Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Error Correction in DNA Data Storage Using a PRISMA Guided Systematic Review and an End to End Retrieval Roadmap

Raj Kumar Thapa

thapaonline@outlook.com

Independent Researcher

Systematic Review

Keywords: DNA data storage, error-correction coding, insertion/deletion errors, PRISMA systematic review, sequencing noise, end-to-end retrieval

Posted Date: March 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8924911/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

DNA data storage has progressed from proof-of-concept demonstrations to early system-level workflows, but robust end-to-end retrieval remains constrained by heterogeneous error processes across synthesis, storage, amplification, and sequencing. This paper presents a PRISMA-guided systematic review of 34 studies and introduces a harmonized analytical layer to strengthen cross-study comparability. We formalize channel error representation using substitution, insertion, and deletion components and apply a CRL (Coding overhead–Retrieval reliability–Latency proxy) framework that compares methods by coding overhead, retrieval reliability, and computational burden proxy. The synthesis shows that substitution-dominant conditions are comparatively tractable with established coding and consensus approaches, whereas insertion/deletion-driven synchronization loss remains the main unresolved barrier, especially in long-read and nanopore-oriented settings. To support deployment decisions, we provide an evidence-to-design workflow that maps error regimes and resource constraints to candidate coding/decoding stacks. We also identify persistent reporting gaps, including inconsistent disclosure of sequencing depth, runtime, and decoding conditions, which limit reproducibility and practical benchmarking. This review contributes a reproducible, implementation-oriented roadmap that moves beyond narrative synthesis toward benchmark-ready method selection for real-world DNA storage systems. [1–34]

1. Introduction

DNA has emerged as a high-density and long-longevity medium for archival information storage. Despite rapid progress in encoding schemes and sequencing technologies, reliable retrieval is still limited by cumulative channel disturbances across wet-lab and computational stages. Error behavior is typically non-uniform: substitution errors can often be reduced by consensus and conventional correction layers, while insertion/deletion events introduce synchronization drift that degrades decoder alignment and final recovery.

Prior review narratives have described major coding families and workflow components, yet cross-study comparability remains weak due to heterogeneous metric definitions, inconsistent reporting granularity, and uneven treatment of computational cost. As a result, method claims are difficult to translate into deployment decisions. A review intended for engineering practice must explicitly connect error regime, correction strategy, and implementation trade-offs.

This study addresses that gap through a PRISMA-guided synthesis of 34 included studies and an added technical comparison layer. The intent is not to introduce a new wet-lab protocol, but to provide a reproducible framework that allows practitioners to compare and select correction pipelines under realistic constraints. [1–34]

Contributions

- A unified cross-study taxonomy linking dominant error regimes to coding/decoding families and sequencing contexts.
- A CRL (Coding overhead-Retrieval reliability-Latency proxy) harmonization framework for quantitative comparison under heterogeneous reporting.
- An evidence-to-design workflow for deployment-oriented method selection and benchmark reporting.

2. Methods

The evidence base summarized in this review comprises the 34 included studies listed in the reference set [1–34].

2.1 Review design and reporting standard

The review followed a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)-guided process including identification, de-duplication, screening, eligibility assessment, and final inclusion. The final corpus contains 34 studies after quality control and removal of retracted dependency. The PRISMA flow diagram is provided as Fig. 1. [35]

2.2 Eligibility and extraction schema

Eligible studies reported DNA storage error-correction methods or empirically grounded evaluation of retrieval robustness. Excluded records were non-DNA-storage contexts, non-technical commentaries, duplicate reports, and items without sufficient methodological detail. For each included study, extraction captured publication year, sequencing context, dominant error type, correction approach, constraints, and reported outcomes. Screening and extraction were managed using Rayyan. [36]

2.3 Analytical model of channel error composition

For each study, channel behavior was represented as e = (p_sub, p_ins, p_del), where p_sub, p_ins, and p_del denote substitution, insertion, and deletion error components, respectively. Synchronization burden was defined as s = p_ins + p_del, reflecting the dominant impact of indels on decoder alignment.

2.4 CRL framework and normalization

Cross-study comparison used three harmonized dimensions: C (coding overhead), R (retrieval reliability), and L (latency proxy). Coding overhead was defined as rho = coded length / payload length when reported. Retrieval reliability used reported decode success indicators (exact or thresholded). Latency proxy used reported runtime, complexity descriptors, or computational-depth indicators when direct runtime was unavailable.

Metrics were normalized to [0, 1] within the included corpus. Composite deployment score S was calculated as: S = w_R * R_hat + w_C * (1 - C_hat) + w_L * (1 - L_hat), with primary weights w_R = 0.5, w_C

= 0.3, w_L = 0.2. Studies with missing values were retained and transparently flagged; sensitivity analysis was used to assess ranking robustness.

Table 1. CRL metric definitions and harmonization rules

2.5 Robustness and sensitivity analysis

To evaluate dependence on weighting assumptions, the composite score was recalculated under three alternative settings: (i) reliability-first, (ii) overhead-constrained, and (iii) compute-constrained priorities. Rank stability across settings was evaluated using Spearman correlation.

