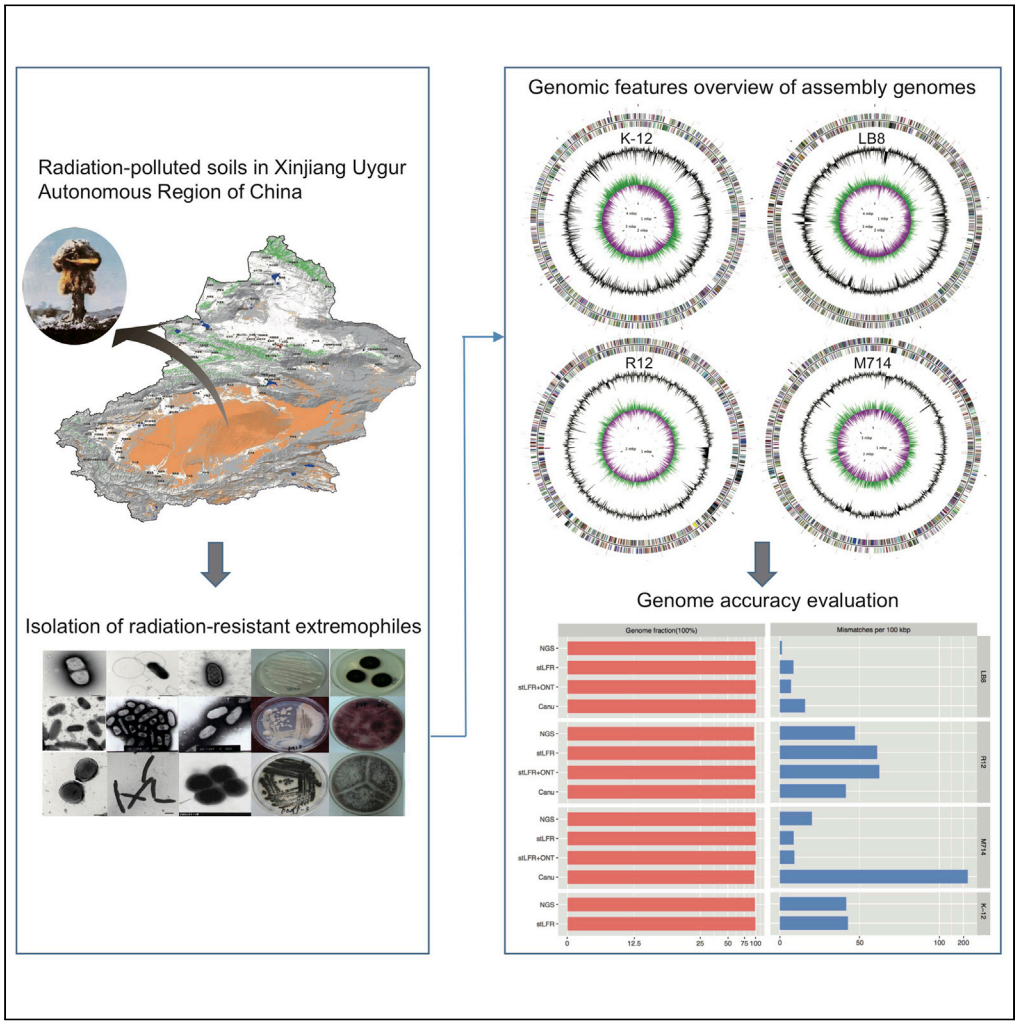
iScience

CellPress
OPEN ACCESS

Article

Comparison of different sequencing strategies for assembling chromosome-level genomes of extremophiles with variable GC content



Zhidong Zhang, Guilin Liu, Yao Chen, ..., He Huang, Ling Jiang, Jianwei Chen

jiangling@njtech.edu.cn (L.J.)
chenjianwei@genomics.cn (J.C.)

HIGHLIGHTS

Assembling and evaluating bacterial chromosome-level genomes by multiple strategies

The stLFR data can assemble chromosome sequences of different GC extremophiles

Zhang et al., iScience 24, 102219
March 19, 2021
https://doi.org/10.1016/j.isci.2021.102219

iScience

CellPress
OPEN ACCESS

Article

Comparison of different sequencing strategies for assembling chromosome-level genomes of extremophiles with variable GC content

Zhidong Zhang,[1,2,10] Guilin Liu,[3,10] Yao Chen,[1] Weizhen Xue,[3] Qianyue Ji,[3] Qiwu Xu,[3] He Zhang,[3] Guangyi Fan,[3,7] He Huang,[4,5] Ling Jiang,[6,*] and Jianwei Chen[3,7,8,9,11,*]

SUMMARY

In this study, six bacterial isolates with variable GC, including *Escherichia coli* as mesophilic reference strain, were selected to compare hybrid assembly strategies based on next-generation sequencing (NGS) of short reads, single-tube long-fragment reads (stLFR) sequencing, and Oxford Nanopore Technologies (ONT) sequencing platforms. We obtained the complete genomes using the hybrid assembler Unicycler based on the NGS and ONT reads; others were *de novo* assembled using NGS, stLFR, and ONT reads by using different strategies. The contiguity, accuracy, completeness, sequencing costs, and DNA material requirements of the investigated strategies were compared systematically. Although all sequencing data could be assembled into accurate whole-genome sequences, the stLFR sequencing data yield a scaffold with more contiguity with more completeness of gene function than NGS sequencing assemblies. Our research provides a low-cost chromosome-level genome assembly strategy for large-scale sequencing of extremophile genomes with different GC contents.

INTRODUCTION

Based on the severe environments to which extremophiles have adapted, they include thermophiles, psychrophiles, alkalophiles, acidophiles, barophiles, and radiation-resistant organisms (Mao et al., 2017; Orellana-Saez et al., 2019; Swarup et al., 2014; Urbieta et al., 2015). These microbes thrive in ecological niches such as deep-sea hydrothermal vents, hot springs, geysers, salt flats, deserts, natural lakes, sulfuric fields, and so on (Brito et al., 2006; DeLong, 2000; Kang et al., 2018; Palmieri et al., 2019; Ziko et al., 2019). Half a century ago, extremophiles received little attention, but they are recently being increasingly explored as sources of basic data as well as useful enzymes for molecular biology and the biotech industry (Merino et al., 2019; Rothschild and Mancinelli, 2001). For example, biocatalysts cloned from extremophiles have had a great impact on the global biotechnological market (Brininger et al., 2018; Mokashe et al., 2018; Schiraldi and De Rosa, 2002; Wang et al., 2019a). The enzymes with the widest applications include Taq DNA polymerase (Chien et al., 1976), heat-tolerant cellulase (Adamiak et al., 2015), alkali-resistant β-D-galactosidase (Wang et al., 2011), etc. Additionally, various CRISPR loci belonging to different CRISPR-Cas systems have been identified in the genomes of extremophiles in recent years, providing a valuable resource for mining efficient gene editing solutions (Makarova et al., 2015). However, the exploitation and utilization of extremophiles is still challenging owing to demanding separation and purification of strains as well as the further mining of their functional genes. What's more, there is still a lack of technology for efficiently mining biological information from extremophiles on a large scale efficiently and at low cost.

Owing to the advances in high-throughput sequencing technology, a large amount of gene sequence data can be acquired in a relatively short time (Metzker, 2010; Niedringhaus et al., 2011). The next-generation sequencing (NGS) technologies, such as Illumina, MGI, Shenzhen, and Ion Proton, have enabled widespread bacterial whole-genome sequencing, producing millions of paired-end reads with a low error rate (0.1%). However, the short reads of 100–300 bp make it challenging to fully reconstruct genomic structures of interest (De Maio et al., 2019). Hybrid assembly based on third-generation sequencing technologies, such as the Oxford Nanopore Technologies (ONT) and SMRT Pacific Biosciences (PacBio) sequencing platforms, combined with NGS short-read sequencing can be used to assemble the complete chromosome and recover plasmid genomes. However, these sequencing strategies require library construction,

[1]College of Biotechnology and Pharmaceutical Engineering, Nanjing Tech University, Nanjing 211816, China

[2]Institute of Applied Microbiology, Xinjiang Academy of Agricultural Sciences/Xinjiang Key Laboratory of Special Environmental Microbiology, Urumqi, Xinjiang 830091, China

[3]BGI-Qingdao, BGI-Shenzhen, Qingdao, Shandong 266555, China

[4]School of Food Science and Pharmaceutical Engineering, Nanjing Normal University, Nanjing 210023, China

[5]School of Pharmaceutical Sciences, Nanjing Tech University, Nanjing 211816, China

[6]College of Food Science and Light Industry, Nanjing Tech University, Nanjing 211816, China

[7]BGI-Shenzhen, Shenzhen, Guangdong 518083, China

[8]Qingdao-Europe Advanced Institute for Life Sciences, BGI-Shenzhen, Qingdao 266555, China

[9]Laboratory of Genomics and Molecular Biomedicine, Department of Biology, University of Copenhagen, Universitetsparken 13, Copenhagen 2100, Denmark

[10]These authors contributed equally

[11]Lead contact

*Correspondence: jiangling@njtech.edu.cn (L.J.), chenjianwei@genomics.cn (J.C.)

https://doi.org/10.1016/j.isci.2021.102219

sequencing on two different platforms, and large amounts of high-quality DNA than NGS sequencing, and are much more costly. When starting a large-scale bacterial whole-genome sequencing (WGS) project, it is a challenge to choose the most cost-effective sequencing strategy and still obtain high-quality genome sequences. Therefore, a new sequencing approach that integrates the low cost and high accuracy with enhanced efficiency for extremophiles is highly desirable but remains challenging.

In the past decades, numerous methods have been developed to capture long-range information with short-read sequencing, including mate-pair (Korbel et al., 2007; Rubin et al., 2007), clonal barcoding methods (e.g., synthetic long reads [Bankevich and Pevzner, 2016; Peters et al., 2012; Voskoboynik et al., 2013], linked reads (Wang et al., 2019b; Zhang et al., 2017; Zheng et al., 2016), and Hi-C]Burton et al., 2013]). Among these, clonal barcoding library technologies (Bankevich and Pevzner, 2016; Peters et al., 2012; Voskoboynik et al., 2013; Wang et al., 2019b; Zhang et al., 2017; Zheng et al., 2016) showed the most promising results in terms of bringing routine long-read capability using second-generation platforms. For example, single-tube long-fragment read (stLFR) technology is a novel WGS library preparation approach that enables efficient WGS, haplotyping, and contig scaffolding on the basis of adding a single unique clonal barcode sequence to sub-fragments of the original DNA in a single-tube process (Wang et al., 2019b). The use of microbeads as miniaturized virtual compartments allows a practically unlimited number of clonal barcodes to be used per sample at a negligible cost. The stLFR method enables short-read NGS systems to generate highly accurate and economical long-read sequencing information for *de novo* genome assembly (Chen et al., 2020).

