**Supplemental Information**

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

Donovan H. Parks, Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, Gene W. Tyson

**Supplementary Results**

## 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 (data not shown). However, the impact on quality estimates can be substantial for genomes undergoing extensive genome reduction. We applied CheckM to the 10 [*Buchnera aphidicola*](http://en.wikipedia.org/wiki/Buchnera_aphidicola), 2 *Mycoplasma genitalium*, 5 [*Rickettsia*](http://en.wikipedia.org/wiki/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*](http://en.wikipedia.org/wiki/Buchnera_aphidicola), accounting for genes loss within this lineage increases the average estimated completeness from 86.2% to 99.4%.

## 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*. 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 S19**; 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 *Gracilibacteria* population genomes identified by Wrighton et al. (2012) along with the single plasmid-like replicon identified as recoding the opal codon (**Supplemental Table S21**), and the two *Gracilibacteria* genomes in the GEBA-MDM dataset (**Supplemental Table S20**). All other genomes considered in this study were identified as using the standard bacterial/archaeal genetic code.

**Supplementary Methods**

## Identification of Trusted Reference Genomes

Bacterial and archaeal genomes along with their associated Pfam (Finn et al. 2014) and TIGRFAMs (Haft et al. 2003) gene annotations were downloaded from IMG (Markowitz et al. 2014) on April 4, 2014. 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 TIGRFAMs 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 high-quality genomes suitable for inferring lineage-specific marker sets, the inferred domain-specific marker sets were used to removed genomes with an estimated completeness <97% or estimated contamination >3% (see Equations 1 and 2). 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).

## Refining Marker Sets for Lineage-specific Gene Loss and Duplication

Marker sets can be refined to account for gene loss and duplication specific to the lineage of a genome (**Supplemental Fig. S14**). A marker gene was considered to be lost (duplicated) 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.

## Determination of Coding Table

Genes are called with Prodigal using both the standard bacterial/archaeal genetic code (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 genes are predicted with the standard bacterial/archaeal table. CheckM uses genes called with table 4 when the coding density under this table is 5% greater than it is under the standard bacterial/archaeal table and the resulting coding density is ≥70%.

## 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 and the number of marker genes in *y* not in *x* will follow a binomial distribution, . The expected number of marker genes in *y* not in *x* is . 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:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  | (5) |

where and are the true completeness and contamination of the query genome. A similar derivation gives the estimated contamination of the query genome as . Supplemental Figure S7 illustrates the degree of this bias. While it is tempting to use Equation 5 to correct for this bias, the function is not invertible and represents an idealized model that does not account for confounding factors such as gene collocation.

**Supplementary References**

[Campbell JH](http://www.ncbi.nlm.nih.gov/pubmed?term=Campbell%20JH%5BAuthor%5D&cauthor=true&cauthor_uid=23509275), [O'Donoghue P](http://www.ncbi.nlm.nih.gov/pubmed?term=O%27Donoghue%20P%5BAuthor%5D&cauthor=true&cauthor_uid=23509275), [Campbell AG](http://www.ncbi.nlm.nih.gov/pubmed?term=Campbell%20AG%5BAuthor%5D&cauthor=true&cauthor_uid=23509275), [Schwientek P](http://www.ncbi.nlm.nih.gov/pubmed?term=Schwientek%20P%5BAuthor%5D&cauthor=true&cauthor_uid=23509275), [Sczyrba A](http://www.ncbi.nlm.nih.gov/pubmed?term=Sczyrba%20A%5BAuthor%5D&cauthor=true&cauthor_uid=23509275), [Woyke T](http://www.ncbi.nlm.nih.gov/pubmed?term=Woyke%20T%5BAuthor%5D&cauthor=true&cauthor_uid=23509275), [Söll D](http://www.ncbi.nlm.nih.gov/pubmed?term=S%C3%B6ll%20D%5BAuthor%5D&cauthor=true&cauthor_uid=23509275), [Podar M](http://www.ncbi.nlm.nih.gov/pubmed?term=Podar%20M%5BAuthor%5D&cauthor=true&cauthor_uid=23509275). 2013. UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proc Natl Acad Sci* *USA* **110**: 5540–5545.

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.

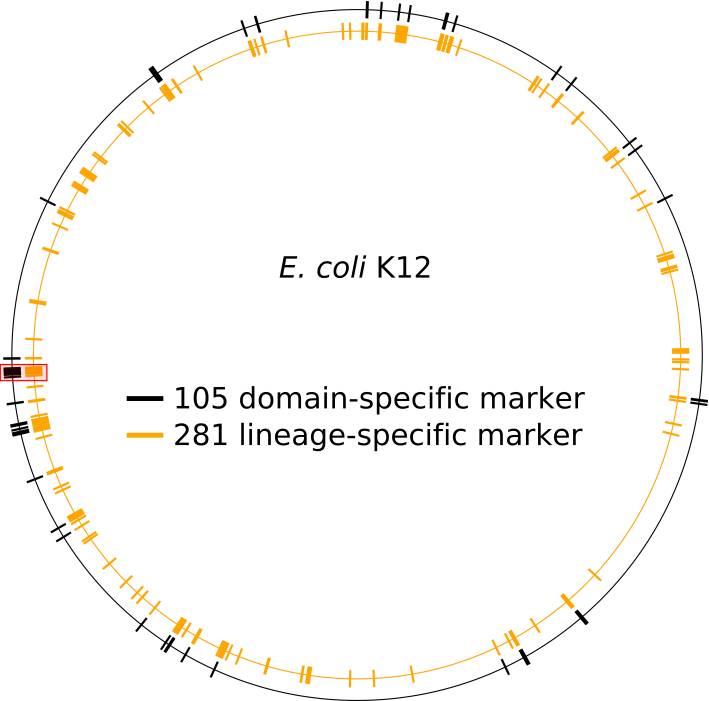
Ivanova NN, Schwientek P, Tripp HJ, Rinke C, Pati A, Huntemann M, Visel A, Woyke T, Kyrpides NC, Rubin EM. 2014. Stop codon reassignments in the wild. *Science* **344**: 909-913.

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.

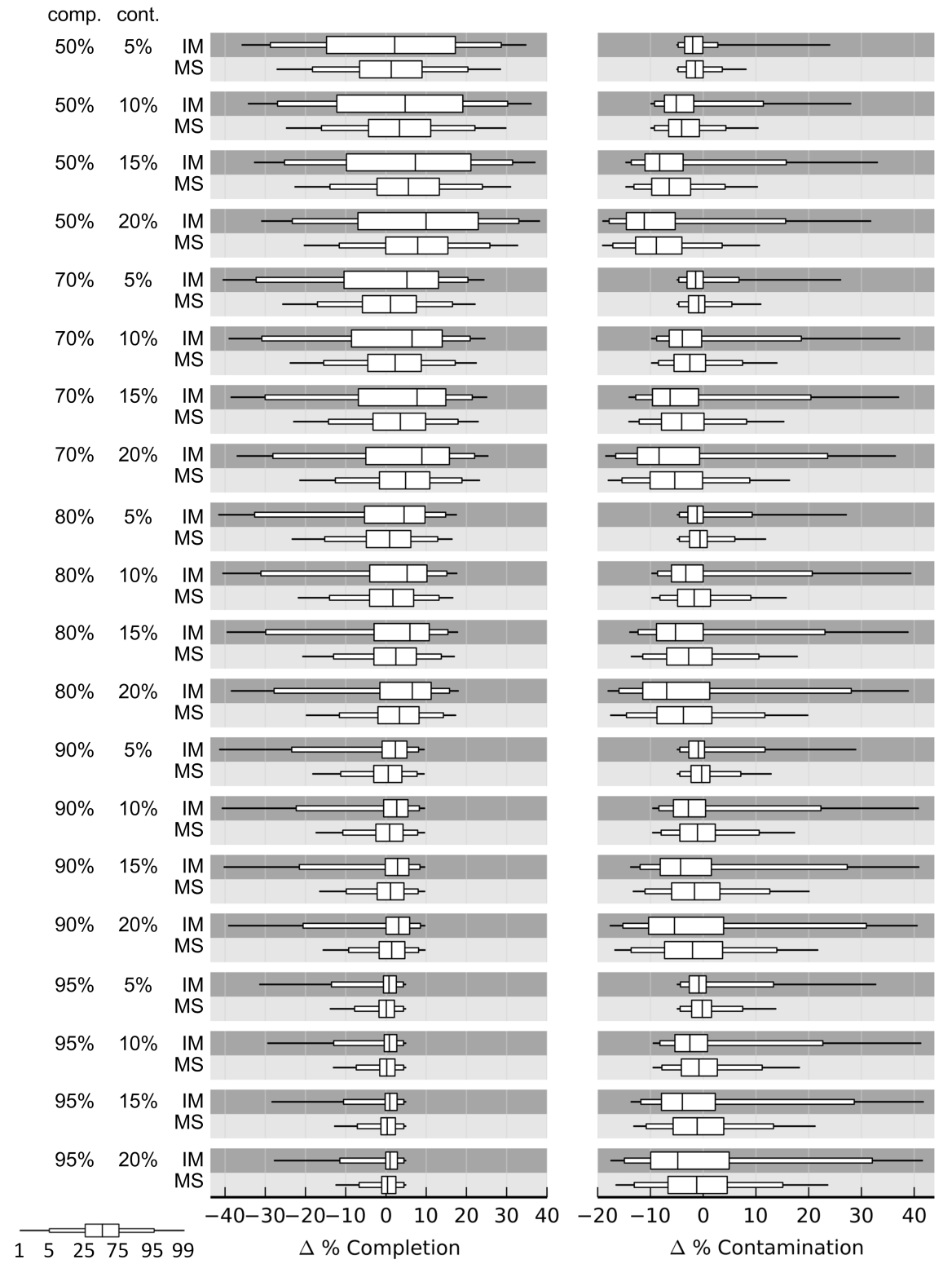
McCutcheon JP and Moran NA. 2011. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* **8**: 13-26.

[Yamao F](http://www.ncbi.nlm.nih.gov/pubmed?term=Yamao%20F%5BAuthor%5D&cauthor=true&cauthor_uid=3887399), [Muto A](http://www.ncbi.nlm.nih.gov/pubmed?term=Muto%20A%5BAuthor%5D&cauthor=true&cauthor_uid=3887399), [Kawauchi Y](http://www.ncbi.nlm.nih.gov/pubmed?term=Kawauchi%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=3887399), [Iwami M](http://www.ncbi.nlm.nih.gov/pubmed?term=Iwami%20M%5BAuthor%5D&cauthor=true&cauthor_uid=3887399), [Iwagami S](http://www.ncbi.nlm.nih.gov/pubmed?term=Iwagami%20S%5BAuthor%5D&cauthor=true&cauthor_uid=3887399), [Azumi Y](http://www.ncbi.nlm.nih.gov/pubmed?term=Azumi%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=3887399), [Osawa S](http://www.ncbi.nlm.nih.gov/pubmed?term=Osawa%20S%5BAuthor%5D&cauthor=true&cauthor_uid=3887399). 1985. UGA is read as tryptophan in Mycoplasma capricolum. *Proc Natl Acad Sci USA* **82**: 2306–2309.

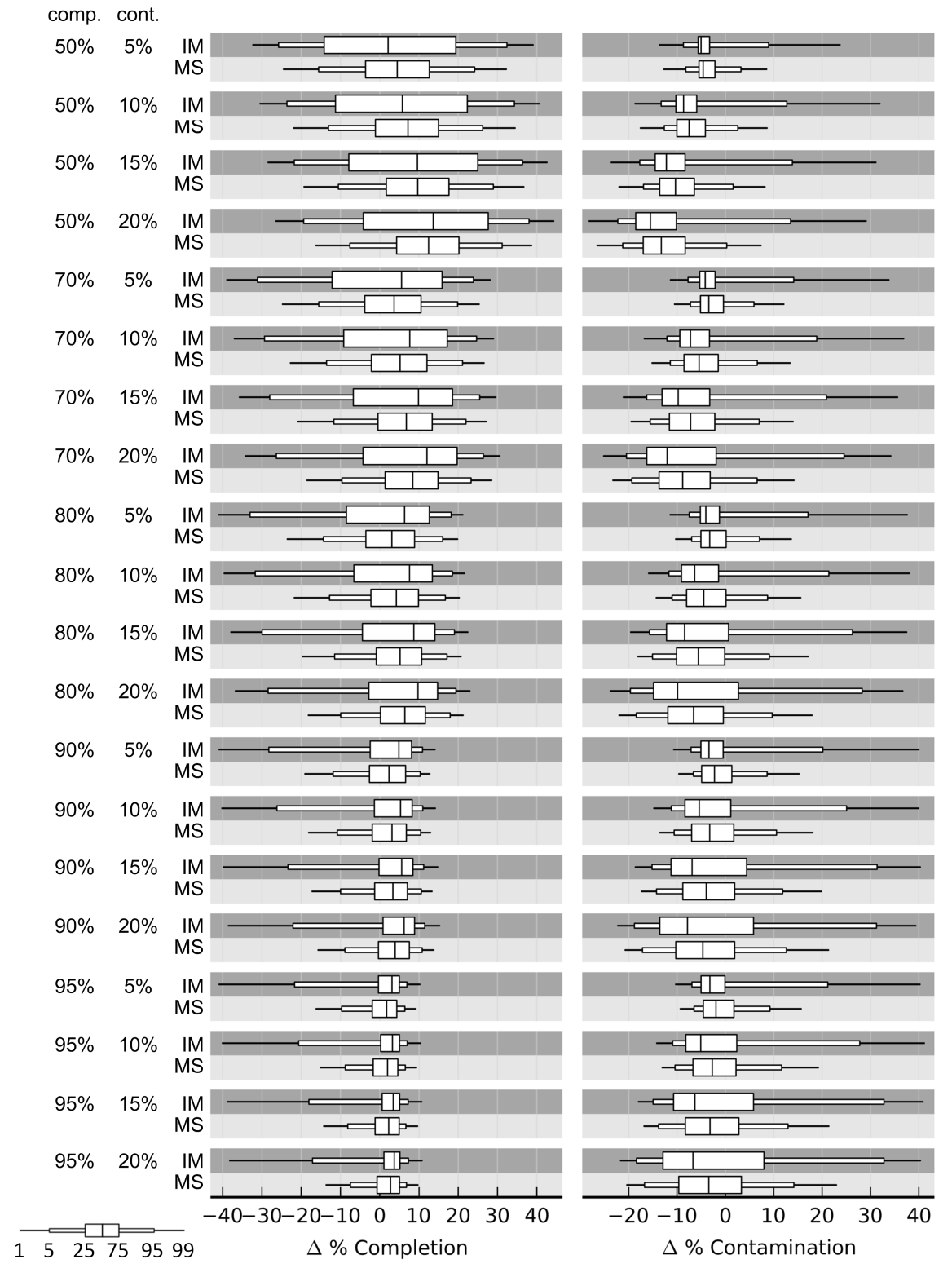
**Supplemental Figures**

****

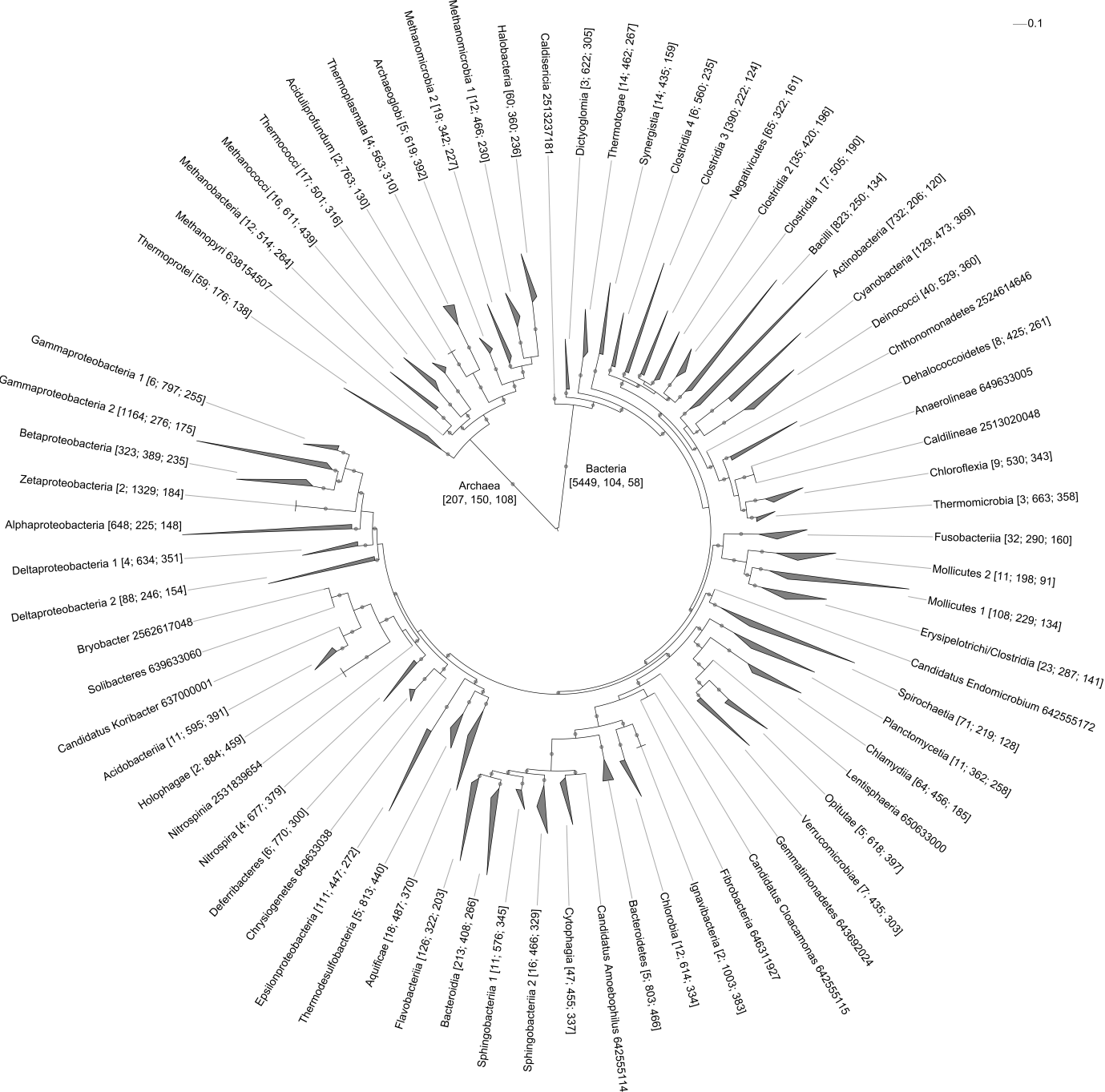
**Supplemental Figure S1**. Distribution of the 104 bacterial and 281 gammaproteobacterial marker genes around the *E. coli* K12 genome. Estimates of completeness and contamination are improved by the more uniform distribution of lineage-specific marker genes around a genome. Marker genes which are consistently collocated within a lineage (e.g., ribosomal operon) are organized into collocated marker sets. This results in 58 bacterial marker sets and 179 gammaproteobacterial marker sets. The red box highlights a set of 27 marker genes within the *E. coli* K12 genome that reside on a 14 kbp region on the genome. Since collocated marker genes do not provide independent information regarding the completeness or contamination of a genome, quality estimates are improved by basing quality estimates on the presence or absence of marker sets as opposed to individual marker genes.



