Fast bootstrap and reliable readout using hidden references for DNA data storage

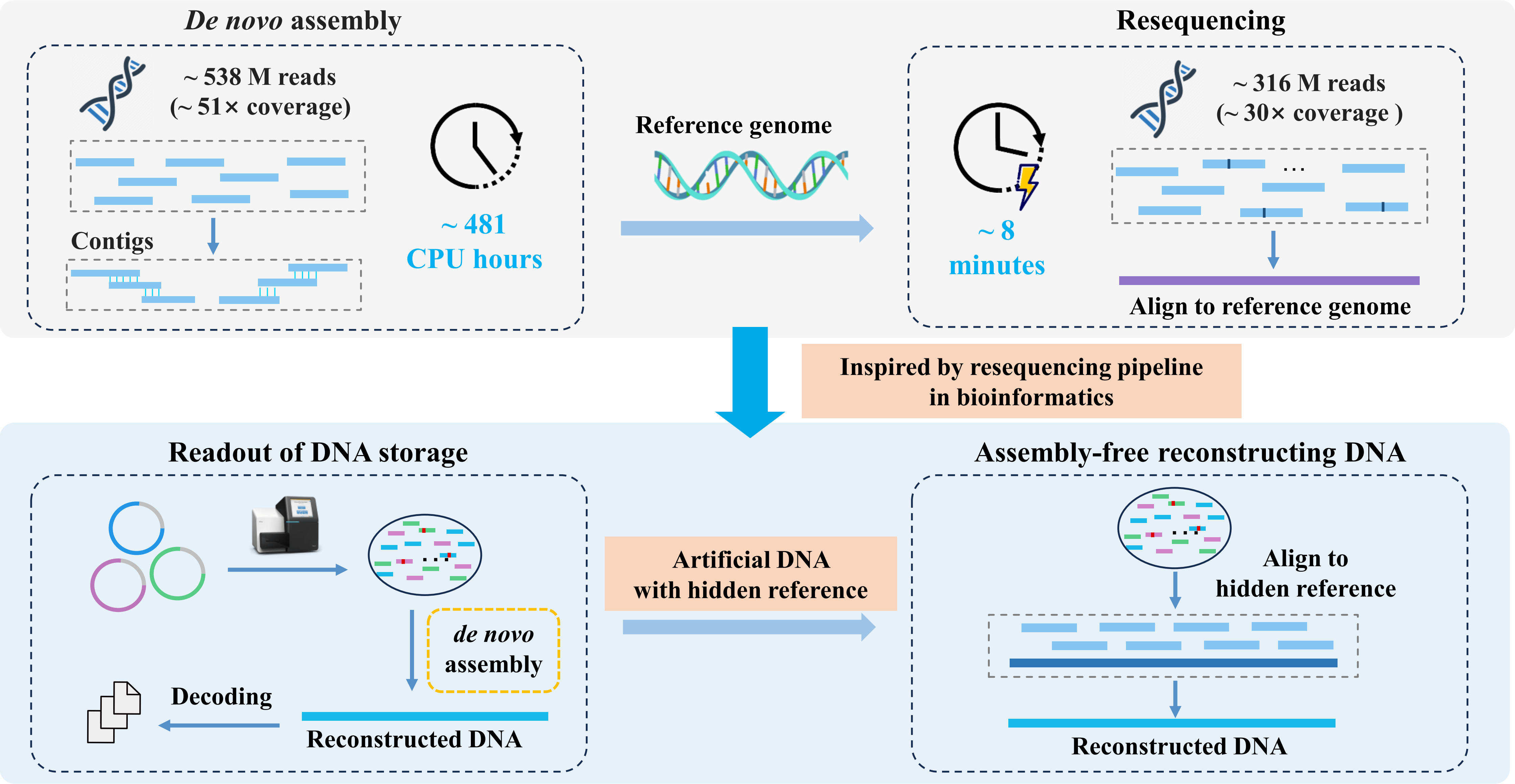
Weigang Chen[1, 2, 3], Shuang Liu[1], Quan Guo[1], Rui Qin[1], Qi Ge[1], Tingting Qi[1], Yingjin Yuan[2, 3]

[1]School of Microelectronics, Tianjin University, Tianjin, China
[2]State Key Laboratory of Synthetic Biology, Tianjin University, Tianjin, China
[3]Frontiers Science Center for Synthetic Biology (Ministry of Education),
School of Synthetic Biology and Biomanufacturing, Tianjin University, Tianjin, China

Weigang Chen, Shuang Liu, Quan Guo, Rui Qin, Qi Ge, Tingting Qi, Yingjin Yuan. 2026. Fast bootstrap and reliable readout using hidden references for DNA data storage. ***iMeta*** 5: e70105.
https://doi.org/10.1002/imt2.70105

Introduction

DNA readout: from *de novo* assembly to resequencing



De novo assembly

~ 538 M reads
(~ 51× coverage)

Contigs

~ 481
CPU hours

Reference genome

Resequencing

~ 316 M reads
(~ 30× coverage)

