



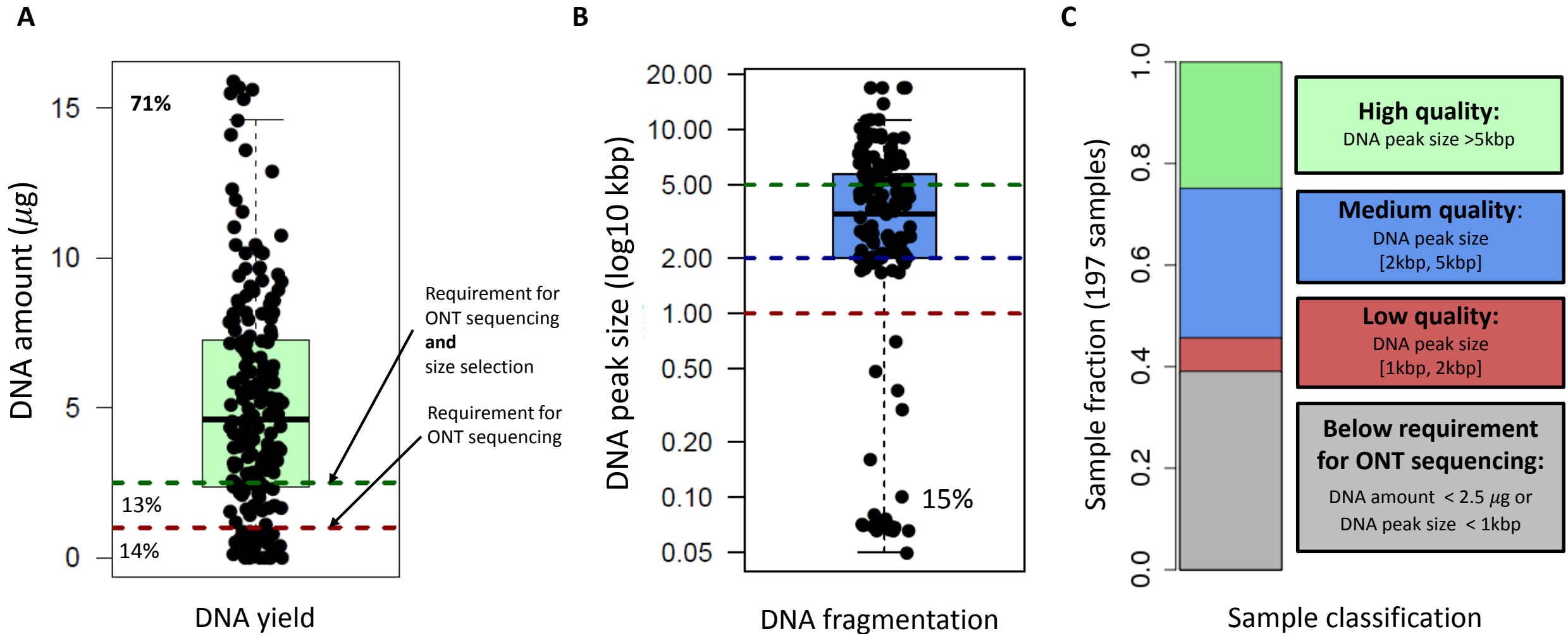


In the format provided by the authors and unedited.

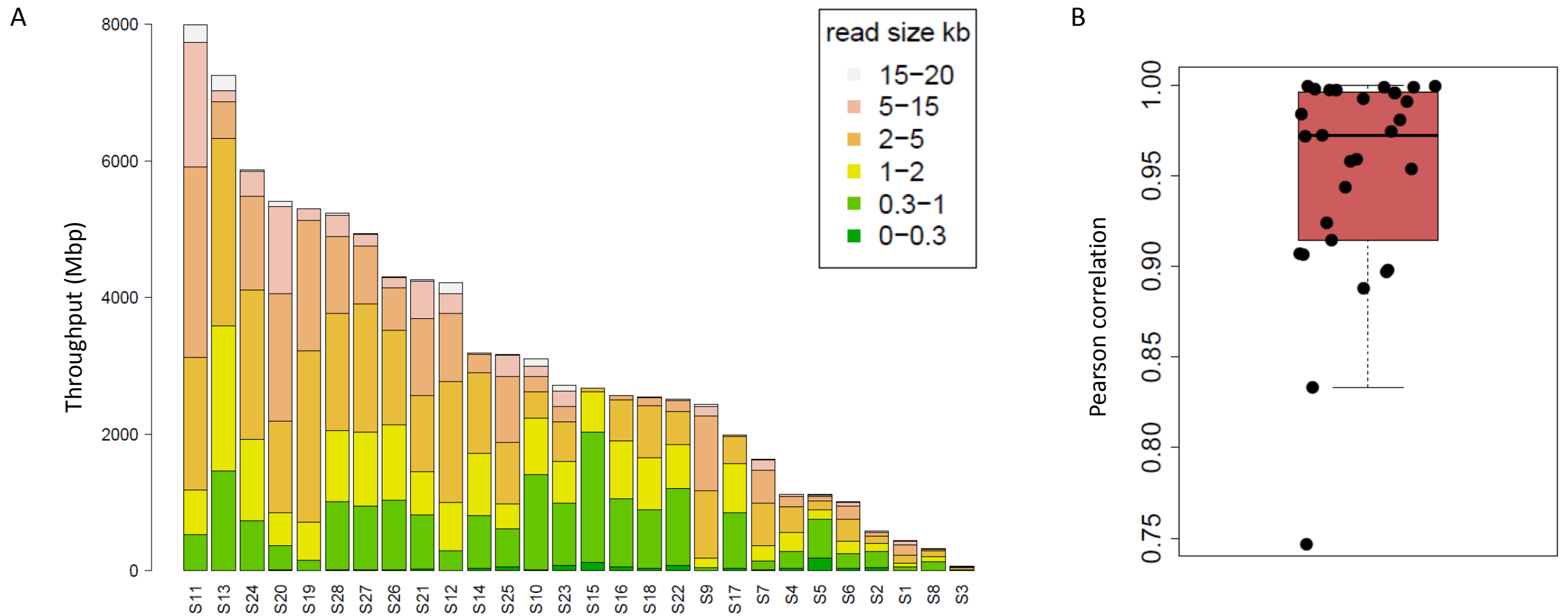
Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes

Denis Bertrand¹, Jim Shaw¹, Manesh Kalathiyappan¹, Amanda Hui Qi Ng¹, M. Senthil Kumar¹, Chenhao Li ¹, Mirta Dvornicic^{1,2}, Janja Paliska Soldo¹, Jia Yu Koh¹, Chengxuan Tong¹, Oon Tek Ng³, Timothy Barkham ⁴, Barnaby Young^{3,5}, Kalisvar Marimuthu^{6,7}, Kern Rei Chng¹, Mile Sikic ² and Niranjan Nagarajan ^{1,7*}

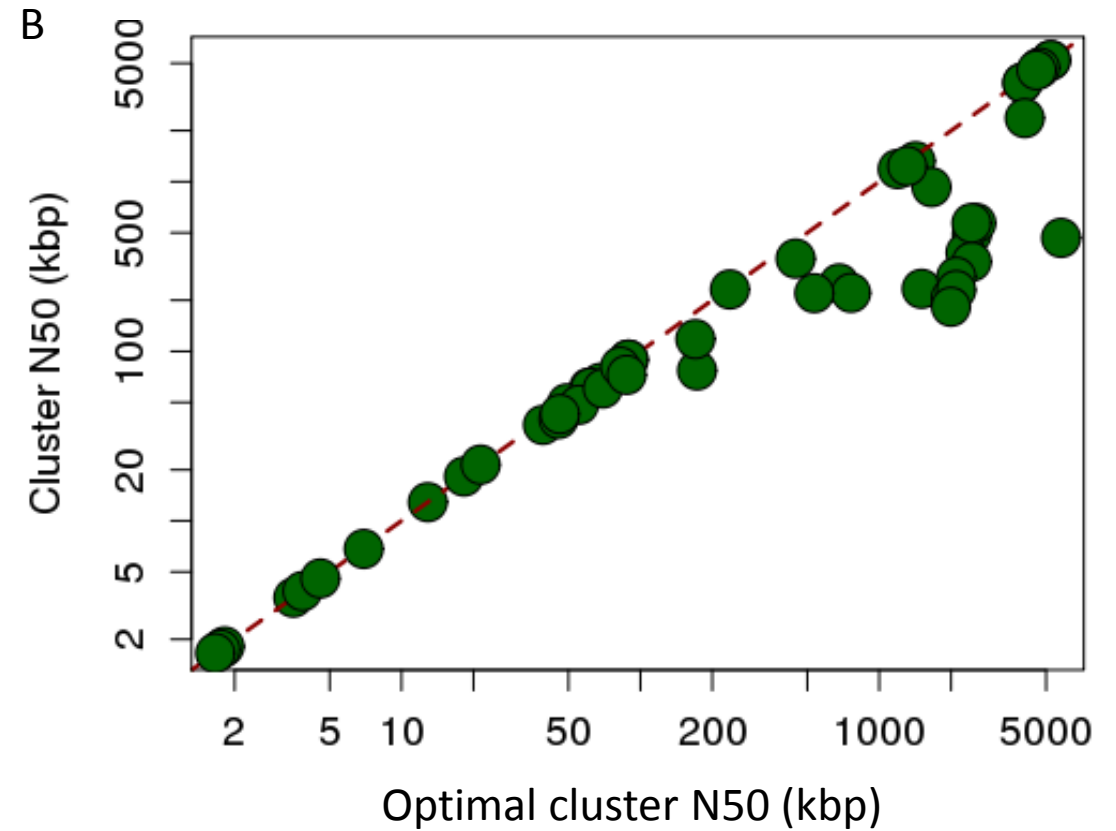
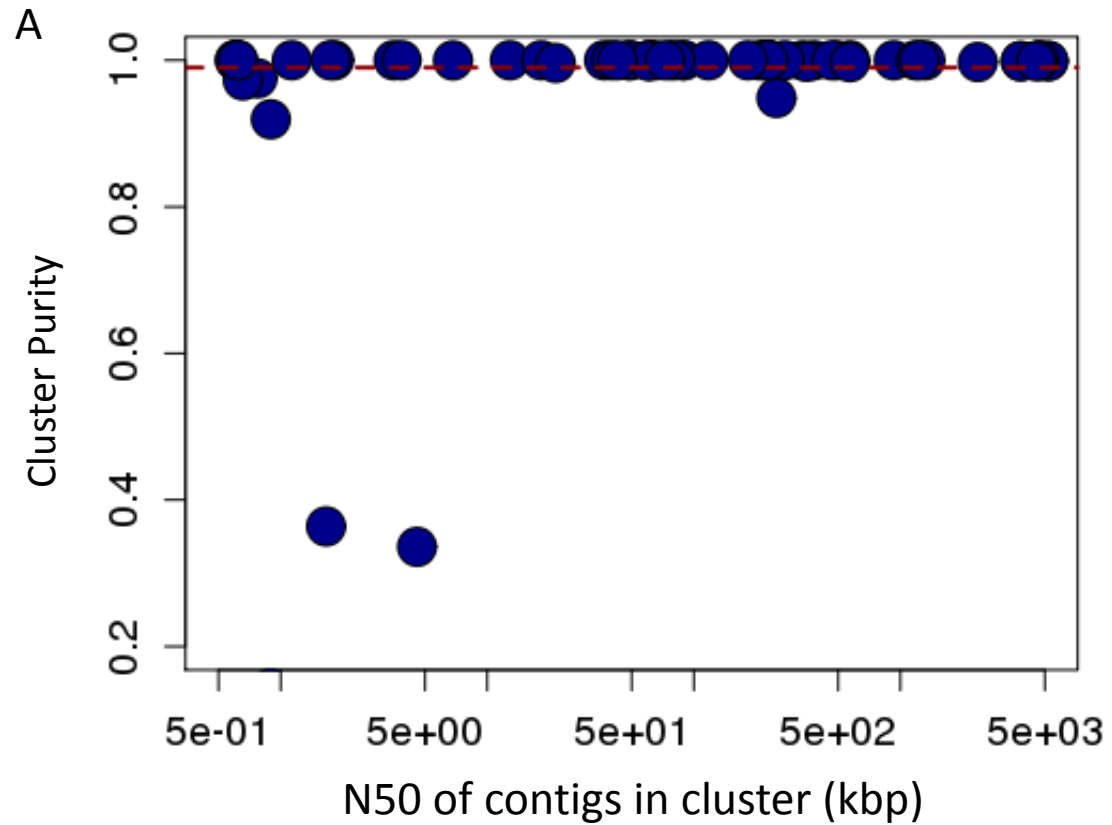
¹Computational & Systems Biology, Genome Institute of Singapore, Singapore, Singapore. ²Faculty of Electrical Engineering and Computing, Department of Electronic Systems and Information Processing, University of Zagreb, Zagreb, Croatia. ³National Centre for Infectious Disease, Tan Tock Seng Hospital, Singapore, Singapore. ⁴Department of Laboratory Medicine, Tan Tock Seng Hospital, Singapore, Singapore. ⁵Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore. ⁶Institute of Infectious Diseases and Epidemiology, Tan Tock Seng Hospital, Singapore, Singapore. ⁷National University of Singapore, Singapore, Singapore. *e-mail: nagarajann@gis.a-star.edu.sg



Supplementary Figure 1: High quality DNA for ONT sequencing extracted from human stool samples ($n=197$). Distribution of (A) DNA yield and (B) DNA fragmentation across samples. Data in B and C is presented as a boxplot (center line, median; box limits, upper and lower quartiles; whiskers, $1.5\times$ interquartile range; points, outliers). (C) classification of samples according to DNA quality. In general, we found that $>60\%$ of samples could be used for ONT sequencing despite working with limited stool quantities.

