Charting and deconvolution of microbial diversity by high-quality single-cell genomic reconstruction utilizing ResolveDNA Microbiome

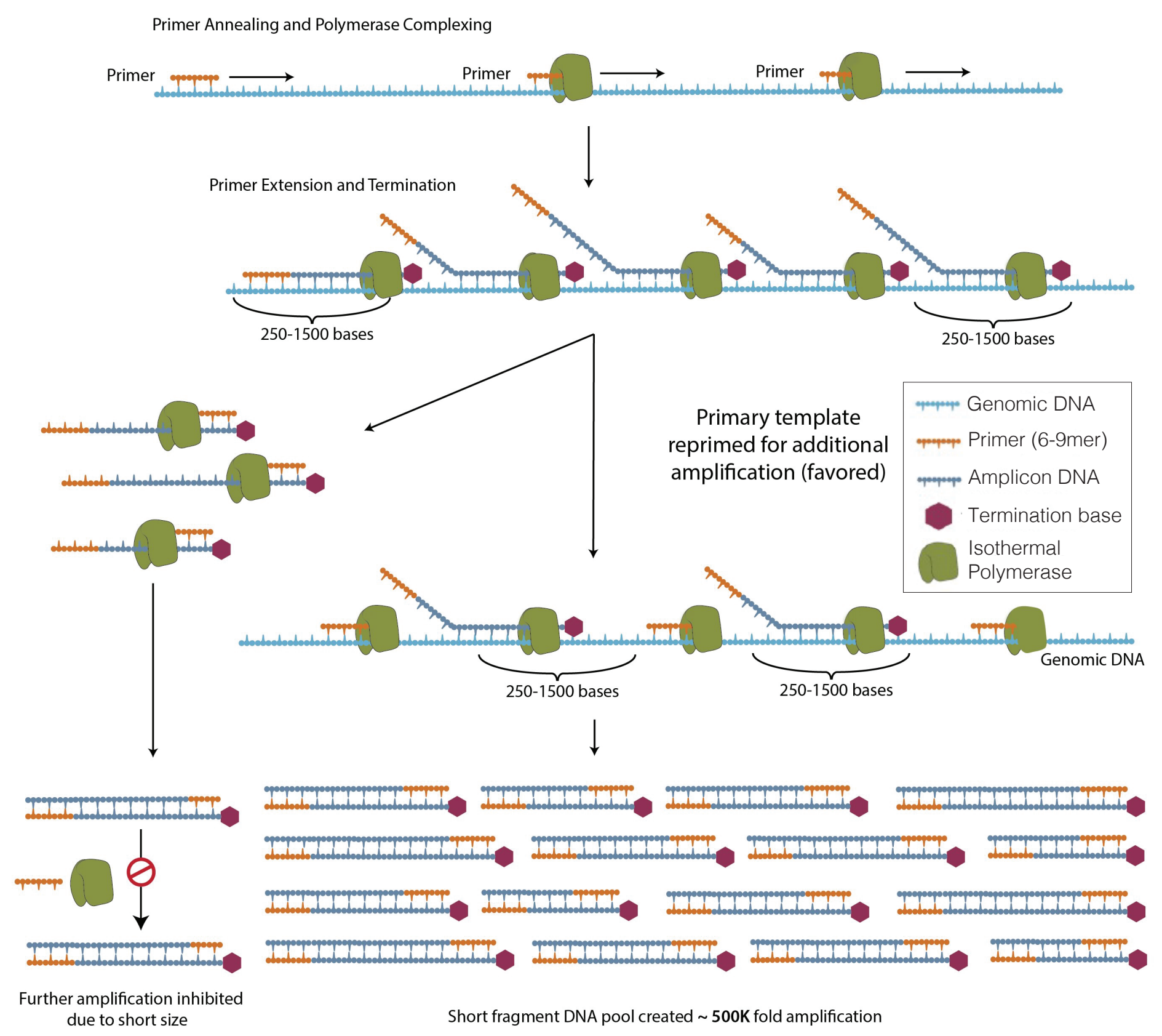
I. Salas-González[1], T. Morozova[1], A. Snediker[1], D. Arvapalli[1], N. Cira[2], J. Zawistowski[1], J. Blackinton[1], J. West[1]
[1]BioSkryb Genomics Inc., Durham, NC. [2]Cornell University., Ithaca, NY.

Abstract

The causes and consequences of microbial variation across biomes have been the subject of intense study for over a century. Specifically, bacterial niche partitioning has been shown to play fundamental roles in the structuring and maintenance of global biogeochemical cycles. Additionally, eukaryotic organisms are inhabited by complex microbial communities, collectively named the microbiota, that have been shown to modulate diverse host physiological traits ultimately impacting the host fitness. A bottleneck to unraveling the vast uncharted bacterial diversity across biomes has been the lack of robust technology to faithfully amplify the femtogram quantities of DNA of a single bacterium. Here we present ResolveDNA , a whole-genome amplification technology, to reconstruct high-quality genomes.

We benchmarked the ability of our protocol to yield high quality reconstructed bacterial genomes by performing whole genome sequencing of a FACS-sorted co-culture experiment between the gram-positive bacteria *B. subtillis* and the gram-negative bacteria *E. coli*. Following sorting, library preparation and sequencing, we filtered low quality reads and performed de novo bacterial assembly followed up by genome deconvolution using differential depth coverage across the assembly contigs. Evaluation of the quality of the assemblies was performed using an unbiased phylogenetic single-copy marker approach whichvalidated that the reconstructed assemblies in our experiment had levels of completeness over 95% and levels of contamination lower than 1%. Finally, we developed a novel computational pipeline, that leverages random fragments of contaminant DNA within an amplified reaction, to estimate that the high-quality assemblies from our experiment were derived from samples with different levels of cells within each reaction, ranging from 1 cell up to 5 cells. The data presented here demonstrates that ResolveDNA can be used to assemble bacterial genomes, at different levels of phylogenetic divergence and cell quantity, with high reliability thus permitting to deconvolute microbial communities into their original constitutes with unseen quality.

