Benchmarking second and third-generation sequencing platforms for microbial metagenomics

Victoria Meslier, Benoit Quinquis, Kévin da Silva, Florian Plaza Oñate, Nicolas Pons, Hugo Roume, Mircea Podar, Mathieu Almeida

scientific data

OPEN

DATA DESCRIPTOR

Benchmarking second and third-generation sequencing platforms for microbial metagenomics

Victoria Meslier[1], Benoit Quinquis[1], Kévin Da Silva[1], Florian Plaza Oñate[1], Nicolas Pons[1], Hugo Roume[1], Mircea Podar[2] ✉ & Mathieu Almeida[1] ✉

Shotgun metagenomic sequencing is a common approach for studying the taxonomic diversity and metabolic potential of complex microbial communities. Current methods primarily use second generation short read sequencing, yet advances in third generation long read technologies provide opportunities to overcome some of the limitations of short read sequencing. Here, we compared seven platforms, encompassing second generation sequencers (Illumina HiSeq 300, MGI DNBSEQ-G400 and DNBSEQ-T7, ThermoFisher Ion GeneStudio S5 and Ion Proton P1) and third generation sequencers (Oxford Nanopore Technologies MinION R9 and Pacific Biosciences Sequel II). We constructed three uneven synthetic microbial communities composed of up to 87 genomic microbial strains DNAs per mock, spanning 29 bacterial and archaeal phyla, and representing the most complex and diverse synthetic communities used for sequencing technology comparisons. Our results demonstrate that third generation sequencing have advantages over second generation platforms in analyzing complex microbial communities, but require careful sequencing library preparation for optimal quantitative metagenomic analysis. Our sequencing data also provides a valuable resource for testing and benchmarking bioinformatics software for metagenomics.

Background & Summary

High throughput metagenomic sequencing has drastically changed our understanding of microbial ecosystems. One of the most popular approach is to use metagenomic sequencing, assembly and binning procedures[1–4] to investigate the structure, functionalities and ecological interactions of microbial communities with their environment or host[5–9]. Most metagenomic studies rely on second generation sequencing providing billions of short sequences in a single run, with the Illumina sequencing platforms being the most widely used[10]. Improvement of third generation sequencing yields millions of long reads per run but are mostly used for genomic assembly procedures[11–13], and further benchmarking is required to evaluate their performance for quantitative metagenomic analysis.

For this purpose, we produced three synthetic uneven DNA mocks, varying in their microbial richness (64 to 87 strains, full composition in Supplementary Table S1) and belonging to 29 prokaryotic phyla (Fig. 1). We particularly focused on combining a large spectrum of genome sizes, GC content and mixing closely related species. These mocks represent to date the most complex synthetic communities for evaluating sequencers performances compared to previous studies[14–19], and not obtained from *in silico* simulated microbial communities[20–23]. We performed five short read sequencing (Ion Proton P1, Ion S5, Illumina HiSeq 3000, DNBSEQ G400, DNBSEQ T7) and two long read sequencing technologies (ONT MinION and PacBio Sequel II), making this study the one covering the widest diversity of sequencing technologies (Table 1).

The mock1 (71 strains) was sequenced using all technologies and mocks 2 and 3 (64 and 87 strains, respectively) were additionally sequenced to estimate the impact of various microbial richness (Table 2). After sequencing and quality control, we were able to align more than 99% of all reads for each technology back to their reference genomes, with almost 100% of uniquely mapped reads for long read technologies, down to 87% for Ion Proton and S5 technologies[15,24]. All technologies provided up to 99% identity, except for the MinION R9 with about 89% identity due to a high in/del errors and substitution errors[25]. The PacBio Sequel II provided the lowest substitution error rate and the DNBSeq G400 and T7 the lowest in/dels rate[26,27].

[1]Université Paris-Saclay, INRAE, MetaGenoPolis, 78350, Jouy-en-Josas, France. [2]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA. ✉e-mail: podarm@ornl.gov; mathieu.almeida@inrae.fr

