nature methods

Article

https://doi.org/10.1038/s41592-023-01940-w

CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning

Received: 6 July 2022

Accepted: 14 June 2023

Published online: 27 July 2023

Check for updates

Alex Chklovski[1], Donovan H. Parks[2], Ben J. Woodcroft[1] & Gene W. Tyson[1]

Advances in sequencing technologies and bioinformatics tools have dramatically increased the recovery rate of microbial genomes from metagenomic data. Assessing the quality of metagenome-assembled genomes (MAGs) is a critical step before downstream analysis. Here, we present CheckM2, an improved method of predicting genome quality of MAGs using machine learning. Using synthetic and experimental data, we demonstrate that CheckM2 outperforms existing tools in both accuracy and computational speed. In addition, CheckM2's database can be rapidly updated with new high-quality reference genomes, including taxa represented only by a single genome. We also show that CheckM2 accurately predicts genome quality for MAGs from novel lineages, even for those with reduced genome size (for example, Patescibacteria and the DPANN superphylum). CheckM2 provides accurate genome quality predictions across bacterial and archaeal lineages, giving increased confidence when inferring biological conclusions from MAGs.

Large-scale sequencing and assembly of genomes directly from environmental samples has led to the recovery of hundreds of thousands of highly diverse metagenome-assembled genomes (MAGs) from metagenomic data[1–3], making it impractical to manually assess the quality of these genomes. The original approach to this problem used by CheckM[4] (hereafter CheckM1)[4], and other similar tools (such as BUSCO[5]), is to identify single-copy, near-universal marker genes associated with specific lineages to predict genome completeness and contamination. However, this approach has a number of limitations.

The single-copy marker gene approach used by CheckM1 relies on comparative genomics to identify lineage-specific marker gene sets to predict the completeness and contamination of a recovered MAG based on their presence, absence and copy number. Well-studied lineages with many high-quality genomes usually have more robust marker sets, which allows for higher accuracy and confidence in genome quality predictions. For novel lineages that lack high-quality genomic representation, only the most general marker sets (for example, domain-level) can be used for genome quality estimates, resulting in reduced accuracy and sensitivity. In addition, this approach typically performs poorly on MAGs from microorganisms with reduced genomes, which lack some 'universal' marker genes[6], and in many instances do not have many high-quality genomic representatives to derive robust marker sets.

An alternative approach to this problem is to use more complex mathematical techniques such as machine learning (ML) to link a wider range of genomic inputs to predict genome quality. ML algorithms can generate insights into complex data and have been used for important biological challenges such as protein folding[5] and metagenomic binning[7]. The application of ML to estimating genome quality has several advantages as it allows the incorporation of additional genomic information such as multi-copy genes, biological pathways and modules, and other genomic features such as amino acid counts and number

[1]Centre for Microbiome Research, School of Biomedical Sciences, Queensland University of Technology, Translational Research Institute, Woolloongabba, Queensland, Australia. [2]Donovan Parks, Bioinformatic Consultant, Castlegar, British Columbia, Canada. e-mail: gene.tyson@qut.edu.au

of coding sequences. Furthermore, it allows for automatic selection of relevant genomic features to use for genome quality predictions without relying on predefined lineage-specific marker sets.

Here we introduce CheckM2, a ML-based tool for predicting isolate, single-cell and MAG genome quality. CheckM2 builds models suitable for predicting bacterial and archaeal genome completeness and contamination without explicitly considering taxonomic information. CheckM2 was trained on simulated genomes with known levels of completeness and contamination, benchmarked against CheckM1 as well as BUSCO, and subsequently applied to MAGs from a range of environments. Overall, CheckM2 outperformed CheckM1 and BUSCO, and performed substantially better on MAGs from unusual lineages characterized by small genome size, such as the Candidate Phyla Radiation (Patescibacteria) and DPANN (an acronym of the names of the first included phyla: Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanohaloarchaeota) superphylum, as well as other lineages with sparse or no genomic representation.

Results

CheckM2 genome simulation, training and benchmarking

To demonstrate that ML can be applied to accurately predict genome quality, synthetic MAGs with known quality were constructed for ML training. A 'random-protein-sampling' method was used to build training MAG sets, where predicted proteins from a subset of 4,978 bacterial and 322 archaeal complete isolate genomes selected from National Center for Biotechnology Information RefSeq[8] Release 89 were randomly sampled to build roughly 700,000 synthetic MAGs at predetermined completeness and contamination percentages (Methods and Fig. 1a). The target completeness for ML training and output was defined as the percentage of MAG length relative to total MAG length, while contamination was defined as the length of the contaminating portion relative to the expected (complete, uncontaminated) genome length. To validate the performance of ML models, two separate MAG simulation approaches were used: (1) a '20 kb-nucleotide-fragmentation' method where the full-length genomes were sheared into roughly 20 kb-long pieces, and (2) a 'MAG-derived-fragmentation' model where full-length genomes were sheared into contig distributions representative of MAGs in the Genome Taxonomy Database (GTDB)[9] (Fig. 1b). In both simulation models used for validation, contigs were randomly sampled to build MAGs with a range of simulated completeness (5–100%) and contamination (0–100%) values.

To train and test different ML models for predicting genome quality, the genome properties of synthetic MAGs were calculated as feature vectors for the ML models, including the genome length, number of coding sequences and individual amino acid counts, as well as annotation of predicted proteins using KEGG (the Kyoto Encyclopedia of Genes and Genomes)[10]. In total, 11 ML methods (Methods) were trained on randomly selected subsets of the simulated MAGs (75% of all MAGs; 'random-protein-sampling') and subsequently validated on the remainder of the MAGs (25%; for both '20-kb-nucleotide-fragmentation' and 'MAG-derived-fragmentation') for an initial assessment of quality prediction performance across diverse bacterial and archaeal phyla. To assess performance of ML models, predictions on simulated genomes were divided into four groups based on MIMAG (minimum information about a metagenome-assembled genome) completeness and contamination standards[11] (high quality, more than 90% complete and less than 5% contaminated; medium quality, 50–90% complete and less than 10% contaminated and low quality, less than 50% complete and less than 10% contaminated), as well as a separate group for high contamination (more than 10% contaminated) (Supplementary Table 1).

Artificial neural networks[12] (NNs) and gradient boosted (GB) decision trees[13] showed the best overall performance (Supplementary Table 2) and were used in further optimization and testing for CheckM2 (Fig. 1c). Both the NN and GB models exhibited higher accuracy when KEGG annotations were considered in the context of their pathways and modules (Methods). In addition, the NN included convolutional layers for feature extraction, leading to an improvement in accuracy (Supplementary Table 3 and Supplementary Note 1). These optimizations to both models were used in subsequent testing.

Using simulated genomes to assess ML performance

To assess the effect of taxonomic novelty on the accuracy of the optimized NN and GB models, an iterative leave-one-out approach was used on the synthetic genome set, where genomes from specific taxa were removed from the training set from phyla to species, models were trained on the remaining genomes, and prediction accuracy tested on the left-out group. The mean average error (MAE) of both models for predicting completeness and contamination was systematically assessed from phylum to species level (Fig. 1c). Separate models were trained for predicting completeness and contamination for all ML models. As expected, removing lineages from the training set with increasing taxonomic level proportionally affects the genome quality estimates (that is, removing all genus-level representation has a substantially lower impact on accurately predicting genome quality than removing class level or phylum-level representation of query genomes). Overall bacterial and archaeal genome quality estimates improve if the training set contains a genome that is more taxonomically related to the query genome (Fig. 2a). However, the two models have different strengths relative to genome novelty and genome completeness. Completeness quality estimates for query genomes representing novel phyla, classes and orders were more accurate with the GB model, while the NN model was on average more accurate for genomes representing novel families, genera and species (Fig. 2a). Additionally, for low-quality (less than 50% complete) genomes the NN model was more accurate at all taxonomic levels, while the GB model accuracy declined with lower MAG quality (Fig. 2a).

The most difficult completeness prediction scenario is likely to be genomes belonging to a new phylum (that is, a phylum without a complete isolate genome). For near-complete genomes from a novel phylum, the MAE for completeness predictions using the GB model is 3.1 ± 3.9% and 5.2 ± 5.7% for the NN model. For medium-quality genomes, the GB model had a MAE of 4.6 ± 4.4%, while the NN model had a MAE of 5.9 ± 5.3%. These results indicate the models have an ability to generalize to phylum-level novelty with relatively good accuracy even as genome quality declines. While it is impossible to reproduce this test for CheckM1, using CheckM1's domain-level bacterial or archaeal marker sets consisting of roughly 120 universal marker genes resulted in roughly equivalent MAEs of 3.4 ± 4.4% for high-quality and 7.2 ± 5.8% for medium-quality genomes.

For genome contamination predictions at all taxonomic levels, the gradient boost model substantially outperformed the NN and was chosen as the model for predicting contamination (Figs. 1d and 2b). For genomes belonging to a novel phylum, the predicted contamination MAE of the GB model is 2.0 ± 2.2% (high quality), contrasting with a NN MAE of 7.3 ± 5.5% (high quality) and a CheckM1 domain-level marker set comparison of 1.9 ± 2.2% (high quality).