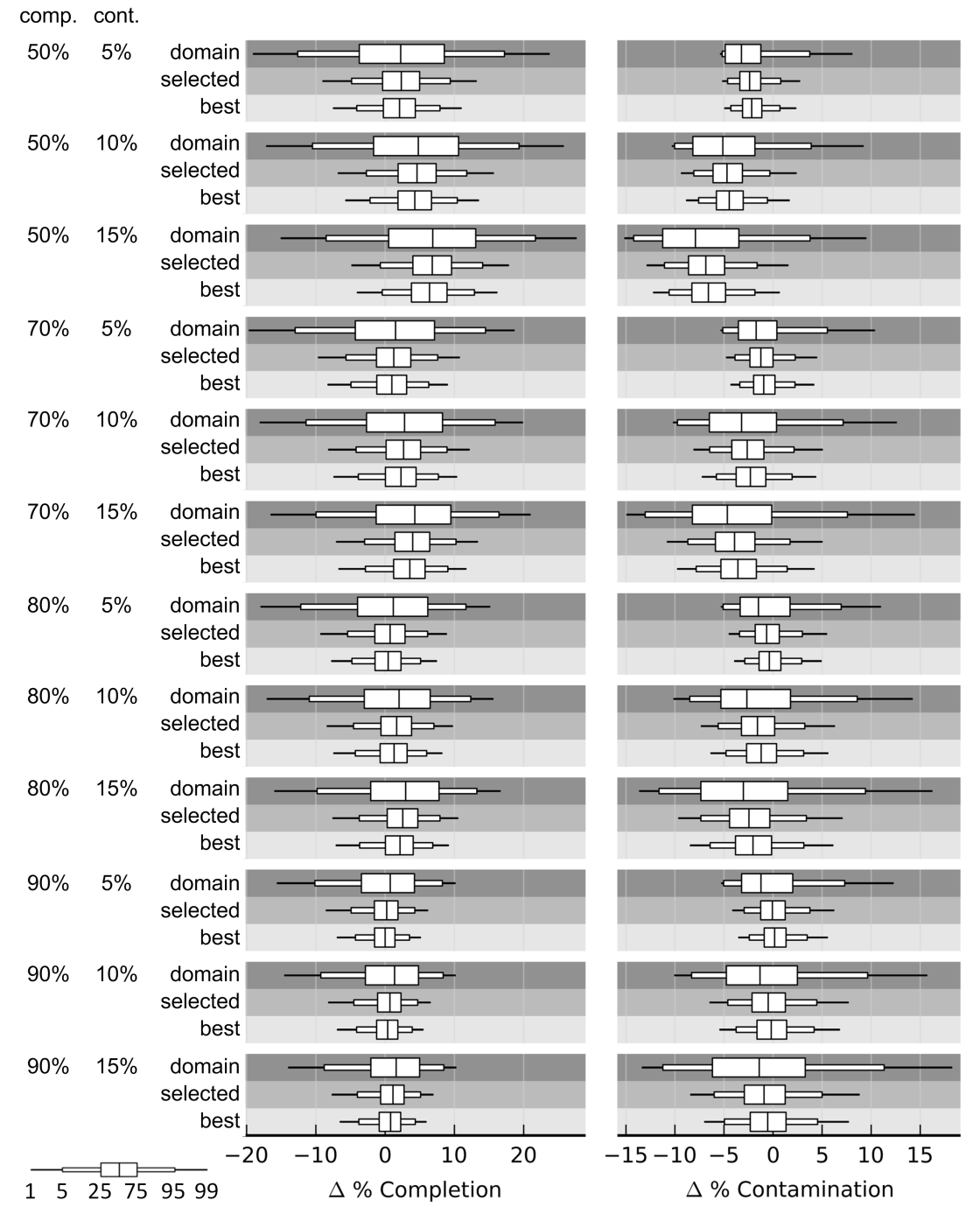
**Supplemental Figure S2**. Error in completeness and contamination estimates on simulated genomes with varying levels of completeness and contamination generated under the random contig model. Quality estimates were determined using domain-level marker genes treated as individual markers (IM) or organized into collocated marker sets (MS). Simulated genomes were generated from 2430 draft genomes. Each draft genome was used to generate 20 simulated genomes and the resulting test cases summarized using box-and-whisker plots.

****

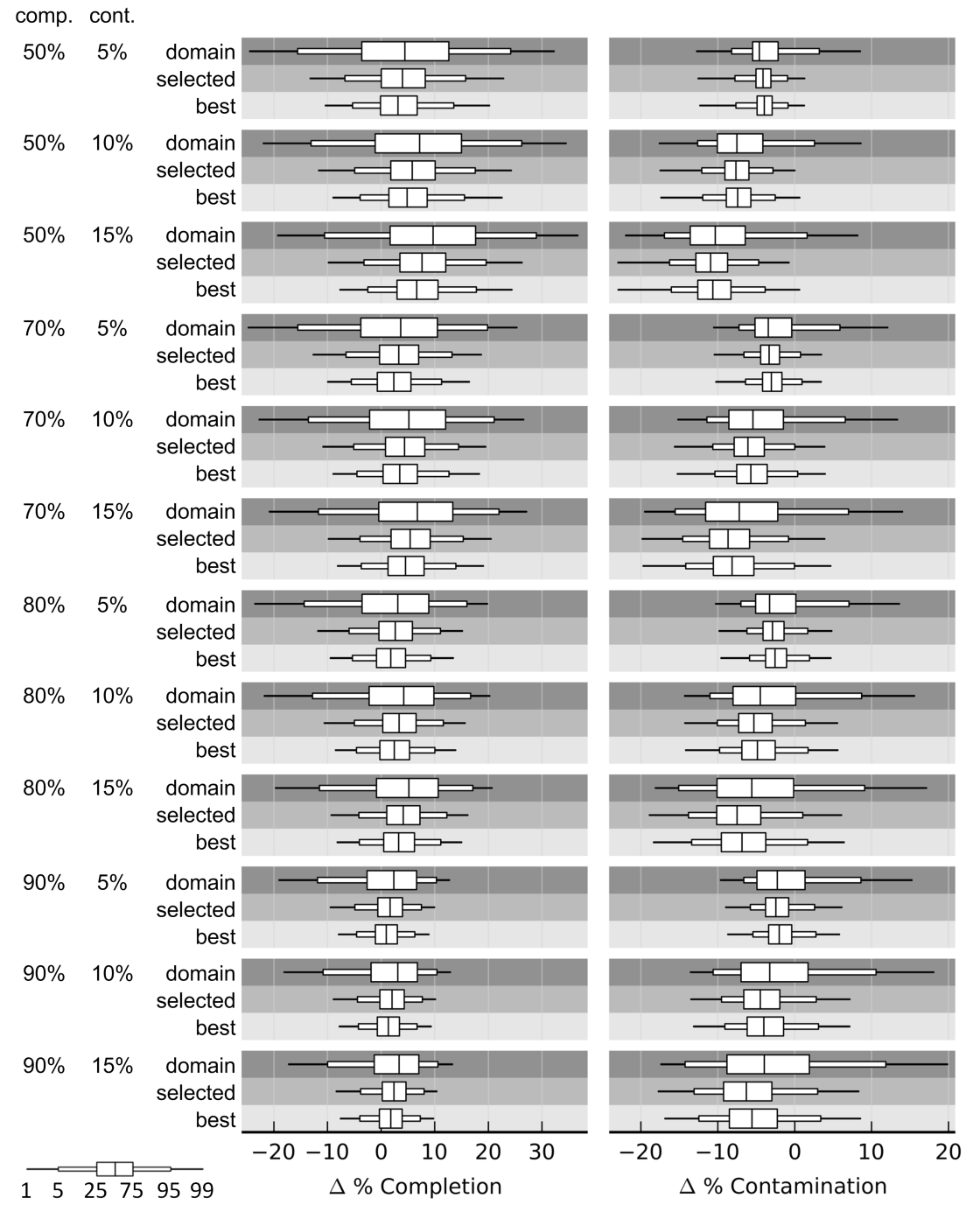
**Supplemental Figure S3**. Error in completeness and contamination estimates on simulated genomes with varying levels of completeness and contamination generated under the inverse length model. Quality estimates were determined using domain-level marker genes treated as individual markers (IM) or organized into collocated marker sets (MS). Simulated genomes were generated from 2430 draft genomes. Each draft genome was used to generate 20 simulated genomes and the resulting test cases summarized using box-and-whisker plots.

****

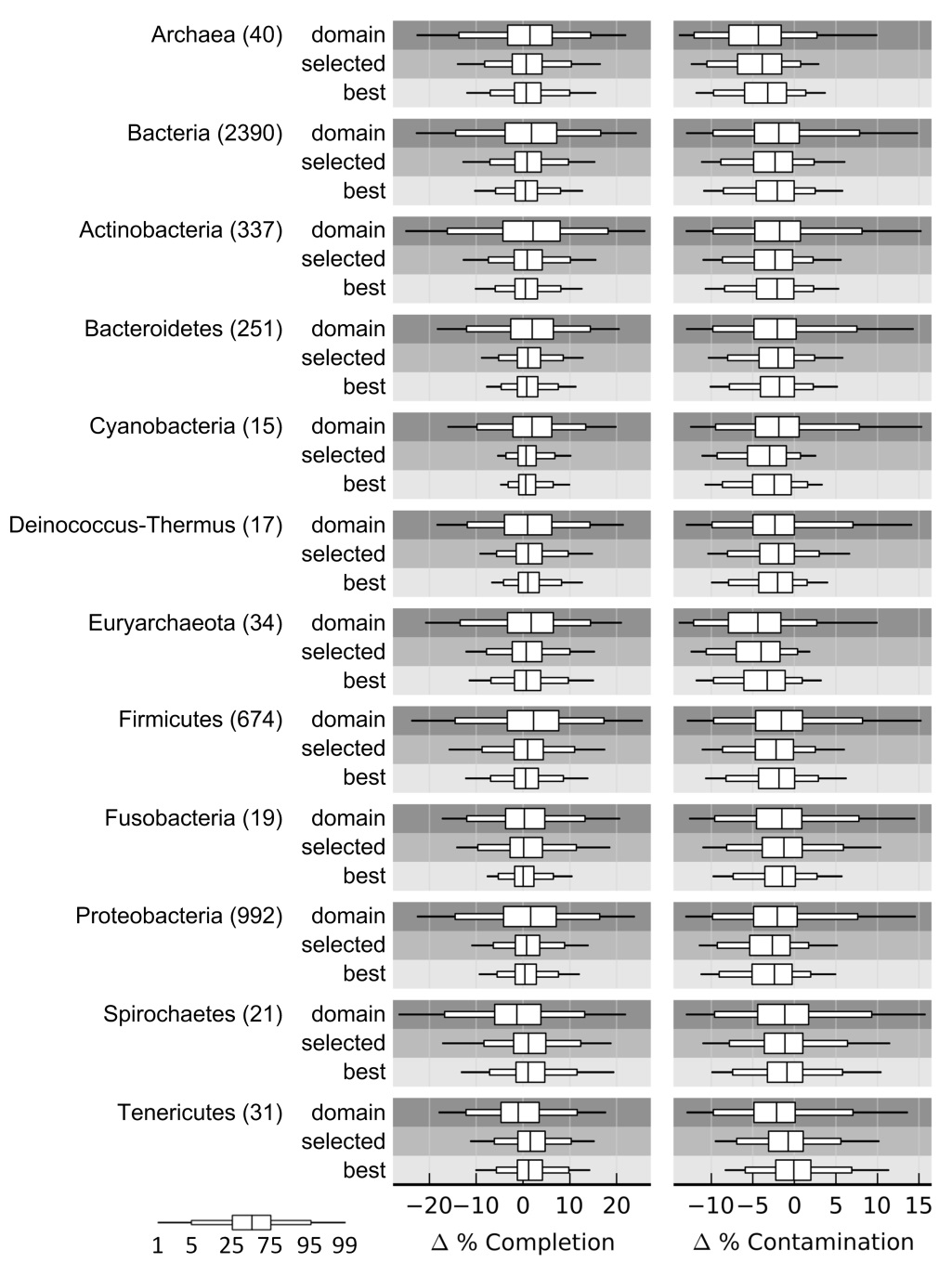
**Supplemental Figure S4**. Maximum-likelihood genome tree inferred from 5656 reference genomes. The tree is collapsed at the rank of class and the number of genomes in each lineage shown in brackets along with the number of identified marker genes and collocated marker sets. Classes represented by a single reference genome are followed by their IMG identifier. Paraphyletic classes are labelled by assigning each lineage a unique number (e.g., Clostridia clades 1 to 4). Grey circles indicate branches with a support value ≥80%.

****

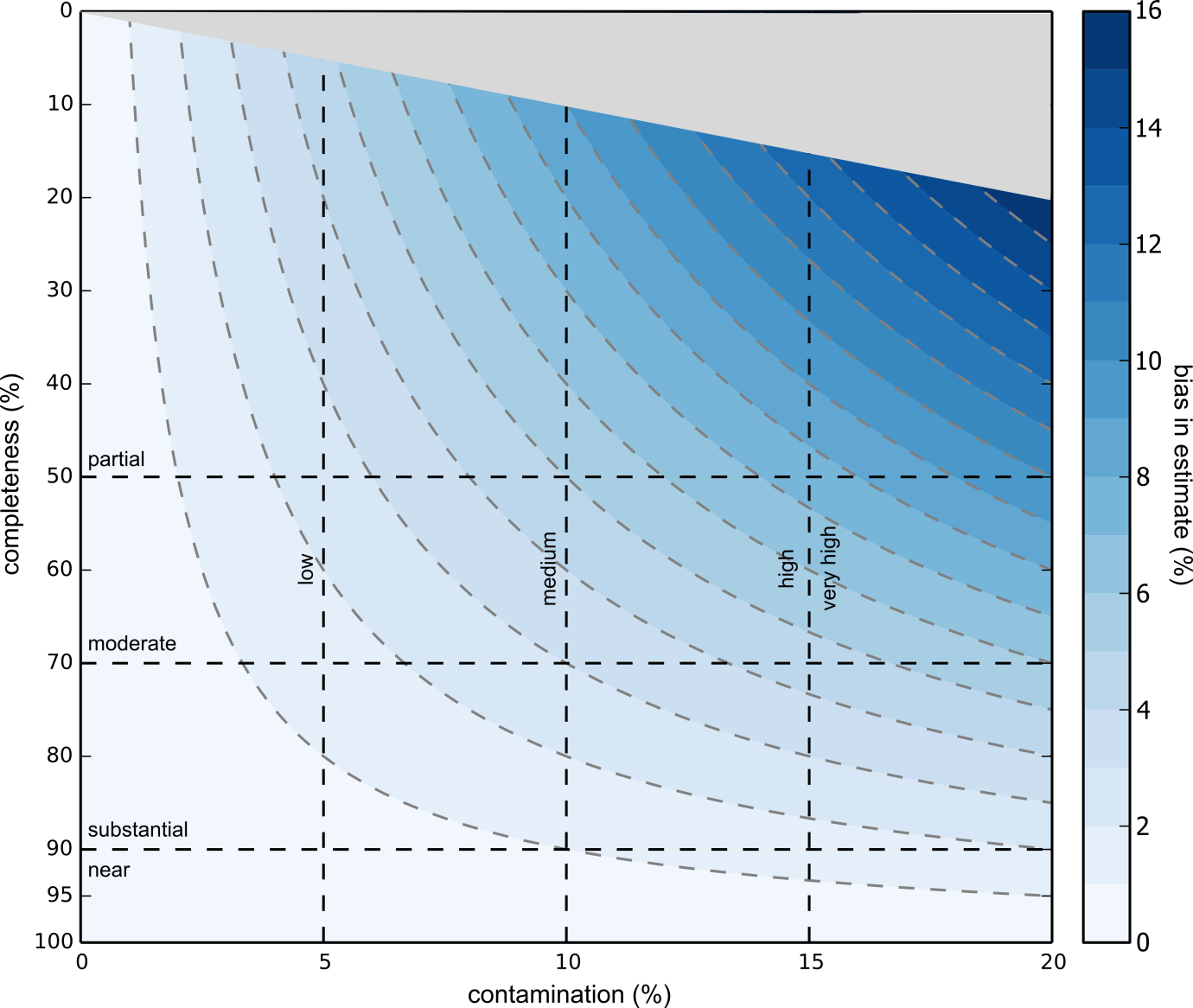
**Supplemental Figure S5**. Error in completeness and contamination estimates on simulated genomes with varying levels of completeness and contamination generated under the random fragment model using a window size of 20 kbp. Quality estimates were determined using i) domain: marker sets inferred across 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 from 3324 draft genomes. Each draft genome was used to generate 20 simulated genomes and the resulting test cases summarized using box-and-whisker plots.

