups per-sample bins into clusters with high taxonomic consistency, allowing precise strain-resolution taxonomic profiling.

Earlier work on metagenomics binning has mainly relied on the principles that DNA sequences originating from the same organism will have high covariance of their abundance signal across samples (coabundance) and that they share similar patterns of *k*-mer usage in their DNA (for example, 2–5 mer)[7–15]. Several attempts have been made to reconstruct thousands of microbial species from massive metagenomics datasets[16–18], independently assembling and binning each sample into genomes. These simple workflows allow for parallel analysis of samples, but do not leverage coabundance. Typical workflows using coabundance deal with sequence redundancy by either coassembling distinct samples or deduplicating sequences before binning[6,8–12]. This leads to intersample chimeric genomes that do not exist, which is especially problematic when strain-level variation can have important biological implications[19]. Furthermore, none of the existing methods leverage deep learning.

The main difference between our method, VAMB, and others is that it utilizes an unsupervised deep learning approach known as a VAE[10,18]. Second, our approach clusters the combined contig dataset from all samples without any preclustering or homology reduction and applies a strategy for splitting genome clusters after clustering (Supplementary Figs. 1–4). When applying this approach, which we term 'multisplit', each cluster should correspond to an organism and each bin in a cluster to a per-sample representation of the genome of that organism. To demonstrate the performance of VAMB compared to other binners, we benchmarked VAMB, Canopy[6], MetaBAT2 (ref. [12]) and MaxBin2 (ref. [11]) on five synthetic datasets from Critical Assessment of Metagenomic Interpretation (CAMI)[20] and one semisynthetic dataset from MetaHIT[21] samples (Supplementary Table 1). We assessed binning performance by counting the number of NC (>90% recall and >95% precision) genomes reconstructed as done in previous work[18,22]. VAMB reconstructed 29–98% more NC genomes at strain level compared to any of the other three binners (Fig. 1a and Supplementary Table 2). Interestingly, the increased performance of VAMB correlated (Pearson correlation coefficient = 0.90, linear regression $P = 0.035$) with the difficulty of the CAMI2 datasets, which we measured as the entropy of the genomes (Fig. 1b). Similarly, we found that VAMB reconstructed more genomes compared to MetaBAT2 at all levels of genome difficulty (Supplementary Fig. 5 and Supplementary Table 3). Additionally, we compared VAMB to ensemble binning, where bins from multiple programs are combined. Using DAS Tool[23], we tried combinations of the other binners and found VAMB to be better compared to all others (Fig. 1c). The addition of VAMB bins to DAS Tool improved the output of DAS Tool by up to 14% compared to VAMB only, but decreased performance on the Airways and MetaHIT datasets. VAMB and MetaBAT2 agreed on 39 NC genomes on average across the datasets and generated 49 and

[1]Department of Health Technology, Technical University of Denmark, Lyngby, Denmark. [2]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [3]Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark. [4]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark. [5]Clinical-Microbiomics A/S, Copenhagen, Denmark. [6]National Food Institute, Technical University of Denmark, Lyngby, Denmark. [7]Center for Genomic Medicine, Copenhagen University Hospital, Copenhagen, Denmark. ✉e-mail: simon.rasmussen@cpr.ku.dk

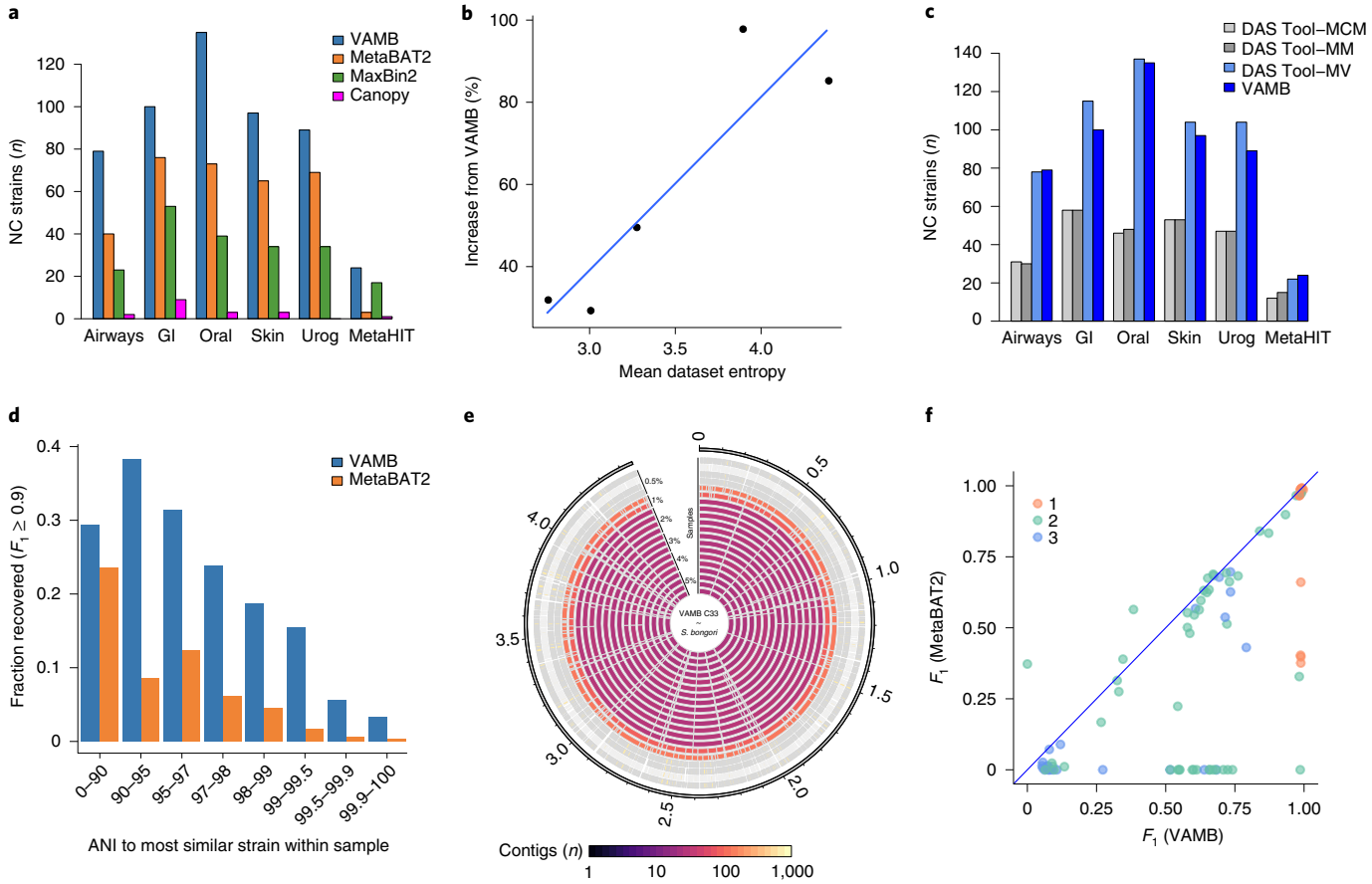

Fig. 1 | Performance of VAMB. a, Number of distinct NC strains recovered from the six benchmark datasets for VAMB (blue), MetaBAT2 (orange), MaxBin2 (green) and Canopy (magenta). **b**, Number of NC strains recovered by VAMB relative to MetaBAT2 as a function of mean sample entropy per dataset. Sample entropy was calculated as the Shannon entropy, with each contig an observation and each strain a class, and was used as a proxy for dataset complexity. **c**, Number of NC strains recovered when using the ensemble binner DAS Tool. We used the binning output from MetaBAT2, MaxBin2 and Canopy (DAS Tool–MCM, light gray), MetaBAT2 and MaxBin2 (DAS Tool–MM, gray), MetaBAT2 and VAMB (DAS Tool–MV, light blue) and VAMB (blue). **d**, Number of genomes recovered with F1 ≥ 0.9, stratified by the ANI to the most similar strain in the same sample across CAMI2 datasets. Blue, VAMB; orange, MetaBAT2. **e**, Alignment of sample-specific genome bins from VAMB cluster 33 to the *S. bongori* reference genome. The rings are ordered according to the number of *S. bongori* reads spiked into the HMP gut microbiome sample from 5% (inner) to 0.5% (outer), and colored according to the number of contigs in the particular sample. **f**, F1 of reconstructed genomes with VAMB and MetaBAT2 in the mixed-strain *Salmonella* spike-in experiment. A total of ten different *Salmonella* genomes were used, and between one and three genomes were added per HMP sample. Each dot represents F1 of a sample–genome pair and the color indicates how many *Salmonella* genomes were added to the particular sample: orange, 1; teal, 2, blue, 3. GI, gastrointestinal; urog, urogenital.

