Briefings in Bioinformatics, 2025, **26(5)**, bbaf504
https://doi.org/10.1093/bib/bbaf504
Problem Solving Protocol

TELLBASE: a novel tool of TELL-seq barcode-assisted scaffold assembler for bacterial genomes

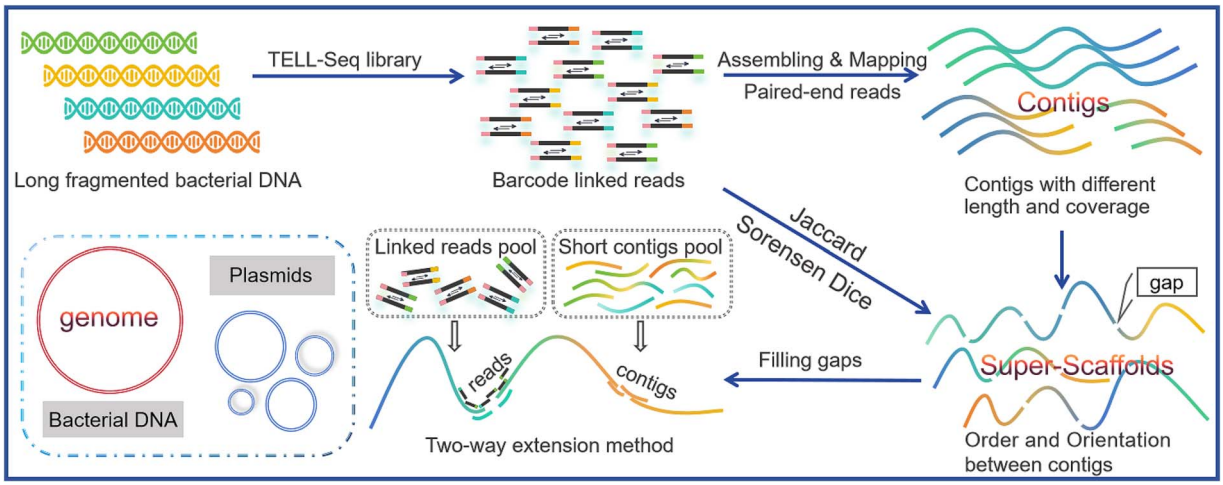
Yutong Li[1,‡], Tianlong Kuang[1,‡], Tao Xu[1,2,3,‡], Hanxiao Du[1], Yi Zhang[2,3], Yu Qian[1], Yiwen Chen[4], Zhenxian Xiao[4], Chen Chen[2,3], Jing Wu[2,3], Wen-Hong Zhang[1,2,3,*], Chenqi Lu[1,*], Ning Jiang[1,2,3,*]

[1]State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, No. 2005 Songhu Road, Yangpu district, Shanghai 200433, China
[2]Shanghai Sci-Tech Inno Center for Infection & Immunity, No. 1688 Guoquan Bei Road, Yangpu district, Shanghai 200433, China
[3]Department of Infectious Diseases, Huashan Hospital, Shanghai Medical College, Fudan University, No. 12 Wulumuqi Zhong Road, Jingan district, Shanghai 200040, China
[4]Department of Bioinformatics, Wuxi Universal Sequencing Co., Ltd., No. 35 South Changjiang Road, Xinwu district, Wuxi 214013, China
*Corresponding authors. Ning Jiang, State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, No. 2005 Songhu Road, Yangpu district, Shanghai 200433, China. E-mail: ningjiang@fudan.edu.cn; Chenqi Lu, State Key Laboratory of Genetic Engineering. School of Life Sciences, Fudan University, No. 2005 Songhu Road, Yangpu district, Shanghai, 200433, China. E-mail: luchenqi@fudan.edu.cn; Wen-Hong Zhang, State Key Laboratory of Genetic Engineering. School of Life Sciences, Fudan University, No. 2005 Songhu Road, Yangpu district, Shanghai, 200433, China. E-mail: zhangwenhong@fudan.edu.cn
‡Yutong Li, Tianlong Kuang, Tao Xu contributed equally.

Abstract

Transposase enzyme linked long-read sequencing (TELL-seq) technology generates barcode-linked reads, facilitating whole-genome sequencing (WGS), and complete assembly with improved accuracy and reduced costs. Unlike mate-pair sequencing technology, TELL-seq employs a near-full-sequence tagging strategy that allows more efficient capture of comprehensive genomic information. However, assembly algorithms and software capable of fully leveraging the characteristics of TELL-seq technology to effectively assemble genomic sequences at the megabase-scale are lacking, particularly for bacteria and their plasmids. In this study, we present TELL-seq barcode-assisted scaffold assembler (TELLBASE), a *de novo* genome assembler designed specifically for assembling bacterial genomes using TELL-seq-derived linked reads. In assembly tests involving bacteria such as *Acinetobacter baumannii, Klebsiella pneumoniae, Mycobacterium tuberculosis,* and *Staphylococcus aureus,* TELLBASE exhibited exceptional efficacy in producing chromosome-level bacterial genomic sequences and successful identification of plasmids present in the sequenced strains. Comparative analysis revealed that TELLBASE significantly outperforms existing assemblers tailored for TELL-seq-derived linked reads, such as TuringAssembler and Ariadne, in terms of the completeness and accuracy of the assembled genomes. Therefore, TELLBASE shows promising potential for refining draft bacterial genomes and further applications in related fields. The package for TELLBASE is freely available on GitHub (https://github.com/sosie1/TELLBASE).

Graphical Abstract



Keywords: complete *de novo* assembly; bacterial genome assembly; assemble algorithm for TELL-seq platform; barcoded NGS for bacterial genome

Received: March 4, 2025. **Revised:** August 3, 2025. **Accepted:** August 22, 2025

Introduction

In recent decades, there has been an unprecedented increase in the scale of bacterial sequencing datasets, stemming from the release of the first bacterial genome sequence, from *Haemophilus influenzae* [1]. Whole-genome sequencing (WGS) has proven clinically useful because it enables prediction of antibiotic resistance and virulence, as well as guides transmission and outbreak identification efforts [2–5]. Furthermore, WGS facilitates the integration of comparative and functional genomics, allowing the translation of fundamental microbiological research to applications in contemporary clinical practices [6]. Fully assembled bacterial genomes, derived via comprehensive and detailed processing of raw sequencing data, provide essential insights into the genetic blueprint that directs bacterial life processes, including metabolism, replication, and interaction with both the environment and hosts. By comparing the genomes of various bacterial species or strains, researchers can gain valuable insights into the mechanism by which bacteria adapt to their environments and hosts [7–8].

Progress in biological research hinges not only on the development and application of new sequencing technologies but also on innovative assembly algorithms that provide high-throughput and high-accuracy genome sequences, facilitating the application of genetic engineering techniques for bacterial manipulation in biotechnological contexts [9]. Nonetheless, the gap between well-assembled genomic data and the vast amounts of raw sequence data that are uploaded—often as original reads or poorly annotated scaffolds—continues to pose challenges. One major hurdle in the processing of raw genomic data is that specific operating systems, software, and programming skills are required to interpret the data, which can be a significant barrier for researchers lacking expertise in bioinformatics. Moreover, the fragmented nature of these datasets limits the ability of users to intuitively grasp the overall genomic structure and quickly locate specific DNA regions or genes. Consequently, there is an urgent need for an efficient, accurate and low-cost solution for assembling bacterial genome sequencing data.

In recent years, multiple methodologies have been developed to extract information from long fragments utilizing short reads from next-generation sequencing. These methods can be experimental or algorithmic and include mate-pair sequencing [10], clonal barcoding methods such as Synthetic Long Read (SLR) [11] and linked-read sequencing [12], Hi-C sequencing [13], and transposase enzyme linked long-read sequencing (TELL-seq) [14], among various other techniques.

The advent of these enhanced next-generation sequencing technologies has prompted the design of assemblers tailored to specific SLR library types. FragScaff is an assembler specifically developed for processing contiguity-preserving transposition sequencing (CPT-seq) reads; it uses the minimum spanning tree (MST) algorithm on scaffold graphs to complete scaffolding on the basis of co-barcoding information [15]. TruSPAdes can be used to assemble TSLR data through a barcode-based assembly approach. Although TSLRs have the potential to increase the contiguity of metagenomic assemblies, the maximum assembly length is capped at 10 kb [11]. Kuleshov et al. developed a *de novo* scaffolder called Architect, which uses Illumina's SLR sequencing technology to create scaffold graphs through the integration of co-barcoding and paired-end read information [16–17]. ARCS reconstructs contigs obtained by ABySS [18] to organize draft genomes into more contiguous assemblies, leveraging the barcoded information found in 10× Genomics Chromium data [19]. Compared with ARCS, ARKS employs a k-mer-based mapping approach that significantly accelerates the assembly process for experimental human genome datasets while accurately estimating gap sizes between contigs using barcoding information from 10×-linked reads [20]. However, ARKS sometimes requires assistance from other methods, such as Hi-C. Tolstoganov *et al.* introduced cloudSPAdes [21], a universal assembler that uses a de Bruijn graph-based approach for scaffold construction; however, it lacks independent modules for scaffolding, reducing its convenience for assembling other sequencing strains. Recently, a standalone scaffolder for stLFRs, known as the SLR-superscaffolder [22], was unveiled; it generates scaffolds with greater contiguity and higher accuracy than other tools for next generation sequencing-based (NGS-based) draft assemblies and requires only a draft assembly of contigs along with an SLR dataset as input. Another novel assembler, Ariadne [23], supports multiple assembly strategies suited for various types of datasets, particularly those from environmental microbial strains with minimal prior characterization, utilizing a new SLR deconvolution algorithm based on assembly graphs to address the challenge posed by 3′ unique molecular identifier deconvolution in metagenomics and fully exploit linkage information from SLR technology. The SpLitteR [24] tool effectively utilizes linked reads to increase the contiguity of HiFi assemblies; unlike other methods, SpLitteR use HiFi assembly graphs and supports diploid assemblies, addressing some repeat-related issues. In summary, complete genome assembly via NGS platforms remains a daunting challenge. The integration of new sequencing technologies and new algorithms for assembly is a crucial requirement.

