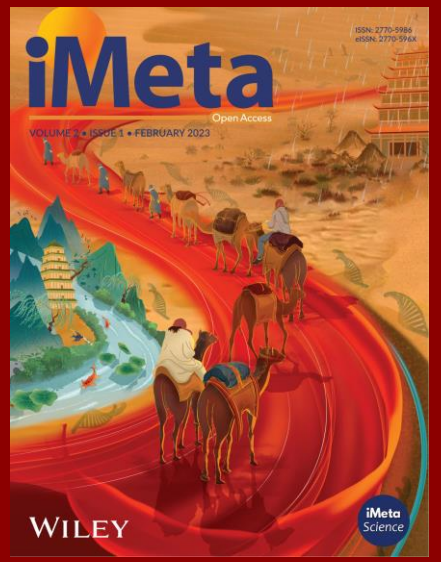
基于隐藏参考序列的DNA数据存储快速自举式可靠读出

陈为刚[1, 2, 3], 刘双[1], 郭全[1], 秦蕊[1], 葛奇[1], 齐婷婷[1], 元英进[2, 3]

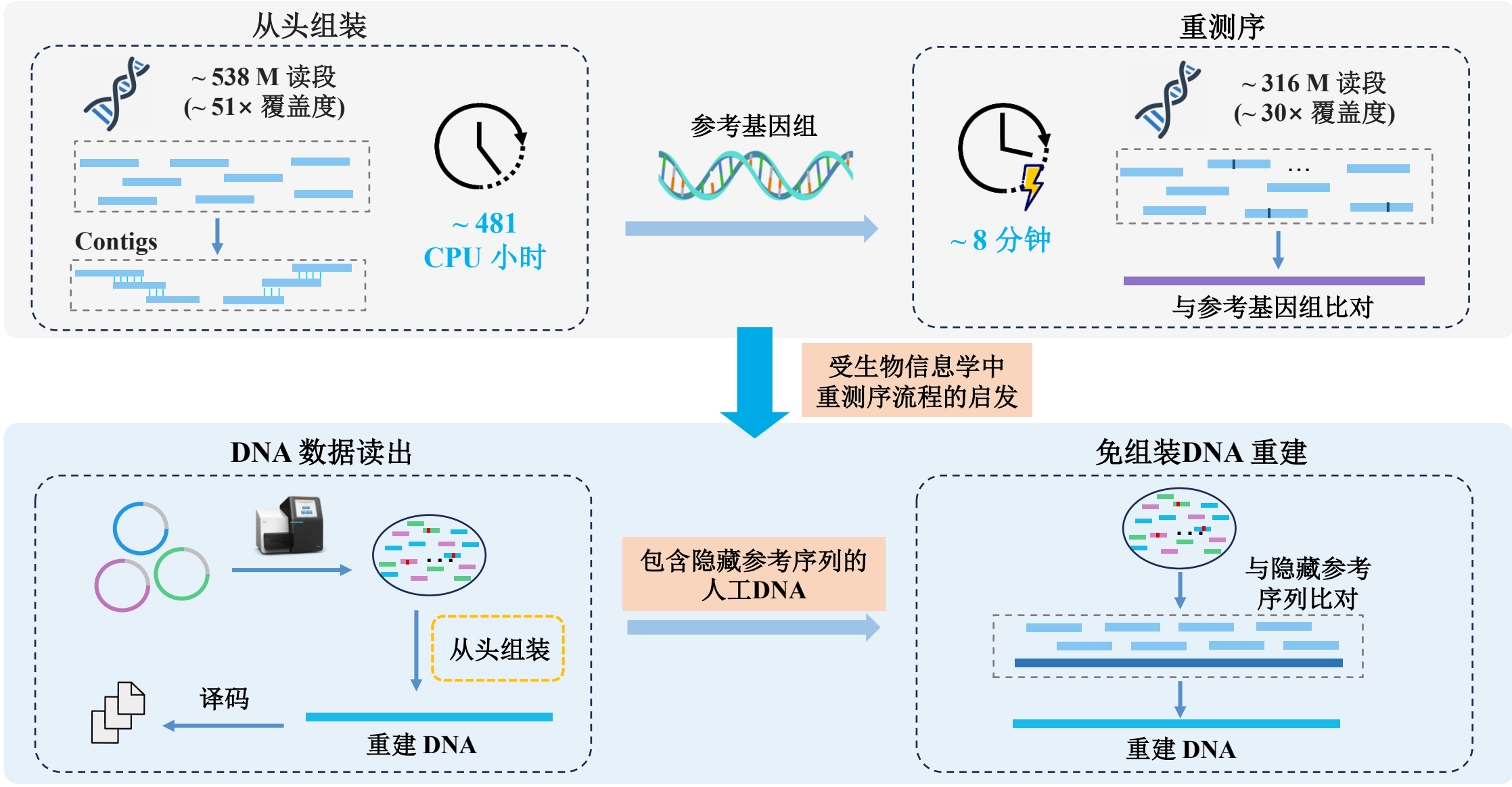
[1]天津大学微电子学院
[2]天津大学合成生物技术全国重点实验室
[3]教育部“珠峰计划”合成生物学前沿科学中心,
天津大学合成生物与生物制造学院