16 unique NC genomes on average, respectively. However, only a few NC genomes were unique to the combination (average, 1.6) and more NC genomes (average, 13) were lost (Supplementary Fig. 6).

To show that the superior performance of VAMB on strains was due to better binning, and not merely that VAMB defaults to a precision–recall tradeoff that happens to fit strain-level binning in our datasets, we tested the performance of VAMB at the species and genus levels. Here, VAMB on average reconstructed 14% more species than the second-best binner, MetaBAT2, which outperformed VAMB on only the CAMI2 Urogenital dataset. At genus level, VAMB and MetaBAT2 had similar performance with the former 4% better across all datasets, but MetaBAT2 outperformed VAMB on the CAMI2 Skin and Urogenital datasets (Supplementary Fig. 7 and Supplementary Tables 4 and 5). Furthermore, we tried to subsample the number of reads used for binning and found that VAMB performed well even with as few as 200,000 read pairs from each dataset (Supplementary Fig. 8).

One particularly difficult aspect of metagenomics binning is when multiple strains are present in a sample simultaneously. We therefore revisited the CAMI2 datasets, which contain a mixture of different strains and community complexities (Supplementary Fig. 9), and analyzed two simulated datasets originally created by Cleary et al.[24]. For the CAMI2 datasets we found that VAMB was able to bin more genomes with a weighted recall and precision ($F1 > 0.9$) across all intervals of strain abundances (Supplementary Fig. 10). Similarly, VAMB showed better performance across all intervals when assessing ANI than the most similar strain in the same sample. Here VAMB reconstructed 38% of all genomes as NC when the most similar strain had between 90–95% ANI, and even 15.5% of all genomes when the most similar strain had 99.0–99.5% ANI (Fig. 1d). For the Cleary et al. spike-in datasets, we first investigated a single spike-in with *Salmonella bongori* to a background of human gut microbiome samples[25]. Here VAMB created a single cluster (C33) where each sample-specific bin had $F1 = 1$ when 200,000 or more read pairs were added. Additionally, no contigs were assigned from any sample where we did not spike-in *S. bongori* reads, highlighting the ability of VAMB to group related strains across samples into a single cluster of sample-specific bins (Fig. 1e,

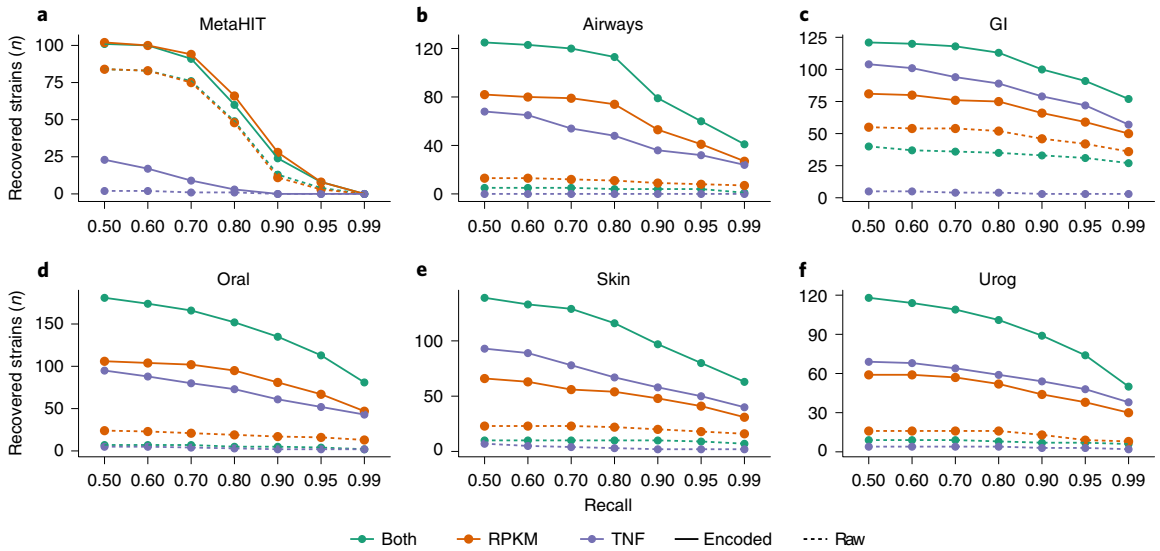

Fig. 2 | Performance of clustering different inputs. a–f, VAMB can effectively integrate coabundance and *k*-mer information (teal solid lines) to a clusterable representation that yields more reconstructed genomes than other combinations of the data, or using raw compared to encoded data. **a**, MetaHIT dataset. **b**, CAMI2 Airways dataset. **c**, CAMI2 Gastrointestinal (GI) dataset. **d**, CAMI2 Oral dataset. **e**, CAMI2 Skin dataset. **f**, CAMI2 Urogenital (Urog) dataset. Purple. *k*-mer frequency (TNF); orange, coabundance (RPKM); teal, concatenation of both *k*-mer and coabundance). Dashed lines, raw data input; solid lines, latent representation from the variational autoencoder in VAMB; *y* axis, number of distinct strains recovered at precision >0.95; *x* axis, increasing recall threshold of genomes.

Supplementary Fig. 11 and Supplementary Table 6). We then performed a second experiment where we spiked-in reads from ten different *Salmonella* genomes, with up to three *Salmonella* genomes per gut microbiome sample (Supplementary Data 1). Here, VAMB and MetaBAT2 were able to reconstruct 19 and 12 *Salmonella* strain–sample pairs, respectively, with $F1 > 0.9$ (Fig. 1f and Supplementary Data 2). As above, we quantified the ability of VAMB to distinguish between within-sample *Salmonella* genomes as a function of ANI. Here 14 genomes (78%) could be reconstructed ($F1 > 0.6$) when the other *Salmonella* genome had 90–91% ANI, eight genomes (57%) at 93–94% ANI and four (27%) at 98–99.5% ANI (Supplementary Fig. 12). Taken together, VAMB is able to distinguish between mixed strains at even 98–99.5% ANI, although the accuracy may be limited by the de novo assembly process for very similar genomes.

To test our hypothesis that the performance of VAMB stemmed in part from the VAE integrating information from both coabundance and *k*-mer composition, we compared the number of NC genomes produced by clustering of the raw coabundance data, raw *k*-mer composition or both raw datasets concatenated. Further, we compared to the bins produced by clustering their VAE latent spaces. For five of the six datasets, clustering the concatenation of raw data did not yield better results than the abundance or *k*-mer composition. However, for all datasets apart from MetaHIT, encoding of the concatenation gave the best results of all six input combinations, yielding 27 and 67% more NC genomes for our two validation datasets compared to the second-best combination (Fig. 2). Integrating the two data types with the VAE therefore results in a latent representation that is more informative than either of the inputs alone, and more amenable to clustering than the simple concatenation of the two raw data types. Furthermore, we investigated the effect of using different sizes of *k*-mers for encoding ($k = 2–5$) and, in line with previous work[7,10,26,27], found that $k = 4$ gave the best performance in three of the datasets (Supplementary Fig. 13). To test the importance of the probabilistic VAE encodings in VAMB, we tested a version of VAMB with the VAE replaced by a deterministic autoencoder. Here we found worse performance for all datasets, with the number of NC genomes from our two validation datasets dropping by 43 and 39%, respectively (Supplementary Fig. 14). We visualized the input space and latent encodings and, in line with our hypothesis, found that the VAE encoding appears to have genomes more clearly separated (Supplementary Fig. 15). Finally, we tested using *k*-means clustering rather than VAMB's iterative medoid clustering method. We found that VAMB's clustering algorithm was superior when using VAE-encoded data, and had the best overall performance compared to any combination of *k*-means clustering (Supplementary Fig. 16).

