CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes

Donovan H. Parks[1], Michael Imelfort[1], Connor T. Skennerton[1], Philip Hugenholtz[1,2], Gene W. Tyson[1,3]

[1]Australian Centre for Ecogenomics, School of Chemistry & Molecular Biosciences, The University of Queensland, St. Lucia, Queensland, Australia.
[2]Institute for Molecular Bioscience, The University of Queensland, St. Lucia, Queensland, Australia.
[3]Advanced Water Management Centre, The University of Queensland, St. Lucia, Queensland, Australia.

Correspondence should be addressed to Donovan Parks (d.parks@uq.edu.au) and Gene Tyson (g.tyson@uq.edu.au)

Running title: Assessing the quality of microbial genomes

Keywords: genome quality, marker genes, isolates, single-cell genomics, metagenomics, population genomes

Abstract

Large-scale recovery of genomes from isolates, single cells, and metagenomic data has been made possible by advances in computational methods and substantial reductions in sequencing costs. While this increasing breadth of draft genomes is providing key information regarding the evolutionary and functional diversity of microbial life, it has become impractical to finish all available reference genomes. Making robust biological inferences from draft genomes requires accurate estimates of their completeness and contamination. Current methods for assessing genome quality are *ad hoc* and generally make use of a limited number of 'marker' genes conserved across all bacterial or archaeal genomes. Here we introduce CheckM, an automated method for assessing the quality of a genome using a broader set of marker genes specific to the position of a genome within a reference genome tree along with information about the collocation of these genes. We demonstrate the effectiveness of CheckM using synthetic data and a wide range of isolate, single cell and metagenome derived genomes. CheckM is shown to provide accurate estimates of genome completeness and contamination, and to outperform existing approaches. Using CheckM, we identify a diverse range of errors currently impacting publicly available isolate genomes and demonstrate that genomes obtained from single cells and metagenomic data vary substantially in quality. In order to facilitate the use of draft genomes, we propose an objective measure of genome quality that can be used to select genomes suitable for specific gene- and genome-centric analyses of microbial communities. CheckM is open source software available at http://ecogenomics.github.io/CheckM.

Introduction

Recent advances in high-throughput sequencing combined with improving computational methods are enabling the rapid, cost effective recovery of genomes from cultivated and uncultivated microorganisms across a wide range of host-associated and environmental samples. Large-scale initiatives such as the Genomic Encyclopedia of Bacteria and Archaea (GEBA; Wu et al. 2009) aim to provide reference genomes from isolated species across the Tree of Life, while targeted efforts such as the Human Microbiome Project (HMP; Turnbaugh et al. 2007) and the GEBA-Root Nodulating Bacteria (GEBA-RNB; http://jgi.doe.gov/) initiatives are providing reference genomes necessary for

understanding the role of microorganisms in specific habitats. These efforts are complemented by initiatives such as the GEBA-Microbial Dark Matter (GEBA-MDM) project which used single-cell genomics to obtain genomes from major uncultivated bacterial and archaeal lineages (Rinke et al. 2013). Several studies have also demonstrated the successful recovery of high-quality population genomes directly from metagenomic data (Tyson et al. 2004; Wrighton et al. 2012; Albertsen et al. 2013; Sharon et al. 2013). Together these initiatives have produced thousands of additional draft genomes, and stand to provide tens of thousands more as sequencing technology and computational methodologies continue to improve. While this rapid recovery of genomes stands to greatly improve our understanding of the microbial world, it is outpacing our ability to manually assess the quality of individual genomes.

In order to make robust inferences from the increasing availability of draft genomes, it is critical to distinguish between genomes of varying quality (Mardis et al. 2002; Chain et al. 2009). In particular, genomes recovered from single cells or metagenomic data require careful scrutiny due to the additional complications inherent in obtaining genomes using these approaches (Dick et al. 2010; Albertsen et al. 2013; Imelfort et al. 2014). The quality of isolate genomes has traditionally been evaluated using assembly statistics such as N50 (Salzberg et al. 2012; Gurevich et al. 2013), but recent single cell and metagenomic studies have also used universal single-copy 'marker' genes for estimating genome completeness (Wrighton et al. 2012; Haroon et al. 2013; Rinke et al. 2013; Sharon et al. 2013). Specifically, completeness is estimated as the fraction of expected marker genes present within a genome. While the accuracy of this estimate has not been evaluated, it is limited by both the uneven distribution of universal marker genes across a genome and their low number, typically accounting for less than 10% of all genes (Sharon and Banfield 2013). These limitations have been partially addressed by identifying genes that are ubiquitous and single copy within a specific phylum, which increases the number of marker genes used in the estimate (Swan et al. 2013). Single-copy marker genes present multiple times within a recovered genome have also been used to estimate potential contamination (Albertsen et al. 2013; Soo et al. 2014).

Here we describe CheckM, an automated method for estimating the completeness and contamination of a genome using marker genes that are specific to its inferred lineage within a reference genome tree. Using simulated genomes of varying degrees of quality, we demonstrate that lineage-specific marker genes provide refined estimates of genome completeness and contamination compared to the universal or domain-level marker genes commonly used. Marker genes that are consistently collocated within a lineage do not provide independent evidence of a genome's quality, so collocated marker genes were grouped into marker sets in order to further refine estimates of genome quality. Our results indicate that lineage-specific marker sets provide robust estimates across all bacterial and archaeal lineages, with completeness and contamination estimates generally having an absolute error of $\leq 6\%$ even when genomes are relatively incomplete (70%) with moderate contaminated (10%).

We applied CheckM to several large datasets of genomes obtained from isolates, single cells, or metagenomic data (**Table 1**). Our results highlight that the quality of draft genomes being produced is highly variable. To address this, we propose a fixed vocabulary for defining genome quality based on estimates of completeness and contamination that is suitable for automated screening of genomes from large-scale sequencing initiatives and for annotating genomes in reference databases. We also identify several public genomes suffering from a range of errors that make them unsuitable as reference genomes. Our results demonstrate that CheckM will help identify problematic genomes before they are deposited in public databases and can be used retrospectively to establish the quality of genomes in existing public repositories. For single-cell genomes and population genomes recovered from metagenomic data, CheckM allows biological inferences to be made in the context of genome quality, and highlights genomes that would benefit from further refinement.

Results

Simulation Models for Evaluating the Accuracy of Quality Estimates

Three models were used to generate simulated genomes suitable for evaluating methods in CheckM designed to improve the robustness of completeness and contamination estimates. Under the 'random fragment' model, 3324 draft genomes were fragmented into non-overlapping windows of 5 to 50 kbp and randomly subsampled to generate genomes with varying degrees of completeness and contamination. This model allows a large number of genomes to be simulated at varying degrees of quality which provides a baseline for assessing the accuracy of completeness and contamination estimates. In order to simulate genomes reflecting the characteristics of assembled contigs, the 2430 draft genomes comprised of ≥20 contigs were used to simulate incomplete genomes contaminated with foreign DNA. Under this 'random contig' model, incomplete genomes were generated by randomly removing contigs to achieve a desired level of completeness and contamination introduced by randomly adding contigs from another draft genome.

The final model simulates population genomes that reflect the limitations of metagenomic binning methods that rely on the statistical properties of assembled contigs (e.g., tetranucleotide signatures, coverages) to determine their source genome. Since the variance of genome statistics increase with decreasing contig length, binning methods are more likely to incorrectly bin shorter contigs (Dick et al. 2010; Albertsen et al. 2013; Imelfort et al. 2014). The 'inverse length' model captures this limitation by generating incomplete and contaminated genomes in a manner similar to the random contig model, but with contigs removed or added with a probability inversely proportional to a contig's length.

For all three models, genomes were generated at 50, 70, 80, 90, 95, and 100% completeness with 0, 5, 10, 15, and 20% contamination unless stated otherwise.

Organizing Marker Genes into Collocated Sets

As marker genes are required to be present in almost all genomes within a lineage (e.g., all bacteria or archaea), they often encode essential functions and are frequently organized into operons (**Supplemental Fig. 1**). Marker genes that are consistently collocated within a lineage do not provide independent information regarding the completeness or contamination of a genome. To address this, we grouped marker genes that were consistently collocated within a lineage into marker sets and used this grouping structure to refine estimates of genome completeness and contamination. Collocated marker genes are common across all taxonomic groups with 36% of marker genes, on average, being grouped into a marker set with one or more other marker genes (**Supplemental Table S1**).

We evaluated the benefit of marker sets for assessing genome quality by applying domain-specific markers (bacteria: 104 markers, 58 sets; archaea: 150 markers, 108 sets) to genomes simulated under the random fragment model. Completeness and contamination estimates calculated with collocated marker sets were superior to estimates determined with individual marker genes regardless of the completeness or contamination of the simulated genomes (**Fig. 1**; **Supplemental Table S2**). As expected, the impact of collocated marker genes increased with the size of the window used to generate the simulated genomes. While this results in a reduction in the accuracy of quality estimates, the loss in accuracy is substantially mitigated by using marker sets as opposed to individual marker genes. The average absolute error in completeness (contamination) estimates across all simulated genomes increased from 4.3% to 5.7% (3.8% to 4.7%) when using marker sets compared to 5.5% to 9.0% (4.7% to 6.8%) when using individual marker genes as the window size was increased from 5 to 50 kbp (**Supplemental Table S2**).

To further evaluate the benefits of using collocated marker sets for estimating genome quality, domain-specific markers were used to estimate the quality of genomes simulated under the random contig and inverse length models. Under the random contig model, the average absolute error in the completeness and contamination estimates across all simulated genomes was reduced from 8.5% to 5.4% and 5.9% to 4.1%, respectively, when genome quality was estimated with marker sets as

opposed to individual marker genes (**Supplemental Fig. S2**; **Supplemental Table S3**). Similar improvements were obtained under the inverse length model though estimates were less accurate for genomes generated under this model (completeness: 10.3% to 6.6%, contamination: 8.2% to 5.6%; **Supplemental Fig. S3**; **Supplemental Table S4**).