ResolveDNA™



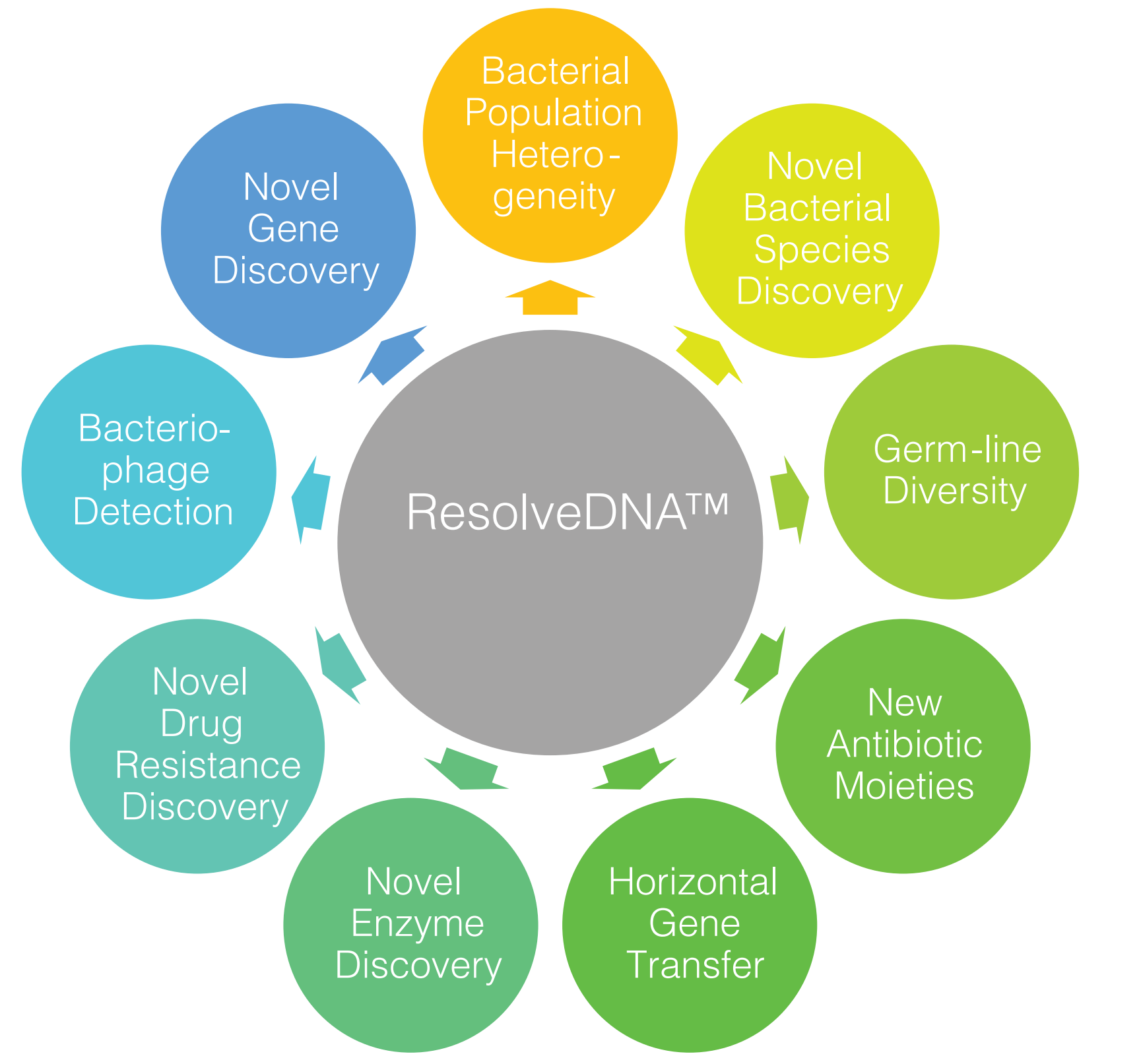
Further amplification inhibited due to short size

Short fragment DNA pool created ~ **500K** fold amplification

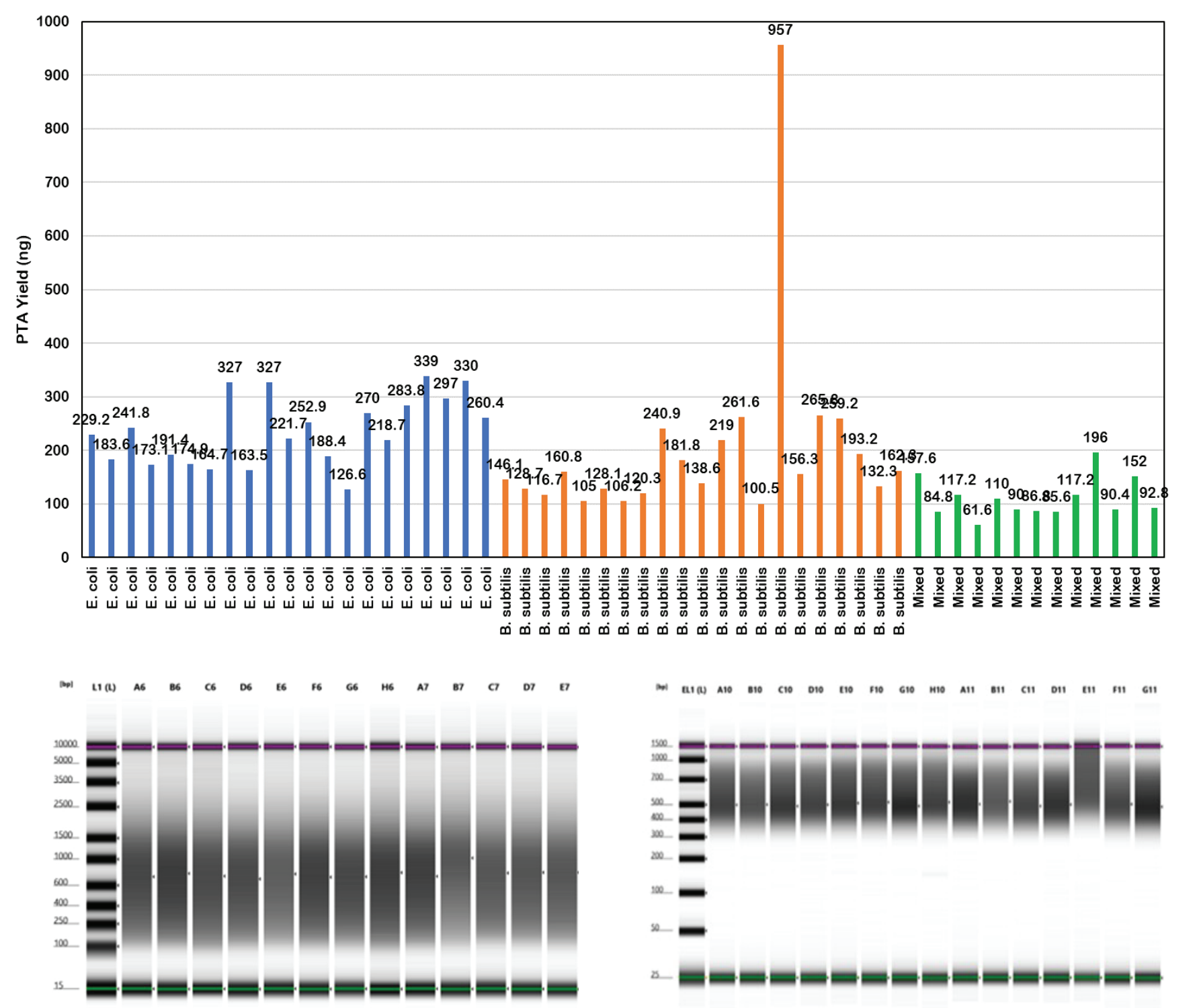
Methods

Single Bacteria Isolation
Three sample types were isolated 1) gram negative bacteria (*E.coli*), 2) gram positive bacteria (*B. subtilis*), and 3) a mixture of these species, to evaluate the performance of the ResolveDNA™ Microbiome WGA Kit. Bacterial stocks of each species were grown in LB broth for 18 hours at 37C. Afterwards, cultures were filtered through a 20 μm mesh filter to remove large cell clusters and counted by O.D. measurement, then an aliquot of each sample was mixed to create the mixed sample at a 1:1 cell ratio. Samples were centrifuged, resuspended in 1X dPBS that was 0.2 μm filter-sterilized.
Single-cell bacteria were sorted
into 96-well plates containing 1 μL ResolveDNA Cell Buffer using a Sony SH800 sorter equipped with a 130um sorting chip. Sorted plates were briefly vortexed and flash frozen on dry ice, followed by -80C storage until ready to perform PTA DNA Amp
lification with the ResolveDNA™ Microbiome WGA Kit.
The ResolveDNA Microbiome protocol was followed, resulting in amplified bacterial DNA. This DNA was then purified with the ResolveDNA™ bead purification kit. We found individual bacteria typically yielded ~1 ug of amplified DNA and had an average size range of 900 bp. Purified, amplified DNA was then transformed into sequencing libraries using the ResolveDNA™ Library Preparation kit. Sequencing libraries were again purified and analyzed by TapeStation 4200 (Agilent) which demonstrated optimal size of PTA amplified single bacterial genome libraries (~500bp). These libraries were then sequenced using the Illumina MiniSeq Platform at 2 million paired end reads per library. Raw sequencing data were analyzed for quality, contig assembly, and alignment.

