

Profile-HMM validation reveals annotation-derived inflation of apparent two-domain NucS occurrence and taxonomically concentrated NucS/MutS-MutL coexistence across 23,435 prokaryotic reference genomes

Kiarash Sadeghian Esfahani

arturestrong@gmail.com

Independent researcher

Research Article

Keywords: DNA mismatch repair, NucS, EndoMS, MutS, MutL, comparative genomics, profile HMM, Pfam, RefSeq

Posted Date: September 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10711921/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Genomics on October 15th, 2026.
See the published version at <https://doi.org/10.1186/s12864-026-13470-8>.

Abstract

Background

EndoMS/NucS mediates a non-canonical mismatch-repair pathway in several archaeal and actinobacterial lineages. A 2017 phylogenomic survey reported that NucS and canonical MutS-MutL were largely mutually exclusive, with coexistence concentrated in Halobacteria and Deinococcus-Thermus. The present study reassessed this distribution in a substantially expanded reference-genome collection while distinguishing annotation labels from multidomain protein-profile evidence.

Results

GFF3 annotations from 23,435 NCBI reference genomes (22,606 Bacteria and 829 Archaea) were screened, exact RefSeq protein accessions for annotation-positive candidates were retrieved, and candidate architectures were evaluated with Pfam profile hidden Markov models. A two-domain NucS call required NucS_N (PF21003) and NucS_C (PF01939) on the same candidate protein. A canonical MutS call required five MutS profiles (PF01624, PF05188, PF05192, PF05190 and PF00488) on the same protein without a same-protein PF01713 (SMR) warning. A MutL call required PF01119 and PF08676; PF13589 was supplementary. Annotation-only screening yielded 1,145 apparent coexistence genomes. Requiring two-domain NucS reduced this set to 473, and the combined NucS/MutS/MutL criteria yielded 470 genomes. Of these, 421 were Halobacteria, 48 were Deinococcota and one was a *Lysobacter korlensis* assembly with a severe contamination warning. The screen recovered two-domain NucS in 5,367 genomes. Explicit *nucS* gene annotations showed 99.63% concordance with the two-domain criterion (5,361/5,381), whereas only 0.73% of genomes classified solely from generic “NucS domain-containing” annotations passed both NucS profiles (5/682).

Conclusions

Annotation-only discovery markedly inflated apparent two-domain NucS occurrence and apparent NucS/MutS-MutL coexistence. Profile validation recovered a taxonomically concentrated coexistence pattern dominated by Halobacteria and Deinococcota, consistent with the qualitative pattern reported in 2017 despite the much larger reference collection. Because annotation-negative genomes were not exhaustively assessed by translated-genome searches, the profile-positive counts represent validated positives recovered by the annotation-first workflow rather than exhaustive prevalence estimates, and non-detection is treated as a screening outcome rather than evidence of biological gene loss.

Background

DNA mismatch repair (MMR) protects genome stability by correcting replication errors before they become fixed mutations. In the canonical bacterial pathway, MutS recognizes mismatched DNA and acts

with MutL to coordinate downstream repair reactions [1].

Many archaeal and actinobacterial lineages lack recognizable canonical MutS-MutL, motivating the search for alternative mutation-avoidance systems [4, 10]. EndoMS, an orthologue of archaeal NucS, was identified as a mismatch-specific endonuclease that cleaves mismatched duplex DNA [2]. Structural work subsequently showed mismatch recognition by an EndoMS dimer and a multidomain architecture with an N-terminal DNA-recognition region and a C-terminal catalytic region [3]. Genetic and biochemical studies in mycobacteria, *Corynebacterium* and *Streptomyces* have since supported roles for NucS/EndoMS in non-canonical MMR and mutation avoidance [4–6, 10].

In a 2017 phylogenomic survey, Castañeda-García et al. screened 3,942 reference proteomes, including 2,709 bacterial and 132 archaeal proteomes [4]. NucS was identified in 370 organisms, and only 28 organisms contained both NucS and MutS-MutL: 21 halobacterial species and seven *Deinococcus-Thermus* species [4]. That study also showed that NucS-associated domains can occur outside the context of the complete protein, providing a strong reason not to treat a single domain annotation as equivalent to two-domain NucS [4].

Public genome resources have expanded substantially since 2017, and NCBI Datasets supports scalable retrieval of genome assemblies, annotations, proteins and metadata [7]. This expansion makes a modern reassessment timely, but it also increases the risk that broad product-name searches will count partial or alternative-domain architectures as complete proteins. Pfam profile HMMs and HMMER provide a family-aware framework for validating multidomain protein architecture using curated profile models and family-specific gathering thresholds [8, 9].

The present study reassessed NucS and canonical MutS-MutL across 23,435 NCBI bacterial and archaeal reference genomes. Annotation calls were used for high-sensitivity candidate discovery, after which exact candidate proteins were retrieved and evaluated with profile HMMs. Three questions were addressed: whether annotation wording predicts two-domain NucS architecture, how profile validation changes apparent coexistence counts, and whether the taxonomic concentration reported in 2017 remains evident in a substantially larger reference collection. Annotation or profile non-detection was not interpreted as definitive biological absence.