Inference of Reference Genome Tree

Estimates of completeness and contamination can be refined by using lineage-specific marker sets (**Supplemental Fig. 1**). Lineage-specific marker sets are determined by placing query genomes into a reference genome tree (**Fig. 2**). The reference tree used by CheckM was inferred from the concatenation of 43 conserved marker genes with largely congruent phylogenetic histories (**Supplemental Table S5 and S6**). It incorporates 2052 finished and 3604 draft genomes obtained from the Integrated Microbial Genomes (IMG; Markowitz et al. 2014) database identified as being nearly complete with minimal contamination (see Methods). The inferred tree (**Supplemental Fig. S4**) shares features in common with recently published genome trees, including the class *Clostridia* being highly paraphyletic (Yutin and Galperin, 2013) and the class *Epsilonproteobacteria* residing outside the *Proteobacteria* phylum (Rinke et al. 2013). These discrepancies between phylogeny and taxonomy will cause marker genes calculated from named lineages within the genome tree to deviate from those determine strictly from assigned taxonomy. More importantly, a reference tree allows lineage-specific marker genes to be established for any internal nodes and not just those representing a named taxonomic group.

Assessment of Lineage-specific Marker Sets

Lineage-specific marker sets were determined for all nodes within the reference genome tree by identifying single-copy genes present in ≥97% of all descendant genomes. The quality of a genome can be estimated using the marker set defined at any parental node between the genome's position in the reference tree and the root. A simulation framework was used to establish the parental lineage with the most favourable set of markers for assessing the quality of genomes placed along any branch in the reference tree (**Fig. 3**). Briefly, finished genomes were used to simulate incomplete and

contaminated genomes placed along a branch, and the parental lineage whose marker genes most accurately estimated the quality of these genomes was determined.

We evaluated the effectiveness of the selected lineage-specific marker sets on genomes generated under all three simulation models. The quality of each simulated genome was estimated using marker sets inferred from genomes within i) the archaeal or bacterial lineage, ii) the lineage selected by our simulation framework, and iii) the parental lineage producing the most accurate estimates. Under all three models, the selected lineage-specific marker sets provided more accurate completeness and contamination estimates than domain-specific marker sets, and produced estimates nearly as accurate as the best performing lineage-specific marker sets (**Fig. 4**; **Table 2**; **Supplemental Figs S5** and **S6**; **Supplemental Tables S7-S9**). The improvement in quality estimates can be substantial with the average absolute error in completeness and contamination being reduced by 44.4% (5.4% to 3.0%) and 19.5% (4.1% to 3.3%) respectively when using selected lineage-specific marker sets instead of the domain-specific sets to estimate the quality of genomes generated with the random contig model. Summarizing results by the taxonomic group affiliated with each simulated genome indicated that the selected lineage-specific sets provided improved estimates compared to the domain-specific sets across all 39 classes (20 phyla) considered in this study, with the exception of the poorly sampled *Synergistetes* lineage where the estimates were largely unchanged (**Fig. 5**; **Supplemental Tables S10-S12**).

Bias in Genome Quality Estimates

Quality estimates based on marker genes or collocated marker sets exhibit a bias resulting in completeness being overestimated and contamination being underestimated (**Figs. 2** and **4**). This bias is the result of marker genes residing on foreign DNA which are absent in an incomplete genome being mistakenly interpreted as an indication of increased completeness as opposed to contamination. This bias approximately follows a binomial distribution suggesting a potential avenue for bias correction (see Methods). Nonetheless, we have elected not to correct for this bias as confounding

factors such as gene collocation make the correction approximate and the bias is small (<2%) when genomes are nearly complete (>80%) with moderate contamination (<10%; **Supplemental Fig. S7**).

Refinement for Gene Loss and Duplication

Marker sets can be refined to account for gene loss and duplication specific to the lineage of a query genome. In general, this refinement has minimal impact on the marker set and consequentially little influence on genome quality estimates. Under the random contig model, refining the marker set for lineage-specific gene loss and duplication changed completeness estimates by only 0.08% and contamination estimates by only 0.05% on average. However, the impact on quality estimates can be substantial for genomes undergoing extensive genome reduction. We applied CheckM to the 10 *Buchnera aphidicola*, 2 *Mycoplasma genitalium*, 5 *Rickettsia prowazekii*, and 7 *Borrelia burgdorferi* genomes within IMG as these species are known to be obligate symbionts with highly reduced genomes (McCutcheon and Moran 2012). These genomes have an average estimated completeness of 86.2%, 99.0%, 99.4%, and 100%, respectively, when using lineage-specific marker sets which have *not* been refined for gene loss or duplication. While this suggests that refining the marker set for gene loss is unnecessary for all these species except *Buchnera aphidicola*, accounting for genes loss within this lineage increases the average estimated completeness from 86.2% to 99.4%.

Assessment of Isolate Genomes

To benchmark CheckM on real world data, we assessed the quality of 2281 isolate genomes from the GEBA, GEBA-KMG, GEBA-PCC, GEBA-RNB, and HMP datasets (**Table 1**). Using lineage-specific marker sets, 2190 (96%) of these genomes were estimated to be ≥95% complete with ≤5% contamination (**Supplemental Table S13**) making them excellent reference genomes for analyses such as assigning taxonomy to anonymous genome fragments (Brady and Salzberg, 2009; Parks et al. 2011) or characterizing metagenomic samples using marker genes (Darling et al. 2014). The remaining 91 (4%) genomes were found to be <95% complete or >5% contaminated making them poor reference genomes for some analyses. A small number of the genomes have an estimated completeness <90% (14 genomes) or an estimated contamination >10% (5 genomes). These genomes

suffer from a diverse range of problems which we illustrate using three public genomes from the HMP available at the time of preparing this manuscript:

- The *Capnocytophaga* sp. oral taxon 329 genome (HMP id: 9074; GenBank id: AFHP00000000; IMG id: 651324019) was estimated as 100% complete and 100% contaminated by CheckM. Investigation of the 157 contigs comprising this genome revealed a bimodal GC-distribution suggesting the presence of two distinct genomes (**Supplemental Fig. S8**). We separated the contigs into two clusters by applying *k*-means clustering with *k*=2 to the tetranucleotide signatures of each contig. Placing the resulting clusters into a genome tree identified one cluster as a novel *Capnocytophaga* genome (99.0% complete, 0.2% contaminated) and the other cluster as closely related to *Paraprevotella clara YIT 11840* (100% complete, 0.4% contaminated; **Supplemental Fig. S9**).
- The least complete HMP genome reported by CheckM was the gastrointestinal *Clostridiales* sp. SM4/1 genome (HMP id: 924; GenBank id: FP929060; IMG id: 2524023221) annotated as finished at IMG and GOLD, but estimated as only 56% complete. CheckM determined the coding density of this genome to be 66% suggesting substantial assembly or gene calling errors. Further investigation revealed that 667 kbp (21.5%) of this 3.1 Mbp genome is comprised of ambiguous base pairs (Ns).
- The *Lactobacillus gasseri* MV-22 genome (HMP id: 515) available from IMG (id: 643886189) consists of 93 contigs comprising 1.89 Mbp with only 193 ambiguous bases. CheckM estimated the completeness of this genome as 90.9% when using the lineage-specific marker sets and 81.2% complete when using the bacterial marker set (**Supplemental Table S14**). While these low completeness estimates could be the result of lineage-specific gene loss, the other three *Lactobacillus gasseri* genomes from HMP are all estimated to be ≥96% complete with only the Leucyl-tRNA synthetase protein family (PF13603) exhibiting lineage-specific gene loss across the bacterial marker genes (**Supplemental Table S15**). This indicates that *Lactobacillus gasseri* MV-22 is incomplete with ≥9% of its genome estimated

to be missing. The incomplete state of this genome is not transparent from its genome size, as available *Lactobacillus gasseri* genomes are between 1.78 and 2.01 Mbp.

The issues exemplified above are not limited to the HMP or large-scale sequencing efforts. For example, the *Paracoccus denitrificans* SD1 genome (Siddavattam et al. 2011) at IMG (id: 2511231195) was estimated to be only 59% complete by CheckM (**Supplemental Table S16**). Comparing this genome to *Paracoccus denitrificans* PD1222 suggests that this species has two chromosomes and a plasmid, and that the SD1 genome is currently missing both a chromosome and its plasmid. CheckM also identified several putative submission errors as exemplified by the type strain *Oligotropha carboxidovorans* OM5 (IMG id: 650716069) which is reported as 99.7% complete and 100.9% contaminated as a result of both draft and finished versions of its chromosome and plasmid being contained in its genome sequence file.

Assessment of Single-Cell Genomes

The GEBA-MDM initiative applied single-cell genomics to novel uncultivated archaeal and bacterial cells (Rinke et al. 2013). While this is the largest single-cell sequencing initiative currently published, other large-scale initiatives are underway and have submitted initial genomes to IMG. To assess the quality of genomes recovered through single-cell genomics, we applied CheckM to i) 201 genomes recovered from individual cells in the GEBA-MDM initiative, ii) 21 genomes co-assembled from GEBA-MDM cells belonging to the same population, and iii) 409 additional genomes from unpublished studies annotated as *uncultured type* or *single cell* in IMG (**Table 1**).

Technical challenges in obtaining single-cell genomes such as low DNA yield and the associated need for genome amplification make it challenging to recover complete genomes. CheckM genomes quality estimates indicate that only 3 of the 201 (1.5%) GEBA-MDM genomes and 17 of the 409 (4.1%) unpublished single-cell genomes have an estimated completeness ≥90%. Combining cells from the same population can substantially improve completeness with the 21 combined assemblies in GEBA-MDM having an average completeness of 64.5% ± 23.3% compared to 34.2% ± 20.5% for the 201 single-cell genomes (**Supplemental Table S17**). Although current techniques for recovering

genomes from single cells result in highly incomplete genomes, these are still valuable reference genomes for analyses such as assigning taxonomy to anonymous genome fragments and resolving phylogenetic relationships (Rinke et al. 2013). However, these reference genomes should be free from substantial contamination as this will be a source of inaccuracy in such analyses. CheckM identified 42 of the 409 (10.2%) unpublished single-cell genomes to have ≥5% contamination. All the GEBA-MDM genomes were found to have <5% contamination, except two combined assemblies which were estimated to be 11.5% and 18.8% contaminated. Comparison of duplicate marker genes within these genomes suggests the contamination is the result of foreign DNA being amplified and not an assembly error.