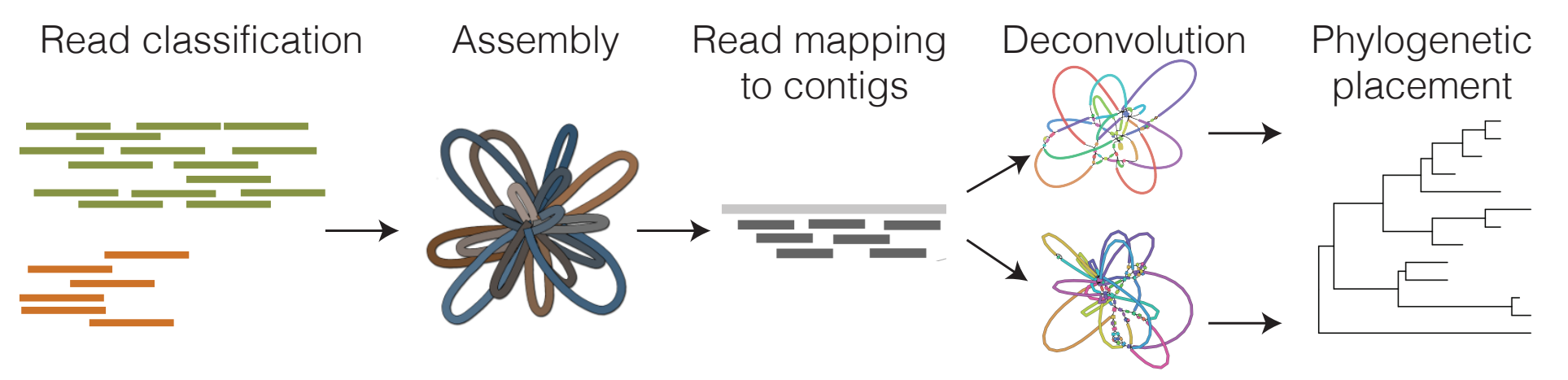
ResolveDNA™ Bacterial Applications



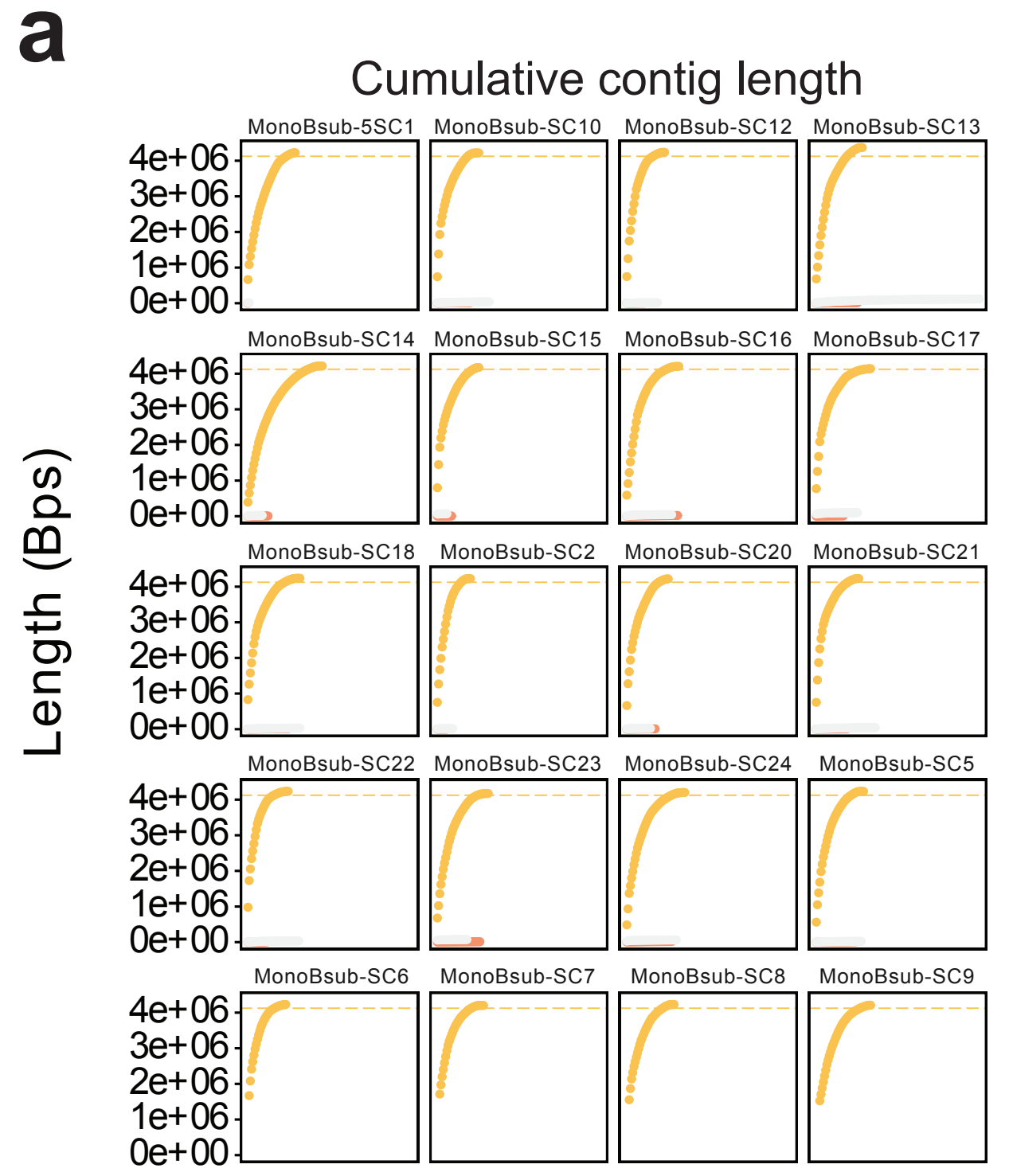
The truest diversity of bacterial species are only characterized at a small fraction of the population. Yet the diversity of bacteria influence and shape the environment and significantly impacts human health.