Because the NN model performed best for less novel genomes and the GB model performed best for more novel genomes, both models were implemented in the final version of CheckM2 for completeness prediction. Only the GB model was implemented to predict contamination. For completeness predictions on novel and more complete genomes, CheckM2 uses a 'general' model based on gradient boost decision tree algorithms, while for genomes more closely related to those in its reference set or less complete genomes it uses a 'specific' model based on artificial NNs (Fig. 1d). A cosine similarity measure was found to correlate well with input genome taxonomic novelty, with a linear relationship between squared cosine similarity and taxonomic distance (Supplementary Fig. 1), enabling CheckM2 to use this measure to select between the 'general' and 'specific' model for each input

https://doi.org/10.1038/s41592-023-01940-w

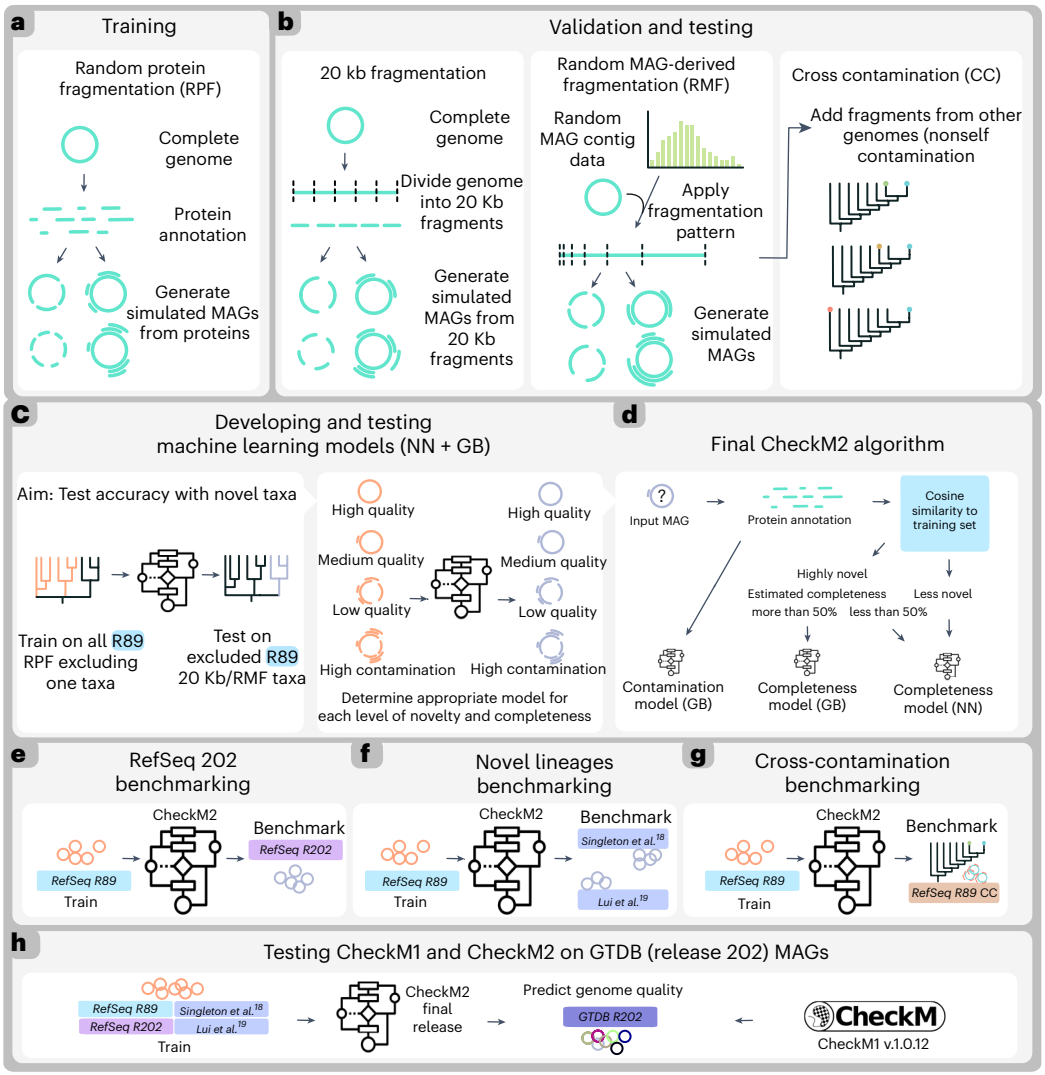

Fig. 1 | Overview of CheckM2 development, benchmarking and validation. **a**,**b**, Simulation of synthetic genomes for training using random protein fragmentation (RPF) (**a**) and for testing using the 20 kb fragmentation (20 kb) and random MAG-derived fragmentation (RMF) (**b**). **c**, Selection of NN and gradient boost models and further testing and refinement. **d**, The final algorithm used by CheckM2 to decide between gradient boost and NN models. **e**, Benchmarking of CheckM2 on RefSeq 202 synthetic genomes. **f**, Benchmarking of CheckM2 on novel and unusual synthetic genomes derived from circular MAGs including Patescibacteria. **g**, Benchmarking CheckM2 on synthetic genomes with nonself-contamination derived from RefSeq r89 genomes. **h**, Comparing CheckM1 and CheckM2 genome quality predictions for all GTDB r202 MAGs.

genome based on predefined cosine similarity cutoffs derived from the leave-one-out approach (Methods, Fig. 1d and Supplementary Table 4).

Benchmarking CheckM2 performance on synthetic RefSeq genomes

The initial CheckM2 ML models were built on genomes from RefSeq release 89, allowing new complete genomes from RefSeq release 202 to be used to test CheckM2's performance, as they were not part of the original training and validation sets (Fig. 1e). In total, this included 2,864 new complete microbial isolate genomes representing six novel phyla, 13 novel classes, 43 novel orders, 87 new families, 439 novel genera and 1,554 novel species according to their GTDB classifications. As these genomes represent the range and types of genomes added to public databases over the course of roughly 2 years, they provide a reasonable indication of how CheckM2 performs when tested against new genomes of varying taxonomic novelty. They also provide suitable complete genomes for simulating new genomes of known completeness and contamination (as in Fig. 1b), allowing benchmarking of CheckM2 against CheckM1 and BUSCO.

When predicting the completeness of 712,880 simulated RefSeq 202-based genomes, CheckM2 was substantially more accurate than CheckM1 with a lower MAE across all genomes (Fig. 3a and Supplementary Note 2). Overall, there was similar performance between CheckM2 and CheckM1 on high-quality genomes (CheckM2 MAE 2.1 ± 2.9%, CheckM1 MAE 2.0 ± 3.2%) with BUSCO being less accurate (BUSCO MAE 4.4 ± 6.8%). CheckM2 was far more accurate for medium, low-quality and highly contaminated genomes then both other tools (Fig. 3a; CheckM2 MAE 2.9 ± 2.9%, CheckM1 MAE 4.7 ± 5.4%, BUSCO MAE 6.4 ± 7.0%). However, as some phyla within RefSeq 202 are highly oversampled, bulk genome MAE underestimates performance across broad taxonomic ranks. When using a phylum-weighted MAE (PW-MAE), CheckM2 outperformed CheckM1 and BUSCO with both substantially higher accuracy and much lower error variance for high-quality genomes (CheckM2 PW-MAE 2.5 ± 2.2%, CheckM1 PW-MAE 5.7 ± 2.9%, BUSCO PW-MAE 10.2 ± 4.5%) as well as medium and low-quality genomes (CheckM2 PW-MAE 3.7 ± 3.2%, CheckM1 PW-MAE 7.1 ± 5.7%, BUSCO PW-MAE 10.2 ± 7.3%). CheckM2 exhibited

https://doi.org/10.1038/s41592-023-01940-w

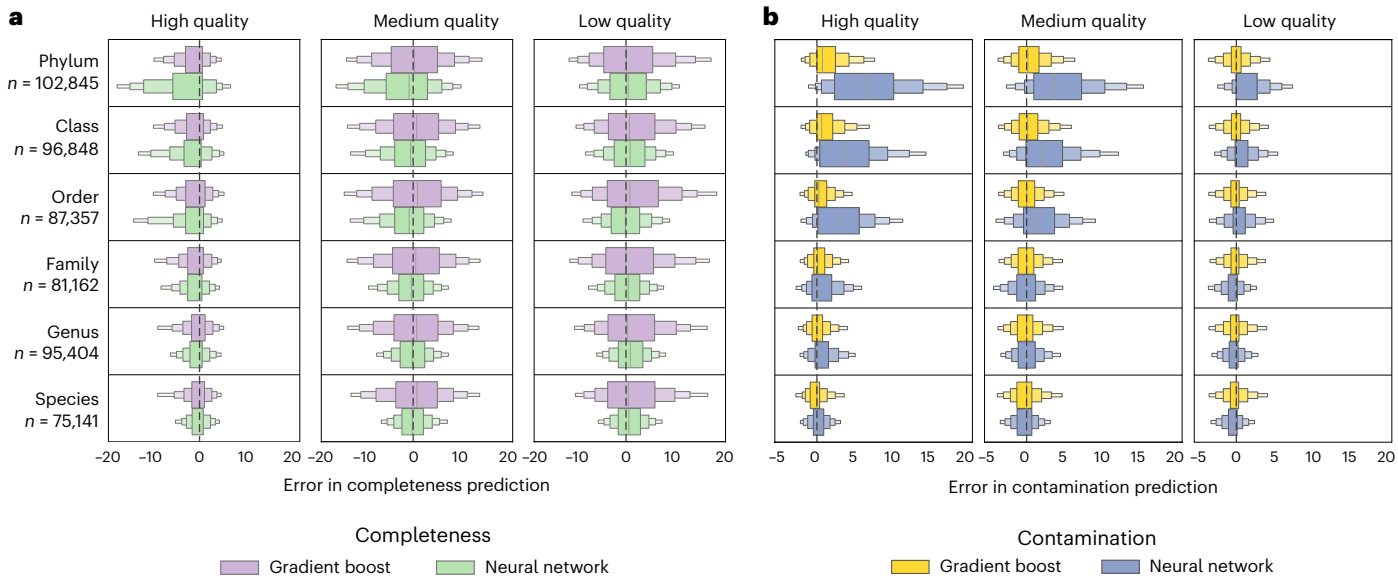

Fig. 2 | Benchmarking ML models on synthetic genomes of varying taxonomic novelty. a, Error in predicting completeness. **b**, Contamination at varying taxonomic levels of novelty. Each taxonomic novelty level is broken into separate error margins across different MIMAG quality cutoffs (high quality, 90–100% completeness and 0–5% contamination; medium quality, 50–90% completeness and 0–10% contamination and low quality, less than 50% completeness and 0–10% contamination). Positive values indicate overestimation, whereas negative values indicate underestimation of true values. The size of each error box in the letter-value plot shows half the remaining data, starting with 50% for the first box, 25% for the second box and so on.

comparable performance across both the '20-kb-fragmentation' and the 'MAG-derived random fragmentation' simulations, suggesting that there is little effect of simulation method on resulting predictions (Supplementary Note 2). The most significant increase in performance of CheckM2 was seen in predicting completeness of genomes from the phyla with very few high-quality genomic representatives such as Iainarchaeota, Nanohaloarchaeota, Dependentiae, Bipolaricaulota and Patescibacteria (high-quality MAE 3.6 ± 2.9%) when compared to CheckM1 (high-quality MAE 26.3 ± 10.8%) or BUSCO (high-quality MAE 34.1 ± 7.1%) (Fig. 3b,c). Notably, there was only a single reference genome for Nanohaloarchaeota, Dependentiae, Bipolaricaulota and Iainarchaeota in the training set for CheckM2, indicating that a single genomic representative of a lineage provides sufficient information for an accurate prediction of genome quality.