Assessment of Population Genomes

Unlike genomes recovered from cultured isolates or single cells, genomes obtained from metagenomic data typically represent a consensus across a non-clonal microbial population. CheckM was applied to 146 population genomes recovered from four metagenomic studies (**Table 1**). As expected, the estimated completeness and contamination of these genomes vary substantially (**Fig. 6**). While population genomes are often incomplete (74 of 146 genomes are between 50% and 95% complete), they can be recovered with relatively little contamination (43 of the 74 partial genomes have ≤5% contamination; **Supplemental Table S18**). In addition to this set of 74 partial genomes, an additional 16 (11%) population genomes were estimated to be ≥95% complete with <5% contamination.

Poor quality estimates are expected for genomic elements such as plasmids or phage as the marker genes used by CheckM are specific to bacterial and archaeal chromosomes. The 10 plasmids and 11 phage identified within the acetate-amended aquifer (Wrighton et al. 2012) and infant gut (Sharon et al. 2013) datasets were estimated to be 0% complete and 0% contaminated, with the exception of two plasmids (CARSEP1P, ACD71) and one phage (ACD33) which were estimated as 4.2%, 2.7%, and 0.15% complete, respectively (**Supplementary Table S17**). The completeness of reduced genomes without representation in the reference genome tree will also be underestimated when genome

reduction has resulted in the loss of marker genes. This is illustrated by the four candidate phylum Saccharibacteria (TM7) genomes obtained from sludge bioreactor metagenomes which were estimated as between 60-70% complete by CheckM, though shown to be ≥85% complete after accounting for lineage-specific gene loss (Albertsen et al. 2013).

We compared the quality estimates obtained for the 90 putative population genomes recovered from the acetate-amended aquifer (Wrighton et al. 2012) community using domain-level marker genes and lineage-specific marker sets (**Supplemental Table S19**). While the completeness and contamination of these population genomes is unknown, these results demonstrate the degree to which quality estimates can change under these two conditions. We have focused on the acetate-amended aquifer dataset as it contains population genomes spanning a wide range of qualities, while other studies have focused exclusively on high-quality population genomes. On average, completeness changed by 13.0% and contamination by 5.1% between these two conditions. Estimates varied substantially for some genomes with completeness estimates changing by ≥15% for 36 genomes and contamination estimates changing by ≥10% for 12 genomes. While completeness estimates with domain-level marker genes and lineage-specific marker sets are highly correlated (R^2=0.84), domain-level estimates tend to overestimate the completeness of genomes relative to lineage-specific estimates (**Supplemental Fig. S10**). The correlation between contamination estimates is weaker (R^2=0.69) and any global trend less clear as the majority of population genomes exhibit <5% contamination (**Supplemental Fig. S11**).

Estimating Strain Heterogeneity

CheckM can distinguish between contamination resulting from the presence of genomic fragments from multiple strains and contamination resulting from the inclusion of genomic fragments from more divergent taxa. This is particularly useful for genomes recovered from metagenomic data as separating strains into individual genomes remains a challenging problem (Imelfort et al. 2014). These two types of contamination are differentiated by using the amino acid identity (AAI) between multi-copy genes as a measure of phylogenetic relatedness (Konstantinidis and Tiedje, 2005). Reanalysis of

the methanotrophic ANME-1 genome recovered from metagenomic data by Meyerdierks et al. (2010) with CheckM illustrates that this population genome is largely a chimera of closely related strains. Of the 229 lineage-specific marker genes used to evaluate the quality of this genome, 42 were identified as being multi-copy within the ANME-1 genome (38 present twice, 2 present three times; 82.3% completeness). While this represents approximately 21% contamination, 82.0% of the comparisons between multi-copy genes have an AAI ≥90% (76.0% at ≥95% AAI; **Supplemental Fig. S12**) revealing that the contamination is largely the result of incorporating genomic fragments from closely related taxa and that multiple ANME-1 strains are likely present within this environment.

Estimates under Opal Stop Codon Recodings

Recoding of stop codons within bacteria appears to be restricted to the opal codon (Ivanova et al. 2014), which is reassigned to either tryptophan (Yamao et al. 1985; McCutcheon et al. 2009) or glycine (Campbell et al. 2013; Rinke et al. 2013) within a few distinct lineages, e.g., *Mollicutes*, *Gracilibacteria*, candidate phylum SR1. CheckM automatically identifies genomes that have recoded the opal stop codon in order to ensure accurate completeness and contamination estimates. Among the finished IMG genomes, only 65 were identified as recoding the opal stop codon and all of these are from genera recognized for this property (**Supplemental Table S16**; e.g., *Mycoplasma*, *Ureaplasma*, *Mesoplasma*) with the exception of the two *Mycobacterium leprae* genomes that have undergone extreme genome reduction and contain high numbers of pseudogenes (Cole et al. 2001). Recoding was also correctly identified for the six population genomes from *Gracilibacteria* and candidate phylum SR1 identified by Wrighton et al. (2012) along with the single plasmid-like replicon identified as recoding the opal codon (**Supplemental Table S18**), and the two *Gracilibacteria* genomes in the GEBA-MDM dataset (**Supplemental Table S17**). All other genomes considered in this study were identified as using the standard genetic code.

Proposed Genome Quality Classification Scheme

Genomes recovered from isolates, single cells, or metagenomic data vary substantially in their quality (**Fig. 6**). To make full use of these genomes, their quality must be reported in reference databases

along with other essential genome information (Field et al. 2008). A vocabulary for discussing genomes of varying quality was proposed by Chain et al. (2009), and here we supplement this effort by broadening their proposed vocabulary and defining completeness and contamination thresholds which permit automated assignment of draft genome quality (**Table 3**). The status of *finished* is reserved for genomes assembled into a single contiguous sequence containing no gaps or ambiguities, where extensive efforts have been made to identify errors (Mardis et al. 2002). Genomes assembled into multiple sequences as a result of repetitive regions, but otherwise of a finished quality may be classified as *noncontiguous finished* (Chain et al. 2009). We propose that all other genomes be designated as *draft* and the quality of the genome qualified based on its estimated completeness and contamination.

Allowing the quality of genomes to be assigned automatically is critical for quality control in large-scale genome sequencing initiatives, and for updating genome databases as new genomes are added or techniques for estimated genome quality improve. Of the 3059 genomes (2281 isolates, 632 single cell, 146 metagenomic) considered in this study, 2216 (72.4%) were classified as being of exception quality with either no detectable (808 genomes; 26.4%) or low (1408 genomes; 46.0%) contamination. These genomes are strong candidates for being classified as *finished* or *noncontiguous finished*, but this designation should only be applied after extensive additional verification. The wide range of quality within the remaining 843 (27.6%) genomes illustrates the need for a verbose vocabulary when discussing draft genomes, e.g., 80 (2.6%) were classified as high-quality drafts with 25 (0.8%) being uncontaminated, 53 (1.7%) having low contamination, and 2 (0.06%) having medium contamination. The presence of metagenomic and single cell genomes was also transparent as 125 (4.1%) of the genomes were classified as poor-quality drafts and 521 (17.0%) as very poor-quality drafts.

Discussion

Here we introduce CheckM, a new tool developed to estimate the completeness and contamination of draft genomes derived from isolates, single cells and metagenomes using lineage-specific marker genes. To evaluate the robustness of genome quality estimates, we simulated genomes under three distinct models: i) a random fragment model where genomic fragments were removed or added uniformly across the genome, ii) a random contig model which accounts for the characteristics of assembled contigs, and iii) an inverse length model reflecting the limitations of metagenomic binning methods. Our results on simulated genomes demonstrate that when lineage-specific marker genes are organized into collocated sets, they are sufficiently spaced throughout a genome to provide accurate estimates of genome quality (**Figs. 2**, **4**, and **5**).

The robust estimates of genome quality provided by CheckM allow for automated quality screening of bacterial and archaeal genomes. Using CheckM, we were able to identify isolate genomes exhibiting a wide range of problems. Incorporation of these low-quality genomes into reference datasets will diminish the accuracy of inferences made in many studies. For example, a study of horizontal gene transfer might incorrectly predict a large number of transfers between *Capnocytophaga* and *Paraprevotella* genomes due to the *Capnocytophaga* sp. oral taxon 329 genome erroneously containing genes from both of these genera. Similarly, a comparative genomics study including the *Clostridiales* sp. SM4/1 genome identified as 56% complete due to an excessive number of ambiguous base pairs (>20% Ns) may incorrectly report the number of core genes among *Clostridiales* genomes or the ubiquity of key metabolic pathways. Comparison of the incomplete *Lactobacillus gasseri* MV-22 genome considered in this study to its GenBank (id: GL531761) counterpart revealed that this issue was localized to the IMG repository which illustrates the potential benefit of independently verify the quality of genomes at different repositories.

Many of the erroneous genomes reported in this study were brought to the attention of IMG and have subsequently been removed from their database in order to ensure IMG is producing the best possible inferences for its users. The *Capnocytophaga* sp. oral taxon 329 has also been retracted from NCBI. Consequently, we have made these genomes available at http://ecogenomic.org/checkm/public-data.

While removal of contaminated or incomplete genomes is warranted, the statistics provided by CheckM can help identify the problems associated with these genomes. In the case of *Capnocytophaga* sp. oral taxon 329, the CheckM statistics directly suggested the presence of two distinct populations which allow for the recovery of two exception-quality genomes.

Incomplete draft genomes are valuable references for many genomic analyses and their use is likely to increase as partial genomes of novel species are recovered from single cells and metagenomic data. While methodologies for handling genomes of varying qualities are currently in their infancy, it is clear many analyses will benefit from accurate estimates of completeness and contamination. The benefit of using highly incomplete genomes for assigning taxonomy to anonymous genome fragments and resolving phylogenetic relationships has already been demonstrated (Rinke et al. 2013), though analyses such as these will often produce poor results if conducted with contaminated genomes. Other analyses such as comparing the metabolic capability of different groups of genomes will likely benefit from restricting the analyses to only near complete genomes in order to ensure confident predictions can be made in regards to differences in their metabolic capabilities. Because the quality of a genome is essential for determining its suitability for different analyses, we recommend public genome repositories and new genome announcements include completeness and contamination estimates (**Table 3**) in addition to traditional assembly statistics.