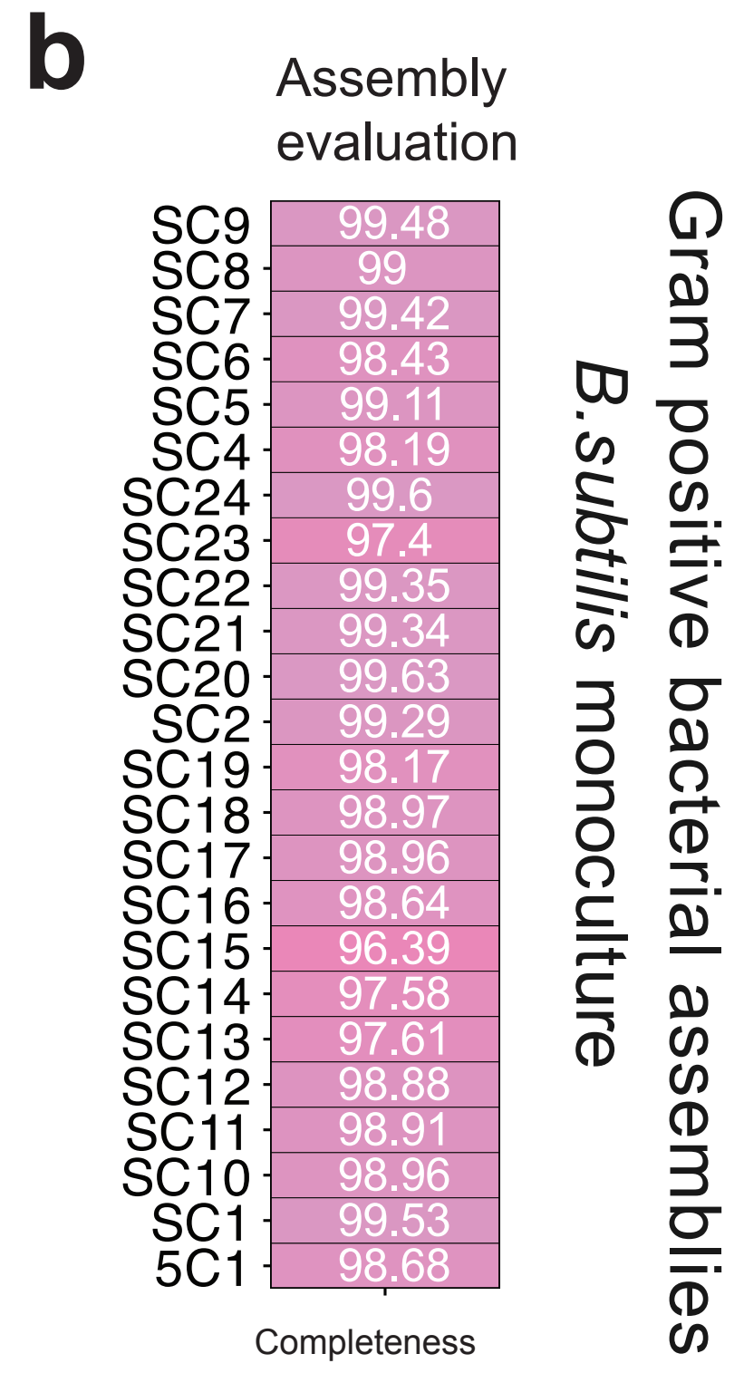


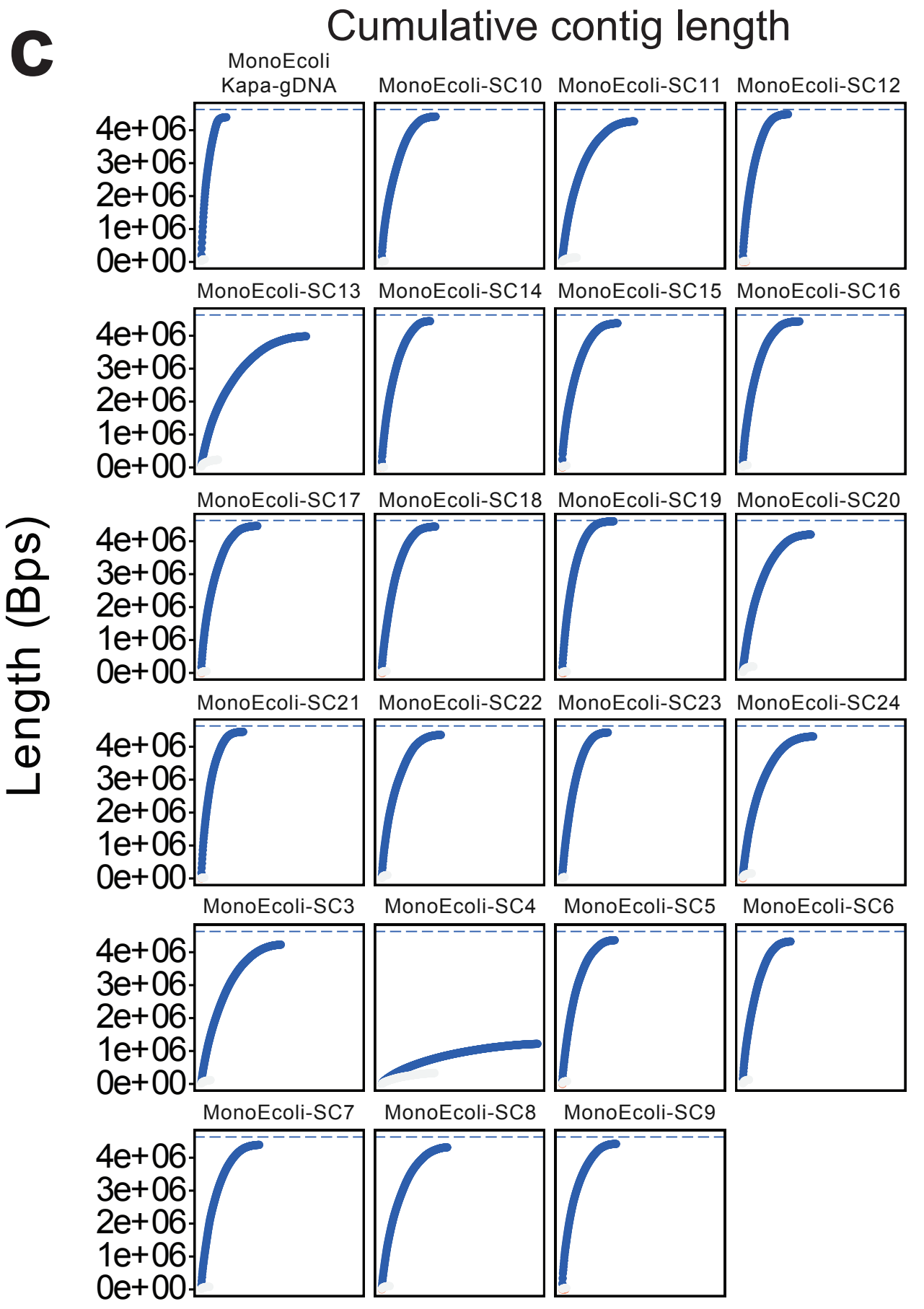
Here we demonstrated single bacterial cells generating 15 ng of amplified genomic DNA(A). Amplification products were subjected to D500 TapeStation analysis for determination of size and range of the amplicons. The range of amplification products centered ~800 bp, while the overall range appeared to extend from 100 to ~1500 BP, slightly smaller than observed for eucaryotic cells. Amplified DNA (~100ng) were then converted to sequencing libraries using BioSkryb's ResolveDNA™ library preparation kit. Insert size was then verified by TapeStation using the D1000 tape cassette. Libraries were then subjected to NGS analysis using the Illumina MiniSeq 300 cycle kit, generating 150 Bp, paired end reads.