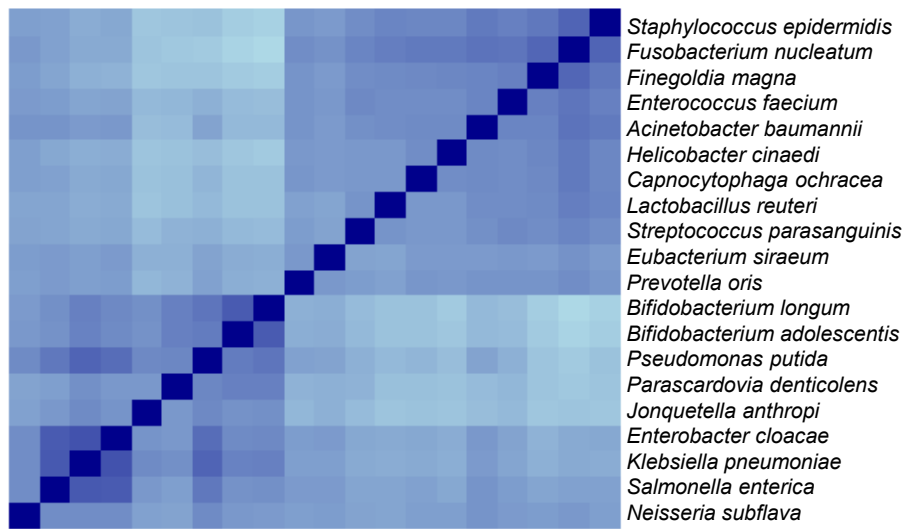


Supplementary Figure 2: Characteristics of long reads for stool metagenomics using ONT sequencing. (A) Throughput as a function of read length across different samples. Reads longer than 1kbp typically provide >80% of the data and in some libraries a sizeable fraction of reads were longer than 5kbp. (B) Correlation between species-level relative abundances using Illumina and ONT data (n=28 samples). The high concordance seen suggests that ONT sequencing does not introduce a significant bias towards some species compared to Illumina sequencing. Abundance profiles for both short and long reads were obtained using Kraken (v0.10.5-beta, default parameters; species with abundance <0.1% were ignored to avoid the introduction of false positives into the correlation analysis). Data is presented as boxplots (center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers).

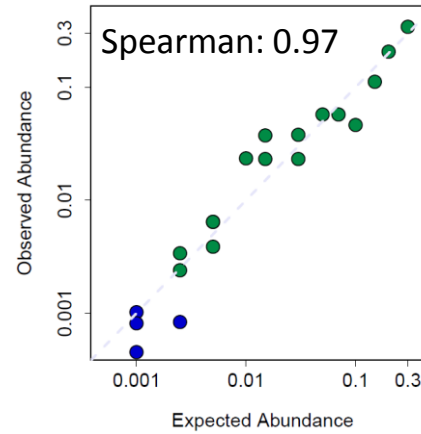


Supplementary Figure 3: Evaluation of OPERA-MS clusters. (A) Purity of clusters (fraction of sequence in cluster that comes from the dominant genome) versus size of contigs in cluster (N50 = size such that >50% of the sequence in the cluster is in longer contigs). Note that for most genomes cluster purity is >99% (dotted red line), except in the case of two genomes from the HMP Mock community with relative abundances <0.01%. (B) Cluster N50 compared to an optimal cluster N50 (when all connected contigs belonging to a genome are in one cluster). Despite providing conservative clusters, OPERA-MS clusters have N50 that is close to optimal values for most genomes.

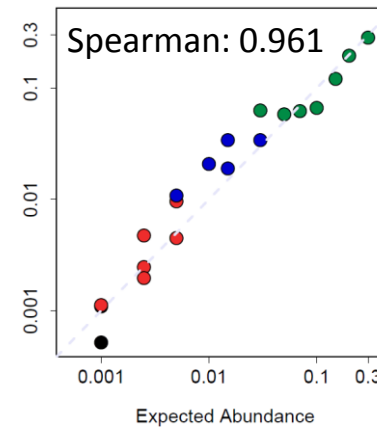
GIS20



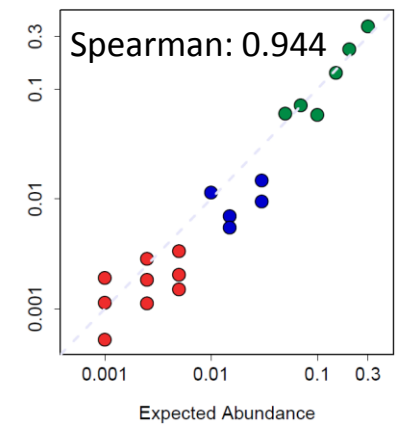
Illumina



ONT



Pacbio

Mash genomic distance

0 0.2 0.4

Species Coverage

- >30x
- 5 to 30x
- 1 to 5x
- <1x

HMP mock
staggered

Mash genomic distance

0 0.2 0.4

- Escherichia coli*
- Neisseria meningitidis*
- Acinetobacter baumannii*
- Helicobacter pylori*
- Bacteroides vulgatus*
- Rhodobacter sphaeroides*
- Pseudomonas aeruginosa*
- Deinococcus radiodurans*
- Propionibacterium acnes*
- Actinomyces odontolyticus*
- Listeria monocytogenes*
- Enterococcus faecalis*
- Staphylococcus epidermidis*
- Staphylococcus aureus*
- Bacillus cereus*
- Streptococcus pneumoniae*
- Streptococcus agalactiae*
- Streptococcus mutans*
- Lactobacillus gasserii*
- Clostridium beijerinckii*

Illumina

Spearman: 0.967

Observed Abundance

Expected Abundance

Illumina Synthetic

Spearman: 0.894

Observed Abundance

Expected Abundance

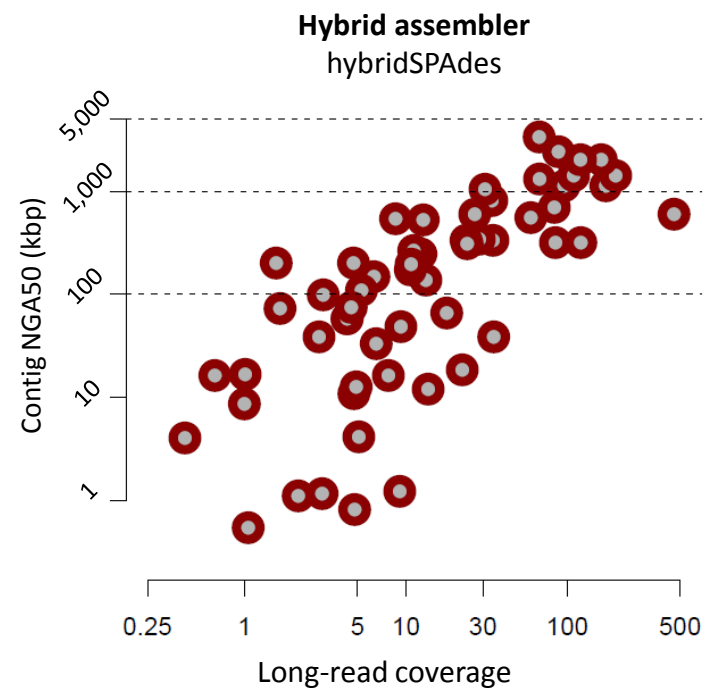
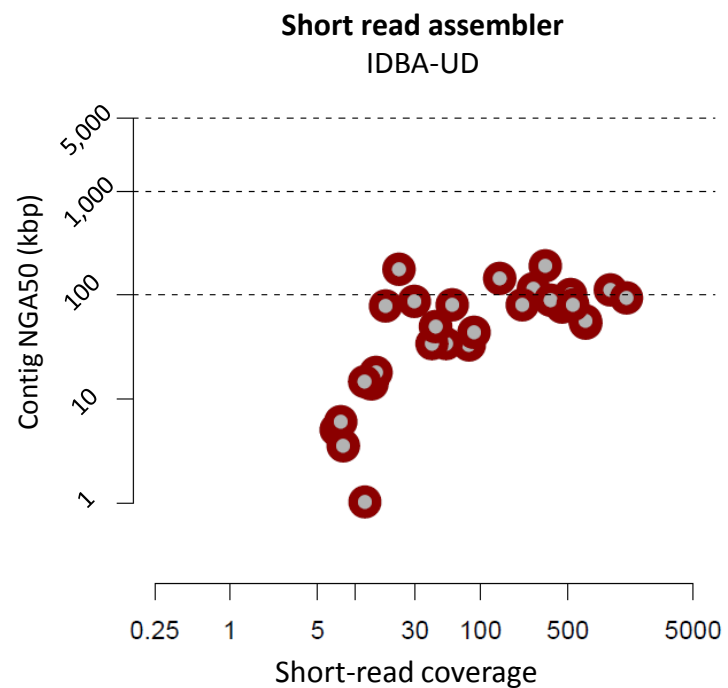
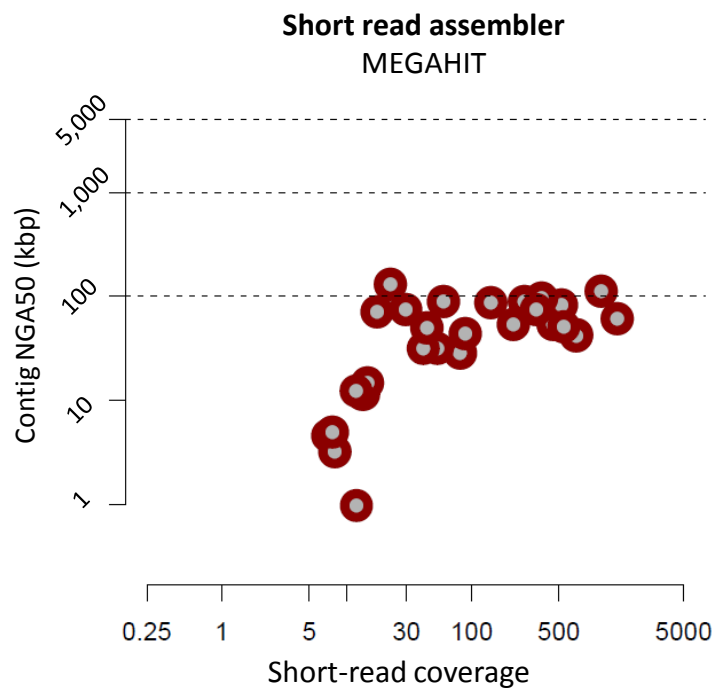
Pacbio