…

~ 8
minutes

Align to reference genome

Inspired by resequencing pipeline
in bioinformatics

Readout of DNA storage

de novo
assembly

Decoding

Reconstructed DNA

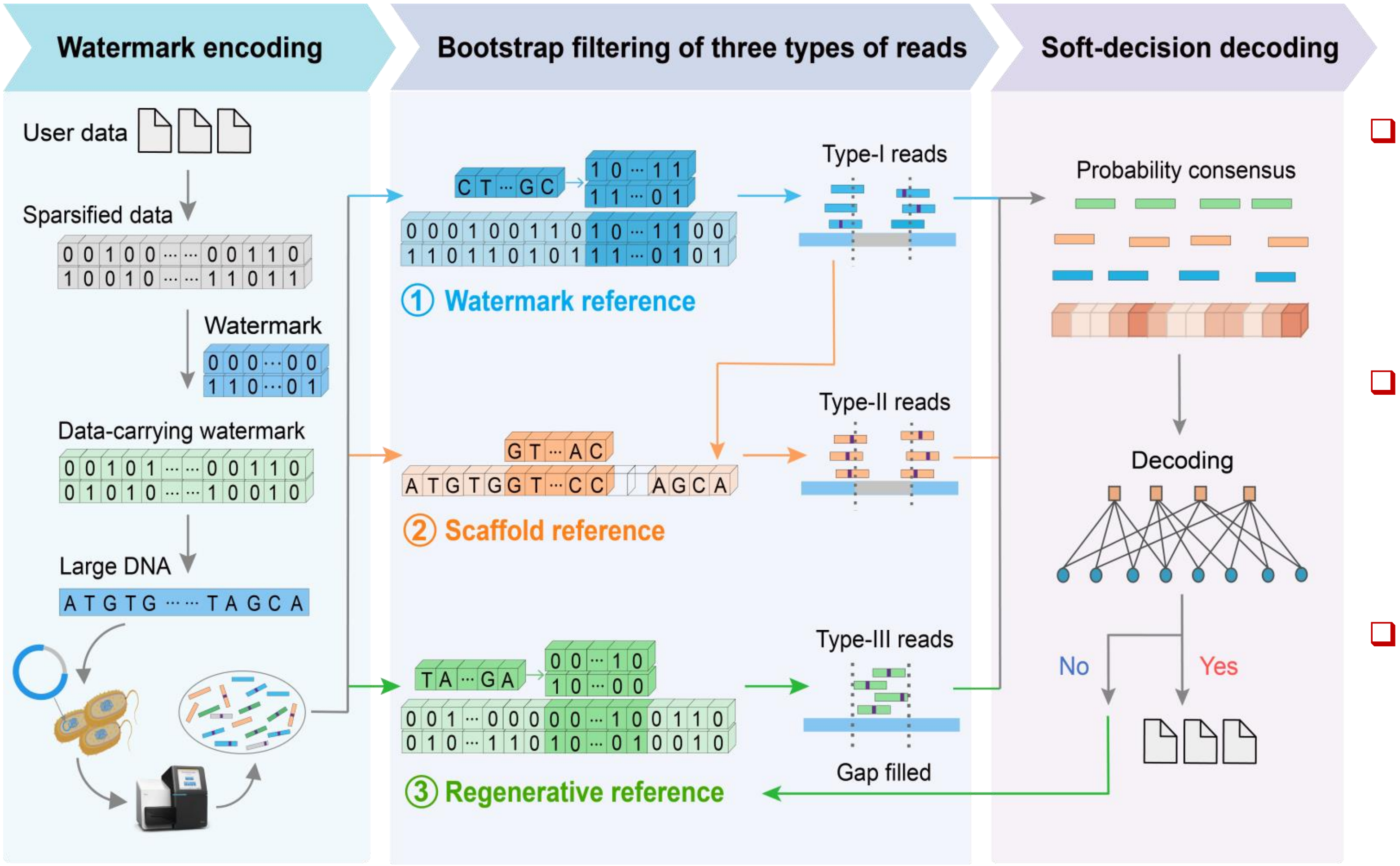
Artificial DNA
with hidden reference

Assembly-free reconstructing DNA

Align to
hidden reference

Reconstructed DNA

Highlights



- **Multiple-fold hidden references enable a fast and reliable readout scheme in a bootstrap manner for large DNA storage, transforming the *de novo* readout into a resequencing-like workflow.**

- **Correlation to the hidden watermark reference identifies low-error-rate reads, and bit-wise consensus generates soft-decision information, enabling fast data recovery.**

- **The read-by-read forward-backward algorithm corrects indel errors, and alignment to the regenerative reference fills in low-coverage regions, enabling reliable data recovery.**

Results

1. Multiple-fold references transform *de novo* readout to resequencing problem

To retrieve data from large DNA fragments, conventional methods rely on *de novo* assembly of noisy reads. These approaches require high sequencing coverage and substantial computational resources, and are further complicated by the diversity of sequencing errors.

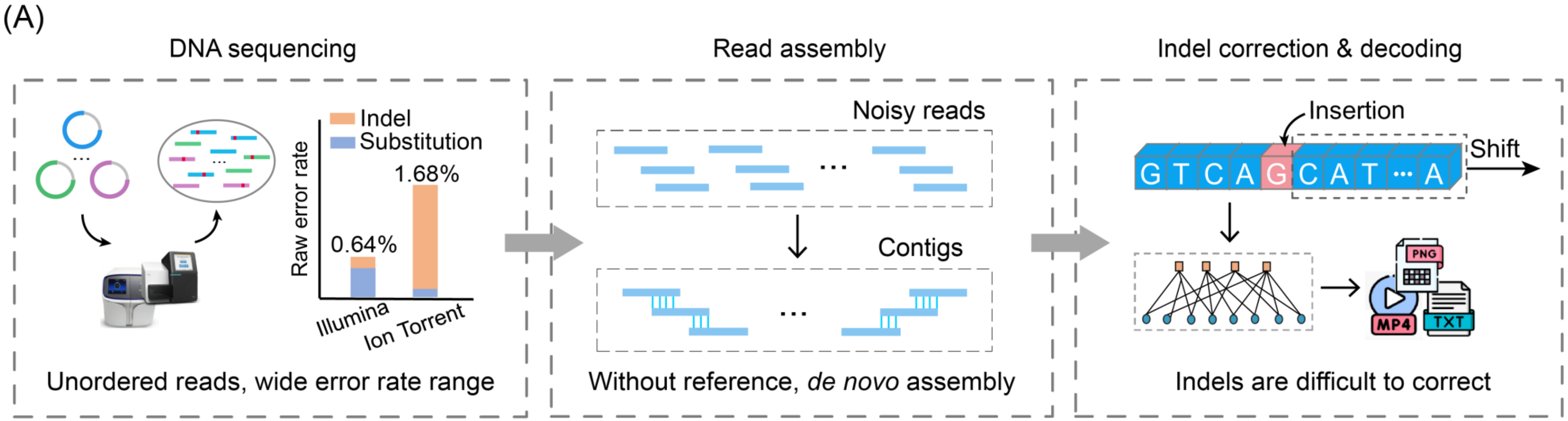
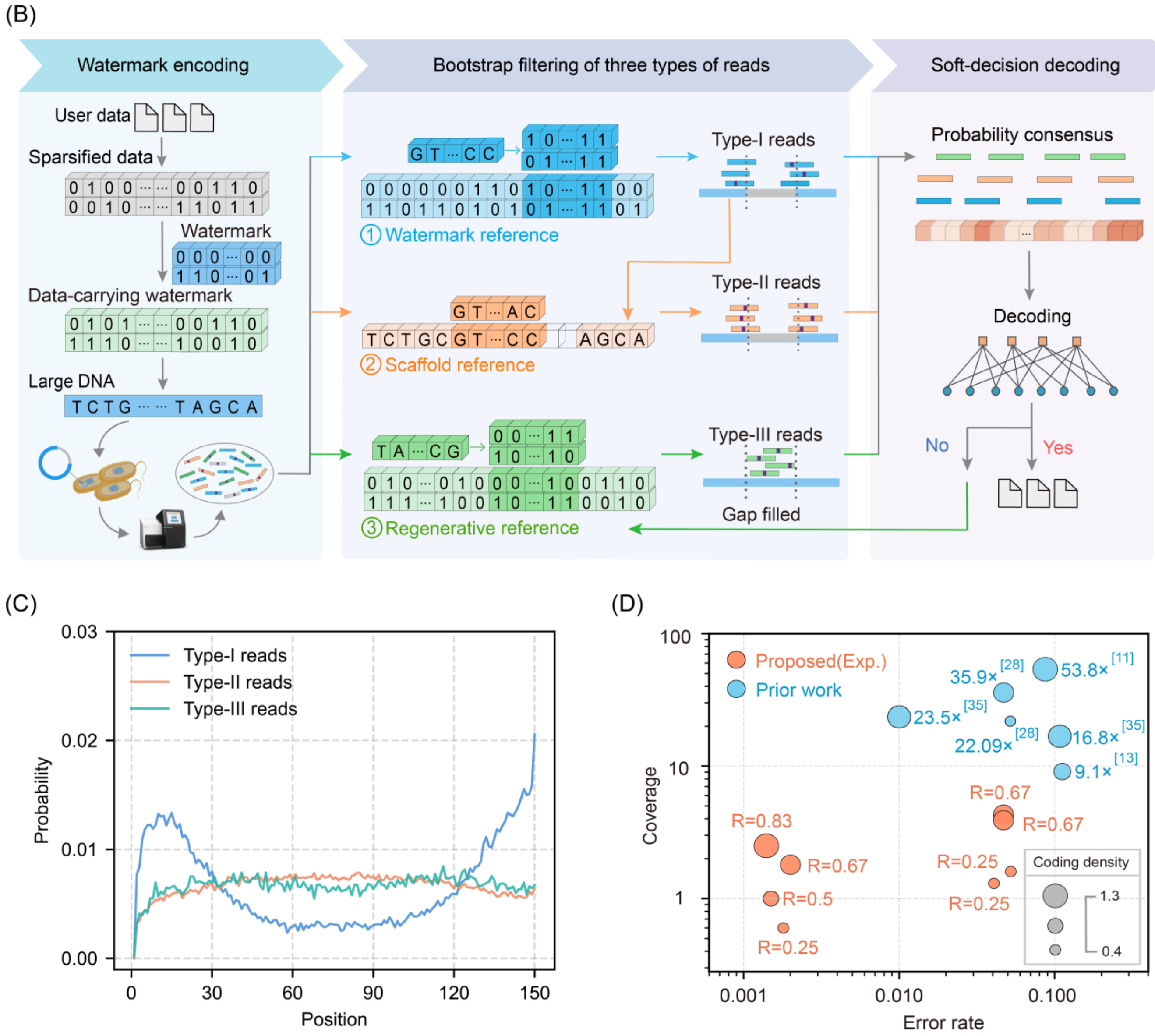