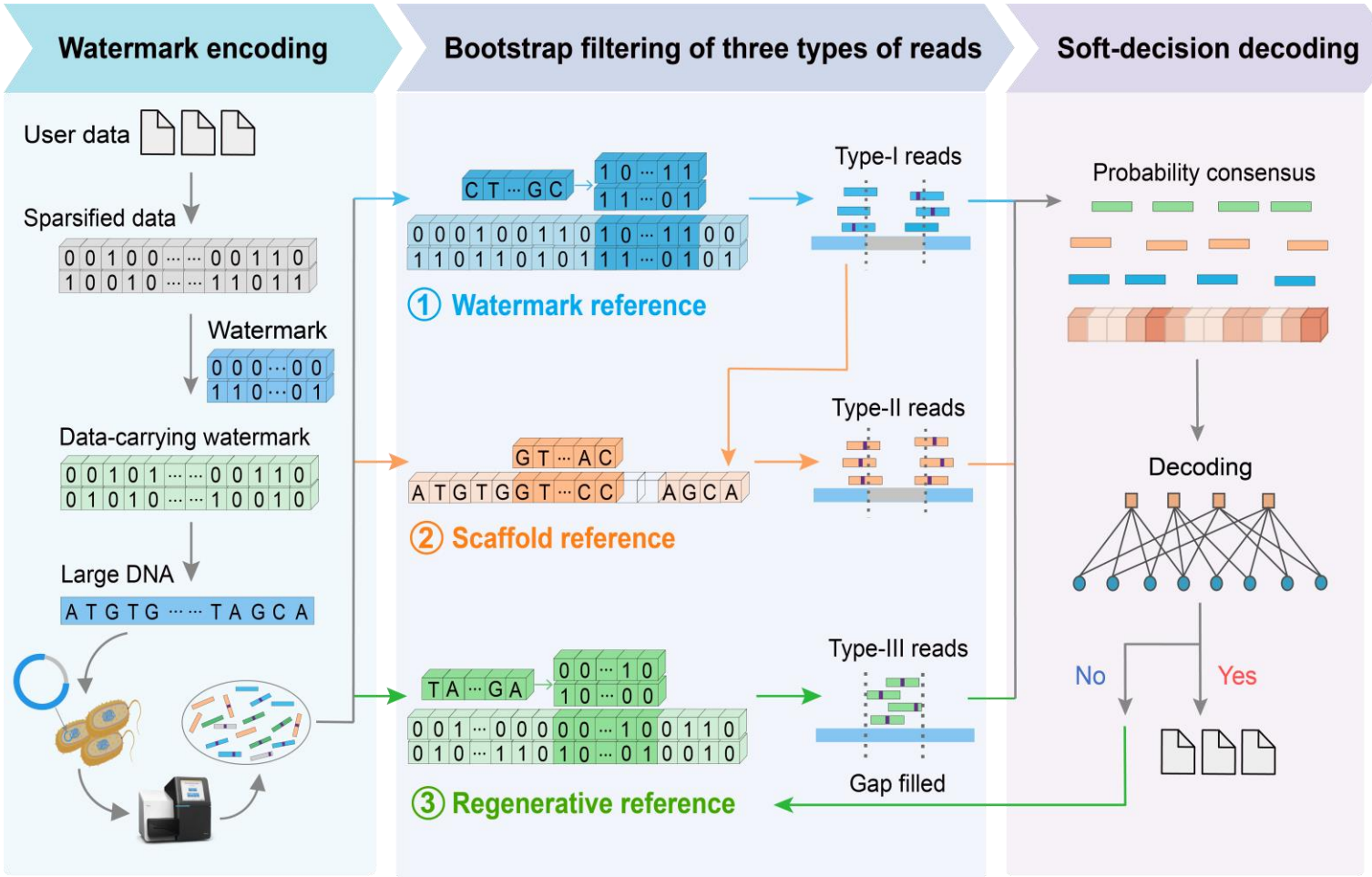
Weigang Chen, Shuang Liu, Quan Guo, Rui Qin, Qi Ge, Tingting Qi, Yingjin Yuan. 2026. Fast bootstrap and reliable readout using hidden references for DNA data storage. ***iMeta*** 5: e70105.
https://doi.org/10.1002/imt2.70105

简介

DNA读出：从头组装到重测序



亮点



- 利用多重隐藏参考序列，实现了面向大片段DNA存储的快速自举式可靠读出，将从头读出转化为类似重测序的工作流程；
- 与水印参考序列滑动相关识别低错误率读段，并利用逐比特共识生成软判决信息，用于快速数据恢复；
- 采用逐读段前向-后向算法纠正插入/删除错误，并与再生参考序列比对填补低覆盖度区域，用于可靠数据恢复。