The limitations of the proposed approach must be considered when interpreting CheckM quality estimates. For example, eukaryotic or phage genomes will be reported as highly incomplete as we have focused on marker sets suitable for evaluating bacterial and archaeal genomes. The quality of plasmids must also be assessed independent of CheckM. When recovering genomes from metagenomic data, the additional assembly statistics reported by CheckM (e.g., GC, coding density, coverage) can be used along with the quality estimates to help distinguish putative genomes representing fragments of an archaeal or bacterial chromosome from phage or plasmids. The estimates for highly incomplete or highly contaminated genomes must be interpreted with regards to the observed systematic bias. This bias is the result of marker genes from foreign genome being misinterpreted as an indicating of additional completeness. For example, if a 50% complete genome is

mixed with 20% contamination, then under an idealized binomial model 50% of the contaminating marker genes will be unique and the resulting genome estimated as 60% complete with 10% contamination (**Supplemental Fig. S7**). The novelty of a genome will also influence the accuracy of CheckM estimates. Estimates for bacterial and archaeal genomes from deep basal lineages with few reference genomes will be determined using domain-level marker sets instead of refined lineage-specific sets which generally provide superior estimates. This limitation is most evident for novel lineages undergoing genome reduction as demonstrated by our reanalysis of the candidate phylum Saccaribacteria (TM7) genomes. Until such genomes can be incorporated into the reference genome tree, a manual assessment of gene loss or duplication across genomes recovered within a novel lineage can be used to improve quality estimates (Albertsen et al. 2013). CheckM provides outputs suitable for performing this refinement.

We anticipate several improvements that will further refine the estimates produced by CheckM. The most substantial impact is likely to be the inclusion of additional reference genomes from lineages that are currently poorly represented. This will mitigate the number of genomes that are evaluated using broad, less accurate marker sets and improve refinements for lineage-specific gene loss and duplication. Incorporation of eukaryotic genomes into the reference tree would also be a substantive benefit when assessing population genomes recovered from environmental samples where fungi and other microbial Eukaryotes may be present. Further exploration of the parameter space of CheckM may also result in improved estimates. For instance, the 97% ubiquity criteria used to delineate marker genes is likely not optimal and the use of a probabilistic model for assessing the presence/absence of a gene across all genomes in a lineage may improve the inferred marker sets (Segata et al., 2013). Ultimately, we expect to adopt a strategy that will allow optimal values for key parameters to be determined independently for each lineage.

CheckM is the first automated tool for estimating the completeness and contamination of isolate, single cell, and population genomes. The need for accurate estimates of genome quality will only grow in importance as we continue to fill out the microbial tree of life and are better able to utilize draft genomes to inform modern gene- and genome-centric analyses of microbial communities.

Methods

Identification of Trusted Reference Genomes

Bacterial and archaeal genomes along with their associated PFAM and TIGRFAM gene annotations were downloaded from IMG (Markowitz et al. 2014) on April 4, 2013. Low-quality genomes consisting of >300 contigs or with an N50 of <10 kbp were removed from the 10,216 (9761 bacterial, 343 archaeal) IMG genomes leaving 9037 bacterial and 333 archaeal genomes. Single-copy PFAM and TIGRFAM genes present in ≥97% of the remaining bacterial or archaeal genomes annotated as finished in IMG were identified using the IMG gene annotations and used to infer domain-specific marker sets (bacteria: 83 marker genes, 42 marker sets; archaea: 140 marker genes, 100 marker sets). To identify near-complete genomes suitable for inferring lineage-specific marker sets, genomes with an estimated completeness <97% or with contamination estimated to be >3% were removed. This filtering resulted in 7820 (7613 bacteria, 207 archaea) genomes being retained of which 2119 were marked as finished in IMG and 5701 as draft. In order to mitigate bias towards specific taxa and to reduce computational requirements, this set of genomes was dereplicated to include a single representative from each strain and at most 20 genomes from each species. Genomes were selected randomly from species with >20 representatives, except preference was first given to genomes marked as finished. Dereplication reduced the set of trusted reference genomes to 5656 (5449 bacteria, 207 archaea; 2052 finished, 3604 draft).

Genome Tree Inference

A genome tree was inferred for the 5656 reference genomes from a set of 43 genes with largely congruent phylogenetic histories. An initial set of 66 universal marker genes was established by taking the intersection between bacterial and archaeal genes determined to be single copy and present in >90% of genomes. From this initial gene set, 18 multi-copy genes with divergent phylogenetic histories in >1% of the reference genomes were removed. A multi-copy gene within a genome was only deemed to have a congruent phylogenetic history if all copies of the gene were situated within a single conspecific clade within its gene tree. Genes were aligned with HMMER v3.1b1

(http://hmmer.janelia.org) and gene trees inferred with FastTree v2.1.3 (Price et al. 2009) under the WAG+GAMMA model. Trees were then modified with DendroPy v3.12.0 (Sukumaran et al. 2010) in order to root the trees between archaea and bacteria unless these groups were not monophyletic in which case midpoint rooting was used. A further five genes found to be incongruent with the IMG taxonomy were also removed as these genes may be subject to lateral transfer. Testing of taxonomic congruency was performed as described in Soo et al. (2014). The final set of 43 phylogenetically informative marker genes (**Supplemental Table S5**) consists primarily of ribosomal proteins and RNA polymerase domains, and is similar to the universal marker sets used by PhyloSift (Darling et al. 2014; **Supplemental Table S6**). A reference genome tree was inferred from the concatenated alignment of 6988 columns with FastTree v2.1.3 under the WAG+GAMMA model and rooted between bacteria and archaea. Local support values were calculated using the Shimodaira-Hasegawa test implemented in FastTree.

Determination of Lineage-specific Marker Genes

Single-copy PFAM and TIGRFAM genes were identified within reference genomes using the annotations provided by IMG. A gene was defined as a linage-specific marker gene if it occur only once in >97% of the genomes within a lineage. For protein families that occur in both the PFAM and TIGRFAM databases, the PFAM HMM is used for identifying the gene when evaluating a genome. PFAM and TIGRFAM families were considered redundant if they matched the same genes in >90% of the finished IMG genomes.

Organization of Marker Genes into Collocated Marker Sets

A pair of marker genes were considered to be collocated within a lineage if they occurred within 5 kbp of each other in >95% of genomes. Sets of collocated markers were then formed from collocated gene pairs by clustering together all pairs with a shared gene (e.g., if genes A and B, and genes B and C are collocated then they are clustered into the collocated set ABC).

Refining Marker Sets for Lineage-specific Gene Loss and Duplication

Marker set can be refined to account for gene loss and duplication specific to the lineage of a genome (**Supplemental Fig. S13**). A marker gene was considered to be lost (duplicates) within a lineage if it was absent (present multiple times) in ≥ 50% of all descendent genomes. Refinement of a marker set was achieved by removing all marker genes identified as lost or duplicated while preserving the collocated set structure.

Estimation of Completeness, Contamination, and Strain Heterogeneity

Genome completeness is estimated as the number of marker sets present in a genome taking into account that only a portion of a marker set may be identified:

$$\frac{\sum_{s \in M} \frac{|s \cap G_M|}{|s|}}{|M|} \quad (1)$$

where s is a set of collocated marker genes, M is the set of all marker sets, and G_M is the set of marker genes identified in a genome. Genome contamination is estimated from the number of multi-copy marker genes identified in each marker set:

$$\frac{\sum_{s \in M} \frac{\sum_{g \in s} C_g}{|s|}}{|M|} \quad (2)$$

where C_g is N-1 for a gene g identified $N \geq 1$ times, and 0 for a missing gene. CheckM also supports estimating completeness and contamination without arranging marker genes into collocated sets. Equations 1 and 2 can be applied to this case by assigning all marker genes to a single set (i.e., |M|=1).

Contamination resulting from multiple strains or closely-related species being binned into a single putative genome is identified by examining the AAI between multi-copy marker genes. Specifically, a strain heterogeneity index is calculated as the fraction of multi-copy gene pairs above a specified AAI threshold:

$$\frac{\sum_{g \in G} \sum_{i=1}^{|g|} \sum_{j=i+1}^{|g|} aai(g_i, g_j, t)}{\sum_{g \in G} \sum_{i=1}^{|g|} \sum_{j=i+1}^{|g|} 1} \quad (3)$$

where $g=\{g_1,g_2,...g_N\}$ is the set of hits to a marker gene, G is the set of all marker genes, and *aai* is 1 if the AAI between g_i and g_j is greater than t (default = 0.9) and 0 otherwise.

Systematic Bias of Completeness and Contamination Estimates

Completeness and contamination estimates determined using equations 1 and 2 exhibit a systematic bias. This bias is the result of treating all marker genes present exactly once as being from the query genome of interest although some of these markers may reside on contaminating contigs. Under the simplifying assumption that all marker genes are independent, this bias can be modelled as a binomial distribution. Let n be the set of marker genes, x the set of marker genes from the query genome of interest, and y the set of marker genes from other genomes. The probability of a marker gene in y not being in x is $p = (|n| - |x|)/|n|$ and the number of marker genes in y not in x will follow a binomial distribution, $X \sim B(|y|, p)$. The expected number of marker genes in y not in x is $E(X) = |y|p$. Marker genes in y not in x introduce a bias as these markers are treated as contributing to the completeness of the query genome. As such, an upper bound on this bias is given by assuming all such marker genes are unique which results in there being $|x|+|y|p$ single-copy marker genes. This gives an estimated completeness of:

$$\frac{|x| + |y|p}{|n|}$$

$$= \frac{|x|}{|n|} + \frac{|y|\,(|n| - |x|)/|n|}{|n|}$$

$$= \frac{|x|}{|n|} + \frac{|y|}{|n|}\left(1 - \frac{|x|}{|n|}\right)$$

$$= comp_t + cont_t(1 - comp_t) \quad (4)$$

where $comp_t$ and $cont_t$ are simply the true completeness and contamination of the query genome. A similar derivation gives the estimated contamination of the query genome as $cont_t - cont_t(1 - comp_t)$. Supplemental Figure 7 illustrates the degree of this bias.