In the present study, we described for the first time the implementation of stLFR technology to resolve the accurate sequencing of complex extremophile genomes. Different strategies for hybrid bacterial genome assembly were selected and compared, including Illumina, ONT, and PacBio data generated from the same DNA extracts. We selected five radiation-resistant extremophiles isolated from the Xinjiang Uygur Autonomous Region of China (*Bacillus cereus* 43-1A, *Brevibacterium frigoritolerans* 44A, *Rufibacter* sp. LB8, *Deinococcus wulumuqiensis* R12, *Janibacter melonis* M714) as well as *Escherichia coli* K-12 as the reference strain. The GC content of the investigated genomes varied from 30% to 70%. Moreover, extremophilic microbes usually have large genomes of 4.3–6.5 Mb as well as varying numbers of plasmids (Carattoli, 2009). The objective of this work was to evaluate and optimize the accuracy of stLFR technology when sequencing the genomes of extremophiles with different GC content, and the analytical results were compared with both NGS and third-generation sequencing. The conclusion paves the way for rapid, cheap, and accurate generation of completely resolved extremophile genomes to become widely accessible.

RESULTS

High GC bacterial stLFR sequencing and assembly

To determine the optimal conditions for the construction of stLFR libraries for bacteria with high genomic GC content, we used five different concentrations of the interrupting enzyme, ranging from 0.4 to 1.2 pmol/10 ng DNA, to construct the libraries of *D. wulumuqiensis* R12. We generated 2 Gb raw sequencing data for each concentration (Table S1). After filtering, we found that the sequencing reads number and barcode frequency distributions of the five concentration clean reads changed at different enzyme concentrations. When the enzyme concentration was low, less transposon insertion resulted in less fragmentation of the DNA, so that many co-barcode reads with lower barcode frequency perhaps had larger insert size and contributed to genome assembly and scaffolding (Figure 1A). The clean reads of the five concentrations were assembled into draft genome using Supernova. We found that the estimated molecule lengths and the assembled genome sizes of libraries constructed with different enzyme concentrations were similar, but the scaffold N50 values were significantly different (Table S1, Figure 1B). The scaffold N50 of 0.4 pmol/10 ng DNA assembly genome was 2,905 kb, which accounted for 80% of the genome length, and was significantly higher than the other concentrations with scaffold N50, about 156–402 kb. These results indicated that this enzyme concentration offers assembly results with the most contiguity.

Draft genome assembly using NGS short reads

We used more than 100X NGS clean data for each sample to assemble the draft bacterial genomes using SPAdes (Table S2). The assembled genome sizes of the 6 bacteria ranged from 3.39 Mb (*D. wulumuqiensis* R12) to 5.52 Mb (*B. frigoritolerans* 44A), with scaffold N50 values ranging from 34.67 kb (*D. wulumuqiensis* R12) to 973 kb (*J. melonis* M714) (Table 1, Figure 2). The CheckM genome quality evaluation showed that the completeness of all genomes was higher than 97% and contamination was lower than 2.5%, reflecting the high quality of each draft genome (Table 1, Figure 2). Accordingly, the estimated genome sizes ranged

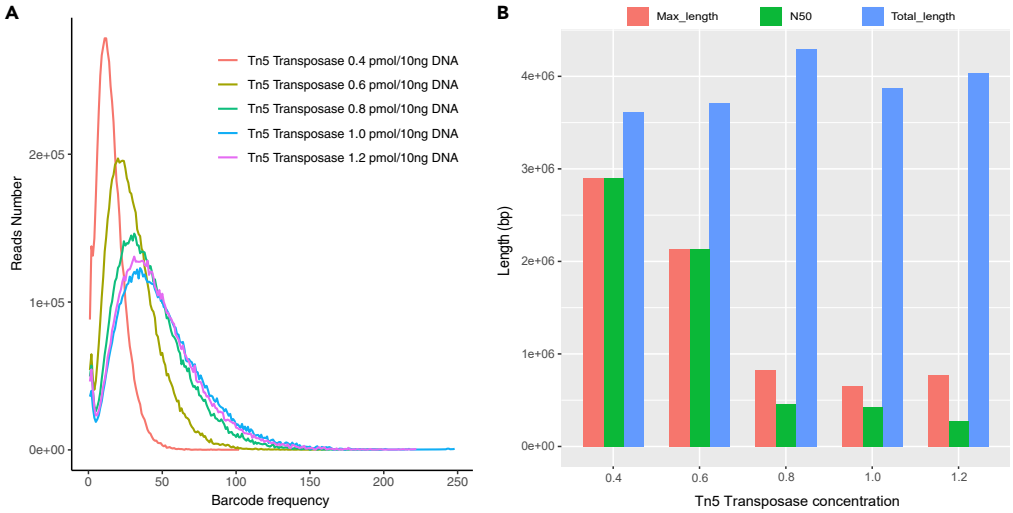


Figure 1. The five different enzyme concentrations/ng DNA using stLFR sequencing of *D. wulumuqiensis* R12

(A) The barcode frequency distribution of five conditions.

(B) The Supernova-assembled statistics of five conditions.

from 3.18 to 6.15 Mb according to the 17 bp k-mer frequency distribution (Figure S2). Thus the assembled genomes were close to the estimated sizes.

Complete genome assembly

For ONT sequencing, we generated 2.22, 1.32, and 2.91 Gb data, with N50 of 20.07, 36.49, and 29.60 kb, and a median read quality of 11 for strains *Rufibacter* sp. LB8, *D.* wulumuqiensis R12, and *J.* melonis M714, respectively (Table S4). We used all ONT reads longer than 8 kb to assemble the bacterial chromosome using Unicycler software based on the NGS reads. Additionally, the ONT reads were directly assembled into the complete genomes using Canu. All the genomes were polished using NGS short reads to fix base errors in Pilon and GATK. For the three ONT Unicycler assembly genomes, the final corrected genomes had circular chromosome sequences and lengths close to the k-mer estimated genome sizes. Additionally, some had circular plasmid sequences. All the genomes had high accuracy, with the genomics completeness >99%, single-base accuracy rates >99.8%, and structural accuracy rates >99.7%. Thus the results were of sufficiently high quality to be regarded as reference genomes of these strains. However, in the high-GC-content genomes assembled using Canu, the genomics completeness was less than Unicycler and also displayed low structural accuracy (91.61% for M714 and 95.89% for R12), which was lower than that of the other genome assemblies, including NGS assemblies (Table 1). This indicated that the quality of genomes assembled from ONT reads with high error rate using Canu was lower than that of the hybrid genome assemblies obtained using Unicycler.

Chromosome-level genome assembly using stLFR

We obtained more than 2.5 Gb clean stLFR data from the six samples (Table S2). To find the best assembly method for bacterial stLFR data, we used SPAdes, cloudSPAdes, Athena, Architect, and Supernova to assemble all clean reads for each genome, and SLR-scaffolder was used to link the scaffolds. Owing to the occurrence of large gaps during scaffolding, the Supernova-assembled genomes of high-GC strains *D. wulumuqiensis* R12 and *J. melonis* M714 were larger than those produced by the other assembly methods. However, the other strains had the same genome sizes with different assembly methods (Figure S1). We found that the scaffold N50 of the Supernova assembly result was higher than that of the other algorithms, and the scaffold N50 length was consistent with the longest scaffold length and accounted for more than 95% of the total assembled genome length (Figure S1), which indicated that it was a chromosome-level scaffold. Among barcoding-based synthetic long-read assembly algorithms, Supernova gave the best assembly results, followed by Athena and cloudSPAdes. We used the best chromosome-level scaffold assemblies for subsequent comparative analysis (Table 1).

Furthermore, in the chromosome-level scaffolds of *Rufibacter* sp. LB8, *D.* wulumuqiensis R12, and *J. melonis* M714, ONT reads were used to close the gaps and obtain the complete genomes. We

Table 1. Statistics of three sequencing strategies assembly genomes

Sample	Method	Scaf (#)	Length (bp)	Gap (bp)	N50 (bp)	N90 (bp)	GC %	ContigN50 (bp)	Mapped rate %	Single base %	Structure %	Completeness %	Contamination %	Gene (#)	Gene AvgL	Gene Check	M16SrRNA	Repeat L (bp)
LB8	NGS	28	4,730,895	100	356,460	127,204	50.28	306,200	92.26	99.99	99.66	99.97	1.04	4,056	985.10	99.88	1	10,640
LB8	stLFR	2	4,746,533	1,025	4,591,406	4,591,406	50.30	2,795,707	92.26	99.98	99.94	99.97	1.04	4,089	981.57	99.80	3	9,909
LB8	stLFR + ONT	2	4,746,090	0	4,590,963	4,590,963	50.30	4,590,963	92.27	100.00	99.92	99.97	1.04	4,086	982.57	99.80	3	9,909
LB8	Unicycler	2	4,746,099	0	4,590,972	4,590,972	50.30	4,590,972	92.27	100.00	99.94	99.97	1.04	4,087	982.35	99.94	3	9,916
LB8	Canu	2	4,875,130	0	4,654,191	4,654,191	50.25	4,654,191	92.25	99.75	99.68	99.97	3.72	4,299	956.03	99.90	3	10,481
M714	NGS	33	3,483,103	300	973,846	862,807	72.89	973,846	97.8	99.77	99.62	99.82	0.18	3,353	955.71	99.28	1	25,845
M714	stLFR	2	3,480,703	10	3,426,494	3,426,494	72.90	1,978,253	97.74	100.00	99.81	99.82	0	3,360	956.86	99.28	2	25,499
M714	stLFT + ONT	2	3,478,886	0	3,426,533	3,426,533	72.90	3,426,533	97.82	100.00	99.81	99.82	0	3,358	957.18	99.28	2	25,374
M714	Unicycler	2	3,481,073	0	3,426,637	3,426,637	72.99	3,426,637	97.85	100.00	99.85	99.82	0	3,359	956.54	99.28	2	25,531
M714	Canu	1	3,357,952	0	3,357,952	3,357,952	72.99	3,357,952	83.8	97.90	91.61	85.99	0	3,321	891.49	85.99	2	28,332
R12	NGS	207	3,392,156	149	34,666	9,895	66.19	33,013	96.48	99.83	95.29	97.88	0.85	3,218	904.53	96.61	1	24,824
R12	stLFR	5	3,577,039	2,430	2,869,672	286,860	66.01	118,540	97.41	99.35	96.52	98.73	2.75	3,430	903.33	98.73	3	26,062
R12	stLFT + ONT	5	3,610,754	0	2,904,890	286,860	65.98	2,904,890	98.26	99.42	97.20	99.58	2.75	3,474	902.14	99.15	3	24,804
R12	Unicycler	5	3,505,947	0	2,857,585	323,544	66.05	2,857,585	98.41	99.89	98.06	99.58	0.21	3,335	913.47	99.15	3	26,574
R12	Canu	4	3,577,988	0	2,874,385	2,874,385	65.90	3,016,002	95.03	99.62	96.50	97.14	0.21	3,701	819.63	97.35	3	34,079
K-12	NGS	861	4,998,809	350	132,349	5,037	49.88	132,349	97.33	91.85	98.65	100.00	2.25	4,520	900.24	99.97	2	26,612
K-12	stLFR	6	4,578,448	1,040	4,561,935	4,561,935	50.77	281,469	97.35	99.97	99.72	99.97	0.04	4,368	928.29	99.97	7	9,426
K-12	Ref	1	4,502,758	0	4,502,758	4,502,758	50.78	4,502,758	96.04	99.98	99.69	99.37	0.04	4,276	932.16	99.37	7	20,057
44A	NGS	98	5,515,358	210	601,314	110,298	40.53	539,757	98.13	99.49	99.47	98.63	1.84	5,410	827.50	98.63	1	46,210
44A	stLFR	8	5,574,407	91,277	4,239,514	264,146	40.43	264,146	97.85	98.36	99.39	98.63	1.39	5,417	828.10	98.63	3	32,708
43-1A	NGS	62	5,442,287	100	429,469	80,246	35.26	372,952	97.03	97.38	99.64	98.61	0.35	5,600	824.33	98.61	1	42,164
43-1A	stLFR	13	5,577,895	146,220	4,637,631	426,869	35.26	282,204	96.77	99.96	99.54	98.61	0.33	5,621	823.66	98.61	3	24,531