3. Results

3.1 Corpus profile

The included 34 studies span diverse sequencing contexts and coding strategies. Reporting completeness varied substantially, particularly for computational runtime and throughput-related metrics. This heterogeneity reinforces the need for explicit harmonization before cross-study comparison.

3.2 Error-regime synthesis

Across the corpus, substitution-dominant conditions were typically associated with stronger decode recovery under conventional ECC and consensus pipelines. In contrast, insertion/deletion-heavy contexts repeatedly showed failure modes linked to synchronization drift and alignment instability. These patterns were consistent across technology generations, with indel management emerging as the key unresolved bottleneck.

3.3 Strategy-performance mapping by regime

Table 2. Regime-based method mapping and deployment implications

3.4 CRL synthesis and sensitivity findings

Under primary and alternative weight configurations, high-reliability methods remained consistently prioritized where synchronization burden was high, while low-overhead methods improved rank primarily in substitution-dominant contexts. Rank-order changes across configurations were moderate rather than disruptive, indicating that major conclusions were robust to reasonable weighting choices.

3.5 Evidence-to-design workflow

Algorithm 1 provides a practical mapping from channel profile and operational constraints to candidate correction stacks.

Algorithm 1: Evidence-to-Design Selection Workflow
Input: (p_sub, p_ins, p_del), platform type, redundancy budget, latency budget
1. Compute synchronization burden s = p_ins + p_del.
2. If s is high, prioritize synchronization-aware or indel-robust strategies.
3. Else if p_sub dominates, prioritize substitution-efficient correction stack.
4. Filter candidates by redundancy and latency constraints.
5. Rank remaining candidates by CRL composite score.
Output: recommended code/decoder profile and trade-off report.

4. Discussion

This review indicates that the field has achieved meaningful progress in substitution-error correction, but deployment-grade performance is still constrained by indel-driven synchronization instability. The main technical implication is that method selection should be regime-aware rather than benchmark-agnostic.

The CRL layer provides a reproducible mechanism to compare methods when studies report non-identical metrics. This shifts the review from descriptive summary toward decision support, enabling practitioners to align correction design with reliability targets and resource constraints.

For implementation practice, three reporting elements should be standardized: (i) explicit channel/error composition assumptions, (ii) transparent redundancy and runtime reporting, and (iii) end-to-end decode criteria under realistic sequencing depth and noise conditions.

5. Limitations and Reproducibility

First, available evidence remains heterogeneous in metric definitions and reporting completeness. Second, runtime and cost are often reported inconsistently, requiring proxy-based harmonization for cross-study comparison. Third, several studies emphasize simulation-heavy settings with limited direct transferability to all laboratory pipelines. To improve reproducibility, this manuscript provides transparent harmonization rules and links all extracted data fields to Supplementary Table S1.

6. Conclusion

A PRISMA-guided synthesis of 34 studies shows that reliable DNA data storage retrieval is primarily limited by insertion/deletion and synchronization challenges, not by substitution handling alone. By adding a harmonized CRL framework, sensitivity analysis, and an evidence-to-design workflow, this review provides an implementation-oriented foundation for method selection and future benchmark standardization. Closing the reproducibility gap will require consistent reporting of channel assumptions, redundancy, runtime burden, and end-to-end recovery criteria.

Declarations

Funding

No specific funding was received for this work.

Competing interests

The author declares no competing interests.

Author contributions

The author was responsible for conceptualization, methodology, screening and data extraction, synthesis, writing (original draft and revisions), and final approval of the manuscript.

Ethics approval

Not applicable. This study is a literature-based systematic review and does not involve human or animal participants.

Clinical trial registration: not applicable.

Consent to Participate declaration: not applicable.

Consent to Publish declaration: not applicable.

Data availability

Not applicable.

AI use disclosure

Language polishing support was used during manuscript preparation. All scientific decisions, screening outcomes, synthesis, and final approval were performed by the author.

References

Note: Study-level extracted fields and the full list of included study titles are provided in Supplementary Table S1.

1. The Essential Component in DNA-Based Information Storage System: Robust Error-Tolerating Module. (2014).
2. Portable and Error-Free DNA-Based Data Storage. (2017).
3. Data storage in DNA with fewer synthesis cycles using composite DNA letters. (2019).
4. Optimized Code Design for Constrained DNA Data Storage With Asymmetric Errors. (2019).
5. Capacity-Approaching Constrained Codes with Error Correction for DNA-Based Data Storage. (2020).