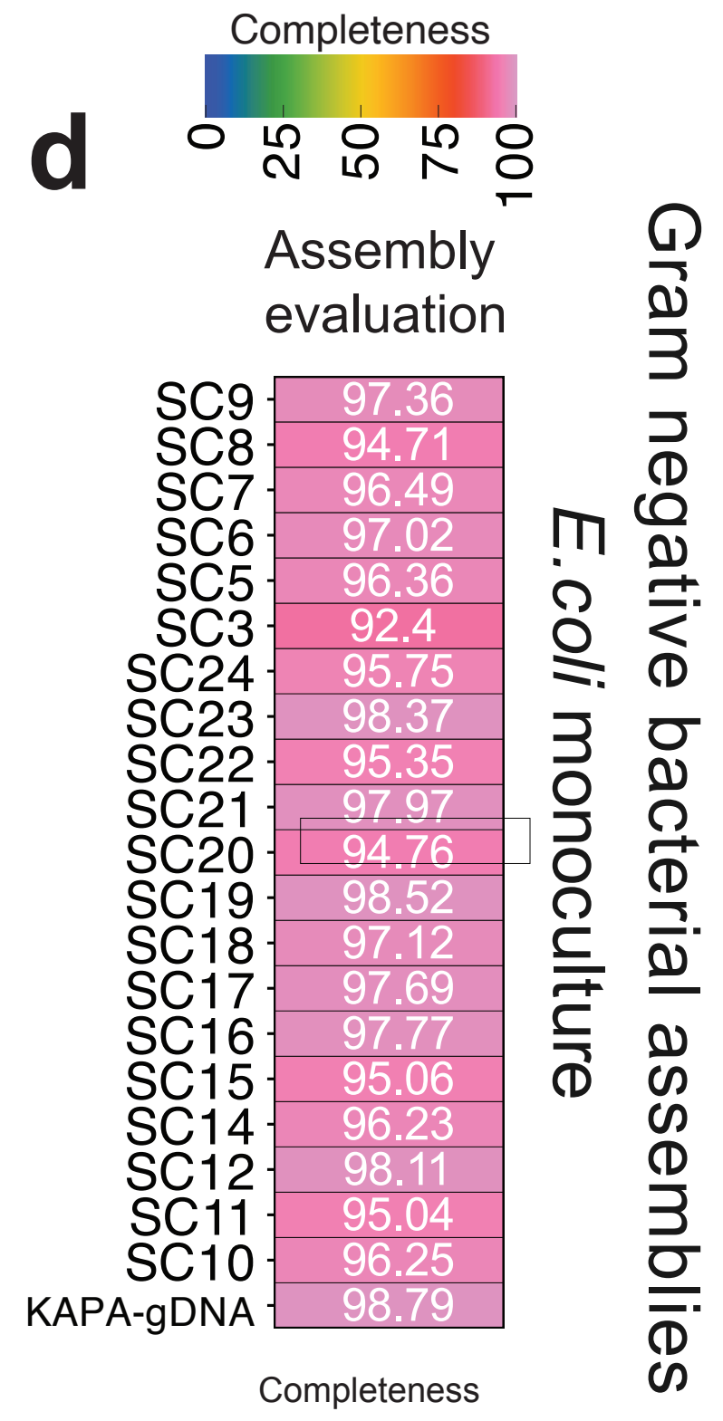


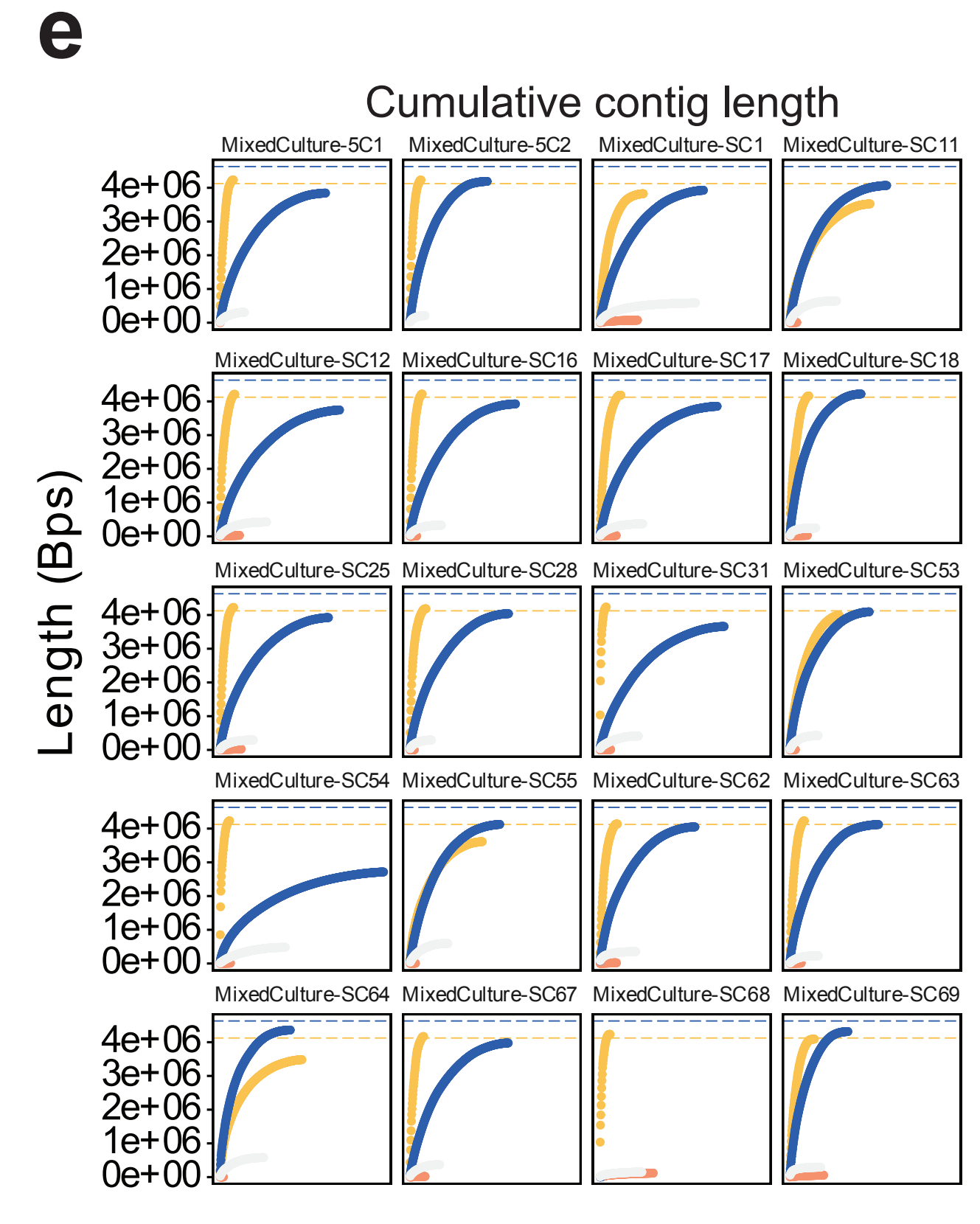
We employ state-of-the-art computational approaches to transform raw-data into high-quality genomes for which genomic mining can be performed. Briefly, raw sequencing reads are filtered and processed using Fastp followed up by an optimized iterative genome assembly using Spades. Raw reads are mapped back into the obtained optimized assembly using BWA-Mem and then deconvoluted using the alignment information via Metabat2. In parallel, read-level taxonomic assigned is performed using Kaiju. Finally, full assembly and deconvoluted bins are taxonomically placed in the GTDB reference tree and evaluated for completeness using Checkm and GTDB-tk.