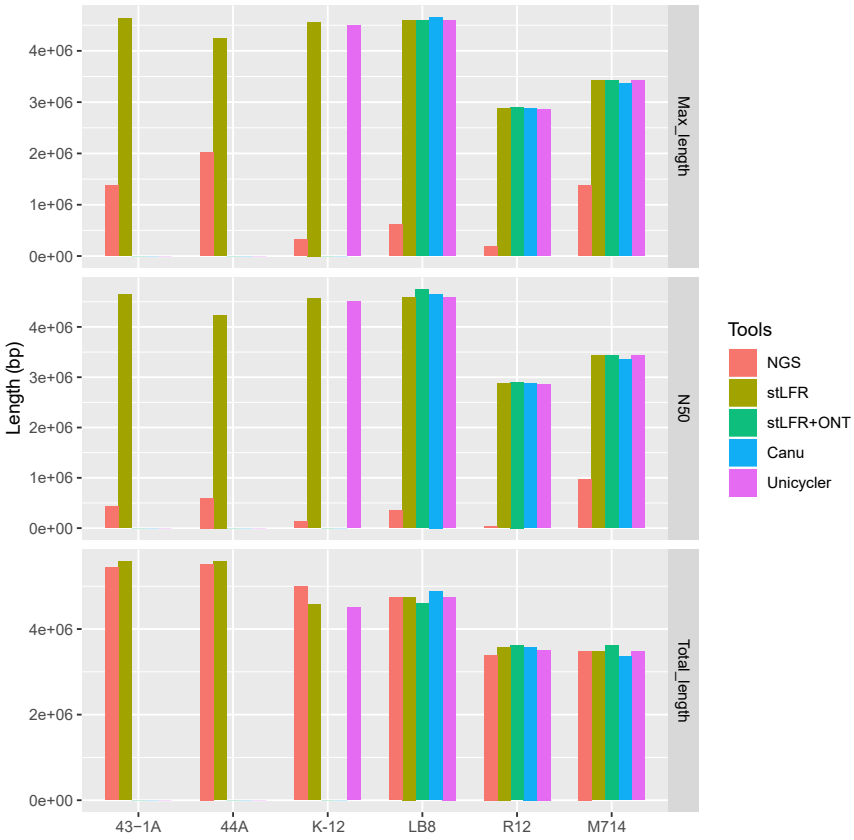


Figure 2. The three sequencing strategies-assembled genome statistics of the six bacterial strains

The total assembly genome length (bottom), N50 length (middle), and maximum scaffold length (top) by using different algorithms are shown. The assembly algorithms of each sample from left to right are NGS draft genome, stLFR chromosome scaffolds, stLFR + ONT complete genome, ONT complete genome assembled by Canu, and Hybrid complete genome assembled by Unicycler using ONT reads and NGS reads.

generated the complete chromosome and plasmid scaffolds for the three samples. The genomes with closed gaps had higher accuracy than the chromosome-level scaffolds, reaching similar genome sizes, completeness, and accuracy as those of the Unicycler assemblies (Table 1).

Comparison of the genome assemblies

We compared the assembly results from the three sequencing strategies using various methods. First, the draft genomes were assembled from NGS data using SPAdes (NGS genomes). Second, the chromosome-level genomes were assembled from stLFR data using Supernova (or Athena) and SLR-scaffolder (stLFR genomes). The hybrid assembled complete genomes were obtained after closing the gaps in TGS-Gapfiller using the ONT reads (stLFR + ONT genomes). Finally, the complete reference genomes were assembled based on the Nanopore and NGS data using Unicycler, except for the downloaded NCBI reference genome CP011124.1, which was assembled from PacBio long sequencing reads as the representative complete genome of *E. coli* K-12 (Table 1) (Tharek et al., 2017). For comparison, we assembled the complete genome in Canu, using only ONT reads. For all assembly results, the genome lengths were consistent with the k-mer estimated results (Figure 2, Table1), and the completeness assessments of the genomes by CheckM were also very close (~99%) with the high mapped rates of NGS reads (~97%), except for the M714 Canu-assembled genomes (Table 1). This indicated that the assembly results of the long-read sequencing strategies (included stLFR, Nanopore, and PacBio) were reasonable and accurate.

In addition, we compared the structural accuracy of the results of the three sequencing strategies for strains *E. coli* K-12, *Rufibacter* sp. LB8, *D. wulumuqiensis* R12, and *J. melonis* M714. There was no significant difference in the single base accuracy rate and structural accuracy rate between stLFR genomes, stLFR + ONT

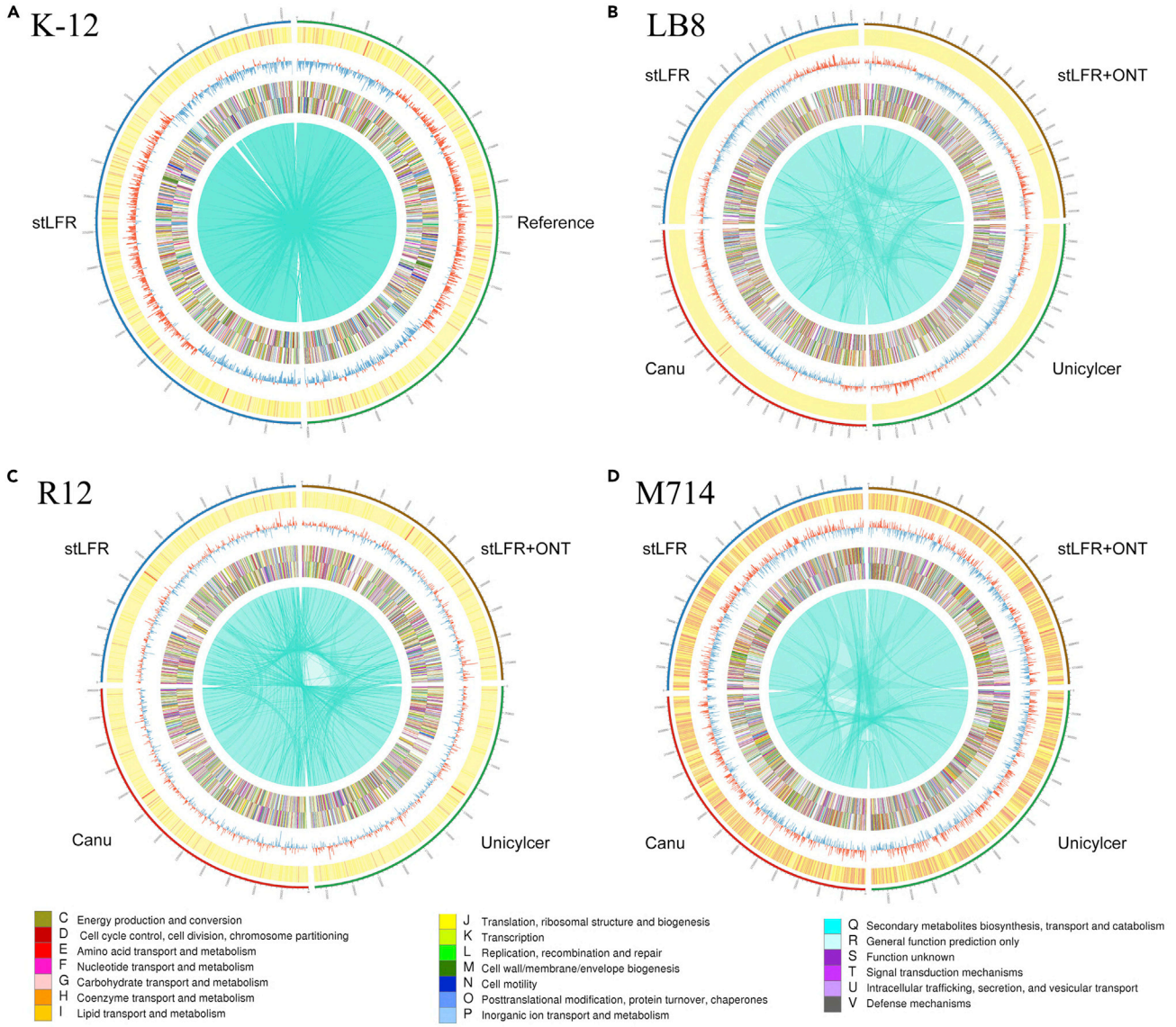


Figure 3. The longest chromosome sequence comparisons for strains

(A–D) (A) *E. coli* K-12 using stLFR-assembled genome and the third-generation sequencing-assembled genome, (B) *Rufibacter* sp. LB8, (C) *D. wulumuqiensis* R12, and (D) *J. melonis* M714 using stLFR-assembled genomes, stLFR + ONT-assembled genomes, ONT + NGS Unicycler-assembled genomes, and ONT Canu-assembled genomes. The outermost circle is GC heatmap, the next circle is the histogram of GC (red: G > C; blue: G < (C), and the middle circle is gene density in chromosomes. The last two circles are the COG positive/negative annotation heatmaps, and the legend is shown at the bottom.