We assessed the current sequencing technologies and selected TELL-seq as the experimental foundation to optimize the following algorithmic procedures. The rapid development of TELL-seq as an SLR sequencing technique warrants attention [25]. While Universal Sequencing Technology (UST)'s TELL-seq approach employs Illumina's NGS platform to produce economical and highly accurate linked reads, it yields ultra-long reads with lengths over 100 kb on short-read NGS platforms, utilizing barcodes for large-scale WGS [14]. A unique feature of TELL-seq, wherein reads from the same DNA fragment carry identical barcodes, renders it particularly advantageous for *de novo* assembly of microbial genomes [26]. Thus, the ability of TELL-seq to generate long reads with high accuracy and its scalability, and compatibility with various bacterial strains positions it as a superior choice for obtaining complete bacterial genomes. TuringAssembler [14], launched by UST, is a specialized assembly algorithm for TELL-seq-generated data, primarily utilizing barcode information to facilitate bacterial strain genome assembly. Although TuringAssembler and Ariadne can effectively process and assemble TELL-seq data, their performance in scaffolding particularly complex bacterial genomes can be inadequate. For genomes with numerous repetitive or highly heterozygous regions, their assembly results may fall short.

To address these challenges, we developed a genome assembler named TELL-seq barcode-assisted scaffold assembler (TELLBASE), which was designed for the complete and precise assembly of bacterial chromosome and plasmid genomic sequences derived from the TELL-seq platform. TELLBASE integrates existing NGS assemblers, alignment software, and barcode information associated with each read to perform efficient scaffolding for sequencing data derived from the TELL-seq platform. Notably, TELLBASE leverages barcode information to improve the determination of contig assembly orientation and order, showing significant

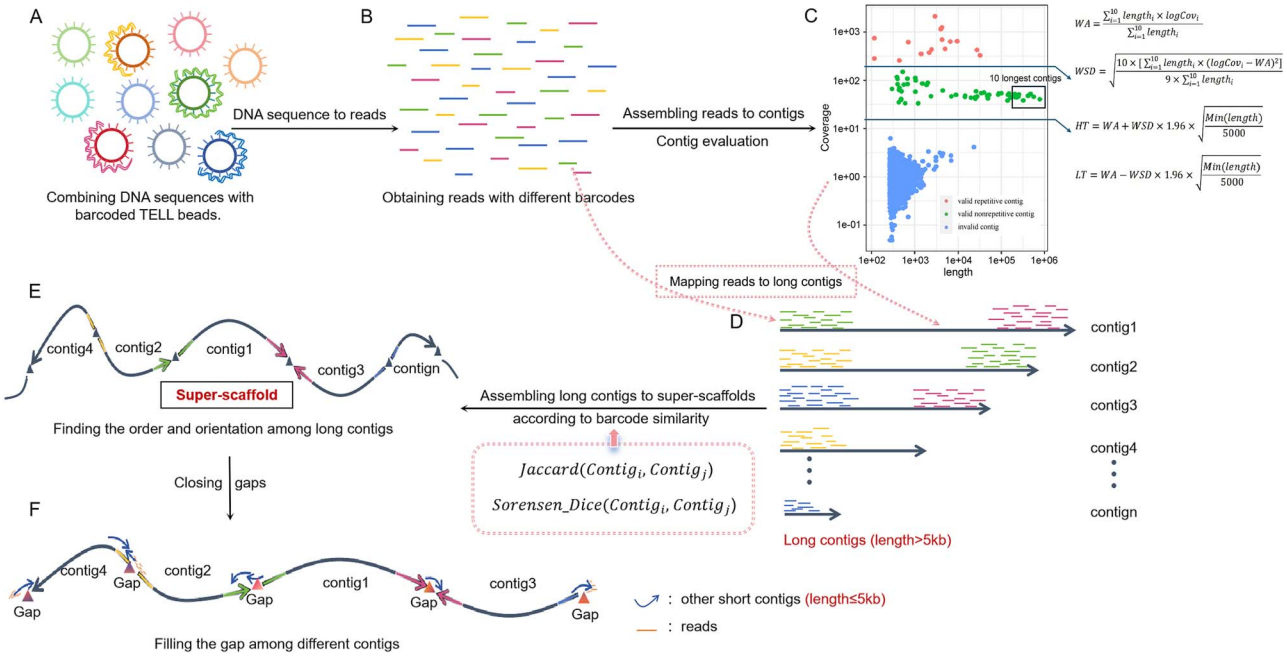


Figure 1. Overview of TELLBASE workflow. (A) Combining DNA sequences with barcoded TELL beads. (B) Paired-end reads with different barcodes. Different colors represent different barcode sequences. (C) Contigs are assembled by SPAdes and classified into three categories. (D) Mapping paired-end reads on heads and tails of long contigs. (E) Identifying the positional relationship among long contigs by using Jaccard correlation and Sorensen-Dice correlation. (F) Utilizing short contigs and pair-end reads to fill the gaps in the super-scaffolds.

advantages in the assembly of strains characterized by a high frequency of repetitive sequences.

Materials and methods

The workflow of TELLBASE algorithm

TELLBASE is a genome assembly algorithm designed for bacterial genome sequencing data obtained via the TELL-seq platform. It has the capability to skillfully and completely utilize barcode information from each DNA fragment for precise completion of bacterial genome assembly. TELLBASE can also accurately distinguish the plasmid genome from the nuclear genome. The workflow for this novel algorithm is shown in Fig. 1, and the details of the algorithm are shown in Fig. 2.

DNA sequence to reads

The strains used in the experiments were platform-stored strains from the Shanghai Sci-Tech Inno Center for Infection & Immunity. The culture conditions of strains, the experimental details of DNA isolation and purification are referred to Chen, Z. *et al.* (2020).

Libraries were prepared via the TELL-seq™ WGS Library Prep Kit in accordance with the TELL-seq™ Microbial WGS Library Prep User Guide. The libraries were submitted to Mingma (Shanghai) Biotechnology Co., Ltd., for sequencing. Library QC was performed via a Qubit fluorometer and an Agilent Bioanalyzer for quantification and fragment analysis, followed by sequencing on the NovaS4 sequencing platform via a custom primer kit, namely, the TELL-seq™ Illumina® Sequencing Primer Kit.

To ensure that the subsequent input reads for the algorithm are as clean as possible, preprocessing of the input data is needed. Data preprocessing consists of several steps: removing duplicate reads, filtering out low-quality reads using fastp (version: 0.23.2) [27] (Phred quality scores <30), discarding low-quality barcode-containing reads (not matching the pattern NNHNNYRNNNNYRNNHNN (IUPAC notation)), removing adapter sequences utilizing Cutadapt (version: 1.18) [28] (minlength = 5), performing fastqc quality inspection employing FastQC (https://github.com/s-andrews/FastQC, Version: 0.12.1), and marking barcode information in paired-end reads.

Assembling reads to contigs (first-round assembly)

TELLBASE uses a two-step assembly algorithm strategy. SPAdes3.15.5 [29] (with the parameter k set to 21, 25, 33, 55, 71, 77, and 115) was finally chosen to assemble paired-end reads into contigs. Then the contigs and their relative barcode information are used as primary inputs for subsequent assembly.

Contig evaluation

TELLBASE classifies all contigs into three categories: valid repetitive contigs (contigs with high coverage and multiple occurrences in the bacterial genome), valid nonrepetitive contigs (contigs with medium coverage and occurring only once in the bacterial genome) and invalid contigs (contigs with low coverage and not present in the bacterial genome) according to the two thresholds: LT and HT (Fig. 1C). TELLBASE utilizes a 95% confidence interval of logarithmic coverage for valid non-repetitive contigs as thresholds. The contigs with coverage below low_threshold are defined as having invalid sequence information and are discarded. Those contigs with coverage higher than high_threshold are considered valid repetitive sequences and are reused in the subsequent assembly process. Additionally, plasmid contigs are present within valid repetitive contigs and valid nonrepetitive contigs. Therefore, TELLBASE employs PlasFlow [30] to filter out the plasmid contigs from all the valid contigs.