When predicting contamination, the MAE of CheckM2 (MAE 1.2 ± 1.3%) was comparable to CheckM1 (MAE 1.5 ± 1.8%) and BUSCO (MAE 1.0 ± 1.4%) on high-quality genomes, and was substantially more accurate for medium- and low-quality genomes (CheckM2 1.7 ± 1.7%, CheckM1 3.0 ± 4.0%, BUSCO 2.9 ± 4.1%). It was also substantially better at predicting contamination in highly contaminated genomes (Fig. 3d). To confirm prediction accuracy with metrics other than MAE, we calculated the R^2 between predicted and actual completeness and contamination metrics for all three tools across a wide range of genome quality values of the synthetic genomes. CheckM2 had a higher R^2 between predicted and actual values for every single group of genome quality cutoffs (Supplementary Table 11).

Benchmarking CheckM2 performance on new lineages

One of the weaknesses of CheckM1 was its poor performance on highly novel genomes relative to the dataset its marker sets were based on, particularly from lineages characterized by organisms with small or highly reduced genomes, such as organisms from the DPANN and Patescibacteria. The archaeal DPANN superphylum and the bacterial Patescibacteria are large radiations of microorganisms comprising a significant fraction of the tree of life[14]. Their high diversity, unusual biology, absence of key genes and often small genomes make predicting their genome quality particularly challenging[15–17]. To assess ability of CheckM2 to predict the quality of genomes from these microorganisms, 57 closed circular genomes were obtained from wastewater (Singleton et al.[18]), including 30 genomes from the Patescibacteria, as well as other highly novel and often small genomes from phyla such as Dependentiae, Iainarchaeota and UBA10199. Additionally, 36 additional circularized Patescibacteria genomes were obtained from Lui et al.[19]. These lineages are poorly represented in the RefSeq release 202, which mostly cover phyla and classes with existing isolate representatives. Together, this dataset represents 25 unique classes and 45 unique orders of curated and circular Patescibacteria genomes along with a number of other circularized MAGs representing novel phyla and classes, providing an excellent opportunity to test CheckM2's performance on novel genomes (Fig. 1f). From these complete genomes, simulated genomes of varying completeness and contamination were created as above (Fig. 1b) to enable tool benchmarking across different levels of genome quality.

Across all classes of Patescibacteria, CheckM2 was far more accurate than CheckM1, with the performance CheckM1 only being improved by using a custom Patescibacteria marker set based on 43 ribosomal genes[6] (Fig. 4c). However, the accuracy of CheckM1 using the custom candidate phyla radiation (CPR) marker set substantially declined on medium- and low-quality Patescibacteria genomes with completeness error rates as high as 30–40%, making the method unreliable. The superior performance of CheckM2 extends to all Patescibacteria classes represented in the Singleton et al.[18] and Lui et al.[19], indicating that CheckM2 can robustly and accurately predict genome quality from highly diverse lineages such as the Patescibacteria, despite only having a few genomic representatives in the training set.

Across other unusual lineages, CheckM2 is more accurate for both high-quality genomes (MAE 2.9 ± 2.6%) and medium- and low-quality genomes (MAE 4.4 ± 3.8%) than CheckM1 (high-quality MAE 4.7 ± 6.0%; medium- and low-quality MAE 6.1 ± 5.8%) or BUSCO (high-quality MAE 10.9 ± 9.0%; medium- and low-quality MAE 10.2 ± 8.3%), showing an ability to generalize well across a range of phyla and classes not represented in its reference set (Fig. 4a). For unusual lineages with reduced genomes such as the phyla Patescibacteria, Dependentiae or Iainarchaeota, where an in-built specific marker set is not

https://doi.org/10.1038/s41592-023-01940-w

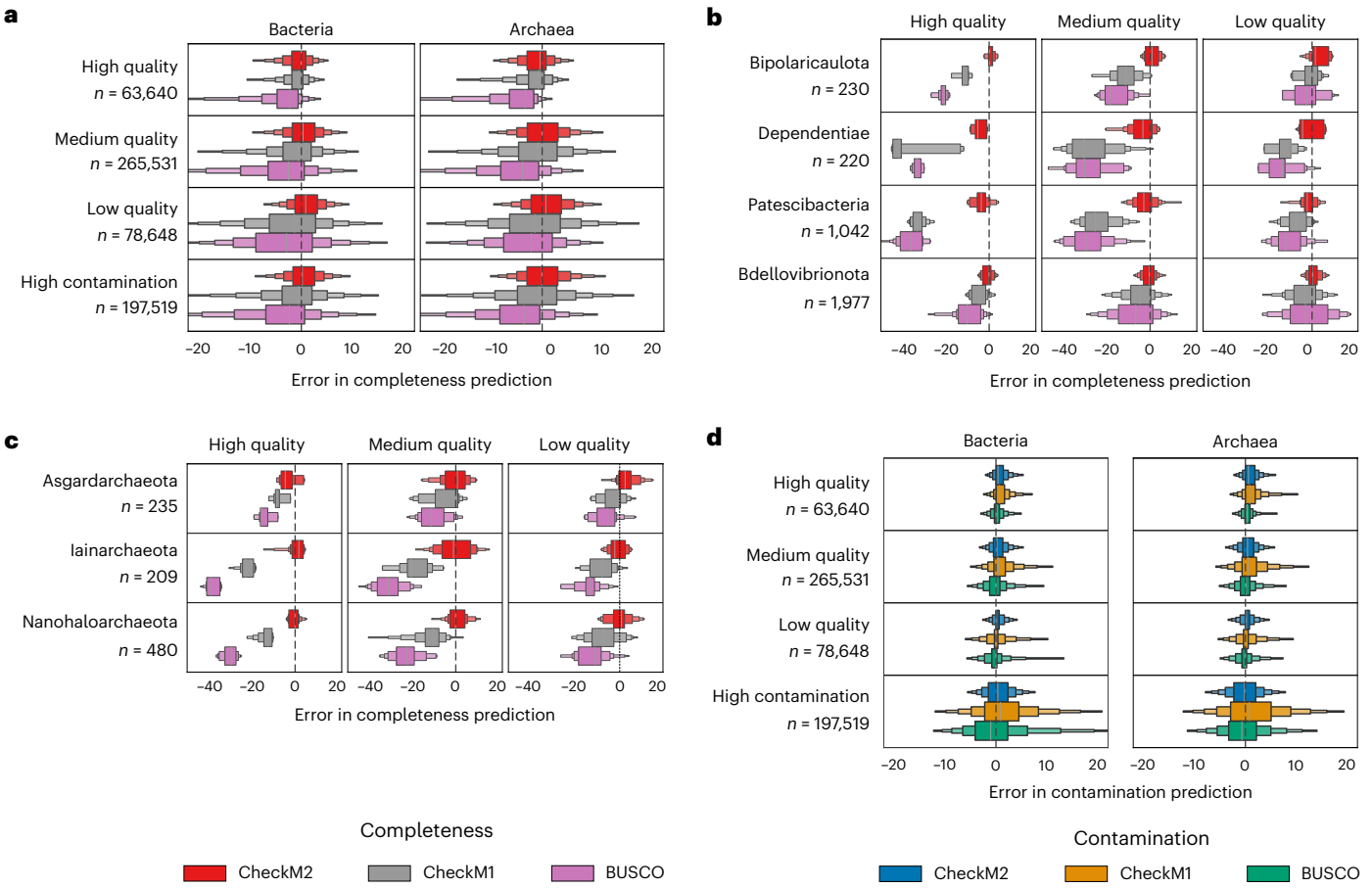

Fig. 3 | Comparison of tools on RefSeq r202 genomes. a, Error in predicting completeness on bacterial and archaeal genomes. **b**, Error in predicting completeness on specific bacterial phyla. **c**, Error in predicting completeness on specific archaeal phyla with substantial differences between CheckM2 and other tools. **d**, Error in predicting contamination on bacterial and archaeal genomes. Results are broken into separate error margins across different MIMAG quality cutoffs (high quality, 90–100% completeness and 0–5% contamination; medium quality, 50–90% completeness and 0–10% contamination; low quality, less than 50% completeness, 0–10% contamination and high contamination, more than 10% completeness). Positive values indicate overestimation while negative values indicate underestimation of true values. The size of each error box in the letter-value plot shows half the remaining data, starting with 50% for the first box, 25% for the second box and so on.

available for CheckM1, CheckM2 completeness predictions are far more accurate (MAE 5.8 ± 5.3%) than CheckM1 (MAE 19.8 ± 10.6%) or BUSCO (MAE 30.3 ± 13.2%; Fig. 4b). As with RefSeq 202 benchmarking, CheckM2 showed similar performance on genomes simulated by both test simulation methods and consistently outperformed CheckM1 (Supplementary Note 3).

CheckM2 also outperformed other tools on most cases of contamination (Fig. 4d). The only partial exception are some high-quality Patescibacteria genomes in which CheckM1's lineage-specific marker sets provide slightly better accuracy (Singleton et al.[18]: CheckM1 MAE 1.4 ± 1.1%, CheckM2 MAE 1.7 ± 1.3%; Supplementary Note 3). In part, this may be due to CheckM2's conservative nature when approaching contamination predictions. However, it is likely that the addition of these new circularized Patescibacteria genomes to CheckM2's final reference set (Fig. 1h) will increase its accuracy.

As with the RefSeq release 202 benchmarking, we calculated the R^2 between predicted and actual completeness and contamination values of simulated genomes and predictions by all tools. As above, CheckM2 outperformed both other tools across every criterion (Supplementary Table 11).