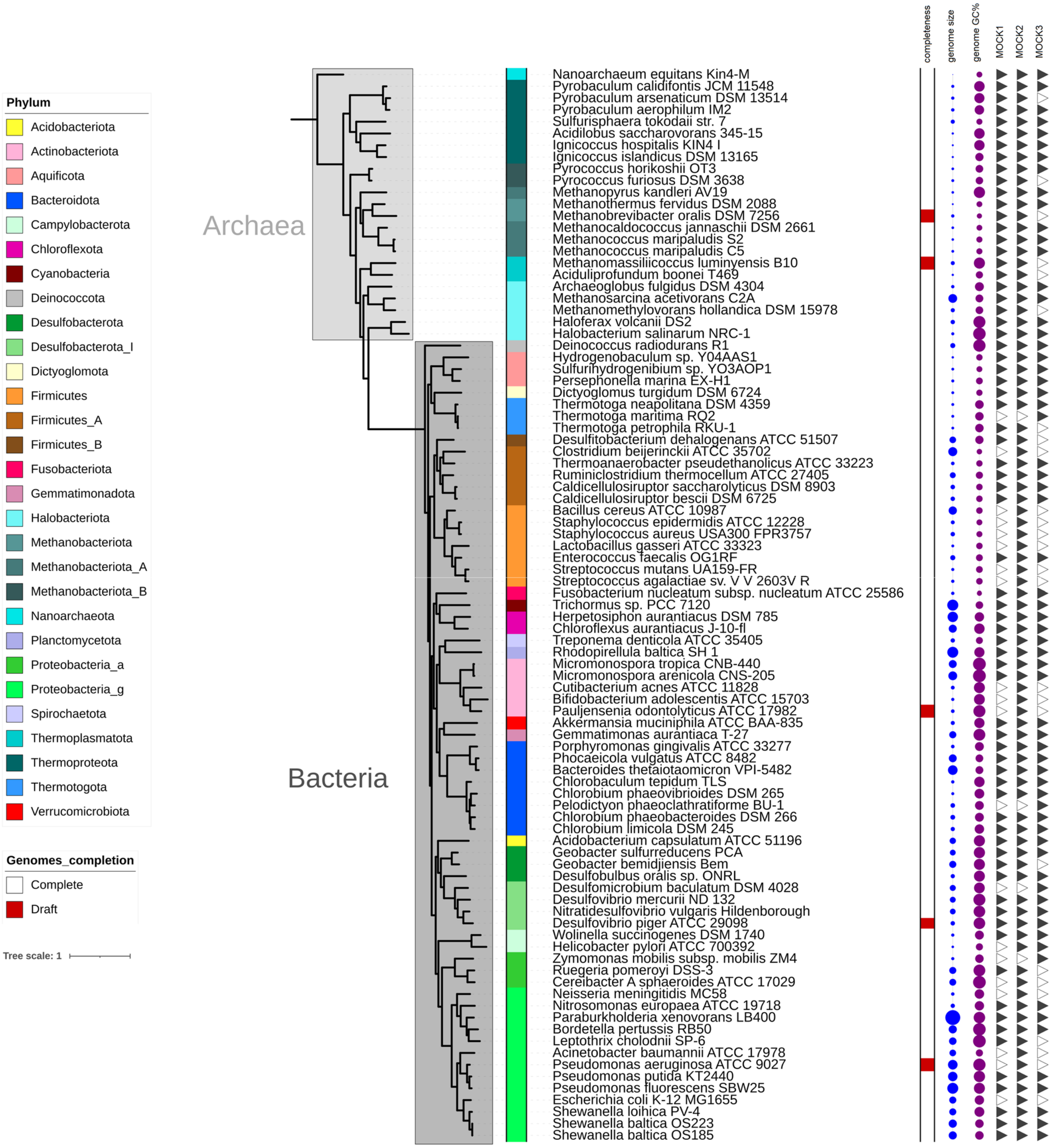
Phylum
Acidobacteriota
Actinobacteriota
Aquificota
Bacteroidota
Campylobacterota
Chloroflexota
Cyanobacteria
Deinococcota
Desulfobacterota
Desulfobacterota_I
Dictyoglomota
Firmicutes
Firmicutes_A
Firmicutes_B
Fusobacteriota
Gemmatimonadota
Halobacteriota
Methanobacteriota
Methanobacteriota_A
Methanobacteriota_B
Nanoarchaeota
Planctomycetota
Proteobacteria_a
Proteobacteria_g
Spirochaetota
Thermoplasmatota
Thermoproteota
Thermotogota
Verrucomicrobiota
Genomes_completion
Complete
Draft
Tree scale: 1
Archaea
Bacteria
completeness
genome size
genome GC%
MOCK1
MOCK2
MOCK3
Nanoarchaeum equitans Kin4-M
Pyrobaculum calidifontis JCM 11548
Pyrobaculum arsenaticum DSM 13514
Pyrobaculum aerophilum IM2
Sulfurisphaera tokodaii str. 7
Acidilobus saccharovorans 345-15
Ignicoccus hospitalis KIN4 I
Ignicoccus islandicus DSM 13165
Pyrococcus horikoshii OT3
Pyrococcus furiosus DSM 3638
Methanopyrus kandleri AV19
Methanothermus fervidus DSM 2088
Methanobrevibacter oralis DSM 7256
Methanocaldococcus jannaschii DSM 2661
Methanococcus maripaludis S2
Methanococcus maripaludis C5
Methanomassiliicoccus luminyensis B10
Aciduliprofundum boonei T469
Archaeoglobus fulgidus DSM 4304
Methanosarcina acetivorans C2A
Methanomethylovorans hollandica DSM 15978
Haloferax volcanii DS2
Halobacterium salinarum NRC-1
Deinococcus radiodurans R1
Hydrogenobaculum sp. Y04AAS1
Sulfurihydrogenibium sp. YO3AOP1
Persephonella marina EX-H1
Dictyoglomus turgidum DSM 6724
Thermotoga neapolitana DSM 4359
Thermotoga maritima RQ2
Thermotoga petrophila RKU-1
Desulfitobacterium dehalogenans ATCC 51507
Clostridium beijerinckii ATCC 35702
Thermoanaerobacter pseudethanolicus ATCC 33223
Ruminiclostridium thermocellum ATCC 27405
Caldicellulosiruptor saccharolyticus DSM 8903
Caldicellulosiruptor bescii DSM 6725
Bacillus cereus ATCC 10987
Staphylococcus epidermidis ATCC 12228
Staphylococcus aureus USA300 FPR3757
Lactobacillus gasseri ATCC 33323
Enterococcus faecalis OG1RF
Streptococcus mutans UA159-FR
Streptococcus agalactiae sv. V V 2603V R
Fusobacterium nucleatum subsp. nucleatum ATCC 25586
Trichormus sp. PCC 7120
Herpetosiphon aurantiacus DSM 785
Chloroflexus aurantiacus J-10-fl
Treponema denticola ATCC 35405
Rhodopirellula baltica SH 1
Micromonospora tropica CNB-440
Micromonospora arenicola CNS-205
Cutibacterium acnes ATCC 11828
Bifidobacterium adolescentis ATCC 15703
Pauljensenia odontolyticus ATCC 17982
Akkermansia muciniphila ATCC BAA-835
Gemmatimonas aurantiaca T-27
Porphyromonas gingivalis ATCC 33277
Phocaeicola vulgatus ATCC 8482
Bacteroides thetaiotaomicron VPI-5482
Chlorobaculum tepidum TLS
Chlorobium phaeovibrioides DSM 265
Pelodictyon phaeoclathratiforme BU-1
Chlorobium phaeobacteroides DSM 266
Chlorobium limicola DSM 245
Acidobacterium capsulatum ATCC 51196
Geobacter sulfurreducens PCA
Geobacter bemidjiensis Bem
Desulfobulbus oralis sp. ONRL
Desulfomicrobium baculatum DSM 4028
Desulfovibrio mercurii ND 132
Nitratidesulfovibrio vulgaris Hildenborough
Desulfovibrio piger ATCC 29098
Wolinella succinogenes DSM 1740
Helicobacter pylori ATCC 700392
Zymomonas mobilis subsp. mobilis ZM4
Ruegeria pomeroyi DSS-3
Cereibacter A sphaeroides ATCC 17029
Neisseria meningitidis MC58
Nitrosomonas europaea ATCC 19718
Paraburkholderia xenovorans LB400
Bordetella pertussis RB50
Leptothrix cholodnii SP-6
Acinetobacter baumannii ATCC 17978
Pseudomonas aeruginosa ATCC 9027
Pseudomonas putida KT2440
Pseudomonas fluorescens SBW25
Escherichia coli K-12 MG1655
Shewanella loihica PV-4
Shewanella baltica OS223
Shewanella baltica OS185

Fig. 1 Phylogenetic tree of the microbial species used in the mock microbial communities. Neighbor joining tree built using 40 universal protein markers and visualized using iTOL. On the left, colored strips referred to Phylum phylogenetic ranks using GTBD. Annotations on the right referred to genome completeness (white square, complete; red square, draft genome), genome size, genome GC percent (circle sizes proportional to the dataset range), and mocks composition (plain triangle, present; empty triangle, absent).