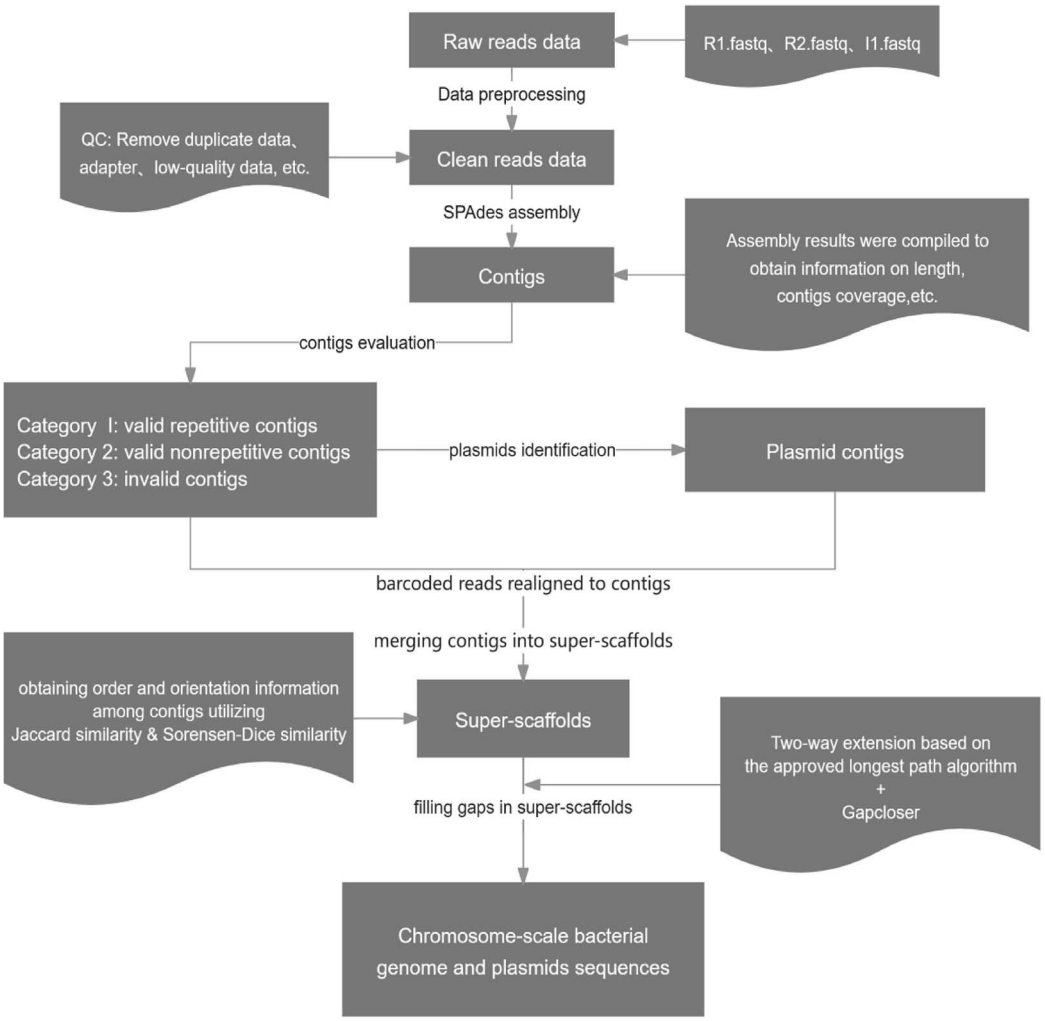


Figure 2. Details of TELLBASE workflow and principle. The input file of TELLBASE is the fastq format data obtained from TELL-seq platform, and the output gets organized plasmids and chromosome-level bacterial genome sequence information respectively.

Assembling long contigs to super-scaffolds (second round assembly)

TELL_seq provides a large amount of barcodes that can be used to link contigs into super-scaffolds. TELLBASE applies BWA-MEM (Version: 0.7.17-r1188) [31] to align all filtered paired-end reads with long valid contigs longer than 5000 bp with the parameters "-q 20 -bS -f 2 -F 256". The position of paired-end reads in contigs was subsequently obtained from alignment bam files via SAMtools (Version: 1.11) [32]. Through the information of shared barcodes, this algorithm uses $\mathbf{Sorensen\_Dice}\ (\mathbf{Contig_i}, \mathbf{Contig_j})$ to obtain the positional relationships among long contigs and then utilizes $\mathbf{Jaccard}\ (\mathbf{Contig_i}, \mathbf{Contig_j})$ to determine the order and orientation information for contigs with adjacent physical positions. Here, $\mathbf{Category}\ (\mathbf{Contig_i})$ represents the total category of valid barcodes that $\mathbf{Contig_i}$ has, and $\mathbf{Counter}\ (\mathbf{Contig_i})$ represents the total number of valid barcodes that $\mathbf{Contig_i}$ has

$$\mathrm{Jaccard}\ (\mathrm{Contig_i}, \mathrm{Contig_j}) = \frac{\mathrm{Category}\ (\mathrm{Contig_i}) \cap \mathrm{Category}\ (\mathrm{Contig_j})}{\mathrm{Category}\ (\mathrm{Contig_i}) \cup \mathrm{Category}\ (\mathrm{Contig_j})}$$

$$\mathrm{Sorensen\_Dice}\ (\mathrm{Contig_i}, \mathrm{Contig_j}) = \frac{\mathrm{Counter}\ (\mathrm{Contig_i}) \cap \mathrm{Counter}\ (\mathrm{Contig_j})}{\mathrm{Counter}\ (\mathrm{Contig_i}) + \mathrm{Counter}\ (\mathrm{Contig_j})}$$

$$(\mathrm{i}, \mathrm{j} \in [1, \mathrm{n}]\,, \mathrm{i}, \mathrm{j} \in \mathrm{Z}, \mathrm{n} = \mathrm{number\_of\_contig})$$

Closing gap

According to barcode similarity matrixes, long contigs continuously extended toward largest similarity and formed larger path based on the alignment results via BLAST (version 2.11.0+) [33]. TELLBASE uses Score (P) to select the optimal extension path when more than two perfect paths are generated. The optimal path between two contigs tends to have the fewest short contigs and the largest length (Fig. 3).

$$Score\ (P) = \frac{L\ (P)}{L_{max}} - 0.1 * \frac{N\ (P)}{N_{max}}$$

- *L (P)*: Total length of path *P*.
- *N (P)*: Number of contigs involved in path *P*.

- L_{max}: Length of the longest path in all perfect complement paths.
- N_{max}: Maximum number of contigs in all perfect hole paths.

If gaps still exist after contig extension, length, alignment consistency score, and the number of short contigs are considered as significant metrics for suboptimal path determination with the optimization of the depth-first search algorithm, albeit with varying priorities. P_{final} is set as the principle.When gaps still exists in contigs, GapCloser [34] is subsequently selected to further fill the gaps using paired-end reads.

$$P_{final} = \underset{P_i}{\text{argmax}} \left(L\ (P_i), -N\ (P_i), avg\_consistency\ (P_i) \right)$$

- P_{final}: Best suboptimal path.
- $L\ (P_i)$: Length of the total extension path P_i.
- $N\ (P_i)$: Number of contigs involved in path P_i.
- $avg\_consistency\ (P_i)$: Average alignment consistency of path P_i.

Results

De novo bacterial genome assembly with ultralow amounts of input DNA material

Barcode-linked reads for each strain were generated using TELL-seq library preparation and sequencing technology. The Sorensen-Dice coefficient matrices for contigs at varying DNA concentrations were illustrated through heatmaps (Fig. 4), which intuitively verifies the rationality of using barcodes to reconstruct complete bacterial genomes, as contigs with more shared barcodes are physically closer in the genome, fitting the principles of TELL-seq library construction. This approach clarifies the rationale for using barcode sequences from TELL-seq data for contig positional restoration on the reference genome.

The DNA loading concentration does affect the assembly results of linked-reads sequencing platforms. In 2020, Chen et al. compared the assembly outcomes of *Escherichia coli* K12 MG1655 strain under different DNA loading concentrations from 100 pg to 1000 ng, concluding that the TELL-Seq assembly results are optimal at a DNA loading concentration of 100 pg for bacterial genomes assembly [14]. On this basis, we systematically tested the assembly results of three bacterial strain samples using a DNA input concentration gradient (10 pg, 20 pg, 50 pg, 100 pg, and 200 pg).

For the five DNA input concentration gradients of the three test strains, *de novo* genome assembly was performed by using TELLBASE, TuringAssembly, and Ariadne, respectively (Supplementary Tables S1–S3). The assembly results are all assessed by QUAST [35] (version: 5.2.0, 3d87c606). Evaluating the assembly results based on the number of scaffolds and genome fraction metrics, the three strains exhibited better assembly outcomes at 20 pg, 50 pg, and 100 pg. The performance at 200 pg was underperformed for all the strains. The assembly results for *A. baumannii ATCC19606* (*AB ATCC19606*) (NCBI Taxonomy ID: 470) with a 20 pg input were superior to those obtained with a 100 pg input, as evidenced by the total aligned length, the size of the largest contig, and the NA50 (Table 1). As for *K. pneumoniae HS11286* (*KP HS11286*) (NCBI Taxonomy ID: 1125630), the assembly results for the 100 pg input exceeded those for the 20 pg input in terms of total aligned length, NA50, and total number of contigs (Table 1).

This could be due to the fact that during TELL-Seq library construction and sequencing, bacterial DNA fragments are combined with millions of TELL beads, each carrying unique barcode information. Therefore, the higher DNA input concentration, the more DNA fragments from different genomic regions will enter the same droplet and share the same barcode. This seriously undermines the basis of secondary assembly in TELLBASE—the similarity of barcodes. Although the assembly results for the strain *KP 2044* at 10 pg DNA input concentration were the best, the assembly quality for the other two strains at this concentration declined compared to the results at 20–100 pg. The reason is that using 10 pg requires the incorporation of the lambda genome DNA during library preparation and sequencing to meet the minimum input concentration threshold, which imposes higher demands on sequencing quality and laboratory equipment. Even minor contamination can significantly degrade the assembly performance.