Methods

Genome collection and checksum validation

The analysis used the NCBI prokaryotic reference-genome collection retrieved on 14 August 2026 with NCBI Datasets CLI v18.35.0 [7]. Reference-genome accessions were enumerated separately for Bacteria and Archaea using the NCBI Datasets filters `--reference`, `--annotated` and `--mag exclude`. The resulting collection contained 23,435 assemblies: 22,606 Bacteria and 829 Archaea. GFF3 annotations were downloaded in resumable batches and validated against the MD5 checksums supplied in each NCBI data package. Six initially incomplete GFF3 files were re-retrieved and successfully checksum-validated

before the annotation screen. Taxonomic metadata were obtained with NCBI Datasets/Dataformat and joined to assembly accessions by NCBI Taxonomy identifiers [13].

Annotation-level discovery screen

Each GFF3 file was parsed for NucS/EndoMS, canonical MutS and MutL annotations. Products explicitly annotated as MutS2, MutS3 or other specialized MutS paralogues were excluded from the canonical MutS discovery rule. Five annotation-level states were recorded: BOTH (NucS plus canonical MutS and MutL), NUCS_ONLY, MUTSL_ONLY, NEITHER and PARTIAL_MUTSL. These labels were used solely as high-sensitivity discovery categories. The exact GFF3 parsing and annotation-matching rules are provided in the publicly available analysis-code repository:

https://github.com/Doran79/NucS_MutSL_Analysis.git.

NucS annotation calls were stratified into explicit gene-level *nucS/endoMS* annotations, explicit product-level NucS/EndoMS annotations and generic NucS domain-containing annotations. This distinction was defined before sequence validation because NucS-associated domains can occur in proteins outside the complete two-domain architecture [4].

Exact protein retrieval and retrieval-integrity checks

All RefSeq protein accessions present in the candidate annotations were downloaded through NCBI Datasets gene packages [7, 12]. NucS retrieval initially yielded 5,976 of 6,270 expected accessions; the 294 missing accessions were recovered by small exact-accession retries and NCBI EFetch [17] fallback. The final accession-bearing NucS candidate FASTA contained sequences for all 6,270 expected accessions. All 17,742 MutS and 17,608 MutL candidate accessions were likewise retrieved, with zero missing and zero unexpected accessions. Annotation-positive loci without a protein accession were retained as unresolved and were not counted as retrieval failures. Profile classification was performed only after these accession-set checks passed.

Taxonomic summaries and descriptive analysis

Taxonomic distributions were summarized from NCBI Taxonomy metadata [13] joined by assembly accession. Percentages are descriptive of the August 2026 NCBI reference collection and are not unbiased estimates of organismal prevalence. Comparisons with the 2017 study emphasize qualitative taxonomic pattern rather than prevalence because sampling units, database composition and annotation environments differ. The term *Deinococcus-Thermus* is retained only when describing the taxonomy used in the 2017 study; contemporary NCBI assignments are reported as *Deinococcota*, class *Deinococci*.

Supplementary quality-control documentation

Additional file 1 records accession-retrieval integrity, the final MutS/SMR classifier correction, unresolved-locus bookkeeping, final profile-state counts, the exceptional assembly quality-control case and analyses deliberately outside the scope of this study. The shell and Python analysis code, including the corrected final classifier and supporting workflow scripts, is publicly available at https://github.com/Doran79/NucS_MutSL_Analysis.git.

Use of large language model assistance

In accordance with BMC Genomics guidance that use of a large language model (LLM) should be documented in the Methods section [16], the following use is disclosed. OpenAI ChatGPT using GPT-5.6 Sol [15] was used as an assistive tool for computational-workflow planning and refinement of operational rules; generation, revision and debugging of shell and Python code; review of intermediate computational outputs and quality-control logic; scripting and presentation checks for data-derived figures; manuscript organization and drafting support across the Background, Methods, Results, Discussion and Conclusions; language and grammar editing; and literature-search support and bibliographic cross-checking. Final numerical results and profile classifications were verified against retained computational outputs, and cited references were checked against the source literature. Responsibility for study design, scientific judgments, interpretation, conclusions and the final manuscript remained with the author. ChatGPT was not treated as an author or as scientific evidence for biological claims; reference [15] identifies the AI tool/model used. No generative-AI imagery was used; all figures are data-derived plots.

Results

Annotation-only screening identified 1,145 apparent coexistence genomes

The initial GFF3 screen classified 16,388 genomes as MUTSL_ONLY, 4,839 as NUCS_ONLY, 1,145 as BOTH, 898 as NEITHER and 165 as PARTIAL_MUTSL (Table 1). Thus, annotation labels alone suggested apparent NucS plus canonical MutS-MutL coexistence in 4.89% of the 23,435 reference genomes. These labels were treated as discovery states rather than final biological classifications. The progressive effect of profile validation on the apparent coexistence set is summarized in Fig. 1.