Identification of Marker Genes in Putative Genomes

Open reading frames (ORFs) are predicted on all contigs comprising a putative genome using Prodigal v2.60 (Hyatt et al. 2012), and annotated using HMMER v3.1b1 (http://hmmer.janelia.org) with model specific cutoff values for both the PFAM (-cut_gc) and TIGRFAM (-cut_tc) HMMs. PFAM annotations are assigned using the same methodology as the Sanger Institute and IMG, which accounts for homologous relationships between PFAM clans (see pfam_scan.pl which is available on the Sanger Institute FTP site). While this requires searching predicted proteins with all PFAM HMMs that are from the same clan as a PFAM marker gene, it can substantially improve the identification of marker genes. ORF calling errors occasionally occur due to ambiguous bases in a contig that can result in adjacent, erroneous ORFs being assigned to the same marker gene. These errors are resolved by checking if such ORFs have a best match to adjacent, non-overlapping portions of a marker gene's HMM.

Determination of Coding Table

ORFs are called with Prodigal using both the standard translation table (i.e., table 11) and with UGA recoded for tryptophan (i.e., table 4). CheckM does not handle the recoding of UGA to glycine (i.e., table 25) though should perform well for any recoding of UGA as the resulting protein sequence will differ only slightly from its true identity ensuring marker genes are still robustly identified. Genomes recoding UGA to an amino acid have a low coding density when ORFs are predicted with the standard table. CheckM uses ORFs called with table 4 when the coding density under this table is 5% greater than it is under the standard table and the resulting coding density is ≥70%.

Selection of Lineage-specific Marker Genes

To assess the quality of a putative genome with lineage-specific marker genes it must first be placed into the reference genome tree. Phylogenetically informative marker genes are identified within a lineage as described above. Identified genes are aligned with HMMER and the concatenated alignment used to place a genome into the reference genome tree using pplacer v2.6.32 (Matsen et al. 2010).

Marker genes can be inferred at all internal nodes in the reference tree along the path from the putative genome to the root (**Fig. 3A**). The most suitable set of marker genes for assessing a genome depends on a number of factors including the novelty of the putative genome relative to the surrounding reference genomes and the breadth of diversity covered by these reference genomes. A simulation framework was used to establish the parent node producing the most suitable marker set for each branch in the reference genome tree. This allows completeness and contamination estimates for new genomes to be assess using the lineage-specific marker genes associated with the position of each genome in the reference tree.

The simulation framework was restricted to the 2052 finished reference genomes, as draft genomes were used for evaluating the performance of CheckM. For each branch, the descendant lineage with the fewest genomes was removed from the reference tree (**Fig. 3B**). These genomes were used as proxies to simulate genomes that would be placed on this branch. Each genome was fragmented into 10 kbp windows and used to simulate 100 independent genomes with completeness randomly selected between 50-100% and contamination randomly selected between 0-20% (**Fig. 3C**). Marker genes were then inferred for each parent node and used to assess the completeness and contamination of the simulated genomes. For the purposes of this simulation, marker genes were not formed into marker sets in order to reduce computational complexity and to allow a fair assessment of how this feature influences genome assessment. The parental node whose inferred marker genes minimize the error in the estimated completeness and contamination over all simulated genomes was assigned to the branch (**Fig. 3D**):

$$\arg\min_{m \in M} = \sum_{g \in G} \sum_{i=1}^{N} |comp_{est}(g_i, m) - comp_t(g_i)| + |cont_{est}(g_i, m) - cont_t(g_i)| \quad (5)$$

where m is a set of marker genes, M is the set of marker gene for each parent node, $comp_{est}(g_i,m)$ is the estimated completeness of simulated genome g_i using m, $comp_t(g_i)$ is the true completeness of g_i, $cont_{est}(g_i,m)$ and $cont_t(g_i)$ are analogous functions for contamination, and N is the number of simulated genomes derived from g. Marker genes associated with each internal node were calculated *de novo* during the simulation to reflect removing the test genomes and then re-calculated afterwards using all available reference genomes in order to produced refined sets of marker genes.

Simulation of Incomplete and Contaminated Isolate and Population Genomes

Incomplete and contaminated genomes were simulated to determine appropriate lineage-specific marker sets and to evaluate the performance of CheckM. To simulate a large number of genomes at different degrees of completeness and contamination, 3324 draft genomes obtained from IMG were randomly subsampled. Under this random fragment simulation scheme, each contig comprising a genome was fragmented into non-overlapping windows of a fixed size between 5 and 50 kbp. Fragments were then sampled without (with) replacement to generated genomes at a desired level of completeness (contamination; **Fig. 3D**). Due to the discrete nature of this simulation, an exact degree of completeness and contamination cannot typically be achieved. Windows were sampled until the genome had a simulated completeness and contamination equal to or just greater than the target values. Generation of simulated genomes was limited to genomes annotated as draft at IMG as finished genomes were used to determine appropriate parental lineages for inferring marker sets as described above.

The 2430 draft reference genomes comprised of ≥20 contigs were used to simulate partial and contaminated genomes which reflected the characteristics of assembled contigs. Under the random contig model, genomes were generated by randomly removing contigs until the simulated genome reached or fell below a target completeness level. Contamination was introduced by randomly adding contigs from a randomly selected draft genome until the simulated genome reached or exceeded the

desired level of contamination. These 2430 draft genomes were also used to generate genomes reflecting the limitations of metagenomic binning tools. As binning methods rely on the statistical properties of contigs (e.g., tetranucleotide signature, coverage) to determine source genomes, they are more likely to incorrectly bin shorter contigs (Dick et al. 2010; Albertsen et al. 2013; Imelfort et al. 2014). To simulate this, partial genomes were generated by randomly removing contigs with a probability inversely proportional to their length until the simulated genome reached or fell below a target completeness level. Contamination was simulated by randomly selecting another draft reference genome and adding contigs from this genome with a probability inversely proportional to length until the simulated genome reached or exceeded the desired level of contamination.

Evaluation using Simulated Genomes

Evaluation of CheckM was performed using the draft reference genomes. To help elevate bias towards well-sampled lineages and highly similar genomes, 280 of the 3604 draft genomes with identical phylogenetic marker genes (strains of the same species) were not considered during evaluation. For each of the remaining 3224 draft genomes, 20 genomes were simulated for all combinations of 50, 70, 80, 90, 95 and 100% completeness with 0, 5, 10, 15, or 20% contamination. Marker genes and marker sets were inferred with the test genome removed from the set of reference genomes (i.e., leave-one-out testing) and their performance evaluated by considering the absolute error in completeness and contamination estimates. To evaluate the performance of the lineage-specific markers selected by the simulation framework described above, results were compared to the lineage-specific markers resulting in the best performance as determined by applying equation 4 independently to each set of simulated genomes generated from a test genome at a specific level of completeness and contamination. This represents a highly idealized case, as it assumes a method capable of selecting different optimal lineage-specific markers for the same genome under varying levels of completeness and contamination. The evaluation with genomes simulated to reflect incorrect metagenomic binning was done for the same set of target completeness and contamination values.

Genome Datasets

Population genomes from the Wrighton et al. (2012) and Sharon et al. (2013) studies were downloaded from ggKbase (http://ggkbase.berkeley.edu/) on March 31, 2014. Tyson et al. (2004) and Meyerdierks et al. (2010) population genomes were obtained from NCBI. The population genomes from the Albertsen et al. (2013) study can be obtained from http://ecogenomic.org/checkm/public-data. Reference genomes at NCBI and IMG are occasionally removed or modified. For posterity, the reference genomes analyzed in this paper have been achieved at http://ecogenomic.org/checkm/public-data. The GEBA, GEBA-KMG, GEBA-PCC, GEBA-RNB, GEBA-MDM, and HMP genomes comprise part of the data downloaded from IMG on April 4, 2014.

Acknowledgements

Many of the genomes considered in this manuscript were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/ in collaboration with the user community.

DHP is supported by the Natural Sciences and Engineering Research Council of Canada. MI is supported by a Great Barrier Reef Foundation Postdoctoral Research Fellowship through the ReFuGe2020 consortium. CTS was supported by an Australian Postgraduate Award from the Australian Research Council. GWT and PH are supported by a Discovery Outstanding Researcher Award (DORA) and Queen Elizabeth II Fellowship from the Australian Research Council, grants DP120103498 and DP1093175, respectively.

Figures

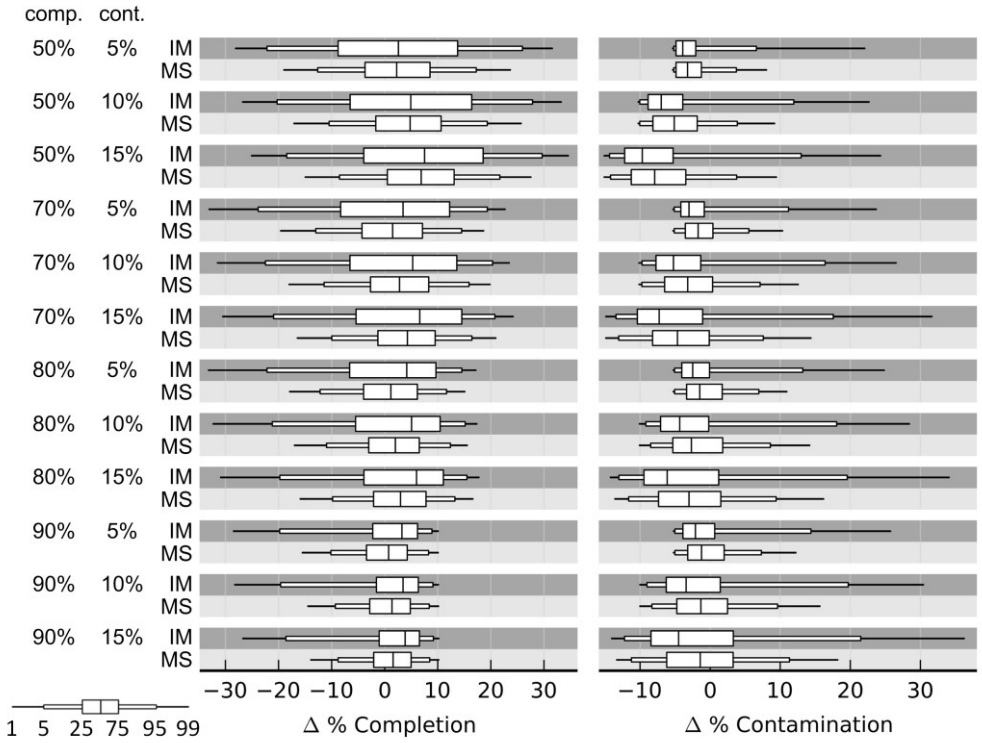


Figure 1. Absolute error in completeness and contamination estimates on simulated genomes with 50%, 70%, 80%, or 90% completeness (comp.) and 5%, 10%, or 15% contamination (cont.). Quality estimates were determined using domain-level marker genes treated as individual markers (IM) or organized into collocated marker sets (MS). Simulated genomes were generated under the random fragment model where genomes were fragmented into non-overlapping windows of 20 kbp which were randomly subsampled to generate genomes with the desired levels of completeness and contamination. Simulated genomes were generated from 3324 draft genomes spanning 39 classes (20 phyla) with each draft genome being used to generate 20 simulated genomes. A systematic bias in the estimates results in completeness being overestimated on average (median value to the right of zero) and contamination being underestimated on average (median value to the left of zero). Results are summarized using box-and-whisker plots showing the 1st (99th), 5th (95th), 25th (75th), and 50th percentiles.