****

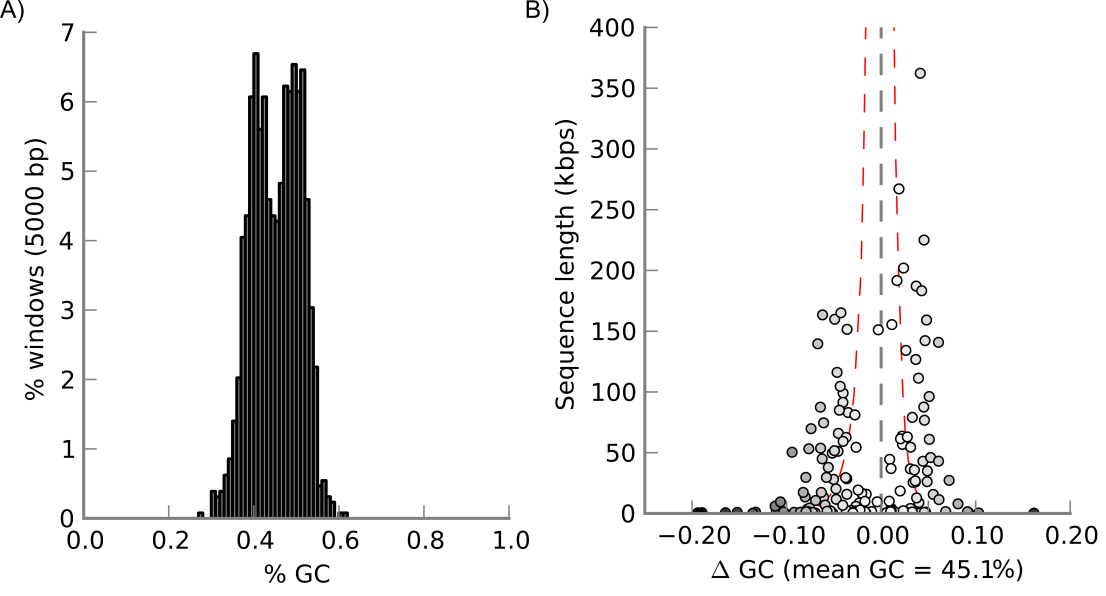
**Supplemental Figure S6**. Error in completeness and contamination estimates on simulated genomes with varying levels of completeness and contamination generated under the inverse length model. Quality estimates were determined using i) domain: marker sets inferred across all archaeal or bacterial genomes, 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 from 2430 draft genomes. Each draft genome was used to generate 20 simulated genomes and the resulting test cases summarized using box-and-whisker plots.

****

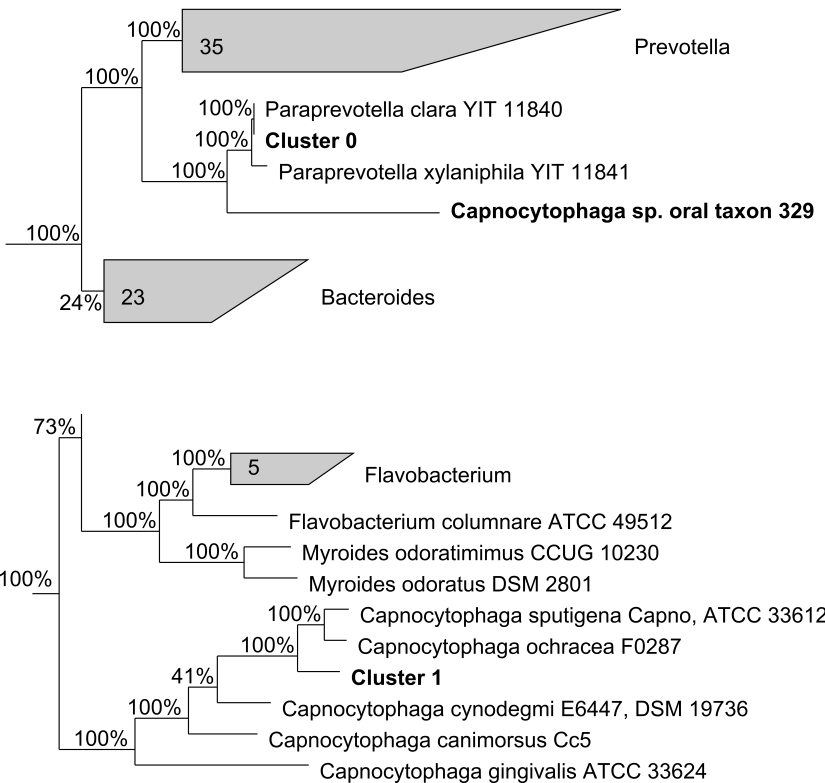
**Supplemental Figure S7**. Error in completeness and contamination estimates on simulated genomes from different phyla. Quality estimates were determined using i) domain: marker sets inferred across all archaeal or bacterial genomes, ii) selected: marker sets inferred from genomes within the lineage 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 from 2430 draft genomes, 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 and given for all phyla with ≥15 representative draft genomes. The number of representative draft genomes for each lineage is given in parentheses.

****

**Supplemental Figure S8**. Bias in completeness and contamination estimates when modelled as a binomial distribution. The bias approaches zero as genomes increase in completeness or decrease in contamination. Curved dashed lines indicate 1% increments in the bias. Straight dashed lines segment the plot into regions related to the proposed vocabulary for defining genome quality.

****

**Supplemental** **Figure S9**. GC-distribution plots of the HMP *Capnocytophaga* sp. oral taxon 329 genome. The GC-distribution of non-overlapping 5 kbp windows from the 157 contigs comprising the genome show a clear bimodal signal (A). Plotting contigs based on the deviation from the mean GC of the genome versus their length indicates two distinct clusters of contigs (B). The red dashed lines indicate the 1st and 99th percentiles of the expected deviation from the mean GC as determined over all finished IMG genomes. Plots produced with CheckM.

****

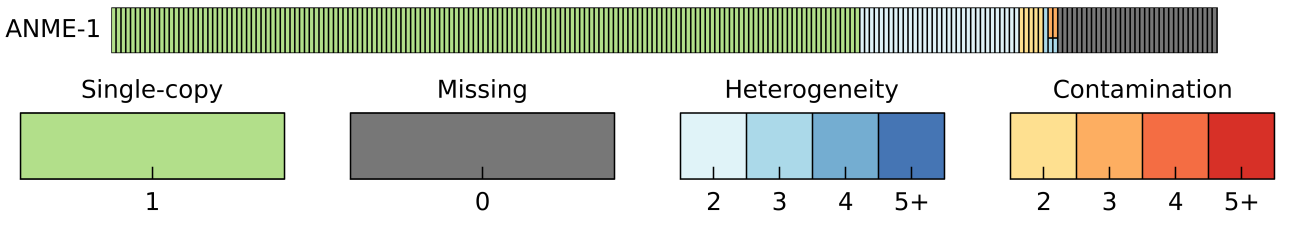
**Supplemental** **Figure S10**. Phylogenetic placement of the two genomes (Cluster 0 and Cluster 1) identified within the HMP *Capnocytophaga* sp. oral taxon 329 genome. The genome tree was inferred from the 86 bacterial marker genes described in Soo et al. (2014).



**Supplemental** **Figure S11**. Completeness estimates for 90 putative population genomes recovered from an acetate-amended aquifer. Estimates along the x-axis were determined using domain-level marker genes, while estimates along the y-axis were determined using lineage-specific marker sets. The best-fit regression line suggests that domain-level estimates which are >50% tend to overestimate the completeness of genomes by >20% relative to lineage-specific estimates.



**Supplemental** **Figure S12**. Contamination estimates for 90 putative population genomes recovered from an acetate-amended aquifer. Estimates along the x-axis were determined using domain-level marker genes, while estimates along the y-axis were determined using lineage-specific marker sets.

****

**Supplemental** **Figure S13**. Identification of the 213 marker genes within the Meyerdierks et al. (2010) ANME-1 genome. Each bar represents a marker gene. Bars in green represent markers identified exactly once, while bars in grey represent missing markers. Markers identified multiple times in the genome are represented by shades of blue or red depending on the AAI between pairs of multi-copy genes and the total number of copies present (2-5+). Pairs of multi-copy genes with an AAI ≥90% are indicated with shades of blue, while genes with less amino acid similarity are shown in red. A gene present 3 or more times may have pairs with an AAI ≥90% and pairs with an AAI < 90%. Plot produced with CheckM.

****

**Supplemental Figure S14**. Refining a marker set for lineage-specific gene loss and duplication. To evaluate a genome *G*, it is placed into a reference genome tree. The completeness and contamination of *G* is evaluated using the lineage-specific marker set defined at a predetermined parental node (see Fig. 3). To improve quality estimates for *G*, this marker set can be refined to account for gene loss and duplication in lineage A. The example illustrates an initial marker set consisting of 5 sets of marker genes. Marker genes *a* and *f* are deemed to be subject to gene loss in lineage *A* because they are absent in ≥ 50% of the genomes in lineage A. Similarly, marker gene *g* is deemed to subject to gene duplication because it is present multiple times in ≥ 50% of the genomes in lineage A. Consequently, marker genes *a*, *f*, and *g* are removed from the marker set.

**Supplemental Tables**

**Supplemental Table S1**. Mean absolute error of completeness (comp.) and contamination (cont.) estimates determined using different universal- and domain-specific marker gene sets. Simulated genomes were generated under the random fragment model from 3324 draft genomes with 20 genomes constructed per reference. Results are given for simulated genomes generated with a window size of 20 kbp over a range of completeness and contamination values.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Universal markers*** | |  | |  | |  | |  | |
|  | | **CheckM** | | **GEBA-MDM** | | **PhyloSift** | | **SpecI** | |
|  | | **(56 markers)** | | **(38 markers; 1e-10)** | | **(37 markers; 1e-10)** | | **(40 markers; 1e-10)** | |
| **comp. (%)** | **cont. (%)** | **comp.** | **cont.** | **comp.** | **cont.** | **comp.** | **cont.** | **comp.** | **cont.** |
| 70 | 5 | 17.8 | 5.7 | 18.6 | 5.0 | 21.2 | 5.9 | 25.7 | 3.6 |
| 70 | 10 | 18.3 | 9.9 | 18.9 | 8.9 | 21.5 | 10.3 | 24.8 | 7.0 |
| 70 | 15 | 18.1 | 13.6 | 19.2 | 12.7 | 21.7 | 14.5 | 24.1 | 10.3 |
| 80 | 5 | 15.3 | 5.7 | 14.8 | 4.9 | 16.7 | 5.8 | 29.9 | 3.5 |
| 80 | 10 | 15.0 | 9.7 | 14.5 | 8.5 | 16.3 | 10.0 | 29.1 | 6.6 |
| 80 | 15 | 15.3 | 13.4 | 14.8 | 12.2 | 16.7 | 14.1 | 28.9 | 9.6 |
| 90 | 5 | 9.6 | 5.8 | 8.9 | 5.1 | 10.0 | 6.0 | 34.1 | 3.5 |
| 90 | 10 | 9.5 | 10.3 | 8.8 | 9.1 | 9.8 | 10.8 | 33.9 | 6.5 |
| 90 | 15 | 9.8 | 13.6 | 9.0 | 12.2 | 10.2 | 14.4 | 34.0 | 9.4 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Bacterial markers*** | |  | |  | |  | |  | |
|  | | **CheckM** | | **Amphora 2** | | **Dupont** | | **GEBA-MDM** | |
|  | | **(104 markers)** | | **(76 markers; 1e-15)** | | **(107 markers)** | | **(139 markers)** | |
| **comp. (%)** | **cont. (%)** | **comp.** | **cont.** | **comp.** | **cont.** | **comp.** | **cont.** | **comp.** | **cont.** |
| 70 | 5 | 11.7 | 4.2 | 27.7 | 6.3 | 11.3 | 3.8 | 10.5 | 3.4 |
| 70 | 10 | 12.0 | 7.0 | 26.8 | 7.1 | 12.0 | 6.4 | 10.7 | 6.0 |
| 70 | 15 | 12.2 | 9.5 | 26.1 | 8.4 | 12.7 | 8.5 | 11.0 | 7.7 |
| 80 | 5 | 10.0 | 4.1 | 32.3 | 7.3 | 10.4 | 3.8 | 9.1 | 3.8 |
| 80 | 10 | 10.0 | 6.8 | 31.6 | 7.8 | 10.8 | 6.2 | 9.1 | 6.2 |
| 80 | 15 | 10.3 | 9.2 | 31.3 | 8.9 | 11.2 | 8.0 | 9.4 | 7.8 |
| 90 | 5 | 6.6 | 4.1 | 36.8 | 8.4 | 7.4 | 3.9 | 6.5 | 4.8 |
| 90 | 10 | 6.6 | 6.9 | 36.6 | 8.7 | 7.6 | 6.3 | 6.6 | 6.9 |
| 90 | 15 | 6.8 | 8.9 | 36.6 | 9.6 | 7.9 | 8.7 | 6.7 | 8.2 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***Archaeal markers*** | |  | |  | |  | |
|  | | **CheckM** | | **Amphora 2** | | **GEBA-MDM** | |
|  | | **(150 markers)** | | **(104 markers; 1e-10)** | | **(162 markers)** | |
| **comp. (%)** | **cont. (%)** | **comp.** | **cont.** | **comp.** | **cont.** | **comp.** | **cont.** |
| 70 | 5 | 7.7 | 3.4 | 7.3 | 2.8 | 7.3 | 2.9 |
| 70 | 10 | 8.0 | 5.3 | 7.2 | 4.6 | 7.6 | 5.2 |
| 70 | 15 | 8.4 | 6.9 | 7.6 | 5.9 | 7.9 | 6.7 |
| 80 | 5 | 6.6 | 3.4 | 6.3 | 3.0 | 6.1 | 3.4 |
| 80 | 10 | 7.1 | 5.0 | 6.5 | 4.2 | 6.5 | 4.8 |
| 80 | 15 | 6.9 | 6.5 | 6.0 | 5.7 | 6.5 | 6.1 |
| 90 | 5 | 4.9 | 3.3 | 4.9 | 3.2 | 4.6 | 4.1 |
| 90 | 10 | 4.9 | 5.2 | 4.5 | 4.8 | 4.4 | 5.3 |
| 90 | 15 | 4.8 | 6.3 | 4.4 | 5.7 | 4.4 | 5.9 |

**Supplemental Table S2**. Number of marker genes and marker sets for taxonomic groups with ≥ 20 reference genomes.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Taxon** | **# reference genomes** | **# marker genes** | **# marker sets** | **Marker sets with**  **≥ 2 marker genes** |
| **Domain** | Archaea | 207 | 150 | 108 | 28% |
|  | Bacteria | 5449 | 104 | 58 | 44% |
|  |  |  |  | *Average*: | 36% |
|  |  |  |  |  |  |
| **Phylum** | Chloroflexi | 20 | 225 | 149 | 34% |
|  | Fusobacteria | 32 | 290 | 160 | 45% |
|  | Deinococcus-Thermus | 40 | 529 | 360 | 32% |
|  | Crenarchaeota | 54 | 219 | 170 | 22% |
|  | Chlamydiae | 64 | 456 | 185 | 59% |
|  | Spirochaetes | 71 | 219 | 128 | 42% |
|  | Tenericutes | 119 | 178 | 106 | 40% |
|  | Cyanobacteria | 129 | 473 | 369 | 22% |
|  | Euryarchaeota | 146 | 189 | 126 | 33% |
|  | Bacteroidetes | 419 | 287 | 196 | 32% |
|  | Actinobacteria | 731 | 204 | 119 | 42% |
|  | Firmicutes | 1349 | 172 | 99 | 42% |
|  | Proteobacteria | 2343 | 183 | 120 | 34% |
|  |  |  |  | *Average*: | 37% |
|  |  |  |  |  |  |
| **Class** | Oscillatoriales | 25 | 546 | 416 | 24% |
|  | Sphingobacteriia | 27 | 335 | 234 | 30% |
|  | Methanomicrobia | 29 | 250 | 166 | 34% |
|  | Fusobacteriia | 32 | 290 | 160 | 45% |
|  | Deinococci | 40 | 529 | 360 | 32% |
|  | Cytophagia | 47 | 444 | 333 | 25% |
|  | Thermoprotei | 54 | 219 | 170 | 22% |
|  | Chroococcales | 55 | 491 | 379 | 23% |
|  | Halobacteria | 59 | 368 | 242 | 34% |
|  | Chlamydiia | 64 | 456 | 185 | 59% |
|  | Negativicutes | 64 | 335 | 168 | 50% |
|  | Spirochaetia | 71 | 219 | 128 | 42% |
|  | Deltaproteobacteria | 93 | 198 | 126 | 36% |
|  | Epsilonproteobacteria | 111 | 447 | 272 | 39% |
|  | Mollicutes | 119 | 178 | 106 | 40% |
|  | Flavobacteriia | 126 | 322 | 203 | 37% |
|  | Bacteroidia | 211 | 403 | 267 | 34% |
|  | Betaproteobacteria | 322 | 389 | 235 | 40% |
|  | Clostridia | 446 | 196 | 110 | 44% |
|  | Alphaproteobacteria | 648 | 225 | 148 | 34% |
|  | Actinobacteria | 729 | 205 | 119 | 42% |
|  | Bacilli | 821 | 250 | 136 | 46% |
|  | Gammaproteobacteria | 1167 | 281 | 179 | 36% |
|  |  |  |  | *Average*: | 37% |