Benchmarking CheckM2 cross-contamination performance

Contamination in MAGs may come from the binning together of closely related strain or species, but may potentially also contain divergent sequences from other lineages or even domains. CheckM1 uses duplicated single-copy marker gene counts to infer contamination, on the assumption that contamination will come from closely related genomes being binned together, and thus will contain duplicated single-copy genes. This is likely to work better when the sources of contamination are highly related and thus likely to share the same distribution of single-copy markers.

However, it is unclear how accurate CheckM1 is when assessing contamination from a different source from the same strain or species, and whether CheckM2's weighted combination of feature vectors is better able to identify foreign contamination compared to only using duplicate single-copy marker genes. Here, the contamination predictions of CheckM1, CheckM2 and BUSCO were benchmarked on simulated genomes with contamination originating from increasingly divergent sources, from species to domain (Fig. 1g). In addition, GUNC[20], was also benchmarked for contamination prediction, as it uses an alternative approach based on the presence of taxonomically discordant contigs.

Our results demonstrate that CheckM2 is accurate at identifying foreign contamination, particularly for high-quality genomes (Fig. 5), although it was less accurate on higher-taxa contamination for medium-quality genomes, as were CheckM1 and especially BUSCO, which substantially underestimated contamination (Fig. 5). CheckM2 is far less likely to overestimate contamination compared to CheckM1,

https://doi.org/10.1038/s41592-023-01940-w

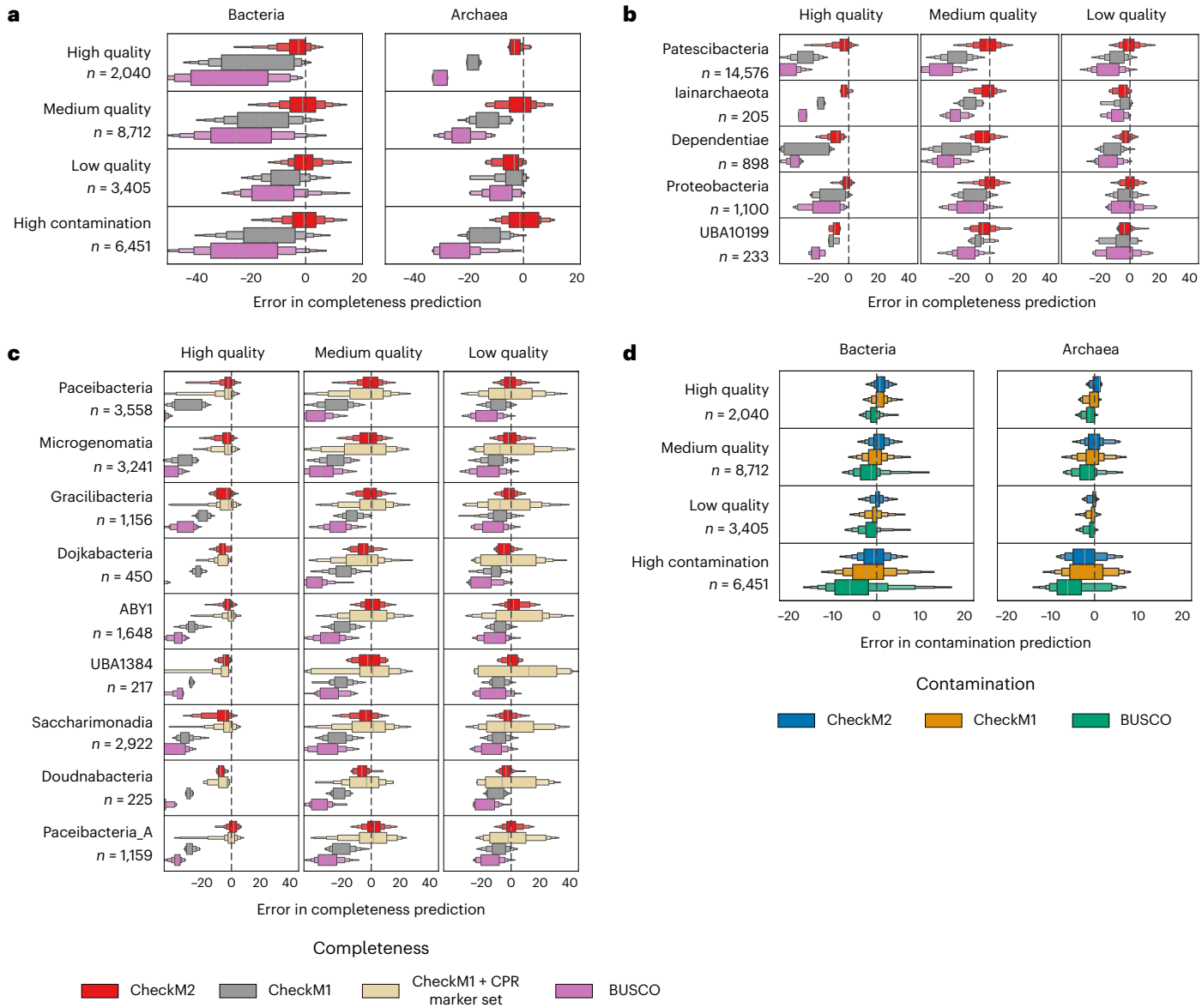

Fig. 4 | Comparison of tools on novel lineages. a, Error in predicting completeness on bacterial and archaeal genomes. **b**, Error in predicting completeness on specific bacterial and archaeal phyla with substantial differences between CheckM2 and other tools. **c**, Error in predicting completeness on all classes of Patescibacteria, including predictions by CheckM1 using the CPR marker set. **d**, Error in predicting contamination on bacterial and archaeal genomes. Results are broken into separate error margins across different MIMAG quality cutoffs (high quality 90–100% completeness and 0–5% contamination; medium quality 50–90% completeness and 0–10% contamination; low quality, less than 50% completeness, 0–10% contamination and high contamination, more than 10% completeness). Positive values indicate overestimation whereas negative values indicate underestimation of true values. The size of each error box in the letter-value plot shows half the remaining data, starting with 50% for the first box, 25% for the second box and so on.

likely due to the fact it does not rely as strongly on single-copy marker genes and does not use small marker sets (Supplementary Note 4).

CheckM2 outperformed CheckM1 and BUSCO at all levels in the mean of the absolute error (AE) predictions for high- and medium-quality genomes, with contamination from the same-species-derived contamination (CheckM2 high-quality AE 1.7 ± 1.6%, CheckM1 AE 2.6 ± 3.1%, BUSCO AE 2.4 ± 3.1%) to contamination derived from the same phylum (CheckM2 AE 2.4 ± 2.3%, CheckM1 AE 3.6 ± 4.7%, BUSCO 3.7 ± 3.2%), a different phylum (CheckM2 AE 2.6 ± 2.4%, CheckM1 AE 4.3 ± 6.2%, BUSCO 4.1 ± 3.1%) or different domain (for example archaeal contamination of bacterial MAGs) (CheckM2 AE 3.1 ± 2.5%, CheckM1 AE 3.2 ± 2.6%, BUSCO 4.3 ± 2.9%). GUNC was substantially more accurate when contamination was derived from a different class or more taxonomically distant contamination but tended to overpredict contamination across other levels, and substantially underestimated same-species and same-genus contamination (Fig. 5). CheckM2 was more accurate than other tools in predicting contamination from the same species, genus or family, which is considerably more difficult to detect with taxonomy-based detection tools such as GUNC[20].

Application of CheckM2 to environmental MAGs

Comparison of CheckM1 versus CheckM2 predictions across all taxa. Following benchmarking on synthetic genomes, CheckM2 was retrained with all complete genomes in RefSeq release 202 to provide a comprehensive reference database for inclusion with the CheckM2 release version. We then used CheckM2 to predict the genome quality across all bacterial and archaeal lineages. As GUNC is unable to

https://doi.org/10.1038/s41592-023-01940-w

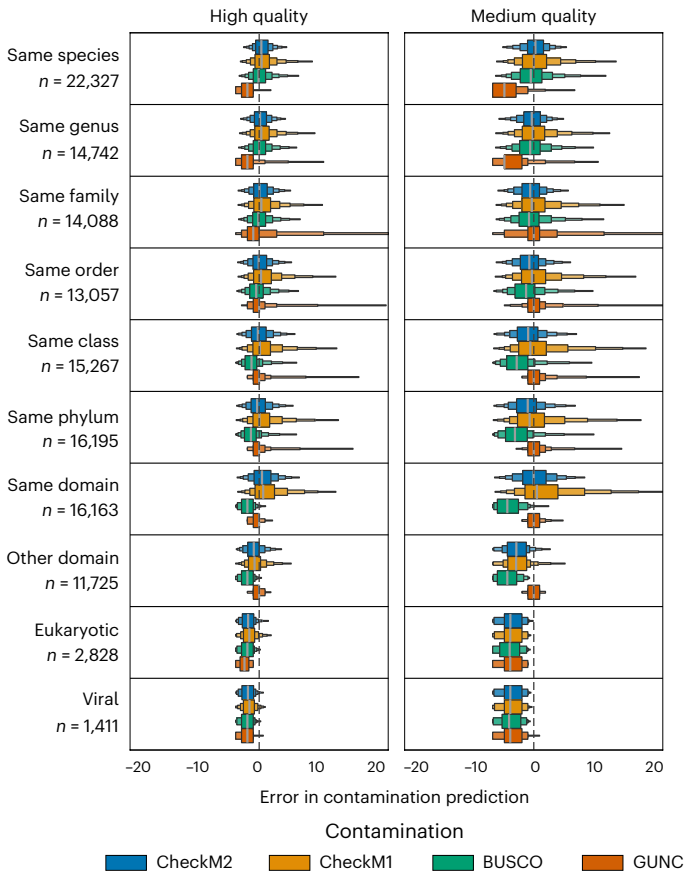

Fig. 5 | Comparison of tools on non-self contamination. Error in contamination prediction is broken by the taxonomic source of the contaminant relative to contaminated genome. Results are broken into separate error margins across different MIMAG quality cutoffs (high quality 90–100% completeness and 0–5% contamination; medium quality, 50–90% completeness and 0–10% contamination). Positive values indicate overestimation while negative values indicate underestimation of true values. Each box shows half the remaining data, starting with 50% for the first and 25% for the second and so on.

predict completeness, and BUSCO was consistently outperformed by both CheckM1 and CheckM2 in benchmarking tests above, CheckM2 predictions were compared to CheckM1 predictions for completeness and contamination of 224,101 bacterial and 3,881 archaeal genomes in the GTDB release 202 (most recent GTDB release available at time of testing) annotated as 'incomplete': that is, not isolate genomes or closed circular MAGs (Fig. 1h).