结果

一、多重隐藏参考将从头读出问题转化为重测序问题

大片段DNA存储数据恢复的传统方法依赖于噪声读段的从头组装，组装过程需要较高的测序覆盖度和大量计算资源，并且由于不同测序平台误差分布差异较大而变得更加复杂。

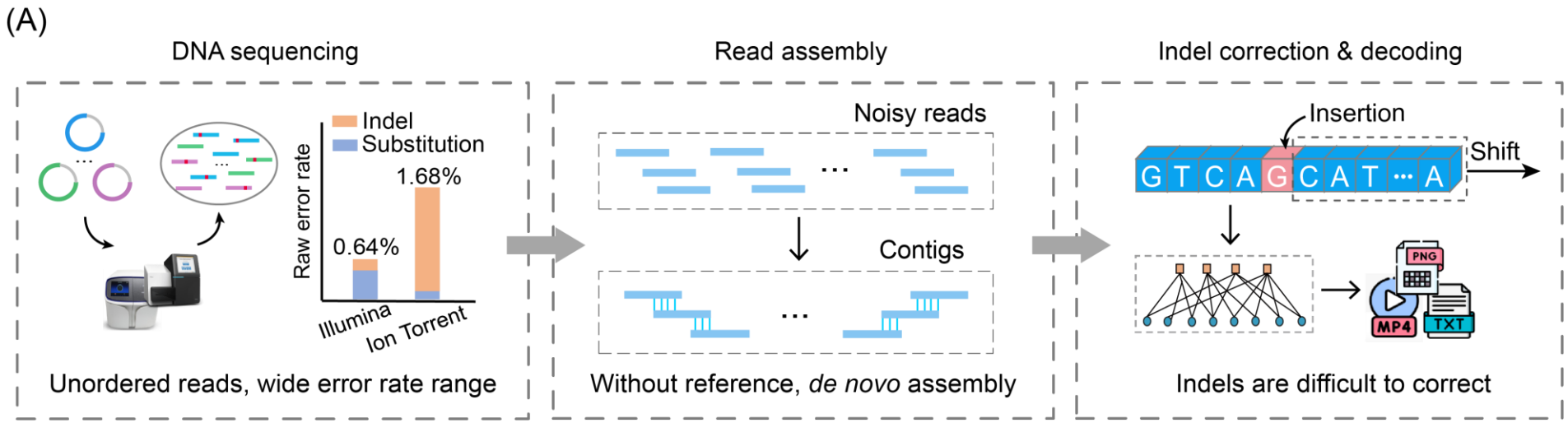


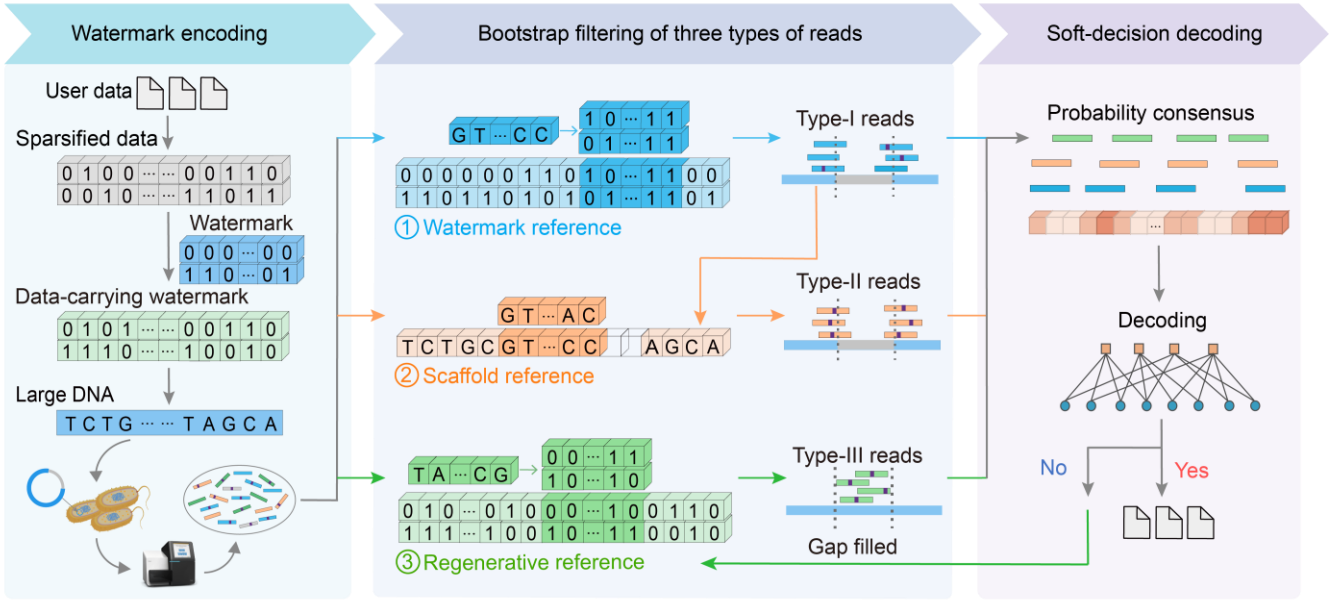
图 1. 基于多重隐藏参考的不同读段过滤与读出方案

(A) 大片段DNA存储数据恢复依赖从头组装

结果

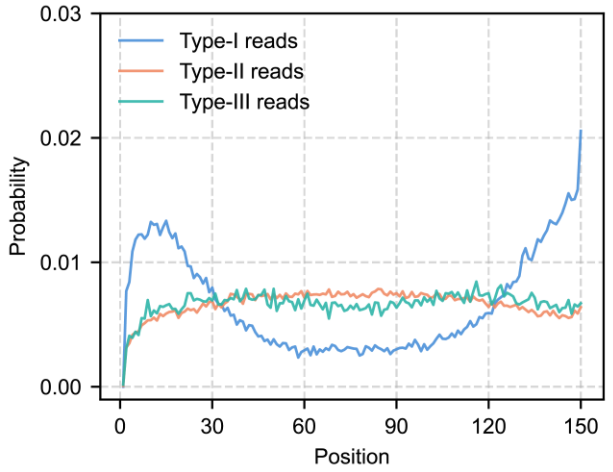
一、多重隐藏参考将从头读出问题转化为重测序问题

(B)



- 针对叠加水印编码的大片段DNA，构建多重参考序列渐进式识别具有不同特征的读段，实现快速自举式读出；
- 与现有最先进方法相比，本方法在较宽泛的错误率范围内实现了低覆盖度读出。

(C)



(D)