genomes, and third-generation sequencing Unicycler-assembled genomes, whereas it was higher than that of NGS sequencing and Canu-assembled genomes (Table 1). The circular synteny analysis of stLFR genomes, stLFR + ONT genomes, Canu assembly genomes, and Unicycler assembly genomes did not reveal any abnormal structural errors or large fragment indels. Moreover, the gene distribution was even and without bias, which also showed that the assembled structures of the genomes were consistent (Figure 3). For the strains *B. cereus* 43-1A and *B. frigoritolerans* 44A, the synteny analysis also showed that the NGS assembly could completely match the stLFR-assembled genomes (Figure S5). Furthermore, the NGS-, stLFR-, stLFR + ONT-, and Canu-assembled genomes were compared with the complete Unicycler-assembled genomes using QUAST to show the sequence alignment and to estimate the base accuracy, genome fraction, mismatches, and indels per 100 kb. The stLFR assemblies and stLFR + ONT assemblies showed higher consistency and fewer misassembled blocks compared with Unicycler assemblies than the NGS assemblies (Figure 4A). The Canu assemblies also had high consistency but contained more misassembled

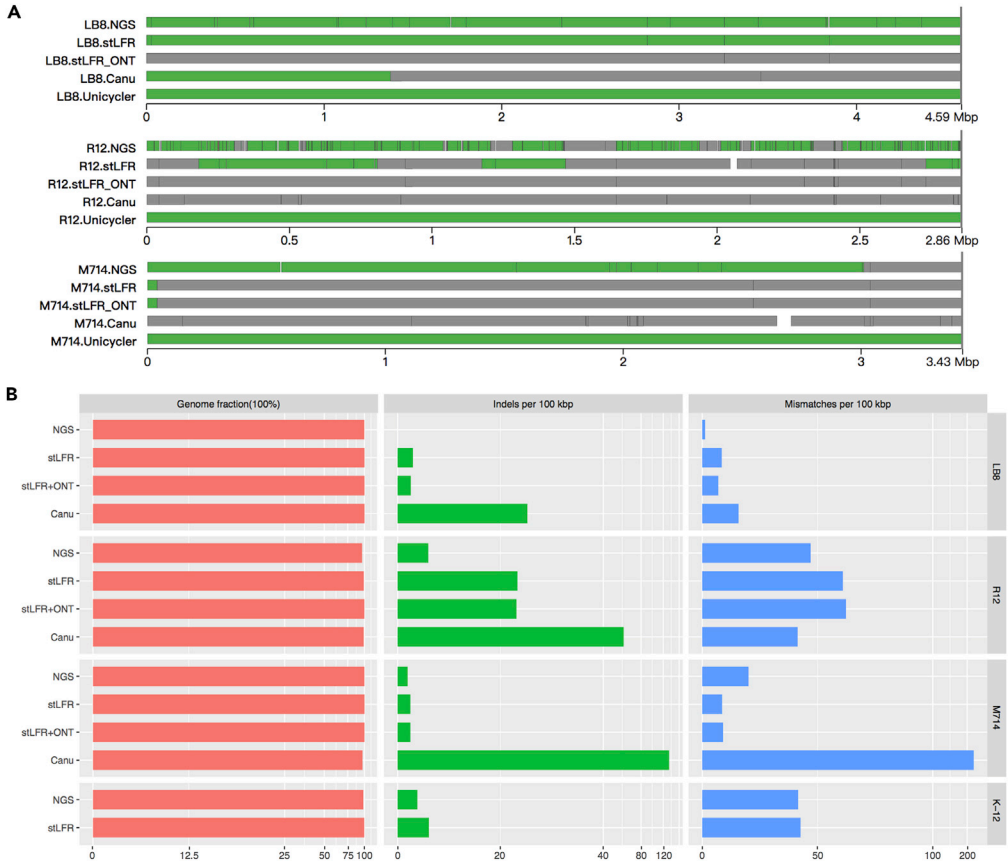


Figure 4. The longest chromosome sequence comparison

(A and B) (A) Sequences alignment viewer and (B) accuracy evaluation of stLFR genomes, stLFR + ONT genomes, Canu-assembled genomes, and NGS genomes when compared with the complete genomes assembled by Unicycler using NGS and ONT reads for the strains *E. coli* K-12, *Rufibacter* sp. LB8, *D. wulumuqiensis* R12, and *J. melonis* M714.

blocks than stLFR assemblies and stLFR + ONT assemblies for the high GC content strains R12 and M714 (Figure 4A). In all the strains, the genome fractions of stLFR assemblies and stLFR + ONT assemblies were higher than those of the NGS assemblies, and except *D. wulumuqiensis* R12, the SNPs and indels percentage between stLFR assemblies, stLFR + ONT assemblies, and NGS assemblies were similar (Figure 4B). Although the Canu assemblies had somewhat higher genome fractions, the stLFR assemblies and stLFR + ONT assemblies had lower mismatches and indels for *Rufibacter* sp. LB8, *D. wulumuqiensis* R12, and *J. melonis* M714. This was especially true for the M714 Canu assemblies, which had more than seven times higher percentages of indels than the other assemblies (Figure 4B).

Finally, we compared the genome components and functional gene annotations of the different assemblies. The gene number and average gene length of stLFR assemblies and stLFR + ONT assemblies were closer to the Unicycler assembly results than the NGS assemblies and Canu assemblies. Furthermore, the stLFR assemblies, stLFR + ONT assemblies, and Canu assemblies detected the same number of 16S rRNA copies as the Unicycler assemblies, whereas the NGS assemblies detected only one 16SrRNA (Table 1). The COG annotation heatmap revealed that the different assembly strategies had the same annotated COG categories for each strain. Compared with the NGS fragment assembly genomes and Canu assemblies, the annotated gene number of each category of stLFR assemblies and stLFR + ONT assemblies was more similar to the Unicycler assemblies (Figures 3 and S3). At the same time, we investigated the KEGG annotation results and found the same patterns. For each strain, the different sequencing strategies produced the same annotated pathways, and the annotated gene number of stLFR assemblies and stLFR + ONT assemblies was similar to the complete reference genomes in each pathway (Figure S4).

DISCUSSION

Current research on extremophiles is mostly focused on the exploration of physiological parameters, such as the extremozymes, to offer an excellent source of replacement for mesophilic ones currently used in biotechnology. The development of rapid and low-cost NGS technologies has cleared the way for exploiting natural genetic diversity and identifying the corresponding functional genes. In this study, we used three different sequencing strategies, including NGS, stLFR technology, and third-generation sequencing (Nanopore or PacBio) to construct libraries, sequence, and assemble the genomes of five extremophilic radiation-resistant strains and *E. coli* K-12. The GC content of the investigated genomes varied from 35% to 72%.

Among the three sequencing strategies, the cost of stLFR is about twice that of NGS, as well as less than one-third the cost of Nanopore sequencing and one-fourth that of PacBio sequencing (Figure S6). In addition, we generally combined the Illumina or MGI short-reads sequencing (NGS) with different long-read sequencing technologies (included Nanopore and PacBio) to obtain the complete accurate assembly of bacterial genomes, which increased the sequencing cost. Additionally, the computational resources required for stLFR are comparable to those needed for NGS and far lower than those needed for third-generation sequencing.

Furthermore, stLFR requires only 1–10 ng high-quality DNA, which is much lower than the 1,500 ng required for Nanopore or PacBio sequencing, and also lower than the 200 ng required for NGS (Tables S2–S4).

We investigated the optimal stLFR library construction conditions for bacteria with a high GC content and found that a low transposase concentration was favorable for sequencing. For all bacteria, we used a transposon that contains the sample index, and one-tenth of the magnetic beads were used for the library construction. Different from previous studies, in which stLFR was applied for animal or plant genomics, our method allowed the pooling and parallel sequencing of a large number of samples, which cannot be achieved by 10X Genomics read cloud sequencing technology (Goodwin et al., 2016). We also investigated the impacts of valid clean data and assembly software on assembly quality. We found that using 2–4 Gb clean data can result in a high-quality assembly, whereby Supernova and Athena were more suitable for the assembly stLFR reads to obtain a complete bacterial genome (Figure S1).

Based on the evaluation of the five genomes assembled using three sequencing strategies, we found that we could obtain chromosome-scale assembly scaffolds from stLFR sequencing data, while achieving the same structural and functional accuracy as the assembly results of third-generation sequencing (Table1, Figure2). Compared with the NGS assembly results, the stLFR assemblies had higher completeness and fewer mismatches. Additionally, we also used stLFR data to assemble the complete plasmid genomes, which was only achieved using third-generation sequencing before (Table1). Furthermore, we used ONT reads to fill the gaps in the stLFR assemblies and obtain the complete genomes. Compared with the ONT Unicycler assemblies, the new assembly method could also generate complete genomes with high accuracy using fewer computational resources.

Many assembly methods were compared for the construction of bacterial complete genomes using Nanopore or PacBio long-read sequencing data (Chin et al., 2016; Danko et al., 2019; De Maio et al., 2019; Koren et al., 2017). Here, we also compared the accuracy of the assembly results produced using Canu to directly assemble ONT data and using the hybrid assembly software Unicycler to assemble ONT data. Owing to the high sequence error rate of ONT reads, we found that the accuracy (included the ratio of mapped reads, genomic completeness, single-base accuracy, and structural accuracy) of the Unicycler hybrid assembly based on the NGS assembly contigs was much higher than that of Canu, especially for bacteria with abnormally high GC content. Similar studies used 10X Genomics long-read cloud sequencing data to assemble microbial genomes (Bishara et al., 2018a, 2018b; Tolstoganov et al., 2019; Weisenfeld et al., 2017). We optimized the conditions for the stLFR library construction and sequencing of bacterial genomes and constructed an stLFR *de novo* assembly pipeline to obtain chromosome-scale bacterial genomes. In conclusion, we have shown that assembling high-quality reference-grade bacterial genomes using stLFR sequencing data is a cost-effective option, especially for bacteria that are difficult to culture or do not yield large amounts of DNA due to challenging extraction. Based on the presented stLFR assembly results, we are confident that it will be possible to fill gaps and optimize ONT sequencing data to obtain the complete genome of any industrially important strain in the future.

Limitations of the study

We used stLFR sequencing technology for the first time to assemble the chromosome-level bacterial genomes, and this approach currently has some limitations. Owing to the short read length of stLFR sequencing, the bacterial chromosome sequences still contained gaps, and we will try to increase the read length using a 200- to 300-bp paired-end sequencing strategy to fill in the gaps caused by short repeats. There are currently few *de novo* assembly algorithms for stLFR data. Although the recently released stLFR *de novo* software is an exception (https://github.com/BGI-biotools/stLFRdenovo), it is based on Supernova and is commonly used in animal or plant genome assembly. It is therefore necessary to develop publicly available tools specifically for the assembly of bacteria stLFR data to obtain better results. In addition, the calling of large variations (such as structural variations, inversions, and copy number variations) in large-scale bacterial sequencing projects based on stLFR data also needs to be addressed further.