Overall, there is good congruence among completeness predictions across most phyla between CheckM2 and CheckM1, with 73% of all completeness predictions being within 1% of each other and 91% being within 5% of each other (Fig. 6a). Similar congruency in results was observed for contamination, with 82% of all genome predictions being within 1% of each other and 99% being within 5% (Fig. 6b). Substantially higher or lower completeness predictions (more than 5% difference) using CheckM2 often occurred across entire lineages (phylum through to genera), while discrepancies in contamination were typically restricted to specific genomes within lineages (that is, were not systematic; Fig. 6, Supplementary Table 5 and Supplementary Figs. 2 and 3). CheckM2 was also able to identify previously undetected contamination in a number of MAGs and isolate genomes and may avoid some potential contamination overestimations by CheckM1 (Supplementary Note 8).

In the Bacteria, the highest divergence in completeness predictions is within the Patescibacteria phylum, where CheckM2 scores are substantially higher than those predicted by CheckM1 (Fig. 6c). Based on benchmarking, the CheckM2 results are likely to be substantially more accurate, enabling much better Patescibacteria MAG curation in the future and giving greater confidence to biological insights derived from these genomes. Other bacterial lineages predicted to be substantially more complete all appear to have common features such as small or reduced genome size, and/or hypothesized endosymbiotic or parasitic lifestyle. This includes the phyla Dependentiae, which are phylogenetically related to the Patescibacteria[21], as well as the orders RF32 within the Proteobacteria, TANB77 within the Firmicutes_A[22], and the actinobacterial orders Actinomarinales and Nanopelagicales[23]. While some families within the Firmicutes_A order Christensenellales have concordant CheckM1 and CheckM2 predictions, other families such as CAG-74 have much higher CheckM2 completeness values. CAG-74 has been hypothesized to lack certain key functions (for example, amino acid biosynthesis pathways) and may be potential symbionts[22]. Members of the family UBA1242, where the average genome size is roughly 1 mega-basepairs also shows substantially higher CheckM2 completeness predictions (on average 11% more complete), indicating that this family may also have a symbiotic or parasitic lifestyle that has not previously been reported (Fig. 6c).

Analysis of manually curated complete bacterial endosymbiont genomes (Supplementary Table 6) demonstrated that CheckM2 markedly outperformed CheckM1 by predicting a much higher completeness, with CheckM2 predicting an average completeness of 71%, compared to CheckM1's 39% average. Notably, CheckM2 was able to achieve this accuracy with little to no endosymbiont representation in its training database (as they are usually excluded from RefSeq) and incorporating the test genomes into the final models will likely substantially improve its accuracy on future endosymbiont cases. It is likely that use of CheckM2 on assembled metagenomic data will lead to the discovery of novel endosymbiont genomes that are highly complete with a small genome size.

Archaeal lineages with substantially higher CheckM2 completeness scores are primarily in the DPANN superphylum, including members of the Nanoarchaeota, Nanohaloarchaeota and Micrarchaeota phyla, which have high-quality genomic representatives in CheckM2, as well as the phyla Huberarchaeota, Aenimatarchaeota and PWEA01 (formerly part of Aenigmatarchaeota), which are not represented in the CheckM2 release reference set (Fig. 6c). These predictions underscore the effectiveness of CheckM2's prediction approach, which generalizes to novel taxa with biological similarities to genomes CheckM2 was trained on. Other lineages that are predicted to be more complete by CheckM2 include the class Poseidoniia_A within the Thermoplasmatota, which is missing several single-copy genes used by CheckM1 (ref. 24) and the Asgardarchaeota order CR-4. The recently isolated and sequenced *Prometheoarchaeum syntrophicum*, which belongs to the order CR-4, was included in the release reference set for CheckM2, which likely contributed to a higher completeness score for this lineage. This highlights the power of including a single genome representative in CheckM2's ML predictions.

In a small number of instances, CheckM2 completeness values were substantially lower than CheckM1 (5.4% of all genomes were more than or equal to 5% lower, 1.6% of genomes were more than or equal to 10% lower or more). The underlying cause for this difference is unclear but is likely due to multiple factors, such as the novelty of genomes, CheckM2's choice of ML model or CheckM1's use of a kingdom-level marker set to assess the completeness of some unusual lineages (Supplementary Note 5). Indel-dominated genomes were also found to have a particularly low CheckM2 score relative to CheckM1 (Supplementary Note 6). Furthermore, as CheckM1 is often used to select MAGs for submission and publication, this can produce an imbalanced selection effect where genomes with overprediction error are retained in databases at higher rates than genomes with underprediction errors

https://doi.org/10.1038/s41592-023-01940-w

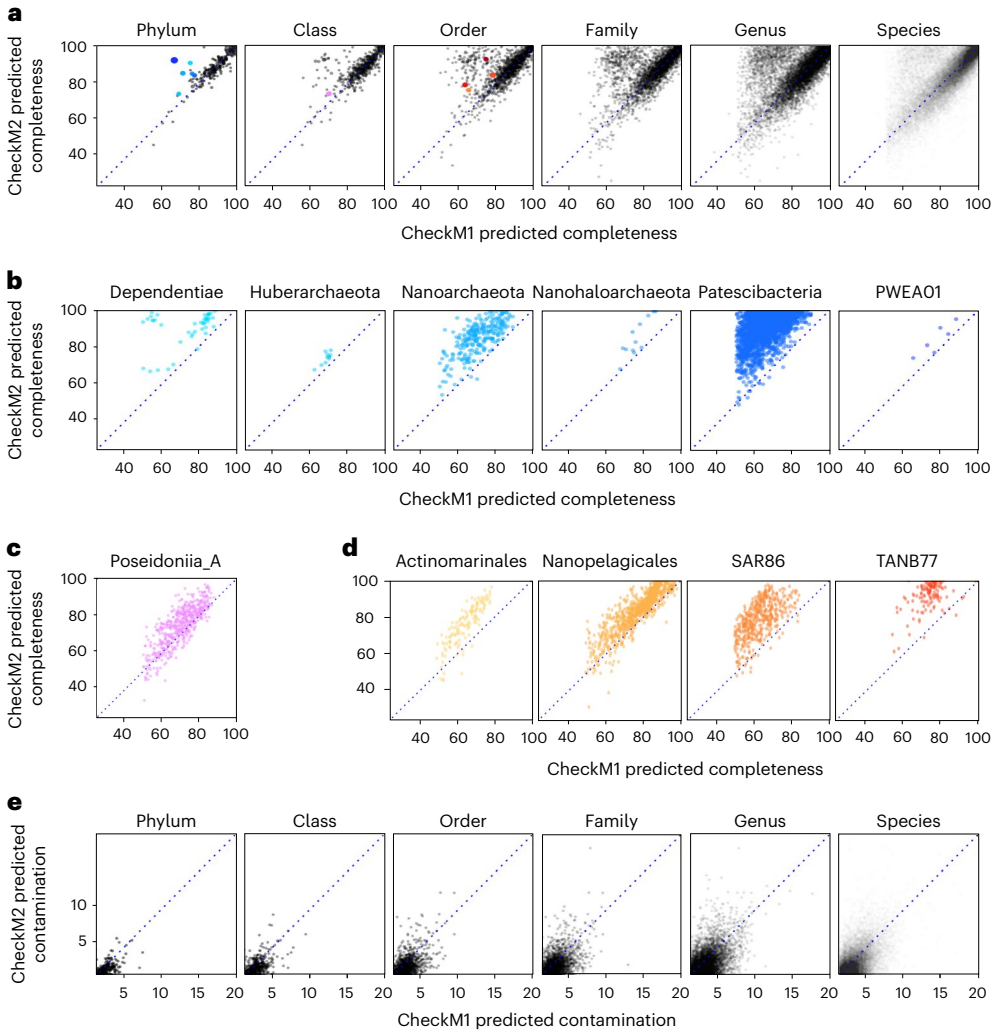

Fig. 6 | CheckM1 versus CheckM2 predictions across GTDB. a, Mean completeness prediction from phylum to species. **b**, Completeness predictions for genomes in the specific phyla. **c**, Completeness predictions for each genome in the class shown. **d**, Completeness predictions for each genome in the orders shown. **e**, Mean contamination prediction from phylum to species. For both, the size of the circle corresponds to the number of genomes in each taxa.

(Supplementary Note 7). Given the benchmarking results and careful investigation of example cases it is likely that CheckM2 is more accurate than CheckM1 in most of these instances. However, for some lineages with only a few MAGs and no complete genomes, it is difficult to assess whether CheckM1 or CheckM2 scores are more accurate. The addition of any single complete representative genome will improve and/or validate the accuracy of CheckM2's predictions for these lineages.

Finally, some binning algorithms may yield MAGs where some contigs (such as those containing single-copy marker genes) are more preferentially recovered than others (for example repeats or plasmid genes)[25]. Our investigation of the effect of binning algorithms on genome prediction accuracy using a predefined CAMI2 dataset shows that any bias is likely slight, and that overall CheckM2 is more accurate than CheckM1 or BUSCO on MAGs derived from a variety of binning algorithms (Supplementary Note 10).

Overall, we see good congruence between CheckM2 and CheckM1, and increased CheckM2 completeness scores in lineages where CheckM1 is known to have poor predictive capacity[16,26]. This gives confidence in the robustness and reliability of both estimates, given the different underlying algorithms behind both tools. Based on these results, the benchmarking data sets and investigation of individual cases (Supplementary Notes 3–8) we believe that in most cases of incongruent predictions, CheckM2 values are likely to be more accurate.