Moreover, as demonstrated in Fig. 4, for the same bacterial strain, lower concentrations yielded clearer diagonal patterns on the heatmap and reduced background noise. Notably, for the same strain, the Jaccard similarity index is higher at low concentrations (Fig. 5).

Therefore, given that bacterial genome sizes predominantly fall within the range of 1–10 Mbp, the DNA input concentration gradient employed for TELL-Seq sequencing should generally not exceed 100 pg. For different bacterial genomes, when using TELL-Seq for library construction and sequencing, 100 pg should first be used as the DNA input concentration. Taking the number of assembled scaffolds as the judgment index, if more than one scaffold is obtained from the assembly, a lower concentration (50 pg, 20 pg) should be selected for sequencing to find the optimal loading concentration. Notably, the optimal input concentration should not be <10 pg.

Comparison of the completeness of genomes assembled by the three assemblers

The following 15 strains were categorized into two groups: five strains with complete genomic reference sequences obtained via Nanopore long-read sequencing (*AB ATCC19606*, *A. baumannii Z4* (*AB Z4*), *Clostridioides difficile 1326* (*C.Diff 1326*), *KP HS11286* and *KP 2044* and 10 strains with highly homologous genomes, the sequences of which were used as reference sequences and downloaded from the NCBI dataset [*Enterococcus faecalis 23HH337VP* (*E.Faecalis 23HH337VP*), *H. influenzae PittEE 23HH197Hi* (*Hi.PittEE 23HH197Hi*), *K. pneumoniae 23HH453KP* (*KP 23HH453KP*), *Proteus mirabilis ZS21–78* (*PM ZS21–78*), *S. aureus CoL* (*Staph.Aureus CoL*), *S. aureus FPR* (*Staph.Aureus FPR*), *S. aureus USA500* (*Staph.Aureus USA500*), *S. epidermidis ZJ23HZ072* (*Staph.Epidermidis ZJ23HZ072*), *S. aureus 23HH236SA* (*Staph.Aureus 23HH236SA*) and *Tubercle bacillus 4–221 716* (*TB 4–221 716*)].

TELLBASE algorithm can be installed on either computer workstation or cluster server, and independently execute parallel multitasking for assembly operations. Moreover, it imposes relatively low configuration requirements on the operating platform. The platform and its configuration used for these tests is as follows: a Linux server (equipped with an Intel Xeon Silver 4210 CPU @ 2.20 GHz), employing 32 CPU cores and 32 GB of available memory. When using 32 cores, our algorithm can complete genome assembly for the vast majority of monoclonal cultured strains within 3–5 h. Furthermore, TELLBASE algorithm enables the adjustment of the number of CPU threads to enhance computing efficiency and shorten the runtime. In Table S4, five strain samples with genome sizes spanning from 1 m to 5 m were selected to compare the running times of three software tools. The execution durations of TELLBASE and TuringAssembler are relatively close, both being slightly longer than that of Ariadne.

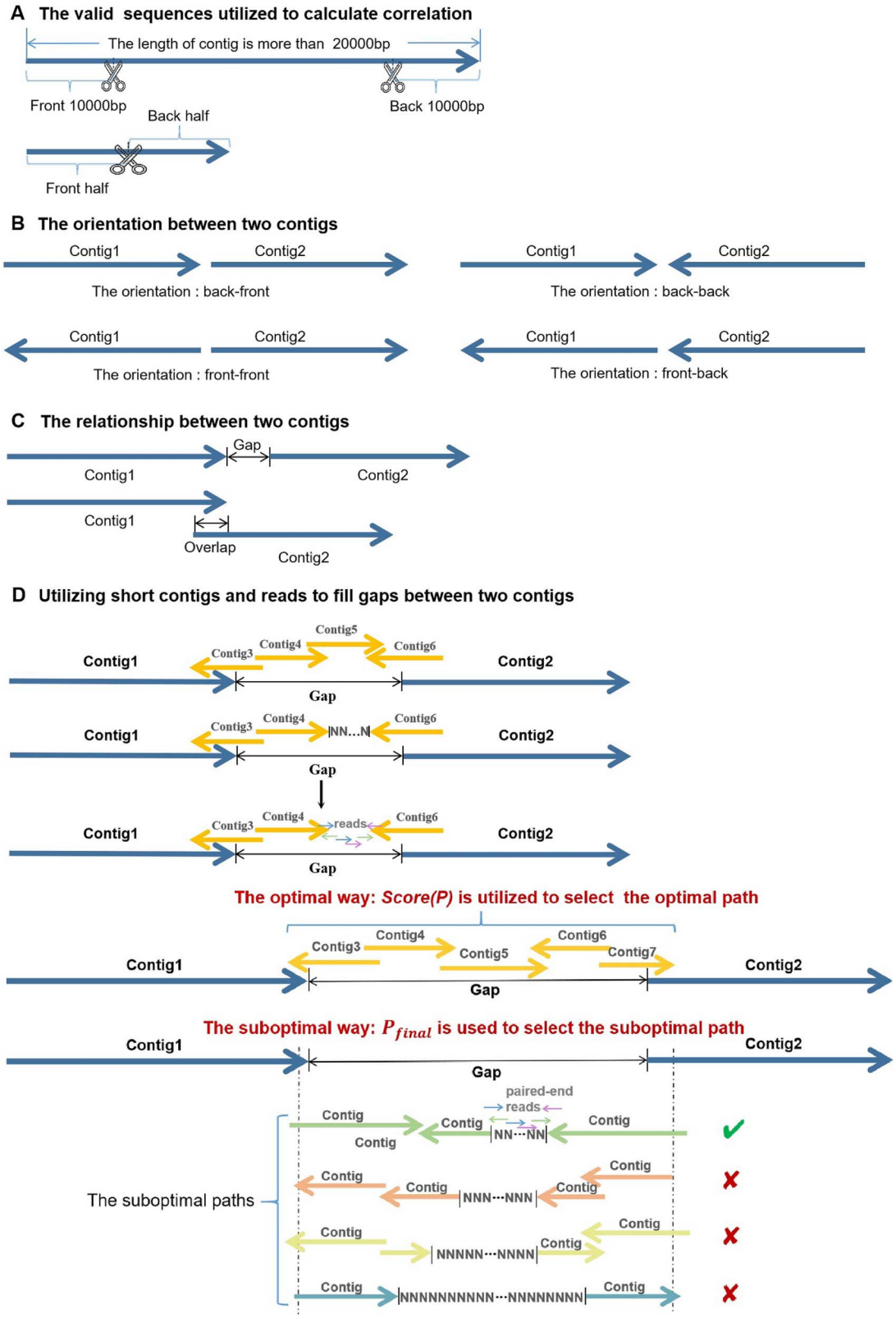


Figure 3. Schematic diagram of contigs extension and gaps filling. (A) When calculating the correlation of contigs, to minimize the influence of length, sequences longer than 20 kb are selected from the barcode information corresponding to each 10 kb site at each end. Contigs shorter than 20 kb are split into two parts. (B) There are only four directional relationships between contigs: back-front, back-back, front-front, and front-back. (C) The relationships between contigs may include gaps in the middle or overlaps at the end or start sequences. (D) Various paths can be used to fill gaps between contigs. Score (P) is utilized to select the optimal path that perfectly fills the gaps. Additionally, P_{final} is formulated to select suboptimal path. Paired-end reads are used to fill gaps if there are still "N" in the suboptimal path.