Spearman: 0.967

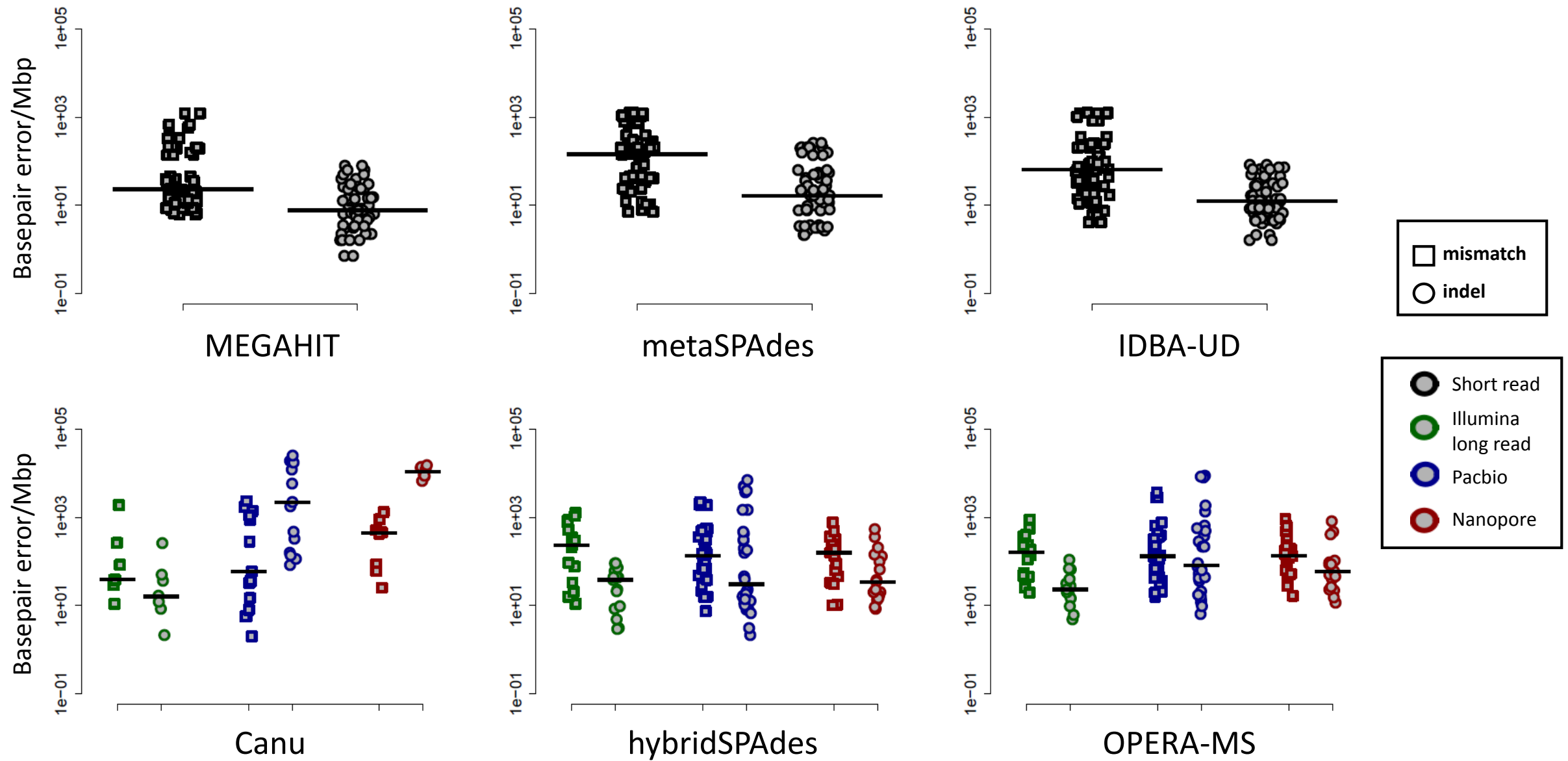
Observed Abundance

Expected Abundance

Supplementary Figure 4: Characteristics of the mock communities and datasets used in this study. Heatmaps on the left show genomic distances between the species ($n=20$ for both datasets) that make up the community. Plots on the right show correlation between observed and expected species abundances using various sequencing technologies. Read coverage of the respective genomes were obtained by mapping short reads (BWA-MEM v0.7.10-r789, default parameters) and long reads (GraphMap v0.2.2, default parameters) to the reference genomes.

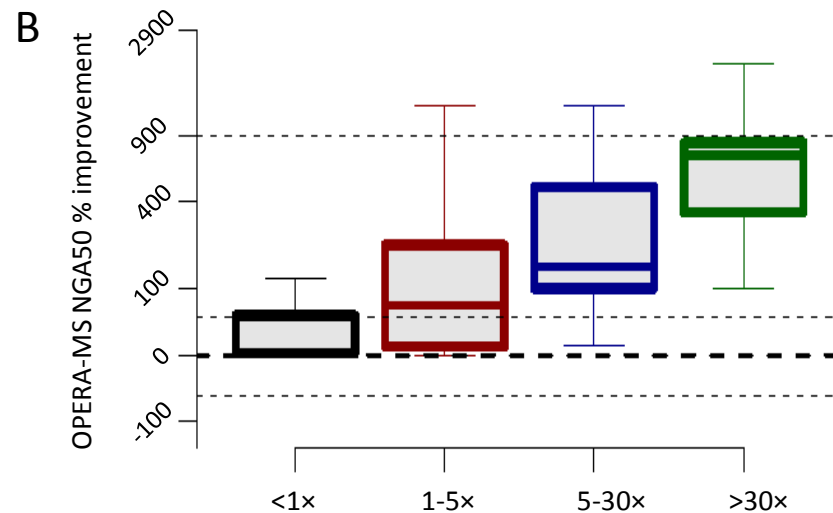
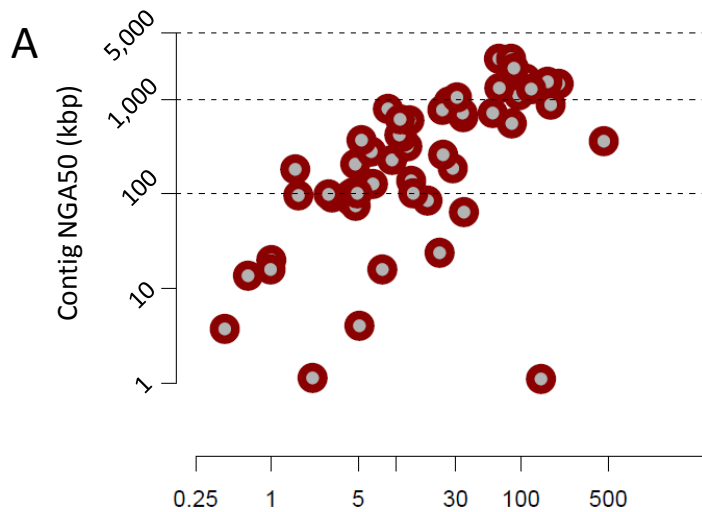


Supplementary Figure 5: Correlation between read coverage and contig NGA50 for the genomes assembled by MEGAHIT (n=37), IDBA-UD (n=37) and hybridSPAdes (n=74).

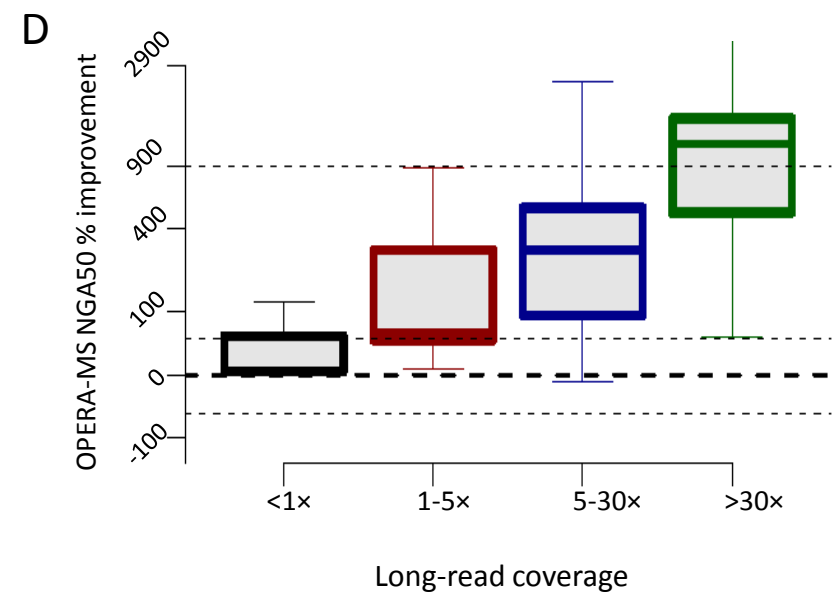
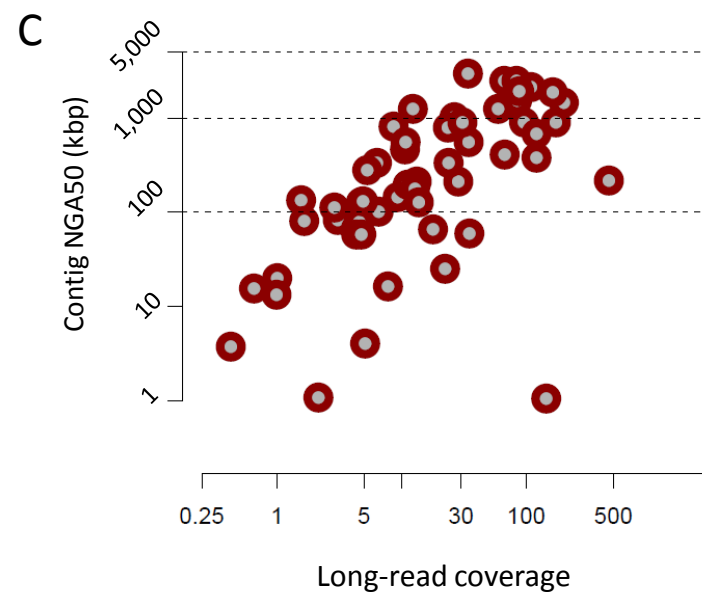


Supplementary Figure 6: Indel and mismatch error rates for assemblers as a function of the different sequencing technologies used. Note that each point represents performance for a genome in the mock community datasets used, and that the centre value represents the median.