Single-sample binning workflows are popular because they are trivial to parallelize and inherently prevent intersample chimeras. We therefore tested the performance of VAMB on single samples of the CAMI2 datasets compared to MetaBAT2 and MaxBin2. While VAMB reconstructed most genomes on average, the differences were not significant (Wilcoxon rank-sum tests, two-tailed, $P > 0.05$) (Supplementary Fig. 17). We then compared the performance of the single-sample and multisplit approaches. Here, we found for all datasets that the multisplit approach was superior because the number of NC genomes rose from 1 to 24 for the MetaHIT dataset and increased by 28–105% for the five CAMI2 datasets (Supplementary Fig. 18). Importantly, using VAMB in multisplit mode was significantly better when measured across all datasets for as few as four samples (Supplementary Table 7 and Supplementary Data 3). Furthermore, we repeated the benchmark after discarding all 'easy' genomes (fewer than five contigs) and found that the improvement gain from multisplit was even more pronounced (109–282%; Supplementary Fig. 19). Because most metagenomics studies compare multiple samples of similar microbial communities with highly fragmented genomes, we expect that much higher-quality genomes can be recovered using VAMB and the multisplit approach.

One advantage of single-sample binning workflows is that they are inherently parallel and allow binning of large datasets[16–18]. To test the scalability of VAMB, we ran it on the entire benchmark dataset of Almeida and coworkers[18], consisting of 1,000 randomly selected human gut microbiome samples and a total of 5.9 million contigs (Supplementary Data 4). We used a single graphical processing unit (GPU) and ran VAMB in 12.4 h, 27 times faster than

running MetaBAT2 in single-sample mode (Supplementary Table 8 and Supplementary Fig. 20). To compare the quality of the resulting bins to those obtained by Almeida et al. using MetaBAT2, we estimated genome completeness and contamination and counted the number of bins estimated as NC. Using VAMB with default parameters, we obtained 5,036 NC bins compared to 3,480 for MetaBAT2, an increase of 1,556 NC bins (45%). We additionally tested VAMB with different hyperparameters and found that a slight decrease in the network yielded 5,288 NC bins, an increase of 252 NC bins (7.2% additional increase). For a fair comparison, we focused on the results of the default run and found that 2,517 of the NC bins were found by both methods whereas 1,019 and 1,500 NC bins generated by VAMB were medium quality (MQ) or missing from MetaBAT2, respectively. Similarly, MetaBAT2 reconstructed 480 and 483 NC bins that were MQ or missing from VAMB, respectively. Additionally, VAMB generated more MQ bins (5,169 versus 4,221) and therefore a larger number of MQ and NC bins in total (10,205 versus 7,701), as well as significantly more NC bins per sample (Wilcoxon signed-rank test, two-tailed, $V = 209{,}930$, $P = 7.1 \times 10^{-92}$) (Supplementary Fig. 21). Additionally, for the common set of 6,017 MQ or NC bins, median completeness, contamination and $F1$ were consistently better for VAMB ($F1 = 0.96$) compared to MetaBAT2 ($F1 = 0.94$) ($F1$, Wilcoxon signed-rank test, two-tailed, $V = 10{,}535{,}000$, $P = 7.8 \times 10^{-110}$) (Supplementary Fig. 22). However, because estimates of completeness based on conserved genes can be overestimated[28], we compared the sequence length of the common set of bins but found no significant difference (Wilcoxon signed-rank test, two-tailed, $V = 9{,}152{,}000$, $P = 0.20$) (Supplementary Fig. 22). Moreover, we compared NC bins that had an assembled genome at the National Center for Biotechnology Information (NCBI), and found that VAMB and MetaBAT2 bins were 10.5 and 14.3% shorter on average, respectively (Supplementary Fig. 23). For these bins we investigated the functional potential of extra contigs binned only by VAMB and found these to be highly enriched in Gene Ontology terms including 'Translation', 'Metal ion binding', 'Transposases' and more. Furthermore, we predicted phage-like contigs in 338 of 959 (35%) of these and identified a significantly increased AT nucleotide content compared to the entire bin (Wilcoxon rank-sum test, one-tailed, $W = 515{,}500$, $P = 2.4 \times 10^{-8}$) (Supplementary Fig. 23). Higher AT content is consistent with previous findings of horizontally transferred regions[29], and presumably reflects the ability of VAMB to recruit mobile genetic elements to the bins. Finally, we investigated taxonomic annotations and found a large overlap between the two sets. However, VAMB bins represented a larger taxonomic diversity from genus to genome level and, on average, reconstructed 97 more NC bins per phylum (Supplementary Fig. 24, Supplementary Tables 9 and 10 and Supplementary Data 5 and 6). Importantly, while VAMB is clearly better than MetaBAT2, Almeida et al. found no difference when using MetaBAT2 in single-sample mode, in coassembly mode or using the information from three different binners combined with MetaWRAP[30], replicating results from DAS Tool on the benchmark datasets.

As mentioned previously, another advantage of the multisplit approach is that a single cluster represents a particular organism across multiple samples. To test the phylogenetic consistency of clusters, we used 40 bacterial marker genes from the 5,036 NC bins to create a phylogeny (Fig. 3a). Here we found that 93.2% of clusters were monophyletic and that for 98.7% of the bins all leaves were extremely close to the cluster's central leaf, corresponding to >99% amino acid identity. Similar to the example with *S. bongori*, this implies that bins split from the same cluster are very closely related and represent different strains of the same species observed across samples. Zooming in on microdiversity, we analyzed the largest cluster, cluster 546, that contained 255 NC, 94 MQ and 115 low-quality bins. We found high taxonomic consistency with 92% of all contigs assigned to *B. vulgatus* and 5.7% to other *Bacteroides* species with slightly lower identity (Supplementary Data 7). These bins therefore represent 349 (NC and MQ) different individually de novo reconstructed *B. vulgatus* genomes in 349 human gut microbiomes. If we compare this to using an approach based on ANI > 95%, such as used in other large-scale, single-sample binning studies[17,18], *B. vulgatus* and *B. dorei* would have been merged into one species (Supplementary Fig. 25) rather than clusters 546 and 94, respectively.

One advantage of VAMB reconstructing more NC bins is increased statistical power when investigating associations with metadata. We therefore reconstructed the phylogeny of the 255 NC bins from the *B. vulgatus* cluster (Fig. 3b and Supplementary Fig. 26). We verified the phylogeny by considering samples ($n = 18$) with multiple sequencing runs ($n = 40$) and found the samples to be placed either monophyletically or with very short distances between them (Fig. 3b and Supplementary Fig. 26). When comparing phylogenetic placement of *B. vulgatus* strains to the recorded metadata, we found phylogenetic distance to be significantly associated with the geographical location of the sample (permutational multivariate analysis of variance (PERMANOVA), adjusted $P = 0.007$, $F = 2.26$, degrees of freedom = 216), although only at a low coefficient of determination ($R^2 = 0.02$). European and North American samples did not cluster exclusively, and Asian *B. vulgatus* strains were interspersed throughout the entire tree. Previous work comparing North America and Europe found a similar trend for *B. vulgatus*[31], although another study investigating Ca. *Cibiobacter qucibialis* found a clade associated with Chinese samples[17]. Furthermore, previous work based on sample taxonomy and community structure has shown a clear association with geographical location of the sample[17,32–34]. We therefore expanded our analysis to all clusters with 20 or more NC bins ($n = 52$ and $n = 2{,}606$ NC bins in total) and found significant association with geographical location for 34 of the 52 clusters (adjusted $P < 0.05$) (Fig. 3c and Supplementary Data 8). However, the effect of geographical location (R^2) was markedly different between the clusters and was not associated with whether they corresponded to known or unknown species. We found clear differences between the families (Fig. 3d) and Bacteroidaceae, which included six different *Bacteroides* species, had the lowest overall association (median $R^2 = 0.03$) (Supplementary Fig. 27). On the contrary, taking species of Bifidobacteriaceae and Lachnospiraceae as examples, these showed much higher variance in R^2, from 0.06 to 0.50 and 0.08 to 0.44, respectively (Fig. 3d and Supplementary Fig. 27). These results indicate that strains of certain gut microbiome species are ambiguously distributed whereas others are geographically restricted. This could be due to either diet or to how well a species adapts to differences in host genetics, but could also be influenced by difference in transmission mode—for instance, vertical transmission (mother–child inheritance)[35,36].