Biological insights into the ML models. It is difficult to identify the contribution of specific genomic features to the predictions of ML models used by CheckM2. Some interpretable ML approaches, such as SHAP[27], use robust mathematical techniques to approximate feature importance. While imperfect, when applied in the context of CheckM2's models, these approaches can highlight the importance of specific genes and pathways that can be further investigated and assessed independently.

According to their SHAP values, key pathways contributing to completeness predictions across most lineages are ribosomal proteins, as well as genes in the DNA processing and tRNA biosynthesis pathways. There are also individual pathways with substantially higher predictive values for only certain lineages. For example, the membrane transporters pathway in the Patescibacteria have much higher importance values compared to most other lineages that are not characterized by genome reduction or streamlining (Supplementary Table 7). Transporters are particularly noteworthy as they are likely to be key to an auxotrophic lifestyle, while also presenting an example of a set of genes that would be missed using only a conserved single-copy marker approach. The high importance placed on these pathways are in line with our biological understanding of microorganisms and give confidence that the ML models are capturing details of underlying biological reality. Average SHAP value contributions across all phyla also show that genomic

features contribute a high degree of predictive power, with eight genomic features being placed in the top 500 features for predicting completeness by the general gradient boost model (Supplementary Table 12), the highest being the number of amino acids in the genome (ranked 76th out of 21,241 input feature vectors) and number of coding sequences in the genome (ranked 143rd).

CheckM2 updates, computational benchmarking and resources. CheckM2 will be updated in line with GTDB releases. Unlike the very computationally heavy simulation method of CheckM1 (ref. 4), the simulation and training with new complete genomes taking less than 1 min per genome (per thread), with KEGG annotation of simulated genomes using DIAMOND forming the only computational bottleneck. This means that a new GTDB release can be updated into CheckM2 within 24–48 h.

During runtime, CheckM2 was consistently faster than CheckM1, processing an average of 1.56 ± 0.83 genomes per minute per thread on an AMD EPYC 7702 64-Core Processor, relative to CheckM1's 0.57 ± 0.19 genomes per minute per thread. As CheckM2 has no taxonomic determination step, its speed is more variable than CheckM1, being substantially faster when predicting the quality of small or low-completeness genomes. CheckM2 is capable of processing hundreds of thousands of genomes at a time with reasonable (less than 90 GB for a batch run of 224,000 genomes) RAM usage.

Future versions of CheckM2 will be iteratively updated and may also include additional annotation databases (for example, STRING[28] and EggNOG[29]) if this leads to significant improvements in genome quality predictions. We may also explore alternative groupings of individual genes into pathways outside of KEGG pathways, such as those provided by for example DRAM[30] or its future versions. Finally, we are exploring alternatives to the UniRef database, such as a dereplicated database of GTDB proteins annotated with the most current KEGG Orthology hidden Markov models.

Discussion

Here we present CheckM2, a ML approach for predicting completeness and contamination of microbial genomes derived from metagenomic, single-cell and isolate sequence data. When benchmarked against CheckM1, we show congruency in genome quality prediction for lineages with good genomic representation but demonstrate that CheckM2 has substantially better accuracy on medium- and low-quality genomes and genomes from lineages with poor genomic representation. We also demonstrate that in most cases it can generate highly accurate predictions for genomes in phyla with only a single genomic representative. Additionally, CheckM2 is substantially more accurate on lineages with small or reduced genomes such as the DPANN, Patescibacteria and Dependentiae, where CheckM1 often produces highly inaccurate predictions. Finally, CheckM2 typically performs better than or equal to CheckM1 on lineages with no genomic representation in the reference database.

The use of genome quality predictions from CheckM2 are likely to have important implications for existing databases and biological interpretations of new or unusual lineages. For example, CheckM2 completeness predictions will allow the inclusion of additional genomes currently excluded from GTDB due to the inaccurate CheckM1-based minimum cutoff (50% completeness), as demonstrated for the Patescibacteria phylum and DPANN.

Improved genome quality predictions by CheckM2 are the result of considering a wide variety of annotation genes in its ML models, as opposed to CheckM1's requirement for single-copy marker gene sets in each lineage. An additional advantage of the CheckM2 approach is that its models can be easily and rapidly updated to incorporate additional high-quality genomic representation for novel lineages, further increasing the accuracy of its genome quality predictions. Additionally, detection of contamination from divergent taxonomic sources may be improved through more complex training data simulation. CheckM2 is a major step forward in our ability to rapidly and accurately predict genome quality across bacterial and archaeal genomes.

Online content

Any methods, additional references, Nature Portfolio 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/s41592-023-01940-w.

References

1. Woodcroft, B. J. et al. Genome-centric view of carbon processing in thawing permafrost. *Nature* **560**, 49–54 (2018).
2. Anantharaman, K. et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* **7**, 13219 (2016).
3. 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).
4. 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).
5. AlQuraishi, M. AlphaFold at CASP13. *Bioinformatics* **35**, 4862–4865 (2019).
6. Brown, C. T. et al. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).
7. Nissen, J. N. et al. Improved metagenome binning and assembly using deep variational autoencoders. *Nat. Biotechnol.* **39**, 555–560 (2021).
8. Haft, D. H. et al. RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.* **46**, 851–860 (2017).
9. 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).
10. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
11. 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).
12. Abadi, M. et al. Tensorflow: a system for large-scale machine learning. In *Proc. 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* 265–283 (2016).
13. Ke, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **30**, 3146–3154 (2017).
14. 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).
15. Castelle, C. J. & Banfield, J. F. Major new microbial groups expand diversity and alter our understanding of the tree of life. *Cell* **172**, 1181–1197 (2018).
16. Castelle, C. J. et al. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat. Rev. Microbiol.* **16**, 629–645 (2018).
17. Méheust, R., Burstein, D., Castelle, C. J. & Banfield, J. F. The distinction of CPR bacteria from other bacteria based on protein family content. *Nat. Commun.* **10**, 4173 (2019).
18. Singleton, C. M. et al. Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. *Nat. Commun.* **12**, 2009 (2021).
19. Lui, L. M., Nielsen, T. N. & Arkin, A. P. A method for achieving complete microbial genomes and improving bins from metagenomics data. *PLoS Comput. Biol.* **17**, e1008972 (2021).

20. Orakov, A. et al. GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol.* **22**, 178 (2021).
21. Yeoh, Y. K., Sekiguchi, Y., Parks, D. H. & Hugenholtz, P. Comparative genomics of candidate phylum TM6 suggests that parasitism is widespread and ancestral in this lineage. *Mol. Biol. Evol.* **33**, 915–927 (2016).
22. Bowerman, K. L. et al. Disease-associated gut microbiome and metabolome changes in patients with chronic obstructive pulmonary disease. *Nat. Commun.* **11**, 5886 (2020).
23. Neuenschwander, S. M., Ghai, R., Pernthaler, J. & Salcher, M. M. Microdiversification in genome-streamlined ubiquitous freshwater Actinobacteria. *ISME J.* **12**, 185–198 (2018).
24. Rinke, C. et al. A phylogenomic and ecological analysis of the globally abundant Marine Group II archaea (Ca. *Poseidoniales* ord. nov.). *ISME J.* **13**, 663–675 (2019).
25. Nelson, W. C., Tully, B. J. & Mobberley, J. M. Biases in genome reconstruction from metagenomic data. *PeerJ* **8**, e10119 (2020).
26. Jarett, J. K. et al. Single-cell genomics of co-sorted Nanoarchaeota suggests novel putative host associations and diversification of proteins involved in symbiosis. *Microbiome* **6**, 161 (2018).
27. Lundberg, S. M., Allen, P. G. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Info. Proc. Syst.* **30**, 4765–4774 (2017).
28. Von Mering, C. et al. STRING 7—recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.* **35**, D358–D362 (2007).
29. Jensen, L. J. et al. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.* **36**, D250–D254 (2007).
30. Shaffer, M. et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* **48**, 8883–8900 (2020).

Methods

Simulating genome completeness and contamination

To construct a training and validation set of genomes with known ranges of completeness and contamination, all genomes from RefSeq release 89 annotated as 'complete' or 'chromosome' were downloaded and dereplicated at 99% ANI using the software package Galah v.0.2.0 (ref. 31). Genes were predicted using Prodigal v.2.6.3 (ref. 32), and resulting predicted proteins were randomly sampled using the BBMap[33] suite v.38.18 ('reformat.sh' with the 'samplerate' option) to create a range of completeness between 5 and 100% at 5% intervals using a custom script written in bash (v.4.3.48(1)-release (x86_64-suse-linux-gnu)) and python (v3.7). To simulate self-contamination, the same proteins were sampled multiple times to generate a range of 0% to 35% contamination. True completeness was calculated as $\frac{\text{amino acids in simulated genome}}{\text{amino acids in complete genome}}$. True contamination percentage was calculated as $\frac{\text{amino acids in proteins sampled>1 time}}{\text{amino acids in complete genome}}$.

For testing, two different simulation methods were used. These were purposefully distinct from the training set to minimize any overfitting to the simulation method. In the '20-kb-fragmentation' method, the input genome was fragmented into approximately 20-kb fragments using the 'shred.sh' option of the BBMap suite with median length 20,000 and variance of 2,000. These were sampled using the BBMap suite ('reformat.sh' with the 'samplereadstarget' option) to create a range of completeness and contamination values as described above. The fragments were not stitched back together to replicate the state of a MAG with multiple contigs. In the 'MAG-derived-length-fragmentation', a database of contig size distributions was created using GTDB release 95 as its base, where the contig size distribution of all MAGs with less than 350 and more than ten contigs was used as a pool to choose from. Each genome had to be at least 65% complete and no more than 5% contaminated as determined using CheckM1. For each test genome, a MAG contig distribution was randomly selected and the full genomic sequence of the genome being simulated was cut into the same length fragments relative to genome size, where the order of contig sizes was randomized. These were then sampled in the same way as in the '20-kb-fragmentation' method. For both, true completeness was calculated as $\frac{\text{bases in simulated genome}}{\text{bases in complete genome}}$. True contamination percentage was calculated as $\frac{\text{bases in contigs sampled>1 time}}{\text{bases in complete genome}}$.