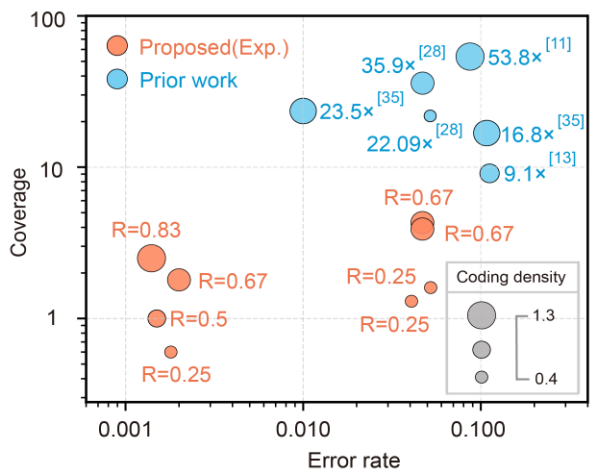


图1. 基于多重隐藏参考的不同读段过滤与读出方案

(B) 叠加水印大片段DNA的多重参考辅助读出工作流程

(C) 120次独立实验中不同类型读段的插入/删除错误位置分布

(D) 与其他DNA数据存储方案的比较

结果

二、与水印参考序列滑动相关实现快速读段定位

在稀疏编码数据和测序错误的两类噪声干扰下，本方法可通过与水印参考序列滑动相关实现读段快速定位。

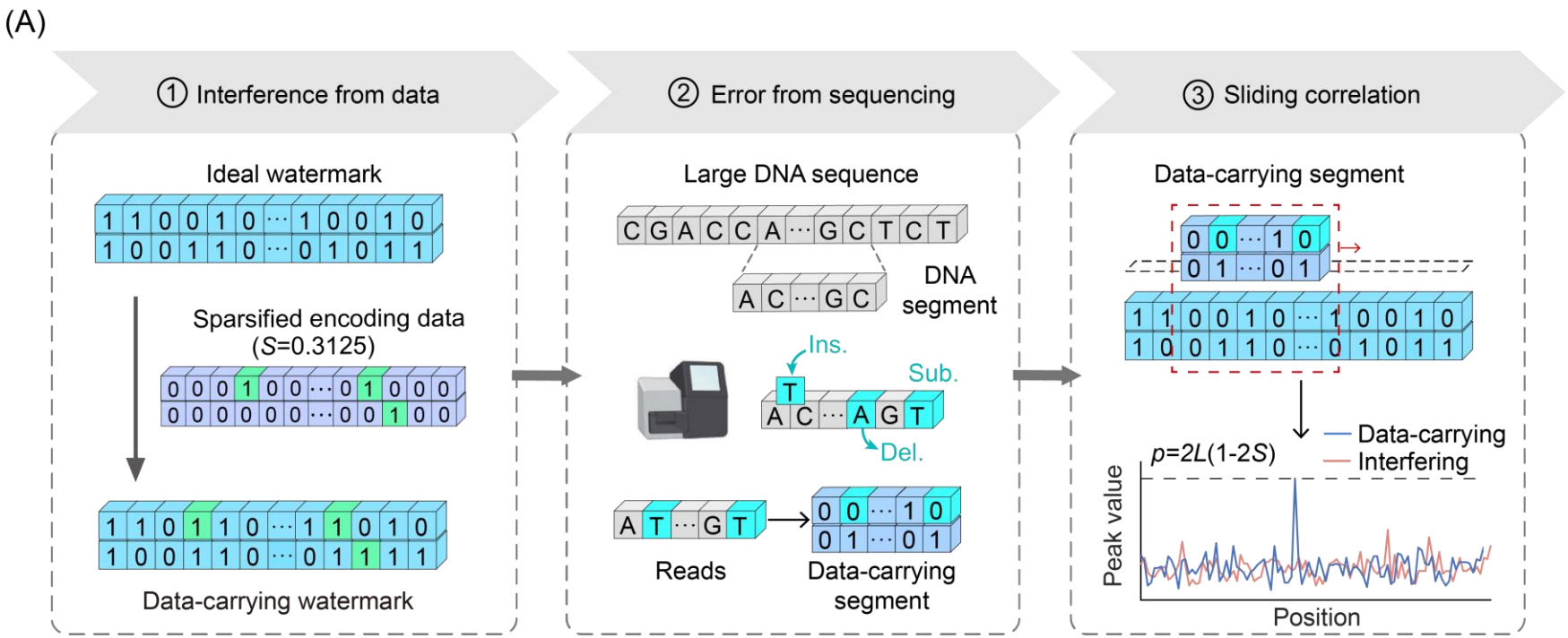


图2. 噪声读段与水印参考序列滑动相关实现读段定位

(A) 基于滑动相关的读段对齐工作流程

结果

二、与水印参考序列滑动相关实现快速读段定位

- 干扰读段的相关峰值显著低于数据读段，可通过设定相关阈值过滤干扰读段；
- 在干扰读段比例比较高的条件下，滑动相关方法可实现~99%的分类准确率。