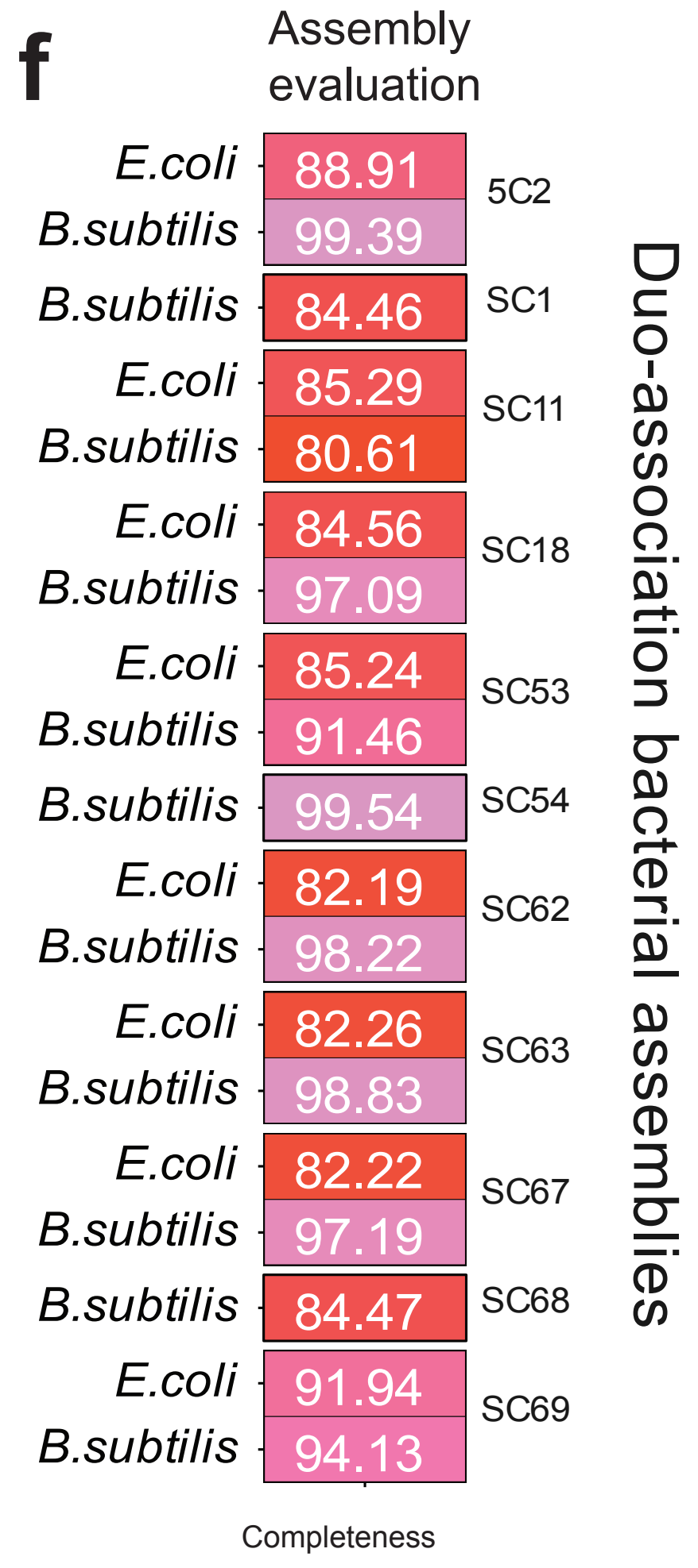


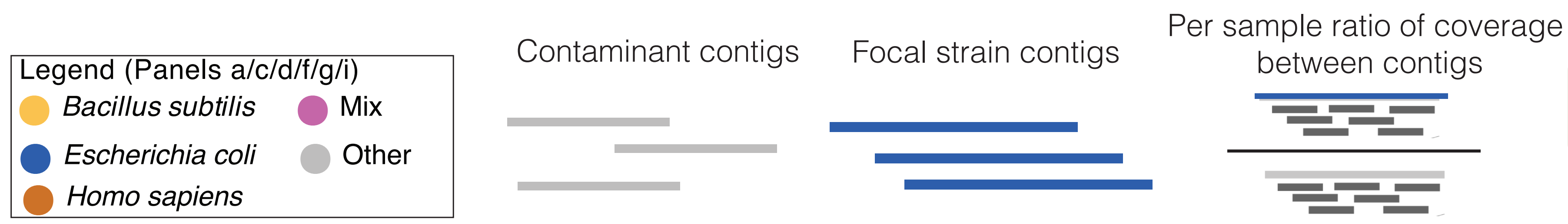


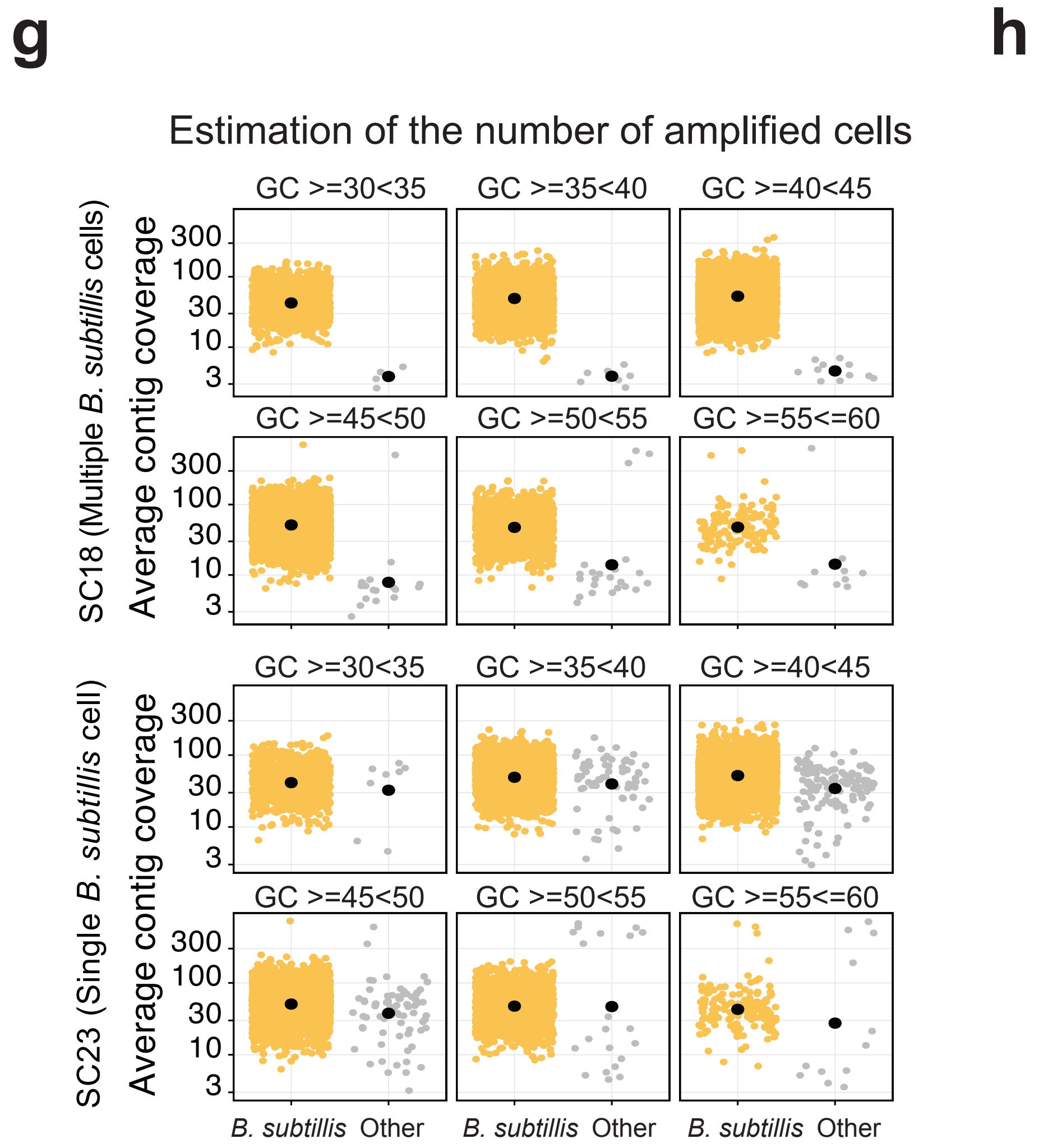


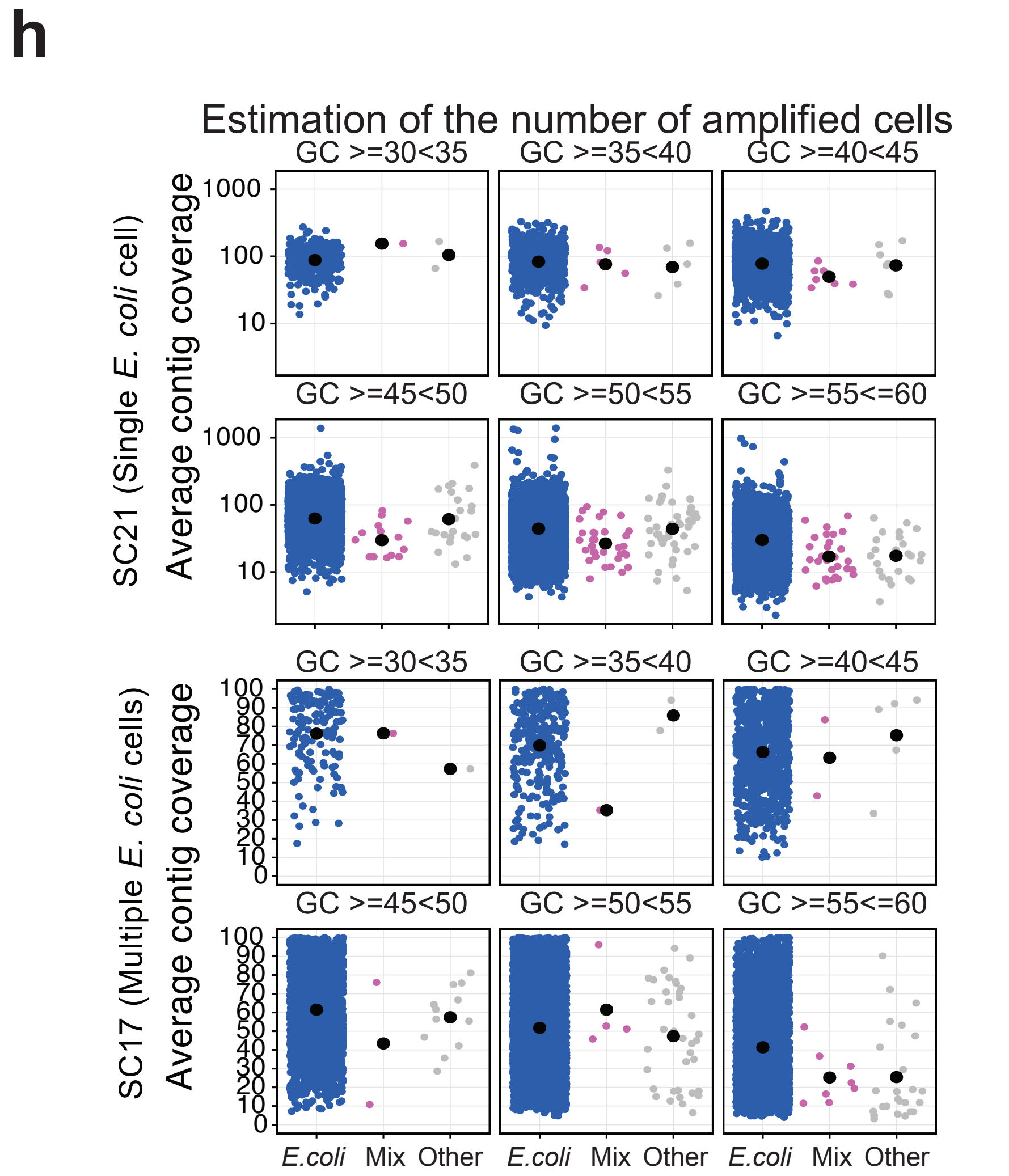


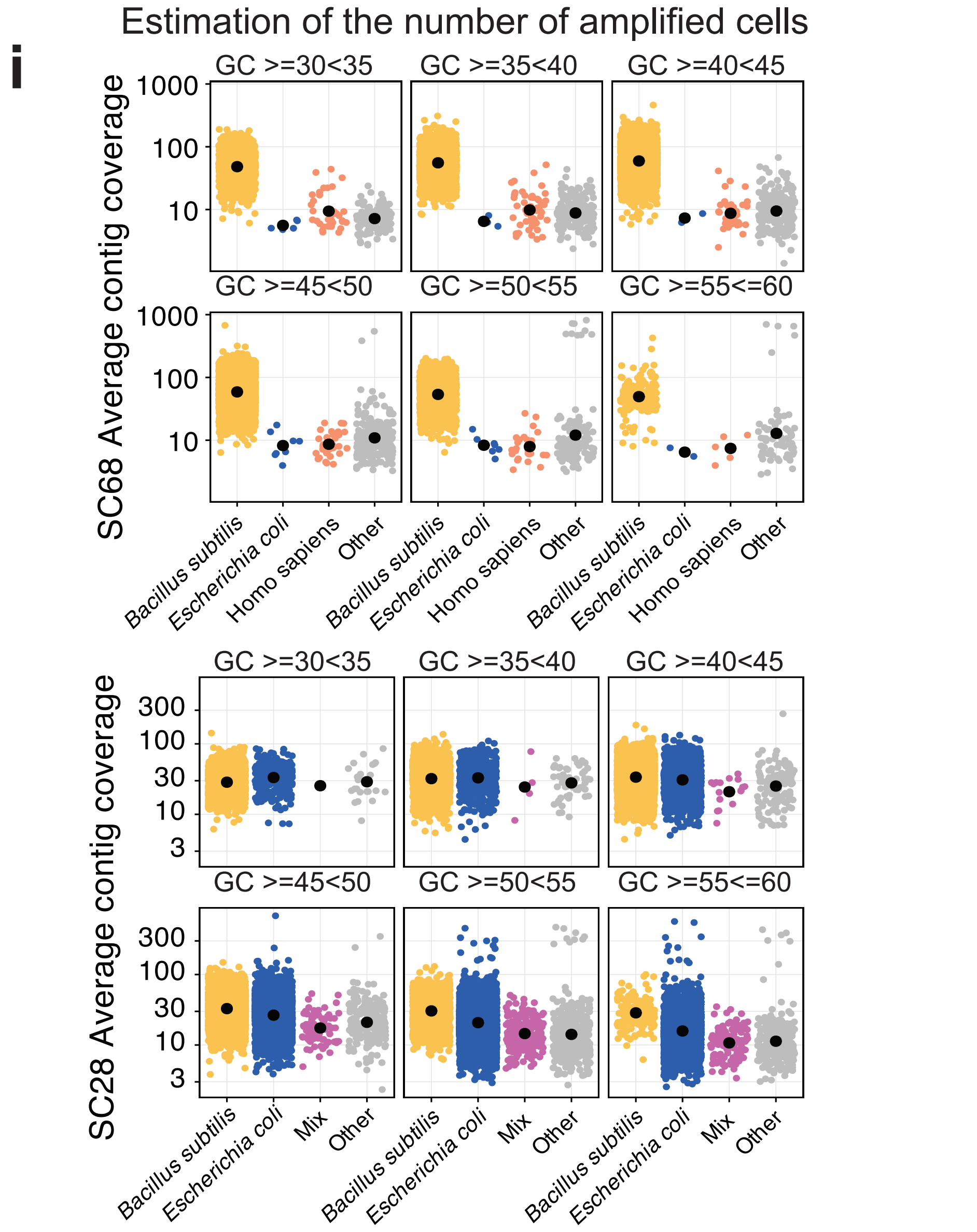












Results

The results from this work demonstrate the feasibility of PTA to recover high quality single-cell bacterial genomes (SAGs) with a high level of completeness (>95% completeness based on the presence of taxon specific single-copy markers).
Analysis of single bacteria demonstrated the ability to detect nearly the entire bacterial genome from each individual cell. We found that the SAGs obtained, from both bacteria species tested, approached the empirically estimated genome size (Panels a,d). For this analysis approximately 20 samples of each bacteria were sequenced (12 representative shown) (We noted minor differences in coverage where single cell genome of *E. coli* typically covered ~90-95% of the genome, where single B. subtilis detected up to 100% of the empirically estimated genome size. The data further detected minor components of other (un-identified) contigs, however these are considered to be minor contaminants, which based on the data do not affect the ability to identify the species/-genus of origin (Panels a,d).