Figure 1. Readout using different filtering schemes according to multiple-fold hidden references.

(A) Existing methods for large DNA fragments after high-throughput sequencing typically require assembly.

Results

1. Multiple-fold references transform *de novo* readout to resequencing problem



- ❑ **For large DNA fragments with embedded watermark, multiple-fold references are constructed to identify reads with distinct features, supporting bootstrap and reliable readout.**
- ❑ **Compared to state-of-the-art methods, our approach achieved low-coverage readout across a wide error rate range.**

Figure 1. Readout using different filtering schemes according to multiple-fold hidden references.

(B) The workflow of multi-reference-assisted readout with watermarked large DNA fragments.

(C) Distribution of indel error positions across different types of reads in 120 independent experiments.

(D) Comparison with other DNA data storage schemes.

Results

2. Correlation to hidden watermark reference supports rapid read positioning

The proposed method effectively identifies the read position under interference from the superimposed sparsified data and sequencing errors.

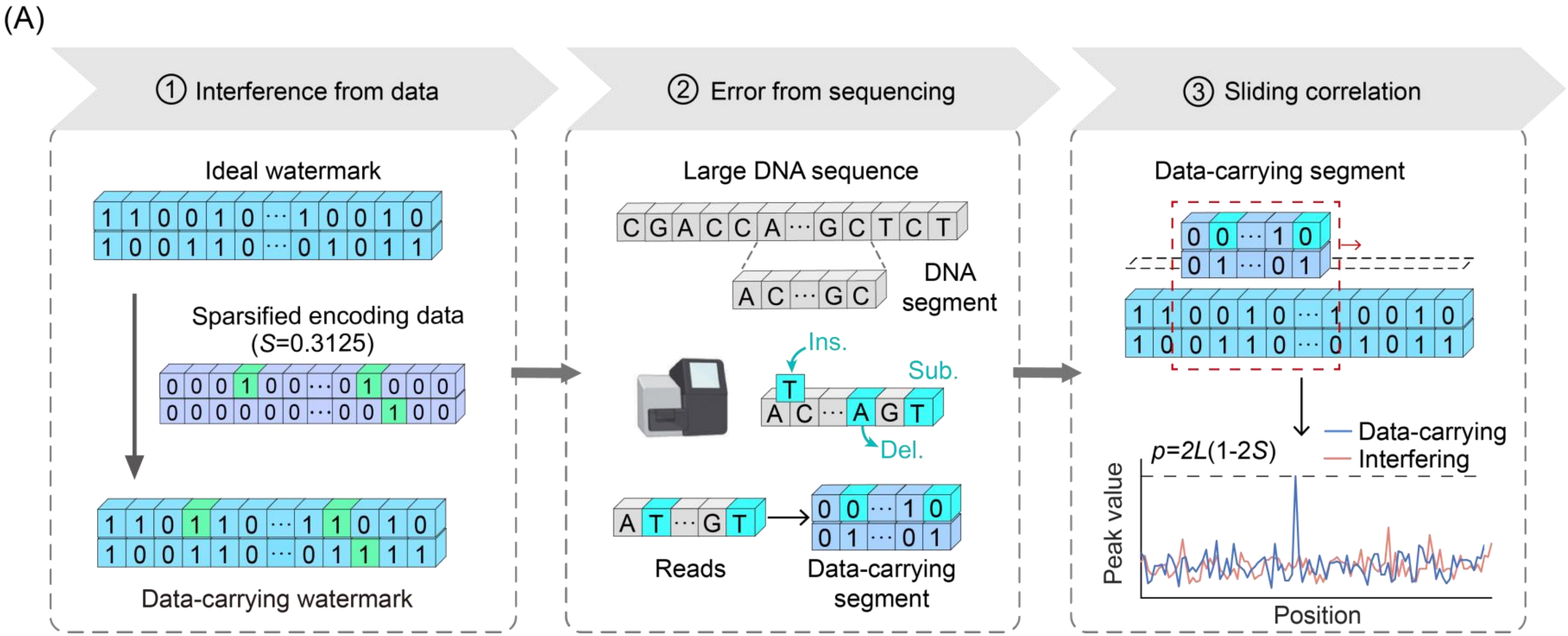


Figure 2. Read positioning using correlation between noisy reads and watermark reference.