To evaluate the impact of sequencing depth, we performed a subsampling analysis and compared observed versus theoretical genome abundances (Fig. 2). In general, Spearman correlations were high for all technologies, reaching values above 0.9 when mapping at least 100,000 reads. Notably, correlations were slightly lower for mock communities with higher microbial richness, partially due to cross-matching events during mapping procedures. Whilst second generation sequencers were equivalent for taxonomic profiling[28], we found more pronounced decreases for MinION and PacBio correlations, even if reads were almost entirely uniquely mapped. Although the PacBio sequencer presented the lowest error rate, these results could be explained by the DNA size filtering step performed during library preparation, which was calibrated to maximize the read length. We hypothesize that the filtering step could remove highly fragmented DNA, thus impacting strains relative abundances[29–31].

	Illumina HiSeq 3000	Ion Proton P1	Ion S5	DNBSEQ-G400	DNBSEQ- T7	ONT MinION R9	PacBio Sequel II
Amplification type	Solid-Phase bridge	emPCR	emPCR	DNB	DNB	—	—
Sequencing principle	SBS	SBS	SBS	SBS	SBS	SMS	SMS
Average reads length ± stdv after trimming (bp)	149 ± 4.24	144.041 ± 28.43	145.76 ± 28.12	99.91 ± 0.96	99.52 ± 2.58	4408.41 ± 2831.95	10289.7 ± 4036.27
Max read length after trimming (bp)	150	373	347	100	100	60869	40278
SE/PE	PE	SE	SE	PE	PE	SE	SE
Average insert size ± stdv (bp)	433.47 ± 92.37	—	—	245.13 ± 51.04	235.56 ± 54.80	—	—
Total Run Time	4d	4 h	4 h	3d	3d	48 h*	30 h

Table 1. Overview of the main sequencing platform characteristics used in this study. emPCR = emulsion PCR; SBS: Sequencing by Synthesis; SMS: Single-Molecule Sequencing; DNB: DNA NanoBall. *Data produced by MinION are accessible few minutes after the sequencing start, however, we run the MinION for 48 h to analyze the maximal throughput. DNBSEQ-G400 platform was formerly named MGISEQ-2000. SE: Single end (Forward read only). PE: Paired End. Run Time indicates time to obtain the maximal throughput.

Sample ID	Sequencing Technology	N. reads before trimming (million)	N. reads after trimming (million)	%Mapped end-to-end	%Uniquely mapped	%Avg end-to-end best mapped identity	%Avg end-to-end best mapped substitutions	%Avg end-to-end best mapped in/dels
MOCK1 (N = 71 species)	Illumina HiSeq 3000	21.38*2	20.59*2	99.62	93.21	99.45	0.46	0.09
	Ion Proton P1	21.23	20.00	99.29	87.13	99.42	0.12	0.46
	Ion S5	30.05	28.51	99.35	87.13	99.61	0.08	0.31
	ONT Minion R9	0.757	0.696	99.75	99.63	89.08	3.37	7.55
	PacBio Sequel II	0.525	0.524	99.65	99.62	99.72	0.06	0.22
	DNBSEQ-G400	36.17*2	35.42*2	99.22	89.16	99.70	0.30	0.003
	DNBSEQ-T7	404.06*2	375.12*2	98.92	88.78	99.42	0.58	0.003
MOCK2 (N = 87 species)	Illumina HiSeq 3000	24.27*2	23.17*2	99.66	89.44	99.43	0.49	0.08
	Ion Proton P1	22.21	20.98	99.46	86.61	99.41	0.12	0.47
	Ion S5	25.36	24.17	99.55	86.66	99.59	0.09	0.32
	ONT Minion R9	0.919	0.831	99.74	99.61	89.05	3.39	7.56
	DNBSEQ-G400	37.92*2	37.14*2	99.43	88.85	99.73	0.27	0.003
	DNBSEQ-T7	404.99*2	376.04*2	99.03	88.32	99.46	0.54	0.003
MOCK3 (N = 64 species)	Illumina HiSeq 3000	63.97*2	62.08*2	99.58	90.36	99.39	0.52	0.09
	Ion Proton P1	21.50	20.26	99.15	88.33	99.44	0.11	0.45
	Ion S5	27.32	25.90	99.31	88.40	99.62	0.08	0.30
	ONT Minion R9	0.865	0.791	99.79	99.61	89.06	3.35	7.59

Table 2. Mapping summary per mock and sequencing technology. The number of reads before and after trimming refer to the sequencing depth (million reads) before and after quality control filtering and trimming. %Mapped end-to-end correspond to the read percentage aligned to a reference genome considering the read full length, while %Uniquely mapped reads correspond to the percentage of reads aligned to only one region of a reference genome. %Avg end-to-end refer to the best hit mean percentage for mapped identity and substitutions and insertions/deletions (in/dels) respectively. See the Method section for trimming and mapping parameters.

By focusing on mock1 individual genome abundances, we found that most genomes were accurately estimated for all technologies (Fig. 3). Over or under abundance estimation for most genomes was not particularly related to sequencing technology, read length, taxonomy, nor by GC-content, genome size and genome completeness, even at a low depth of 500,000 reads. These results suggest promising opportunities for affordable alternatives to high depth metagenomic sequencing, by using a limited number of reads- the so-called shallow shotgun sequencing- to explore the composition of complex microbiota[32], even with third generation sequencers[19].