Annotation of genomes

For all simulated genomes, genes were predicted using Prodigal v.2.6.3 and annotated with KEGG[34] ids using diamond v.2.0.4 'blastp' command against uniref100 (downloaded on 3 June 2018) containing KEGG Orthology (KO) annotations. To filter annotations, a query_cover of 80, a subject_cover of 80, a value of 1×10^{-5} and a percent_id of 30 was used, taking only the top hit for each gene. Annotations were converted to a frequency matrix containing all existing KEGG IDs with rows representing a simulated genome and columns representing counts of annotation detected. KOs found in the same pathway were grouped next to each other to allow sliding convolutional windows of the NN to extract useful information from this grouping. KOs present in multiple pathways were assigned only to the first pathway based on pathway alphabetical order. KEGG definitions of modules, pathways and categories were downloaded from KEGG on 26 November 2018.

After annotation, KEGG pathway, module and category completeness was calculated based on the definitions for each downloaded from KEGG on 26 November 2018 where completeness was defined as the fraction of genes present in a genome out of all genes defined in a module, pathway or category. Each module, pathway and category completeness feature vector was encoded as an additional column with fractional value between zero and one. Nested modules (modules containing other modules) were not used. Only the 'general' gradient boost model used these additional completeness feature vectors.

Additionally, the frequency counts of each amino acid, number of coding sequences and the total amino acid length of each genome was calculated and added to the protein annotations.

For testing, all genes were predicted and annotated de novo for each simulated genome.

Selection of additional genomes

To widen the scope of the trained model and reduce the uniformity of the training dataset, a small number of potentially complete non-RefSeq[35] genomes were identified using a 'repeated-quality-metrics' strategy: CheckM1 and CheckM2 (trained only on RefSeq release 89) were used to assess all GenBank genomes part of release 89. Those that had at least five or more members of a species, and had the same completeness and contamination scores from CheckM1 for more than three-quarters of them as well as a genome size within 5% of each other were selected as potentially complete. A genome of the same length assembled repeatedly to yield the same completeness and contamination statistics was used as potential evidence for completeness. As these genomes were often in multiple contigs, the contigs were sampled randomly and completeness calculated based as $\frac{\text{bases in contigs sampled}}{\text{bases in complete genome}}$. From these generated synthetic genomes, only those with less than 85% or less completeness were used for training to avoid incorrect bias for highly complete genomes in CheckM2. These new genomes were added to the training pool with a 50% lower sample weight relative to known complete RefSeq genomes. Even if some of these genomes are not complete, the addition of potential noise was also a desirable part of this process as a potential regularization constraint on the NN model. Moreover, these genomes also introduce a more accurate example of low-completeness genomes compared to genomes generated using the 'random-protein-sampling' method used for the bulk of the training set, further increasing the NN model's accuracy in these cases. These genomes (Supplementary Table 9) were added to the training pool used to train NN models but not to the gradient boost models (completeness or contamination). NN models trained with these additional genomes included those used to benchmark RefSeq release 202, benchmark the novel circular MAGs from Singleton et al.[18] and Lui et al.[19], as well as used in the final release of CheckM2 (v.0.1.2).

Training ML models

To train the 'general' gradient boost models, annotations of genomes in the training set were used as feature vectors, with contamination and completeness values being the predictor targets for the 'general' completeness and contamination models. The lightgbm[13] package (v3.2.1) was used to train a regression model with the following parameters: 'boosting type': 'gbdt', 'objective': regression, 'num_leaves': 11, 'min_data_in_leaf': 150, 'learning_rate': 0.2, 'feature_fraction': 0.5, 'bagging_fraction': 0.5, 'baging_freq': 3, 'reg_sqrt': True, 'min_child_weight': 180 for completeness and 'boosting type': 'gbdt', 'objective': regression, 'num_leaves': 211, 'learning_rate': 0.2, 'feature_fraction': 0.9, 'bagging_fraction': 0.8, 'baging_freq': 5, 'reg_sqrt': True for contamination. Both models were boosted for 450 iterations.

To train the 'specific' NN model, tensorflow[12] v.2.2.0 was used. The model architecture was encoded using the keras API and consists of a sequential model with three one-dimensional convolutional layers (kernel_size = 10, strides = 10, activation = 'relu') with size 180, followed by another with size 100. These are flattened and followed by a dense layer (size = 100, activation = 'relu') connected to an output layer (activation = 'sigmoid'). A BatchNormalisation layer was added after each convolutional layer to standardize and normalize network weights and feature vector input. The keras-specific loss_weight parameter was used with values of the completeness labels multiplied by 500. This was done to penalize the model errors harsher for more complete genomes (thus aiming for higher accuracy on higher-quality genomes). The loss

function was 'mse' (mean squared error), the optimizer was 'adam' and the learning rate was not changed from the default. The addition of the batch normalization layers added substantial improvement in accuracy, but also caused validation loss to fluctuate substantially during training. Therefore, the model was trained with checkpoints on complete RefSeq genomes simulated using 'random-protein-sampling' as well as non-RefSeq genomes identified using the 'repeated-quality metrics' strategy outlined below, for 5–15 iterations, while using validation accuracy on a subset of the training set (high-, medium- and low-quality simulated MAGs, Supplementary Table 1) to identify and select the model iteration with best validation loss across all quality levels. This process was repeated for both the validation and final CheckM2 NN models. For more details, see Supplementary Note 1.

Filtering out low-quality genomes

While most genomes in RefSeq are likely to be complete if annotated as such, there will always be exceptions. To remove potentially incomplete genomes, an intermediate NN model was trained for two epochs on all training genomes, then used to predict their completeness and contamination. Those with high deviations (more than 10% difference) between predicted and assumed completeness or contamination were removed from the training set. Notably, most of these genomes also had inferior CheckM1 metrics. Generally, they were found to belong to species that had at least one other complete genomic representative in RefSeq that did not show high deviation, indicating an issue with the genomes and not with the biology of particular lineages. A complete list is included in Supplementary Table 10.

Benchmarking performance

To benchmark CheckM2, CheckM1 and BUSCO were used as key software tools of comparison. CheckM1 v.1.0.12 was run with the flag 'lineage_wf' for automatic lineage selection. Error in prediction of completeness or contamination was defined predicted value − true value, with negative errors representing underprediction and positive errors representing overestimation of the true value. MIMAG completeness and contamination standards[11] were used to divide all benchmarking results (Supplementary Table 1). MIMAG standards are often used to report the quality of MAGs and make for a useful way to visualize CheckM2 performance across varying genome quality. As CheckM2 does not detect nor rely on other MIMAG factors such as full-length 16S RNA or the presence of tRNAs, only the completeness and contamination information was used. BUSCO[36] v.5.0.0 was run offline with reference database downloaded on 21 February 2021 with the option '–auto-lineage-prok –mode genome'. Results from BUSCO were parsed to select the specific output if it was generated, otherwise the generic output file was used. In BUSCO's output 'C' was used as completeness, while 'D' was used as contamination for benchmarking against CheckM1 and CheckM2. Fragmented genes reported by BUSCO were not included as part of the completeness percent calculation, as these represent ambiguous results. As a result, some BUSCO results may show less completeness if single-copy genes were fragmented as a result of the simulations. For the section entitled Benchmarking CheckM2 cross-contamination performance, GUNC[20] v.1.0.5 was run using the 'run' command with default settings and the database downloaded on 12 November 2022. For all tools, input consisted of nucleotide FASTA files of generated synthetic genomes.

Benchmarking calculations and visualization

For all benchmarking on synthetic genomes, completeness error (defined as predicted completeness % − actual completeness %) and contamination error (defined as predicted contamination % − actual contamination %) was calculated for each genome, and the results graphed using a letter-value plot in the Seaborn[37] package. For error calculations and reporting, only synthetic genomes with an actual contamination of less than 25% and actual completeness more than 5% were used. MAE values were calculated as $\frac{\sum_{i=1}^{n}|y_i-x_i|}{n}$ where y is the predicted value and x the true value. Phylum-wide MAE value was calculated as $\bar{x}$ across $\frac{\sum_{i=1}^{n}|y_i-x_i|}{n}$ for each phylum where y is the predicted completeness value and x the true completeness value. The pandas package (v.1.1.3) was used to analyze the prediction results and seaborn (v.0.11.0) was used to visualize results and generate figures.

Evaluation of taxonomic novelty effects on accuracy

To test the accuracy of both models on taxonomically novel groups, separate models were trained where one phylum was left out, the models were trained on all remaining phyla and the models were then tested on the omitted group to determine error when predicting completeness and contamination. This was repeated for all phyla, then repeated for all subsequent taxonomic levels of relatedness (class, order, family, genus, species) with the left-out group doubling in scope every level (one group left out per phylum, two per class, four per order and so on) to account for increasing diversity and number of iterations necessary to cover all levels. Lineages representing multiple taxonomic levels of novelty (for example, one class only containing one family) were tested only at the level of highest taxonomic difference (for example, in this case at the class level). In all cases, GTDB r89 taxonomy was used to determine leave-out groups, and only genomes in the RefSeq release 89 dataset were used for simulation and benchmarking.

Cosine similarity measure and model selection

To determine the appropriate model to use, CheckM2 uses the cosine similarity calculation ($\frac{\mathbf{A}\cdot\mathbf{B}}{||\mathbf{A}||||\mathbf{B}||}$) (where **A** and **B** are vector arrays) as cosine similarity correlates well with taxonomic novelty of query genome relative to the closest genome in the reference dataset. Rough taxonomic similarity enables selection between the general and specific completeness prediction models without the need to compute taxonomy, which would require substantially more computational time and resources.

As cosine similarity declined with completeness (Supplementary Fig. 1), a stable 'novelty ratio' was calculated by dividing completeness predicted by the general model by the squared cosine similarity, and subsequently used for selection between the NN and gradient boost models.

Input for cosine similarity calculations is identical to the input into the NN model. Based on the results from novelty testing, the median cosine similarity for novel phyla, classes, orders and some families were assigned to be predicted with the 'general' model and all other with 'specific' model, including all genomes with mean completeness prediction below 50% (Supplementary Fig. 1). As completeness declined, a slightly higher share novel genomes was assigned to the NN model to take advantage of its superior performance at lower completeness levels (see Supplementary Table 4 for the exact calculations).