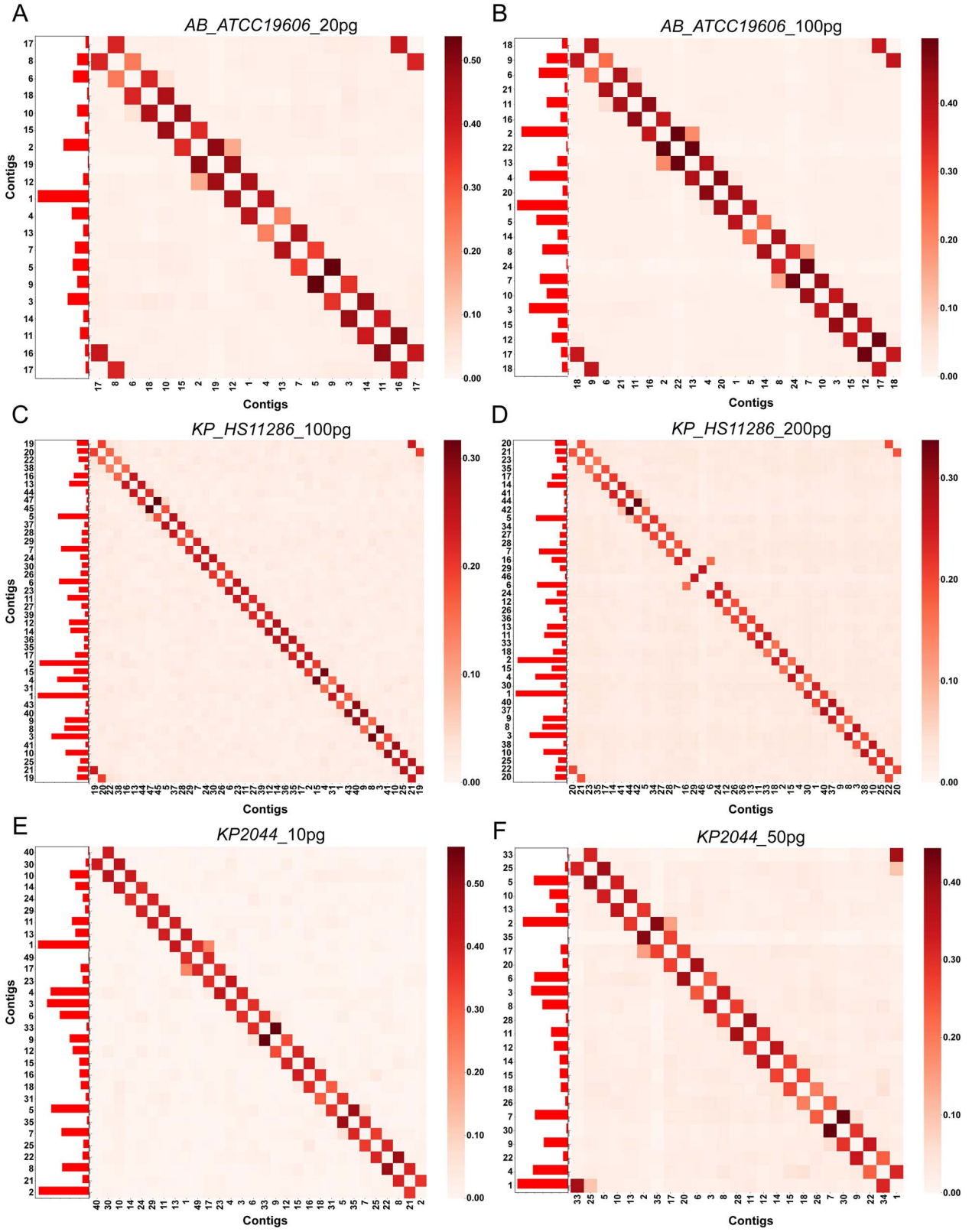


Figure 4. Heatmaps illustrating the correlation between contigs at various input DNA concentrations. Heatmap for the *AB ATCC19606* strain at 20 pg (A), the *AB ATCC19606* strain at 100 pg (B), the *KP HS11286* strain at 100 pg (C), the *KP HS11286* strain at 200 pg (D), the *KP 2044* strain at 10 pg (E), and the *KP 2044* strain at 50 pg (F). All the correlations of contigs are calculated according to formula *Sorensen_Dice* ($Contig_i$, $Contig_j$). The leftmost column displays the length information of the corresponding position contigs.

Table 1. Summary of *de novo* assembly results using TELLBASE on different bacterial strains with different gDNA input.

Strain	*AB ATCC19606*	*AB ATCC19606*	*KP HS11286*	*KP HS11286*	*KP 2044*	*KP 2044*
gDNA input (pg)	100	20	200	100	50	10
Genome size (bp)	3 998 039	3 998 039	5 682 322	5 682 322	5 477 886	5 477 886
Total aligned length (bp)	3 980 400	3 975 957	5 668 047	5 719 973	5 476 578	5 486 880
Genome fraction (%)	99.34	99.44	99.64	99.44	99.52	99.54
NA50 (bp)	1 245 502	1 257 189	1 098 718	3 724 555	1 669 485	636 776
Largest scaffold (bp)	2 381 023	3 948 516	5 232 020	5 355 164	5 262 762	5 280 516
Total length (> = 1000 bp)	3 980 530	3 976 087	5 672 437	5 723 500	5 644 816	5 589 538
N50 (bp)	2 381 023	3 948 516	5 232 020	5 355 164	5 262 762	5 280 516
# scaffolds (> = 5000 bp)	5	4	8	5	3	4
# scaffolds (> = 10 000 bp)	4	2	7	5	3	4
# scaffolds (> = 50 000 bp)	3	1	5	4	3	3
Ns per 100kbp	0.05	0.08	0.21	0.10	0.18	0.32
GC (%)	39.15	39.14	57.15	57.08	57.22	57.28

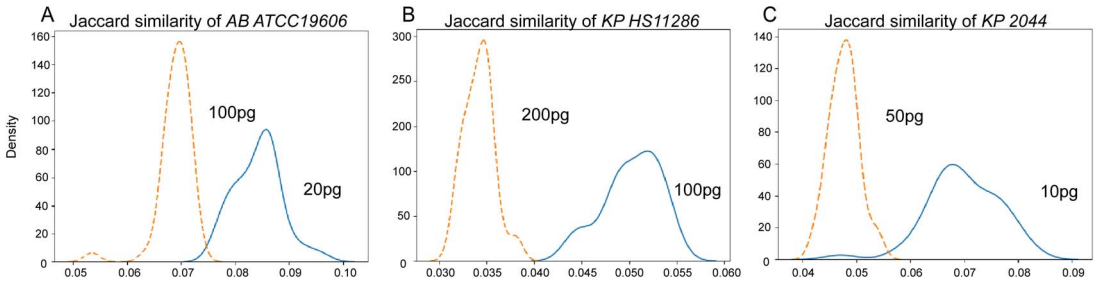


Figure 5. The distribution of Jaccard similarity for *AB ATCC19606* (A), *KP HS11286* (B), and *KP 2044* (C) at different DNA input concentration. For the *AB ATCC19606* strain, the orange curve (dashed curve) corresponds to distribution at 50 pg DNA input concentration, and the blue curve (solid line) corresponds to results at 10 pg DNA input concentration. For the *KP HS11286* strain, the orange curve (dashed curve) corresponds to result at 200 pg DNA input concentration, and the blue curve (solid line) corresponds to results at 100 pg DNA input concentration. For the KP 2044 strain, the orange curve (dashed curve) corresponds to distribution at 50 pg DNA input concentration, and the blue curve (solid line) corresponds to results at 10 pg DNA input concentration.

The heatmaps of all the strains demonstrate the reasonable of using barcodes to determine the physical order and orientation of each contigs (Fig. 4, Fig. S1 and S2). Three different assemblers—Ariadne, TuringAssembler (the official TELL-seq software), and TELLBASE—were employed to assemble genomic sequences for the five bacterial strains with complete reference genome (Table 2) and 10 bacterial strains with the genome of a highly homologous strain used as a reference (Table S5). For all the tested strains, Ariadne generated short contigs and numerous scaffolds, which were not ideal for subsequent analyses.

For bacterial strains with a complete reference genome (Table 2, Fig. 6A), TELLBASE provided the largest genomic fraction for most strains, especially the *K. pneumoniae 2044 (KP 2044)* (NCBI Taxonomy ID: 484021) strain, the value for which reached 99.860%. Compared with TELLBASE, TuringAssembler underperformed for strain *AB ATCC19606*, strain *C.Diff 1326*, and strain *KP HS11286*, but matched TELLBASE for strain *KP 2044* and strain *AB Z4*. TELLBASE exhibited superior performance in generating longer NA50 and N50 lengths across a majority of the strains, particularly for strain *KP 2044*, the N50 length for which reached an impressive 5 280 498 bp. For the *AB ATCC19606* strain and the *C.Diff 1326* strain, the NA50 lengths obtained through assembly with TELLBASE were ∼3 times and 3.65× greater than those produced by TuringAssembler, respectively. Moreover, TELLBASE consistently assembled the largest contigs and the fewest long contigs exceeding 50 kb for nearly all the evaluated strains (Table 1, Fig. 6A), with strain *C.Diff 1326* exemplifying this trend with the largest contig measuring 4 265 300 bp—1.5× longer than the largest contig assembled by TuringAssembler.

For bacterial strains with highly homologous genome sequences available as reference sequences in the NCBI dataset, the TELLBASE assembly results surpassed those of TuringAssembler, featuring significantly fewer scaffolds, higher N50 values, and longer largest contigs (Table S5, Fig. 6A), indicating its effectiveness in maintaining high assembly quality and completeness for bacterial genomes that may feature complex repetitive regions.

In conclusion, compared with TuringAssembler and Ariadne, the sequences generated by TELLBASE are more closely aligned with the reference genome, resulting in a more comprehensive and accurate representation of bacterial genomic sequence information.

Accuracy of bacterial genome assembly at the single-base level

After assessing the overall integrity of the assembly, we conducted a collinearity analysis of the genomic sequences using two software tools, Mauve [36] and BRIG [37], to further evaluate the accuracy and consistency of the assembly at the single-base level.

We performed a collinearity analysis at the single-base level for all 15 strains to evaluate the structural accuracy of the assembly outcomes (Figs 6B–E, Figs S3–S15). In particular, according to comparative genomic analysis of the *C.Diff 1326* strain and the *Staph. Aureus FPR* strain, as presented by Mauve (Fig. 6B and C), the genome sequence generated by the TELLBASE assembler corresponded perfectly to the reference genome sequence in the same order. In contrast, the genome fragments obtained by TuringAssembler presented significant inversions and rearrangements. The assembly results for the *C.Diff 1326*

Table 2. Comparison of *de novo* assembly results of five strains with third-generation sequence reference genome using sequencing data from TELL-seq Illumina library with different assemblers—TELLBASE, TuringAssembler, and Ariadne.