Table 1
Annotation-screen states across 23,435 reference genomes

Annotation-screen state	Genomes	Percent of 23,435	Discovery-stage meaning
MUTSL_ONLY	16,388	69.93%	MutS and MutL annotated; NucS not annotated
NUCS_ONLY	4,839	20.65%	NucS annotated; MutS and MutL not both annotated
BOTH	1,145	4.89%	NucS, canonical MutS and MutL annotated
NEITHER	898	3.83%	None annotated; biological absence not inferred
PARTIAL_MUTSL	165	0.70%	Mixed/partial annotation state

Two-domain NucS profile support was recovered in 5,367 genomes

The NucS candidate set comprised 6,270 accession-bearing RefSeq proteins. After recovery of an initially incomplete download, all 6,270 expected accessions were represented, with zero missing and zero unexpected accessions. Across the 23,435 genomes, 5,367 (22.90%) contained at least one annotation-derived candidate protein that passed both NucS_N and NucS_C profiles. Among the remaining NucS-annotation-positive genomes, 680 contained tested candidate proteins but no two-domain pass, whereas 22 contained at least one annotation-positive locus without a retrievable protein accession and were retained as unresolved. The remaining 17,366 genomes were annotation-negative for NucS and remained absence-unvalidated (Table 2). Accordingly, 5,367 represents the number of validated positives recovered by annotation-first candidate discovery rather than an exhaustive estimate of genomic prevalence.

Table 2
NucS evidence status after two-domain profile validation

NucS evidence status	Genomes	Percent
Two-domain NucS profile-positive	5,367	22.90%
Candidate protein tested; no two-domain pass	680	2.90%
Annotation-positive unresolved NucS locus	22	0.09%
Annotation-negative; biological absence untested	17,366	74.10%

Annotation wording showed marked differences in two-domain NucS support

Two-domain profile support differed markedly among NucS annotation tiers. Among 5,381 genomes in the explicit *nucS* gene tier, 5,361 (99.63%) contained a two-domain NucS protein. By contrast, only 5 of 682 genomes (0.73%) classified through a generic “NucS domain-containing” annotation passed both NucS profiles. The small explicit-product-only tier contained six genomes, of which one (16.67%) passed. These descriptive differences identify generic domain-only annotation as a major source of annotation-level inflation without requiring an inferential prevalence comparison (Fig. 2).

MutS and MutL profile validation

Retrieval was complete for all accession-bearing canonical MutS and MutL candidates: 17,742/17,742 unique MutS accessions and 17,608/17,608 unique MutL accessions were recovered, with zero missing or unexpected accessions. The corrected final MutS classifier assigned 17,023 genomes to the strong canonical MutS category, defined by a five-domain canonical MutS protein without a same-protein SMR warning. A further 293 genomes contained a four-of-five partial MutS architecture, 74 had weaker zero-to-three-of-five profile support, 200 were annotation-positive but lacked a protein accession in the final mutually exclusive genome classification, and 5,845 were annotation-negative and therefore absence-unvalidated. The SMR domain was treated as a MutS2-like warning because it is characteristic of the C-terminal region of bacterial MutS2 proteins [11].

For MutL, 17,077 genomes contained the principal PF01119/PF08676 pair together with supplementary PF13589 support, and 59 contained the principal pair without PF13589, yielding 17,136 strong MutL-positive genomes. Three hundred forty-two annotation-positive genomes had weak or failing MutL profile support, 163 lacked a protein accession in the final mutually exclusive genome classification, and 5,794 were annotation-negative and absence-unvalidated. The three profile criteria and their strong-positive genome counts are summarized in Table 3.

Table 3
Profile-supported NucS, canonical MutS and MutL calls

Validated signal	Strong-positive genomes	Percent of 23,435	Operational criterion
Two-domain NucS	5,367	22.90%	PF21003 + PF01939 on at least one candidate protein
Canonical MutS	17,023	72.64%	Five canonical MutS profiles on one protein; no same-protein PF01713 (SMR) warning
MutL	17,136	73.12%	PF01119 + PF08676 on at least one candidate protein; PF13589 supplementary

Combined profile criteria identified 470 coexistence genomes

Combining the three profile criteria yielded 16,166 genomes with strong MutS-MutL but no two-domain NucS profile, 4,895 with two-domain NucS but without simultaneous strong MutS and MutL support, 887

mixed-profile genomes, 1,017 genomes meeting none of the three strong-positive criteria, and 470 genomes meeting all three criteria (Table 4; Fig. 3). The 470 coexistence genomes represented 2.01% of the reference collection. This percentage is a descriptive screening fraction and is not interpretable as species-level prevalence because the collection is not a balanced taxonomic sample.

Table 4
Final combined profile-state distribution across 23,435 reference genomes

Final profile state	Genomes	Percent	Directly supported evidence
MutS-MutL strong; NucS not profile-positive	16,166	68.98%	Strong MutS and MutL profiles
NucS profile-positive; MutS-MutL not both strong	4,895	20.89%	Two-domain NucS profile
Mixed profile state	887	3.78%	A subset of strong-positive components
No strong profile-positive system	1,017	4.34%	No strong positive criterion; absence not proven
All three criteria met	470	2.01%	Two-domain NucS + canonical MutS + MutL criteria

Coexistence was concentrated in Halobacteria and Deinococcota

Of the 470 genomes meeting all three profile criteria, 421 (89.57%) were archaeal Halobacteria and 48 (10.21%) were bacterial Deinococcota, class Deinococci. Thus, 469/470 (99.79%) belonged to the two groups historically associated with NucS/MutS-MutL coexistence [4]. The sole other case was a Pseudomonadota assembly assigned to *Lysobacter korlensis* (GCF_042434515.1). During targeted quality control, the NCBI assembly report for this genome contained a CheckM contamination estimate of 76.26% [7, 14]. The genome was therefore retained as an unresolved assembly-level candidate rather than as evidence of established cross-phyla coexistence (Fig. 4).