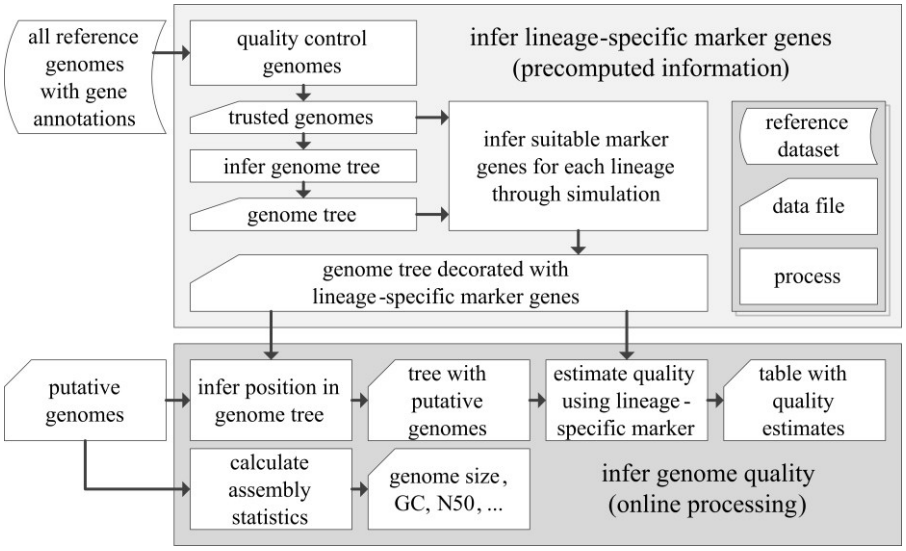


Figure 2. CheckM consists of a workflow for precomputing lineage-specific marker genes for each branch within a reference genome tree (top box) and an online workflow for inferring the quality of putative genomes (bottom box). Starting with a set of annotated reference genomes, the quality of these genomes is assessed in order to produce a set of near complete genomes suitable for inferring marker genes. These genomes form the bases of a reference genome tree. A simulation framework is then used to associate every branch in the reference genome tree with a parental node which spans enough genomic diversity to produce marker genes suitable for robustly estimating the quality of genomes placed along this branch. This computation is expensive, but only needs to be performed once. To determine the quality of a putative genome, its position within the reference genome tree is inferred in order to establish the set of marker genes suitable for assessing its quality. These marker genes are identified within the putative genome and the presence/absence of these genes used to estimate its completeness and contamination. CheckM also calculates standard assembly statistics for each putative genome (e.g., genome size, GC, N50, # scaffolds, coding density).

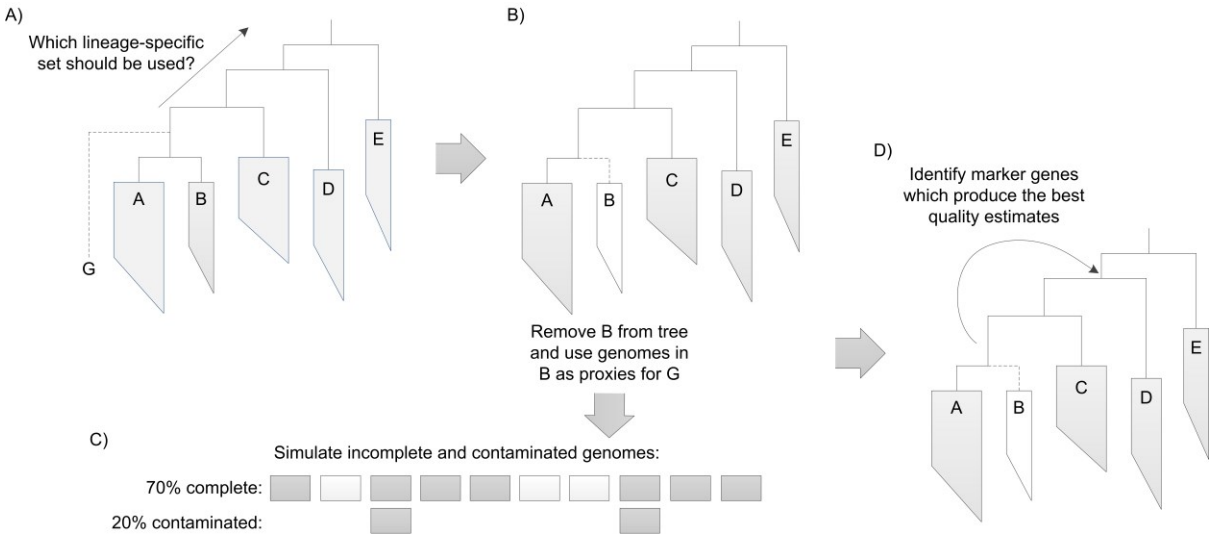
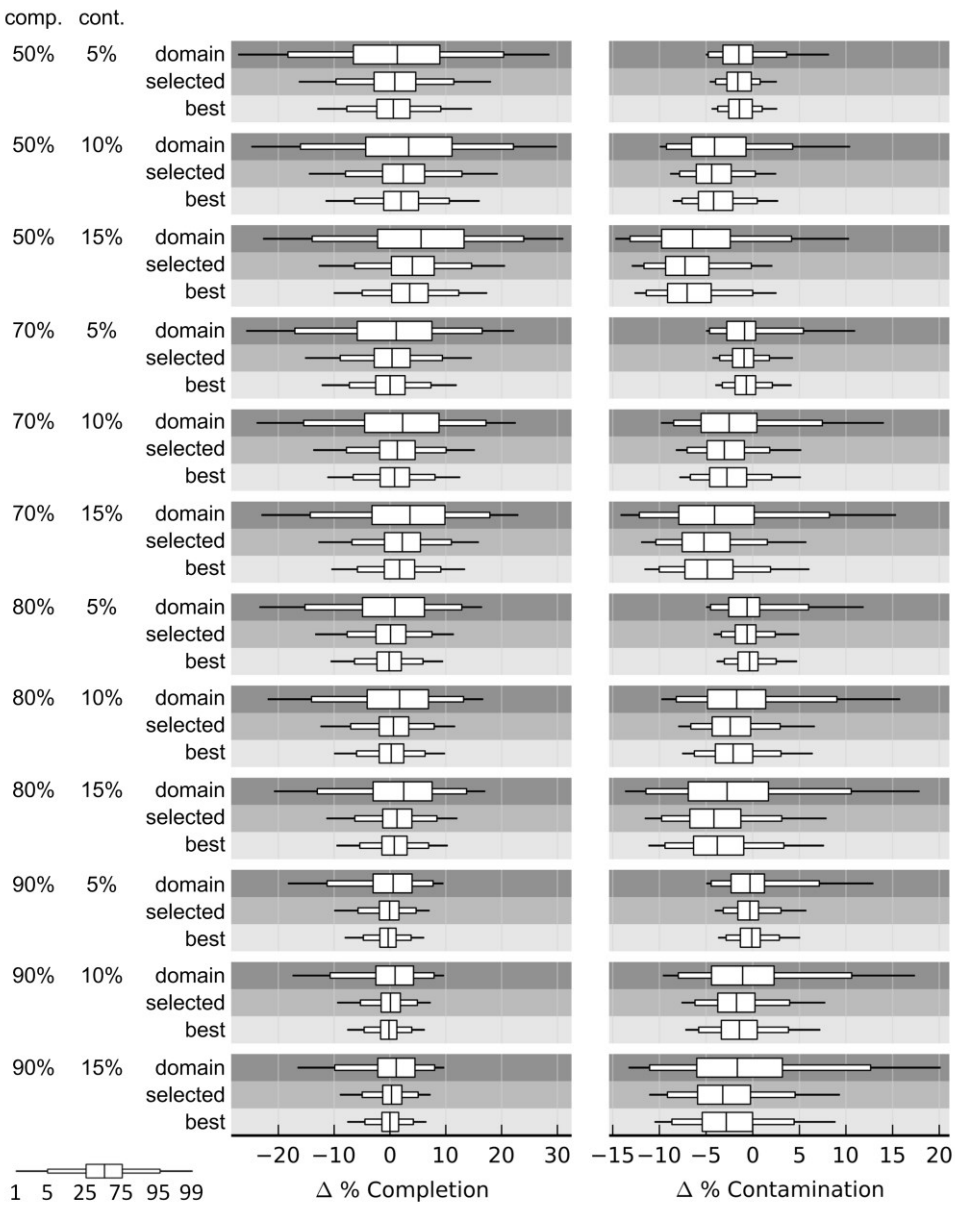
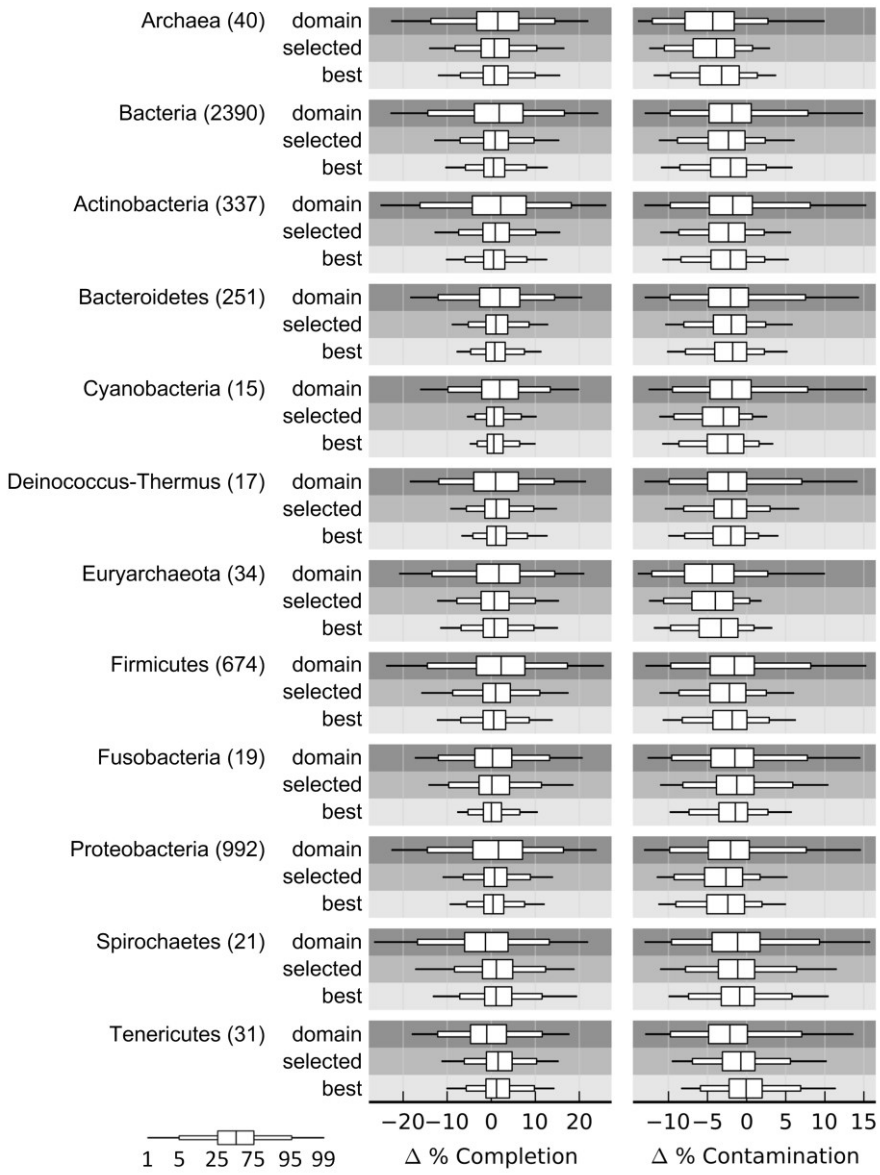