Strain	AB ATCC19606			AB Z4			C. Diff 1326			KP HS11286			KP 2044		
Assembler	TELLBASE	Turing-Assembler	Ariadne	TELLBASE	Turing-Assembler	Ariadne	TELLBASE	Turing-Assembler	Ariadne	TELLBASE	Turing-Assembler	Ariadne	TELLBASE	Turing-Assembler	Ariadne
gDNA input (pg)	20	20	20	50	50	50	100	100	100	100	100	100	10	10	10
Genome size without plasmids(bp)	3 980 844	3 980 844	3 980 844	3 977 193	3 977 193	3 977 193	4 313 749	4 313 749	4 313 749	5 333 942	5 333 942	5 333 942	5 023 660	5 023 660	5 023 660
Total aligned length (bp)	3 948 371	3 944 019	3 916 821	4 099 048	3 924 114	3 874 001	4 235 196	4 227 456	4 170 921	5 353 464	5 267 535	5 256 580	5 041 686	5 052 061	4 972 181
Genome fraction (%)	99.19	98.68	98.31	97.35	97.48	97.287	97.98	97.05	96.61	99.27	98.64	98.48	99.86	99.39	98.88
NA50 (bp)	1 257 189	419 447	103 304	268 145	103 486	43 327	801 314	219 878	130 136	3 724 555	219 858	97 811	636 776	409 469	126 799
Largest scaffold (bp)	3 948 500	3 900 868	408 022	4 067 300	3 766 218	308 422	4 265 300	2 802 262	441 439	5 355 100	5 236 518	360 592	5 280 498	5 435 352	252 546
Total length (> = 1000 bp) (bp)	3 948 500	3 991 158	3 925 155	4 067 300	4 140 481	4 102 072	4 265 300	4 218 877	4 148 647	5 355 100	5 914 162	5 623 050	5 280 498	5 627 113	5 626 976
N50 (bp)	3 948 500	3 900 868	103 304	4 067 300	3 766 218	48 231	4 265 300	2 802 262	130 136	5 355 100	5 236 518	99 382	5 280 498	5 435 352	133 196
# scaffolds (> = 5000 bp)	1	2	68	1	4	106	1	5	68	1	5	92	1	5	84
# scaffolds (> = 10 000 bp)	1	2	58	1	4	89	1	5	53	1	5	81	1	4	66
# scaffolds (> = 50 000 bp)	1	2	26	1	3	19	1	4	25	1	4	41	1	2	35
Ns per 100 kbp	0.05	2.53	0	2.16	4.76	0	0.28	1.44	0	0.07	8.09	0	0	2.11	0
GC (%)	39.18	39.06	39.06	39.08	39.08	39.5	29.38	29.21	29.08	57.41	56.83	57.13	57.64	57.2	57.17

strain were highly fragmented, featuring severe inversion within a long contig. Similarly, the *Staph.Aureus FPR* strain also displayed fragmentation, with seven significant inversions within a long contig.

In terms of the BRIG results (Fig. 6D and E) for the *C.Diff 1326* strain and the *Staph.Aureus FPR* strain, TELLBASE performed almost perfectly. Nevertheless, the first and fourth assembled fragments obtained by TuringAssembler were discontinuous in the reference genome. For all the strains, according to the analysis results obtained through Mauve and BRIG shown in Figs S3–S6 and Figs S7–S15, the assembly results of TELLBASE did not show severe structural errors. TuringAssembler did not perform well for the *Staph.Aureus CoL* strain and the *Staph.Aureus UAS500* strain. Moreover, in the high-GC-content regions of the bacterial genome, there were either sequence breaks or sequences that had been incorrectly assembled on the basis of the TuringAssembler algorithm.

From the above analysis, TELLBASE achieves the most complete assembly results and can successfully overcome these essential problems. To validate this, we further compared the assembly results of the strain *C. Diff 1326* and *Staph. Aureus FPR*, specifically analyzing high-GC regions in their reference genomes (Table S6). Among 24 high-GC regions across two samples, TELLBASE successfully assembled 22 regions, while TuringAssembler assembled only two regions. In the high-GC regions where TuringAssembler failed to complete assembly, 19 gaps were identified alongside three critical misassemblies (Fig. 6). Above result indicates that the assembly accuracy rate of TELLBASE in high-GC regions (91.67%) is significantly higher than that of TuringAssembler (8.33%) ($\chi^2 = 30.08$, $df = 1$, $P = 4.1e\text{-}08$). These unsuccessfully assembled high-GC regions contain five tandem repeats—structural features that present particular challenges for assembly. TELLBASE harnesses to overcome the disadvantages posed by traditional assembly methods, particularly in regions with high GC content and repetitive sequences, ultimately leading to a more complete and accurate assembly.

The above comparison results show that TELLBASE exhibits considerably improved accuracy, especially in regions of the bacterial genome with high GC content. TELLBASE effectively addresses gaps by fully leveraging shared barcode information, short contigs and paired-end reads, accounting for the impact of repeated sequences. Consequently, TELLBASE has overcome the disadvantages posed by traditional assembly methods, particularly in regions with high GC content and repetitive sequences, ultimately leading to a more complete and accurate assembly.

Availability and accuracy in the assembly of plasmid sequences

Plasmids play crucial roles in microbiology, genetics, and biotechnology, but distinguishing plasmids from genome contigs can be challenging owing to the strong similarity of some plasmid regions to the nuclear genome. TELLBASE utilizes PlasFlow as its core tool for extracting plasmid contigs from chromosomal contigs, enabling the identification of a vast majority of plasmid sequences. In contrast, tools such as TuringAssembler and Ariadne cannot recognize plasmids.

Among the strains tested in this study, except strain *C.Diff 1326*, the remaining four strains all possessed varying numbers of plasmids. The assembly results for the plasmids of the *AB ATCC19606* strain, the *AB Z4* strain, the *KP HS11286* strain, and the *KP 2044* strain were summarized in Table 3. TELLBASE could accurately

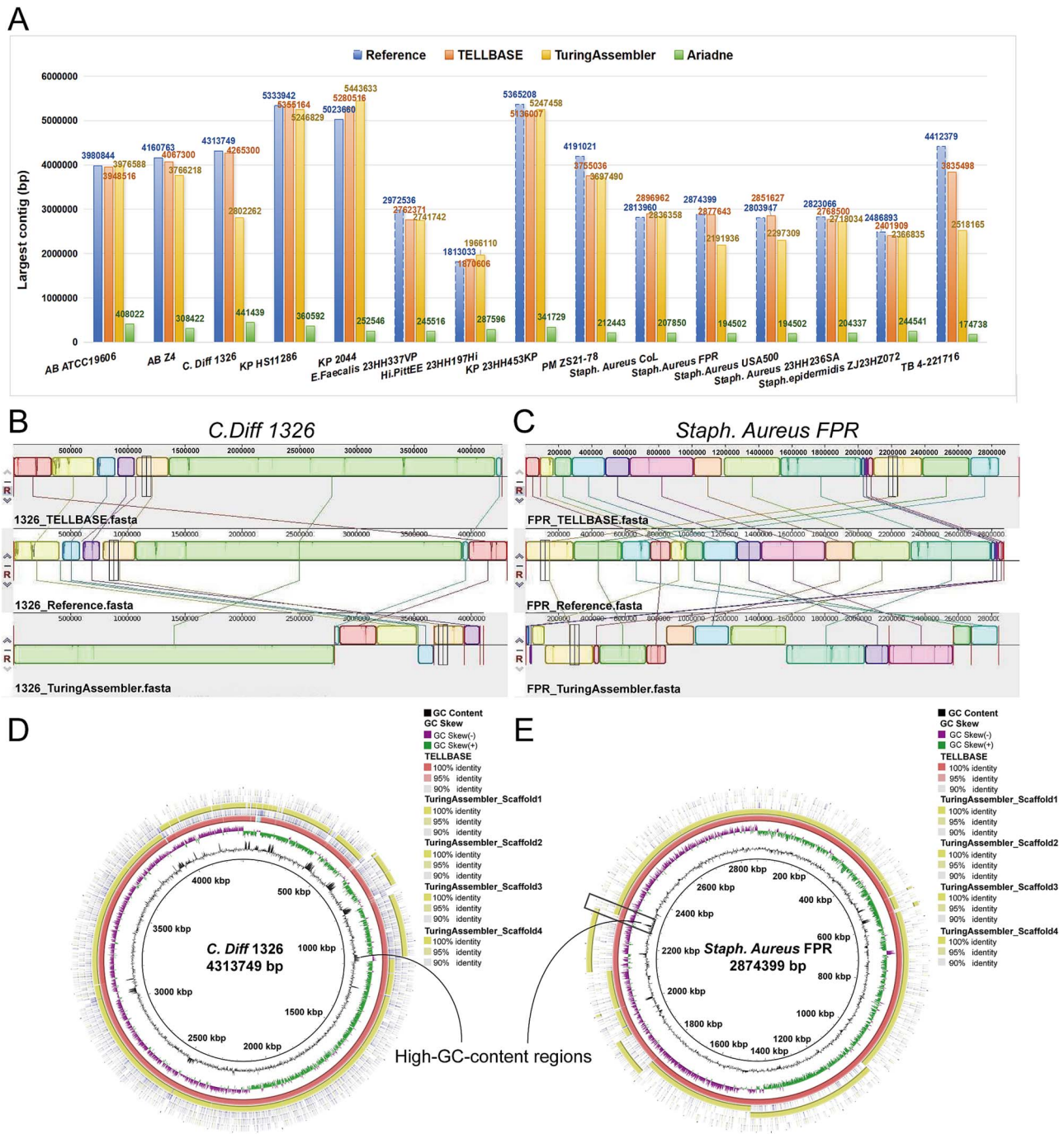


Figure 6. Evaluation of the assembly results between TELLBASE, TuringAssembler, and Ariadne. (A) Bar chart of the largest contigs of all drawn bacterial strains. For each strain, from left to right, the first bar represents the reference genome size, the second barrepresents the TELLBASE assembled genome size, the third bar represents the TuringAssembler assembled genome size, and the fourth bar represents the Ariadne assembled genome size. The dashed portion of the figure represents the size of the genome of a highly homologous strain found at NCBI, serving as a reference genome. The Mauve collinear plots for the strain *C.Diff 1326* (B) and the strain *Staph.Aureus FPR* (C). The BRIG ring image for the strain *C.Diff 1326* (D) and the strain *Staph.Aureus FPR* (E). In the BRIG Ring Image, from inside to outside, the first circle represents the reference genome, the second circle is GC content, the third circle is GC Skew, the fourth circle is the genome assembled by TELLBASE, and the rest are the broken scaffolds assembled by TuringAssembler.