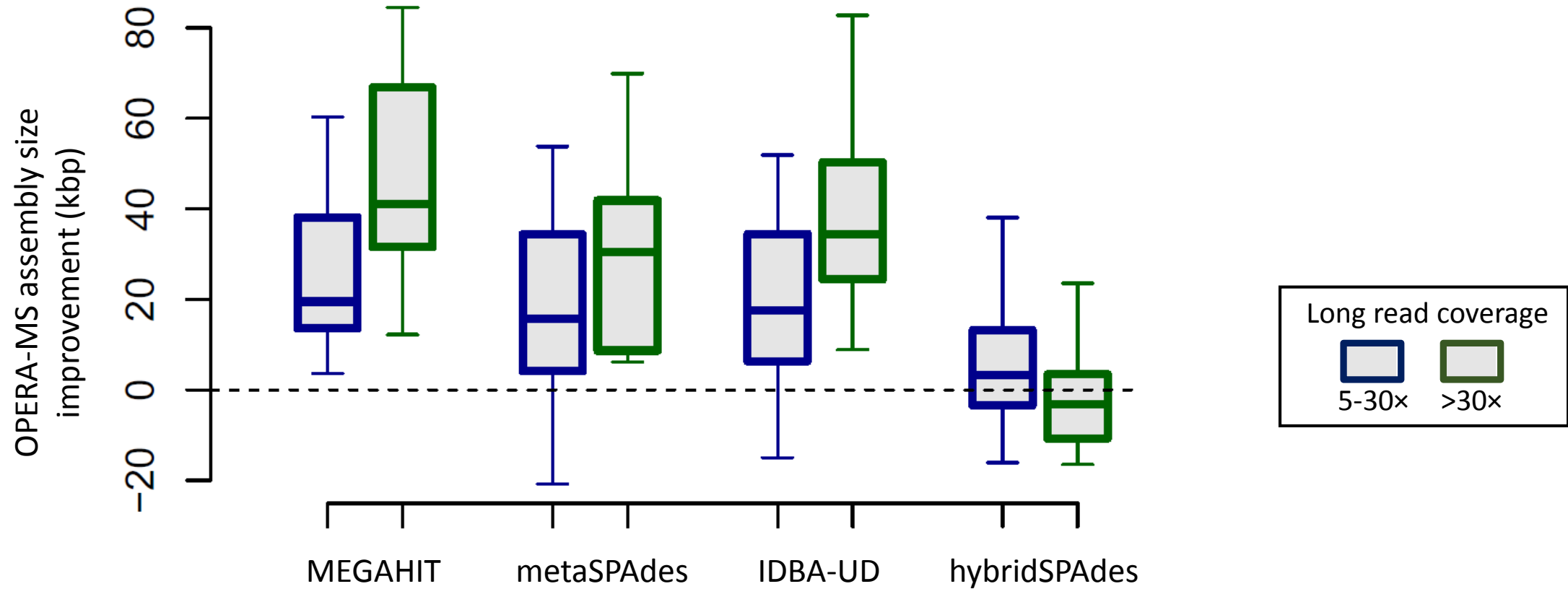
**OPERA-MS w/
contigs from
metaSPAdes**



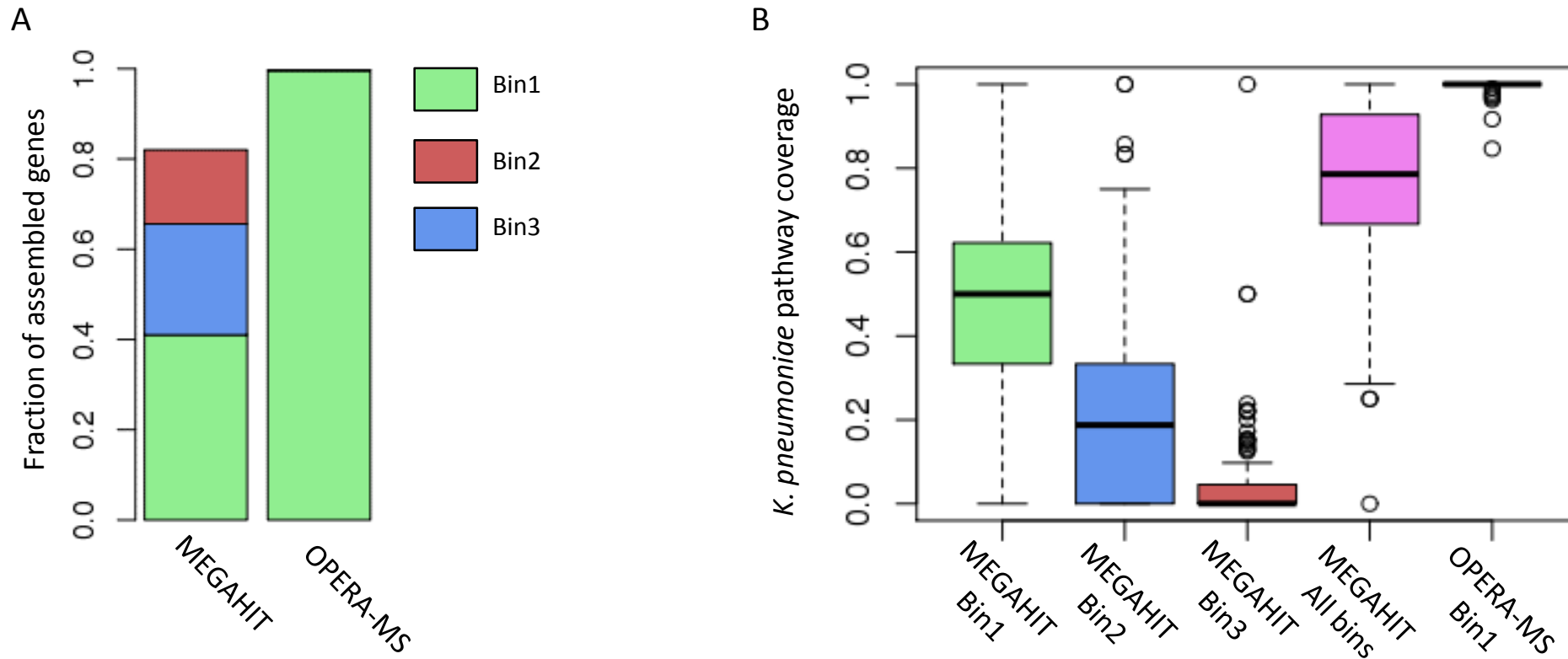
**OPERA-MS w/
contigs from
IDBA-UD**



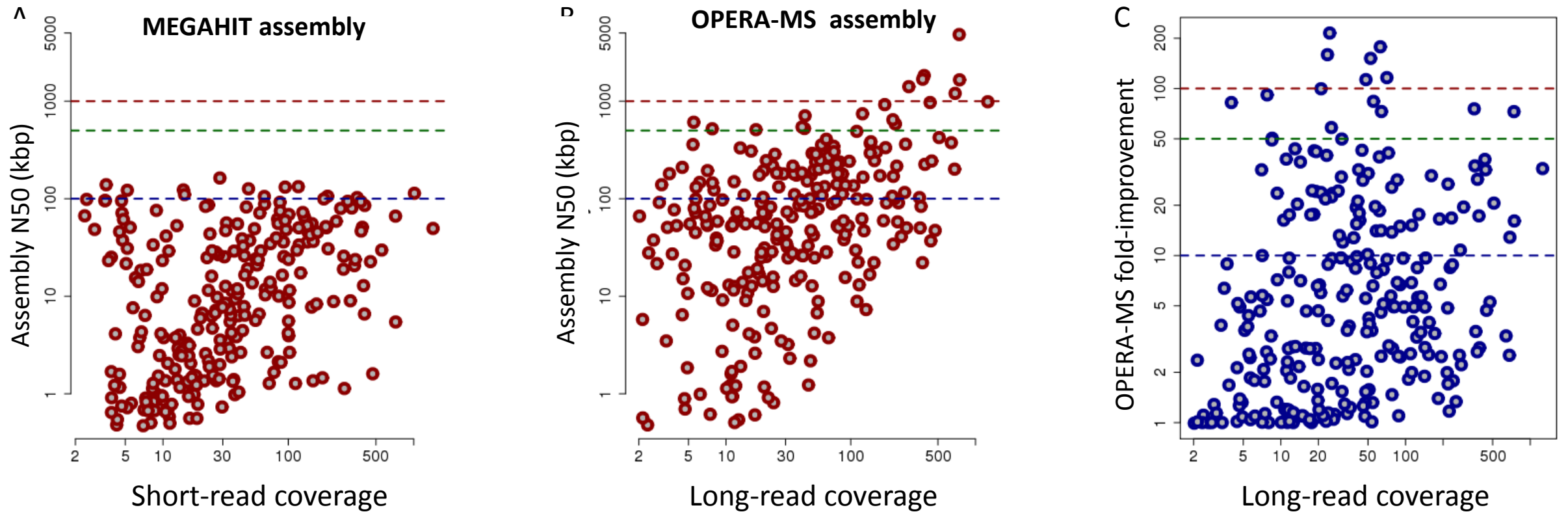
Supplementary Figure 7: OPERA-MS assemblies using contigs from other short-read metagenomic assemblers as inputs. (A,C) Scaling of OPERA-MS NGA50 values as a function of long-read coverage (n=74 genomes). (B,D) Improvement in NGA50 values obtained using OPERA-MS when compared to the starting assemblies from metaSPAdes and IDBA-UD, respectively. Data is presented as boxplots (center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers). The number of genomes in each boxplot, in ascending order of coverage, is 3, 12, 20 and 19.



Supplementary Figure 8: Evaluation of assembly completeness. OPERA-MS provides more complete assemblies than short read only methods for long read coverage $>5\times$. Similar advantages are seen using hybridSPAdes but with a higher misassembly rate (**Figure 2E**). Data is presented as boxplots (center line, median; box limits, upper and lower quartiles; whiskers, $1.5\times$ interquartile range; points, outliers). The number of genomes for coverages $5-30\times$ and $>30\times$ is 21 and 19 respectively.



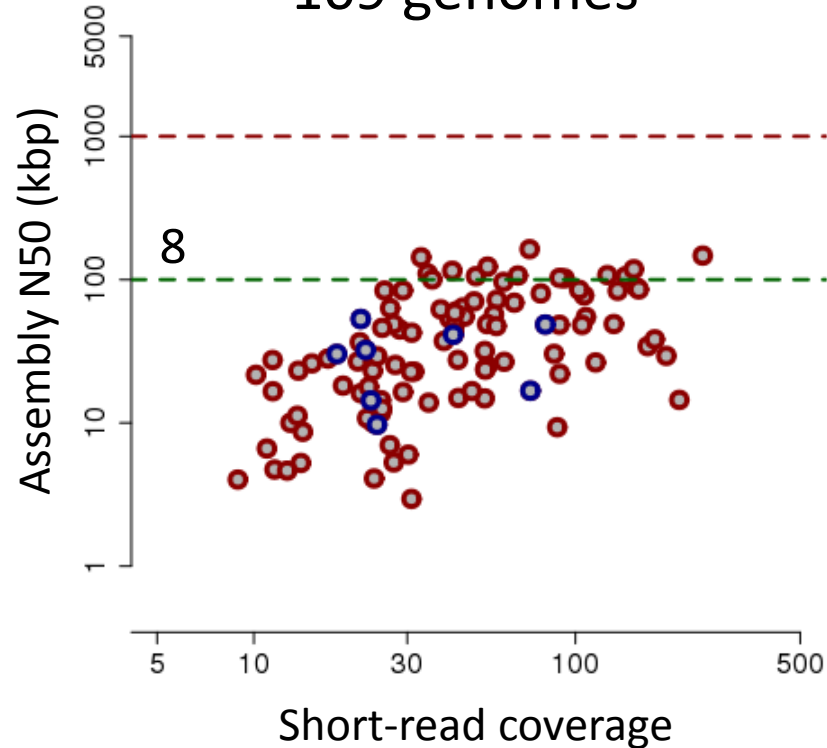
Supplementary Figure 9: Comparison of assembly and binning performance between Illumina-only (MEGAHIT) and hybrid (OPERA-MS) assemblies in the presence of multiple strains of a species. (A) Fraction of genes from *Klebsiella pneumoniae* that are assembled and present in various contig bins using the two approaches. (B) Coverage of *Klebsiella pneumoniae* pathways (n=156) in various assembly bins. Note that despite the binning of Illumina-only contigs, the corresponding bins only cover a fraction of the genes and pathways present in *K. pneumoniae*, while the hybrid approach with OPERA-MS recovers the complete genome despite the presence of multiple *K. pneumoniae* strains in the metagenome. Data is presented as boxplots (center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers)



Supplementary Figure 10: Overview of genome assemblies obtained for 28 gut microbiomes with nanopore sequencing data. N50 of assembled species (those with >1Mbp of sequence with Kraken hits) as a function of read coverage for (A) MEGAHIT and (B) OPERA-MS. (C) Assembly improvement provided by OPERA-MS compared to MEGAHIT as a function of read coverage. As was observed in the case of the mock communities, short-read assemblies plateau at an N50 of ~100kbp. In contrast, hybrid assemblies continue to scale and enable highly contiguous assemblies, with >90 genomes with N50 >100kbp and the assembly of 6 near complete genomes (N50 >1Mbp).