The 2017 study reported 21 halobacterial species and seven Deinococcus-Thermus species with coexistence [4]. Absolute counts are not directly comparable because the unit of analysis in the present study is the reference assembly and the database composition has changed. Nevertheless, the present collection contains 23,435 prokaryotic reference assemblies compared with 2,841 bacterial plus archaeal reference proteomes in the earlier screen, an approximately 8.25-fold difference in the number of prokaryotic entries examined.

Discussion

The expanded reference collection preserves the previously reported coexistence pattern

The principal result is the persistence of a taxonomically concentrated coexistence pattern after multidomain protein-profile validation. Annotation-only screening suggested 1,145 coexistence genomes, whereas 470 met the defined profile criteria for two-domain NucS, canonical MutS and MutL. Of these, 469 belonged to Halobacteria or Deinococcota. This pattern closely recapitulates the qualitative distribution reported by Castañeda-García et al., who found coexistence in Halobacteria and a small set of Deinococcus-Thermus species [4], despite the much larger contemporary reference collection.

This result does not constitute an evolutionary reconstruction. Ancestral states, lineage-specific losses and new horizontal-transfer events were not inferred. Instead, the analysis demonstrates that the previously reported taxonomic concentration of profile-supported coexistence remains evident when a much larger reference-genome collection is screened with explicit multidomain criteria.

Generic domain-containing annotations can inflate apparent two-domain NucS counts

The largest methodological effect involved annotation wording. Nearly every explicit *nucS* gene call was supported by both NucS_N and NucS_C, whereas fewer than 1% of generic NucS domain-containing genome calls passed both profiles. This pattern is consistent with the multidomain organization of NucS established by structural and phylogenomic studies [3, 4] and with the general principle that Pfam domains can occur in multiple protein architectures [8]. A domain-containing product label is therefore unsuitable as a stand-alone proxy for two-domain NucS.

This distinction is directly relevant to large-scale comparative genomics. High-sensitivity annotation searches are useful for candidate discovery, but unusual taxonomic distributions can be dominated by partial-domain or alternative-context annotations unless candidates are subsequently validated at the protein-architecture level. The reduction from 1,145 annotation-level coexistence candidates to 473 after NucS validation is concordant with the very low two-domain validation rate of generic domain-only NucS annotations.

The study provides stronger evidence for presence than for absence

The workflow was intentionally asymmetric: positive annotation calls were subjected to exact protein retrieval and profile validation, whereas annotation-negative genomes were not exhaustively searched at the translated-genome level. This distinction matters because the 2017 study used tblastn searches to confirm selected apparent absence states [4]. An equivalent genome-wide absence-validation step was not applied across all 23,435 assemblies in the present survey.

Consequently, states such as “NucS profile-positive; MutS-MutL not both strong” cannot be interpreted as proof that MutS or MutL is biologically absent. Non-detection can result from annotation failure, assembly fragmentation, pseudogenization, incomplete gene models or sequence divergence below the selected profile thresholds. The same limitation applies to the 1,017 genomes with no strong profile-positive system under this screen. These categories are therefore retained as screening states rather than gene-loss calls. A translated-genome or equivalent sequence-level search would be required to distinguish true absence from annotation failure or highly divergent systems. Profile-positive counts should therefore be interpreted as validated positives recovered from annotation-derived candidates and may underestimate true genomic presence when relevant genes are unannotated or highly divergent.

The exceptional non-Deinococcota bacterial case requires assembly-level caution

The only non-Deinococcota bacterial genome in the final coexistence set was assigned to *Lysobacter korlensis*. Its NucS candidate passed both NucS profiles, and its MutS and MutL candidates passed their respective criteria, but the assembly carried a very high CheckM contamination estimate [14] in the NCBI quality metadata [7]. The assembly therefore does not support a robust claim of cross-phyla coexistence without independent confirmation of scaffold provenance. Protein-family validation and assembly-quality validation address complementary sources of error in large-scale genome surveys.

Biological context and implications

Experimental studies establish that EndoMS/NucS can act as a mismatch-specific endonuclease in Archaea and Actinobacteria [2, 4–6, 10]. In *Corynebacterium glutamicum*, bacterial EndoMS/NucS acts with the sliding clamp, and its disruption increases replication-associated mutagenesis [5]. Work in *Streptomyces* has likewise shown increased spontaneous mutation in *nucS*-deficient backgrounds and a role for NucS in shaping chromosomal mutational bias [6]. These observations support the biological relevance of the profile-supported NucS family examined here, while protein presence alone does not establish identical substrate preference or pathway activity in every lineage.