Here, by combining metagenomics binning with unsupervised deep learning, we show improvements compared to state-of-the-art methods across datasets of different types and sizes. We also show that the VAE automatically learns how to integrate two distinct data types—in this case, coabundance and *k*-mer composition—and that the resulting latent representation clusters better than either of the inputs. This is, in principle, not limited to two input data types and it is possible to add additional data as input to the VAE. For VAMB we avoided using more complex models such as, for example, Gaussian mixture VAEs[37,38], and designed our method to be feasible for standard users. For instance, a standard laptop without GPU acceleration could process the six benchmark datasets each in <6 h with 1 GB random-access memory (Supplementary Table 8). Finally, we believe that the importance of our findings is not limited to the field of microbiome and metagenomics, because data integration is a central process in many fields of life science research. Future discoveries within precision medicine will be greatly enhanced by

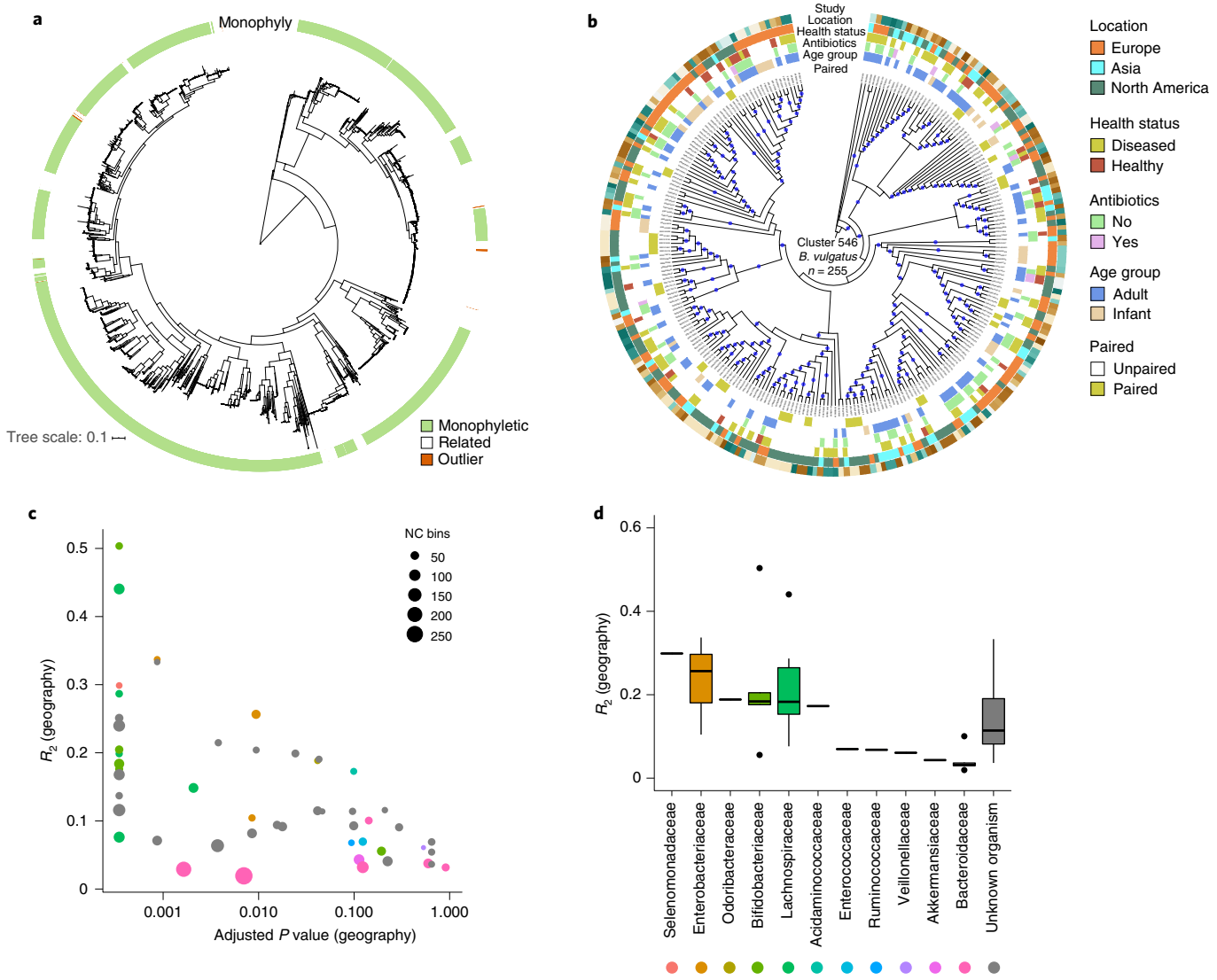

Fig. 3 | Phylogeny of bins across 1,000 human gut microbiome samples. a, Amino acid-level bacterial marker gene maximum likelihood tree for all 5,036 NC bins. Despite VAMB having no phylogenetic information, a large majority of bins are monophyletic and misplaced bins are generally very similar to neighboring clades. Green, leaf is in a monophyletic or extremely closely related bin (>99% aa identity); white, leaf is in a cluster with one or more outliers; orange, leaf is an outlier compared to the medoid of the cluster. **b**, Cladogram of ASTRAL species tree generated from 2,433 gene trees from cluster 546 containing 255 NC bins of *B. vulgatus*. ASTRAL local posterior probabilities branch support is indicated as a blue circle when support is >0.95. The tree is rooted on sample SRR341600, which is the most basal *B. vulgatus* in the CheckM tree (**a**). Rings, from inner to outer: 1, paired samples are in the same clade (green), not paired (white) or paired and not in same clade (red); 2, age group—infant (beige) or adolescent/adult (blue); 3, individual used antibiotics (pink) or not (green); 4, individual has a disease (green) or is healthy (red), 5, geographical origin: Europe (orange), Asia (turquoise), Americas (teal); 6, study origin (multiple colors). White in rings 2–6 (no color) indicates missing data. **c**, Association of phylogenetic distance and geographical location of sample. PERMANOVA *P* values adjusted using Benjamini–Hochberg for geography are shown on the *x* axis, and the coefficient of determination (R^2) is shown on the *y* axis. Point sizes indicate the number of NC bins in the cluster, and colors indicate family (see **d** for color coding). Unannotated clusters (<50% annotated with BLAST) are set as unknown (gray). Values left-most on the *x* axis indicate adjusted $P < 4 \times 10^{-4}$. **d**, We summarized R^2 from **c**, showing that geography explains different amounts of variation for families. Color coding for each family is shown as points below the *x* axis. **c**,**d**, Exact *P* values, *F*-statistic, degrees of freedom and number of observations (*n*) are available in Supplementary Data 8. The lower and upper hinges correspond to the first and third quartiles (25th and 75th percentiles). The upper and lower whiskers extend from the hinge to the highest and lowest values, respectively, but no further than 1.5× interquartile range (IQR) from the hinge. IQR is the distance between the first and third quartiles. Data beyond the ends of whiskers are outliers and are plotted individually.

data integration across several omics datasets. To achieve this, deep learning methods such as VAEs or other models represent promising approaches.

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/s41587-020-00777-4.