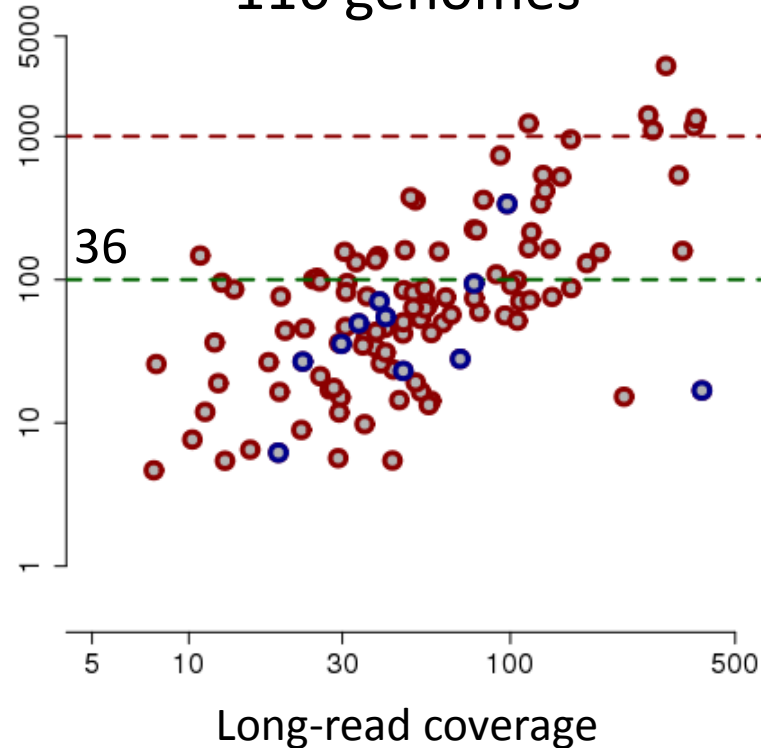
MEGAHIT

109 genomes



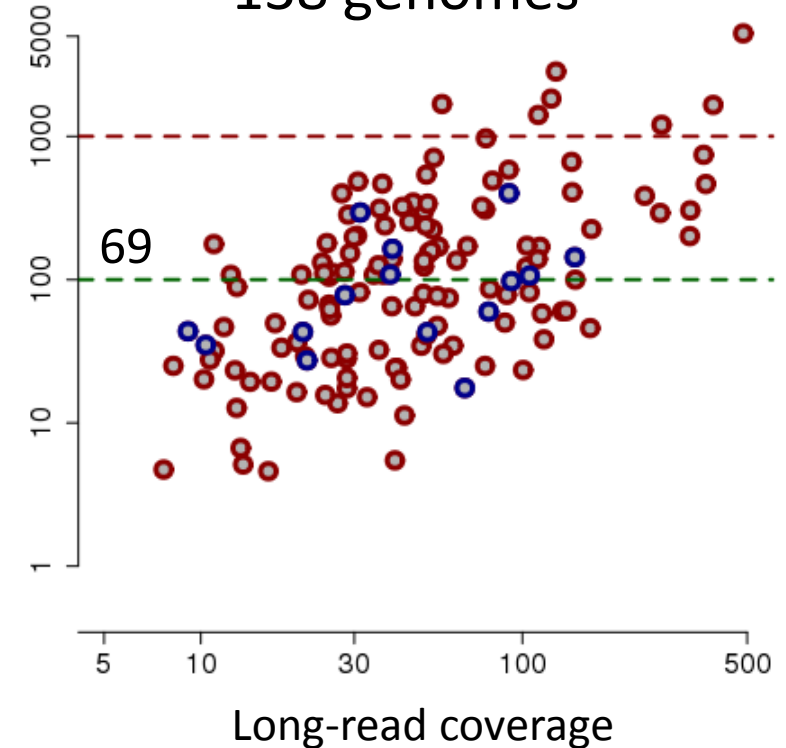
hybridSPAdes

116 genomes



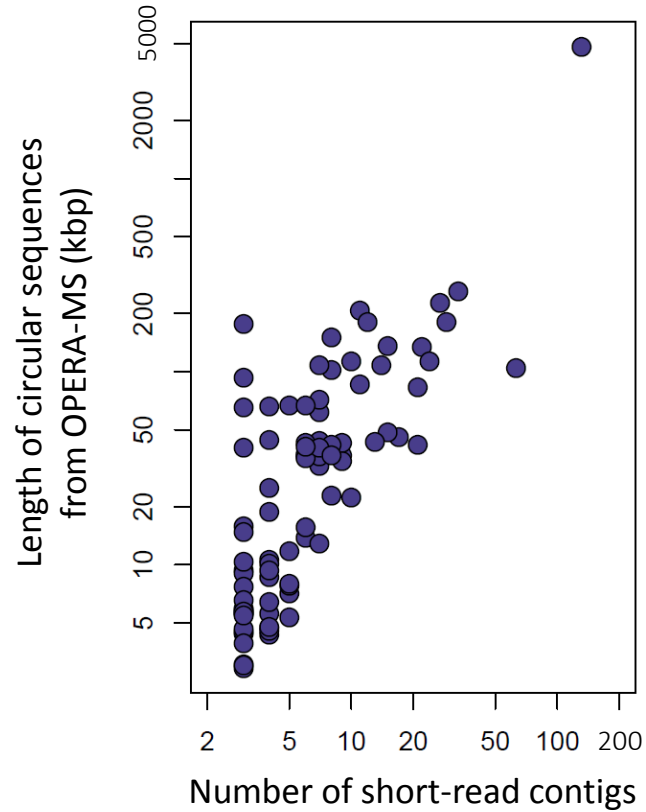
OPERA-MS

138 genomes

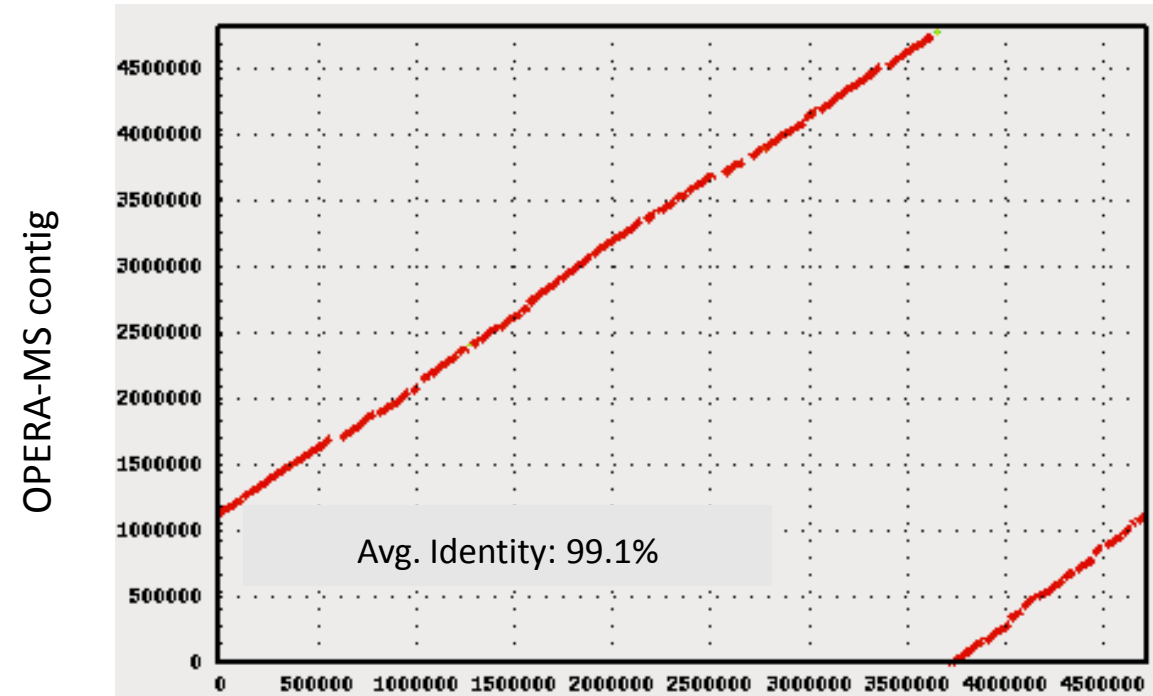


Supplementary Figure 11: Assembly of high quality draft genomes from 28 gut microbiomes with nanopore sequencing data. The graphs show N50 of assembled genomes (after binning) as a function of read coverage. Assembly bins from MaxBin2 were evaluated using CheckM (v1.0.7, --reduced_tree) and bins with completeness >90% and contamination <5% were considered high quality genomes. Pink dots represent species where multiple genomes are present in the metagenome. Overall, OPERA-MS assembles the most high-quality genomes (n=138), many of which have N50 >100kbp (n=69). Also, in cases where there are multiple genomes for a species, OPERA-MS produces more genomes with high N50 (n=8) compared to hybridSPAdes (n=1).

A

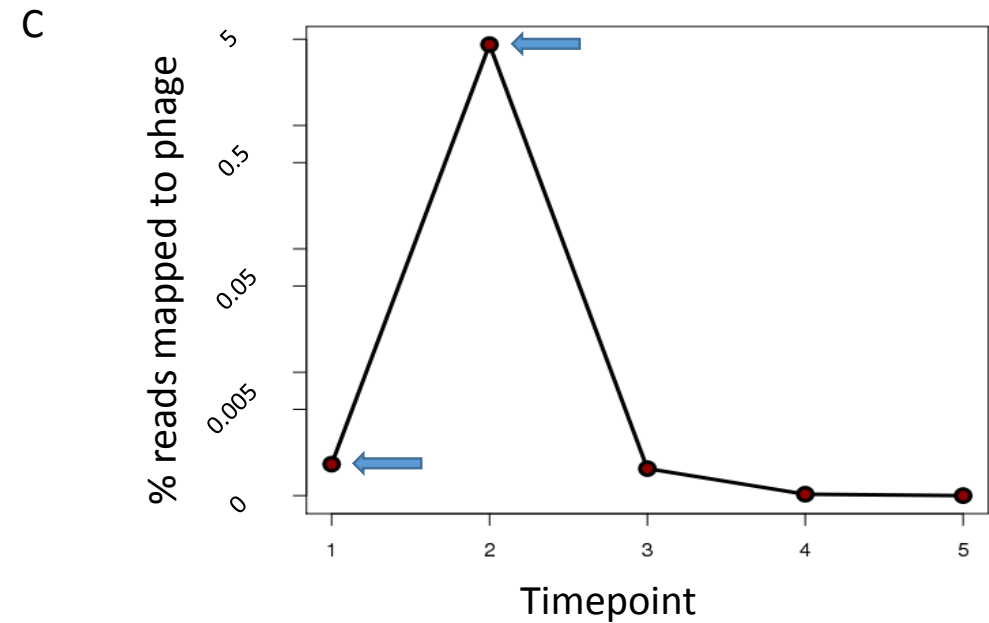
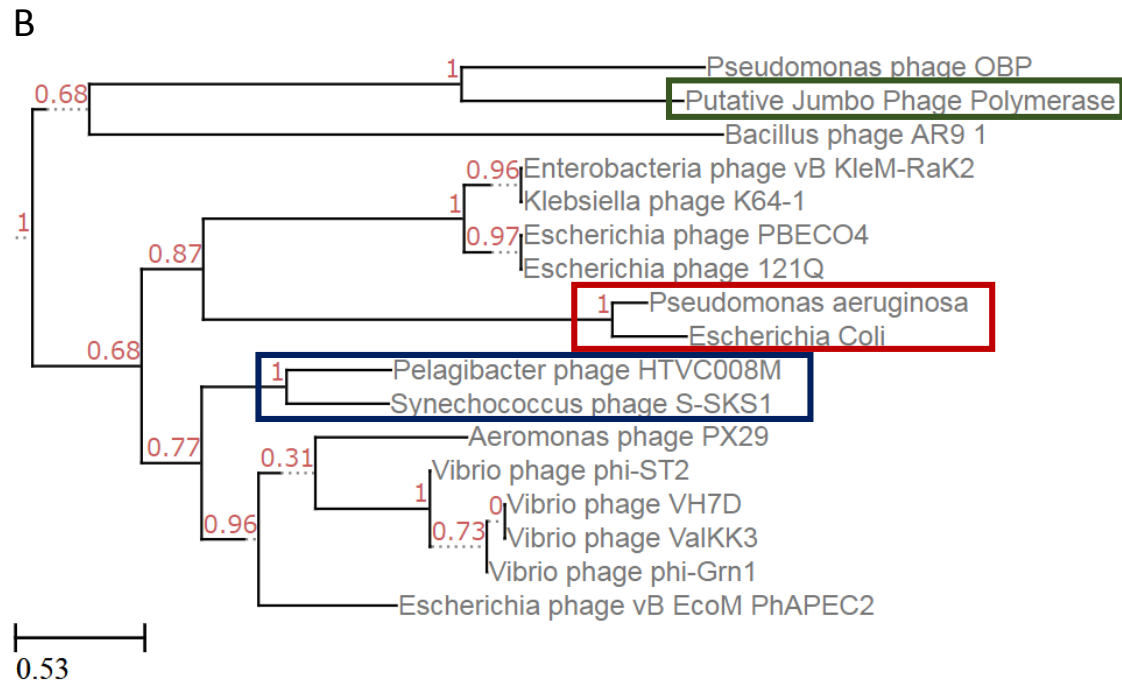
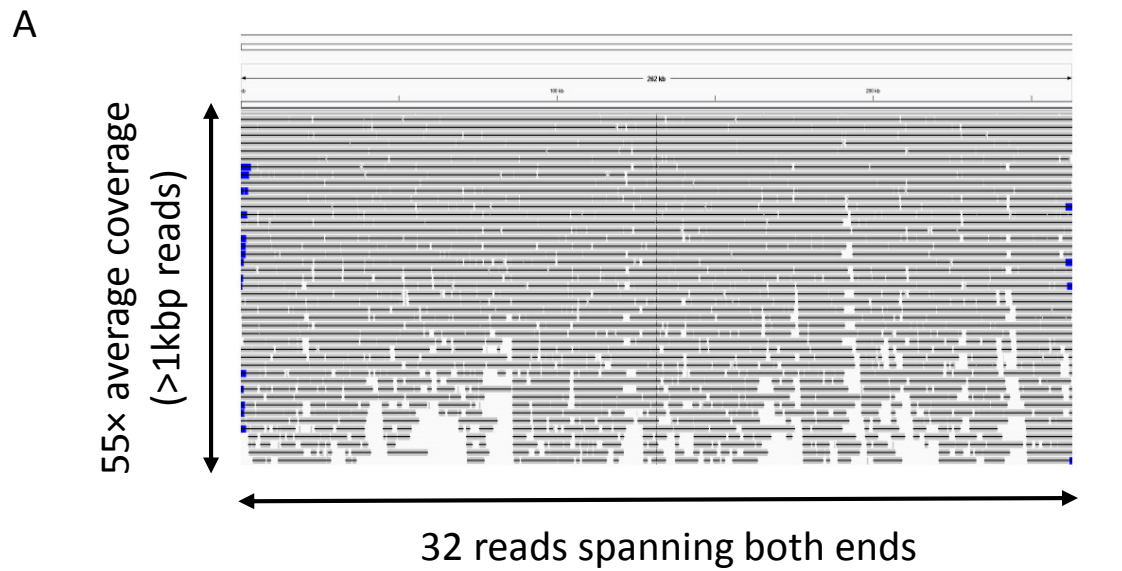