Resource availability

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Jianwei Chen (chenjianwei@genomics.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The sequencing data and assembly genomes that support the findings of this study have been deposited into CNSA (CNGB Sequence Archive) of CNGBdb with accession number CNP0001196 and under NCBI BioProject Accession PRJNA665116.

METHODS

All methods can be found in the accompanying Transparent methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102219.

ACKNOWLEDGMENTS

The work supported by the National Natural Science Foundation of China (31922070, 32060004), the National Science and Technology Major Project of China (2017ZX10303406), the Natural Science Foundation of Jiangsu Province (BK20180038), Tianshan Pine Plan (2017XS26), and Basic Scientific R&D Program for Public Welfare Institutes in Xinjiang (KY2019023, KY 2019019).

AUTHOR CONTRIBUTIONS

Z.Z., writing – original draft preparation and investigation; G.L., application and calculation analysis; Y.C., data curation and software; H.H., conceptualization; W.X. and Q.J., methodology; Q.X., validation; H.Z. and G.F., formal analysis; L.J. and J.C., writing, editing, and funding acquisition.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We worked to ensure diversity in experimental samples through the selection of the genomic datasets. The author list of this paper includes contributors from the location where the research was conducted who participated in the data collection, design, analysis, and/or interpretation of the work.

Received: October 25, 2020
Revised: January 20, 2021
Accepted: February 18, 2021
Published: March 19, 2021

REFERENCES

Adamiak, J., Otlewska, A., and Gutarowska, B. (2015). Halophilic microbial communities in deteriorated buildings. World J.Microbiol.Biotechnol. *31*, 1489–1499.

Bankevich, A., and Pevzner, P.A. (2016). TruSPAdes: barcode assembly of TruSeq synthetic long reads. Nat. Methods *13*, 248–250.

Bishara, A., Moss, E.L., Kolmogorov, M., Parada, A.E., Weng, Z., Sidow, A., Dekas, A.E., Batzoglou, S., and Bhatt, A.S. (2018a). High-quality genome sequences of uncultured microbes by assembly of read clouds. Nat.Biotechnol. *36*, 1067–1075.

Bishara, A., Moss, E.L., Tkachenko, E., Kang, J.B., Zlitni, S., Culver, R.N., Andermann, T.M., Weng, Z., Wood, C., Handy, C., et al. (2018b). De novo assembly of microbial genomes from human gut metagenomes using barcoded short read sequences. bioRxiv, 125211.

Brininger, C., Spradlin, S., Cobani, L., and Evilia, C. (2018). The more adaptive to change, the more likely you are to survive: protein adaptation in extremophiles. Semin.CellDev. Biol. *84*, 158–169.

Brito, J.A., Bandeiras, T.M., Teixeira, M., Vonrhein, C., and Archer, M. (2006). Crystallisation and preliminary structure determination of a NADH: quinoneoxidoreductase from the extremophile Acidianusambivalens. Biochim.Biophys.Acta *1764*, 842–845.

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O., and Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat.Biotechnol. *31*, 1119–1125.

Carattoli, A. (2009). Resistance plasmid families in Enterobacteriaceae. Antimicrob.Agents Chemother. *53*, 2227–2238.

Chen, Z., Pham, L., Wu, T.-C., Mo, G., Xia, Y., Chang, P.L., Porter, D., Phan, T., Che, H., Tran, H., et al. (2020). Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate highly accurate and economical long-range sequencing information. Genome Res. *30*, 898–909.

Chien, A., Edgar, D.B., and Trela, J.M. (1976). Deoxyribonucleic acid polymerase from the extreme thermophile Thermusaquaticus. J.Bacteriol. *127*, 1550–1557.

Chin, C.S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Methods *13*, 1050–1054.

Danko, D.C., Meleshko, D., Bezdan, D., Mason, C., and Hajirasouliha, I. (2019). Minerva: an alignment- and reference-free approach to deconvolve Linked-Reads for metagenomics. Genome Res. *29*, 116–124.

De Maio, N., Shaw, L.P., Hubbard, A., George, S., Sanderson, N.D., Swann, J., Wick, R., AbuOun, M., Stubberfield, E., Hoosdally, S.J., et al. (2019). Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. Microb.Genom. *5*, e000294.

DeLong, E.F. (2000). Extreme genomes. Genome Biol. *1*, reviews1029.1021.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. *17*, 333–351.

Kang, J.-E., Kim, H.-D., Park, S.-Y., Pan, J.-G., Kim, J.H., and Yum, D.-Y. (2018). Dietary supplementation with a Bacillus superoxide dismutase protects against γ-Radiation-induced oxidative stress and ameliorates dextran sulphate sodium-induced ulcerative colitis in mice. J.Crohns Colitis *12*, 860–869.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., et al. (2007). Paired-end mapping reveals extensive structural variation in the human genome. Science *318*, 420–426.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. *27*, 722–736.

Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J., Charpentier, E., Haft, D.H., et al. (2015). An updated evolutionary classification of CRISPR-Cas systems. Nat. Rev.Microbiol. *13*, 722–736.

Mao, D., Grogan, D.W., and de Boer, P.A.J. (2017). How a genetically stable extremophile evolves: modes of genome diversification in the archaeonsulfolobusacidocaldarius. J.Bacteriol. *199*, e00177–17.

Merino, N., Aronson, H.S., Bojanova, D.P., Feyhl-Buska, J., Wong, M.L., Zhang, S., and Giovannelli, D. (2019). Living at the extremes: extremophiles and the limits of life in a planetary context. Front.Microbiol. *10*, 780.

Metzker, M.L. (2010). Sequencing technologies - the next generation. Nat. Rev. Genet. *11*, 31–46.

Mokashe, N., Chaudhari, B., and Patil, U. (2018). Operative utility of salt-stable proteases of halophilic and halotolerant bacteria in the biotechnology sector. Int. J.Biol.Macromol *117*, 493–522.

Niedringhaus, T.P., Milanova, D., Kerby, M.B., Snyder, M.P., and Barron, A.E. (2011). Landscape of next-generation sequencing technologies. Anal. Chem. *83*, 4327–4341.

Orellana-Saez, M., Pacheco, N., Costa, J.I., Mendez, K.N., Miossec, M.J., Meneses, C., Castro-Nallar, E., Marcoleta, A.E., and Poblete-Castro, I. (2019). In-depth genomic and phenotypic characterization of the antarcticpsychrotolerantstrain Pseudomonas sp. MPC6reveals unique metabolic features, plasticity, and biotechnological potential. Front.Microbiol. *10*, 1154.

Palmieri, G., Arciello, S., Bimonte, M., Carola, A., Tito, A., Gogliettino, M., Cocca, E., Fusco, C., Balestrieri, M., Colucci, M.G., et al. (2019). The extraordinary resistance to UV radiations of a manganese superoxide dismutase of Deinococcusradiodurans offers promising potentialities in skin care applications. J.Biotechnol. *302*, 101–111.

Peters, B.A., Kermani, B.G., Sparks, A.B., Alferov, O., Hong, P., Alexeev, A., Jiang, Y., Dahl, F., Tang, Y.T., Haas, J., et al. (2012). Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature *487*, 190–195.

Rothschild, L.J., and Mancinelli, R.L. (2001). Life in extreme environments. Nature *409*, 1092–1101.

Rubin, E.M., Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., et al. (2007). The diploid genome sequence of an individual human. PLoS Biol. *5*, e254.

Schiraldi, C., and De Rosa, M. (2002). The production of biocatalysts and biomolecules from extremophiles. Trends Biotechnol. *20*, 515–521.

Swarup, A., Lu, J., DeWoody, K.C., and Antoniewicz, M.R. (2014). Metabolic network reconstruction, growth characterization and 13C-metabolic flux analysis of the extremophile ThermusthermophilusHB8. Metab. Eng. *24*, 173–180.

Tharek, M., Sim, K.-S., Khairuddin, D., Amir Hamzah, G., and Najimudin, N. (2017). Whole-genome sequence of endophyticplant growth-promoting Escherichia coli USML2. Genome Announc. *5*, e00305–17.

Tolstoganov, I., Bankevich, A., Chen, Z., and Pevzner, P.A. (2019). cloudSPAdes: assembly of synthetic long reads using de Bruijn graphs. Bioinformatics *35*, i61–i70.

Urbieta, M.S., Donati, E.R., Chan, K.-G., Shahar, S., Sin, L.L., and Goh, K.M. (2015). Thermophiles in the genomic era: biodiversity, science, and applications. Biotechnol. Adv. *33*, 633–647.

Voskoboynik, A., Neff, N.F., Sahoo, D., Newman, A.M., Pushkarev, D., Koh, W., Passarelli, B., Fan, H.C., Mantalas, G.L., Palmeri, K.J., et al. (2013). The genome sequence of the colonial chordate, Botryllusschlosseri. Elife *2*, e00569.

Wang, H., Gong, Y., Xie, W., Xiao, W., Wang, J., Zheng, Y., Hu, J., and Liu, Z. (2011). Identification and characterization of a novel thermostablegh-57 gene from metagenomicfosmid library of the Juan de Fuca Ridge hydrothemal vent. Appl.Biochem.Biotechnol. *164*, 1323–1338.

Wang, J., Salem, D.R., and Sani, R.K. (2019a). Extremophilicexopolysaccharides: a review and new perspectives on engineering strategies and applications. Carbohydr.Polym. *205*, 8–26.

Wang, O., Chin, R., Cheng, X., Wu, M.K.Y., Mao, Q., Tang, J., Sun, Y., Anderson, E., Lam, H.K., Chen, D., et al. (2019b). Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. Genome Res. *29*, 798–808.

Weisenfeld, N.I., Kumar, V., Shah, P., Church, D.M., and Jaffe, D.B. (2017). Direct determination of diploid genome sequences. Genome Res. *27*, 757–767.

Zhang, F., Christiansen, L., Thomas, J., Pokholok, D., Jackson, R., Morrell, N., Zhao, Y., Wiley, M., Welch, E., Jaeger, E., et al. (2017). Haplotype phasing of whole human genomes using bead-based barcode partitioning in a single tube. Nat.Biotechnol. *35*, 852–857.

Zheng, G.X.Y., Lau, B.T., Schnall-Levin, M., Jarosz, M., Bell, J.M., Hindson, C.M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D.A., Merrill, L., Terry, J.M., et al. (2016). Haplotypinggermline and cancer genomes with high-throughput linked-read sequencing. Nat.Biotechnol. *34*, 303–311.