Finally, we performed *de novo* metagenomic assembly and confronted assemblies with their reference genomes (Table 3). PacBio Sequel II generated the most contiguous assemblies with 36 full genomes out of 71 in mock1, followed by MinION (22 genomes), making third generation sequencers more adapted for genome reconstruction. When considering the mismatches per 100kbps, PacBio Sequel II was also providing the most accurate assemblies, followed by Illumina HiSeq 3000 and DNBSeq G400 (Table 3). However, the lower indels rates obtained with DNBSeq G400 and Illumina HiSeq suggests that hybrid procedures may provide more

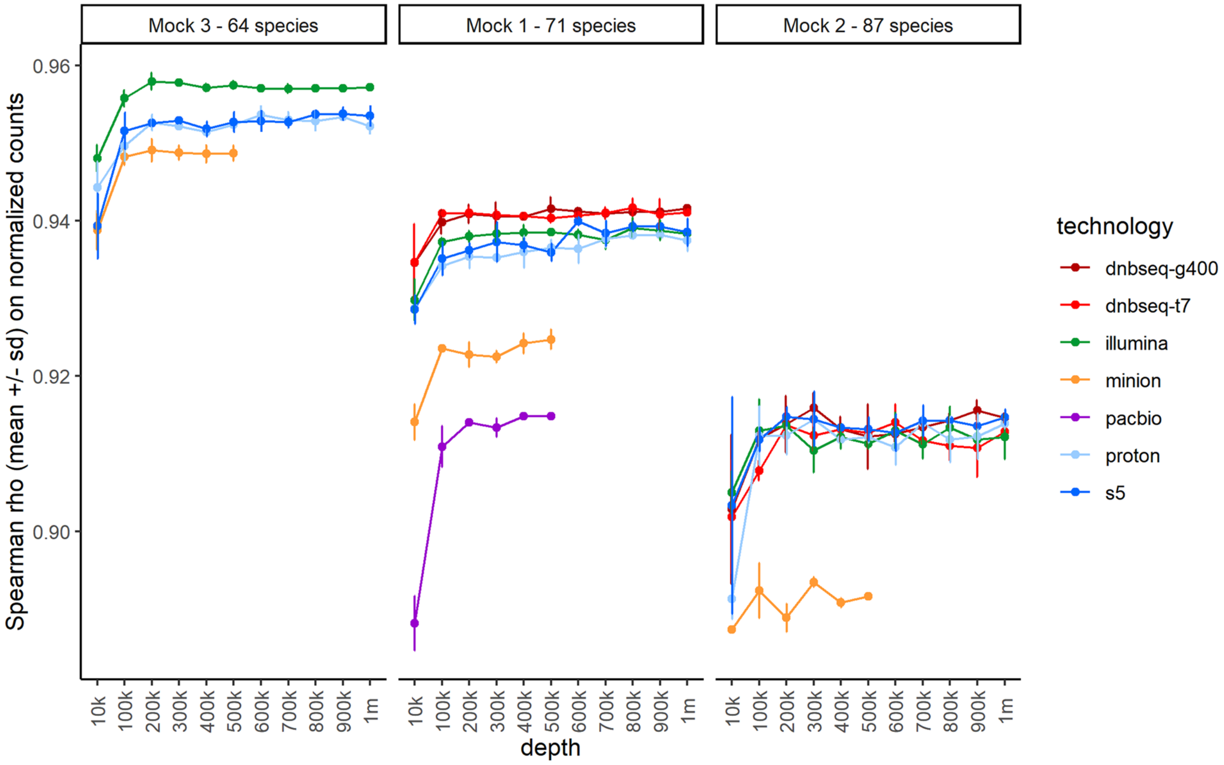

Fig. 2 Overall comparison between observed and excepted mock compositions per technology. After read mapping to the mock reference genome, a subsampling was performed 3 times at multiple sequencing depth from 10,000 to 1 million reads, except for ONT MinION and PacBio for which maximum depth was 500k. Spearman rank correlations were calculated between observed genomes abundances normalized by genome size (expressed in %) and the expected mock composition (%). Means ± SD are reported based on the 3 iterations performed per depth. PacBio sequencing was not performed on mock3 and mock2, DNBSEQ-T7 and DNBSEQ-G400 were not performed on mock3.

accurate assemblies than those obtained using long reads alone. We tested this hypothesis by generating hybrid assemblies (Supplementary Information S2) for each technology. For MinION, the hybrid assemblies improved notably the genome fraction recovery compared to MinION assembly only, while reducing the number of fully unaligned contigs, confirming our initial hypothesis. For PacBio, the hybrid assembly did not improve assembly metrics, except for Illumina and DNBSeq with a lower indels rate per 100 kpbs and an improvement in genome fraction recovery with Illumina.

By this work, we provide a new resource with highly complex synthetic mock samples and extensive metagenomic sequencing data, using the most popular second and third generation sequencing platforms. These data could be used to benchmark or improve metagenomic assemblers, binning software and taxonomic profilers[33,34].

Methods

Synthetic microbial communities' construction. A total of 91 different strains were used in this study. For 58 strains of Archaea and Bacteria, we used archived gDNA from the Shakya *et al.* study[14]. To further increase the complexity of the constructed community, we cultured nine additional microbes for which the genomic sequence was available. High molecular weight DNA was isolated and quantified as described previously[14]. Purified DNA from 4 other bacteria and archaea was obtained from the laboratory of Dr. Cynthia Gilmour (Smithsonian Research Institute, Edgewater, Maryland, USA). We also used the 20 Strain Even Mix Genomic Material from ATCC (MSA-1002). Three different genomic microbial synthetic communities were assembled by mixing individual, purified DNAs. The composition of each community aimed to provide variation in the number and relative abundance of individual microbe and their represented taxonomic category. The communities achieved a diversity ranging from 64 to 87 strains, representing 29 phyla of Archaea and Bacteria, with a relative abundance distribution spanning over three orders of magnitude. The genome size distribution ranged from 0.49 to 9.7 Mbp and the G + C content was between 27 and 69%. Within the 91 strains, 21 have extrachromosomal DNA such as plasmids or additional chromosomes (Supplementary Table S1). A phylogenetic tree for all strains was constructed using 40 universal protein markers as previously described[1] and taxonomic ranks were updated using gtdbtk Release 07-RS207[35]. The tree was visualized and annotated using iTOL[36].