B

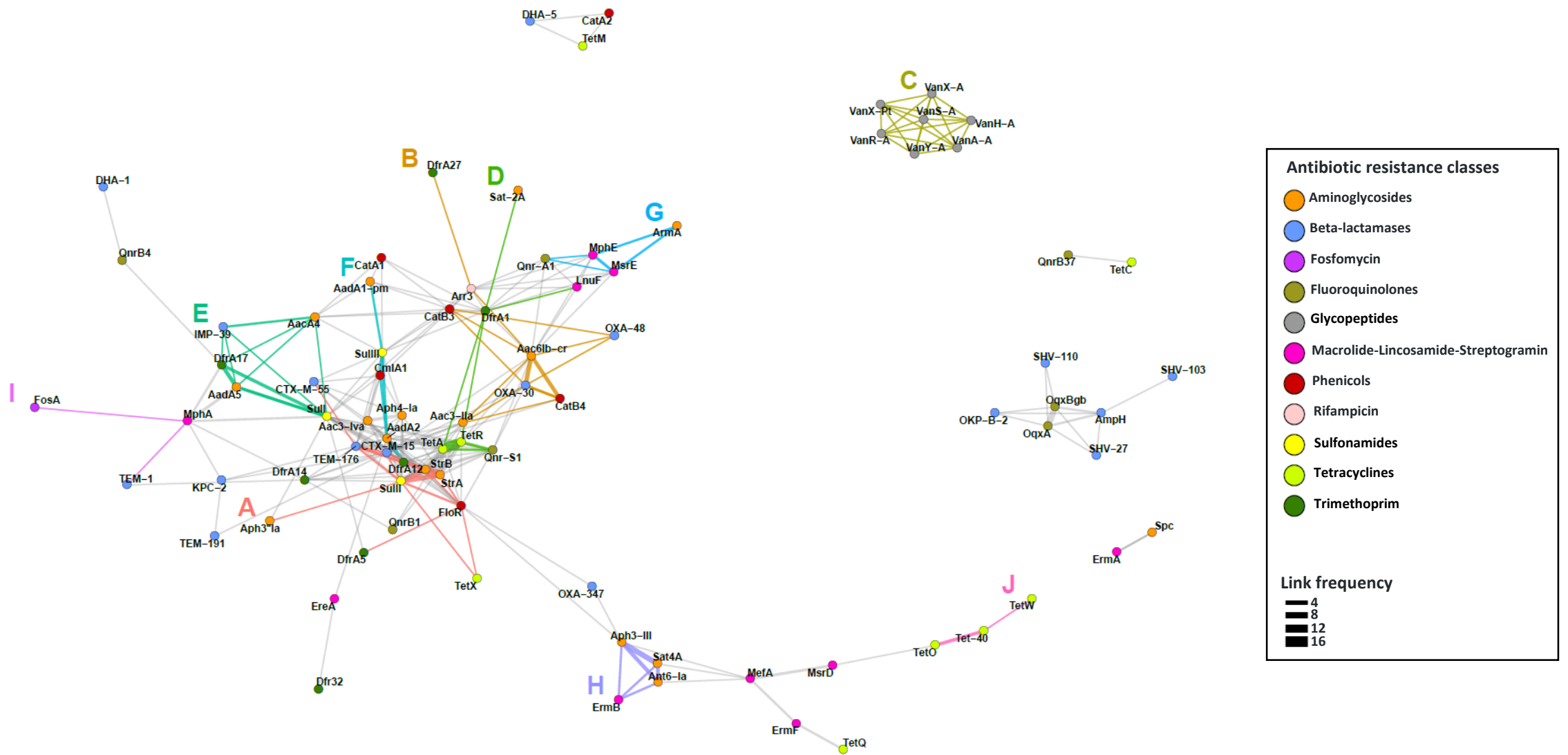


Enterobacter cloacae subsp. *cloacae* strain ENHKU01

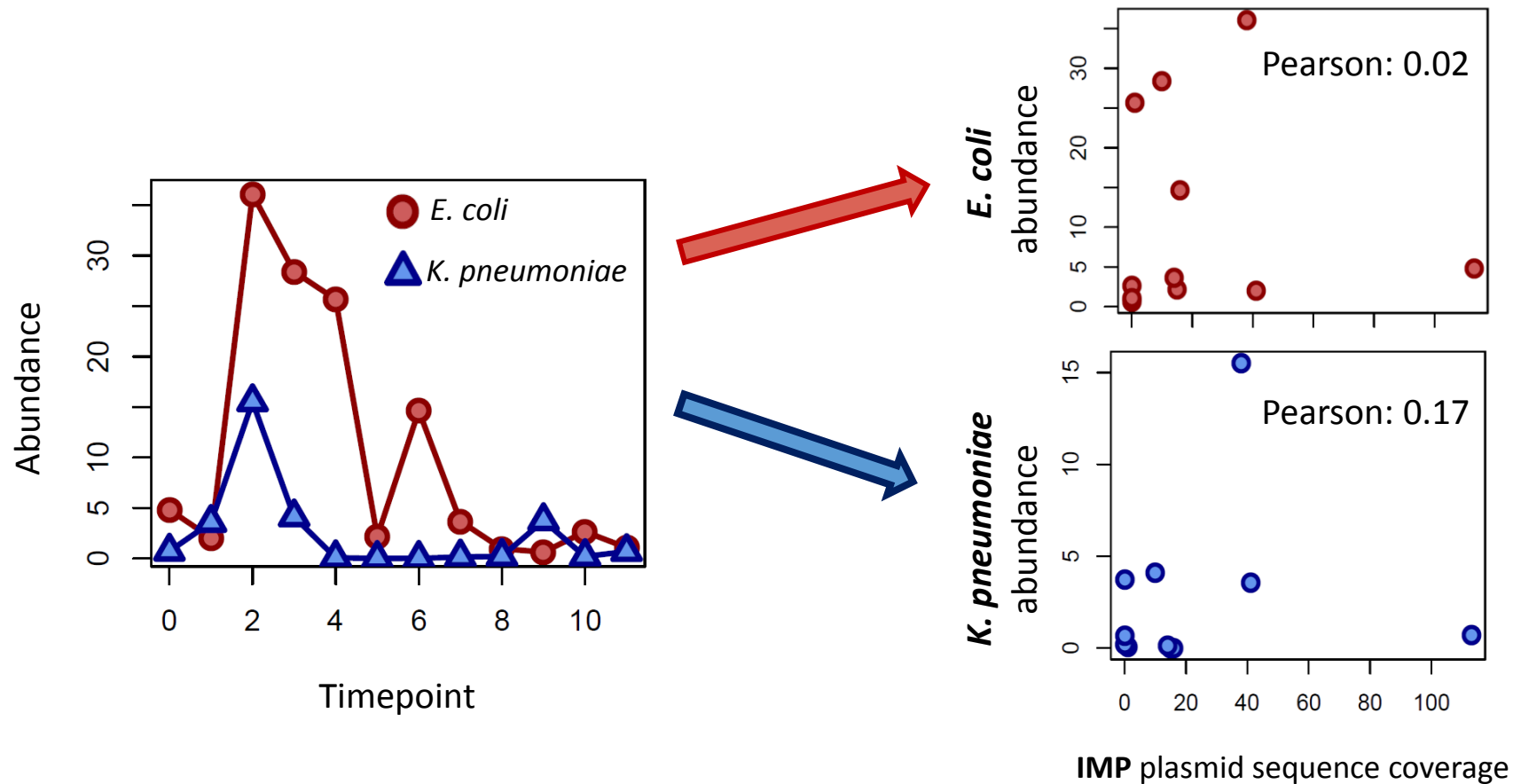
Supplementary Figure 12: OPERA-MS assembles closed circular sequences from stool metagenomic data. (A) The closed circular contigs assembled by OPERA-MS are frequently quite fragmented in the MEGAHIT assembly. (B) The longest circular sequence assembled by OPERA-MS seems to be the complete genome for an *Enterobacter cloacae* strain in the metagenome that shows high structural similarity and sequence identity to the reference genome for *Enterobacter cloacae* subsp. *cloacae* strain ENHKU01. Note that this genome was part of our reference database which enabled rescue of 13 assembly graph edges (out of 263 in total) and helped merge 8 original clusters.



Supplementary Figure 13: (A) Uniform coverage of >1kbp long ONT reads across the 263kbp sequence of the putative jumbo phage. (B) Maximum likelihood phylogenetic tree based on DNA polymerase group B protein sequences from bacteria (red box), small phages (blue box), and jumbo phages (not boxed). The DNA polymerase B protein from the putative jumbo phage genome assembled by OPERA-MS (green box) clustered with other jumbo phage proteins. Numbers on the branches indicate bootstrap confidence values based on 500 replicates. (C) Fraction of metagenomic reads that mapped to the jumbo phage genome. Start and end days for antibiotic treatment are marked with arrows.

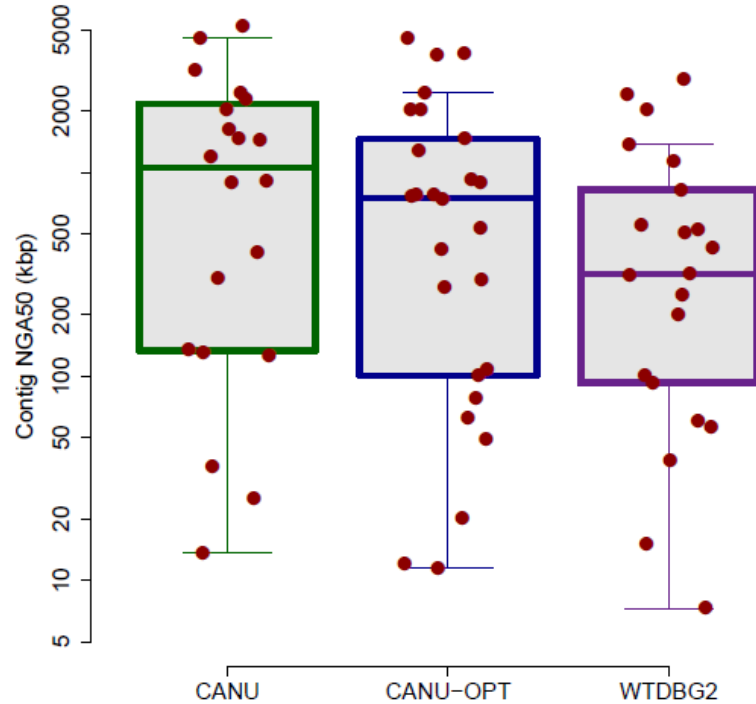


Supplementary Figure 14: Genomic linkage of AR genes and identification of common cassettes in the gut microbiome. Edges between genes indicate genetic linkage (<5kbp apart; grey otherwise) based on the 28 gut metagenomes analyzed. Markov clustering on the network was used to identify AR gene cassettes that tend to co-occur. Edges are color-coded according to their respective clusters (A-J) while resistance genes are color-coded according to the antibiotic classes that they can confer resistance to.