Conclusions

Across 23,435 NCBI bacterial and archaeal reference genomes, annotation-only screening suggested 1,145 apparent NucS/MutS-MutL coexistence cases. Exact candidate-protein retrieval and multidomain profile-HMM validation reduced this to 470 genomes meeting the defined criteria for two-domain NucS, canonical MutS and MutL. The final coexistence set was overwhelmingly composed of Halobacteria and Deinococcota, preserving the dominant taxonomic pattern reported in 2017. The analysis also identified a major annotation pitfall: explicit *nucS* gene annotations were highly concordant with two-domain NucS support, whereas generic NucS domain-containing annotations rarely represented the complete two-domain architecture. These findings support profile-based validation in large-scale surveys of multidomain DNA-repair proteins. Because annotation-negative states were not exhaustively validated at

genome level, the reported positive counts represent profile-supported positives recovered by an annotation-first workflow and should not be interpreted as exhaustive prevalence estimates or as an evolutionary reconstruction of gene gain, loss or horizontal transfer.

Abbreviations

GFF3

General Feature Format version 3

HMM

hidden Markov model

LLM

large language model

MMR

mismatch repair

NCBI

National Center for Biotechnology Information

SMR

small MutS-related

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The author declares that there are no competing interests.

Funding

No funding was received for this research.

Author Contribution

KSE conceived and designed the study, developed the methodology, performed the computational analyses and quality-control validation, curated and interpreted the data, prepared the figures, and

drafted and revised the manuscript. KSE approved the final version of the manuscript and accepts responsibility for the integrity of the work.

Acknowledgements

Not applicable.

Data Availability

All source genomes and RefSeq annotations analyzed in this study are publicly available from NCBI and are identified by assembly accession in the final row-level analysis table. The analysis scripts used for genome acquisition, annotation screening, protein retrieval, profile-HMM validation, quality control and final profile classification are publicly available in the GitHub repository

https://github.com/Doran79/NucS_MutSL_Analysis.git. The final genome-state matrix, candidate accession lists and supporting quality-control files are deposited in Zenodo under DOI <https://doi.org/10.5281/zenodo.21950397>. Additional file 1 provides supplementary methods and quality-control documentation.

References

1. Acharya S, Foster PL, Brooks P, Fishel R. The coordinated functions of the *E. coli* MutS and MutL proteins in mismatch repair. *Mol Cell*. 2003;12:233–46. 10.1016/S1097-2765(03)00219-3.
2. Ishino S, Nishi Y, Oda S, Uemori T, Sagara T, Takatsu N, et al. Identification of a mismatch-specific endonuclease in hyperthermophilic Archaea. *Nucleic Acids Res*. 2016;44:2977–86. 10.1093/nar/gkw153.
3. Nakae S, Hijikata A, Tsuji T, Yonezawa K, Kouyama KI, Mayanagi K, et al. Structure of the EndoMS-DNA complex as mismatch restriction endonuclease. *Structure*. 2016;24:1960–71. 10.1016/j.str.2016.09.005.
4. Castañeda-García A, Prieto AI, Rodríguez-Beltrán J, Alonso N, Cantillon D, Costas C, et al. A non-canonical mismatch repair pathway in prokaryotes. *Nat Commun*. 2017;8:14246. 10.1038/ncomms14246.
5. Takemoto N, Numata I, Su'etsugu M, Miyoshi-Akiyama T. Bacterial EndoMS/NucS acts as a clamp-mediated mismatch endonuclease to prevent asymmetric accumulation of replication errors. *Nucleic Acids Res*. 2018;46:6152–65. 10.1093/nar/gky481.
6. Dagva O, Thibessard A, Lorenzi JN, Labat V, Piotrowski E, Rouhier N, et al. Correction of non-random mutational biases along a linear bacterial chromosome by the mismatch repair endonuclease NucS. *Nucleic Acids Res*. 2024;52:5033–47. 10.1093/nar/gkae132.
7. O'Leary NA, Cox E, Holmes JB, Anderson WR, Falk R, Hem V, et al. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Sci Data*. 2024;11:732.

- 10.1038/s41597-024-03571-y.
8. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. *Nucleic Acids Res.* 2021;49:D412–9. 10.1093/nar/gkaa913.
 9. Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol.* 2011;7:e1002195. 10.1371/journal.pcbi.1002195.
 10. Cebrián-Sastre E, Martín-Blecua I, Gullón S, Blázquez J, Castañeda-García A. Control of genome stability by EndoMS/NucS-mediated non-canonical mismatch repair. *Cells.* 2021;10:1314. 10.3390/cells10061314.
 11. Fukui K, Kosaka H, Kuramitsu S, Masui R. Nuclease activity of the MutS homologue MutS2 from *Thermus thermophilus* is confined to the Smr domain. *Nucleic Acids Res.* 2007;35:850–60. 10.1093/nar/gkl735.
 12. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 2016;44:D733–45. 10.1093/nar/gkv1189.
 13. Schoch CL, Ciufo S, Domrachev M, Hottel CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford).* 2020;2020:baaa062. 10.1093/database/baaa062.
 14. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 2015;25:1043–55. 10.1101/gr.186072.114.
 15. OpenAI. GPT-5.6: frontier intelligence that scales with your ambition. OpenAI. 2026. <https://openai.com/index/gpt-5-6/>. Accessed 15 Aug 2026.
 16. BMC Genomics. Research article submission guidelines. Springer Nature. <https://link.springer.com/journal/12864/submission-guidelines/research-article>. Accessed 15 Aug 2026.
 17. Sayers E, A General Introduction to the E-utilities. 2009 May 26 [Updated 2022 Nov 17]. In: Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-. <https://www.ncbi.nlm.nih.gov/books/NBK25497/>. Accessed 15 Aug 2026.