**Supplemental Table S3**. Mean absolute error of completeness (comp.) and contamination (cont.) estimates determined using domain-specific marker genes treated individually (IM) or organized into collocated marker sets (MS). Simulated genomes were generated under the random fragment model from 3324 draft genomes with 20 genomes constructed per reference. Results are given for simulated genomes generated with a window size of 5, 20, and 50 kbp.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **5 kb windows** | | | | **20 kbp windows** | | | | **50 kbp windows** | | | |
| **Comp (%)** | **Cont (%)** | **IM comp.** | **MS comp.** | **IM cont.** | **MS cont.** | **IM comp.** | **MS comp.** | **IM cont.** | **MS cont.** | **IM comp.** | **MS comp.** | **IM cont.** | **MS cont.** |
| 50 | 0 | 8.3 | 6.2 | 0.1 | 0.1 | 12.4 | 7.3 | 0.1 | 0.1 | 14.8 | 8.1 | 0.1 | 0.1 |
| 50 | 5 | 8.5 | 6.4 | 3.6 | 3.0 | 12.5 | 7.5 | 4.2 | 3.4 | 15.0 | 8.4 | 4.5 | 3.8 |
| 50 | 10 | 9.1 | 7.2 | 6.0 | 5.2 | 13 | 8.2 | 7.4 | 5.6 | 15.3 | 9.1 | 7.9 | 6.1 |
| 50 | 15 | 10.0 | 8.4 | 8.1 | 7.3 | 13.6 | 9.3 | 10.0 | 7.7 | 15.8 | 10.0 | 10.8 | 8.1 |
| 50 | 20 | 11.1 | 9.8 | 10.2 | 9.3 | 14.3 | 10.5 | 12.4 | 9.7 | 16.5 | 11.2 | 13.4 | 10.1 |
| 70 | 0 | 7.6 | 5.7 | 0.2 | 0.3 | 11.4 | 6.7 | 0.2 | 0.3 | 13.3 | 7.5 | 0.2 | 0.3 |
| 70 | 5 | 7.6 | 5.7 | 3.4 | 2.7 | 11.4 | 6.8 | 4.1 | 3.1 | 13.4 | 7.6 | 4.4 | 3.5 |
| 70 | 10 | 7.8 | 6.0 | 5.3 | 4.2 | 11.6 | 7.0 | 6.9 | 4.8 | 13.5 | 7.8 | 7.5 | 5.3 |
| 70 | 15 | 8.1 | 6.4 | 6.8 | 5.5 | 11.8 | 7.3 | 9.1 | 6.2 | 13.7 | 8.2 | 10.0 | 6.8 |
| 70 | 20 | 8.6 | 7.0 | 8.2 | 6.8 | 12.1 | 7.9 | 11.0 | 7.5 | 13.9 | 8.6 | 12.2 | 8.1 |
| 80 | 0 | 6.6 | 5.0 | 0.3 | 0.3 | 9.8 | 5.9 | 0.3 | 0.3 | 11.2 | 6.6 | 0.3 | 0.3 |
| 80 | 5 | 6.6 | 4.9 | 3.4 | 2.7 | 9.8 | 5.9 | 4.1 | 3.1 | 11.2 | 6.6 | 4.4 | 3.5 |
| 80 | 10 | 6.7 | 5.0 | 5.1 | 3.9 | 9.8 | 5.9 | 6.7 | 4.6 | 11.3 | 6.7 | 7.3 | 5.1 |
| 80 | 15 | 6.8 | 5.2 | 6.5 | 5.0 | 9.9 | 6.1 | 8.9 | 5.8 | 11.4 | 6.9 | 9.8 | 6.4 |
| 80 | 20 | 7.0 | 5.5 | 7.6 | 6.0 | 10 | 6.3 | 10.6 | 6.8 | 11.5 | 7.1 | 11.9 | 7.5 |
| 90 | 0 | 4.9 | 3.8 | 0.4 | 0.4 | 6.7 | 4.4 | 0.4 | 0.4 | 7.6 | 5.0 | 0.4 | 0.4 |
| 90 | 5 | 4.9 | 3.7 | 3.4 | 2.7 | 6.6 | 4.4 | 4.1 | 3.1 | 7.6 | 5.0 | 4.4 | 3.5 |
| 90 | 10 | 4.9 | 3.7 | 5.0 | 3.9 | 6.6 | 4.4 | 6.7 | 4.5 | 7.5 | 5.0 | 7.4 | 5.0 |
| 90 | 15 | 4.9 | 3.7 | 6.3 | 4.8 | 6.6 | 4.4 | 8.7 | 5.6 | 7.6 | 5.0 | 9.6 | 6.2 |
| 90 | 20 | 4.9 | 3.8 | 7.3 | 5.6 | 6.6 | 4.5 | 10.4 | 6.5 | 7.6 | 5.1 | 11.6 | 7.2 |
| 95 | 0 | 3.4 | 2.8 | 0.4 | 0.4 | 4.3 | 3.3 | 0.4 | 0.4 | 5.0 | 3.8 | 0.4 | 0.4 |
| 95 | 5 | 3.4 | 2.8 | 3.4 | 2.7 | 4.3 | 3.3 | 4.1 | 3.2 | 5.0 | 3.7 | 4.5 | 3.5 |
| 95 | 10 | 3.4 | 2.7 | 5.1 | 3.9 | 4.2 | 3.2 | 6.7 | 4.5 | 5.0 | 3.7 | 7.4 | 5.0 |
| 95 | 15 | 3.3 | 2.7 | 6.2 | 4.8 | 4.2 | 3.2 | 8.7 | 5.6 | 4.9 | 3.7 | 9.6 | 6.2 |
| 95 | 20 | 3.3 | 2.7 | 7.3 | 5.5 | 4.2 | 3.2 | 10.3 | 6.4 | 4.9 | 3.7 | 11.6 | 7.1 |
| 100 | 0 | 0.6 | 0.7 | 0.4 | 0.5 | 0.8 | 0.9 | 0.4 | 0.5 | 1.4 | 1.3 | 0.4 | 0.5 |
| 100 | 5 | 0.6 | 0.7 | 3.5 | 2.8 | 0.8 | 0.9 | 4.2 | 3.2 | 1.4 | 1.3 | 4.5 | 3.5 |
| 100 | 10 | 0.6 | 0.7 | 5.1 | 3.9 | 0.8 | 0.9 | 6.7 | 4.6 | 1.4 | 1.3 | 7.3 | 5.1 |
| 100 | 15 | 0.6 | 0.7 | 6.3 | 4.8 | 0.8 | 0.9 | 8.7 | 5.6 | 1.4 | 1.3 | 9.7 | 6.2 |
| 100 | 20 | 0.6 | 0.7 | 7.3 | 5.6 | 0.8 | 0.9 | 10.3 | 6.5 | 1.4 | 1.3 | 11.5 | 7.1 |
|  | *Average* | 5.5 | 4.3 | 4.7 | 3.8 | 7.7 | 5.0 | 6.2 | 4.3 | 9.0 | 5.7 | 6.8 | 4.7 |

**Supplemental Table S4**. Mean absolute error and standard deviation of completeness (comp.) and contamination (cont.) estimates determined using domain-specific marker genes treated individually (IM) or organized into collocated marker sets (MS). Simulated genomes were generated under the random contig model from 2430 draft genomes with 20 genomes constructed per reference.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Comp (%)** | **Cont (%)** | **IM comp.** | **MS comp.** | **IM cont.** | **MS cont.** |
| 50 | 0 | 16.2 ± 9.55 | 9.4 ± 7.19 | 0.1 ± 0.45 | 0.2 ± 0.54 |
| 50 | 5 | 16.1 ± 9.48 | 9.3 ± 7.16 | 2.8 ± 4.01 | 2.3 ± 1.95 |
| 50 | 10 | 16.2 ± 9.51 | 9.6 ± 7.29 | 6.1 ± 5.01 | 4.6 ± 2.99 |
| 50 | 15 | 16.5 ± 9.69 | 10.2 ± 7.55 | 9.1 ± 5.46 | 6.8 ± 4.03 |
| 50 | 20 | 17.0 ± 9.90 | 11.1 ± 7.92 | 11.8 ± 5.86 | 9.0 ± 5.05 |
| 70 | 0 | 13.8 ± 9.56 | 8.2 ± 6.39 | 0.3 ± 0.61 | 0.3 ± 0.72 |
| 70 | 5 | 13.7 ± 9.23 | 8.1 ± 6.21 | 3.0 ± 4.74 | 2.3 ± 2.23 |
| 70 | 10 | 13.7 ± 8.91 | 8.1 ± 6.11 | 6.0 ± 6.15 | 4.3 ± 3.18 |
| 70 | 15 | 13.7 ± 8.50 | 8.3 ± 6.04 | 8.7 ± 6.77 | 5.9 ± 4.05 |
| 70 | 20 | 13.8 ± 8.20 | 8.6 ± 6.08 | 10.9 ± 6.97 | 7.3 ± 4.85 |
| 80 | 0 | 10.9 ± 9.30 | 6.8 ± 5.43 | 0.3 ± 0.69 | 0.4 ± 0.81 |
| 80 | 5 | 10.8 ± 8.95 | 6.8 ± 5.27 | 3.0 ± 5.09 | 2.4 ± 2.38 |
| 80 | 10 | 10.7 ± 8.50 | 6.7 ± 5.06 | 6.1 ± 6.82 | 4.2 ± 3.38 |
| 80 | 15 | 10.7 ± 8.17 | 6.7 ± 4.95 | 8.7 ± 7.42 | 5.7 ± 4.21 |
| 80 | 20 | 10.7 ± 7.79 | 6.8 ± 4.87 | 10.8 ± 7.77 | 6.9 ± 4.90 |
| 90 | 0 | 6.4 ± 7.99 | 4.4 ± 3.94 | 0.4 ± 0.78 | 0.5 ± 0.91 |
| 90 | 5 | 6.3 ± 7.81 | 4.4 ± 3.87 | 3.1 ± 5.54 | 2.5 ± 2.60 |
| 90 | 10 | 6.2 ± 7.38 | 4.3 ± 3.72 | 6.2 ± 7.41 | 4.3 ± 3.66 |
| 90 | 15 | 6.1 ± 7.18 | 4.3 ± 3.57 | 8.7 ± 8.14 | 5.7 ± 4.46 |
| 90 | 20 | 6.0 ± 6.79 | 4.2 ± 3.40 | 10.9 ± 8.65 | 6.8 ± 5.18 |
| 95 | 0 | 3.3 ± 5.90 | 2.6 ± 2.81 | 0.4 ± 0.81 | 0.5 ± 0.95 |
| 95 | 5 | 3.3 ± 5.87 | 2.6 ± 2.77 | 3.2 ± 5.62 | 2.5 ± 2.66 |
| 95 | 10 | 3.2 ± 5.56 | 2.5 ± 2.65 | 6.3 ± 7.73 | 4.4 ± 3.81 |
| 95 | 15 | 3.1 ± 5.36 | 2.5 ± 2.53 | 8.8 ± 8.52 | 5.8 ± 4.64 |
| 95 | 20 | 3.1 ± 5.26 | 2.4 ± 2.43 | 10.9 ± 8.93 | 6.8 ± 5.27 |
| 100 | 0 | 0.5 ± 1.04 | 0.6 ± 1.25 | 0.5 ± 0.84 | 0.5 ± 0.99 |
| 100 | 5 | 0.5 ± 1.02 | 0.6 ± 1.22 | 3.3 ± 5.84 | 2.6 ± 2.77 |
| 100 | 10 | 0.5 ± 1.00 | 0.6 ± 1.19 | 6.3 ± 7.74 | 4.5 ± 3.92 |
| 100 | 15 | 0.5 ± 0.97 | 0.5 ± 1.15 | 8.8 ± 8.71 | 5.8 ± 4.74 |
| 100 | 20 | 0.4 ± 0.92 | 0.5 ± 1.11 | 10.9 ± 9.24 | 6.9 ± 5.45 |
|  | *Average* | 8.5 ± 9.37 | 5.4 ± 5.85 | 5.9 ± 7.35 | 4.1 ± 4.37 |

**Supplemental Table S5**. Mean absolute error and standard deviation of completeness (comp.) and contamination (cont.) estimates determined using domain-specific marker genes treated individually (IM) or organized into collocated marker sets (MS). Simulated genomes were generated under the inverse length model from 2430 draft genomes with 20 genomes constructed per reference.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Comp (%)** | **Cont (%)** | **IM comp.** | **MS comp.** | **IM cont.** | **MS cont.** |
| 50 | 0 | 16.7 ± 9.45 | 9.7 ± 7.47 | 0.3 ± 1.08 | 0.3 ± 0.75 |
| 50 | 5 | 16.6 ± 9.78 | 10.3 ± 7.72 | 5.7 ± 4.37 | 4.6 ± 2.68 |
| 50 | 10 | 17.0 ± 10.25 | 11.2 ± 8.17 | 9.4 ± 4.84 | 7.4 ± 3.84 |
| 50 | 15 | 17.8 ± 10.88 | 12.4 ± 8.83 | 12.5 ± 5.56 | 10.1 ± 5.19 |
| 50 | 20 | 18.8 ± 11.58 | 14.1 ± 9.39 | 15.3 ± 6.33 | 12.8 ± 6.17 |
| 70 | 0 | 15.7 ± 9.83 | 8.9 ± 6.75 | 0.4 ± 1.24 | 0.4 ± 0.89 |
| 70 | 5 | 15.5 ± 9.25 | 9.1 ± 6.68 | 5.7 ± 5.39 | 4.2 ± 2.76 |
| 70 | 10 | 15.4 ± 8.86 | 9.3 ± 6.69 | 9.0 ± 6.10 | 6.3 ± 3.86 |
| 70 | 15 | 15.6 ± 8.66 | 9.9 ± 6.87 | 11.6 ± 6.53 | 8.1 ± 4.97 |
| 70 | 20 | 15.8 ± 8.46 | 10.6 ± 7.05 | 13.7 ± 7.11 | 9.8 ± 6.00 |
| 80 | 0 | 13.4 ± 9.71 | 7.7 ± 5.82 | 0.5 ± 1.43 | 0.5 ± 0.97 |
| 80 | 5 | 13.2 ± 9.04 | 7.8 ± 5.65 | 5.8 ± 6.10 | 4.1 ± 2.91 |
| 80 | 10 | 13.1 ± 8.57 | 7.9 ± 5.62 | 9.0 ± 6.85 | 6.0 ± 3.93 |
| 80 | 15 | 13.1 ± 8.08 | 8.2 ± 5.60 | 11.4 ± 7.25 | 7.5 ± 4.90 |
| 80 | 20 | 13.2 ± 7.70 | 8.5 ± 5.66 | 13.2 ± 7.64 | 8.7 ± 5.74 |
| 90 | 0 | 9.5 ± 8.88 | 5.8 ± 4.49 | 0.5 ± 1.54 | 0.6 ± 1.06 |
| 90 | 5 | 9.3 ± 8.28 | 5.7 ± 4.24 | 5.9 ± 6.71 | 4.1 ± 3.13 |
| 90 | 10 | 9.0 ± 7.68 | 5.7 ± 4.07 | 9.1 ± 7.70 | 5.8 ± 4.11 |
| 90 | 15 | 9.0 ± 7.35 | 5.7 ± 3.97 | 11.5 ± 8.23 | 7.1 ± 4.97 |
| 90 | 20 | 8.9 ± 6.97 | 5.8 ± 3.91 | 13.2 ± 8.50 | 8.1 ± 5.72 |
| 95 | 0 | 6.2 ± 7.56 | 4.2 ± 3.43 | 0.6 ± 1.61 | 0.6 ± 1.12 |
| 95 | 5 | 6.0 ± 7.05 | 4.2 ± 3.26 | 6.0 ± 7.03 | 4.1 ± 3.21 |
| 95 | 10 | 5.9 ± 6.79 | 4.1 ± 3.12 | 9.2 ± 8.08 | 5.8 ± 4.22 |
| 95 | 15 | 5.8 ± 6.39 | 4.1 ± 3.00 | 11.5 ± 8.71 | 7.0 ± 5.07 |
| 95 | 20 | 5.7 ± 6.18 | 4.1 ± 2.88 | 13.2 ± 8.98 | 8.0 ± 5.76 |
| 100 | 0 | 0.7 ± 1.87 | 0.8 ± 1.42 | 0.6 ± 1.72 | 0.7 ± 1.20 |
| 100 | 5 | 0.7 ± 1.73 | 0.7 ± 1.36 | 6.2 ± 7.43 | 4.2 ± 3.44 |
| 100 | 10 | 0.6 ± 1.71 | 0.7 ± 1.31 | 9.4 ± 8.70 | 5.9 ± 4.41 |
| 100 | 15 | 0.6 ± 1.62 | 0.6 ± 1.26 | 11.8 ± 9.28 | 7.0 ± 5.18 |
| 100 | 20 | 0.6 ± 1.55 | 0.6 ± 1.21 | 13.6 ± 9.77 | 8.0 ± 6.04 |
|  | *Average* | 10.3 ± 9.81 | 6.6 ± 6.54 | 8.2 ± 8.09 | 5.6 ± 5.26 |

**Supplemental Table S6**. Phylogenetically informative marker genes used to infer the reference genome tree   
along with matching PhyloSift genes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Pfam Id** | **Length** | **Description** | **PhyloSift Id** | **Description** |
| PF00164 | 122 | Ribosomal protein S12/S23 | DNGNGWU00026 | Ribosomal protein S12/S23 |
| PF00177 | 148 | Ribosomal protein S7p/S5e | DNGNGWU00017 | Ribosomal protein S7 |
| PF00181 | 77 | Ribosomal proteins L2, RNA binding domain | DNGNGWU00010 | Ribosomal protein L2 |
| PF00189 | 85 | Ribosomal protein S3, C-terminal domain | DNGNGWU00028 | Ribosomal protein S3 |
| PF00203 | 81 | Ribosomal protein S19 | DNGNGWU00016 | Ribosomal protein S19 |
| PF00237 | 105 | Ribosomal protein L22p/L17e | DNGNGWU00007 | Ribosomal protein L22 |
| PF00238 | 122 | Ribosomal protein L14p/L23e | DNGNGWU00014 | Ribosomal protein L14b/L23e |
| PF00252 | 133 | Ribosomal protein L16p/L10e | DNGNGWU00018 | Ribosomal protein L16/L10e |
| PF00276 | 92 | Ribosomal protein L23 | DNGNGWU00022 | Ribosomal protein L25/L23 |
| PF00281 | 56 | Ribosomal protein L5 | DNGNGWU00025 | Ribosomal protein L5 |
| PF00297 | 263 | Ribosomal protein L3 | DNGNGWU00012 | Ribosomal protein L3 |
| PF00298 | 69 | Ribosomal protein L11, RNA binding domain | DNGNGWU00024 | Ribosomal protein L11 |
| PF00312 | 83 | Ribosomal protein S15 | DNGNGWU00034 | Ribosomal protein S15p/S13e |
| PF00318 | 211 | Ribosomal protein S2 | DNGNGWU00001 | Ribosomal protein S2 |
| PF00333 | 67 | Ribosomal protein S5, N-terminal domain | DNGNGWU00015 | Ribosomal protein S5 |
| PF00366 | 69 | Ribosomal protein S17 | DNGNGWU00036 | Ribosomal protein S17 |
| PF00380 | 121 | Ribosomal protein S9/S16 | DNGNGWU00011 | Ribosomal protein S9 |
| PF00410 | 129 | Ribosomal protein S8 | DNGNGWU00031 | Ribosomal protein S8 |
| PF00411 | 110 | Ribosomal protein S11 | DNGNGWU00029 | Ribosomal protein S11 |
| PF00466 | 100 | Ribosomal protein L10 | DNGNGWU00030 | Ribosomal protein L10 |
| PF00562 | 386 | RNA polymerase Rpb2, domain 6 |  |  |
| PF00572 | 128 | Ribosomal protein L13 | DNGNGWU00037 | Ribosomal protein L13 |
| PF00573 | 192 | Ribosomal protein L4/L1 | DNGNGWU00009 | Ribosomal protein L4/L1e |
| PF00623 | 166 | RNA polymerase Rpb1, domain 2 |  |  |
| PF00673 | 95 | Ribosomal protein L5P, C-terminus | DNGNGWU00025 | Ribosomal protein L5 |
| PF00687 | 220 | Ribosomal protein L1p/L10e | DNGNGWU00003 | Ribosomal protein L1 |
| PF00831 | 58 | Ribosomal protein L29 | DNGNGWU00027 | Ribosomal protein L29 |
| PF00861 | 119 | Ribosomal protein L18p/L5e | DNGNGWU00033 | Ribosomal protein L18p/L5e |
| PF01192 | 57 | RNA polymerase Rpb6 |  |  |
| PF01509 | 149 | TruB family pseudouridylate synthase | DNGNGWU00032 | tRNA pseudouridine synthase B |
| PF02978 | 104 | Signal peptide binding domain |  |  |
| PF03719 | 74 | Ribosomal protein S5, C-terminal domain | DNGNGWU00015 | Ribosomal protein S5 |
| PF03946 | 60 | Ribosomal protein L11, N-terminal domain | DNGNGWU00024 | Ribosomal protein L11 |
| PF03947 | 130 | Ribosomal Proteins L2, C-terminal domain | DNGNGWU00010 | Ribosomal protein L2 |
| PF04560 | 82 | RNA polymerase Rpb2, domain 7 |  |  |
| PF04561 | 190 | RNA polymerase Rpb2, domain 2 |  |  |
| PF04563 | 203 | RNA polymerase beta subunit |  |  |
| PF04565 | 68 | RNA polymerase Rpb2, domain 3 |  |  |
| PF04997 | 337 | RNA polymerase Rpb1, domain 1 |  |  |
| PF05000 | 108 | RNA polymerase Rpb1, domain 4 |  |  |
| PF11987 | 109 | Translation-initiation factor 2 | DNGNGWU00005 | translation initiation factor IF-2 |
| TIGR00344 | 847 | Alanine – tRNA ligases |  |  |
| TIGR00422 | 863 | Valine – tRNA ligase |  |  |