(A) Workflow of the sliding correlation peak detection for alignment.

Results

2. Correlation to hidden watermark reference supports rapid read positioning

- ❑ **Interfering reads show much lower correlation peaks than valid data reads and can be excluded by applying a threshold.**
- ❑ **With a high proportion of interfering reads, the correlation-method achieves a classification accuracy of ~99%.**

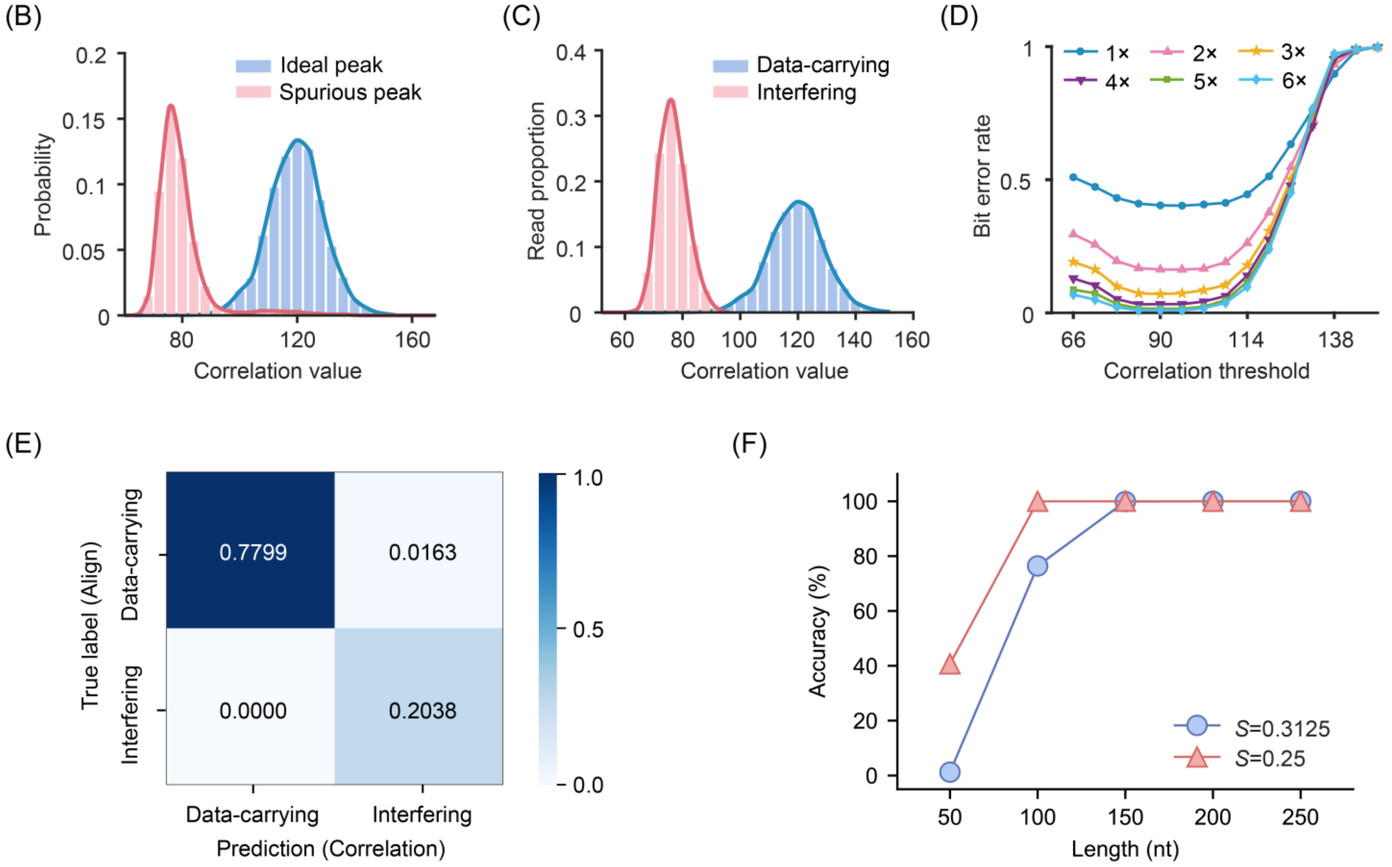


Figure 2. Read positioning using correlation between noisy reads and watermark reference.

(B) Distribution of ideal and spurious correlation peaks of data-carrying segments.

(C) Comparison of correlation peak values between data-carrying reads and interfering reads.

(D) Bit error rate (BER) of consensus sequences as a function of correlation threshold.

(E) The accuracy of the correlation-based method to distinguish between data-carrying and interfering reads.

(F) Correlation accuracy for sequencing reads of varying lengths.

Results

3. Bit-wise consensus and probability generation for soft-decision recovery

This study proposes a rapid recovery method based on bit-wise consensus and probability generation. Bit-wise consensus is obtained through low-complexity majority voting, providing soft-input information to the LDPC decoder and improving data recovery accuracy.