Figures

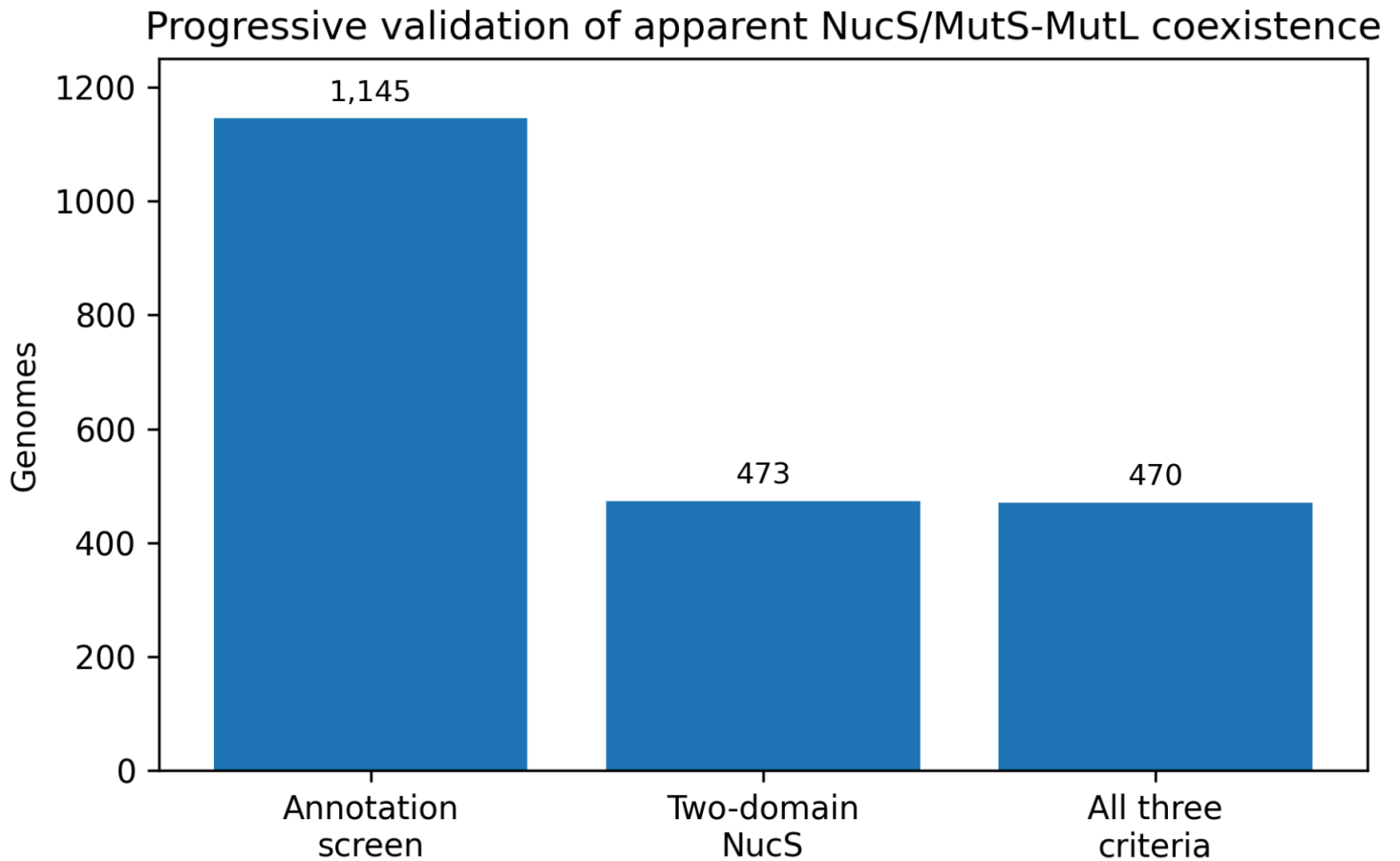


Figure 1

Progressive validation of apparent NucS/MutS-MutL coexistence. Annotation-only screening identified 1,145 BOTH genomes. Requiring two-domain NucS reduced this set to 473, and requiring the defined NucS, canonical MutS and MutL profile criteria yielded 470 genomes.