6. HEDGES error-correcting code for DNA storage corrects indels and allows sequence constraints. (2020).
7. Overcoming High Nanopore Basecaller Error Rates for DNA Storage via Basecaller-Decoder Integration and Convolutional Codes. (2020).
8. A Quaternary Code Correcting a Burst of at Most Two Deletion or Insertion Errors in DNA Storage. (2021).
9. Correcting a Single Indel/Edit for DNA-Based Data Storage: Linear-Time Encoders and Order-Optimality. (2021).
10. Limited-Magnitude Error Correction for Probability Vectors in DNA Storage. (2021).
11. An Efficient Bucket-allocation Decoding Method Based on Forward Error Correction Codes for Deoxyribo Nucleic acid Storage. (2022).
12. DeSP: a systematic DNA storage error simulation pipeline. (2022).
13. A Segmented-Edit Error-Correcting Code with Re-Synchronization Function for DNA-Based Storage Systems. (2023).
14. DNA-Aeon provides flexible arithmetic coding for constraint adherence and error correction in DNA storage. (2023).
15. Integrated Error Correction to Enhance Efficiency of Digital Data Storage Based on DNA Nanostructures. (2023).
16. RepairNatrix: a Snakemake workflow for processing DNA sequencing data for DNA storage. (2023).
17. Study of the error correction capability of multiple sequence alignment algorithm (MAFFT) in DNA storage. (2023).
18. Using Soft Information to Improve Error Tolerance of Motif-Based DNA Storage Systems. (2023).
19. Challenges for error-correction coding in DNA data storage: photolithographic synthesis and DNA decay. (2024).
20. Composite Hedges Nanopores codec system for rapid and portable DNA data readout with high INDEL-Correction. (2024).
21. DNA Bloom Filter enables anti-contamination and file version control for DNA-based data storage. (2024).
22. DNA-Correcting Codes: End-to-End Correction in DNA Storage Systems. (2024).
23. Efficient DNA-based data storage using shortmer combinatorial encoding. (2024).
24. Error-Correcting Codes for Nanopore Sequencing. (2024).
25. Overcoming the High Error Rate of Composite DNA Letters-Based Digital Storage through Soft-Decision Decoding. (2024).
26. Protecting the Future of Information: LOCO Coding with Error Detection for DNA Data Storage. (2024).
27. Turbo autoencoders for the DNA data storage channel with Autoturbo-DNA. (2024).

28. Adaptive Arithmetic Coding-Based Encoding Method Toward High-Density DNA Storage. (2025).
29. Codes for Limited-Magnitude Probability Error in DNA Storage. (2025).
30. DNA palette code for time-series archival data storage. (2025).
31. Dna-storalator: a computational simulator for DNA data storage. (2025).
32. INNSE: Invertible neural network-based DNA image storage with self-correction encoding. (2025).
33. Key technologies for DNA storage: encoding, error correction, random access, and security. (2025).
34. LOCO Codes Can Correct as Well: Error-Correction Constrained Coding for DNA Data Storage. (2025).
35. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. https://doi.org/10.1136/bmj.n71
36. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Systematic Reviews. 2016;5:210. https://doi.org/10.1186/s13643-016-0384-4

Figures

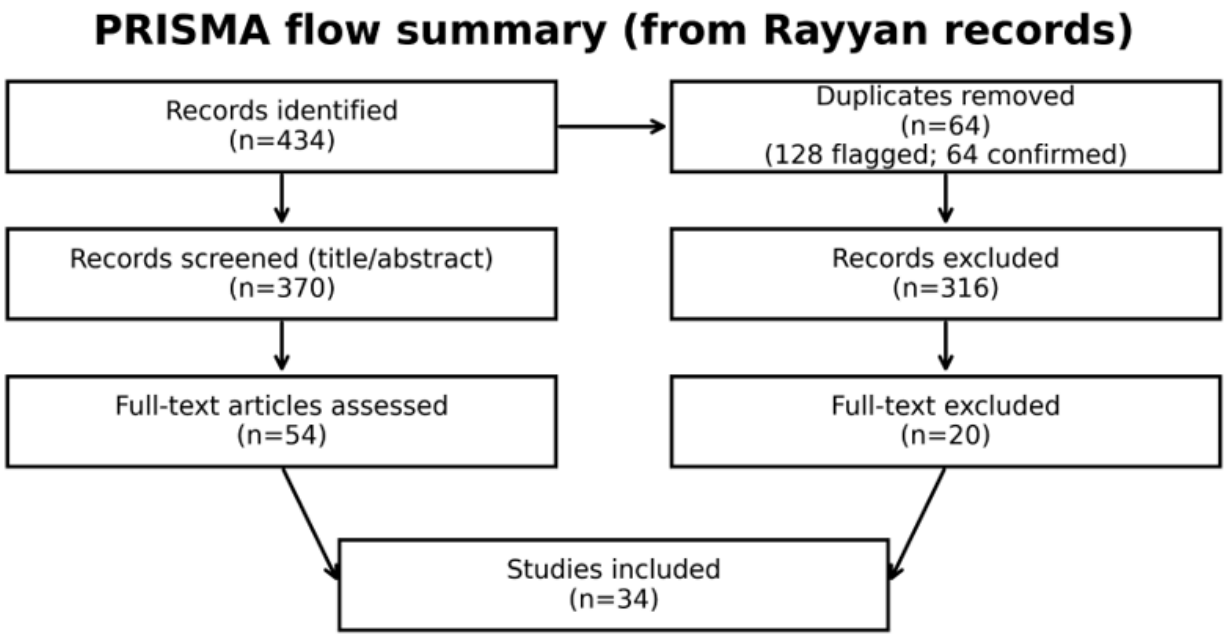


Figure 1

PRISMA flow diagram of study selection and inclusion (final included studies: 34)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTableS1Cleaned.xlsx