Library preparation and Sequencing. *ThermoFisher Ion Proton P1 and Ion GeneStudio S5 library preparation and sequencing.* Ion Proton P1 and Ion GeneStudio S5 libraries were built using Ion Plus Fragment Library kit (Thermo Fisher Scientific, Waltham, MD, USA). 500 ng of High Molecular Weight (HMW) DNA was sheared

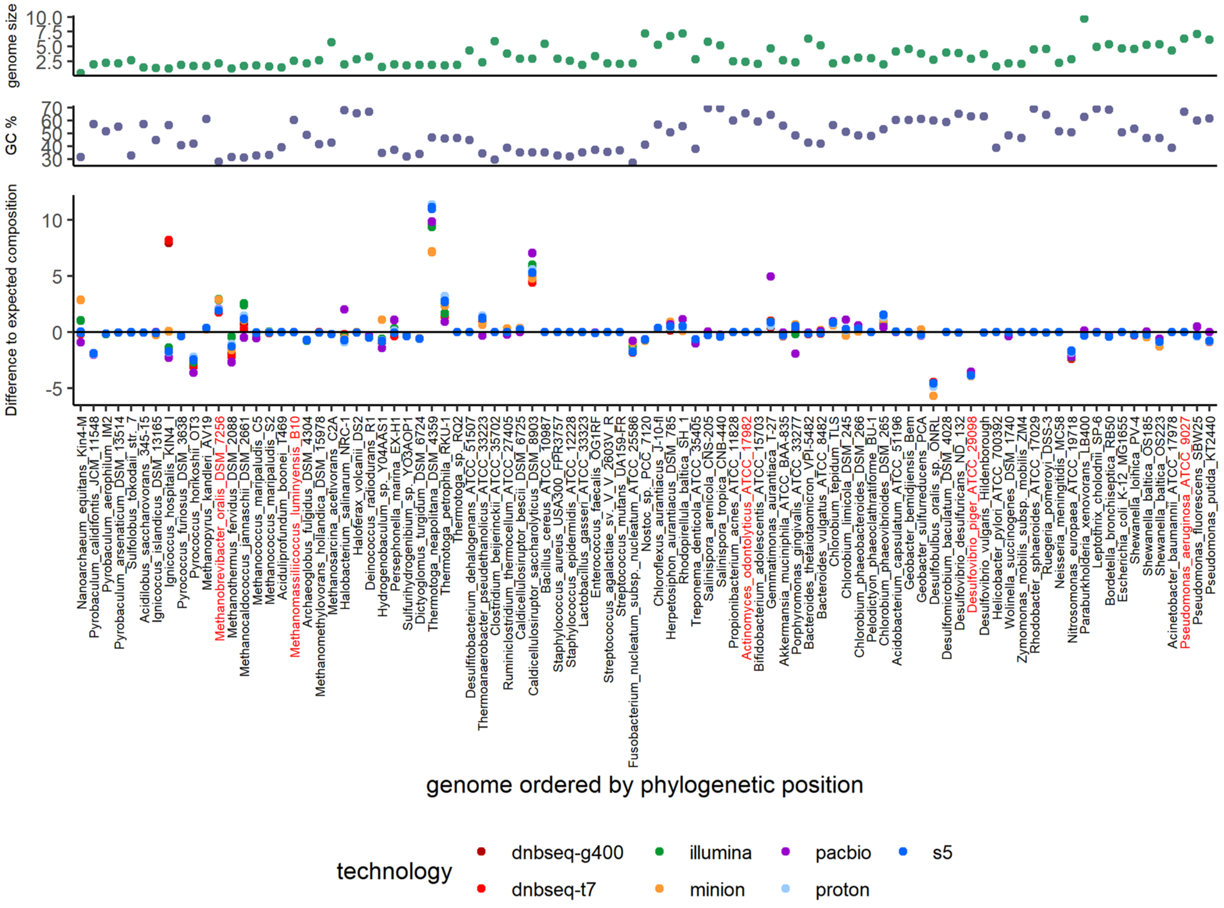

Fig. 3 Differential plot between observed and excepted species abundances in mock1. Abundances (%) for each genome were calculated at 500k depth for each sequencing platform and normalized by genome size. Differential abundance was determined by subtracting the excepted abundances (%) to the observed normalized abundances (%). Positive values, genomes are over-estimated; Negative values, under-estimated. Genomes colored in red indicate draft genomes. Genome size and GC percent are reported for each species.

Sequencer	Ion Proton P1 (spades)	Ion S5 (spades)	Illumina HiSeq 3000 (spades)	DNBSeq G400 (spades)	DNBSeq T7 (spades)	ONT MinION R9 (metaflye)	PacBio Sequel II (metaflye)
Nb Reads (M)	20	20	2 × 10	2 × 10	2 × 10	0.696	0.524
Nb Contigs	45,510	43,879	40,147	44,887	44,603	1,283	**437**
Largest Contig (bp)	384,996	794,907	1,599,668	1,063,396	1,002,925	4,324,150	**7,147,004**
N50 (bp)	7,847	9,089	13,707	8,519	8,184	759,940	**2,013,697**
Genome Fraction(%)	54.767	55.257	61.897	49.397	47.365	44.955	**68.197**
Mismatches per 100kbps	83.29	89.12	47.55	77.22	107.52	339.99	**18.3**
Indels Per 100kbps	77.8	50.03	3.53	3.23	3.67	764.45	11.76
Fully Unaligned Contigs	1,497	1,339	975	735	1,368	231	**6**
Fully Unaligned Length (bp)	900,150	821,545	620,805	426,856	711,992	6,279,694	**134,713**
NB full genome*	5	5	12	7	7	22	**36**

Table 3. Mock1 metaquast assembly report. Only contigs >= 500nt were aligned to the mock1 reference genomes. Number of reads, contigs and the size of the largest contig after each assembly are reported, along with: N50 (bp), a common statistic to evaluate the assembly quality; Genome fraction (%), corresponding to the mean percentage of the genome reconstructed during the assembly; the means for Mismtaches per 100kbps and Indels (Insertions/Deletions) per 100kbps, to evaluate the distance of the reconstructed genome to the reference genome). The number of fully unaligned contigs and the respective length (pb) are reported. * Full genome: More than 99% genome recovery.

using Covaris E220 sonicator and AFA microtubes (Covaris, Brighton, UK) in 100 µL to achieve maximum distribution at 150pb. After shearing, Ampure XP purification and Qubit quantification were performed. Sheared DNA (100 ng) was submitted to enzymatic treatment steps (End repair, barecode ligation with IonXpress Barcode Adaptors kit and final 9 cycles of PCR amplification). Ampure XP beads were used for size selection to 150 pb after End repair reaction and for purification after the other enzymatic treatment steps. Libraries were quantified and size controlled by using High Sensitivity Small Fragment kit and Fragment Analyzer (Agilent Technologies Inc., Santa Clara, CA, USA). Libraries' molarity was between 8.000 and 10.000 pM, before normalization at 95 pM and multiplexing for sequencing. Pipetting was performed with Biomek Fxp or Biomek 3000 Liquid handling (Beckman Coulter Inc., Brea, CA, USA). Pre-sequencing step was performed by Ion Chef (Thermo Fisher) for each sequencing device. A first sequencing was performed by Ion Proton with Ion PI HiQ Chef kit (Thermo Fisher) and Ion PI chip kit v3 (Thermo Fisher). Several run were performed including first path, any path with rebalance libraries and final path to get up to 20 million raw reads for each multiplexed sample. A second sequencing was performed by Ion GeneStudio S5 Prime with Ion 550 chef kit and Ion 550 chip kit. A single run was sufficient to obtain up to 20 million raw reads per multiplexed sample.