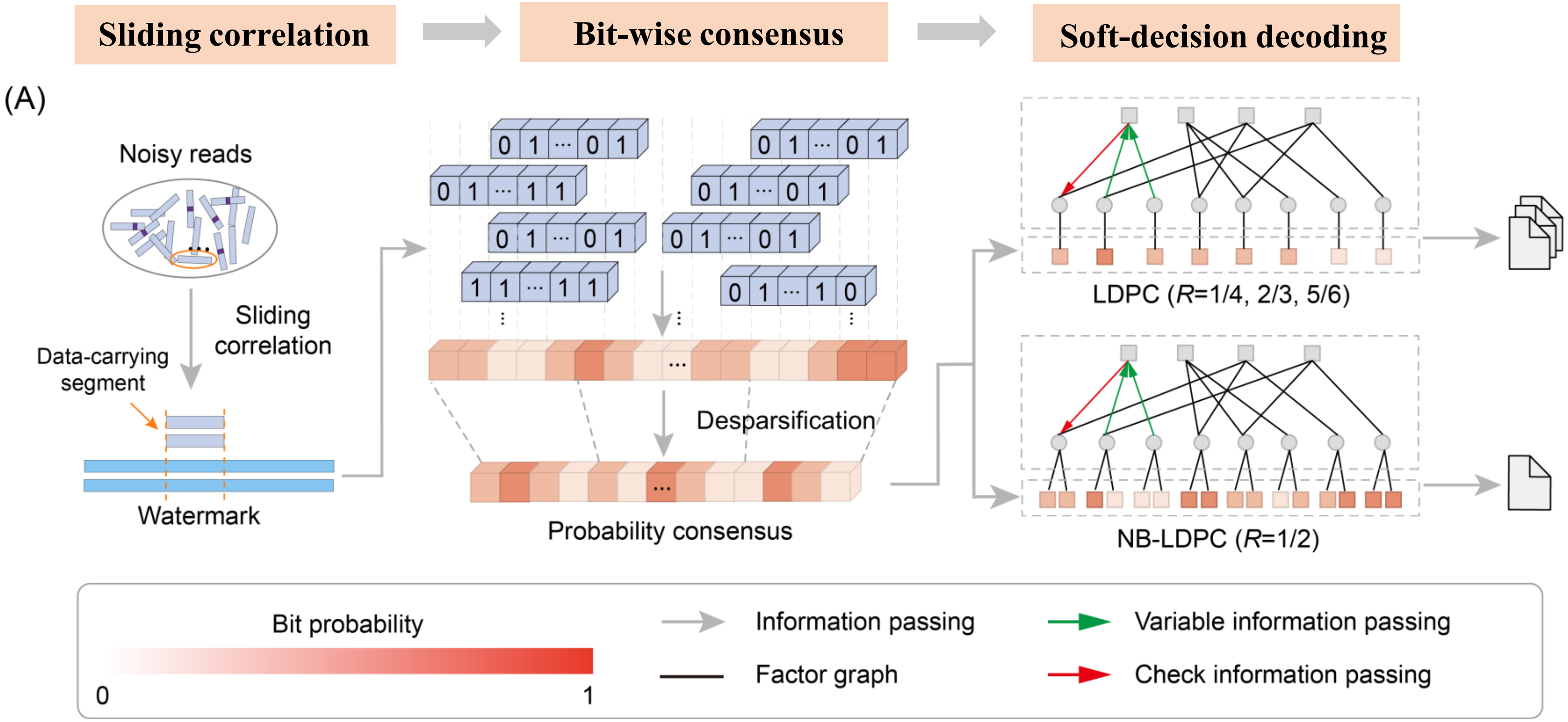


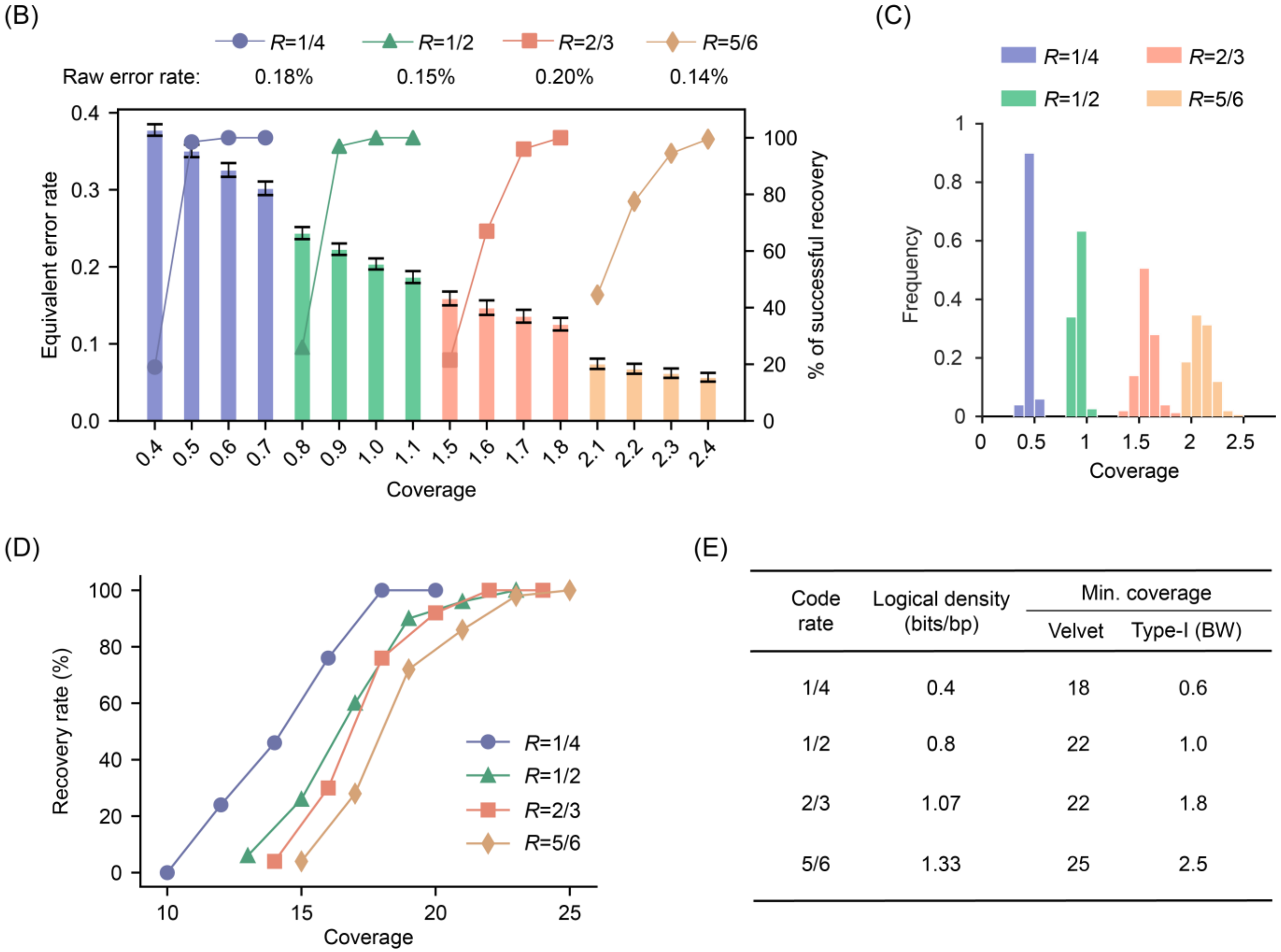
Figure 3. Soft-decision data recovery using bit-wise consensus.

(A) Workflow of the proposed data recovery method.

Results

3. Bit-wise consensus and probability generation for soft-decision recovery

Rapid data recovery was achieved on Illumina sequencing data at 0.6–2.5× coverage, much lower than the ~20× typically required by assembly-based methods.



Code rate	Logical density (bits/bp)	Min. coverage	
		Velvet	Type-I (BW)
1/4	0.4	18	0.6
1/2	0.8	22	1.0
2/3	1.07	22	1.8
5/6	1.33	25	2.5

Figure 3. Soft-decision data recovery using bit-wise consensus.