Ziko, L., Adel, M., Malash, M.N., and Siam, R. (2019). Insights into red sea brine pool specialized metabolism gene clusters encoding potential metabolites for biotechnological applications and extremophile survival. Mar. Drugs *17*, 273.

iScience, Volume *24*

Supplemental information

Comparison of different sequencing strategies for assembling chromosome-level genomes of extremophiles with variable GC content

Zhidong Zhang, Guilin Liu, Yao Chen, Weizhen Xue, Qianyue Ji, Qiwu Xu, He Zhang, Guangyi Fan, He Huang, Ling Jiang, and Jianwei Chen

Comparison of different sequencing strategies for assembling chromosome-level genomes of extremophiles with variable GC content

Zhidong Zhang[1,2,10], Guilin Liu[3,10], Yao Chen[1], Weizhen Xue[3], Qianyue Ji[3], Qiwu Xu[3], He Zhang[3], Guangyi Fan[3,7], He Huang[4,5], Ling Jiang[6,*], Jianwei Chen[3,7,8,9,11,*]

[1]College of Biotechnology and Pharmaceutical Engineering, Nanjing Tech University, Nanjing 211816, China

[2]Institute of Applied Microbiology, Xinjiang Academy of Agricultural Sciences/Xinjiang Key Laboratory of Special Environmental Microbiology, Urumqi, Xinjiang 830091, China

[3]BGI-Qingdao, BGI-Shenzhen, Qingdao, Shandong 266555, China

[4]School of Food Science and Pharmaceutical Engineering, Nanjing Normal University, Nanjing 210023, China

[5]School of Pharmaceutical Sciences, Nanjing Tech University, Nanjing 211816, China

[6]College of Food Science and Light Industry, Nanjing Tech University, Nanjing 211816, China

[7]BGI-Shenzhen, Shenzhen, Guangdong 518083, China

[8]Qingdao-Europe Advanced Institute for Life Sciences, BGI-Shenzhen, Qingdao 266555, China

[9]Laboratory of Genomics and Molecular Biomedicine, Department of Biology, University of Copenhagen, Universitetsparken 13, Copenhagen 2100, Denmark

[10]These authors contributed equally

[11]Lead Contact

[*]Correspondence: jiangling@njtech.edu.cn (L.J.), chenjianwei@genomics.cn (J.C.)

Transparent Methods

Sample Collection and DNA Extraction

Competent cells of *E. coli* K-12 DH5α (Cat.No. CB101) were purchased from Tiangen Biotech (Beijing). The other strains *B. cereus* 43-1A, *B. frigoritolerans* 44A, *R.* sp. LB8, *D. wulumuqiensis* R12 and *J. melonis* M714 were isolated in Xinjiang Uygur Autonomous Region of China and properly stored in our laboratory at Nanjing Tech University.

For next generation sequencing (NGS), genomic DNA was extracted from 2×10^6 cells using the CTAB (cetyltrimethylammonium bromide) method. Extracts were treated with DNase-free RNase to eliminate RNA contamination. Then the DNA quantity was determined using a Qubit 3.0 fluorometer, and DNA integrity was evaluated by gel electrophoresis. Finally, the DNA was sheared into fragments ranging in size from 50 to 800 bp using an E220 ultrasonicator (Covaris, Brighton, UK) as described before (Zhang et al., 2019).

For stLFR sequencing or Oxford Nanopore Technologies (ONT) platform, genomic DNA was extracted from 3×10^8 cells using the Blood & Cell Culture DNA Midi Kit (13343; Qiagen, Germany), according to the manufacturer's instructions. Genomic DNA was quantified using the dsDNA BR assay on a Qubit fluorometer (Thermo Fisher Scientific, USA) and measured by pulsed-field gel electrophoresis (PFGE). DNA with a main band of more than 40 kb and OD_{260}/OD_{280} between 1.6-2.2 was considered to have sufficiently high quality for sequencing.

NGS Library Construction and Sequencing

DNA fragments between 200 and 500 bp were selected using AMPure XP beads (Agencourt, Beverly, MA, USA) and then repaired using T4 DNA polymerase (ENZYMATICS, Beverly, MA, USA). These DNA fragments were ligated at both ends to T-tailed adapters and amplified for eight cycles, after which the amplification products were used to generate a single-strand circular DNA library. The NGS

libraries of LB8, 43-1A and K-12 were sequenced on a BGISEQ-500 platform (BGI, Qingdao, China) to obtain 100 bp paired-end raw reads, and the libraries of 44A, R12 and M714 were sequenced on an Illumina X-Ten platform (Majorbio-Shanghai China) to obtain 150 bp paired-end raw reads.

stLFR Library Construction and Sequencing

The stLFR technology uses Tn5 transposase for the co-barcoding of DNA libraries. The stLFR library was constructed following the standard protocol using the MGIEasy stLFR Library Prep kit v1.1 (PN: 1000005622) (Wang et al., 2019) with some process improvement for better assembly of bacterial genomes. In detail, instead of pooling different libraries for the final sequence, 1/10 of the magnetic beads were resuspended and mixed for subsequent digestion and library construction. To optimize the procedure for bacteria with a high GC content, we construct five stLFR libraries for *D. wulumuqiensis* R12 using different concentrations of the interrupting enzyme ranging from 0.4 to 1.2 pmol/10 ng DNA, and we reduced the input concentration of the interrupting enzyme from 1.0 to 0.40 pmole/ 10 ng DNA for *J. melonis* M714.

After ligation reaction Ⅱ, as described in the protocol, the beads were collected on the side of the tube and washed with 180 μL Wash buffer Ⅱ. The beads were then gently mixed with Wash buffer Ⅱ using a pipette and 18 μL was transferred to a fresh tube. Beads from different samples containing different sample indices were mixed together to a final 180 μL mixture. The final stLFR library was constructed following the protocol and sequenced on a BGISEQ-500 platform in 100 bp pair-end model (BGI-Qingdao, China).

Oxford Nanopore Library Construction and Sequencing

Library preparations and sequencing were carried out using the Oxford Nanopore Ligation Sequencing Kit SQK-LSK109 according to the manufacturer's instructions. Briefly, 1 μg of genomic DNA was fragmented using a Covaris g-TUBE via centrifugation at 4,000 g in an Eppendorf 5424 tube. DNA repair, end-repair and

A-tailing were combined using NEBNext FFPE DNA Repair Mix (M6630, New England BioLabs), and the NEBNext Ultra II End Repair/dA-tailing Module (E7546, New England BioLabs). DNA was subsequently purified using AMPureXP beads (A63882, Beckman Coulter, Ireland). Adapters were then ligated to the DNA using NEBNext Quick T4 DNA Ligase (E6056, New England BioLabs). Library loading onto R9.4.1 flow-cells was performed as stated in the manufacturer's protocol, followed by sequencing on a GridION instrument for 48 h. The ONT reads which quality higher than 7.0 were included.

NGS Assembly

After filtering low-quality reads, adapter contamination, and duplicate reads using SOAPnuke (v1.5.2) (Chen et al., 2018) with the parameters "-q 0.2 -l 0.2 -n 0.05 -d", the clean reads were assembled into contigs and scaffolds using SPAdes (v3.11.1) (Bankevich et al., 2012) with k-mer range of 43 to 83 and a step size of 10.

Nanopore Assembly

Nanopore reads longer than 8 kb were selected to assemble the complete genome. First, we used Unicycler (v0.4.8) (Phillippy et al., 2017) with default parameters to assemble the genomes using Nanopore reads and NGS sequencing reads as the reference genomes. At the sae time, the Nanopore reads were independently assembled into complete genomes using Canu (v1.6.0) (Koren et al., 2017) with the parameters "MhapSensitivity=high, corMinCoverage=4, and minReadLength=2000". To fix the INDEL and SNP errors, all the assembled genomes were successively polished twice with Illumina X-Ten or BGISEQ-500 NGS clean data using Pilon (v 1.23) and GATK (v 3.4-0) with the parameters; "-fix indels -nostrays" for Pilon and "-stand_call_conf 50 -stand_emit_conf 10.0 --filterExpression 'MQ0 >= 4 && ((MQ0 / (1.0 *DP)) > 0.1) && DP < 4'" for GATK.

stLFR Assembly

We constructed a bacterial genome assembly analysis pipeline, which could assemble

chromosome-level genomes from stLFR data alone or assemble complete genomes by adding ONT sequencing data. Firstly, raw sequencing data was filtered using SOAPnuke (v1.5.2) (Chen et al., 2018). The stLFR clean data were split into paired-end 100 bp short reads and their corresponding barcode information. The stLFR corresponding barcodes were transformed to generate barcodes compatible with the 10X Genomics format. The corresponding scripts were made available on GitHub (https://github.com/BGI-Qingdao/stlfr2supernova_pipeline).

We then assembled the draft genomes using various methods. SPAdes (v3.11.1) (Bankevich et al., 2012) and Architect (v0.1) (Kuleshov et al., 2016) were used to assemble the stLFR clean data with the parameters; "-k 43,53,63,73,83" and "--pe-abs-thr 3 --pe-rel-thr 0.15 --pe-rc-rel-thr 0.1 --rc-abs-thr 3 --rc-rel-edge-thr 0.15 --rc-rel-prun-thr 0.1", respectively. cloudSPAdes (v3.12.0-dev) (Tolstoganov et al., 2019), Athena (v1.3) (Bishara et al., 2018) and Supernova (v 2.1.1) (Weisenfeld et al., 2017) were also used to assemble the stLFR clean data with the default parameters. Finally, to make full use of the diversity of the stLFR barcode information, SLR-superscaffolder (v 0.9.0) (Deng et al., 2019) was further applied to improve the scaffolds and obtain chromosome-level genomes for all strains.

The chromosome-level genomes with the longest scaffolds and largest contig N50 values were selected. Then, we enhanced the chromosome-level genome to obtian a complete genome using TGS-GapCloser (v 1.1.1) (Xu et al., 2020) based on the Nanopore sequencing long reads with default parameters. Pilon (v 1.23) and GATK (v 3.4-0) were used to fix the sequencing errors based on the stLFR sequencing reads as described before.

Genome Evaluation and Annotation

To estimate the genome size for each sample according to the NGS data, the 15 bp k-mer frequency distribution was plotted using R (v3.4.1). Additionally, to assess the quality of each genome assembly, NGS reads were mapped to the genome using

SOAPaligner (v2.22) and BWA (v0.7.12-r1039). The ratio of mapped reads base coverage depth and number of mapped paired-end reads were calculated. The proportion of bases with >5× coverage among the total bases was calculated to assess the single base accuracy, and the proportion of mapped bases in more than 6 paired-end reads with normal insert size among total bases (normal insert size mapped and abnormal insert size mapped reads) was calculated to assess the structural accuracy. Subsequently, CheckM (v1.0.13) (Parks et al., 2015) software was used to evaluate the genome completeness and contamination.