Figure 3. Overview of simulation framework for selecting lineage-specific marker genes. To evaluate a genome *G*, it is placed into a reference genome tree (A). Each parental node from the point of insertion to the root of the tree defines a lineage-specific marker set which may be used to estimate the completeness and contamination of this genome. To select a suitable set of lineage-specific marker genes for evaluating *G*, the genomes in the child lineage of *G* with the fewest genomes were used as proxies for *G* (B). Genomes at different levels of completeness and contamination were simulated from these proxy genomes by subsampling and duplicating fixed sized genomic fragments (C). Each parental marker set was then used to estimate the completeness and contamination of these simulated genomes, and the marker set resulting in the best average performance over all simulated genomes identified so it can be used to assess any genomes subsequently inserted along this branch (D).



705 **Figure 4**. Absolute error in completeness and contamination estimates on simulated genomes with 50%, 70%, 80%, or 90% completeness and 5%, 10%, or 15% contamination. Quality estimates were determined using i) domain: marker sets inferred across all archaeal or bacterial genomes, ii) selected: marker sets inferred from genomes within the lineage selected by CheckM, and iii) best: marker sets inferred from genomes within the lineage producing the most accurate estimates. Simulated genomes
710 were generated under the random contig model where draft genomes comprised of ≥20 contigs were randomly subsampled to achieve a desired level of completeness and contamination introduced by randomly adding contigs from another draft genome. Simulated genomes were generated from 2430 draft genomes spanning 31 classes (18 phyla) with each draft genome being used to generate 20

simulated genomes. A systematic bias in the estimates results in completeness being overestimated on average (median value to the right of zero) and contamination being underestimated on average (median value to the left of zero). Results are summarized using box-and-whisker plots showing the 1st (99th), 5th (95th), 25th (75th), and 50th percentiles.



720 **Figure 5**. Absolute error in completeness and contamination estimates on simulated genomes from different phyla. Quality estimates were determined using i) domain: marker sets inferred across all archaeal or bacterial genomes, ii) selected: marker sets inferred from genomes within the lineage selected by CheckM, and iii) best: marker sets inferred from genomes within the lineage producing the most accurate estimates. Simulated genomes were generated under the random contig model 725 where draft genomes comprised of ≥20 contigs are randomly subsampled to achieve a desired level of completeness and contamination introduced by randomly adding contigs from another draft genome. Simulated genomes were generated from 2430 draft with each draft genome being used to generate 20 simulated genomes with a completeness of 50%, 70%, 80%, or 90% and contamination of 5%, 10%,

or 15%. Results are summarized using box-and-whisker plots showing the 1st (99th), 5th (95th), 25th (75th), and 50th percentiles.

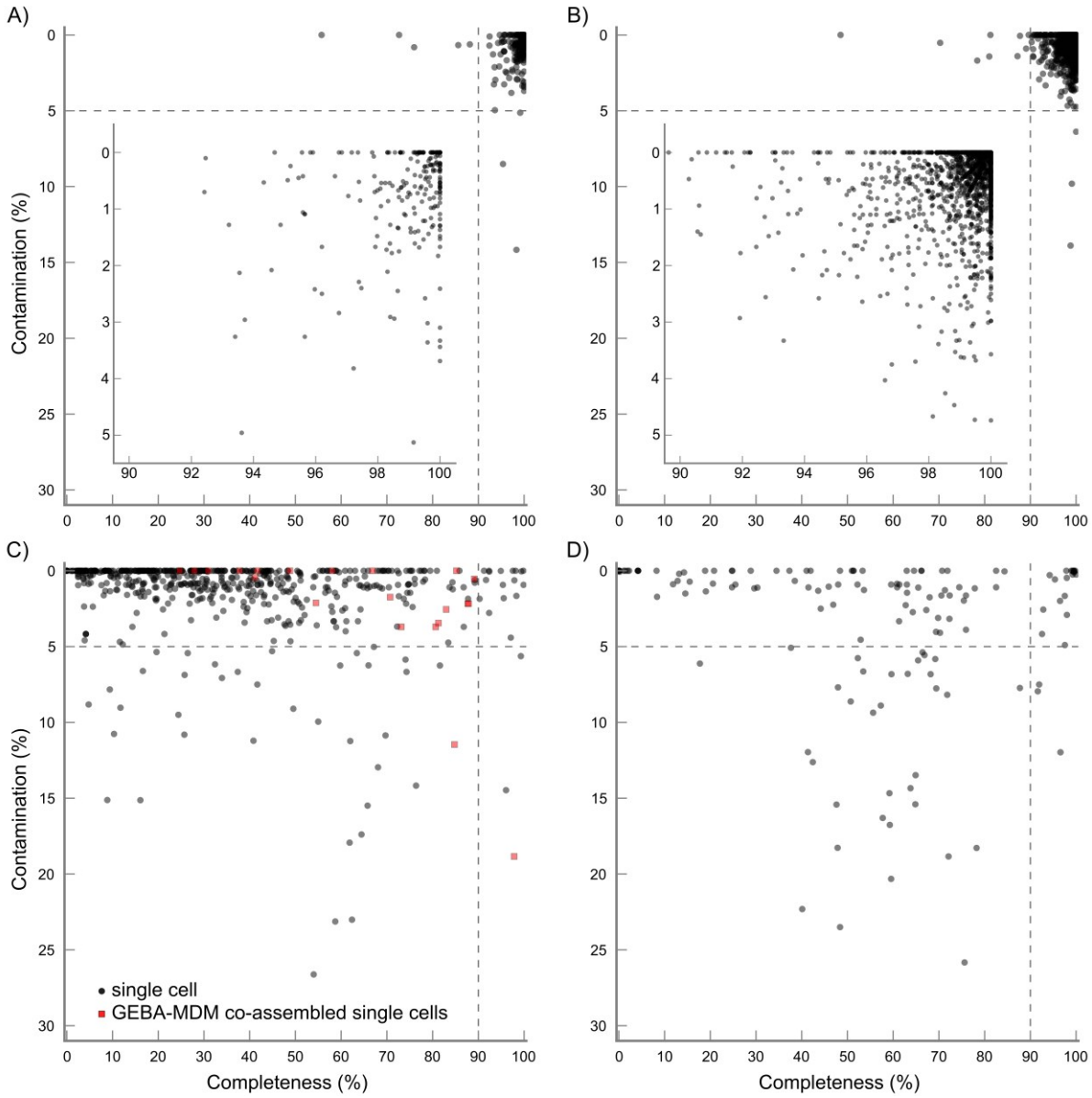


Figure 6. Lineage-specific completeness and contamination estimates for 262 isolates annotated as finished in IMG (A), 2019 isolates annotated as draft in IMG (B), 632 genomes recovered using single cell genomics (C), and 146 population genomes recovered from metagenomic data (D). Dashed lines indicate the criteria required for a genome to be considered a high-quality draft with low contamination. Insets give a more detailed view of the quality of the isolate genomes. The 2281 isolate genomes were obtained from IMG and sequenced as part of the GEBA, GEBA-KMG, GEBA-PCC, GEBA-RNB, or HMP initiatives.