MGI DNBSEQ-G400 and DNBSEQ-T7 library preparation and sequencing. DNBSEQ-G400 and DNBSEQ-T7 libraries were constructed from 500 ng of HMW DNA and fragmented using Covaris sonicator E220 (Covaris, Brighton, UK). Sheared DNA underwent End repair and A-tailing steps as described in the MGI Easy Universal DNA Library Prep Set User Manual v1 (MGI Tech Co., Shenzen, China). Adapters ligation was performed following the instructions of the MGIEasy DNA Adapters kit, and cleaned up with the provided DNA Clean Beads. PCR amplification was carried out on purified adapter-ligated DNA and cleaned-up again using magnetic beads. After quality control using Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, Waltham, MD, USA), purified PCR products were denaturated and ligated to generate single-strand circular DNA libraries. Barcode libraries were pooled in equal amounts to make DNA Nanoballs (DNB), and sequenced using DNBSEQ-G400 and DNBSEQ-T7 sequencer technologies following the manufacturer's recommendations.

Illumina library preparation and HiSeq 3000 sequencing. DNA libraries have been prepared using the Illumina TruSeq PCR-free HT (Illumina, San Diego, CA, USA), following the manufacturer protocol. Briefly, 2 µg of HMW genomic DNA was fragmented by sonication using Covaris sonicator (Covaris, Brighton, UK). Sheared fragments were cleaned using the Sample Purification Beads provided in the kit, before Ends repair and size selection procedures. Adapters were ligated and libraries underwent an additional cleaning step with magnetic beads. Library quality was assessed using an Advanced Analytical Fragment Analyser (Agilent Technologies Inc., Santa Clara, CA, USA) and libraries were quantified by q-PCR using the Kapa Library Quantification Kit (Illumina, San Diego, CA, US) following the manufacturer's recommendations. Prior to multiplexing, libraries were normalized to 4 nM and equal volumes were pooled together. Final libraries were sequenced on Illumina HiSeq 3000 using a paired-end read length of 2 × 150 pb with the Illumina HiSeq 3000 Reagent Kits.

Oxford Nanopore MinION R9 library preparation and sequencing. Libraries were built with 1D Native barcoding genomic DNA kit (SQK-LSK109 rev E) from Oxford Nanopore Technologies. To increase sequencing yield, 1.5 µg DNA samples were sheared using G-Tube (Covaris) for 2 times 30 seconds at 7,200 rpm. Sheared Fragments (1 µg), of length comprised between 8 and 9 kb, underwent end repair and A-Tailing (New England Biolabs M6630L and E7546L kits). Next, 500 ng of repaired DNA was ligated with adapter barcode using Native Barcoding Expansion 1–12 kit (EXP-NBD104) and Blunt/TA Ligase (New England Biolabs, M0367L). Native barcode ligated DNA was quantified with Qubit and Fragment Analyzer. Equimolar libraries were pooled for a total quantity of 700 ng to ligate to the sequencing adapters. At this step, we choose the Long Fragment Buffer (LFB) from SQK-LSK109 kit to increase the recovery of 3 kb or longer fragments. Pooled libraries were loaded onto R9 FLO-MIN106 flowcell and sequencing was performed during 48 hours.

Pacific Biosciences Sequel II library preparation and sequencing. For Pacific Biosciences Sequel II sequencing, 500 ng of HMW genomic DNA was used to make unamplified libraries using the SMRTbell® Express Template Prep Kit 2.0. First, gDNA was sheared to a targeted fragment size of 12 kb using Megaruptor and Long Hydropores (Diagenode, Denville, NJ, USA). Sheared gDNA were concentrated using AMPure PB Beads according to the manufacturer recommendations (Pacific Biosciences, Menlo Park, CA, USA) and underwent two treatment procedures for DNA damage repair and end-repair. Barcoded overhang Hairpins adapters from the manufacturer were ligated to the fragment ends to create SMRTbell templates used for sequencing. SMRTbell templates were purified using an exonuclease procedure to remove any free ends molecules or no adapter templates. Then, size-selection was conducted using Ampure PB beads at a concentration of 0.45X to ensure the removing of short fragments. Our mock SMRTbell template was multiplexed with tree additional samples (not included in this study) to equal molarity. On the resulting template, fragments < 3 kb were removed using an additional diluted Ampure PB beads procedure. PacBio primer v2 annealing to the SMRTbell template and polymerase binding to the annealed template were achieved before being sequenced with Sequel II sequencer using Chemistry 2.0 and 30-hour movie.

Sequence QC. The raw reads were quality trimmed using software tools with similar trimming parameters to improve technical comparisons. Illumina HiSeq 3000 and DNBSeq G400 and T7 paired-end reads were trimmed with FASTP v.0.20.0[37], using Illumina TruSeq adapaters for the Illumina HiSeq 300 sequencer and DNBSeq adapters for the DNBSeq G400 and T7. The minimum read length after trimming was 45nt and all reads with a single N nucleotide or unpaired reads after trimming were discarded. The Ion S5 and Ion Proton reads were trimmed using AlienTrimmer v2.0[38] by providing the Ion S5 and Proton contaminants and using the

	MOCK1 (71 species)	MOCK2 (87 species)	MOCK3 (64 species)
Illumina HiSeq 3000	ERR9765446	ERR9765447	ERR9765448-49
Ion Proton P1	ERR9765780-58	ERR9765759-67	ERR9765768-76
Ion S5	ERR9765777	ERR9765778	ERR9765779
ONT Minion R9	ERR9765780	ERR9765781	ERR9765782
PacBio Sequel II	ERR9765783	NA	NA
DNBSEQ-G400	ERR9765742	ERR9765743	NA
DNBSEQ-T7	ERR9765744	ERR9765745	NA

Table 4. Shotgun metagenomic datasets description. Run accession numbers were reported for each sample and technology. Metagenomic data have been deposited under BioProject number PRJEB52977.

following parameters for trimming: "-k 10 -l 45 -m 5 -p 40 -q 20". The minION R9 reads were base called and quality trimmed using Guppy v2.3.1 + 9514fbc[39] with the kit SQK-LSK109 and the barcoding kit EXP-NBD103. The PacBio CSS reads were processed through PacBio custom pipeline. Finally, all PacBio and MinION reads shorter than 500nt were discarded.