Received: 6 December 2019; Accepted: 17 November 2020;
Published online: 4 January 2021

References

1. Turaev, D. & Rattei, T. High definition for systems biology of microbial communities: metagenomics gets genome-centric and strain-resolved. *Curr. Opin. Biotechnol.* **39**, 174–181 (2016).
2. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
3. Wang, J. & Jia, H. Metagenome-wide association studies: fine-mining the microbiome. *Nat. Rev. Microbiol.* **14**, 508–522 (2016).
4. Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. Preprint at https://arxiv.org/abs/1312.6114 (2014).
5. Rezende, D. J., Mohamed, S. & Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *Proc. Mach. Learn. Res.* **32**, 1278–1286 (2014).
6. Nielsen, H. B. et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).
7. Teeling, H., Meyerdierks, A., Bauer, M., Amann, R. & Glöckner, F. O. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.* **6**, 938–947 (2004).
8. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
9. Albertsen, M. et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**, 533–538 (2013).
10. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ.* **3**, e1165 (2015).
11. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
12. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ.* **7**, e7359 (2019).
13. Plaza Oñate, F. et al. MSPminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. *Bioinformatics* **35**, 1544–1552 (2019).
14. Lin, H. H. & Liao, Y. C. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Sci. Rep.* **6**, 24175 (2016).
15. Chatterji, S., Yamazaki, I., Bai, Z. & Eisen, J. A. CompostBin: A DNA Composition-Based Algorithm for Binning Environmental Shotgun Reads. in Research in Computational Molecular Biology (eds. Vingron, M. & Wong, L.) 17–28 (Springer, 2008).
16. Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
17. Pasolli, E. et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176**, 649–662 (2019).
18. Almeida, A. et al. A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).
19. Brooks, B. et al. Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome. *Nat. Commun.* **8**, 1–7 (2017).
20. Sczyrba, A. et al. Critical Assessment of Metagenome Interpretation – a benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).
21. Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
22. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
23. Sieber, C. M. K. et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).
24. Cleary, B. et al. Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nat. Biotechnol.* **33**, 1053–1060 (2015).
25. Huttenhower, C. et al. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
26. Saeed, I., Tang, S.-L. & Halgamuge, S. K. Unsupervised discovery of microbial population structure within metagenomes using nucleotide base composition. *Nucleic Acids Res.* **40**, e34 (2012).
27. Pride, D. T., Meinersmann, R. J., Wassenaar, T. M. & Blaser, M. J. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res.* **13**, 145–156 (2003).
28. Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. Accurate and complete genomes from metagenomes. *Genome Res.* **30**, 315–333 (2020).
29. Daubin, V., Lerat, E. & Perrière, G. The source of laterally transferred genes in bacterial genomes. *Genome Biol.* **4**, R57 (2003).
30. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).
31. Schloissnig, S. et al. Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–50 (2013).
32. Yatsunenko, T. et al. Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
33. Deschasaux, M. et al. Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography. *Nat. Med.* **24**, 1526–1531 (2018).
34. He, Y. et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat. Med.* **24**, 1532–1535 (2018).
35. Asnicar, F. et al. Studying vertical microbiome transmission from mothers to infants by strain-level metagenomic profiling. *mSystems* **2**, e00164–16 (2017).
36. Ferretti, P. et al. Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. *Cell Host Microbe* **24**, 133–145 (2018).
37. Grønbech, C. H. et al. scVAE: variational auto-encoders for single-cell gene expression data. Bioinformatics 36, 4415–4422 (2020).
38. Dilokthanakul, N. et al. Deep unsupervised clustering with Gaussian mixture variational autoencoders. Preprint at https://arxiv.org/abs/1611.02648 (2017).

Methods

Overview of VAMB. The input to the VAMB pipeline is (1) a catalog of metagenomic sequences to be binned and (2) their abundances. The VAMB pipeline consists of three major steps (Supplementary Fig. 1). For each sequence in the catalog, the per-sequence tetranucleotide frequencies (TNF) for all possible canonical tetramers are calculated and the abundance of each sequence is estimated based on read mappings. These tables are concatenated and used to train a VAE tabula rasa (Supplementary Fig. 2). After training, the DNA sequences and coabundance information of the sequence catalog are encoded to the mean of their latent distributions. This latent representation is then clustered through an online iterative medoid clustering algorithm that dynamically estimates clustering threshold in cosine distance space. VAMB can be run in three different workflows: single-sample approach, where each sample is binned independently, a multisample approach on a coassembly or a multisplit approach (Supplementary Fig. 3). For the single-sample approach, normalized abundances are used whereas intersample abundance ratios (coabundance) are used for the multisample approaches. Finally, in the multisplit approach each cluster is split into sample-specific bins of the particular organism (Supplementary Fig. 4).

Computation of abundance and TNF. For each sequence, the frequencies of each tetramer not containing ambiguous bases were calculated to obtain TNFs[7]. TNFs were projected into a 103-dimensional orthonormal space as done in other work[39]. Thus, for n sequences the output was a table, $n \times 103$. To determine abundance, we counted the number of individual reads mapped to each sequence. If a read was mapped to n sequences, it counted $1/n$ towards each. The read counts were normalized by sequence length and total number of mapped reads, such that abundance was given in reads per kilobase sequence per million mapped reads (RPKM). With s samples and n sequences, the abundance output was a table, $n \times s$. Abundance values were normalized across samples to sum to 1, mimicing a probability distribution that was reconstructed from the final VAE by applying softmax to the abundance output neurons. Finally, TNFs were normalized by z-scaling each tetranucleotide across the sequences to increase the relative intersequence variance.

Architecture of the VAE. Each sequence was input to the VAE as an abundance vector $\mathbf{A}_{\text{in}}$ of length s and a TNF vector $\mathbf{T}_{\text{in}}$ of length 103 (Supplementary Fig. 2). These were concatenated to a vector of length $s+103$ before being passed through the hidden encoding layers consisting of two fully connected layers, each using batch normalization[40] and dropout[41] ($P=0.2$). The output of the last layer was passed to two different, fully connected layers of length N_{L}, termed the μ and σ layers. The latent layer, l, is of length N_{L} obtained by sampling the Gaussian distribution using the μ and σ layers as parameters—for example, $l_i \sim N(\mu_i, \sigma_i)$ for each neuron $i = 1,\ldots,N_{\text{L}}$. The sampled latent representation was then passed through the hidden decoding layers, identical in size to the hidden encoding layers except arranged in reverse order. Finally, the last hidden decoding layer was connected to a $s+103$ fully connected layer, which was split into two output vectors, $\mathbf{A}_{\text{out}}$ and $\mathbf{T}_{\text{out}}$, of length s and 103, respectively. We used leaky rectified linear units[42] as activation functions except for the μ and σ layers, which used linear and softplus activation, respectively. Furthermore, for the last layer generating the reconstructions we used softmax activation for generating $\mathbf{A}_{\text{out}}$ mimicking a probability distribution and linear activation for $\mathbf{T}_{\text{out}}$. After training, the input sequences were encoded by passing them through the VAE and extracting the values of the μ layer. The VAE models were trained using the Adam optimizer[43] and one Monte Carlo sample of the Gaussian latent representation. The VAE was implemented using PyTorch[44] (v.1.2.0), and CUDA (v.10.1.243) was used when running on a GPU.

Loss function. When training the VAE with s samples and N_{L} hidden neurons, the failure to reconstruct the input was penalized by the reconstruction error, consisting of an abundance error (E_{ab}) and a TNF error (E_{TNF}), defined as

$$E_{\text{ab}} = \sum \ln\left(\mathbf{A_{out}} + 10^{-9}\right)\mathbf{A_{in}},\ E_{\text{TNF}} = \sum\left(\mathbf{T_{out}} - \mathbf{T_{in}}\right)^2$$

that is, using cross-entropy (CE) and the sum of squared errors (SSE), respectively. When running on a single sample, E_{ab} was defined using SSE, because CE on a single normalized value trivially is zero. To regularize the model, the distribution given by the μ and σ layers was constrained by a prior $N(0,\text{I})$, by penalizing the deviance from this distribution with the Kullback–Leibler divergence:

$$D_{\text{KL}}(\text{latent}|\text{prior}) = -\sum\frac{1}{2}\left(1 + \ln(\sigma) - \mu^2 - \sigma\right)$$

Finally, the combined model loss was then

$$L = w_{\text{ab}}E_{\text{ab}} + w_{\text{TNF}}E_{\text{TNF}} + w_{\text{KLD}}D_{\text{KL}}$$

where the weighting terms are defined as $w_{\text{ab}} = (1-\alpha)\ln(s)^{-1}$, $w_{\text{TNF}} = \alpha/103$ and $w_{\text{KLD}} = (N_{\text{L}}\beta)^{-1}$. The parameters α and β were set to 0.15 and 200, respectively. For values of loss, E_{ab}, E_{TNF} and D_{KL} represent the six benchmark datasets (Supplementary Fig. 20).