Having sequenced these individual bacteria, a third set of samples was prepared where a combination of both the *E. coli* and *B. subtilis* were sorted into wells (Panels g,h,i). Again, greater than 20 samples were processed (12 representative samples shown). We found both bacterial species were detectable, with cumulative contigs lengths that approximated the empirically estimated genome sizes for each species. In many cases, it appears we dispensed more than one cell into each well, as both *E. coli* and *B. subtilis* genomes were detected, however, in the last panel only *B. subtilis* was detected, suggesting assay specificity. Trace amounts of human DNA and unidentified species DNA was detected in these samples, however, the short contigs make these reads easy to identify and remove (Panel g).
In addition, we employed an unbiased *de novo* phylogenetic approach that places assemblies in a representative bacterial phylogeny and estimates completeness of an assembly using taxon-specific single copy markers (Panels b,e,h). This second approach quantitatively validated the high-quality nature of the SAGs observed employing the cumulative contig length approach.
Finally, we developed a computational approach that utilized contaminant (Other, Homo sapiends) DNA material within a reaction to estimate the total number of focal cells within a reaction using the ratio of coverage between focal strain contigs and contaming DNA contigs (Panels c,f,i). This analysis allowed us to estimate that our assemblies were obtained from reactions containing a range of cells between 1-6 cells.