Two-domain NucS support by annotation tier

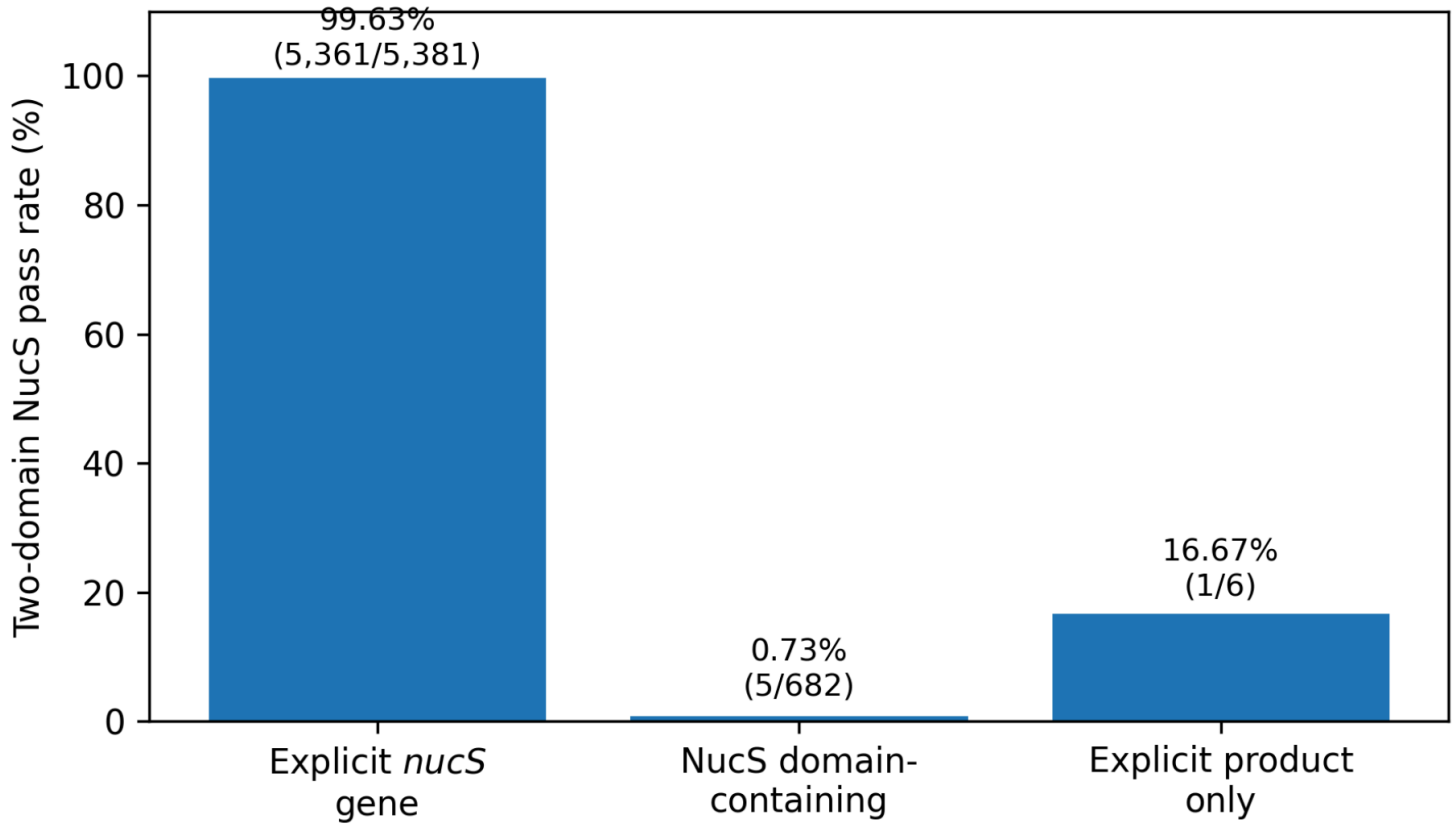


Figure 2

Genome-level two-domain NucS support by annotation tier. Explicit *nucS* gene annotations were highly concordant with NucS_N plus NucS_C support, whereas generic NucS domain-containing annotations rarely passed both profiles.