RNAmmer (v 1.2) (Lagesen et al., 2007) was used to predicted the rRNAs. All the coding sequences (CDS) were predicted using Prokka (v1.13) (Seemann, 2014). Functional gene annotation was carried out using Diamond (v0.8.23.85), and aligned against the KEGG (v84) and COG (20141110) public database. We aligned the chromosome-level genomes to each other using Blast (v2.2.31) using the criteria E-value < 10^{-5} and match length >10,000 bp, and the consistency and genome features were visualized using Circos (v0.69-6). Finally, QUAST (v5.0.2) (Gurevich et al., 2013) was used to compare the NGS, stLFR and Nanopore assemblies for each sample, while the Unicycler assembly genomes were used as references. the assembly errors including SNVs, InDels and genome fraction were calculated.

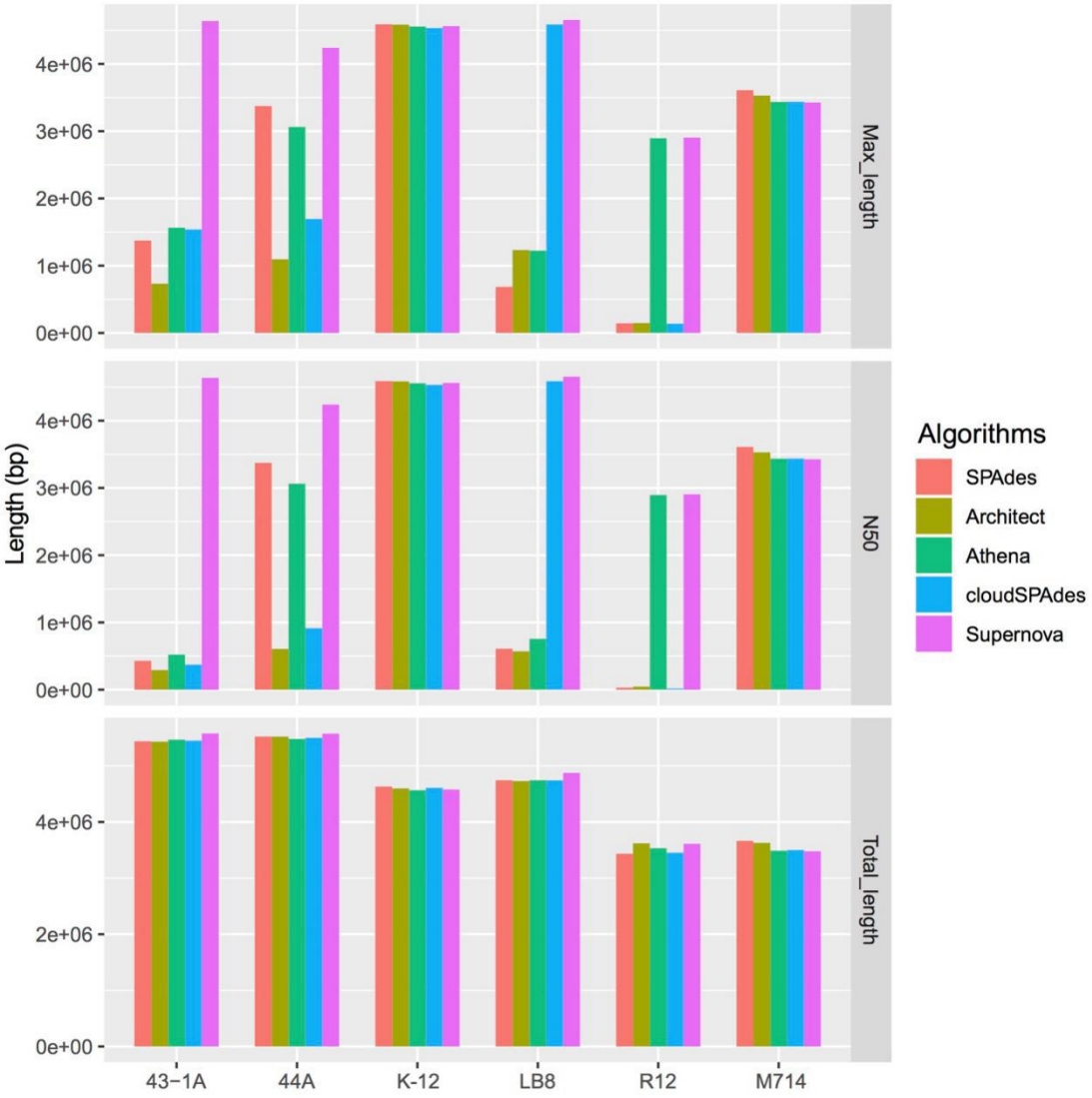


Figure S1. The stLFR sequencing data assembly statistics by using five different algorithms. Related to Figure 2.

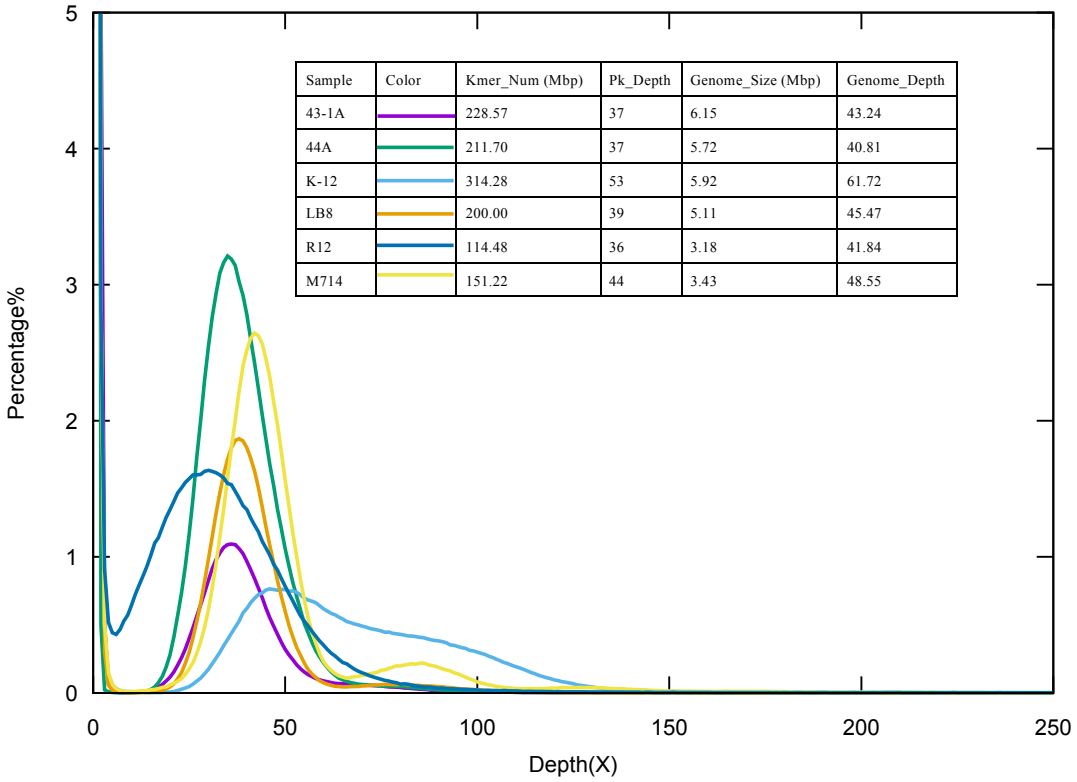


Figure S2. K-mers frequency distribution of six strains using NGS data. The peaks and estimated genome sizes are showed on the top of figure. Related to Figure 2 and Table 1.

Figure S3. The COG categories annotation in the six strains using different sequencing strategies for assembly the genomes. For each strain, we divided the annotation gene number of each assembly genome by the average number of all strategies to calculate the relative value for every COG class. Related to Figure 4.

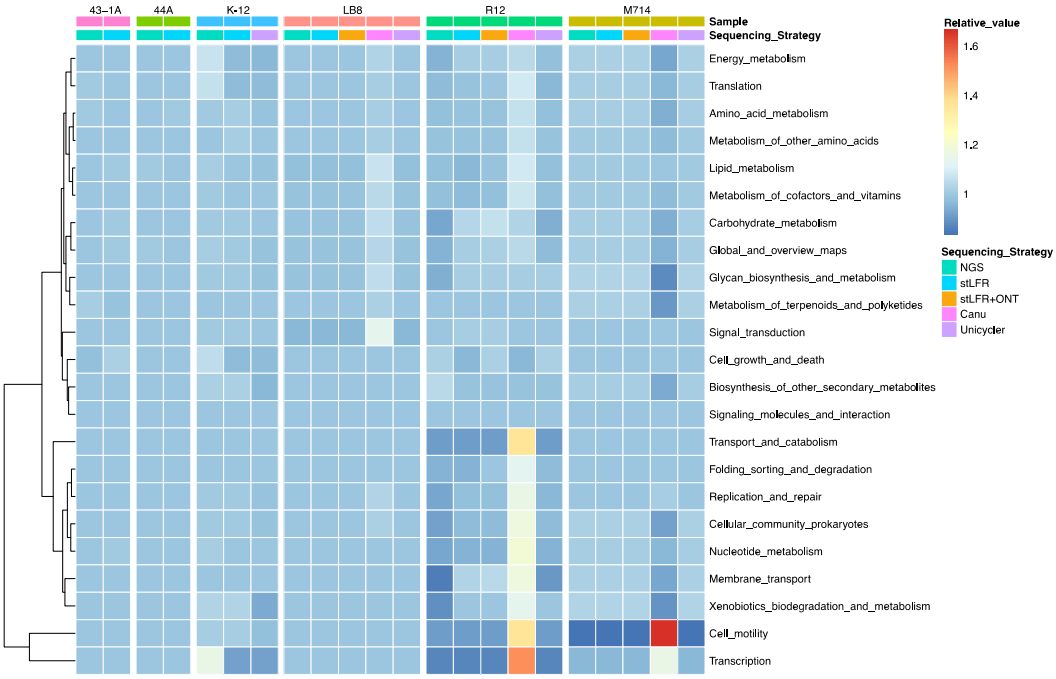


Figure S4. The KEGG pathway classification in the 6 strains of different sequencing strategies assembled genomes. For each strain, we divided the annotation gene number of each assembly genome by the average number of all strategies to calculate the relative value for every pathway. Related to Figure 4.