(B) Recovery performance without genome interference.

(C) Minimum sequencing coverage required for error-free recovery across 150 trials at different low-density parity-check (LDPC) code rates.

(D) Recovery performance of the Velvet-based method.

(E) Comparison of the proposed recovery scheme with an assembly-based approach (Velvet).

Results

4. Read-by-read forward-backward algorithm corrects insertions and deletions

- ❑ A scaffold reference sequence constructed from Type-I reads is used to partially align correlation-failed reads, effectively rescuing indel-containing reads.
- ❑ The forward–backward algorithm corrects indel errors and generates soft information for decoding.

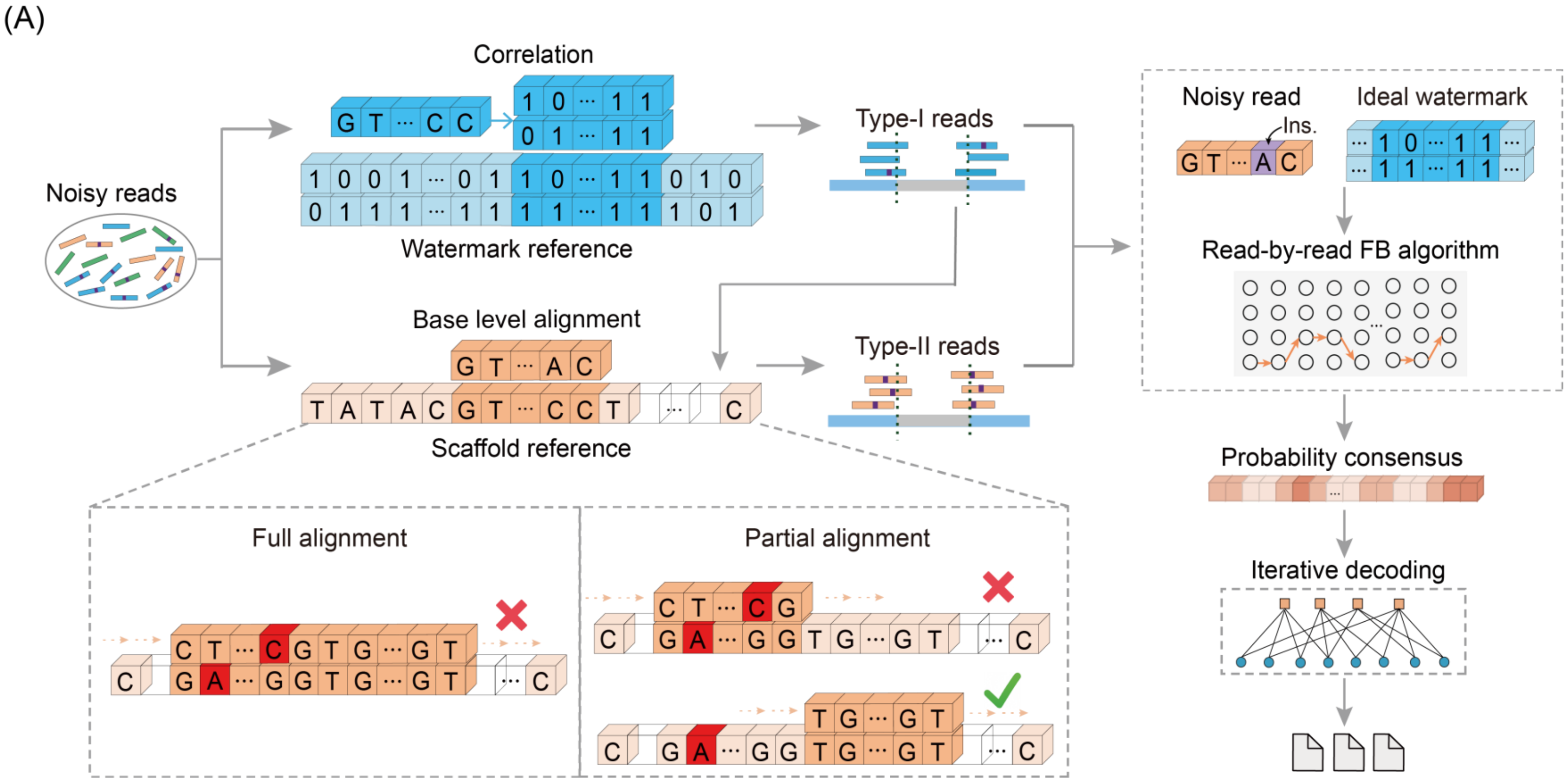


Figure 4. The forward-backward algorithm (FBA) corrects indel errors.

(A) Workflow of indel correction. Partial-length alignment improves the proportion of reads identified.

结果

4. Read-by-read forward-backward algorithm corrects insertions and deletions

The combination of FBA and the incorporation of Type-II reads progressively improves recovery performance, enabling error-free recovery at 0.8-3.5× coverage.