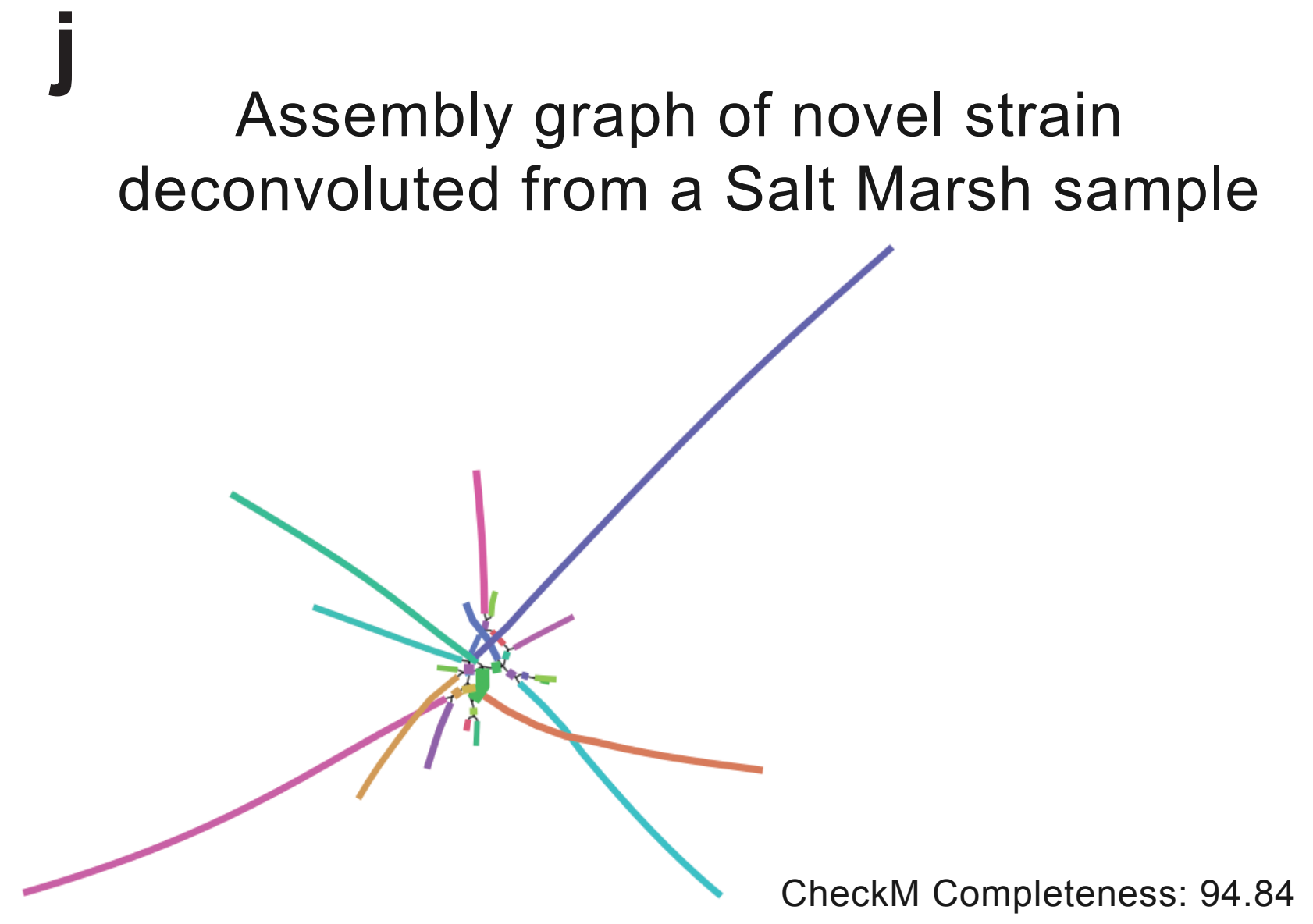
Legends

Panels (a,c,e): Contig cumulative length plots showing the number of contigs presents in each sequenced well. Colors of the contig denote their taxonomic classification based on taxonomic read assignation. Dashed horizontal lines in the plot denote the empirically estimated genome size of the bacterial species in the experiment based on representative complete genomes downloaded from NCBI RefSeq. Note that the levels of contamination per well is minimal.
Panels (b,d,f): Heatmaps showing the estimated level of completeness for the assembled genomes using a single copy marker phylogenomic approach.

Legends

Panels (g,h,i): We developed a computational pipeline that leverages read-level taxonomic assignation and contig-coverage to estimate the number of focal cells amplified within a reaction. Briefly, we assigned a taxonomic classification to each assembled contig based on read-level taxonomic assignment. Next, we estimate coverage across all contigs in the dataset and perform a ratio of coverages between focal contigs (E.g *E. coli*) against contaminant contigs (assuming they are present at 1 copy).

Assembly graph of novel strain deconvoluted from a Salt Marsh sample



CheckM Completeness: 94.84
Number of bins: 344
Largest contig: 100916 bps
N50: 27,787

Legends

Panel J: Assembly graph of the reconstructed genome for an uncharacterized strain isolated from Salt marsh utilizing BioSkryb's ResolveDNA.

The novel discovered strain is placed in a genus belonging to the family Cyclobacteriaceae.Note the overall completeness and contiguity (Number of bins) of the obtained .assembly

Acknowledgments

The BioSkryb team would like to acknowledge the efforts of these key people and groups in the execution of this work.
- Charles Gawad and the Gawad Laboratory at Stanford University for the collaboration for this work.
- The BioLabs organization for their support of BioSkryb and the use of common laboratories in the conduction of the sample preparation and sequencing of the samples.
- BioSkryb would like to thank Angela Jones at the VANTAGE Group for her collaborative efforts for deep sequencing services.

References

1)Hamidreza Chitsaz & Roger S. Lasken etal. De novo assembly of bacterial genomes from single cells. Nat Biotechnol. ; 29(10): 915–921. doi:10.1038/nbt.
2) Zhixin Ma, Pan M. Chu, Yingtong Su, Yue Yu, Hui Wen, Xiongfei Fu, Shugiang Huang. Applications of single-cell technology on bacterial analysis. Quantitative Biology 2019, 73): 171–181
3) Ozbudak E. M., Thattai M., Kurtser I., Grossman A. D. and van Oudenaarden A. (2002) Regulation of noise in the expression of a single gene. Nat. Genet., 31, 69–73.
4) Rosenfeld N., Young J.W., Alon U., Swain P. S. and Elowitz M. B. (2005) Gene regulation at the single-cell level. Science, 307, 1962– 1965 5) Wang P., Robert L., Pelletier J., Dang W. L., Taddei F., Wright A. and Jun S. (2010) Robust growth of Escherichia coli. Curr. Biol., 20, 1099–1103
6) Robert L., Ollion J.,Robert J., Song X., Matic I. and Elez M. (2018) Mutation dynamics and fitness effects followed in single cells. Science, 359, 1283–1286
7)Jones D. L., Leroy P., Unoson C., Fange D., Ćurić V., Lawson M. J. and Elf J. (2017) Kinetics of dCas9 targetsearch in Escherichia coli. Science, 357, 1420–1424