**Supplemental Table S7**. Phylogenetically informative genes used in PhyloSift without a matching CheckM gene.

|  |  |  |  |
| --- | --- | --- | --- |
| PhyloSift Id | Description | Matching Pfam Id | Note (with regards to Pfam model) |
| DNGNGWU00002 | Ribosomal protein S10 | PF00338 | Identified as ubiquitous, but not single copy in archaea |
| DNGNGWU00006 | metalloendopeptidase | no clear match | No clear match to either a Pfam or TIGRFAMs model |
| DNGNGWU00013 | phenylalanyl-tRNA synthetase beta subunit | PF01409 | Identified as ubiquitous, but not single copy in archaea |
| DNGNGWU00019 | Ribosomal protein S13 | PF00416 | Identified as a multi-copy gene with divergent phylogenetic histories |
| DNGNGWU00020 | phenylalanyl-tRNA synthetase alpha subunit | PF01409 | Identified as ubiquitous, but not single copy in archaea |
| DNGNGWU00021 | Ribosomal protein L15 | PF00828 | Identified twice in all reference archaea |
| DNGNGWU00023 | Ribosomal protein L6 | PF00347 | Identified as incongruent by taxonomic congruency test |
| DNGNGWU00035 | Porphobilinogen deaminase | PF01379, PF03900 | Identified in 84.5% of archaea and 77.8% of bacteria |
| DNGNGWU00039 | ribonuclease HII | PF01351 | Identified as ubiquitous, but not single copy in bacteria |
| DNGNGWU00040 | Ribosomal protein L24 | no clear match | No clear matching Pfam model and split into domain-specific TIGRFAMs models |

**Supplemental Table S8**. Mean absolute error of completeness (comp.) and contamination (cont.) estimates determined using domain-specific marker sets (dms), the lineage-specific marker set selected by CheckM (sms), and the best performing lineage-specific marker set (bms). Simulated genomes were generated under the random fragment model from 3324 draft genomes with 20 genomes constructed per reference. Results are given for simulated genomes generated with a window size of 20 and 50 kbp.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **20 kbp** | | | | | | **50 kbp** | | | | | |
| **Comp (%)** | **Cont (%)** | **dms comp.** | **sms**  **comp.** | **bms comp.** | **dms cont.** | **sms cont.** | **bms cont.** | **dms comp.** | **sms**  **comp.** | **bms comp.** | **dms cont.** | **sms cont.** | **bms cont.** |
| 50 | 0 | 7.3 | 3.3 | 2.7 | 0.1 | 0.2 | 0.3 | 8.1 | 3.8 | 3.2 | 0.1 | 0.2 | 0.3 |
| 50 | 5 | 7.5 | 3.9 | 3.4 | 3.4 | 2.4 | 2.2 | 8.4 | 4.5 | 3.8 | 3.8 | 2.7 | 2.5 |
| 50 | 10 | 8.2 | 5.3 | 4.8 | 5.6 | 4.6 | 4.4 | 9.1 | 5.8 | 5.2 | 6.1 | 4.8 | 4.5 |
| 50 | 15 | 9.3 | 7.1 | 6.6 | 7.7 | 6.8 | 6.5 | 10 | 7.4 | 6.9 | 8.1 | 6.9 | 6.6 |
| 50 | 20 | 10.5 | 8.9 | 8.4 | 9.7 | 8.8 | 8.5 | 11.2 | 9.2 | 8.7 | 10.1 | 8.9 | 8.6 |
| 70 | 0 | 6.7 | 3.1 | 2.6 | 0.3 | 0.4 | 0.4 | 7.5 | 3.6 | 3.0 | 0.3 | 0.4 | 0.4 |
| 70 | 5 | 6.8 | 3.3 | 2.8 | 3.1 | 1.8 | 1.6 | 7.6 | 3.8 | 3.2 | 3.5 | 2.1 | 1.9 |
| 70 | 10 | 7.0 | 3.8 | 3.4 | 4.8 | 3.0 | 2.7 | 7.8 | 4.4 | 3.8 | 5.3 | 3.4 | 3.0 |
| 70 | 15 | 7.3 | 4.6 | 4.2 | 6.2 | 4.2 | 3.8 | 8.2 | 5.2 | 4.6 | 6.8 | 4.6 | 4.1 |
| 70 | 20 | 7.9 | 5.6 | 5.1 | 7.5 | 5.4 | 5.0 | 8.6 | 6.0 | 5.4 | 8.1 | 5.7 | 5.2 |
| 80 | 0 | 5.9 | 2.7 | 2.3 | 0.3 | 0.5 | 0.5 | 6.6 | 3.2 | 2.7 | 0.3 | 0.5 | 0.5 |
| 80 | 5 | 5.9 | 2.8 | 2.4 | 3.1 | 1.6 | 1.4 | 6.6 | 3.3 | 2.8 | 3.5 | 1.9 | 1.7 |
| 80 | 10 | 5.9 | 3.1 | 2.7 | 4.6 | 2.5 | 2.1 | 6.7 | 3.6 | 3.1 | 5.1 | 2.9 | 2.5 |
| 80 | 15 | 6.1 | 3.5 | 3.1 | 5.8 | 3.3 | 2.9 | 6.9 | 4.0 | 3.5 | 6.4 | 3.7 | 3.2 |
| 80 | 20 | 6.3 | 4.0 | 3.6 | 6.8 | 4.0 | 3.5 | 7.1 | 4.5 | 4.0 | 7.5 | 4.5 | 3.9 |
| 90 | 0 | 4.4 | 2.2 | 1.9 | 0.4 | 0.6 | 0.5 | 5.0 | 2.5 | 2.1 | 0.4 | 0.6 | 0.5 |
| 90 | 5 | 4.4 | 2.2 | 1.8 | 3.1 | 1.6 | 1.4 | 5.0 | 2.6 | 2.2 | 3.5 | 1.8 | 1.6 |
| 90 | 10 | 4.4 | 2.2 | 1.9 | 4.5 | 2.2 | 1.9 | 5.0 | 2.7 | 2.3 | 5.0 | 2.6 | 2.2 |
| 90 | 15 | 4.4 | 2.3 | 2.0 | 5.6 | 2.7 | 2.3 | 5.0 | 2.8 | 2.4 | 6.2 | 3.2 | 2.7 |
| 90 | 20 | 4.5 | 2.5 | 2.2 | 6.5 | 3.2 | 2.7 | 5.1 | 2.9 | 2.6 | 7.2 | 3.7 | 3.1 |
| 95 | 0 | 3.3 | 1.7 | 1.5 | 0.4 | 0.6 | 0.4 | 3.8 | 2.0 | 1.7 | 0.4 | 0.6 | 0.4 |
| 95 | 5 | 3.3 | 1.7 | 1.4 | 3.2 | 1.6 | 1.4 | 3.7 | 2.0 | 1.7 | 3.5 | 1.9 | 1.6 |
| 95 | 10 | 3.2 | 1.7 | 1.4 | 4.5 | 2.2 | 1.9 | 3.7 | 2.0 | 1.7 | 5.0 | 2.5 | 2.2 |
| 95 | 15 | 3.2 | 1.7 | 1.5 | 5.6 | 2.6 | 2.3 | 3.7 | 2.1 | 1.8 | 6.2 | 3.1 | 2.6 |
| 95 | 20 | 3.2 | 1.8 | 1.5 | 6.4 | 3.0 | 2.6 | 3.7 | 2.1 | 1.9 | 7.1 | 3.5 | 3.0 |
| 100 | 0 | 0.9 | 0.8 | 0.3 | 0.5 | 0.7 | 0.1 | 1.3 | 1.0 | 0.6 | 0.5 | 0.7 | 0.1 |
| 100 | 5 | 0.9 | 0.8 | 0.4 | 3.2 | 1.7 | 1.5 | 1.3 | 1.0 | 0.6 | 3.5 | 1.9 | 1.7 |
| 100 | 10 | 0.9 | 0.8 | 0.5 | 4.6 | 2.2 | 2.0 | 1.3 | 1.0 | 0.7 | 5.1 | 2.5 | 2.2 |
| 100 | 15 | 0.9 | 0.8 | 0.6 | 5.6 | 2.7 | 2.4 | 1.3 | 1.0 | 0.7 | 6.2 | 3.0 | 2.6 |
| 100 | 20 | 0.9 | 0.8 | 0.6 | 6.5 | 3.1 | 2.7 | 1.3 | 1.0 | 0.7 | 7.1 | 3.5 | 3.0 |
|  | *Average* | 5.0 | 3.0 | 2.6 | 4.3 | 2.7 | 2.4 | 5.7 | 3.4 | 2.9 | 4.7 | 2.9 | 2.6 |

**Supplemental Table S9**. Mean absolute error and standard deviation of completeness (comp.) and contamination (cont.) estimates determined using domain-specific marker sets (dms), the lineage-specific marker set selected by CheckM (sms), and the best performing lineage-specific marker set (bms). Simulated genomes were generated under the random contig model from 2430 draft genomes with 20 genomes constructed per reference.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Comp (%)** | **Cont (%)** | **dms comp.** | **sms comp.** | **bms comp.** | **dms cont.** | **sms cont.** | **bms cont.** |
| 50 | 0 | 9.4 ± 7.19 | 4.9 ± 4.38 | 4.0 ± 3.63 | 0.2 ± 0.54 | 0.2 ± 0.48 | 0.3 ± 0.53 |
| 50 | 5 | 9.3 ± 7.16 | 4.9 ± 4.39 | 3.9 ± 3.52 | 2.3 ± 1.95 | 1.8 ± 1.31 | 1.7 ± 1.23 |
| 50 | 10 | 9.6 ± 7.29 | 5.3 ± 4.52 | 4.3 ± 3.64 | 4.6 ± 2.99 | 4.2 ± 2.34 | 4.0 ± 2.27 |
| 50 | 15 | 10.2 ± 7.55 | 6.1 ± 4.76 | 5.1 ± 4.02 | 6.8 ± 4.03 | 6.9 ± 3.20 | 6.7 ± 3.16 |
| 50 | 20 | 11.1 ± 7.92 | 7.1 ± 5.07 | 6.2 ± 4.42 | 9.0 ± 5.05 | 9.6 ± 3.97 | 9.3 ± 4.00 |
| 70 | 0 | 8.2 ± 6.39 | 4.2 ± 3.83 | 3.5 ± 3.15 | 0.3 ± 0.72 | 0.4 ± 0.67 | 0.5 ± 0.64 |
| 70 | 5 | 8.1 ± 6.21 | 4.2 ± 3.74 | 3.4 ± 3.02 | 2.3 ± 2.23 | 1.5 ± 1.25 | 1.4 ± 1.14 |
| 70 | 10 | 8.1 ± 6.11 | 4.3 ± 3.71 | 3.5 ± 3.00 | 4.3 ± 3.18 | 3.3 ± 2.19 | 3.1 ± 2.08 |
| 70 | 15 | 8.3 ± 6.04 | 4.5 ± 3.74 | 3.8 ± 3.09 | 5.9 ± 4.05 | 5.3 ± 3.10 | 5.0 ± 3.03 |
| 70 | 20 | 8.6 ± 6.08 | 4.9 ± 3.84 | 4.2 ± 3.34 | 7.3 ± 4.85 | 7.3 ± 3.95 | 6.9 ± 3.88 |
| 80 | 0 | 6.8 ± 5.43 | 3.5 ± 3.18 | 3.0 ± 2.68 | 0.4 ± 0.81 | 0.5 ± 0.80 | 0.5 ± 0.67 |
| 80 | 5 | 6.8 ± 5.27 | 3.5 ± 3.13 | 2.9 ± 2.55 | 2.4 ± 2.38 | 1.5 ± 1.26 | 1.4 ± 1.13 |
| 80 | 10 | 6.7 ± 5.06 | 3.5 ± 3.04 | 2.9 ± 2.51 | 4.2 ± 3.38 | 3.0 ± 2.15 | 2.8 ± 2.01 |
| 80 | 15 | 6.7 ± 4.95 | 3.5 ± 3.00 | 3.0 ± 2.54 | 5.7 ± 4.21 | 4.7 ± 3.00 | 4.4 ± 2.89 |
| 80 | 20 | 6.8 ± 4.87 | 3.7 ± 3.02 | 3.2 ± 2.65 | 6.9 ± 4.90 | 6.3 ± 3.87 | 5.9 ± 3.72 |
| 90 | 0 | 4.4 ± 3.94 | 2.4 ± 2.25 | 2.1 ± 1.96 | 0.5 ± 0.91 | 0.6 ± 0.90 | 0.5 ± 0.64 |
| 90 | 5 | 4.4 ± 3.87 | 2.3 ± 2.20 | 2.0 ± 1.82 | 2.5 ± 2.60 | 1.5 ± 1.32 | 1.3 ± 1.17 |
| 90 | 10 | 4.3 ± 3.72 | 2.3 ± 2.17 | 2.0 ± 1.78 | 4.3 ± 3.66 | 2.8 ± 2.14 | 2.6 ± 1.96 |
| 90 | 15 | 4.3 ± 3.57 | 2.3 ± 2.06 | 2.0 ± 1.77 | 5.7 ± 4.46 | 4.2 ± 2.95 | 3.9 ± 2.80 |
| 90 | 20 | 4.2 ± 3.40 | 2.3 ± 2.02 | 2.0 ± 1.81 | 6.8 ± 5.18 | 5.7 ± 3.81 | 5.1 ± 3.56 |
| 95 | 0 | 2.6 ± 2.81 | 1.6 ± 1.66 | 1.3 ± 1.46 | 0.5 ± 0.95 | 0.7 ± 0.96 | 0.3 ± 0.55 |
| 95 | 5 | 2.6 ± 2.77 | 1.5 ± 1.65 | 1.2 ± 1.23 | 2.5 ± 2.66 | 1.5 ± 1.36 | 1.3 ± 1.19 |
| 95 | 10 | 2.5 ± 2.65 | 1.5 ± 1.59 | 1.3 ± 1.23 | 4.4 ± 3.81 | 2.8 ± 2.16 | 2.5 ± 1.98 |
| 95 | 15 | 2.5 ± 2.53 | 1.5 ± 1.52 | 1.3 ± 1.24 | 5.8 ± 4.64 | 4.1 ± 2.94 | 3.7 ± 2.74 |
| 95 | 20 | 2.4 ± 2.43 | 1.4 ± 1.47 | 1.3 ± 1.25 | 6.8 ± 5.27 | 5.4 ± 3.74 | 4.7 ± 3.42 |
| 100 | 0 | 0.6 ± 1.25 | 0.8 ± 1.42 | 0.1 ± 0.56 | 0.5 ± 0.99 | 0.7 ± 1.00 | 0.1 ± 0.50 |
| 100 | 5 | 0.6 ± 1.22 | 0.7 ± 1.40 | 0.3 ± 0.56 | 2.6 ± 2.77 | 1.5 ± 1.42 | 1.4 ± 1.38 |
| 100 | 10 | 0.6 ± 1.19 | 0.7 ± 1.37 | 0.4 ± 0.71 | 4.5 ± 3.92 | 2.7 ± 2.17 | 2.5 ± 2.04 |
| 100 | 15 | 0.5 ± 1.15 | 0.7 ± 1.32 | 0.4 ± 0.75 | 5.8 ± 4.74 | 4.0 ± 2.93 | 3.6 ± 2.73 |
| 100 | 20 | 0.5 ± 1.11 | 0.7 ± 1.28 | 0.4 ± 0.78 | 6.9 ± 5.45 | 5.3 ± 3.74 | 4.5 ± 3.40 |
|  | *Average* | 5.4 ± 5.85 | 3.0 ± 3.47 | 2.5 ± 2.90 | 4.1 ± 4.37 | 3.3 ± 3.43 | 3.1 ± 3.27 |