Read mapping procedures. All read mapping procedures were performed on reference genomes corresponding to the expected mock composition. For Illumina, DNBSEQ G400 and T7 platforms, mapping was done with bowtie2 v2.3.5.1[40] using paired-end best hit end-to-end match and sensitive presets parameters. For Ion Proton and S5, bowtie2 with single-end best-hit end-to-end match and sensitive presets parameters. For MinION and PacBio, mapping was performed with minimap2 version 2.15-r915-dirty[41] using default parameters, soft clipping activated and by keeping only the best hit.

Read subsampling. Subsampling was performed by a python script using the random library and differential analysis in Figs. 2 and 3 between the observed and expected mock composition at different depth, from 10k to 1 M reads were performed under R version 3.6.0 using *stats*, *ggplot2*, *data.table* and *reshape2* packages.

Metagenomic assembly. The Illumina HiSeq 3000, DNBSeq G400 and T7 paired-end reads were assembled with SPAdes v3.14.1[42] with "--meta" presets and kmer iteration "--k 21,33,55" for DNBSeq, and "--k 21,33,55,77" for Illumina to account for their respective maximal read length. The Ion Proton and Ion S5 single reads were also assembled with SPADES, using the "--iontorrent" and "--careful" flag, as the "--meta" flag is not available for single reads, and a kmer iteration "--k 21,33,55,77". MinION and Pacbio were assembled with metaFlye v2.8.1-b1688[43] using the "--meta" preset, "--plasmids" to recover short unassembled plasmids, a minimum overlap of 2000nt, "--pacbio-hifi--hifi-error 0.003" for Pacbio and "--nano-raw" for MinION reads. Finally, the assemblies' quality was assessed using metaquast v4.6.3[44].

Hybrid metagenomic assembly. Hybrid assemblies were generated using SPAdes v3.14.1[42] with the same parameters previously described and by adding –pacbio and –nanopore parameters when combining with PacBio reads or MinION reads respectively.

Data Records

Shotgun metagenomes are publicly available without restriction in the EMBL-EBI European Nucleotide Archive (ENA) under accession number PRJEB52977[45]. All binning and taxonomy assignment results and parameters are available as a publicly shared KBase narrative (https://narrative.kbase.us/narrative/125743) and can also be seen at Figshare[46].

Technical Validation

Library QC checks. Before library preparation by the different sequencing platforms, gDNA mock samples were required to pass quality and quantity controls. Initial DNA quality control included DNA quantification using Quant-iT™ dsDNA Assay Kit broad range (Q33130) reading by FiltermaxF3 (Molecular Devices, Sunnyvale, CA, USA), Qubit dsDNA Assay Kit (Q32853) and fragment analysis using HS Genomic DNA kit (DNG-488-500) on Fragment Analyzer (Agilent Technologies Inc., Santa Clara, CA, USA). The size of initial DNA peak was between 14 and 17 kb without major degradation smear. Additional specific technical validations for DNA integrity were required during each sequencing library preparation to ensure high quality of the final libraries on each platform. Depending on the sequencing technology, these validation steps typically included QC checks after DNA shearing, size selections, purifications on magnetic stands and on the pooled final libraries.

Usage Notes

Run accession numbers for all metagenomic samples, accessible in the ENA website (PRJEB52977), are fully described in Table 4.

The protocols and datasets we are presenting in this work can be reused for different applications, in particular to benchmark and improve metagenomic assemblers, taxonomic profilers and binning software. As an example for binning applications, we used three binning software (CONCOCT[47] v.1.1, MetaBAT2[48] v1.7 and MaxBin2[49] v2.2.4) by importing the mock1 Illumina HiSeq 3000 reads, assemblies and hybrid assemblies into KBase[50] (https://narrative.kbase.us/narrative/125743), with minimum contig size of 1500 nt (Supplementary

Table S3). The bins were then optimized using DAS Tool[51] v1.1.2. We observed comparable number of binned MAGs corresponding to reference genomes using hybrid assembly and higher than using the Illumina short reads dataset alone. With all assemblies and datasets, the recovery of high quality MAGs was not successful for very low abundance genomes, present at less than 0.1% of the mock1 community (Supplementary Table S3).

Code availability

All reference genomes and scripts for mapping, assembly, genome coverage estimation, subsampling and correlation calculations associated with tables and figures are available at https://forgemia.inra.fr/metagenopolis/benchmark_mock.

Received: 2 June 2022; Accepted: 4 October 2022;
Published online: 11 November 2022