Combined profile-state distribution across 23,435 genomes

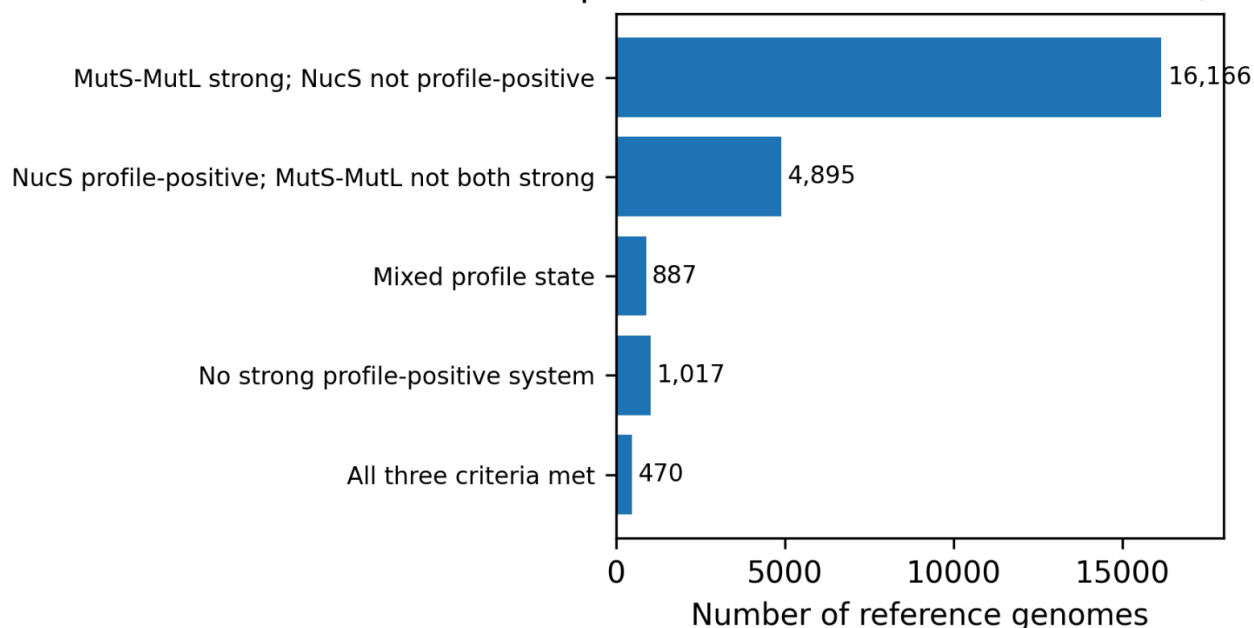


Figure 3

Profile-state distribution across all 23,435 reference genomes. “Not profile-positive” denotes failure to meet the defined strong-positive criterion and should not be interpreted as proof of biological absence.

Taxonomic composition of the 470 coexistence genomes

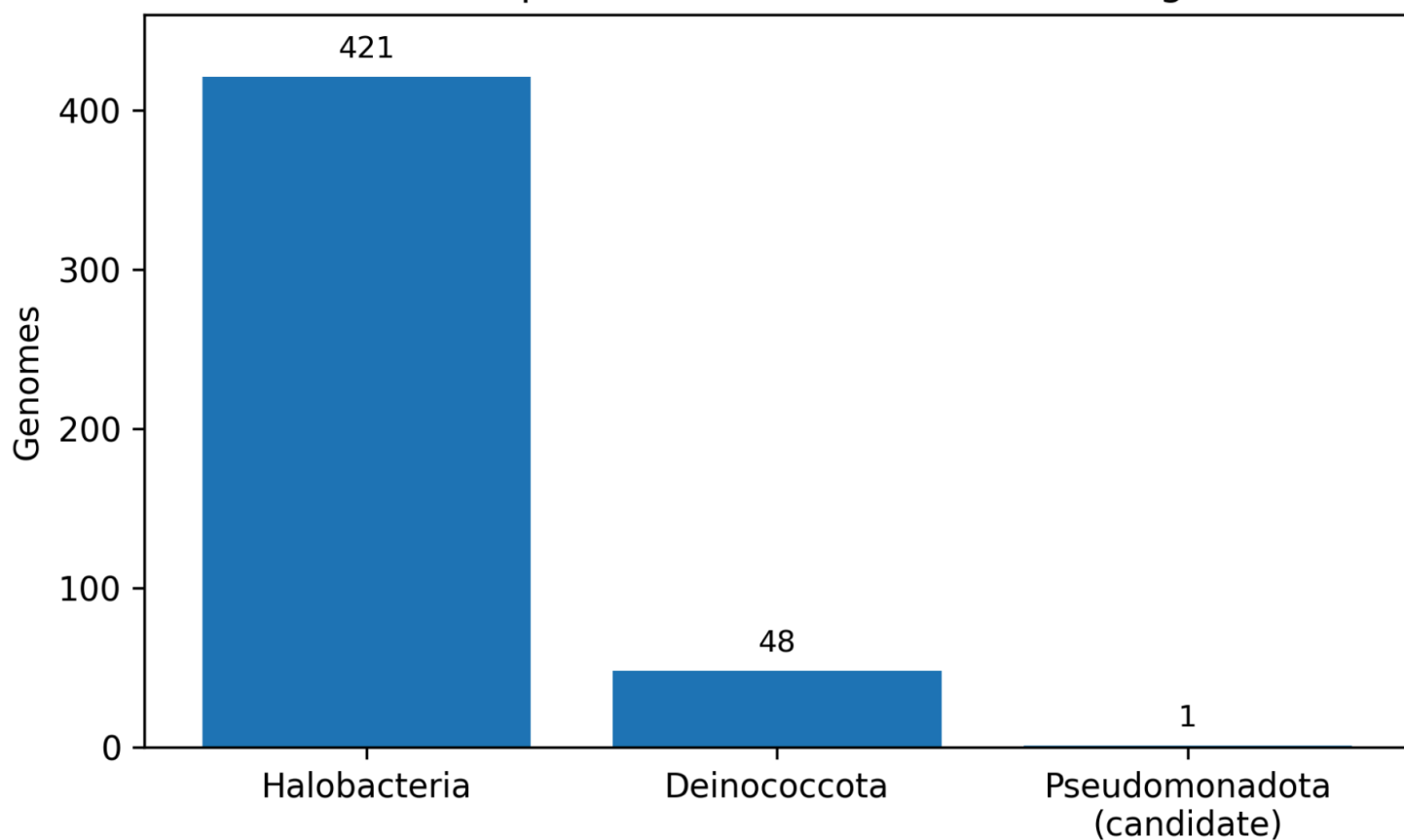


Figure 4

Taxonomic composition of the 470 genomes meeting all three profile criteria. The single *Pseudomonadota* case is retained as an unresolved candidate because its NCBI assembly-quality report contained a severe contamination warning.

Supplementary Files

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

- [Additionalfile1.docx](#)