**Supplemental Table S10**. Mean absolute error and standard deviation of completeness (comp.) and contamination (cont.) estimates determined using domain-specific marker sets (dms), the lineage-specific marker set selected by CheckM (sms), and the best performing lineage-specific marker set (bms). Simulated genomes were generated under the inverse length model from 2430 draft genomes with 20 genomes constructed per reference.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Comp (%)** | **Cont (%)** | **dms comp.** | **sms comp.** | **bms comp.** | **dms cont.** | **sms cont.** | **bms cont.** |
| 50 | 0 | 9.7 ± 7.47 | 5.5 ± 4.87 | 4.5 ± 4.14 | 0.3 ± 0.75 | 0.3 ± 0.59 | 0.3 ± 0.59 |
| 50 | 5 | 10.3 ± 7.72 | 6.4 ± 5.21 | 5.3 ± 4.56 | 4.6 ± 2.68 | 4.3 ± 2.38 | 4.2 ± 2.41 |
| 50 | 10 | 11.2 ± 8.17 | 7.4 ± 5.53 | 6.3 ± 4.95 | 7.4 ± 3.84 | 7.6 ± 3.20 | 7.4 ± 3.24 |
| 50 | 15 | 12.4 ± 8.83 | 8.7 ± 5.93 | 7.7 ± 5.50 | 10.1 ± 5.19 | 10.9 ± 4.29 | 10.6 ± 4.41 |
| 50 | 20 | 14.1 ± 9.39 | 10.3 ± 6.35 | 9.3 ± 5.98 | 12.8 ± 6.17 | 14.1 ± 4.87 | 13.6 ± 5.07 |
| 70 | 0 | 8.9 ± 6.75 | 5.1 ± 4.33 | 4.2 ± 3.75 | 0.4 ± 0.89 | 0.4 ± 0.71 | 0.5 ± 0.69 |
| 70 | 5 | 9.1 ± 6.68 | 5.5 ± 4.42 | 4.5 ± 3.78 | 4.2 ± 2.76 | 3.5 ± 2.29 | 3.3 ± 2.30 |
| 70 | 10 | 9.3 ± 6.69 | 5.9 ± 4.54 | 5.0 ± 4.03 | 6.3 ± 3.86 | 6.1 ± 3.24 | 5.8 ± 3.23 |
| 70 | 15 | 9.9 ± 6.87 | 6.6 ± 4.71 | 5.8 ± 4.33 | 8.1 ± 4.97 | 8.6 ± 4.31 | 8.1 ± 4.32 |
| 70 | 20 | 10.6 ± 7.05 | 7.4 ± 4.92 | 6.7 ± 4.67 | 9.8 ± 6.00 | 11.0 ± 5.18 | 10.2 ± 5.18 |
| 80 | 0 | 7.7 ± 5.82 | 4.4 ± 3.62 | 3.6 ± 3.10 | 0.5 ± 0.97 | 0.5 ± 0.78 | 0.5 ± 0.72 |
| 80 | 5 | 7.8 ± 5.65 | 4.6 ± 3.68 | 3.8 ± 3.10 | 4.1 ± 2.91 | 3.2 ± 2.22 | 2.9 ± 2.20 |
| 80 | 10 | 7.9 ± 5.62 | 4.9 ± 3.75 | 4.1 ± 3.24 | 6.0 ± 3.93 | 5.4 ± 3.13 | 5.1 ± 3.12 |
| 80 | 15 | 8.2 ± 5.60 | 5.3 ± 3.83 | 4.6 ± 3.48 | 7.5 ± 4.90 | 7.5 ± 4.27 | 7.0 ± 4.21 |
| 80 | 20 | 8.5 ± 5.66 | 5.7 ± 3.95 | 5.2 ± 3.74 | 8.7 ± 5.74 | 9.6 ± 5.17 | 8.8 ± 5.01 |
| 90 | 0 | 5.8 ± 4.49 | 3.2 ± 2.71 | 2.7 ± 2.34 | 0.6 ± 1.06 | 0.6 ± 0.89 | 0.5 ± 0.73 |
| 90 | 5 | 5.7 ± 4.24 | 3.3 ± 2.63 | 2.7 ± 2.21 | 4.1 ± 3.13 | 2.9 ± 2.13 | 2.6 ± 2.07 |
| 90 | 10 | 5.7 ± 4.07 | 3.4 ± 2.63 | 2.8 ± 2.28 | 5.8 ± 4.11 | 4.9 ± 3.16 | 4.5 ± 3.10 |
| 90 | 15 | 5.7 ± 3.97 | 3.5 ± 2.66 | 3.1 ± 2.37 | 7.1 ± 4.97 | 6.7 ± 4.31 | 6.2 ± 4.19 |
| 90 | 20 | 5.8 ± 3.91 | 3.7 ± 2.70 | 3.4 ± 2.51 | 8.1 ± 5.72 | 8.4 ± 5.11 | 7.5 ± 4.80 |
| 95 | 0 | 4.2 ± 3.43 | 2.3 ± 2.02 | 2.0 ± 1.76 | 0.6 ± 1.12 | 0.6 ± 0.93 | 0.5 ± 0.68 |
| 95 | 5 | 4.2 ± 3.26 | 2.3 ± 1.97 | 2.0 ± 1.62 | 4.1 ± 3.21 | 2.8 ± 2.10 | 2.5 ± 2.04 |
| 95 | 10 | 4.1 ± 3.12 | 2.4 ± 1.94 | 2.0 ± 1.65 | 5.8 ± 4.22 | 4.6 ± 3.07 | 4.2 ± 2.95 |
| 95 | 15 | 4.1 ± 3.00 | 2.4 ± 1.92 | 2.1 ± 1.69 | 7.0 ± 5.07 | 6.3 ± 4.11 | 5.7 ± 3.86 |
| 95 | 20 | 4.1 ± 2.88 | 2.4 ± 1.89 | 2.3 ± 1.77 | 8.0 ± 5.76 | 7.9 ± 5.08 | 6.9 ± 4.67 |
| 100 | 0 | 0.8 ± 1.42 | 0.8 ± 1.42 | 0.3 ± 0.70 | 0.7 ± 1.20 | 0.7 ± 1.02 | 0.3 ± 0.67 |
| 100 | 5 | 0.7 ± 1.36 | 0.8 ± 1.39 | 0.5 ± 0.77 | 4.2 ± 3.44 | 2.7 ± 2.09 | 2.4 ± 1.97 |
| 100 | 10 | 0.7 ± 1.31 | 0.7 ± 1.34 | 0.5 ± 0.81 | 5.9 ± 4.41 | 4.4 ± 3.04 | 3.9 ± 2.82 |
| 100 | 15 | 0.6 ± 1.26 | 0.7 ± 1.29 | 0.5 ± 0.82 | 7.0 ± 5.18 | 6.0 ± 4.06 | 5.1 ± 3.69 |
| 100 | 20 | 0.6 ± 1.21 | 0.7 ± 1.26 | 0.5 ± 0.80 | 8.0 ± 6.04 | 7.5 ± 5.06 | 6.2 ± 4.48 |
|  | *Average* | 6.6 ± 6.54 | 4.2 ± 4.38 | 3.6 ± 3.91 | 5.6 ± 5.26 | 5.3 ± 4.92 | 4.9 ± 4.71 |

**Supplemental Table S11**. Mean absolute error and standard deviation of completeness (comp.) and contamination (cont.) estimates determined using domain-specific marker sets (dms) and the lineage-specific marker set selected by CheckM (sms). Simulated genomes were generated under the random fragment model from 3324 draft genomes with 20 genomes constructed per reference with a completeness of 50%, 70%, 80%, or 90% and contamination of 5%, 10%, or 15%. Results are given for simulated genomes generated with a window size of 20 kbp.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Taxon** | **# draft genomes** | **dms  comp. (%)** | **sms comp. (%)** | **dms cont. (%)** | **sms cont. (%)** |
| **Domain** | Archaea | 56 | 6.2 ± 4.77 | 4.1 ± 3.38 | 4.6 ± 3.44 | 3.3 ± 2.56 |
|  | Bacteria | 3268 | 6.4 ± 5.04 | 3.7 ± 3.07 | 4.8 ± 3.53 | 3.1 ± 2.43 |
|  |  |  |  |  |  |  |
| **Phylum** | Acidobacteria | 9 | 6.1 ± 4.84 | 4.0 ± 3.23 | 4.6 ± 3.34 | 3.4 ± 2.65 |
|  | Actinobacteria | 484 | 6.5 ± 5.06 | 3.6 ± 3.00 | 4.8 ± 3.56 | 3.0 ± 2.39 |
|  | Aquificae | 5 | 6.5 ± 5.16 | 3.5 ± 2.87 | 4.7 ± 3.38 | 3.0 ± 2.38 |
|  | Bacteroidetes | 316 | 5.9 ± 4.61 | 3.5 ± 2.84 | 4.5 ± 3.25 | 2.9 ± 2.29 |
|  | Chlamydiae | 15 | 6.7 ± 5.12 | 3.8 ± 3.10 | 5.1 ± 3.62 | 3.4 ± 2.56 |
|  | Chloroflexi | 3 | 6.1 ± 4.72 | 5.0 ± 3.95 | 4.6 ± 3.30 | 3.5 ± 2.79 |
|  | Crenarchaeota | 8 | 6.5 ± 5.16 | 5.2 ± 4.14 | 4.8 ± 3.64 | 3.9 ± 2.95 |
|  | Cyanobacteria | 49 | 5.8 ± 4.50 | 3.1 ± 2.55 | 4.3 ± 3.25 | 2.7 ± 2.14 |
|  | Deinococcus-Thermus | 20 | 5.6 ± 4.50 | 3.6 ± 2.97 | 4.6 ± 3.38 | 3.0 ± 2.38 |
|  | Euryarchaeota | 46 | 6.2 ± 4.72 | 3.9 ± 3.18 | 4.5 ± 3.42 | 3.1 ± 2.46 |
|  | Firmicutes | 853 | 6.5 ± 5.10 | 4.0 ± 3.30 | 4.7 ± 3.56 | 3.2 ± 2.55 |
|  | Fusobacteria | 26 | 6.0 ± 4.68 | 5.0 ± 3.99 | 4.5 ± 3.33 | 3.8 ± 2.87 |
|  | Nitrospirae | 2 | 6.4 ± 5.02 | 4.8 ± 3.74 | 4.9 ± 3.50 | 3.4 ± 2.80 |
|  | Planctomycetes | 5 | 6.3 ± 4.82 | 4.9 ± 3.81 | 4.7 ± 3.52 | 3.7 ± 2.92 |
|  | Proteobacteria | 1353 | 6.6 ± 5.15 | 3.4 ± 2.84 | 5.0 ± 3.58 | 2.9 ± 2.32 |
|  | Spirochaetes | 31 | 6.3 ± 5.13 | 4.5 ± 3.68 | 4.9 ± 3.63 | 3.6 ± 2.83 |
|  | Synergistetes | 9 | 6.8 ± 5.19 | 6.8 ± 5.17 | 4.9 ± 3.58 | 4.9 ± 3.57 |
|  | Tenericutes | 65 | 5.6 ± 4.40 | 4.5 ± 3.64 | 4.6 ± 3.29 | 3.6 ± 2.85 |
|  | Thermodesulfobacteria | 3 | 6.0 ± 4.71 | 3.7 ± 3.03 | 4.6 ± 3.38 | 2.8 ± 2.16 |
|  | Verrucomicrobia | 7 | 6.3 ± 4.97 | 4.3 ± 3.37 | 4.8 ± 3.56 | 3.4 ± 2.62 |
|  |  |  |  |  |  |  |
| **Class** | Acidobacteriia | 6 | 6.0 ± 4.76 | 3.9 ± 3.22 | 4.7 ± 3.30 | 3.4 ± 2.57 |
|  | Actinobacteria | 482 | 6.5 ± 5.06 | 3.6 ± 3.00 | 4.8 ± 3.56 | 3.0 ± 2.39 |
|  | Alphaproteobacteria | 383 | 6.6 ± 5.12 | 3.5 ± 2.90 | 4.9 ± 3.54 | 2.9 ± 2.32 |
|  | Aquificae | 5 | 6.5 ± 5.16 | 3.5 ± 2.87 | 4.7 ± 3.38 | 3.0 ± 2.38 |
|  | Bacilli | 496 | 6.5 ± 5.12 | 3.8 ± 3.19 | 4.8 ± 3.56 | 3.2 ± 2.50 |
|  | Bacteroidia | 168 | 5.9 ± 4.61 | 3.5 ± 2.81 | 4.5 ± 3.26 | 2.9 ± 2.29 |
|  | Betaproteobacteria | 193 | 6.3 ± 4.92 | 3.5 ± 2.84 | 4.8 ± 3.47 | 3.0 ± 2.33 |
|  | Chlamydiia | 15 | 6.7 ± 5.12 | 3.8 ± 3.10 | 5.1 ± 3.62 | 3.4 ± 2.56 |
|  | Chroococcales | 21 | 5.7 ± 4.50 | 3.1 ± 2.59 | 4.3 ± 3.28 | 2.8 ± 2.19 |
|  | Clostridia | 283 | 6.5 ± 5.08 | 4.1 ± 3.40 | 4.7 ± 3.57 | 3.3 ± 2.59 |
|  | Cytophagia | 35 | 6.0 ± 4.75 | 3.5 ± 2.82 | 4.5 ± 3.30 | 2.8 ± 2.24 |
|  | Deinococci | 20 | 5.6 ± 4.50 | 3.6 ± 2.97 | 4.6 ± 3.38 | 3.0 ± 2.38 |
|  | Deltaproteobacteria | 48 | 6.8 ± 5.24 | 4.3 ± 3.41 | 5.1 ± 3.71 | 3.4 ± 2.58 |
|  | Epsilonproteobacteria | 52 | 5.5 ± 4.38 | 3.4 ± 2.77 | 4.4 ± 3.19 | 2.9 ± 2.33 |
|  | Erysipelotrichi | 17 | 6.6 ± 5.08 | 5.1 ± 4.05 | 4.9 ± 3.62 | 3.8 ± 2.88 |
|  | Flavobacteriia | 94 | 5.8 ± 4.57 | 3.5 ± 2.86 | 4.5 ± 3.22 | 2.9 ± 2.30 |
|  | Fusobacteriia | 26 | 6.0 ± 4.68 | 5.0 ± 3.99 | 4.5 ± 3.33 | 3.8 ± 2.87 |
|  | Gammaproteobacteria | 675 | 6.8 ± 5.27 | 3.3 ± 2.76 | 5.1 ± 3.65 | 2.9 ± 2.28 |
|  | Halobacteria | 33 | 6.2 ± 4.69 | 3.6 ± 2.92 | 4.5 ± 3.38 | 3.0 ± 2.32 |
|  | Holophagae | 2 | 6.3 ± 5.04 | 3.9 ± 3.10 | 4.3 ± 3.45 | 3.3 ± 2.73 |
|  | Methanobacteria | 4 | 6.7 ± 5.08 | 5.2 ± 3.99 | 4.8 ± 3.65 | 3.9 ± 2.90 |
|  | Methanomicrobia | 4 | 6.4 ± 4.83 | 4.7 ± 3.62 | 4.5 ± 3.61 | 3.4 ± 2.67 |
|  | Mollicutes | 65 | 5.6 ± 4.40 | 4.5 ± 3.64 | 4.6 ± 3.29 | 3.6 ± 2.85 |
|  | Negativicutes | 57 | 6.4 ± 4.99 | 4.2 ± 3.37 | 4.7 ± 3.46 | 3.4 ± 2.61 |
|  | Nitrospira | 2 | 6.4 ± 5.02 | 4.8 ± 3.74 | 4.9 ± 3.50 | 3.4 ± 2.80 |
|  | Nostocales | 6 | 5.8 ± 4.54 | 3.0 ± 2.51 | 4.3 ± 3.24 | 2.6 ± 2.09 |
|  | Oscillatoriales | 13 | 5.7 ± 4.39 | 3.0 ± 2.49 | 4.3 ± 3.14 | 2.5 ± 2.09 |
|  | Planctomycetia | 5 | 6.3 ± 4.82 | 4.9 ± 3.81 | 4.7 ± 3.52 | 3.7 ± 2.92 |
|  | Pleurocapsales | 2 | 6.2 ± 4.56 | 3.0 ± 2.54 | 4.8 ± 3.35 | 2.5 ± 2.14 |
|  | Solirubrobacterales | 2 | 6.7 ± 5.23 | 4.9 ± 3.98 | 4.8 ± 3.54 | 3.5 ± 2.85 |
|  | Sphingobacteriia | 17 | 5.9 ± 4.58 | 3.7 ± 3.08 | 4.5 ± 3.19 | 2.9 ± 2.29 |
|  | Spirochaetia | 31 | 6.3 ± 5.13 | 4.5 ± 3.68 | 4.9 ± 3.63 | 3.6 ± 2.83 |
|  | Stigonematales | 5 | 6.1 ± 4.61 | 3.2 ± 2.52 | 4.5 ± 3.41 | 2.5 ± 2.05 |
|  | Synergistia | 9 | 6.8 ± 5.19 | 6.8 ± 5.17 | 4.9 ± 3.58 | 4.9 ± 3.57 |
|  | Thermococci | 2 | 5.5 ± 4.22 | 3.7 ± 2.95 | 4.2 ± 3.19 | 3.3 ± 2.51 |
|  | Thermodesulfobacteria | 3 | 6.0 ± 4.71 | 3.7 ± 3.03 | 4.6 ± 3.38 | 2.8 ± 2.16 |
|  | Thermoprotei | 8 | 6.5 ± 5.16 | 5.2 ± 4.14 | 4.8 ± 3.64 | 3.9 ± 2.95 |
|  | Verrucomicrobiae | 6 | 6.4 ± 5.02 | 4.4 ± 3.42 | 4.8 ± 3.52 | 3.5 ± 2.62 |
|  | Zetaproteobacteria | 2 | 6.3 ± 4.47 | 4.5 ± 3.67 | 5.0 ± 3.31 | 3.7 ± 2.65 |