Clustering. Clustering of the latent space was done using an iterative medoid clustering algorithm inspired by Nielsen et al.[6] based on cosine distances between encodings. The algorithm works in two steps (Supplementary Fig. 4): (1) an arbitrary point is chosen to be medoid. The medoid's 'neighbors' are defined as any points within a distance of 0.05 in cosine distance space. VAMB then randomly samples points from the neighbors and, if any point has more neighbors than the medoid, this becomes the new medoid. When VAMB has futilely sampled 25 neighbors in a row or tried all neighbors, go to step 2. (2) The distances from the medoid to all other points are calculated and a histogram is created. A heuristic function checks whether the histogram is composed of a 'near' peak of close points and a 'far' peak of further points separated by a deep valley with fewer points in intermediate distance from the medoid. 'Deep' is initially defined as the valley minimum being <0.1× the maximum of the small peak. If a deep valley is found, all points closer than the valley minimum are removed as a cluster; if not, the medoid is ignored. VAMB checks how often a medoid has been ignored: if >185 of the last 200 tries, the definition of 'deep' is increased by 0.1; if 'deep' is already 0.6, VAMB will ignore the valley's minimum and instead remove all points within an adaptive cosine distance as a cluster. This distance is determined as the median distance from all previous clusters. The method was implemented for both central processing unit (CPU) and GPU usage.

Benchmarking datasets. We used four training and two holdout datasets. One training dataset was the MetaHIT 'error-free' dataset ($n=264$) originally created by Kang and coworkers[10] while the other three were datasets from CAMI[20], where we used the sample-specific assemblies from three of the five CAMI2 'toy' human short-read datasets: CAMI2 Airways ($n=10$), CAMI2 Oral ($n=10$) and CAMI2 Urogenital ($n=9$). Our holdout datasets were the other two, CAMI2 Skin ($n=10$) and CAMI2 Gastrointestinal ($n=10$). We originally also tested VAMB on the CAMI High dataset ($n=5$) but, due to an unrealistic contig size distribution (Supplementary Fig. 28) influencing both abundance and TNF estimation, we discarded the dataset (see Supplementary Table 1 for an overview of the datasets). For all datasets we used only contigs >2,000 base pairs (bp) as input to VAMB. For the MetaHIT error-free dataset we used the abundance table supplied from Kang and coworkers (originally created using the script jgi_summarize_bam_contig_depths from MetaBAT) and the contigs as input to VAMB with default parameters. For each of the CAMI2 datasets we aligned the synthetic short paired-end reads from each sample using bwa-mem (v.0.7.15)[45] to the concatenation of per-sample contigs from the particular dataset. BAM files were sorted using samtools (v.1.7)[46] and abundances calculated using jgi_summarize_bam_contig_depths from MetaBAT2 (v.2.10.2)[12]. The jgi-abundance table and contig sequences were input to VAMB and run using default parameters with bin splitting enabled.

Benchmarking. When benchmarking a set of bins against a set of genomes, we matched each bin with each genome and defined the number of nucleotides in the genome covered by any contig from the bin as true positives. The total number of covered nucleotides of other genomes from contigs in that bin represented the false positives, and number of nucleotides in the genome that were covered by any contig in the dataset, but not by any contig in the bin, represented the false negatives. A genome was considered recovered at a particular recall–precision threshold pair if any bin matched with the genome reached or exceeded those precision and recall thresholds. For the CAMI2 datasets we used their definitions of strain, species and genus taxonomic levels[20]. Sczyrba et al. aligned extracted marker genes for each genome and aligned these to a 16S RNA alignment, clustered the alignment and then assigned a taxonomy based on that clustering (see Supplementary Note 1 of Sczyrba et al.'s paper for details)[20]. For the MetaHIT dataset, strain was defined as the individual reference genomes that were used to create the dataset, while species and genus levels were defined using NCBI taxonomy of the given reference genome. When comparing the performance of VAMB with other binners, we used Canopy from the original version (published in 2014) and ran it with default parameters. MetaBAT2 (v.2.10.2)[12] was run with default parameters, except setting–minClsSize=1, so that it would not discard small but accurate bins. MaxBin2 (v.2.2.4)[11] was run with default parameters. For all runs we used default parameters of VAMB as determined in the hyperparameter searches. For DAS Tool (v.1.1.1)[23] we used a combination of bins as input and default parameters. Benchmarking of subsampling reads for the CAMI2 datasets was done by randomly sampling reads for each sample to between 200,000 and 10 million read pairs and then running VAMB and benchmarking as described above. For comparison to k-means we used minibatch k-means[47] implemented in scikit-learn: 'MiniBatchKMeans (n_clusters=750, random_state=0, batch_size=4096, max_iter=25, init_size=20000, reassignment_ratio=0.02)'.

Hyperparameter search. To identify the best hyperparameters of the VAE, we developed it using four training datasets (MetaHIT, CAMI2 Oral, CAMI2 Airways and CAMI2 Urogenital) and used two other datasets (CAMI2 Skin and CAMI2 Gastrointestinal) as held-out test sets. We first varied each hyperparameter while keeping the others fixed and assessed the resulting bins. In the second round of optimization, we tested various hyperparameter combinations to select the final, best-performing values (Supplementary Figs. 29–32 and Supplementary Data 9). For single-sample analyses we used 256 neurons in two hidden layers with

32 latent neurons, no dropout, minibatch size of 128 and doubling after 25, 75, 150 and 300 epochs to a final total of 2,048, a learning rate of 10^{-3} and trained for 500 epochs. For the multisample approaches we similarly used default parameters, which were 512 neurons in two hidden layers with 32 latent neurons, 0.2 dropout, minibatch size of 128 that doubled after 25, 75, 150 and 300 epochs to a final total of 2,048, a learning rate of 10^{-3} and trained for 500 epochs. For the effect of the number of epochs, see Supplementary Fig. 33.

Strain-mixing datasets. We replicated the two *Salmonella* spike-in simulation experiments from Cleary et al.[24]. For the spike-in of a single genome we used *S. bongori* NCTC 12419 (NC_015761), where we simulated Illumina paired-end reads using ART (v.2.5.8)[48]. The reads were simulated as 100-nt error-free pairs with 300-nt insert size and a standard deviation of 10 nt. The reads were added in amounts of between 100,000 and 1 million read pairs (0.5–5% of total reads) to 30 different Human Microbiome Project (HMP) human gut microbiome samples, so that the total was 20 million read pairs (Supplementary Table 6). For the mixed-strain spike-in dataset we used three *Salmonella enterica* subsp. *enterica* serovar *Typhimurium* genomes (NC_003197, NC_016810, NC_022544), five non-Typhimurium *S. enterica* subsp. *enterica* serovar genomes (NC_010067, NC_011094, NC_021812, NC_021902, NC_022221) and two *S. bongori* genomes (NC_015761, NC_021870). We simulated read pairs as described above and added them to 19 million read pairs from 50 different HMP human gut microbiome samples (Supplementary Data 1). For both experiments, each sample was de novo assembled individually using SPAdes (v.3.9.0)[49] with the --meta flag, and scaffolds from each sample >2,000 bp were added to a combined scaffold set for each dataset. Reads were then mapped to the combined scaffold set using Minimap2 (v.2.15r905)[50], sorted using samtools (v.1.7)[46] and abundances calculated using jgi_summarize_bam_contig_depths from MetaBAT2 (v.2.10.2)[12]. These were combined into one file and used as input to VAMB, with default parameters and bin splitting enabled. MetaBAT2 (v.2.10.2)[12] was run in single-sample mode using default parameters on all samples. To assign scaffolds to reference genomes we used blat (v.385)[51] with default parameters and the *Salmonella* genomes as database and accepted all hits of length ≥500 nt, and with ≥99.5% identity for the single-strain spike-in experiment and 99.9% identity for the mixed-strain spike-in. When assessing whether *Salmonella* genomes could be reconstructed, we used $F1 > 0.9$ and 0.6 for the single- and mixed-strain spike-in experiments, respectively. We used the lower threshold for the mixed-sample spike-in to account for de novo assembly, creating chimeric sequences when strains were very similar. To assign true positives we used the length of the blat hit and, to determine false positives, we used the entire length of the contig (Benchmarking). The plot of *Salmonella* alignments to the reference genome was done using Circos (v.0.69.9)[52]. For genome comparisons of the *Salmonella* and CAMI2 datasets we used FastANI (v.1.1)[53] with default parameters to calculate ANI. Within each sample we then determined the distance to the most similar genome. For abundance of each genome we used that supplied with the CAMI2 datasets. The two *Salmonella* datasets (30 and 50 gut microbiome samples) were run in 4 and 8 h, respectively, using 24 CPU cores.