Evaluation of nonredundant contamination effects on accuracy

To simulate nonredundant contamination, simulated genomes were created from RefSeq release 89. Different levels of completeness were generated by removing a random subsection of the genome to generate completeness between 50 and 100%. A contaminant fragment was then added to each of these simulated genomes, which was chosen from a taxonomic source defined by GTDB release 89 taxonomic assignment as follows: a randomly chosen isolate genome was chosen from the same species, same genus and so on, up to different domain, which included prokaryotic as well as viral and eukaryotic genomes. A random subsection of that genome was used as the contaminating contig. The contaminating fraction did not exceed a maximum of 10% contamination, where contamination was calculated as $\frac{\text{bases in contaminant fragment}}{\text{bases in complete base genome}}$. For each synthetic cross-contaminated species, only one other species was used, after which the same species could not be used as a source for

contamination at that taxonomic level. For eukaryotic genomes, a random section of either human or fungal DNA was randomly chosen, using all human or fungal sequences available in RefSeq release 89. For viral contamination, all viral sequences in RefSeq release 89 was used for the pool of candidates. If a virus was too small to make up the required 1–10% contamination relative to the genome being contaminated, a different virus was randomly chosen from the entire pool without replacement until the criteria were met.

Benchmarking speed of CheckM1 and CheckM2

Five replicate metagenomic sets were used to benchmark the speed of CheckM1 and CheckM2, consisting of an average of 450 genomes per set. Three sets consisted of MAGs randomly sampled from publicly available metagenomes, one set comprised a random selection of RefSeq r89 genomes and one set comprised a random sample of synthetic MAGs belonging to the DPANN superphylum or Patescibacteria phylum derived from MAGs in RefSeq r202. CheckM1 was run in the 'lineage_wf' mode with 45 threads and 45 pplacer_threads, while CheckM2 was run in the 'predict' mode with 45 threads. All benchmarking was done on an AMD EPYC 7702 64-Core Processor and time was determined using the 'time' bash command, where the 'real' time was for comparison. During runtime, threads were not shared with any other processes. Time taken per minute per thread was calculated as $\frac{\frac{\text{real runtime}}{\text{number of genomes}}}{45 \text{ threads}}$. Peak RAM usage for a large batch job (225,000 GTDB release 202 genomes in a single folder) was determined from 'maximum resident set size' using the command /usr/bin/time -v and visual verification using the htop command.

SHAP value calculations

SHAP values for the gradient boost models were calculated using the SHAP[27] package (v.0.39.0). To calculate the ten feature vectors contributing most toward completeness predictions by phylum (Supplementary Table 7), a TreeExplainer was used to generate SHAP values for the gradient boost completeness model in the CheckM2 release version. Aggregate sums across all 21,241 feature vectors were calculated, and the top ten were included in the Supplementary Table 7 with a mean per phylum.

Reporting summary

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

Data availability

Additional analyses supporting the conclusions of this study have been supplied as Supplementary Information. Supplementary scripts required to generate all synthetic genomes used in training and testing can be accessed from Zenodo (https://doi.org/10.5281/zenodo.6861629). Benchmarking data can be accessed from Zenodo (https://doi.org/10.5281/zenodo.8024307). A full list of feature vectors used by CheckM2 and their order can be accessed on GitHub (https://github.com/chklovski/checkm2_supplementary). The annotation vectors of all synthetic genomes used to train CheckM2, as well as completeness/contamination labels, are available as part of this repository in sparse vector format, formatted for both the NN and gradient boost models. Source data are provided with this paper.

Code availability

CheckM2 is available on GitHub (https://github.com/chklovski/CheckM2) and is released under the GNU General Public License v.3. The script required to update CheckM2 with new high-quality genomes is also available on GitHub (https://github.com/chklovski/checkm2_supplementary), although this will be carried out centrally by the CheckM2 team.

References

31. Woodcroft, B. J. Galah. *GitHub* https://github.com/wwood/galah (2020).
32. Hyatt, D. et al. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinforma.* **11**, 119 (2010).
33. Bushnell, B. BBMap: a fast, accurate, splice-aware aligner (OSTI, US DoE, 2014).
34. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
35. Benson, D. A. et al. GenBank. *Nucleic Acids Res.* **46**, D41 (2018).
36. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness. *Methods Mol. Biol.* **1962**, 227–245 (2019).
37. Waskom, M. L. Seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).

Acknowledgements

We thank E. McMaster (Queensland University of Technology, Translational Research Institute, Woolloongabba, Queensland, Australia) for her help in refining the figures. This work was supported by the National Science Foundation Biology Integration Institute – EMERGE (GRT00059410). A.C. is supported by Australian Government Research Training Program Scholarships. G.W.T. is supported by Australian Research Council (ARC) (grant no. FT170100070). B.J.W. is supported by ARC Discovery Early Career Research (grant no. DE160100248).

Author contributions

A.C. and G.W.T. designed the overall workflow and planned the key steps. A.C. generated the synthetic genomes, trained ML models, performed benchmarking and wrote the final code base of CheckM2. B.J.W. and D.H.P. guided code improvements and optimizations. G.W.T., D.H.P and B.J.W. helped interpret the result and made further suggestions for future directions and improvements. A.C. and G.W.T. drafted and wrote the manuscript. All authors edited the manuscript before submission.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-023-01940-w.

Correspondence and requests for materials should be addressed to Gene W. Tyson.

Peer review information *Nature Methods* thanks Stephen Nayfach, Mads Albertsen and C. Titus Brown for their contribution to the peer review of this work. Primary Handling Editor: Lei Tang and Lin Tang, in collaboration with the *Nature Methods* team.

nature research

Corresponding author(s): Gene Tyson, Alex Chklovski

Last updated by author(s): Jul 19, 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	No software was used for data collection - publicly available RefSeq and GenBank genomes were downloaded from public repositories.
Data analysis	To dereplicate downloaded genomes, Galah v0.2.0 was used. To generate synthetic genomes for the 'random-protein-sampling' and '20-kb-fragmentation' methods referenced in the manuscript, the BBMap suite v.38.18 was used in combination with custom bash (v4.3.48(1)-release (x86_64-suse-linux-gnu)) and python (v 3.7) scripts available from Zenodo with the DOI identifier 10.5281/zenodo.6861629. For visualisation of CheckM1 and CheckM2 performance, the data was processed with the pandas package (1.1.3), and visualised using the Seaborn package (0.11.0). CheckM2 software used for benchmarking is available from https://github.com/chklovski/CheckM2 (v 0.1.3). CheckM1 software used for benchmarking was v 1.0.12 and is available from https://github.com/Ecogenomics/CheckM/. The BUSCO software used for benchmarking was v5.0.0, run offline with reference database downloaded on 21-02-21. For all simulated genomes, genes were predicted using Prodigal v.2.6.3 and annotated with KEGG ids using diamond v2.0.4 'blastp' command against uniref100 (released 26/11/2018) containing KO annotations. To train the neural network model, tensorflow v2.2.0 was used. To train the gradient boost model, lightgbm 3.2.1 was used. SHAP values for the gradient boost models were calculated using the SHAP package (0.39.0). The GUNC software used for benchmarking was v 1.0.5.

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

The benchmarking data that support the findings of this study are available from the corresponding author upon reasonable request. (?)
The uniref100 fasta files used to annotate with KO annotations were downloaded from the UniRef website on 03-06-2018.
KEGG definitions of modules, pathways and categories were downloaded from KEGG on 26-11-2018.
All genomes used in benchmarking, testing and training were downloaded from NCBI RefSeq and Genbank releases 89 and 202.
All genome metadata and taxonomic information used for analysis was downloaded from GTDB release 202.

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	Number of complete bacterial and archaeal genomes in RefSeq releases 89, 202, as well as those publications referenced in the manuscript determined the sample size for all training and testing. Fragmentation of each complete genome to synthetic genomes with completeness ranges from 5%-100% and 0%-35% contaminated at 5% completeness intervals on average determined final sample sizes. Contamination above 35% was not simulated as these are far outside MIMAG (1) standards and accuracy in contamination prediction above 35% was unlikely to provide useful biological insights relative to the processing power required to generate such samples. Using 5% intervals in completeness were a middle ground between covering the entire range of possible levels of completeness (0-100%) and having small enough sample sizes to prototype and retrain machine learning models in a computationally tractable timeframe. 1) Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nature Biotechnology vol. 35 725–731 (2017).
Data exclusions	Some genomes were excluded due to suspected poor quality despite being annotated as complete (accession codes available in the Supplementary Tables).
Replication	For initial testing using 11 machine learning methods, 6 rounds of randomised cross-validation were performed, where the data was split into a 75%/25% train/validation sets using k-fold cross-validation by selecting GTDB-taxonomy-derived species and re-sampling if necessary ensuring at least 5 phyla/classes/orders were unique to the validation set. For novelty testing, a leave-one-out approach was used for Archaea and a leave-one/two/four/eight/sixteen/thirty-two-out approach for Bacteria from phylum-level to species-level (see Methods). For cross-contamination, for each genome in RefSeq release 89 a foreign contaminant contig was added at taxonomic levels from same species to different domain, where a genome used to supply the contig was not used again at that taxonomic level. For all benchmarking, two different methods of creating synthetic genomes (20-kb-fragmentation and random-MAG-derived fragmentation) were used on all complete genomes to produce two separate benchmarking datasets, for which results were averaged. Individual results are available in the Supplementary Data.
Randomization	Randomisation for cross-validation were picked using random numbers generated the python (v 3.6) 'random' package. Randomisation for novelty testing was determined using the same package. Benchmarking for RefSeq was determined by availability of complete genomes between software development (RefSeq release 89) and software benchmarking (RefSeq release 202 as well as additional genomes from studies cited in the manuscript).
Blinding	Blinding during cross-validation consisted of random selection of genomes, with re-sampling if necessary (see Replication). Blinding during benchmarking consisted of data released between RefSeq release 89 and RefSeq release 202 representing data not available during training, as well as genomes representing novel lineages from additional studies, many of which were not represented or sparsely represented in RefSeq release 89. Benchmarking genomes were not selected before completing the training on base models, ensuring that no bias towards these lineages was encoded.

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