**Supplemental Table S12**. Mean absolute error and standard deviation of completeness (comp.) and contamination (cont.) estimates determined using domain-specific marker sets (dms) and the lineage-specific marker sets selected by CheckM (sms). Simulated genomes were generated under the random contig model from 2430 draft genomes with 20 genomes constructed per reference with a completeness of 50%, 70%, 80%, or 90% and contamination of 5%, 10%, or 15%.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Taxon** | **# draft genomes** | **dms  comp. (%)** | **sms comp. (%)** | **dms cont. (%)** | **sms cont. (%)** |
| **Domain** | k\_\_Archaea | 40 | 6.5 ± 5.85 | 4.3 ± 4.29 | 5.3 ± 3.76 | 4.5 ± 3.26 |
|  | k\_\_Bacteria | 2390 | 7.2 ± 6.06 | 3.9 ± 3.66 | 4.2 ± 3.67 | 3.4 ± 2.83 |
|  |  |  |  |  |  |  |
| **Phylum** | Acidobacteria | 2 | 6.8 ± 5.64 | 5.0 ± 4.54 | 4.0 ± 3.59 | 3.3 ± 2.96 |
|  | Actinobacteria | 337 | 8.1 ± 6.59 | 4.1 ± 3.66 | 4.3 ± 3.73 | 3.3 ± 2.75 |
|  | Bacteroidetes | 251 | 6.2 ± 5.08 | 3.3 ± 2.95 | 4.2 ± 3.62 | 3.0 ± 2.58 |
|  | Chlamydiae | 9 | 4.3 ± 4.01 | 2.8 ± 2.52 | 4.2 ± 3.83 | 3.2 ± 3.37 |
|  | Chloroflexi | 2 | 6.1 ± 5.08 | 4.9 ± 4.03 | 4.4 ± 3.79 | 3.4 ± 2.84 |
|  | Crenarchaeota | 5 | 6.9 ± 8.55 | 5.9 ± 7.47 | 5.1 ± 3.63 | 3.8 ± 2.97 |
|  | Cyanobacteria | 15 | 5.6 ± 4.60 | 2.5 ± 2.20 | 4.2 ± 3.68 | 3.7 ± 2.89 |
|  | Deinococcus-Thermus | 17 | 6.3 ± 5.12 | 3.7 ± 3.27 | 4.2 ± 3.59 | 3.1 ± 2.57 |
|  | Euryarchaeota | 34 | 6.5 ± 5.38 | 4.1 ± 3.57 | 5.4 ± 3.78 | 4.6 ± 3.29 |
|  | Firmicutes | 674 | 7.5 ± 6.34 | 4.5 ± 4.30 | 4.2 ± 3.68 | 3.3 ± 2.77 |
|  | Fusobacteria | 19 | 5.5 ± 4.98 | 4.6 ± 4.25 | 4.1 ± 3.54 | 3.4 ± 2.90 |
|  | Planctomycetes | 4 | 7.4 ± 6.06 | 5.3 ± 4.21 | 4.2 ± 3.65 | 3.6 ± 2.95 |
|  | Proteobacteria | 992 | 7.2 ± 5.95 | 3.5 ± 3.24 | 4.3 ± 3.67 | 3.6 ± 2.93 |
|  | Spirochaetes | 21 | 7.0 ± 6.12 | 4.8 ± 4.54 | 4.3 ± 3.75 | 3.3 ± 2.93 |
|  | Synergistetes | 3 | 7.4 ± 5.80 | 7.5 ± 5.85 | 4.1 ± 3.47 | 4.1 ± 3.45 |
|  | Tenericutes | 31 | 5.4 ± 4.54 | 4.0 ± 3.57 | 4.1 ± 3.49 | 2.9 ± 2.62 |
|  | Thermodesulfobacteria | 3 | 6.1 ± 4.93 | 3.5 ± 2.98 | 4.3 ± 3.70 | 3.0 ± 2.57 |
|  | Verrucomicrobia | 2 | 6.7 ± 6.41 | 4.2 ± 3.47 | 4.1 ± 3.55 | 3.2 ± 2.58 |
|  |  |  |  |  |  |  |
| **Class** | Actinobacteria | 335 | 8.1 ± 6.60 | 4.1 ± 3.66 | 4.3 ± 3.73 | 3.3 ± 2.75 |
|  | Alphaproteobacteria | 227 | 7.5 ± 6.14 | 4.1 ± 3.63 | 4.3 ± 3.69 | 3.5 ± 2.88 |
|  | Bacilli | 392 | 7.7 ± 6.63 | 4.3 ± 4.26 | 4.2 ± 3.70 | 3.3 ± 2.79 |
|  | Bacteroidia | 149 | 6.2 ± 5.04 | 3.2 ± 2.86 | 4.2 ± 3.62 | 2.9 ± 2.52 |
|  | Betaproteobacteria | 144 | 6.8 ± 5.65 | 3.8 ± 3.56 | 4.2 ± 3.65 | 3.4 ± 2.83 |
|  | Chlamydiia | 9 | 4.3 ± 4.01 | 2.8 ± 2.52 | 4.2 ± 3.83 | 3.2 ± 3.37 |
|  | Chroococcales | 5 | 5.1 ± 4.25 | 2.8 ± 2.60 | 4.3 ± 3.93 | 3.5 ± 2.86 |
|  | Clostridia | 220 | 7.2 ± 5.88 | 4.5 ± 4.20 | 4.2 ± 3.65 | 3.3 ± 2.76 |
|  | Cytophagia | 23 | 6.9 ± 5.80 | 3.9 ± 3.30 | 4.2 ± 3.69 | 3.2 ± 2.68 |
|  | Deinococci | 17 | 6.3 ± 5.12 | 3.7 ± 3.27 | 4.2 ± 3.59 | 3.1 ± 2.57 |
|  | Deltaproteobacteria | 34 | 7.0 ± 5.73 | 4.5 ± 3.70 | 4.3 ± 3.72 | 3.2 ± 2.75 |
|  | Epsilonproteobacteria | 49 | 5.4 ± 4.46 | 2.7 ± 2.33 | 4.2 ± 3.60 | 2.7 ± 2.44 |
|  | Erysipelotrichi | 15 | 7.8 ± 6.35 | 6.2 ± 5.04 | 4.2 ± 3.53 | 3.2 ± 2.73 |
|  | Flavobacteriia | 67 | 5.9 ± 4.90 | 3.3 ± 2.97 | 4.2 ± 3.60 | 3.2 ± 2.66 |
|  | Fusobacteriia | 19 | 5.5 ± 4.98 | 4.6 ± 4.25 | 4.1 ± 3.54 | 3.4 ± 2.90 |
|  | Gammaproteobacteria | 536 | 7.4 ± 6.06 | 3.2 ± 2.94 | 4.3 ± 3.66 | 3.8 ± 3.01 |
|  | Halobacteria | 26 | 6.7 ± 5.53 | 4.1 ± 3.61 | 5.4 ± 3.77 | 4.5 ± 3.18 |
|  | Methanobacteria | 4 | 5.9 ± 4.86 | 4.4 ± 3.71 | 5.3 ± 3.84 | 4.9 ± 3.68 |
|  | Mollicutes | 31 | 5.4 ± 4.54 | 4.0 ± 3.57 | 4.1 ± 3.49 | 2.9 ± 2.62 |
|  | Negativicutes | 47 | 6.9 ± 5.75 | 5.2 ± 4.55 | 4.1 ± 3.69 | 3.1 ± 2.68 |
|  | Nostocales | 4 | 5.9 ± 4.79 | 2.5 ± 1.97 | 4.2 ± 3.63 | 3.8 ± 2.87 |
|  | Oscillatoriales | 4 | 5.8 ± 4.62 | 2.2 ± 1.92 | 4.0 ± 3.43 | 3.7 ± 2.90 |
|  | Planctomycetia | 4 | 7.4 ± 6.06 | 5.3 ± 4.21 | 4.2 ± 3.65 | 3.6 ± 2.95 |
|  | Solirubrobacterales | 2 | 7.9 ± 6.37 | 5.5 ± 4.45 | 4.4 ± 3.73 | 3.4 ± 2.77 |
|  | Sphingobacteriia | 11 | 6.1 ± 4.87 | 3.8 ± 3.17 | 4.3 ± 3.61 | 3.0 ± 2.56 |
|  | Spirochaetia | 21 | 7.0 ± 6.12 | 4.8 ± 4.54 | 4.3 ± 3.75 | 3.3 ± 2.93 |
|  | Synergistia | 3 | 7.4 ± 5.80 | 7.5 ± 5.85 | 4.1 ± 3.47 | 4.1 ± 3.45 |
|  | Thermodesulfobacteria | 3 | 6.1 ± 4.93 | 3.5 ± 2.98 | 4.3 ± 3.70 | 3.0 ± 2.57 |
|  | Thermoprotei | 5 | 6.9 ± 8.55 | 5.9 ± 7.47 | 5.1 ± 3.63 | 3.8 ± 2.97 |
|  | Verrucomicrobiae | 2 | 6.7 ± 6.41 | 4.2 ± 3.47 | 4.1 ± 3.55 | 3.2 ± 2.58 |
|  | Zetaproteobacteria | 2 | 5.4 ± 4.30 | 3.5 ± 3.06 | 4.3 ± 3.64 | 3.4 ± 3.14 |

**Supplemental Table S13**. Mean absolute error and standard deviation of completeness (comp.) and contamination (cont.) estimates determined using domain-specific marker sets (dms) and the lineage-specific marker sets selected by CheckM (sms). Simulated genomes were generated under the inverse length model from 2430 draft genomes with 20 genomes being constructed per reference with a completeness of 50%, 70%, 80%, or 90% and contamination of 5%, 10%, or 15%.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Taxon** | **# draft genomes** | **dms  comp. (%)** | **sms comp. (%)** | **dms cont. (%)** | **sms cont. (%)** |
| **Domain** | Archaea | 40 | 9.1 ± 7.32 | 6.7 ± 6.12 | 8.2 ± 4.60 | 7.6 ± 4.20 |
|  | Bacteria | 2390 | 8.6 ± 6.70 | 5.4 ± 4.53 | 6.3 ± 4.40 | 6.0 ± 4.07 |
|  |  |  |  |  |  |  |
| **Phylum** | Acidobacteria | 2 | 9.4 ± 6.56 | 6.5 ± 5.19 | 6.2 ± 4.76 | 5.3 ± 4.20 |
|  | Actinobacteria | 337 | 9.4 ± 7.17 | 5.5 ± 4.45 | 6.2 ± 4.42 | 5.8 ± 4.01 |
|  | Bacteroidetes | 251 | 7.7 ± 5.78 | 4.9 ± 3.88 | 6.3 ± 4.35 | 5.4 ± 3.76 |
|  | Chlamydiae | 9 | 7.5 ± 5.38 | 5.8 ± 4.08 | 6.6 ± 4.67 | 5.0 ± 4.03 |
|  | Chloroflexi | 2 | 8.0 ± 6.09 | 7.2 ± 5.17 | 5.9 ± 4.07 | 4.4 ± 3.29 |
|  | Crenarchaeota | 5 | 13.3 ± 11.58 | 12.0 ± 10.24 | 8.2 ± 4.82 | 6.9 ± 4.50 |
|  | Cyanobacteria | 15 | 7.5 ± 5.71 | 3.8 ± 3.15 | 6.3 ± 4.48 | 6.9 ± 4.17 |
|  | Deinococcus-Thermus | 17 | 7.1 ± 5.52 | 5.4 ± 4.09 | 6.1 ± 4.19 | 5.1 ± 3.56 |
|  | Euryarchaeota | 34 | 8.5 ± 6.28 | 5.9 ± 4.83 | 8.2 ± 4.56 | 7.7 ± 4.14 |
|  | Firmicutes | 674 | 9.0 ± 7.02 | 6.0 ± 5.08 | 6.2 ± 4.41 | 5.7 ± 4.00 |
|  | Fusobacteria | 19 | 7.5 ± 6.08 | 6.5 ± 5.22 | 6.3 ± 4.53 | 5.6 ± 4.13 |
|  | Planctomycetes | 4 | 8.1 ± 6.31 | 5.4 ± 4.58 | 5.7 ± 4.37 | 4.5 ± 3.79 |
|  | Proteobacteria | 992 | 8.4 ± 6.54 | 5.1 ± 4.18 | 6.3 ± 4.40 | 6.4 ± 4.19 |
|  | Spirochaetes | 21 | 8.8 ± 7.78 | 7.8 ± 6.97 | 6.2 ± 4.48 | 5.4 ± 3.89 |
|  | Synergistetes | 3 | 8.0 ± 6.07 | 8.0 ± 6.09 | 6.3 ± 4.55 | 6.2 ± 4.53 |
|  | Tenericutes | 31 | 6.6 ± 5.45 | 6.4 ± 4.67 | 6.1 ± 4.35 | 4.5 ± 3.55 |
|  | Thermodesulfobacteria | 3 | 7.0 ± 5.20 | 5.5 ± 4.07 | 6.2 ± 3.99 | 4.0 ± 3.09 |
|  | Verrucomicrobia | 2 | 7.3 ± 6.14 | 5.0 ± 3.93 | 6.3 ± 4.25 | 4.5 ± 3.49 |
|  |  |  |  |  |  |  |
| **Class** | Actinobacteria | 335 | 9.4 ± 7.18 | 5.5 ± 4.45 | 6.2 ± 4.42 | 5.8 ± 4.01 |
|  | Alphaproteobacteria | 227 | 9.0 ± 6.95 | 6.3 ± 4.97 | 6.4 ± 4.49 | 6.3 ± 4.19 |
|  | Bacilli | 392 | 9.3 ± 7.24 | 5.8 ± 4.99 | 6.2 ± 4.43 | 5.8 ± 4.04 |
|  | Bacteroidia | 149 | 7.7 ± 5.70 | 4.8 ± 3.80 | 6.3 ± 4.37 | 5.3 ± 3.72 |
|  | Betaproteobacteria | 144 | 7.8 ± 6.17 | 5.1 ± 4.22 | 6.3 ± 4.35 | 6.1 ± 4.05 |
|  | Chlamydiia | 9 | 7.5 ± 5.38 | 5.8 ± 4.08 | 6.6 ± 4.67 | 5.0 ± 4.03 |
|  | Chroococcales | 5 | 8.0 ± 6.09 | 4.9 ± 3.68 | 6.8 ± 4.83 | 7.3 ± 4.49 |
|  | Clostridia | 220 | 8.6 ± 6.65 | 6.0 ± 5.09 | 6.1 ± 4.39 | 5.8 ± 3.96 |
|  | Cytophagia | 23 | 8.7 ± 6.29 | 5.8 ± 4.45 | 6.2 ± 4.34 | 5.6 ± 3.84 |
|  | Deinococci | 17 | 7.1 ± 5.52 | 5.4 ± 4.09 | 6.1 ± 4.19 | 5.1 ± 3.56 |
|  | Deltaproteobacteria | 34 | 8.4 ± 6.35 | 6.5 ± 4.80 | 6.2 ± 4.34 | 5.3 ± 3.73 |
|  | Epsilonproteobacteria | 49 | 7.7 ± 5.65 | 5.1 ± 3.61 | 6.0 ± 4.26 | 4.7 ± 3.54 |
|  | Erysipelotrichi | 15 | 9.4 ± 6.96 | 7.1 ± 5.57 | 6.0 ± 4.23 | 4.8 ± 3.64 |
|  | Flavobacteriia | 67 | 7.6 ± 5.81 | 4.9 ± 3.85 | 6.2 ± 4.35 | 5.7 ± 3.83 |
|  | Fusobacteriia | 19 | 7.5 ± 6.08 | 6.5 ± 5.22 | 6.3 ± 4.53 | 5.6 ± 4.13 |
|  | Gammaproteobacteria | 536 | 8.4 ± 6.52 | 4.4 ± 3.63 | 6.4 ± 4.39 | 6.7 ± 4.25 |
|  | Halobacteria | 26 | 8.9 ± 6.50 | 6.2 ± 5.06 | 8.2 ± 4.55 | 7.6 ± 4.05 |
|  | Methanobacteria | 4 | 7.7 ± 5.47 | 5.8 ± 4.09 | 8.3 ± 4.62 | 8.4 ± 4.44 |
|  | Mollicutes | 31 | 6.6 ± 5.45 | 6.4 ± 4.67 | 6.1 ± 4.35 | 4.5 ± 3.55 |
|  | Negativicutes | 47 | 8.7 ± 6.73 | 7.1 ± 5.45 | 6.2 ± 4.39 | 5.5 ± 3.87 |
|  | Nostocales | 4 | 7.4 ± 5.45 | 3.9 ± 2.98 | 6.0 ± 4.24 | 6.7 ± 4.16 |
|  | Oscillatoriales | 4 | 7.2 ± 5.47 | 3.0 ± 2.46 | 5.9 ± 4.11 | 6.6 ± 3.71 |
|  | Planctomycetia | 4 | 8.1 ± 6.31 | 5.4 ± 4.58 | 5.7 ± 4.37 | 4.5 ± 3.79 |
|  | Solirubrobacterales | 2 | 8.9 ± 6.48 | 6.2 ± 5.09 | 6.3 ± 4.64 | 4.6 ± 3.85 |
|  | Sphingobacteriia | 11 | 6.8 ± 5.40 | 4.6 ± 3.65 | 6.2 ± 4.18 | 4.9 ± 3.57 |
|  | Spirochaetia | 21 | 8.8 ± 7.78 | 7.8 ± 6.97 | 6.2 ± 4.48 | 5.4 ± 3.89 |
|  | Synergistia | 3 | 8.0 ± 6.07 | 8.0 ± 6.09 | 6.3 ± 4.55 | 6.2 ± 4.53 |
|  | Thermodesulfobacteria | 3 | 7.0 ± 5.20 | 5.5 ± 4.07 | 6.2 ± 3.99 | 4.0 ± 3.09 |
|  | Thermoprotei | 5 | 13.3 ± 11.58 | 12.0 ± 10.24 | 8.2 ± 4.82 | 6.9 ± 4.50 |
|  | Verrucomicrobiae | 2 | 7.3 ± 6.14 | 5.0 ± 3.93 | 6.3 ± 4.25 | 4.5 ± 3.49 |
|  | Zetaproteobacteria | 2 | 6.6 ± 5.03 | 4.3 ± 3.33 | 6.4 ± 4.29 | 5.3 ± 3.71 |

**Supplemental Table S14**. Taxonomic rank of the selected lineage-specific marker set used for evaluating the quality of genomes at different degrees of taxonomic novelty. Results for genomes from species with multiple strains are reported under the label *other*.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Taxonomic novelty** |  | **lineage of selected marker set** | | | | | | |
| **# genomes** | **domain** | **phylum** | **class** | **order** | **family** | **genus** | **species** |
| phylum | 3 | 3 (100%) | - | - | - | - | - | - |
| class | 3 | 2 (67%) | 1 (33%) | - | - | - | - | - |
| order | 12 | 3 (25%) | - | 9 (75%) | - | - | - | - |
| family | 33 | 6 (18%) | 8 (24%) | 9 (27%) | 10 (30%) | - | - | - |
| genus | 313 | 21 (7%) | 36 (12%) | 102 (33%) | 109 (35%) | 45 (14%) | - | - |
| species | 1037 | 63 (6%) | 103 (10%) | 236 (23%) | 390 (38%) | 164 (16%) | 81 (8%) | - |
| *other* | 1923 | 52 (3%) | 86 (4%) | 206 (11%) | 568 (30%) | 489 (25%) | 495 (26%) | 27 (1%) |