Calculation of dataset entropy and genome difficulty. We determined entropy of the datasets by calculating the Shannon diversity (SD) of individual samples in each of the CAMI2 datasets. Here SD was calculated based on the number of contigs per strain in a sample. We defined the entropy of a dataset as the mean SD across all its samples. MetaHIT was excluded because no per-sample annotation was available (coassembly data). Genome difficulty was determined as the minimum number of contigs needed to reconstruct the genome at 90% recall.

Binning a large dataset of the human gut microbiome. We obtained de novo assemblies of 1,000 human gut microbiome samples from Almeida et al.[18]. These samples had been randomly selected across datasets in the European Nucleotide Archive (ENA), and we obtained the exact assemblies that were used in that particular work. The assemblies had been created using SPAdes (v.3.10.0)[49] with the flag --meta. Similar to their approach, we used only contigs >2,000 bp and accepted only bins ≥200 kb. We downloaded the reads from each sample from ENA and verified that we had precisely the same number of reads as reported for each sample in Almeida et al. Hereafter we used Minimap2 (v.2.15r905)[50] to map reads from each sample to the pooled set of contigs from all samples and sorted the alignments using samtools (v.1.7)[46]. We then calculated the abundances of each sample with jgi_summarize_bam_contig_depths from MetaBAT2 (v.2.10.2)[12] and combined the abundance information into one file. This abundance information was used as input to VAMB, together with the combined fasta of the contigs and run with default settings. Training and clustering were done on a NVIDIA Tesla V100 GPU. When running with a smaller network than default, we used 24 latent neurons and 384 hidden neurons. We used CheckM (v.1.0.18)[54] to estimate the completeness and contamination of each bin and compared these to the results of Almeida et al. (Supplementary Data 10). Because Almeida et al. did not include archaeal bins in their data, we downloaded the samples where VAMB had produced an archaeal genome of any quality ($n = 27$) and ran MetaBAT2 using the same parameters as in Almeida et al. This yielded 11 NC and nine medium-quality (MQ) archaeal bins, which we added to the MetaBAT2 set; VAMB generated 15 NC and six MQ archaeal bins from the same samples. For comparison we used the definition of NC bins from their work as >0.9 completeness and <0.05 contamination, MQ as >0.5 completeness and <0.1 contamination and we defined low-quality bins as those not passing NC or MQ criteria. $F1$ was calculated as $2\times(\text{precision}\times\text{recall})/(\text{precision}+\text{recall})$, where recall was set as CheckM completeness and precision as 1 – CheckM contamination. Annotation of cluster 546 was done using ncbi-blastn (v.2.8.1)[55] against the nonredundant nucleotide database (nt), and filtered for 90% identity and 500-nt alignment length. Abundance of each cluster was determined from jgi_summarize_bam_contig_depths, calculated above from the alignments. First the weighted average was determined for each bin in a cluster weighting the read abundance with contig length. Hereafter, the abundance of the clusters was determined as the sum of each bin in the particular cluster. The abundance matrix, CheckM results and bins in fasta format for all clusters are available for download (Data Availability). To identify which bins overlapped each other from VAMB and MetaBAT2 runs, we used MASH (v.2.0)[56] with 10,000 sketches per bin to compare the two sets. We then assigned corresponding bins between VAMB and MetaBAT2 from MASH distance ≤0.01 and confirmed that the bins were from the same sample. Using this approach, we could match 6,017 bins between the two datasets. NCBI-assembled genome lengths were obtained from ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/ASSEMBLY_REPORTS/ANI_report_bacteria.txt. For the analysis of functional differences between VAMB and MetaBAT2 bin pairs, we predicted protein sequences using Prodigal (v.2.6.3)[57] and annotated them using InterProScan (v.5.36–75.0)[58]. We counted gene ontologies in the common and unique sets of each bin and used a two-sided Fisher's test with Benjamini–Hochberg correction[59] to determine VAMB unique contig-enriched GOs (adjusted $P < 0.005$). We scanned for phages using CheckV (v.0.4.0)[60] and accepted a hit if CheckV completeness was >40%. Additionally, we used DeepVirFinder (v.1.0)[61] and accepted hits with $P < 0.01$. Annotation of bins using GTDB (release 89)[62] was done using GTDB-TK (v.1.1.0)[63] and the function classify_wf. We estimated run time for MetaBAT2 on the Almeida dataset by running it in single-sample mode on 50 random samples and extrapolating to 1,000 samples.

Phylogeny. For the provisional taxonomic assignment of the Almeida et al. dataset clusters, contigs were aligned with ncbi-blastn (v.2.8.1)[55] against nonredundant nt_v5, retaining for each contig the best hit with >90% nucleotide identity over 500 nucleotides. The cluster was assigned the species with most hits. For the marker gene tree, we concatenated the core gene amino acid alignments created by CheckM (v.1.0.18)[54] and ran IQ-TREE (v.1.6.8)[64] with the LG model and one partition per gene. For the individual trees of cluster 546 (*B. vulgatus*) and the 51 other clusters with at least 20 NC bins, we first used Prodigal (v.2.6.3)[57] to infer genes of the bins and then used SonicParanoid (v.1.3.0)[65] on protein sequences using the 'fast' mode to identify orthologous groups per cluster. Here we accepted all orthologous groups of proteins when they had the same number of proteins as the number of seed orthologous down to 90% of the number of NC bins in that cluster. In other words, for cluster 546 with 255 NC bins we accepted all orthologous groups of proteins when there were 255–230 proteins and the same number of seed orthologous in the group. We then extracted the DNA sequence for the genes and aligned each gene using MAFFT (v.7.453)[66] and the '—auto' option. We then reconstructed a tree for each gene using IQ-TREE (v.1.6.8)[64] using automated model selection for each gene[67]. For each cluster we then used all gene trees as input to ASTRAL-III (v.5.7.3)[68] to build a species tree, and calculated branch lengths on the tree using IQ-TREE (v.1.6.8)[64] where the ASTRAL tree was input as constrained topology. Here the gene sequences were concatenated into a supermatrix and we set a partition for each gene with automated model selection. Bootstrap support were calculated by UFBoot2 (ref. [69]) using ASTRAL. Trees were visualized using ITOL (v.5.2)[70]. Association between phylogenetic placement and metadata (location and study) was done using PERMANOVA implemented in the R package vegan (v.2.5–6)[71] using the function adonis2. If multiple NC strains originated from different sequencing runs of the same sample, one was randomly selected as the representative. Leaf distances were extracted from each phylogenetic tree using the R package ape (v.5.3)[72] with the function 'cophenetic.phylo', and the model used for adonis2 was '$d \approx$ Location + Study', where d is the phylogenetic distance, Location is Asia, North America and Europe and Study is given in Supplementary Table 1 from Almeida et al.

Statistics. For our analyses we used Wilcoxon signed-rank tests (paired data) and Wilcoxon rank-sum tests (unpaired data) implemented in R as the function wilcox.test to test for statistical significance between distributions. When investigating for GO enrichment we used a two-tailed Fisher's test implemented in R as the function fisher.test. Furthermore, we used PERMANOVA with 9,999 permutations implemented in the R package vegan as the function adonis2. Adjustment for multiple testing was done using Benjamini–Hochberg correction implemented in R and the function p.adjust. Sample sizes (n) and test statistics for all tests are given either in the text or the respective Supplementary figure, table or data.