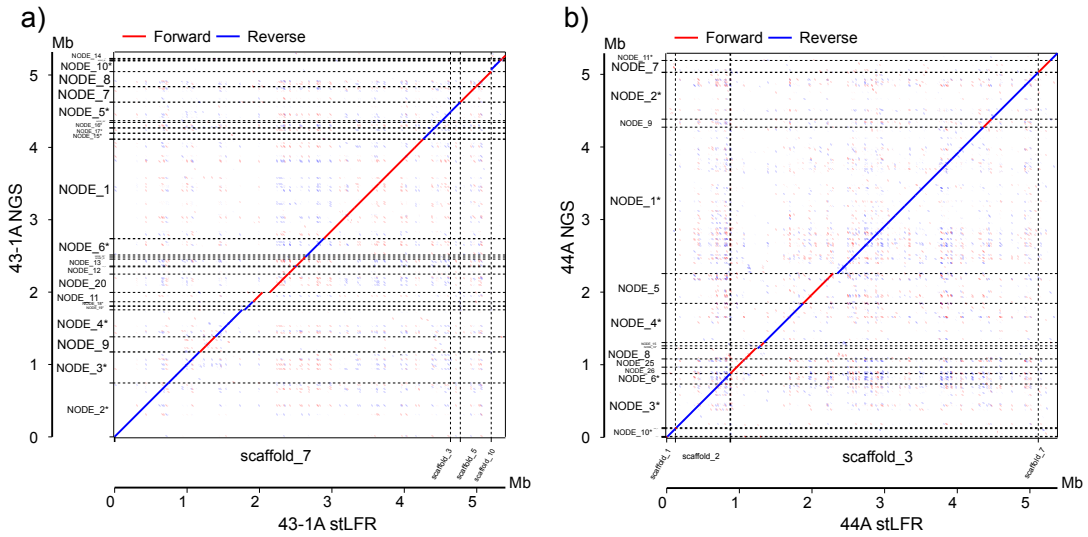


Figure S5. Genome comparisons of the stLFR assembled genome and the NGS assembled genome for strains a) *B. cereus* 43-1A and b) *B. frigoritolerans* 44A. Related to Figure 3.

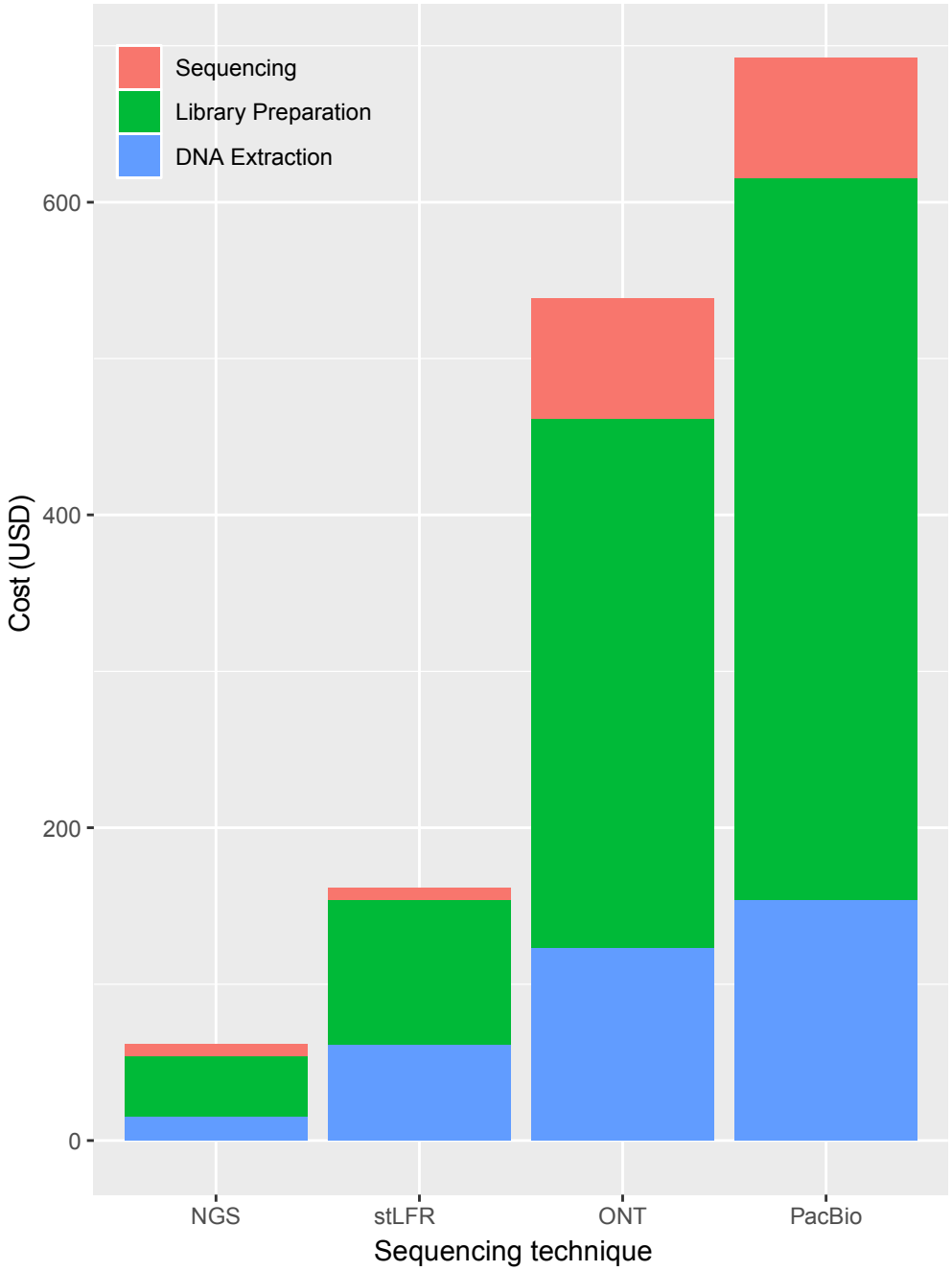


Figure S6. The sequencing costs of NGS, stLFR, ONT and PacBio for bacteria genomes. For each sequencing strategy, the cost includes DNA extraction, library preparation and sequencing of 1 Gb cleandata. Related to Table 1.

Table S1. The sequences statistics of the five concentrations stLFR sequencing of *D. wulumuqiensis* R12. Related to Figure 1.

Concentrations	Raw Data (Mb)	Low Qual (%)	Adapter (%)	Dup (%)	Clean Data (Mb)	Clean Q20	Molecule length (Kb)	DNA (ng)
0.4	2,000	7.06	8.14	9.91	1,498	97.88	60.37	10
0.6	2,000	6.60	10.20	9.24	1,479	96.96	57.19	10
0.8	2,000	6.29	7.97	10.25	1,510	97.23	56.06	10
1.0	2,000	6.41	10.54	9.19	1,477	95.96	64.44	10
1.2	2,000	6.05	9.35	9.86	1,495	97.00	58.92	10

Table S2. The NGS sequences statistics of the six strains. Related to Figure 2 and Table 1.

Sample	Raw Data (Mb)	Low Qual (%)	Adapter (%)	Dup (%)	Clean Data (Mb)	Clean Q20	DNA amount (ng)
43-1A	6,537	34.47	3.84	26.89	6,329	98.89	500
44A	8,787	60.94	0.54	38.40	7,380	98.49	500
LB8	9,830	9.43	1.82	11.99	9,300	94.43	500
R12	1,084	97.69	1.52	0.70	556	95.37	500
M714	904	13.72	0.31	85.96	667	98.63	500
K-12	7,008	63.29	4.93	20.36	6,674	98.190	500

Table S3. The stLFR sequences statistics of the six strains. Related to Figure 2 and Table 1.

Sample	Raw Data (Mb)	Low Qual (%)	Adapter (%)	Dup (%)	Clean Data (Mb)	Clean Q20	Molecule length (Kb)	DNA amount (ng)
43-1A	5,225.42	10.52	21.45	3.55	3,368.98	96.37	33.09	10
44A	4,716.93	11.33	12.94	3.39	3,411.92	95.53	47.34	10
LB8	6,053.28	10.57	16.23	4.97	4,130.41	94.85	60.71	10
R12	4,450.48	6.67	30.00	4.27	2,628.44	97.25	54.23	10
M714	6,000.00	1.88	6.70	52.46	2,337.39	95.87	51.26	10
K-12	2,828.84	10.22	13.78	5.30	2,000.00	93.24	42.86	10

Table S4. The Nanopore sequences statistics of strains *R.* sp. LB8, *D. wulumuqiensis* R12 and *J. melonis* M714. Related to Figure 2 and Table 1.

Sample	Raw Reads	Raw Data (Gb)	N50 (bp)	N90 (bp)	Quality score	DNA amount (ng)
LB8	325,092	2.22	20,067	3,032	10.06	1,500
R12	55,821	1.32	36,493	14,553	12.03	1,500
M714	162,546	2.91	29,596	9,482	9.42	1,500

Supplemental References

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D*., et al.* (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology *19*, 455-477.

Bishara, A., Moss, E.L., Kolmogorov, M., Parada, A.E., Weng, Z., Sidow, A., Dekas, A.E., Batzoglou, S., and Bhatt, A.S. (2018). High-quality genome sequences of uncultured microbes by assembly of read clouds. Nature Biotechnology *36*, 1067-1075.

Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., Li, Y., Ye, J., Yu, C., Li, Z*., et al.* (2018). SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. GigaScience *7*.

Deng, L., Guo, L., Xu, M., Wang, W., GU, S., Zhao, X., Chen, F., Wang, O., Xu, X., Fan, G*., et al.* (2019). SLR-superscaffolder: a de novo scaffolding tool for synthetic long reads using a top-to-bottom scheme. bioRxiv.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics *29*, 1072-1075.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res *27*, 722-736.

Kuleshov, V., Snyder, M.P., and Batzoglou, S. (2016). Genome assembly from synthetic long read clouds. Bioinformatics *32*, i216-i224.

Lagesen, K., Hallin, P., Rodland, E.A., Staerfeldt, H.H., Rognes, T., and Ussery, D.W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res *35*, 3100-3108.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res *25*, 1043-1055.

Phillippy, A.M., Wick, R.R., Judd, L.M., Gorrie, C.L., and Holt, K.E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLOS Computational Biology *13*, e1005595.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068-2069.

Tolstoganov, I., Bankevich, A., Chen, Z., and Pevzner, P.A. (2019). cloudSPAdes: assembly of synthetic long reads using de Bruijn graphs. Bioinformatics *35*, i61-i70.

Wang, O., Chin, R., Cheng, X., Wu, M.K.Y., Mao, Q., Tang, J., Sun, Y., Anderson, E., Lam, H.K., Chen, D*., et al.* (2019). Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. Genome Res *29*, 798-808.

Weisenfeld, N.I., Kumar, V., Shah, P., Church, D.M., and Jaffe, D.B. (2017). Direct determination of diploid genome sequences. Genome Research *27*, 757-767.

Xu, M., Guo, L., Gu, S., Wang, O., Zhang, R., Peters, B.A., Fan, G., Liu, X., Xu, X., Deng, L*., et al.* (2020). TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. Gigascience *9*.

Zhang, D., Yang, Y., Qin, Q., Xu, J., Wang, B., Chen, J., Liu, B., Zhang, W., and Qiao, L. (2019). MALDI-TOF Characterization of Protein Expression Mutation During Morphological Changes of

Bacteria Under the Impact of Antibiotics. Anal Chem.