**Supplemental Table S15**. Mean absolute error and standard deviation of completeness (comp.) and contamination (cont.) estimates for simulated genomes at different degrees of taxonomic novelty. Simulated genomes were generated under the random fragment model from 3324 draft genomes with 20 genomes constructed per reference and a window size of 20 kbp. Results for genomes from species with multiple strains are reported under the label *other*.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Taxonomic novelty** |  | **comp. = 70%, cont. = 5%** | | **comp. = 80%, cont. = 5%** | | **comp. = 90%, cont. = 5%** | |
| **# genomes** | **comp.** | **cont.** | **comp.** | **cont.** | **comp.** | **cont.** |
| phylum | 3 | 5.2 ± 4.60 | 2.1 ± 1.38 | 5.5 ± 3.57 | 2.3 ± 2.16 | 6.0 ± 3.35 | 2.3 ± 1.57 |
| class | 3 | 5.7 ± 3.60 | 2.0 ± 1.52 | 4.8 ± 3.52 | 2.3 ± 1.87 | 3.5 ± 2.39 | 2.1 ± 1.95 |
| order | 12 | 3.8 ± 2.81 | 1.7 ± 1.30 | 3.3 ± 2.59 | 2.0 ± 1.68 | 2.5 ± 1.95 | 2.1 ± 1.69 |
| family | 33 | 3.8 ± 2.91 | 1.9 ± 1.42 | 3.2 ± 2.54 | 1.8 ± 1.58 | 2.7 ± 2.38 | 2.0 ± 1.71 |
| genus | 313 | 3.6 ± 2.87 | 1.9 ± 1.33 | 3.2 ± 2.64 | 1.8 ± 1.37 | 2.5 ± 2.23 | 1.8 ± 1.54 |
| species | 1037 | 3.5 ± 2.78 | 1.8 ± 1.30 | 3.0 ± 2.42 | 1.7 ± 1.32 | 2.3 ± 1.98 | 1.7 ± 1.42 |
| *other* | 1923 | 3.1 ± 2.48 | 1.8 ± 1.21 | 2.6 ± 2.14 | 1.6 ± 1.21 | 2.0 ± 1.71 | 1.5 ± 1.27 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Taxonomic novelty** |  | **comp. = 70%, cont. = 10%** | | **comp. = 80%, cont. = 10%** | | **comp. = 90%, cont. = 10%** | |
| **# genomes** | **comp.** | **cont.** | **comp.** | **cont.** | **comp.** | **cont.** |
| phylum | 3 | 4.5 ± 3.42 | 3.6 ± 2.10 | 5.1 ± 3.69 | 2.9 ± 2.08 | 5.4 ± 3.36 | 3.1 ± 2.37 |
| class | 3 | 4.6 ± 3.14 | 3.5 ± 2.06 | 4.8 ± 3.80 | 2.8 ± 1.90 | 3.9 ± 3.05 | 3.5 ± 3.38 |
| order | 12 | 4.5 ± 3.19 | 3.0 ± 1.99 | 3.3 ± 2.27 | 2.6 ± 2.01 | 2.5 ± 1.88 | 2.6 ± 2.17 |
| family | 33 | 4.1 ± 3.03 | 3.3 ± 2.13 | 3.4 ± 2.72 | 2.5 ± 1.95 | 2.6 ± 2.20 | 2.6 ± 2.09 |
| genus | 313 | 4.0 ± 3.09 | 3.1 ± 2.06 | 3.4 ± 2.67 | 2.6 ± 1.93 | 2.6 ± 2.19 | 2.4 ± 1.97 |
| species | 1037 | 4.0 ± 2.98 | 3.1 ± 2.02 | 3.3 ± 2.50 | 2.6 ± 1.88 | 2.4 ± 1.96 | 2.3 ± 1.91 |
| *other* | 1923 | 3.7 ± 2.72 | 3.0 ± 1.89 | 2.9 ± 2.24 | 2.4 ± 1.77 | 2.1 ± 1.72 | 2.1 ± 1.69 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Taxonomic novelty** |  | **comp. = 70%, cont. = 15%** | | **comp. = 80%, cont. = 15%** | | **comp. = 90%, cont. = 15%** | |
| **# genomes** | **comp.** | **cont.** | **comp.** | **cont.** | **comp.** | **cont.** |
| phylum | 3 | 4.3 ± 3.53 | 4.8 ± 2.71 | 4.0 ± 2.84 | 4.7 ± 2.98 | 6.2 ± 3.90 | 3.8 ± 2.69 |
| class | 3 | 5.5 ± 4.13 | 3.8 ± 3.10 | 4.3 ± 3.93 | 4.0 ± 2.66 | 4.5 ± 3.08 | 3.7 ± 2.90 |
| order | 12 | 5.4 ± 3.79 | 4.0 ± 2.63 | 4.1 ± 2.93 | 3.5 ± 2.54 | 2.7 ± 1.92 | 2.9 ± 2.22 |
| family | 33 | 4.6 ± 3.22 | 4.2 ± 2.70 | 3.7 ± 2.62 | 3.5 ± 2.54 | 2.6 ± 2.14 | 3.2 ± 2.52 |
| genus | 313 | 4.7 ± 3.28 | 4.3 ± 2.68 | 3.7 ± 2.73 | 3.4 ± 2.45 | 2.6 ± 2.15 | 3.0 ± 2.42 |
| species | 1037 | 4.7 ± 3.26 | 4.2 ± 2.64 | 3.6 ± 2.64 | 3.4 ± 2.40 | 2.5 ± 1.94 | 2.8 ± 2.29 |
| *other* | 1923 | 4.6 ± 2.99 | 4.2 ± 2.47 | 3.4 ± 2.42 | 3.2 ± 2.26 | 2.2 ± 1.72 | 2.6 ± 2.07 |

**Supplemental Table S16**.Lineage-specific completeness and contamination estimates for isolate genomes from large-scale sequencing initiatives.

(see Excel file)

**Supplemental Table S17**.Completeness and contamination estimates of the *Lactobacillus gasseri* MV-22 genome for increasingly basal lineage-specific marker sets.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Lineage** | **# genomes** | **# marker genes** | **# marker sets** | **Completeness (%)** | **Contamination (%)** |
| s\_\_gasseri | 3 | 758 | 112 | 90.27 | 0.99 |
| s\_\_gasseri | 4 | 726 | 131 | 91.32 | 1.49 |
| s\_\_gasseri | 5 | 694 | 131 | 92.92 | 1.30 |
| g\_\_Lactobacillus | 9 | 630 | 137 | 93.45 | 1.06 |
| g\_\_Lactobacillus | 25 | 442 | 113 | 91.49 | 0.97 |
| g\_\_Lactobacillus\* | 58 | 451 | 129 | 90.92 | 0.90 |
| g\_\_Lactobacillus | 59 | 442 | 130 | 90.58 | 0.77 |
| g\_\_Lactobacillus | 90 | 433 | 154 | 90.76 | 1.19 |
| g\_\_Lactobacillus | 91 | 432 | 155 | 90.82 | 1.08 |
| g\_\_Lactobacillus | 92 | 429 | 157 | 91.56 | 1.06 |
| o\_\_Lactobacillales | 177 | 352 | 164 | 90.73 | 1.83 |
| o\_\_Lactobacillales | 471 | 352 | 192 | 92.61 | 2.08 |
| o\_\_Lactobacillales | 490 | 338 | 185 | 92.33 | 2.16 |
| c\_\_Bacilli | 515 | 331 | 184 | 92.29 | 2.17 |
| c\_\_Bacilli | 586 | 327 | 182 | 92.23 | 2.20 |
| c\_\_Bacilli | 725 | 279 | 151 | 90.46 | 1.99 |
| c\_\_Bacilli | 739 | 276 | 152 | 90.53 | 1.97 |
| c\_\_Bacilli | 750 | 276 | 152 | 90.53 | 1.97 |
| c\_\_Bacilli | 762 | 275 | 151 | 90.46 | 1.99 |
| c\_\_Bacilli | 764 | 275 | 151 | 90.46 | 1.99 |
| c\_\_Bacilli | 765 | 275 | 151 | 90.46 | 1.99 |
| c\_\_Bacilli | 811 | 261 | 145 | 90.07 | 1.38 |
| c\_\_Bacilli | 813 | 261 | 145 | 90.07 | 1.38 |
| c\_\_Bacilli | 814 | 261 | 145 | 90.07 | 1.38 |
| c\_\_Bacilli | 815 | 260 | 144 | 90.00 | 1.39 |
| c\_\_Bacilli | 823 | 250 | 134 | 90.00 | 1.49 |
| p\_\_Firmicutes | 830 | 245 | 130 | 89.69 | 0.77 |
| p\_\_Firmicutes | 930 | 213 | 118 | 88.64 | 0.85 |
| p\_\_Firmicutes | 1318 | 179 | 104 | 88.10 | 0.96 |
| p\_\_Firmicutes | 1324 | 176 | 102 | 87.86 | 0.98 |
| k\_\_Bacteria | 2248 | 125 | 78 | 84.34 | 1.28 |
| k\_\_Bacteria | 5415 | 105 | 59 | 81.51 | 1.69 |
| k\_\_Bacteria | 5429 | 105 | 59 | 81.51 | 1.69 |
| k\_\_Bacteria | 5443 | 105 | 59 | 81.51 | 1.69 |
| k\_\_Bacteria | 5446 | 105 | 59 | 81.51 | 1.69 |
| k\_\_Bacteria | 5449 | 104 | 58 | 81.19 | 1.72 |

\* selected lineage-specific marker set

**Supplemental Table S18**.Bacterial marker genes identified within the HMP *Lactobacillus gasseri* genomes. Markers missing from a genome or present in multiple copies are highlighted with a grey background.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Gene Annotation** | ***Lactobacillus gasseri*  MV-22** | ***Lactobacillus gasseri*  JV-V03** | ***Lactobacillus gasseri*  202-4** | ***Lactobacillus gasseri*  224-1** |
| PF02367 | Uncharacterised P-loop hydrolase UPF0079 | **2** | 1 | 1 | 1 |
| PF00162 | Phosphoglycerate kinase | 1 | 1 | 1 | 1 |
| PF00238 | Ribosomal protein L14p/L23e | 1 | 1 | 1 | 1 |
| PF00276 | Ribosomal protein L23 | 1 | 1 | 1 | 1 |
| PF00281 | Ribosomal protein L5 | 1 | 1 | 1 | 1 |
| PF00298 | Ribosomal protein L11, RNA binding domain | 1 | 1 | 1 | 1 |
| PF00312 | Ribosomal protein S15 | 1 | 1 | 1 | **0** |
| PF00366 | Ribosomal protein S17 | 1 | 1 | 1 | 1 |
| PF00380 | Ribosomal protein S9/S16 | 1 | 1 | 1 | 1 |
| PF00410 | Ribosomal protein S8 | 1 | 1 | 1 | 1 |
| PF00411 | Ribosomal protein S11 | 1 | 1 | 1 | 1 |
| PF00416 | Ribosomal protein S13/S18 | 1 | 1 | 1 | 1 |
| PF00453 | Ribosomal protein L20 | 1 | 1 | 1 | 1 |
| PF00466 | Ribosomal protein L10 | 1 | 1 | 1 | 1 |
| PF00562 | RNA polymerase Rpb2, domain 6 | 1 | 1 | 1 | 1 |
| PF00573 | Ribosomal protein L4/L1 family | 1 | 1 | 1 | 1 |
| PF00673 | ribosomal L5P family C-terminus | 1 | 1 | 1 | 1 |
| PF00687 | Ribosomal protein L1p/L10e family | 1 | 1 | 1 | 1 |
| PF01000 | RNA polymerase Rpb3/RpoA insert domain | 1 | 1 | 1 | 1 |
| PF01018 | GTP1/OBG | 1 | 1 | 1 | 1 |
| PF01121 | Dephospho-CoA kinase | 1 | 1 | 1 | 1 |
| PF01193 | RNA polymerase Rpb3/Rpb11 domain | 1 | 1 | 1 | 1 |
| PF01195 | Peptidyl-tRNA hydrolase | 1 | 1 | 1 | 1 |
| PF01196 | Ribosomal protein L17 | 1 | 1 | 1 | 1 |
| PF01250 | Ribosomal protein S6 | 1 | 1 | 1 | 1 |
| PF01281 | Ribosomal protein L9, N-terminal domain | 1 | 1 | 1 | 1 |
| PF01509 | TruB family pseudouridylate synthase | 1 | 1 | 1 | 1 |
| PF01632 | Ribosomal protein L35 | 1 | 1 | 1 | 1 |
| PF01649 | Ribosomal protein S20 | 1 | 1 | 1 | 1 |
| PF01668 | SmpB protein | 1 | 1 | 1 | 1 |
| PF01765 | Ribosome recycling factor | 1 | 1 | 1 | 1 |
| PF01795 | MraW methylase family | 1 | 1 | 1 | 1 |
| PF02033 | Ribosome-binding factor A | 1 | 1 | 1 | 1 |
| PF02130 | Uncharacterized protein family UPF0054 | 1 | 1 | 1 | 1 |
| PF03484 | tRNA synthetase B5 domain | 1 | 1 | 1 | 1 |
| PF03946 | Ribosomal protein L11, N-terminal domain | 1 | 1 | 1 | 1 |
| PF03948 | Ribosomal protein L9, C-terminal domain | 1 | 1 | 1 | 1 |
| PF04560 | RNA polymerase Rpb2, domain 7 | 1 | 1 | 1 | 1 |
| PF04561 | RNA polymerase Rpb2, domain 2 | 1 | 1 | 1 | 1 |
| PF04563 | RNA polymerase beta subunit | 1 | 1 | 1 | 1 |
| PF04565 | RNA polymerase Rpb2, domain 3 | 1 | 1 | 1 | 1 |
| PF04998 | RNA polymerase Rpb1, domain 5 | 1 | 1 | 1 | 1 |
| PF05000 | RNA polymerase Rpb1, domain 4 | 1 | 1 | 1 | 1 |
| PF05491 | Holliday junction DNA helicase ruvB C-terminus | 1 | 1 | 1 | 1 |
| PF06071 | Protein of unknown function (DUF933) | 1 | 1 | 1 | 1 |
| PF06421 | GTP-binding protein LepA C-terminus | 1 | 1 | 1 | 1 |
| PF08459 | UvrC Helix-hairpin-helix N-terminal | 1 | 1 | 1 | 1 |
| PF08529 | NusA N-terminal domain | 1 | 1 | 1 | 1 |
| PF10385 | RNA polymerase beta subunit external 1 domain | 1 | 1 | 1 | 1 |
| PF11987 | Translation-initiation factor 2 | 1 | 1 | 1 | 1 |
| PF12344 | Ultra-violet resistance protein B | 1 | 1 | 1 | 1 |
| PF13184 | NusA-like KH domain | 1 | 1 | 1 | 1 |
| TIGR00084 | Holliday junction DNA helicase RuvA | 1 | 1 | 1 | 1 |
| TIGR00250 | RNAse H domain protein, YqgF family | 1 | 1 | 1 | 1 |
| TIGR00329 | Metallohydrolase, glycoprotease/Kae1 family | 1 | 1 | 1 | 1 |
| TIGR00344 | Alanine--tRNA ligase | 1 | 1 | 1 | 1 |
| TIGR00392 | Isoleucine--tRNA ligase | 1 | 1 | 1 | 1 |
| TIGR00459 | Aspartate--tRNA ligase | 1 | 1 | 1 | 1 |
| TIGR00460 | Methionyl-tRNA formyltransferase | 1 | 1 | 1 | 1 |
| TIGR00615 | Recombination protein RecR | 1 | 1 | 1 | 1 |
| TIGR00755 | rRNA small subunit methyltransferase A | 1 | 1 | 1 | 1 |
| TIGR00810 | Preprotein translocase, SecG subunit | 1 | 1 | 1 | 1 |
| TIGR00855 | Ribosomal protein L7/L12 | 1 | 1 | 1 | 1 |
| TIGR00922 | Transcription termination factor NusG | 1 | 1 | 1 | 1 |
| TIGR00967 | Preprotein translocase, SecY subunit | 1 | 1 | 1 | 1 |
| TIGR01079 | Ribosomal protein L24 | 1 | 1 | 1 | 1 |
| TIGR02075 | UMP kinase | 1 | 1 | 1 | 1 |
| TIGR02432 | tRNA(Ile)-lysidine synthetase | 1 | 1 | 1 | 1 |
| TIGR03263 | Guanylate kinase | 1 | 1 | 1 | 1 |
| TIGR03723 | tRNA threonylcarbamoyl adenosine protein | 1 | 1 | 1 | 1 |
| PF00164 | Ribosomal protein S12/S23 | **0** | 1 | 1 | 1 |
| PF00177 | Ribosomal protein S7p/S5e | **0** | 1 | 1 | 1 |
| PF00181 | Ribosomal Proteins L2, RNA binding domain | **0** | 1 | 1 | 1 |
| PF00189 | Ribosomal protein S3, C-terminal domain | **0** | 1 | 1 | 1 |
| PF00203 | Ribosomal protein S19 | **0** | 1 | 1 | 1 |
| PF00237 | Ribosomal protein L22p/L17e | **0** | 1 | 1 | 1 |
| PF00252 | Ribosomal protein L16p/L10e | **0** | 1 | 1 | 1 |
| PF00297 | Ribosomal protein L3 | **0** | 1 | 1 | 1 |
| PF00318 | Ribosomal protein S2 | **0** | 1 | 1 | 1 |
| PF00333 | Ribosomal protein S5, N-terminal domain | **0** | 1 | 1 | 1 |
| PF00338 | Ribosomal protein S10p/S20e | **0** | 1 | 1 | 1 |
| PF00347 | Ribosomal protein L6 | **0** | 1 | 1 | 1 |
| PF00572 | Ribosomal protein L13 | **0** | 1 | 1 | 1 |
| PF00623 | RNA polymerase Rpb1, domain 2 | **0** | 1 | 1 | 1 |
| PF00828 | Ribosomal protein L18e/L15 | **0** | 1 | 1 | 1 |
| PF00829 | Ribosomal prokaryotic L21 protein | **0** | 1 | 1 | 1 |
| PF00831 | Ribosomal L29 protein | **0** | 1 | 1 | 1 |
| PF00861 | Ribosomal L18p/L5e family | **0** | 1 | 1 | 1 |
| PF00886 | Ribosomal protein S16 | **0** | 1 | 1 | 1 |
| PF00889 | Elongation factor TS | **0** | 1 | 1 | 1 |
| PF01016 | Ribosomal L27 protein | **0** | 1 | 1 | 1 |
| PF01245 | Ribosomal protein L19 | **0** | 1 | 1 | 1 |
| PF01409 | tRNA synthetases class II core domain (F) | **0** | 1 | 1 | 1 |
| PF01746 | tRNA (Guanine-1)-methyltransferase | **0** | 1 | 1 | 1 |
| PF02912 | Aminoacyl tRNA synthetase class II | **0** | 1 | 1 | 1 |
| PF02978 | Signal peptide binding domain | **0** | 1 | 1 | 1 |
| PF03719 | Ribosomal protein S5, C-terminal domain | **0** | 1 | 1 | 1 |
| PF03947 | Ribosomal Proteins L2, C-terminal domain | **0** | 1 | 1 | 1 |
| PF04983 | RNA polymerase Rpb1, domain 3 | **0** | 1 | 1 | 1 |
| PF04997 | RNA polymerase Rpb1, domain 1 | **0** | 1 | 1 | 1 |
| PF05697 | Bacterial trigger factor protein (TF) | **0** | 1 | 1 | 1 |
| PF13603 | Leucyl-tRNA synthetase, Domain 2 | **0** | **0** | **0** | **0** |
| TIGR00019 | Peptide chain release factor 1 | **0** | 1 | 1 | 1 |
| TIGR03594 | Ribosome-associated GTPase EngA | **0** | 1 | 1 | 1 |

**Supplemental Table S19**.Lineage-specific completeness and contamination estimates for genomes annotated as finished at IMG, along with predicted translation tables and calculated coding density.

(see Excel file)

**Supplemental Table S20**. Lineage-specific completeness and contamination estimates for single-cell genomes from the GEBA-MDM initiative along with traditional assembly statistics.

(see Excel file)

**Supplemental Table S21**.Lineage-specific completeness and contamination estimates for population genomes, plasmids, and phage recovered from metagenomic datasets along with traditional assembly statistics.

(see Excel file)

**Supplemental Table S22**.Completeness and contamination estimates for population genomes recovered from an acetate-amended aquifer determined using domain-level marker genes and lineage-specific marker sets.

(see Excel file)