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

Data availability

The sequence data used in this study are publicly available from either the respective studies or ENA. The semisynthetic MetaHIT dataset was downloaded from https://portal.nersc.gov/dna/RD/Metagenome_RD/MetaBAT/Files/ as the files depth.txt.gz and assembly-filtered.fa.gz. The simulated CAMI High and CAMI2 datasets were downloaded from https://data.cami-challenge.org/participate from 'Toy Test Dataset High_Complexity' and '2nd CAMI Toy Human Microbiome Project Dataset', respectively. The de novo assemblies of the Almeida dataset were obtained through personal communication with A. Almeida and R. D. Finn, and the reads downloaded from ENA as specified in their publication. The data and results of binning the MetaHIT, CAMI2 and Almeida datasets, as well as the source data for Figs. 1–3, are available on figshare at https://figshare.com/projects/VAMB/72677. A CodeOcean capsule of VAMB v.3.0.1, including the six training and test datasets for reproducing benchmarking results, is available from https://doi.org/10.24433/CO.2518623.v1. Source data are provided with this paper.

Code availability

All code can be found on GitHub at https://github.com/RasmussenLab/vamb and is freely available under the permissive MIT license. All analyses were performed using VAMB v.3.0.1. Additionally, code are available as a CodeOcean capsule at https://doi.org/10.24433/CO.2518623.v1.

References

39. Kislyuk, A., Bhatnagar, S., Dushoff, J. & Weitz, J. S. Unsupervised statistical clustering of environmental shotgun sequences. *BMC Bioinform.* **10**, 316 (2009).
40. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Preprint at https://arxiv.org/abs/1502.03167 (2015).
41. Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. Preprint at https://arxiv.org/pdf/1207.0580.pdf (2012).
42. Maas, A. L., Maas, A. L., Hannun, A. Y. & Ng, A. Y. Rectifier nonlinearities improve neural network acoustic models. Preprint at https://arxiv.org/pdf/1207.0580.pdf (2013).
43. Kingma, D. P. & Ba, J. L. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2017).
44. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**, 8026–8037 (2019).
45. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997?upload=1 (2013).
46. Li, H. et al. The sequence alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
47. Sculley, D. Web-Scale k-Means Clustering. in *Proc. 19th International Conference on World Wide Web* 1177–1178 (ACM Press, 2010).
48. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).
49. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. MetaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
50. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
51. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
52. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
53. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat. Commun. 9, 5114 (2018).
54. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
55. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421 (2009).
56. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
57. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 119 (2010).
58. Mitchell, A. L. et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **47**, D351–D360 (2019).
59. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B* **57**, 289–300 (1995).
60. Nayfach, S., Pedro Camargo, A., Eloe-Fadrosh, E. & Roux, S. CheckV: assessing the quality of metagenome-assembled viral genomes. Preprint at *bioRxiv* https://doi.org/10.1101/2020.05.06.081778 (2020).
61. Ren, J. et al. Identifying viruses from metagenomic data using deep learning. *Quant. Biol.* **8**, 64–77 (2020).
62. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
63. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).
64. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
65. Cosentino, S. & Iwasaki, W. SonicParanoid: fast, accurate and easy orthology inference. *Bioinformatics* **35**, 149–151 (2018).
66. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
67. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
68. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinform.* **19**, 153 (2018).
69. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation.*Mol. Biol. Evol.* **35**, 518–522 (2018).
70. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
71. Oksanen, J. et al. Package 'vegan'. Community Ecology Package v.2.5-6. R Package version 3.4.0 1–296. https://cran.r-project.org/src/contrib/Archive/vegan/vegan_2.5-6.tar.gz (2019).
72. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2018).

Acknowledgements

We thank A. Almeida and R. D. Finn for sharing de novo assemblies of the 1,000 gut microbiome samples that we used for benchmarking VAMB. We thank C. Titus Brown for his source code contribution to the VAMB software package. J.N.N., J.J., R.L.A., L.J.J. and S.R. were supported by the Novo Nordisk Foundation (grant NNF14CC0001). S.R. was supported by the Jorck Foundation Research Award.

Author contributions

S.R. conceived the study and guided the analysis. J.N.N., S.R., J.J. and R.L.A. performed the analyses. J.N.N. wrote the software. C.K.S., J.J.A.A., C.H.G., T.N.P., L.J.J., H.B.N. and O.W. provided guidance and input for the analysis. J.N.N., L.J.J. and S.R. wrote the manuscript with contributions from all coauthors. All authors read and approved the final version of the manuscript.

Competing interests

H.B.N. is employed at Clinical-Microbiomics A/S. The remaining authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-020-00777-4.

Correspondence and requests for materials should be addressed to S.R.

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

nature research

Corresponding author(s): Simon Rasmussen

Last updated by author(s): 13/11/20

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 Authors & Referees 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	The simulated read data was generated using ART v2.5.8. The reads were simulated as 100 nt error-free pairs with 300 nt insert size and a standard deviation of 10 nt
Data analysis	Python v.3.7 PyTorch v.1.2.0 CUDA v.10.1.243 VAMB v.3.0.1 MetaBAT2 v.2.10.2 MaxBin v.2.2.4 Canopy v.2014 DAS Tool v.1.1.1 sklearn v. 0.21.3 SPAdes v.3.9.0 SPAdes v.3.10.0 BWA v.0.7.15 Minimap2 v.2.15r905 Samtools v.1.7 blat v.385 MASH v.2.0 CheckM v.1.0.18 NCBI-blastn v.2.8.1 Circos v.0.69.9 FastANI v.1.1 GTDB-TK v.1.1.0 Prodigal v.2.6.3

SonicParanoid v.1.3.0
MAFFT v.7.453
ASTRAL-III v.5.7.3
IQ-TREE v.1.6.8
ITOL v.5.2
InterProScan v.5.36-75.0
CheckV v.0.4.0
DeepVirFinder v.1.0
R-package ape v.5.3
R-package vegan v.2.5-6
R-package tidyverse v.1.2.1
R-package ggplot2 v.3.1.0

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/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

The sequence data used in this study is publicly available, from either the respective studies or from the European Nucleotide Archive (ENA). The semi-synthetic MetaHIT dataset was downloaded from https://portal.nersc.gov/dna/RD/Metagenome_RD/MetaBAT/Files/ as the files depth.txt.gz and assembly-filtered.fa.gz. The simulated CAMI High and CAMI2 datasets were downloaded from https://data.cami-challenge.org/participate from the "Toy Test Dataset High_Complexity" and "2nd CAMI Toy Human Microbiome Project Dataset", respectively. The de novo assemblies of the Almeida dataset were obtained through personal communication with Alexandre Almeida and Robert D. Finn and the reads downloaded from the ENA as specified in their publication. The data and results of binning the MetaHIT, CAMI2 and Almeida datasets as well as Figure 1-3 source data are available on figshare at the follow link: https://figshare.com/projects/VAMB/72677. A CodeOcean capsule of VAMB v.3.0.1 including the six training and test datasets for reproducing benchmarking results are available from http://doi.org/10.24433/CO.2518623.v1.

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

Life sciences study design

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

Sample size	No calculation of sample sizes were made. We used the available datasets that have sample sizes spanning from 9 (CAMI2 Urogenital) to 1000 (Almeida).
Data exclusions	We originally also tested VAMB on the CAMI High dataset (n=5), but due to an unrealistic contig size distribution (Supplementary Figure 28) influencing both abundance and TNF estimates, we discarded the dataset.
Replication	No experimental replication was performed. The method was developed based on four training datasets (MetaHit, CAMI2 Airways, Oral and Urogenital) and tested on two other datasets (CAMI2 Skin and Gastrointestinal). Further it was validated using two spike-in datasets from Cleary et al., and finally an external dataset (Almeida et al.)
Randomization	No randomization was performed
Blinding	Investigators were not blind to the datasets, but the development of the method was done using the training datasets that were kept separate from the test datasets

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
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

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