identify almost all the plasmids. For the *AB ATCC19606* strain, complete plasmid sequence information was obtained, with a genome fraction of up to 95.342%. Fortunately, some plasmid sequences hidden in long contigs could also be identified. On the basis of the BLAST alignment results, we separated the plasmids from the long contigs, with the genomic fraction of the reference sequence reaching 100% (Fig. S16). Additionally, the assembly results for the identified plasmids were highly accurate, with no structural rearrangements or inversions observed in ~90% of the plasmids (Fig. S17). In conclusion, TELLBASE exhibits a significant advantage over TuringAssembler and Ariadne in that it effectively recognizes and assembles plasmids in a near-circular state.

Discussion

WGS has been widely used to predict drug resistance, identify virulence factors, determine sequence types, and detect transmission clusters [38] in bacteria. Numerous tools have

Table 3. Identified plasmids information of ***AB ATCC19606***, ***AB Z4***, ***KP HS11286***, ***771*** and ***KP 2044***.

Strains	***AB ATCC19606***	***AB Z4***		***KP HS11286***			***KP 2044***		
Number of plasmids	**1**	**2**		**3**			**3**		
plasmids	**Plasmid1**	**Plasmid1**	**Plasmid2**	**Plasmid1**	**Plasmid2**	**Plasmid3**	**Plasmid1**	**Plasmid2**	**Plasmid3**
[A]Reference length (bp)	17 195	112 293	71 277	122 799	111 195	105 974	223 649	146 332	73 002
If unique plasmids	√	×	√	√	√	√	√	×	×
Genome fraction (%)	95.342	100	85.482	97.746	96.100	90.107	95.554	100.000	100.000
Largest scaffold (bp)	16 394	112 293	60 748	120 033	118 127	97 119	216 213	146 332	76 477
# scaffolds (> = 5000 bp)	1	1	1	1	1	1	1	1	1
N50 (bp)	**16 394**	**112 293**	**60 748**	**120 033**	**118 127**	**97 119**	**216 209**	**146 332**	**76 477**

[A]complete plasmids obtained via nanopore long-reads sequencing

been developed to assemble bacterial genomes using WGS data. Nanopore sequencing also plays a pivotal role in WGS, especially for the sequencing of monoclonal culture strains, plasmid sequencing, complete genome assembly, and full-length viral sequencing. Although NGS offers high basic quality, the short read lengths obtained impede the complete assembly of plasmids, which hinders our understanding of plasmid evolution and acquisition pathways. Consequently, combining Nanopore long-read and Illumina short-read data is the currently suitable strategy for comprehensive bacterial genome assembly [38–39]. The original sequencing accuracy for Nanopore platform is ∼96%–98% (error rate 2%–4%) which is not of the same order of magnitude compared with the over 99.9% sequencing accuracy obtained from the current illumine platform. Recently, using high-precision model based on R10.4.1+ chip and Kit 14+ reagent, the original single accuracy can reach 98%–99% (error rate 1%–2%), which is closer to Illumina's Q30 (error rate 0.1%). However, the throughput may decrease by ∼30%, and the cost will be higher. This high error rate of Nanopore sequencing data, with the main error types being insertions/deletions (Indels), could lead to inaccurate assembly during the process of genomic assembly analysis, which commonly requires correction based on data from Illumina high-quality short-read sequencing.

TELLBASE offers significant advantages by integrating NGS with plasmid identification, providing a cost-effective method with a short turnaround time for clinical and diagnostic applications. Furthermore, by combining TELLBASE with the TELL-seq platform, with advantages in terms of precision and cost reduction, TELLBASE can be utilized for plasmid sequencing to track nosocomial transmission and *in vivo* microbial evolution [40].

On the basis of the principles of TELL-seq library construction and sequencing, we explored the optimal input concentration for microorganisms such as bacteria. DNA fragments were ligated to magnetic beads harbouring barcodes using transposase, after which the DNA was randomly fragmented, ensuring that each fragment was linked to the corresponding barcode. The assembly process was then used to reconstruct the complete genome from these smaller fragments by identifying the associations among the barcodes. Thus, TELL-seq library construction effectively transforms short-read sequencing into high-precision long-read sequencing, providing valuable information for subsequent assembly that surpasses the capabilities of standard NGS.

According to the mean coverage of strain contigs obtained by SPAdes, TELLBASE categorizes the contigs into three groups and focuses on the valid contigs. We fully utilized linked-read barcode information to determine the order and orientation of ultra-long contigs whose length was greater than the threshold (5000 bp by default) with high accuracy. Then, relying on the optimized longest path algorithm, we design a new metric, Scope (P), to choose the optimal path to fill the gaps between ultra-long contigs. It is possible that these gaps cannot be filled due to low mapping depth or complex genome structure. Thus, it is necessary to use GapCloser software to eliminate these gaps.

This study aimed to develop a novel algorithm for assembling bacterial genomes from TELL-seq Illumina libraries and to assess the performance of the innovative assembler TELLBASE in genome assembly. Our findings revealed that TELLBASE achieved chromosome-level assembly with an overall assembly completeness rate exceeding 99%, surpassing TuringAssembler and Ariadne in key metrics such as N50, the largest contig length, and the number of contigs longer than 10 kb. Furthermore, an analysis of the results revealed that TuringAssembler performed poorly on strains such as *AB Z4*, *C.Diff 1326*, and *Staph.Aureus FPR*, which was particularly evident in the errors of assembly in high-GC regions. High-repeat and high-GC regions have been among the greatest challenges encountered in genome assembly. This is because regions with high GC content may have high homology in the genome, which increases the risk of incorrectly splicing these regions together during assembly. Fortunately, under TELL-seq library sequencing, the fragments of repetitive regions carry different barcodes, and TELLBASE software leverages valid barcode information with the Jaccard coefficient and Sorensen-Dice coefficient to link contigs and correctly assemble them into super-scaffolds, even at the chromosome level. The improvements in contig length suggest that TELLBASE is particularly effective at resolving repetitive regions, which is crucial for accurately reconstructing complex genomes such as those of *K. pneumoniae* and *M. tuberculosis*.

In this study, all these strains were monoclonal cultured bacteria and had a good performance without severe degradation. In future studies in clinical samples, we will use clinical isolates from culture and screen the single colony to ensure the DNA qualities. TELLBASE employs a length-weighted logarithmic coverage classification method to categorize contigs from a round of assembly, where the classification thresholds vary according to the overall coverage profile of each sequencing sample, adapting dynamically to changes in the sample's coverage distribution. For critically important strain samples requiring complete genome resolution, enhancing the sequencing depth of short-read (Illumina) data can improve assembly completeness. For samples with ultra-low coverage or uneven barcode distribution, supplementing with low-coverage long-read (e.g. PacBio or Nanopore) sequencing data is recommended to obtain comprehensive genomic information.

Additionally, in the assembly process for TELLBASE, plasmids of the selected bacterial strains were identified using PlasFlow

software. Following the completion of plasmid identification by PlasFlow, the assembly of the bacterial nuclear genome and plasmids is performed in parallel. Even if certain plasmid contig sequences remain unidentified, it does not impede the assembly of nuclear genome sequences; however, the resulting plasmid sequences may be incomplete. Looking ahead, should more robust plasmid identification software with higher accuracy and completeness emerge, we plan to update this algorithm by integrating the new plasmid identification tool to achieve more complete and precise genome assembly results for monoclonal cultured strains.

The GapCloser gap-filling procedure is performed after the second-round assembly and bidirectional extension gap-filling, aiming to further utilize Pair-end reads to fill gaps in the genomic sequence. Its application does not affect the preceding assembly results, serving solely as a step to polish the assembly outcome. If subsequent genomic analysis requires explicit determination of the content in genomic gap regions, low-coverage long-read (third-generation) sequencing can be employed.

Although TELLBASE was initially optimized for single-strain bacterial genome assembly, we are adapting it for the complete assembly of genome samples from certain screenings that might be from multiple strains. For example, colonies that can grow on antibiotic containing plates might be from one or multiple species of bacteria in clinical or environmental samples. We are continuing this work to validate TELLBASE's performance in metagenomic datasets and enhance its utility in microbiome research.

In summary, this study highlights the effectiveness and advantages of TELLBASE in achieving high-quality genome assembly and its potential implications for bioinformatics. This approach has not only enhanced the quality of bacterial genome assembly but also deepened our understanding of data obtained from the TELL-seq Illumina library. Also, the core algorithmic of TELLBASE is inherently adaptable to other linked-read technologies (e.g. 10× Genomics Chromium, Bionano Genomics optical mapping, or Hi-C). Key considerations for such extensions include standardization of sequencing data formats across diverse platforms and integration of disparate linking modalities. We anticipate the emergence of new methodologies for high-quality genome assembly from the extensive sequences produced by NGS. Moreover, the TELL-seq-TELLBASE platform may prove to be a more practical option, in terms of both cost and efficiency, for those looking to either construct a bacterial genome sequence from scratch or expand an existing one.