References

1. Almeida, M. *et al*. Construction of a dairy microbial genome catalog opens new perspectives for the metagenomic analysis of dairy fermented products. *BMC Genomics* **15** (2014).
2. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci. Data* **5** (2017).
3. Pasolli, E. *et al*. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649–662.e20 (2019).
4. Almeida, A. *et al*. A new genomic blueprint of the human gut microbiota. *Nature* **568**, 199–504 (2019).
5. Venter, J. C. *et al*. Environmental Genome Shotgun Sequencing of the Sargasso Sea J. Craig Venter. *Science (80-.).* **304**, 66–74 (2004).
6. Tringe, S. G. *et al*. Comparative metagenomics of microbial communities. *Science* **308**, 554–7 (2005).
7. Qin, N. *et al*. Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59–64 (2014).
8. Uritskiy, G. & DiRuggiero, J. Applying Genome-Resolved Metagenomics to Deconvolute the Halophilic Microbiome. *Genes (Basel)* **10**, 220 (2019).
9. Fromentin, S. *et al*. Microbiome and metabolome features of the cardiometabolic disease spectrum. *Nat. Med.* **28**, 303–314 (2022).
10. Segerman, B. The Most Frequently Used Sequencing Technologies and Assembly Methods in Different Time Segments of the Bacterial Surveillance and RefSeq Genome Databases. *Front. Cell. Infect. Microbiol* **10**, 1–7 (2020).
11. Van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. The Third Revolution in Sequencing Technology. *Trends Genet.* **34**, 666–681 (2018).
12. Athanasopoulou, K. *et al*. Third-Generation Sequencing: The Spearhead towards the Radical Transformation of Modern Genomics. *Life* **12**, 30 (2021).
13. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
14. Shakya, M. *et al*. Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Env. Microbiol* **15**, 1882–1899 (2014).
15. Sevim, V. *et al*. Shotgun metagenome data of a defined mock community using Oxford Nanopore, PacBio and Illumina technologies. *Sci. Data* **6**, 285 (2019).
16. Singer, E. *et al*. Next generation sequencing data of a defined microbial mock community. *Sci. Data* **3**, 160081 (2016).
17. Nicholls, S. M., Quick, J. C., Tang, S. & Loman, N. J. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *Gigascience* **8**, giz043 (2019).
18. Tourlousse, D. M. *et al*. Characterization and Demonstration of Mock Communities as Control Reagents for Accurate Human Microbiome Community. *Microbiol Spectr* **10**, e0191521 (2022).
19. Hu, Y., Fang, L., Nicholson, C. & Wang, K. Implications of Error-Prone Long-Read Whole- Genome Shotgun Sequencing on Characterizing Reference Microbiomes. *iScience* **23** (2020).
20. Yang, C., Chu, J., Warren, L. & Inanc, B. NanoSim: nanopore sequence read simulator based on statistical characterization. *Gigascience* **6**, 1–6 (2017).
21. Moss, E. L., Maghini, D. G. & Bhatt, A. S. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat. Biotechnol.* **38**, 701–707 (2020).
22. Peabody, M. A., Van Rossum, T., Lo, R. & Brinkman, F. S. L. Evaluation of shotgun metagenomics sequence classification methods using in silico and *in vitro* simulated communities. *BMC Bioinformatics* **16**, 1–19 (2015).
23. Alili, R. *et al*. Exploring Semi-Quantitative Metagenomic Studies Using Oxford Nanopore Sequencing: A Computational and Experimental Protocol. *Genes (Basel)* **12**, 1496 (2021).
24. Lahens, N. F. *et al*. A comparison of Illumina and Ion Torrent sequencing platforms in the context of differential gene expression. *BMC Genomics* **18**, 602 (2017).
25. Tyler, A. D. *et al*. Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. *Sci. Rep.* **8**, 10931 (2018).
26. Hon, T. *et al*. Highly accurate long-read HiFi sequencing data for five complex genomes. *Sci. Data* **7**, 399 (2020).
27. Kim, H. *et al*. Comparative analysis of 7 short-read sequencing platforms using the Korean Reference Genome: MGI and Illumina sequencing benchmark for whole-genome sequencing. *Gigascience* **10**, 1–9 (2021).
28. Anslan, S. *et al*. Highly comparable metabarcoding results from MGI-Tech and Illumina sequencing platforms. *PeerJ* **9**, e12254 (2021).
29. Bowers, R. M. *et al*. Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. *BMC Genomics* **16** (2015).
30. Jones, M. B. *et al*. Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proc Natl Acad Sci USA* **112**, 14024–14029 (2015).
31. Costea, P. I. *et al*. Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.* **35**, 1069–1076 (2017).
32. Hillmann, B. *et al*. Evaluating the information content of shallow shotgun metagenomics. *mSystems* **3**, e00069–18 (2018).
33. Sczyrba, A. *et al*. Critical Assessment of Metagenome Interpretation — a benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).
34. Meyer, F. *et al*. Critical Assessment of Metagenome Interpretation: the second round of challenges. *Nat. Methods* **19**, 429–440 (2022).
35. Parks, D. H. *et al*. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, 785–794 (2022).
36. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, 256–259 (2019).
37. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, 884–890 (2018).

38. Criscuolo, A. & Brisse, S. ALIENTRIMMER: A tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. *Genomics* **102**, 500–506 (2013).
39. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**, 1–10 (2019).
40. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–360 (2012).
41. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
42. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
43. Kolmogorov, M. *et al.* metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).
44. Mikheenko, A., Saveliev, V. & Gurevich, A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* **32**, 1088–1090 (2016).
45. *ENA European Nucleotide Archive* https://identifiers.org/ena.embl:PRJEB52977 (2022).
46. Meslier, V. *et al.* Benchmarking second and third-generation sequencing platforms for microbial metagenomics, *Figshare*, https://doi.org/10.6084/m9.figshare.21261396.v1 (2022).
47. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
48. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
49. Wu, Y., Simmons, B. A. & Singer, S. W. MaxBin 2. 0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
50. Arkin, A. P. *et al.* KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat Biotechnol.* **36**, 566–569 (2018).
51. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).

Acknowledgements

This work was supported by the European FP7 Marie Skłodowska-Curie actions AgreenSkillsPlus PCOFUND-GA-2013-609398 grant to MA. Additional funding was from the Metagenopolis grant ANR-11-DPBS-0001. MP was supported by the U.S. Department of Energy, Office of Science, Biological, and Environmental Research as part of the Genomic Science Program (the Plant Microbe Interfaces SFA), by the Subsurface Biogeochemical Research Program (Biogeochemical Transformations at Critical Interfaces SFA) and by grant R01DE024463 from the National Institute of Dental and Craniofacial Research of the US National Institutes of Health. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725. We thank Dr. Cynthia Gilmour (Smithsonian Research Institute, Edgewater, Maryland, USA) for providing some of the purified gDNA used in this study. We also thank David Stucki, Deborah Moine and Jules Bourgon (Pacific Biosciences, Menlo Park, California, USA) for PacBio sequencing, Jihua Sun and Yong Hou (BGI, Shenzhen, Guangdong, China) for DNBseq sequencing and Clemence Genthon (Plateforme Génomique GeT INRAE Transfert, Toulouse, France) for Illumina sequencing.

Author contributions

M.A., M.P. and H.R. conceived the project. M.P. generated mock communities. M.A. and V.M. analysed the data and generated the figures. B.Q. performed library preparation and sequencing for Ion proton, DNBseq and Minion technologies. N.P. deposited the data. F.P.O., M.P. and K.D.S. contributed to data analysis. V.M., M.A., B.Q. and M.P. wrote the paper. All authors revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-022-01762-z.

Correspondence and requests for materials should be addressed to M.P. or M.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.