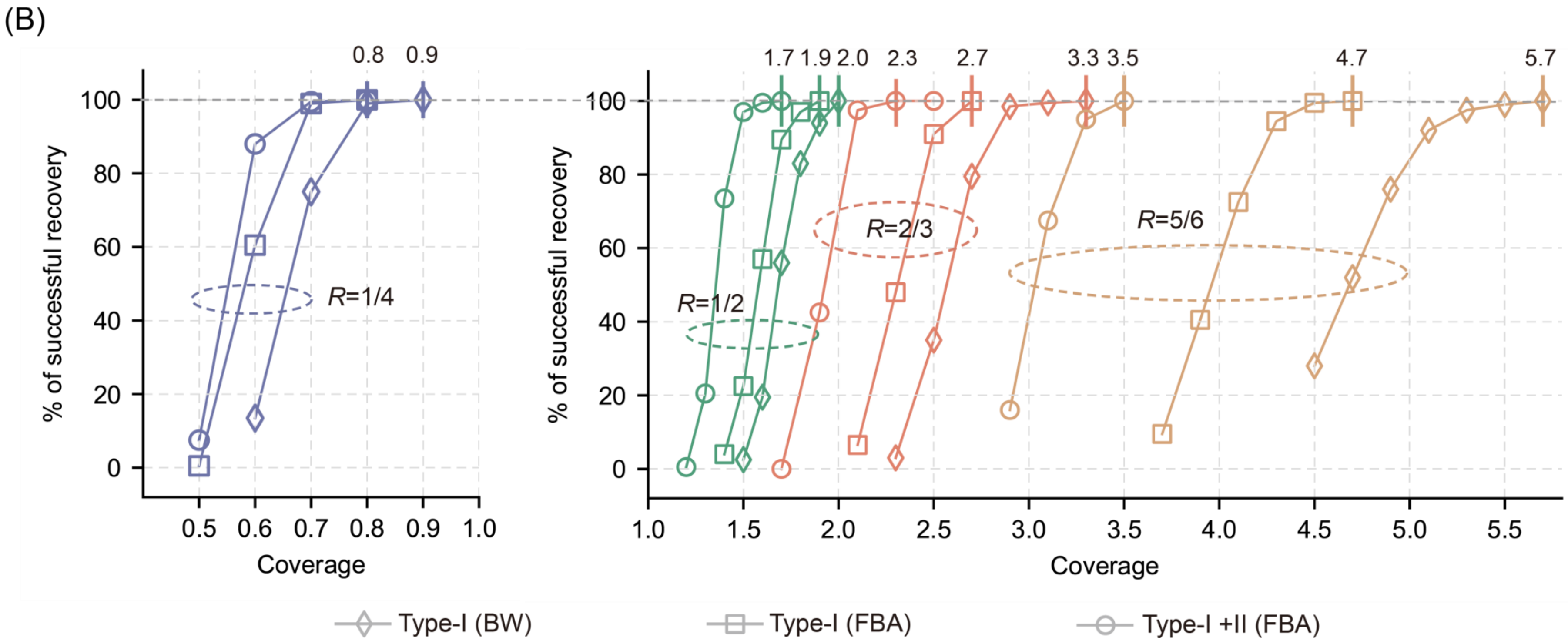


Figure 4. The forward-backward algorithm (FBA) corrects indel errors.

(B) Comparison of recovery results using different types of reads and soft-decision information generation schemes.

(BW: Bit-wise probability generation; FBA: Probability generation with FBA)

Results

5. Iterative alignment to regenerative reference fills in low-coverage gap

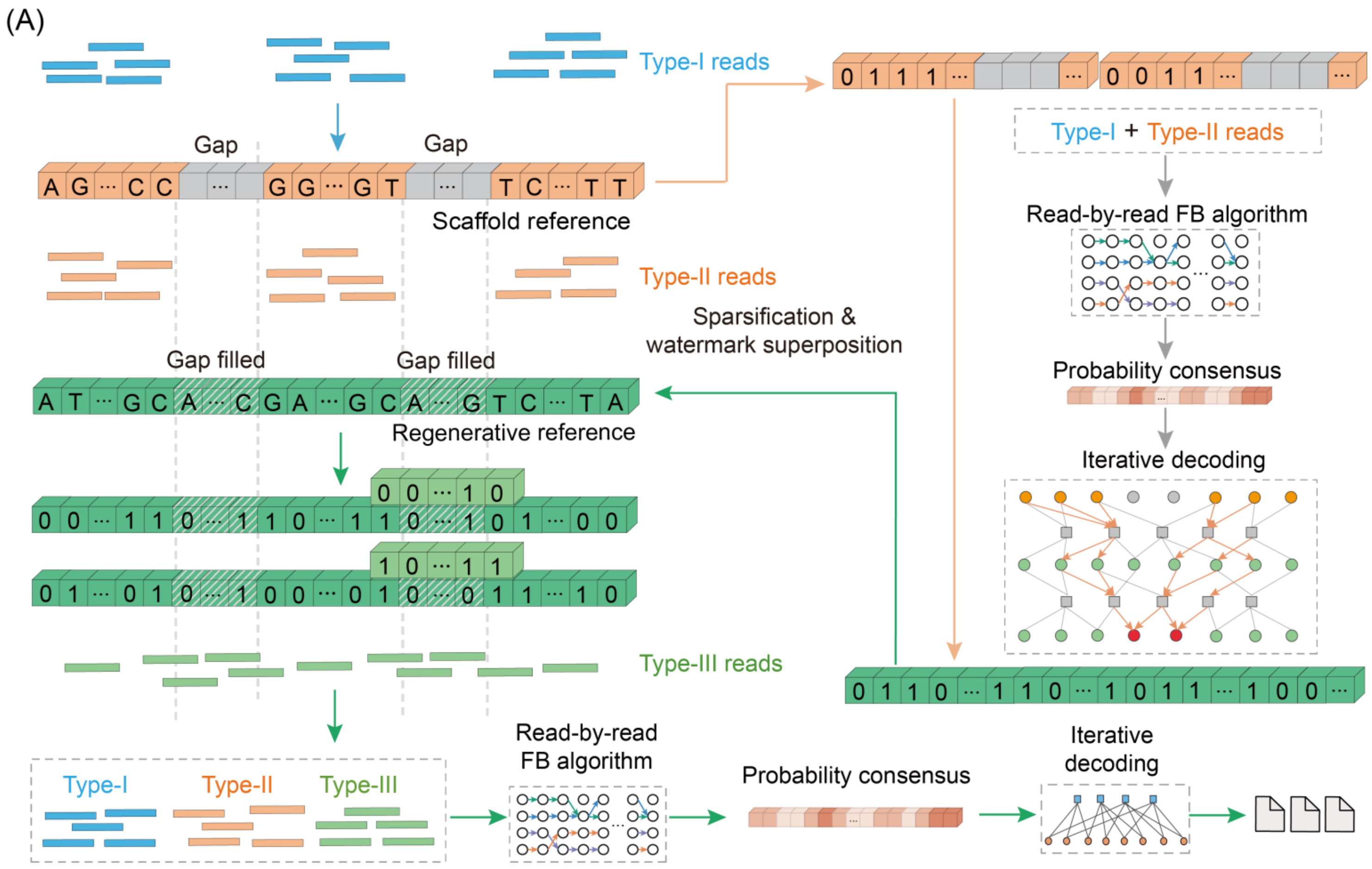


Figure 5. Gap-filling scheme with regenerative reference.

(A) Workflow of the gap-filling scheme.