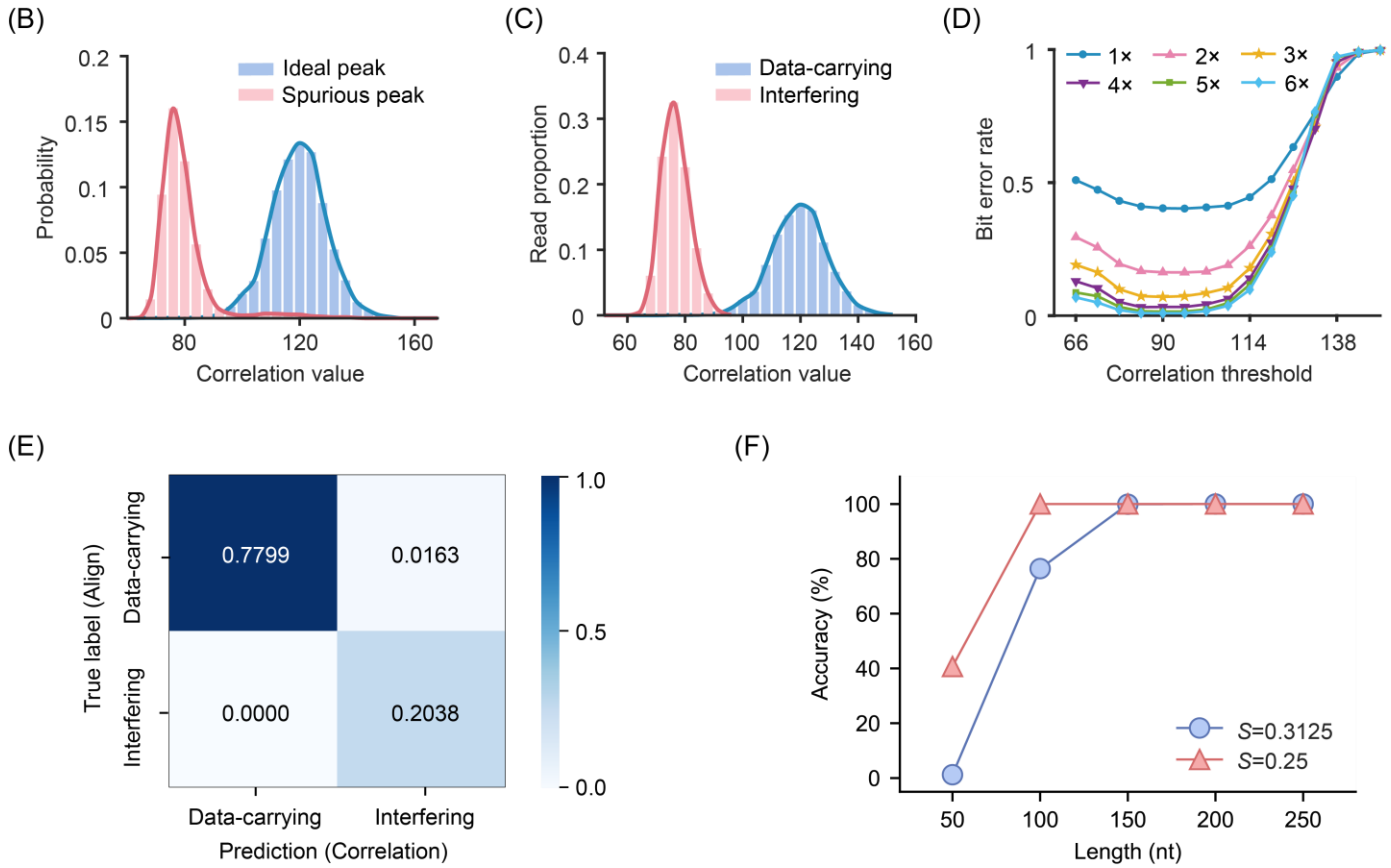


图2. 噪声读段与水印参考序列滑动相关实现读段定位

(B) 数据读段中理想相关峰值与杂散相关峰值分布

(C) 数据读段与干扰读段的相关峰值分布

(D) 不同相关阈值下共识序列的比特错误率

(E) 滑动相关方法的读段分类性能

(F) 不同读段长度下滑动相关定位准确率

结果

三、基于逐比特共识与概率生成的软判决数据恢复

本研究提出一种基于逐比特共识和概率生成的快速恢复方法，通过逐比特多数投票生成概率级共识，为LDPC译码器提供软信息输入，提升数据恢复的准确性。

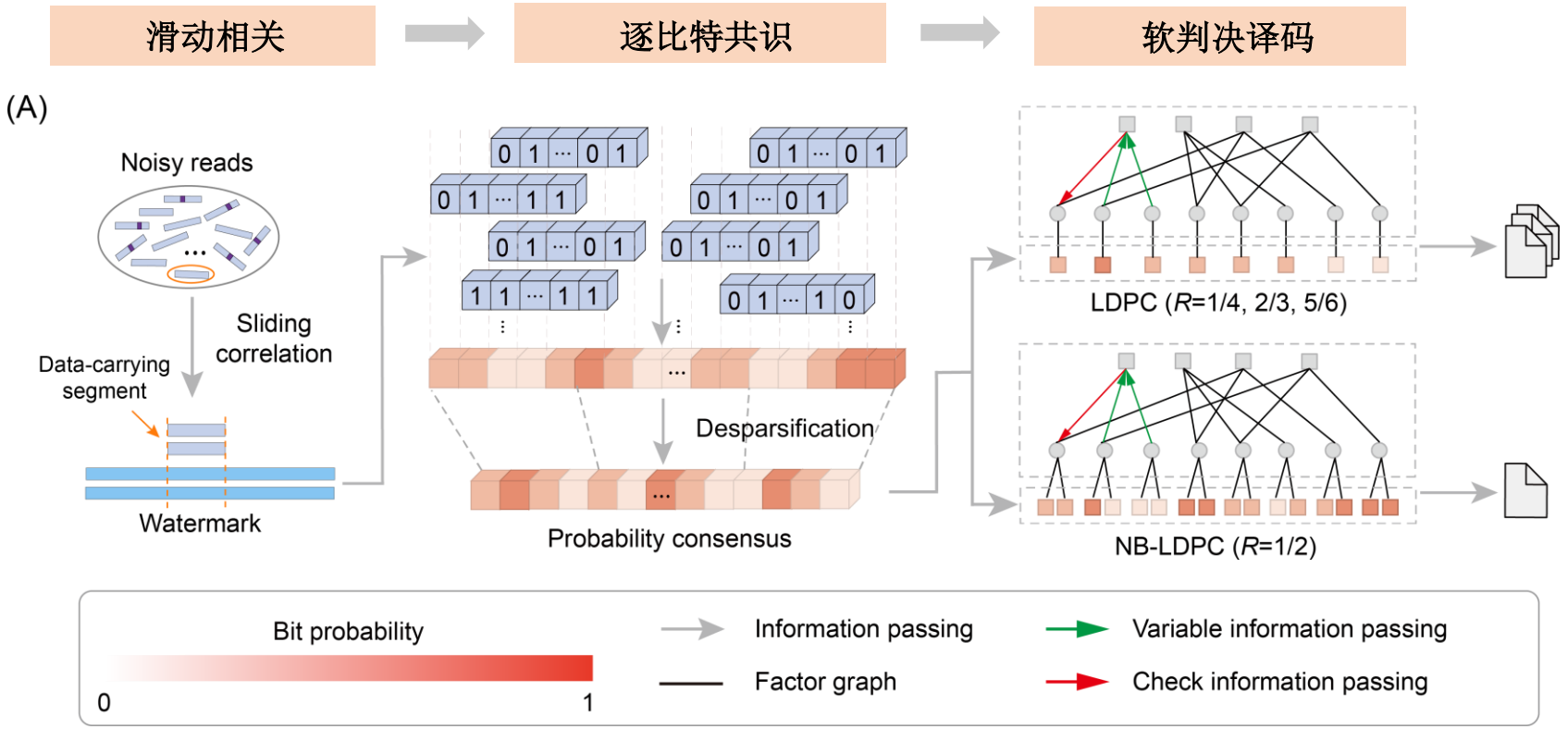


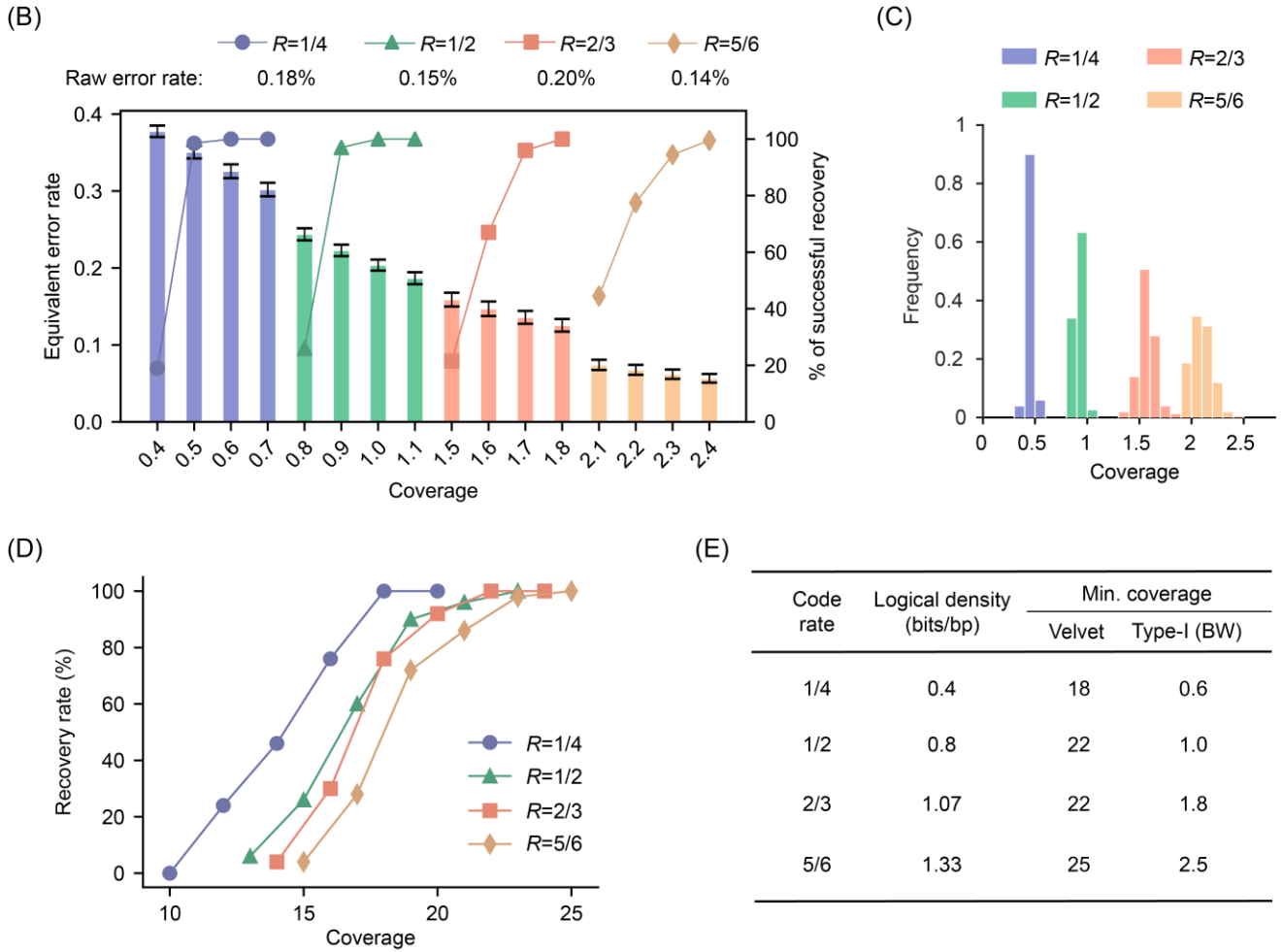
图 3. 基于逐比特共识的软判决数据恢复

(A) 逐比特共识的软判决数据恢复方法流程

结果

三、基于逐比特共识与概率生成的软判决数据恢复

对于四种码率的Illumina测序数据，本方法仅需0.6–2.5×覆盖度即可实现快速数据恢复，显著低于组装方法通常所需的~20×覆盖度。



(E)

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

图 3. 基于逐比特共识的软判决数据恢复

(B) 无基因组干扰条件下的数据恢复性能

(C) 不同码率下实现无错恢复所需的最小测序覆盖度

(D) 基于Velvet组装方法的数据恢复性能

(E) 所提出的数据恢复方案与基于Velvet组装方法的比较

结果

四、逐读段前向-后向算法纠正插入/删除错误

- 相关失败的读段与由第I类读段构建的骨架参考序列进行部分比对，识别包含插入/删除错误的读段；
- 通过前向-后向算法纠正插入/删除错误，生成概率信息进行软判决译码。