Key Points

- We designed the assembly process for the bacterial genome sequencing data obtained from the TELL-seq library-building technology, and developed the corresponding algorithm TELLBASE.
- Based on our developed TELLBASE algorithm, we obtained nearly perfect assembly results from both accuracy and completeness of bacterial genomic and plasmids sequences, and our assembly results significantly outperformed the results get through TuringAssembler which was developed by Universal Sequencing Technology (UST) company.
- We discovered the appropriate DNA input concentration could yield better assembly results, and demonstrated the advantages of TELL-seq technology in bacterial genome assembly.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

This work was supported by National Key Research and Development Program of China (2022YFA0806200, 2022YFC2009802 and 2021YFC2501800) and the National Natural Science Foundation of China (32000097).

Data availability

The sequencing data generated in this study have been submitted to the CNCB database (https://ngdc.cncb.ac.cn/gsub/submit/gsa/) under accession number PRJCA033007.

Supporting information

The online version contains supplementary figures and tables available.

References

1. Fleischmann RD, Adams MD, White O. *et al.* Whole-genome random sequencing and assembly of haemophilus influenzae Rd. *Science* 1995;**269**:496–512. https://doi.org/10.1126/science.7542800
2. Ferreira S, Queiroz JA, Oleastro M. *et al.* Insights in the pathogenesis and resistance of arcobacter: a review. *Crit Rev Microbiol* 2016;**42**:364–83. https://doi.org/10.3109/1040841X.2014.954523
3. Baker S, Thomson N, Weill FX. *et al.* Genomic insights into the emergence and spread of antimicrobial-resistant bacterial pathogens. *Science* 2018;**360**:733–8. https://doi.org/10.1126/science.aar3777
4. Rozman V, Mohar Lorbeg P, Treven P. *et al.* Genomic insights into antibiotic resistance and mobilome of lactic acid bacteria and bifidobacteria. *Life Sci Alliance* 2023;**6**:e202201637. https://doi.org/10.26508/lsa.202201637
5. Salamzade R, McElheny CL, Manson AL. *et al.* Genomic epidemiology and antibiotic susceptibility profiling of uropathogenic Escherichia coli among children in the United States. *mSphere* 2023;**8**:e0018423. https://doi.org/10.1128/msphere.00184-23
6. Kobras CM, Fenton AK, Sheppard SK. Next-generation microbiology: from comparative genomics to gene function. *Genome Biol* 2021;**22**:123. https://doi.org/10.1186/s13059-021-02344-9
7. Zhu Q, Mai U, Pfeiffer W. *et al.* Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains bacteria and archaea. *Nat Commun* 2019;**10**:5477. https://doi.org/10.1038/s41467-019-13443-4
8. Vincent AT, Schiettekatte O, Goarant C. *et al.* Revisiting the taxonomy and evolution of pathogenicity of the genus Leptospira through the prism of genomics. *PLoS Negl Trop Dis* 2019;**13**:e0007270. https://doi.org/10.1371/journal.pntd.0007270
9. de Lorenzo V. 15 years of microbial biotechnology: the time has come to think big-and act soon. *Microb Biotechnol* 2022;**15**:240–6. https://doi.org/10.1111/1751-7915.13993
10. Korbel JO, Urban AE, Affourtit JP. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* 2007;**318**:420–6. https://doi.org/10.1126/science.1149504

11. Bankevich A, Pevzner PA. TruSPAdes: barcode assembly of TruSeq synthetic long reads. *Nat Methods* 2016;**13**:248–50. https://doi.org/10.1038/nmeth.3737
12. Wang O, Chin R, Cheng X. *et al.* 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* 2019;**29**:798–808. https://doi.org/10.1101/gr.245126.118
13. Burton JN, Adey A, Patwardhan RP. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol* 2013;**31**:1119–25. https://doi.org/10.1038/nbt.2727
14. Chen Z, Pham L, Wu TC. *et al.* 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* 2020;**30**:898–909. https://doi.org/10.1101/gr.260380.120
15. Adey A, Kitzman JO, Burton JN. *et al.* In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome Res* 2014;**24**:2041–9. https://doi.org/10.1101/gr.178319.114
16. Kaper F, Swamy S, Klotzle B. *et al.* Whole-genome haplotyping by dilution, amplification, and sequencing. *Proc Natl Acad Sci* 2013;**110**:5552–7. https://doi.org/10.1073/pnas.1218696110
17. Kuleshov V, Snyder MP, Batzoglou S. Genome assembly from synthetic long read clouds. *Bioinformatics* 2016;**32**:i216–24. https://doi.org/10.1093/bioinformatics/btw267
18. Simpson JT, Wong K, Jackman SD. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res* 2009;**19**:1117–23. https://doi.org/10.1101/gr.089532.108
19. Yeo S, Coombe L, Warren RL. *et al.* ARCS: scaffolding genome drafts with linked reads. *Bioinformatics* 2018;**34**:725–31. https://doi.org/10.1093/bioinformatics/btx675
20. Coombe L, Zhang J, Vandervalk BP. *et al.* ARKS: chromosome-scale scaffolding of human genome drafts with linked read kmers. *BMC Bioinformatics* 2018;**19**:234. https://doi.org/10.1186/s12859-018-2243-x
21. Tolstoganov I, Bankevich A, Chen Z. *et al.* cloudSPAdes: assembly of synthetic long reads using de Bruijn graphs. *Bioinformatics* 2019;**35**:i61–70. https://doi.org/10.1093/bioinformatics/btz349
22. Guo L, Xu M, Wang W. *et al.* SLR-superscaffolder: a de novo scaffolding tool for synthetic long reads using a top-to-bottom scheme. *BMC Bioinformatics* 2021;**22**:158. https://doi.org/10.1186/s12859-021-04081-z
23. Mak L, Meleshko D, Danko DC. *et al.* Ariadne: synthetic long read deconvolution using assembly graphs. *Genome Biol* 2023;**24**:197. https://doi.org/10.1186/s13059-023-03033-5
24. Tolstoganov I, Chen Z, Pevzner PA. *et al.* SpLitteR: diploid genome assembly using TELL-Seq linked-reads and assembly graphs. 2022, bioRxiv 2022.12.08.519233. https://doi.org/10.1101/2022.12.08.519233
25. Chen T. Simple and scalable genome analysis with transposase enzyme linked long-read sequencing (TELL-Seq): from haplotype phasing to De novo assembly in a tube. *J Biomol Tech* 2019;**30**:S37. https://api.semanticscholar.org/CorpusID:209671812
26. Höjer P, Frick T, Siga H. *et al.* BLR: a flexible pipeline for haplotype analysis of multiple linked-read technologies. *Nucleic Acids Res* 2023;**51**:e114. https://doi.org/10.1093/nar/gkad1010
27. Chen S, Zhou Y, Frank M. *et al.* Fastp: a fast, ultra-fast quality control tool for FastQ files. *Bioinformatics* 2018;**34**:2637–9. https://doi.org/10.1093/bioinformatics/bty560
28. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* 2011;**17**:10–2. https://doi.org/10.14806/ej.17.1.200
29. Bankevich A, Nurk S, Antipov D. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;**19**:455–77. https://doi.org/10.1089/cmb.2012.0021
30. Krawczyk PS, Lipinski L, Dziembowski A. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res* 2018;**46**:e35. https://doi.org/10.1093/nar/gkx1321
31. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013, arXiv:1303.3997v2. https://doi.org/10.48550/arXiv.1303.3997
32. Li H, Handsaker B, Wysoker A. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9. https://doi.org/10.1093/bioinformatics/btp352
33. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res* 2004;**32**:W20–5. https://doi.org/10.1093/nar/gkh435
34. Luo R, Liu B, Xie Y. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 2012;**1**:18. https://doi.org/10.1186/2047-217X-1-18
35. Gurevich A, Saveliev V, Vyahhi N. *et al.* QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;**29**:1072–5. https://doi.org/10.1093/bioinformatics/btt086
36. Darling AC, Mau B, Blattner FR. *et al.* Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* 2004;**14**:1394, 10.1101/gr.2289704–403.
37. Alikhan NF, Petty NK, Ben Zakour NL. *et al.* BLAST ring image generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 2011;**12**:402. https://doi.org/10.1186/1471-2164-12-402
38. Chen C, Zhang Y, Yu SL. *et al.* Tracking carbapenem-producing Klebsiella pneumoniae outbreak in an intensive care unit by whole genome sequencing. *Front Cell Infect Microbiol* 2019;**9**:281. https://doi.org/10.3389/fcimb.00281
39. Zhang Y, Chen C, Wu J. *et al.* Sequence-based genomic analysis reveals transmission of antibiotic resistance and virulence among Carbapenemase-producing Klebsiella pneumoniae strains. *mSphere* 2022;**7**:3. $ $.00143-22
40. Conlan S, Park M, Deming C. *et al.* Plasmid dynamics in KPC-positive Klebsiella pneumoniae during long-term patient colonization. *MBio* 2016;**7**:e00742–16. https://doi.org/10.1128/mbio.00742-16

Briefings in Bioinformatics, 2025, **26(5)**, bbaf504
https://doi.org/10.1093/bib/bbaf504
Problem Solving Protocol