❑ **Since scaffold references often contain gaps, a regenerative reference is constructed from the decoded codeword.**

❑ **Remaining reads are aligned to the regenerative reference to identify Type-III reads for filling low-coverage regions.**

Results

5. Iterative alignment to regenerative reference fills in low-coverage gap

(B)

Dataset: DNA-40.5kb-MC-Sim-2 (*R* = 5/6)

Code rate	*R* = 5/6
Ins.	0.3%
Del.	0.3%
Sub.	0.6%
Total	1.2%

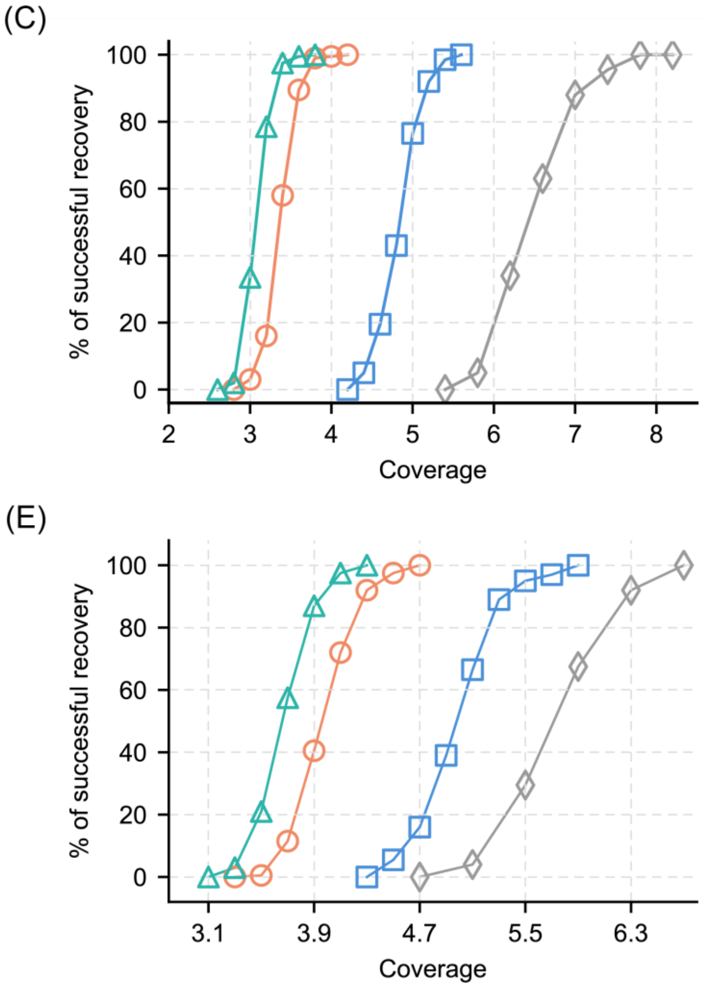
Simulated dataset representing a high-indel error profile (Ins.: del.: sub. = 1: 1: 2)

(D)

Dataset: DNA-40.5kb-EM-ONT-1 (*R* = 2/3)

Code rate	*R* = 2/3
Ins.	1.4%
Del.	1.5%
Sub.	1.8%
Total	4.7%

Nanopore sequencing on an R10.4.1 flow cell, with super-accurate basecalling (Guppy v7.0.9)



- ❑ **For simulated data with an error rate of 1.2%, error-free recovery was achieved at a coverage of 3.7×.**
- ❑ **For nanopore sequencing data with an error rate of 4.7%, error-free recovery was achieved at a coverage of 4.3× coverage.**

Figure 5. Gap-filling scheme with regenerative reference.

(B) Error profile of the simulated dataset (DNA-40.5kb-MC-Sim-2, *R* = 5/6).

(C) Recovery performance under different sequencing coverages for the dataset in (B) .

(D) Error profile of the nanopore sequencing dataset (DNA-40.5kb-EM-ONT-1, *R* = 2/3).

(E) Recovery performance under different sequencing coverages for the dataset in (D).

Summary

We demonstrate a multi-stage alignment and error correction strategy, transforming the *de novo* readout into a resequencing-like workflow.

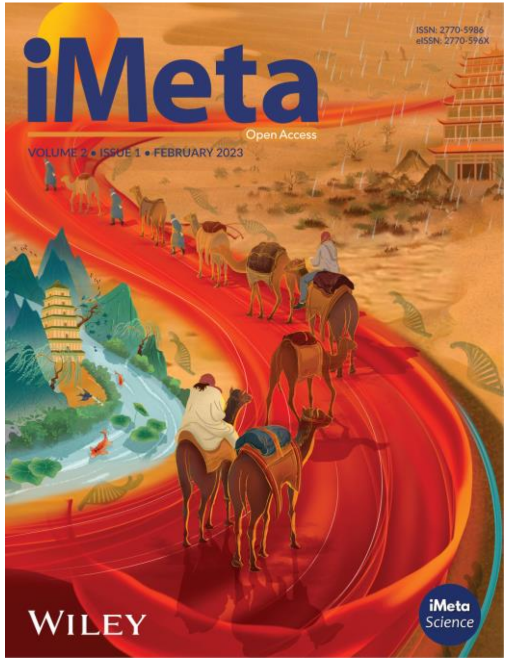
- ❑ Correlation to the hidden watermark reference identifies low-error-rate reads, enabling rapid data recovery.
- ❑ Scaffold reference and FBA rescue reads with indels improving read utilization and consensus accuracy.
- ❑ Alignment to the regenerative reference fills in low-coverage regions, reducing consensus erasure.

Leveraging multiple-fold hidden references, our method enables fast bootstrap and reliable readout for large DNA fragment storage, demonstrated across both Illumina and ONT sequencing platforms.

Weigang Chen, Shuang Liu, Quan Guo, Rui Qin, Qi Ge, Tingting Qi, Yingjin Yuan. 2026. Fast bootstrap and reliable readout using hidden references for DNA data storage. ***iMeta*** 5: e70105.
https://doi.org/10.1002/imt2.70105

iMeta: To be top journals in biology and medicine

WILEY



"***iMeta***" launched in 2022 by iMeta Science Society, impact factor (IF) **33.2**, ranking top 65/22249 in world and 2/161 in the microbiology. It aims to publish innovative and high-quality papers with broad and diverse audiences. **Its scope is similar to *Cell***, *Nature Biotechnology/Methods/Microbiology/Medicine/Food*. Its unique features include video abstract, bilingual publication, and social media with 600,000 followers. Indexed by **SCIE/ESI**, **PubMed**, **Google Scholar** etc.

"***iMetaOmics***" launched in 2024, with a **target IF>10, and its scope is similar to *Nature Communications***, ***Cell Reports***, *Microbiome, ISME J, Nucleic Acids Research, Briefings in Bioinformatics,* etc.

"***iMetaMed***" launched in 2025, with a **target IF>15, similar to *Med, Cell Reports Medicine***, *eBioMedicine, eClinicalMedicine* etc.

Society: http://www.imeta.science
Publisher: https://wileyonlinelibrary.com/journal/imeta
iMeta: https://wiley.atyponrex.com/journal/IMT2
Submission: iMetaOmics: https://wiley.atyponrex.com/journal/IMO2
iMetaMed: https://wiley.atyponrex.com/journal/IMM3

iMetaScience

iMetaScience

office@imeta.science
imetaomics@imeta.science

Promotion Video

Update
2025/7/6