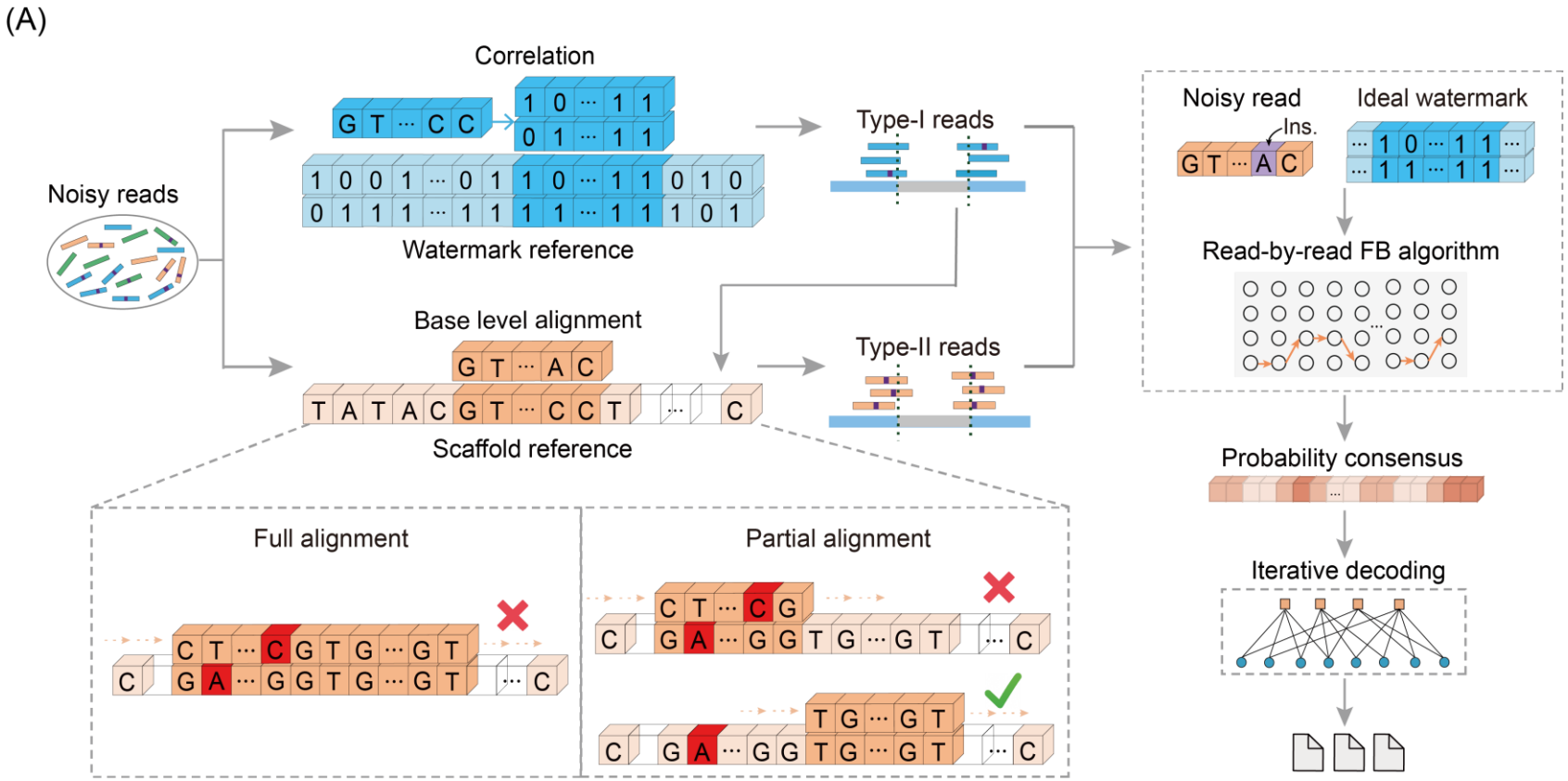


图 4. 前向–后向算法纠正插入/删除错误

(A) 插入/删除错误纠正的工作流程

结果

四、逐读段前向-后向算法纠正插入/删除错误

结合逐读段前向-后向算法和引入第II类读段，可逐步提高恢复性能，在0.8-3.5×的覆盖度范围内实现无错恢复。

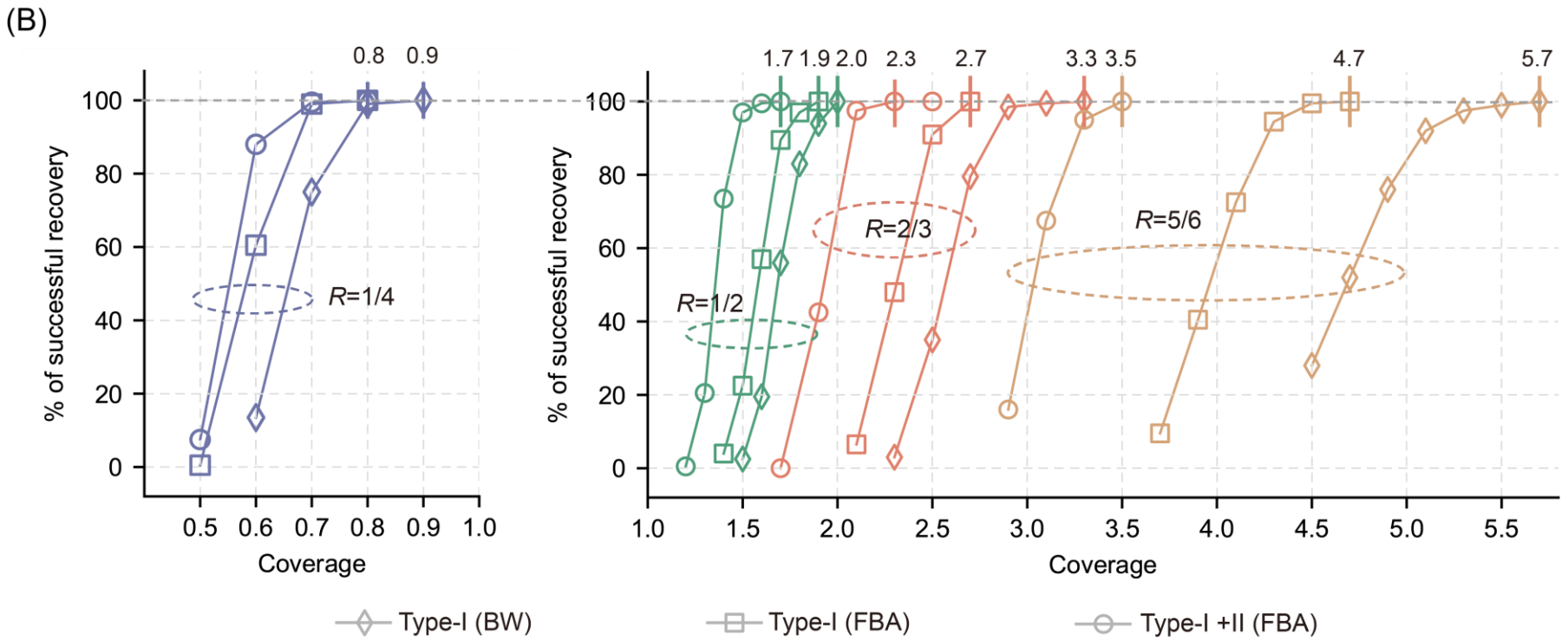


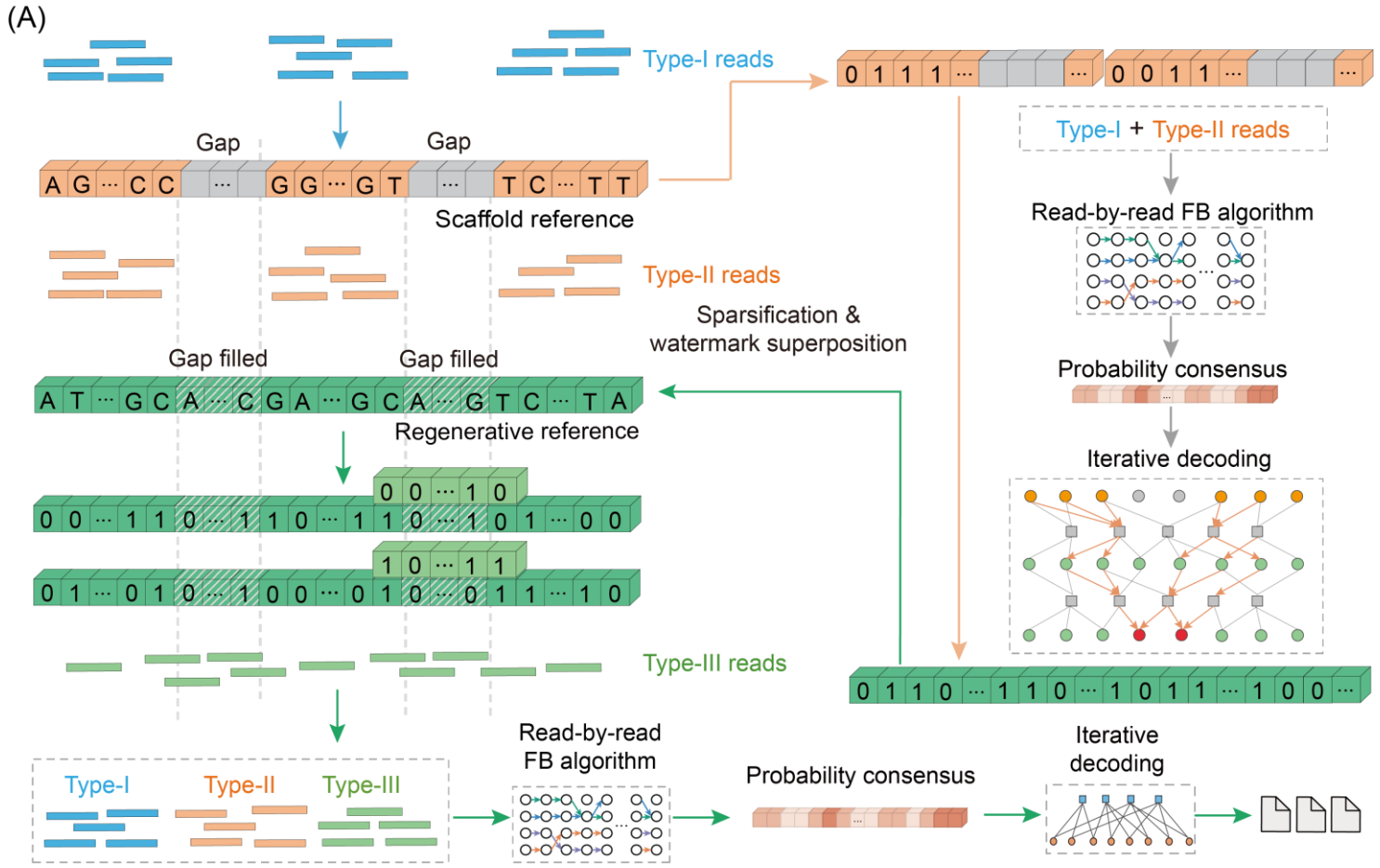
图 4. 前向–后向算法纠正插入/删除错误

(B)不同读段类型和软信息生成方法的数据恢复性能比较

（BW：逐比特概率生成；FBA：基于前向–后向算法的概率生成）

结果

五、再生参考序列比对填补低覆盖度空缺



- 由于骨架参考序列中通常存在空缺，利用译码结果反馈构建再生参考序列；
- 剩余读段与再生参考序列进行比对，识别第III类读段以填补低覆盖度区域。

图 5. 基于再生参考序列的空缺填补方案

(A) 空缺填补方案的工作流程

结果

五、再生参考序列比对填补低覆盖度空缺

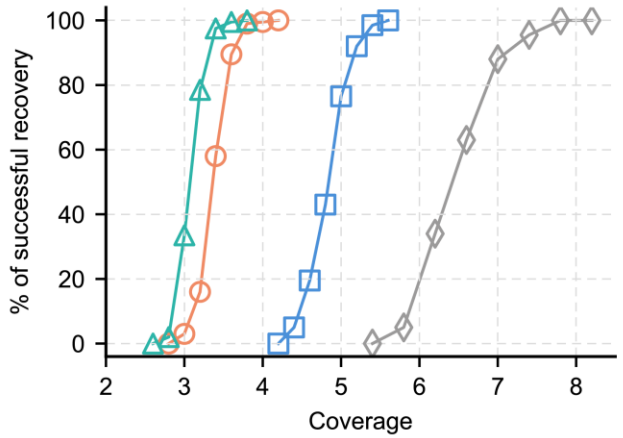
(B)

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)

(C)



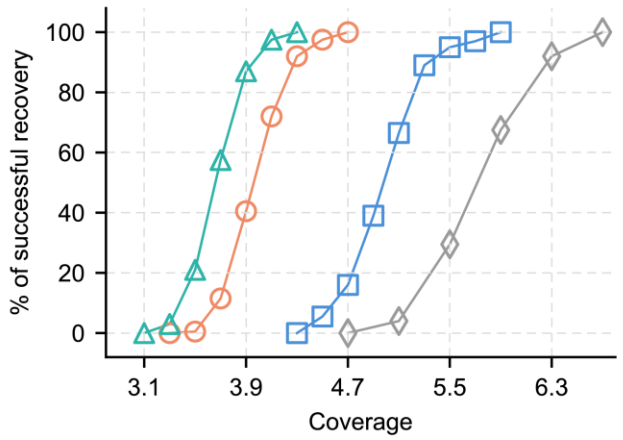
(D)

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)

(E)



Type-I (BW)　Type-I (FBA)　Type-I +II (FBA)　Type-I +II +III (FBA)

- 对于仿真数据（错误率为**1.2%**），在**3.7×**覆盖度下实现无错恢复；
- 对于纳米孔测序数据（错误率为**4.7%**），在**4.3×**覆盖度下实现无错恢复。

图 5. 基于再生参考序列比对的空缺填补方案

(B) 仿真数据集DNA-40.5kb-MC-Sim-2（*R*=5/6）的错误特性

(C) 对于(B)所示数据集，不同测序覆盖度下的数据恢复性能

(D) 纳米孔测序数据集DNA-40.5kb-EM-ONT-1（*R*=2/3）的错误特性