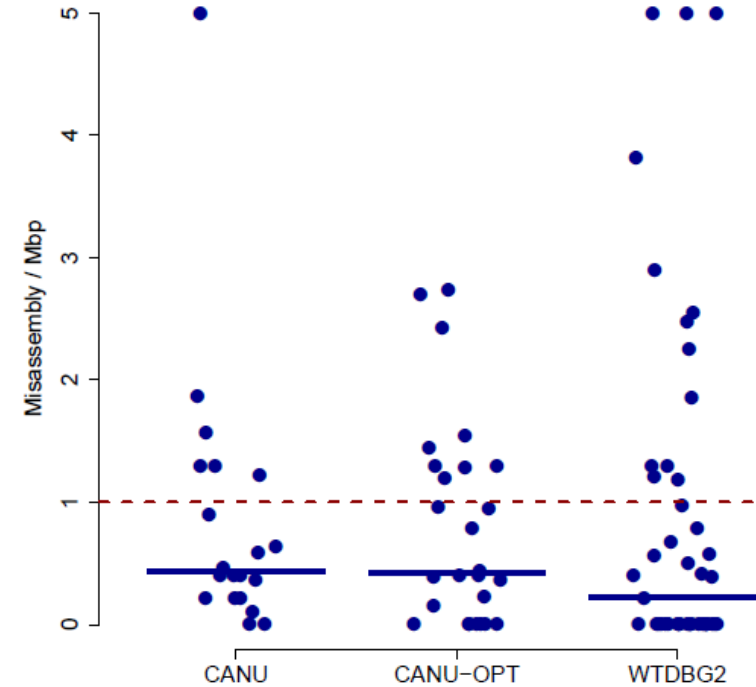


Supplementary Figure 15: Species level analysis (n=12 samples) detects no correlation between abundance of two *Enterobacteriaceae* species (*E. coli* and *K. pneumoniae*) over time and the abundance of a plasmid containing a beta-lactamase gene (IMP).

A



B



Supplementary Figure 16: Performance of long read assemblers for metagenomic assembly. We compared Canu with default settings as used for the results in the manuscript, with alternate parameter settings (CANU-OPT, recommended for metagenomic assembly; corOutCoverage=10000 corMhapSensitivity=high corMinCoverage=0), and WTDBG2 (default setting), and observed that it has (A) higher assembly contiguity (median NGA50) where data is presented as boxplots (center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers), and (B) slightly higher misassembly rate (centre value represents the median). In each figure, each data point represents one assembled genome from the mock communities (n=20, 26 and 21 for CANU, CANU-OPT and WTDBG2 respectively).

Assembly method	Single or multi-sample analysis	Sequence data	Metagenomics specific	Assembly contiguity (NGA50)	Strain deconvolution	Software website
MGS Canopy	Multi-sample	Illumina	Yes	<100kbp	No	https://bitbucket.org/HeyHo/mgs-canopy-algorithm/wiki/Home
MEGAHIT	Single	Illumina	Yes	<100kbp	No	https://github.com/voutcn/megahit
metaSPAdes	Single	Illumina	Yes	<100kbp	No	http://cab.spbu.ru/software/spades/
IDBA-UD	Single	Illumina	Yes	<100kbp	No	https://github.com/loneknightpy/idba
Canu	Single	PacBio/Nanopore/Illumina -SLR	No	100kbp-1Mbp	No	https://github.com/marbl/canu
WTDBG2	Single	PacBio/Nanopore/Illumina -SLR	No	100kbp-1Mbp	No	https://github.com/ruanjue/wtdbg2
hybridSPAdes	Single	Illumina + PacBio/Nanopore/Illumina -SLR	No	100kbp-1Mbp	No	http://cab.spbu.ru/software/spades/
Athena	Single	Illumina +10x read cloud	No	100kbp-1Mbp	No	https://github.com/abishara/athena_meta/
OPERA-MS	Single	Illumina + PacBio/Nanopore/Illumina -SLR	Yes	100kbp-1Mbp	Yes	https://github.com/CSB5/OPERA-MS

Supplementary Table 1: Attributes of various assembly algorithms that have been used for metagenomic assembly. Assembly contiguity values give an indication of expected performance assuming sufficient read coverage for the genome of interest (typically >10-30×).

Sample ID	Patient ID	Throughput (Mbp)	Number of reads (thousands)	Read N50 (kbp)	ONT sequencer	ONT flow cell	Technique used	Kit	Basecaller
S1	818VB-3	444	28	4.9	MinION	R9	HMW	NSK007	RNN SQK007 1.107
S2	631RZ-6	582	124	1.0	MinION	R9 spoton	HMW	SQK-LSK208	2D flo106 250bps 1.125
S3	631RZ-5	70	5	4.2	MinION	R9 spoton	WGA	SQK-LSK208	2D flo106 250bps 1.125
S4	135EA6	1124	150	2.0	MinION	R9 spoton	HMW	SQK-LSK208	2D flo106 250bps 1.125
S5	948BA1	1115	426	0.6	MinION	R9 spoton	HMW	SQK-LSK208	2D flo106 250bps 1.125
S6	V01-T-0506-S02	1006	106	2.5	MinION	R9 spoton	Normal	SQK-LSK208	Albacore
S7	V03-T-0506-S04	1632	123	4.0	MinION	R9 spoton	Normal	SQK-LSK208	Albacore
S8	V02-S-0510-S03	325	60	1.3	MinION	R9.5 spoton	Normal	SQK-LSK208	Albacore
S9	V06-T-0502-S07	2438	172	5.1	MinION	R9.5 spoton	BluePippin	SQK-LSK208	Albacore
S10	V05-T-0502-S06	3106	665	1.1	MinION	R9.5 spoton	Normal	SQK-LSK208	Albacore
S11	V03-T-0508-S04	7996	393	6.4	MinION	R9.5 spoton	Normal	SQK-LSK208	Albacore
S12	V00-S-0509-S01	4218	333	3.6	MinION	R9.5 spoton	Size Select	SQK-LSK208	Albacore
S13	V02-T-0504-S03	7255	1032	2.0	MinION	R9.5 spoton	Size Select	SQK-LSK208	Albacore
S14	V06-T-0501-S07	3190	511	1.8	GridION	R9.5 spoton	Normal	SQK-LSK208	Guppy v0.3.0 for live 1D basecalling
S15	V00-S-0511-S01	2677	1345	0.7	GridION	R9.5 spoton	Size Select	SQK-LSK208	Guppy v0.3.0 for live 1D basecalling
S16	V04-S-0509-S04	2572	629	1.2	GridION	R9.5 spoton	Size Select	SQK-LSK208	Guppy v0.3.0 for live 1D basecalling
S17	V05-S-0512-S05	1990	540	1.1	GridION	R9.5 spoton	Normal	SQK-LSK208	Guppy v0.3.0 for live 1D basecalling
S18	V05-T-0513-S05	2541	514	1.4	GridION	R9.5 spoton	Size Select	SQK-LSK208	Guppy v0.3.0 for live 1D basecalling
S19	V07-S-0512-S07	5307	430	4.3	GridION	R9.5 spoton	Normal	SQK-LSK208	Guppy v0.3.0 for live 1D basecalling
S20	V02-T-1664-S03	5412	275	6.2	GridION	R9.5 spoton	Normal	SQK-LSK208	Guppy v0.3.0 for live 1D basecalling
S21	V02-T-1665-S03	4260	308	3.7	GridION	R9.5 spoton	Normal	SQK-LSK208	Guppy v0.3.0 for live 1D basecalling
S22	V03-S-1663-S04	2517	618	1.0	GridION	R9.5 spoton	Normal	SQK-LSK208	Guppy v0.3.0 for live 1D basecalling
S23	V03-S-0457-S04	2714	394	1.5	GridION	R9.5 spoton	Normal	SQK-LSK208	Guppy v0.3.0 for live 1D basecalling
S24	V03-T-0504-S04	5870	537	3.1	GridION	R9.5 spoton	Size Select	SQK-LSK208	Guppy v0.3.0 for live 1D basecalling
S25	V04-T-0508-S05	3167	233	4.0	GridION	R9.5 spoton	Size Select	SQK-LSK208	Guppy v0.3.0 for live 1D basecalling
S26	V07-T-0504-S08	4308	559	2.0	GridION	R9.5 spoton	Size Select	SQK-LSK208	Guppy v0.3.0 for live 1D basecalling
S27	V07-S-0510-S08	4940	561	2.5	GridION	R9.5 spoton	Size Select	SQK-LSK208	Guppy v0.3.0 for live 1D basecalling
S28	V08-S-0510-S09	5233	507	2.8	GridION	R9.5 spoton	Size Select	SQK-LSK208	Guppy v0.3.0 for live 1D basecalling

Supplementary Table 2: ONT sequencing protocols and statistics for stool metagenomics. Technique used: HMW (High Molecular Weight Extraction protocol), WGA (Whole Genome Amplification), Normal (extraction protocol used for shotgun libraries), BluePippin (size selection from 3-50kb), Size Select (size selection with 0.45X AMPure XP).

Strain ID	Name	Mass (in μg)	Relative Abundance (%)	Reference Genome
ATCC 39213	<i>Pseudomonas putida</i>	69	30.00%	NA
ATCC 700721	<i>Klebsiella pneumoniae</i>	46	20.00%	NC_009648
ATCC 17978	<i>Acinetobacter baumannii</i>	34.5	15.00%	GCF_001077675.1_ASM107767v1_genomic
ATCC 12228	<i>Staphylococcus epidermidis</i>	23	10.00%	NC_004461
ATCC BAA-472	<i>Enterococcus faecium</i>	16.1	7.00%	NC_017960
ATCC 13311	<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar <i>typhimurium</i>	11.5	5.00%	GCF_000743055.1_ASM74305v1_genomic
DSM17610	<i>Neisseria subflava</i>	6.9	3.00%	NA
DSM 7271	<i>Capnocytophaga ochracea</i>	6.9	3.00%	NC_013162
DSM20016	<i>Lactobacillus reuteri</i>	3.45	1.50%	NC_010609
DSM 20219	<i>Bifidobacterium longum</i>	3.45	1.50%	NC_015067
DSM 22815	<i>Jonquetella anthropi</i>	2.3	1.00%	GCF_000237805.1_ASM23780v1_genomic
DSM20472	<i>Fingoldia magna</i>	1.15	0.50%	NC_010376
DSM 30054	<i>Enterobacter cloacae</i>	1.15	0.50%	NC_014121
DSM 15702	<i>Eubacterium siraeum</i>	1.15	0.50%	NA
DSM 20083	<i>Bifidobacterium adolescentis</i>	0.575	0.25%	NC_008618
DSM 6778	<i>Streptococcus parasanguinis</i>	0.575	0.25%	NC_015678
DSM 20482	<i>Fusobacterium nucleatum</i> subsp. <i>polymorphum</i>	0.575	0.25%	GCF_000153625.3_ASM15362v1_genomic
DSM 18711	<i>Prevotella oris</i>	0.23	0.10%	GCF_000377685.1_ASM37768v1_genomic
DSM5359	<i>Helicobacter cinaedi</i>	0.23	0.10%	NC_020555
DSM 10105	<i>Parascardovia denticolens</i>	0.23	0.10%	GCF_000191785.1_ASM19178v1_genomic

Supplementary Table 3: GIS20 mock community composition.

Sequencing run	Throughput (Mbp)	Number of reads (thousands)	N50 (kbp)	ONT flow cell	Kit	Basecaller
1	238	9	9.4	R7	MAP006	RNN SQK007 1.69
2	415	17	8.6	R7	MAP006	RNN SQK007 1.69
3	267	7	12.1	R7	MAP006	RNN SQK007 1.69
4	188	4	14.7	R7	MAP006	RNN SQK007 1.69
5	480	22	8.0	R9	NSK007	RNN SQK007 1.97
6	859	34	9.3	R9	NSK007	RNN SQK007 1.99
TOTAL	2,446	91	9.2			

Supplementary Table 4: MinION sequencing statistics for the GIS20 mock community.