Tables

Table 1. Completeness and contamination of genomes from large-scale sequencing projects.

		Completeness				Contamination			
Isolates	**Genomes**	**100%**	**≥95%**	**≥90%**	**<90%**	**0%**	**≤5%**	**≤10%**	**>10%**
GEBA	244	34.0	60.7	4.5	0.8	28.3	70.5	0.4	0.8
GEBA-KMG	724	35.5	62.8	1.7	0	31.6	67.8	0.3	0.3
GEBA-PCC	55	20.0	78.2	1.8	0	20.0	78.2	1.8	0
GEBA-RNB	92	55.4	44.6	0	0	23.9	76.1	0	0
HMP	1166	26.1	71.6	1.5	0.8	36.3	63.2	0.3	0.2
Single cells									
GEBA-MDM	201	0	0	1.5	98.5	51.2	48.3	0.5	0.0
GEBA-MDM (combined)	21	4.8	0	4.8	90.5	28.6	52.4	9.5	9.5
IMG single cell	410	0	3.4	1.0	95.6	31.5	53.3	8.1	7.1
Metagenomics									
Sludge bioreactor	13	7.7	61.5	0	30.8	30.8	61.5	7.7	0
Acid mine drainage	5	0	0	20.0	80.0	0	40.0	40.0	20.0
Infant gut	16	0	43.8	0	56.2	50.0	43.8	0	6.2
Acetate-amended aquifer	90	0	1.1	2.2	96.7	15.6	44.4	13.3	26.7
Acetate-amended aquifer*	22	0	0	13.6	86.4	13.6	68.2	9.1	9.1
Mixed									
'Finished' IMG genomes	2360	26.0	68.4	2.6	3.0	37.4	62.0	0.5	0.1

* re-binning of select Wrighton et al. (2012) bins by Albertsen et al. (2013)

References: GEBA (Wu et al. 2009), GEBA-PCC (Shih et al. 2012), HMP (Turnbaugh et al. 2007), GEBA-MDM (Rinke et al. 2013), IMG (Markowitz et al. 2014), Sludge bioreactor (Albertsen et al. 2013), Acid mine drainage (Tyson et al. 2004), Infant gut (Sharon et al. 2013), Acetate-amended aquifer (Wrighton et al. 2012), IMG (Markowitz et al. 2014). GEBA-RNB genomes were produced by the US Department of Energy Joint Genome Institute.

Table 2. Average absolute error in completeness (comp.) and contamination (cont.) estimates for i) domain: marker sets inferred across all archaeal or bacterial genomes, ii) selected: marker sets inferred from genomes within the lineage selected by CheckM, and iii) best: marker sets inferred from genomes within the parental lineage producing the most accurate estimates.

	Domain		Selected		Best	
Simulation model	**Comp. (%)**	**Cont. (%)**	**Comp. (%)**	**Cont. (%)**	**Comp. (%)**	**Cont. (%)**
random fragment, 5 kbp	4.3 ± 4.29	3.8 ± 3.73	2.6 ± 2.75	2.4 ± 2.49	2.3 ± 2.51	2.2 ± 2.37
random fragment, 20 kbp	5.0 ± 4.89	4.3 ± 4.23	3.0 ± 3.06	2.7 ± 2.73	2.6 ± 2.75	2.4 ± 2.54
random fragment, 50 kbp	5.7 ± 5.37	4.7 ± 4.65	3.4 ± 3.41	2.9 ± 3.01	2.9 ± 3.04	2.6 ± 2.77
random contig	5.4 ± 5.85	4.1 ± 4.37	3.0 ± 3.47	3.3 ± 3.43	2.5 ± 2.90	3.1 ± 3.27
inverse length	6.6 ± 6.54	5.6 ± 5.26	4.2 ± 4.38	5.3 ± 4.92	3.6 ± 3.91	4.9 ± 4.71

Table 3. Controlled vocabulary of draft genome quality based on estimated genome completeness and contamination.

Completeness	**Classification**	**Contamination**	**Classification**
≥ 95%	Exceptional quality	0%	Uncontaminated
≥ 90%	High quality	≤ 5%	Low
≥ 80%	Near complete	≤ 10%	Moderate
≥ 70%	Standard quality	≤ 15%	High
≥ 50%	Poor quality	> 15%	Excessive
< 50%	Very poor quality		

References

Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnol* **31**: 533-538.

Brady A, Salzberg SL. 2009. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods* **6**: 673-676.

Campbell JH, O'Donoghue P, Campbell AG, Schwientek P, Sczyrba A, Woyke T, Söll D, Podar M. 2013. UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proc Natl Acad Sci USA* **110**: 5540–5545.

Chain PS, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C, et al. 2009. Genome project standard in a new era of sequencing. *Science* **326**: 236-237.

Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honoré N, Garnier T, Churcher C, Harris D, et al. 2001. Massive gene decay in the leprosy bacillus. *Nature* **22**: 1007-1011.

Darling AE, Jospin G, Lowe E, Matsen FA, Bik HM, Eisen JA. 2014. PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* **2**: e243, doi: 10.7717/peerj.243.

Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, Banfield JF. 2010. Community-wide analysis of microbial genome sequences signatures. *Genome Biol* **10**: R85, doi: 10.1186/gb-2009-10-8-r85.

Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, et al. (2008). The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnol* **26**: 541-547.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al. 2014. Pfam: the protein families database. *Nucleic Acids Res* **42**: D222-230.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **15**: 1072-1075.

Haft DH, Selengut JD, White O. 2003. The TIGRFAMs database of protein families. *Nucleic Acids Res* **31**: 371-373.

Haroon MF, Hu S, Shi Y, Imelfort M, Keller J, Hugenholtz P, Yuan Z, Tyson GW. 2013. Anaerobic oxidation of methane coupled to nitrate reduction in a novel archaeal lineage. *Nature* **500**: 567-570.

Hyatt D, Locascio PF, Hauser LJ, Uberbacher EC. 2012. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **28**: 2223-2230.

Imelfort M, Parks DH, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW. 2014. GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ PrePrints* **2**: ev409v1 http://dx.doi.org/10.7287/peerj.preprints.409v1.

Ivanova NN, Schwientek P, Tripp HJ, Rinke C, Pati A, Huntemann M, Visel A, Woyke T, Kyrpides NC, Rubin EM. 2014. Stop codon reassignments in the wild. *Science* **344**: 909-913.

Konstantinidis KT and Tiedje JM. 2005. Towards a genome-based taxonomy for prokaryotes. *J Bacteriol* **187**: 6258-6264.

Mardis E, McPherson J, Martienssen R, Wilson RK, McCombie WR. 2002. What is finished, and why does it matter. *Genome Res* **12**: 669-671.

Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Pillay M, Ratner A, Huang J, Woyke T, Huntemann M, et al. 2014. IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucl Acids Res* **42**: D560-D567.

Matsen FA, Kodner RB, Armbrust EV. 2010. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11**: doi:10.1186/1471-2105-11-538.

McCutcheon JP and Moran NA. 2011. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* **8**: 13-26.

McCutcheon JP, McDonald BR, Moran NA. 2009. Origin of an alternative genetic code in the extremely small and GC–rich genome of a bacterial symbiont. *PLoS Genet* **5**: e1000565, doi: 10.1371/journal.pgen.1000565.

Meyerdierks A, Kube M, Kostadinov I, Teeling H, Glöckner FO, Reinhardt R, Amann R. 2010. Metagenome and mRNA expression analyses of anaerobic methanotropic archaea of the ANME-1 group. *Environ Microbiol* **12**: 422-439.

Parks DH, MacDonald N, Beiko R. 2011. Classifying short genomic fragments from novel lineages using composition and homology. *BMC Bioinformatics* **12**: 328, doi: 10.1186/1471-2105-12-328.

Price MN, Dehal PS, Arkin AP. 2009. FastTree: Computing large minimum-evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**: 1641-1650.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA, et al. 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**: 431-437.

Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, et al. 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res* **22**: 557-567.

Sharon I, Banfield JF. 2013. Genome from metagenomics. *Science* **342**: 1057-1058.

Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF. 2013. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res* **23**: 111-120.

Segata N, Börnigen D, Morgan XC, Huttenhower C. 2013. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat Commun* **4**: doi:10.1038/ncomms3304.

Shih PM, Wu D, Latifi A, Axen SD, Fewer DP, Talla E, Calteau A, Cai F, Tandeau de Marsac N, Rippka R, et al. 2013. Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc Natl Acad Sci USA* **110**: 1053-1058.

Siddavattam D, Karegoudar TN, Modde SK, Kumar N, Baddam R, Avasthi TS, Ahmed N. 2011. Genome of a novel isolate of *Paracoccus denitrificans* capable of degrading *N,N*-dimethylformamide. *J Bacteriol* **193**: 5598-5599.

Soo RM, Skennerton CT, Sekiguchi Y, Imelfort M, Paech SJ, Dennis PG, Steen JA, Parks DH, Tyson GW, Hugenholtz P. 2014. An expanded genomic representation of the phylum Cyanobacteria. *Genome Biol and Evol* **6**: 1031-1045.

Sukumaran J and Holder MT. 2010. DendroPy: A Python library for phylogenetic computing. *Bioinformatics* **26**: 1569-1571.

Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, González JM, Luo H, Wright JJ, Landry ZC, Hanson NW, et al. 2013. Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc Natl Acad Sci USA* **110**: 11463-11468.

Turnbaugh PJ, Ley RE, Hamady M, Frader-Liggett CM, Knight R, Gordon JI. 2007. The Human Microbiome Project. *Nature* **449**: 804-810.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37-43.

Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH, et al. 2012. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* **337**: 1661-1665.

Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, et al. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**: 1056-1060.

Yamao F, Muto A, Kawauchi Y, Iwami M, Iwagami S, Azumi Y, Osawa S. 1985. UGA is read as tryptophan in Mycoplasma capricolum. *Proc Natl Acad Sci USA* **82**: 2306–2309.

Yutin N, Galperin MY. 2013. A genomic update on clostridial phylogeny: Gram-negative spore formers and other misplaced clostridia. *Environ Microbiol* **15**: 2631-2641.