(E) 对于(D)所示数据集，不同测序覆盖度下的数据恢复性能

总结

我们提出了一种多阶段比对与纠错策略，将从头读出转化为类似重测序的工作流程。

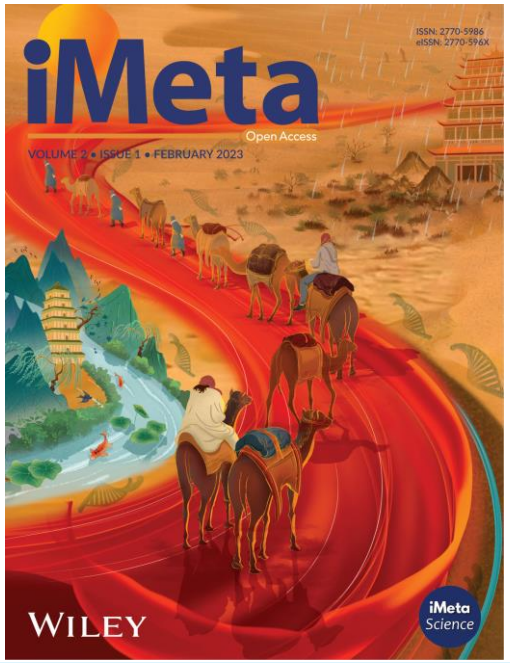
- 与隐藏水印参考序列滑动相关识别低错误率读段，实现快速数据恢复；
- 骨架参考序列结合前向-后向算法挽救含插入/删除错误的读段，提高读段利用率和共识准确性；
- 与再生参考序列比对识别来自于空缺区域的读段，填补低覆盖度区域以降低共识序列擦除率。

本方法通过构建多重隐藏参考序列，实现了大片段DNA数据存储的快速自举式可靠读出，并经实验验证可有效应用于Illumina和纳米孔测序平台的读出恢复。

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(宏): 整合宏组学重新认识生物和医学

iMeta WILEY



iMeta*(宏)**期刊是由宏科学、千名华人科学家和威立共同出版，对标Cell***的生物/医学类综合期刊，主编刘双江和傅静远教授，欢迎高影响力的研究、方法和综述投稿，重点关注生物技术、大数据和组学等前沿交叉学科。已被SCIE、PubMed等收录，最新IF 33.2，位列全球SCI期刊第65位(前千分之三)，中国第5位，微生物学研究类全球第一，中科院生物学双1区Top。外审平均21天，投稿至发表中位数87天。
子刊***iMetaOmics*** (宏组学)、***iMetaMed*** (宏医学)定位IF>10和15的生物、医学综合期刊，欢迎投稿!

主页: http://www.imeta.science
出版社: https://wileyonlinelibrary.com/journal/imeta

投稿: iMeta: https://wiley.atyponrex.com/journal/IMT2
iMetaOmics: https://wiley.atyponrex.com/journal/IMO2
iMetaMed: https://wiley.atyponrex.com/journal/IMM3

office@imeta.science
imetaomics@imeta.science

宣传片

iMeta

更新日期
2025/7/6