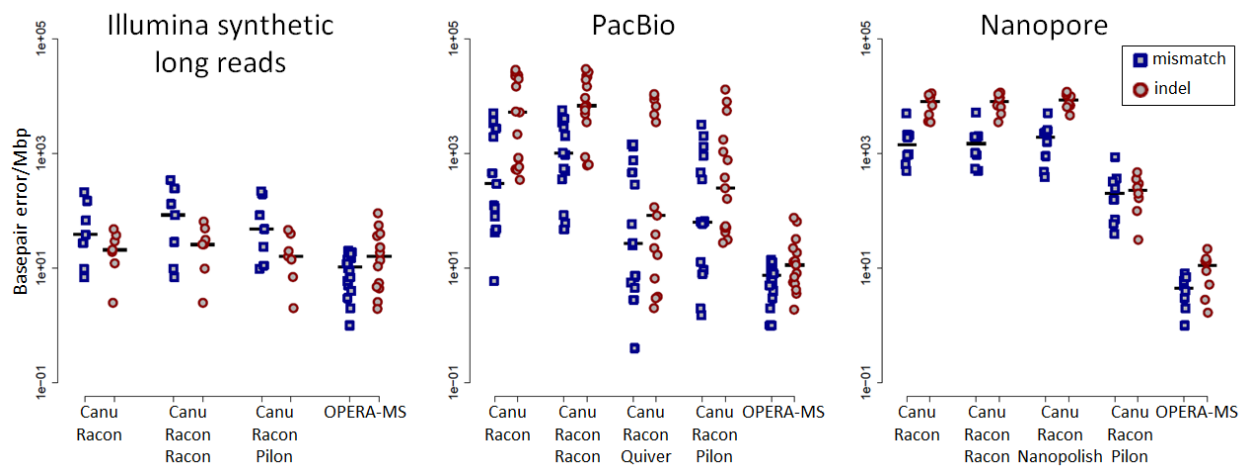
Gene Name	Associated Antibiotic Resistance
Aac3-Iva Aph4-Ia TEM-176 CmlA1 SulIII AadA2 CTX-M-55	aminoglycoside(3), cephalosporin(2), phenicol, sulfone
DfrA17 Sull AadA5 AacA4 IMP-39	diaminopyrimidine, sulfone, aminoglycoside(2), carbapenem
Sull CmlA1 Arr3 SulIII	sulfone(2), phenicol, rifamycin
OXA-48 OXA-30 CatB3 Aac6Ib-cr	carbapenem/cephalosporin, penam, phenicol, fluoroquinolone, aminoglycoside
Aph3-III ErmB Ant6-Ia Sat4A	aminoglycoside(2), lincosamide, nucleoside
OXA-30 Aac6Ib-cr Qnr-S1	penam, fluoroquinolone(2), aminoglycoside
OXA-30 Aac6Ib-cr CatB4	penam, fluoroquinolone, aminoglycoside, phenicol
CmlA1 AadA2 SulIII	phenicol, aminoglycoside, sulfone
SulIII CmlA1 AadA2	sulfone, phenicol, aminoglycoside
Aph3-III Ant6-Ia Sat4A	aminoglycoside(2), nucleoside
OXA-30 Aac6Ib-cr CatB4	penam, fluoroquinolone, aminoglycoside, phenicol
KPC-2 TEM-191 CTX-M-15	carbapenem, cephalosporin(2)
OXA-30 Aac6Ib-cr CatB4	penam, fluoroquinolone, aminoglycoside, phenicol
Aph3-III OXA-347 FloR	aminoglycoside, penam, phenicol
KPC-2 MphA	carbapenem, macrolide
SulIII TEM-176	sulfone, cephalosporin
Aac6Ib-cr CatB4	fluoroquinolone, aminoglycoside, phenicol
ErmA Spc	lincosamide, aminoglycoside
TetO Tet-40	tetracycline(2)
TetX FloR	tetracycline, phenicol
Tet-40 TetO	tetracycline(2)
TetW Tet-40	tetracycline(2)
QnrB4 DfrA17	fluoroquinolone, diaminopyrimidine
Aph3 ^{III} Sull	aminoglycoside, sulfone

Supplementary Table 5 : Novel antibiotic resistance gene combinations identified.

Supplementary Note 1: Impact of technology-specific and Illumina polishing on base-pair level accuracy of long-read assembly

We evaluated the base-pair level accuracy of Canu long-read only assemblies using different polishing approaches. In addition to the default approach of using Racon which is technology agnostic, we evaluated the use of Quiver¹ for PacBio reads and Nanopolish² for nanopore reads. As expected, using Quiver improved the accuracy of the Pacbio assemblies further but many genomes still continue to have low quality, likely due to shallow coverage for these genomes (**Supplementary Note Figure 1**). Improvement using Nanopolish was limited for nanopore assemblies, likely due to the lack of coverage in low abundance genomes. Finally, polishing with Illumina reads (using Pilon) provided accuracy that was closer to hybrid assemblies but OPERA-MS assemblies still had 1/5th of the errors (**Supplementary Figure 6**).

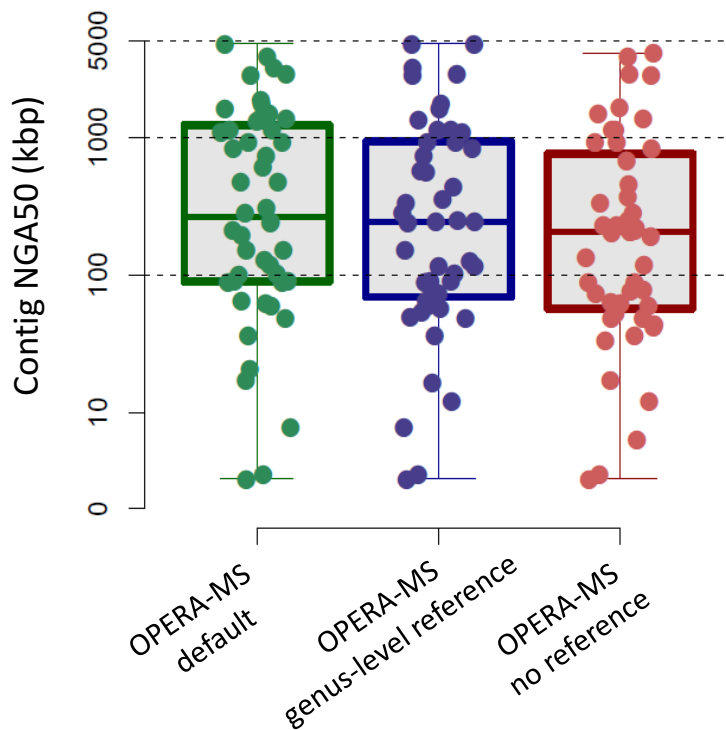
Note that a potential advantage of long-read assembly is its ability to resolve repeat regions better. We evaluated this property, using MUMmer self-alignments to identify repeat regions in the genome, and observed that while OPERA-MS improved notably over Canu assemblies for nanopore data (1/3rd of the errors), it showed slight improvements for Illumina synthetic long read and PacBio datasets (20% fewer base-pair errors) in such regions.



Supplementary Note Figure 1: Evaluation of base-pair level accuracy for Canu assemblies with different polishing steps (polishing steps are listed in order of use in the legend). Each dot represents results for a genome from the mock communities. Quiver (v2.2.2) and Nanopolish (v0.9.2; R9 model) were run with default settings.

Supplementary Note 2: Utility of reference genomes for metagenomic assembly with OPERA-MS

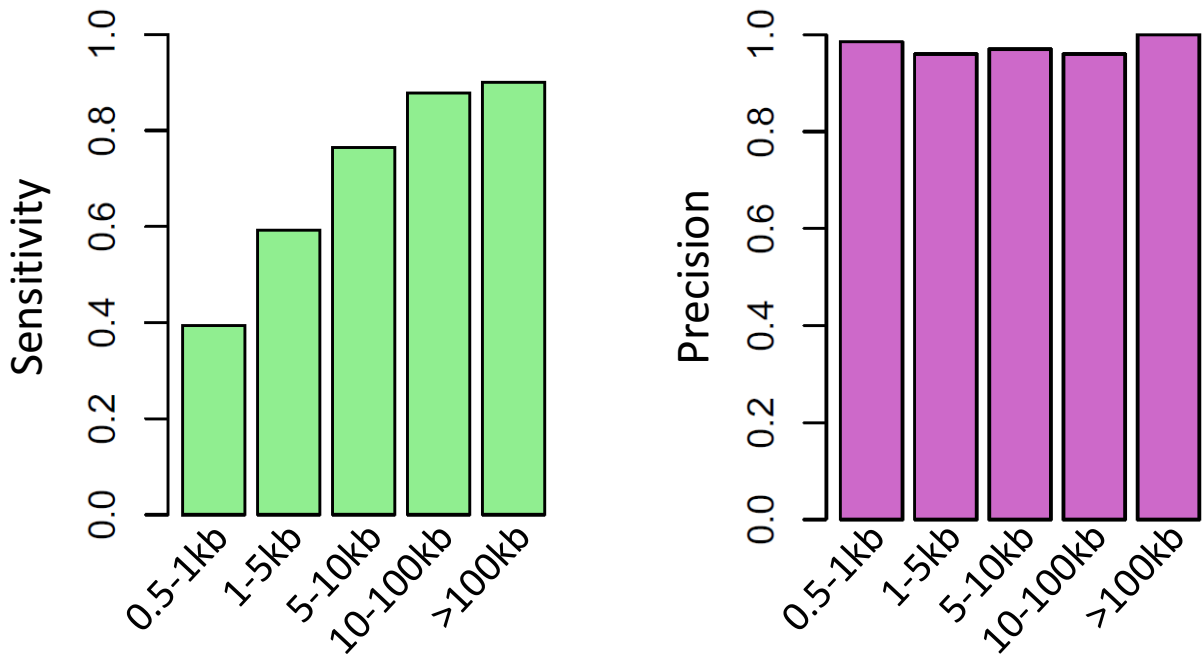
To evaluate the utility of the reference-guided step in OPERA-MS, we ran OPERA-MS without this optional step (no reference), as well as with a database where species-level references were removed (genus-level reference), before analyzing the mock communities. In all settings, OPERA-MS provided assemblies with high median NGA50 compared to Illumina-only assemblies (**Supplementary Note Figure 2**). Notably, switching from having all references to only those at the genus-level resulted in no NGA50 reduction, while using no references led to ~20% reduction (median across genomes). The impact of references was felt under two conditions: (i) low read coverage (<5×), where references provided a 65% boost in median NGA50, or (ii) high coverage skew (>1.2), as can be seen in rapidly dividing bacteria³, with median NGA50 improvement of 25%. With the sequencing coverage generated in this study, we estimated that these conditions primarily impact less than half of the species with relative abundance <0.5% in the community. Additionally, we did not see an impact on the ability to assemble strain genomes in the presence of multiple strains in the virtual gut community (e.g. the dominant *K. pneumoniae* strain was still assembled into a 4.3Mbp contig with only 2 small errors).



Supplementary Note Figure 2: Evaluation of utility of the reference-guided step for metagenomic assembly with OPERA-MS. We evaluated results on the mock communities with 3 settings: (i) default reference database, (ii) using a filtered genome database with no corresponding species-level references (genus-level reference), and (iii) using no references. Each dot in the figure represents an assembled genome in the mock communities.

Supplementary Note 3: Sensitivity of Mash distance for identifying links between clusters

The Mash toolkit was used to estimate distances between contig clusters and reference genomes due to its computational efficiency⁴. As Mash uses kmer occurrences to measure Jaccard similarities, for small clusters we can expect a drop in sensitivity. This is indeed what we observed on the mock community, with sensitivity being as low as 40% for clusters shorter than 1kbp (**Supplementary Note Figure 3**). However, for clusters longer than 10kbp, sensitivity was greater than 90% and precision was consistently high in all cases (~95%). Correspondingly, we noted that hundreds of new links were correctly discovered for the mock communities (GIS20 Nanopore – 601, Pacbio – 454, HMP staggered mock Pacbio – 936, Illumina synthetic long read - 1037), especially for low coverage or high coverage skew genomes (see **Supplementary Note 2**).



Supplementary Note Figure 3: Performance of Mash distance for identifying links between clusters. Evaluation of the (A) sensitivity and (B) precision of the Mash distance based approach in OPERA-MS to rescue links between clusters, as a function of cluster size (minimum size among the two clusters for a link). The evaluation was done on the GIS20 nanopore dataset and true links were identified by mapping of contigs to the reference genomes using MUMmer.

References

1. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
2. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* **12**, 733–735 (2015).
3. Brown, C. T., Olm, M. R., Thomas, B. C. & Banfield, J. F. Measurement of bacterial replication rates in microbial communities. *Nat. Biotechnol.* **34**, 1256–1263 (2016).
4. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).