

From fragmented ITS references to an auditable, phylogeny-validated taxonomic backbone for plant metabarcoding studies

Why reference databases need curation

Reference quality sets the ceiling for metabarcoding accuracy

Different taxonomies
Sequences and taxonomies distributed across databases with different contents and standards [1].

[1] Patterson et al. Challenges with using names to link digital biodiversity information; Biodiversity Data Journal; DOI: 10.3897/BDJ.4.e8080

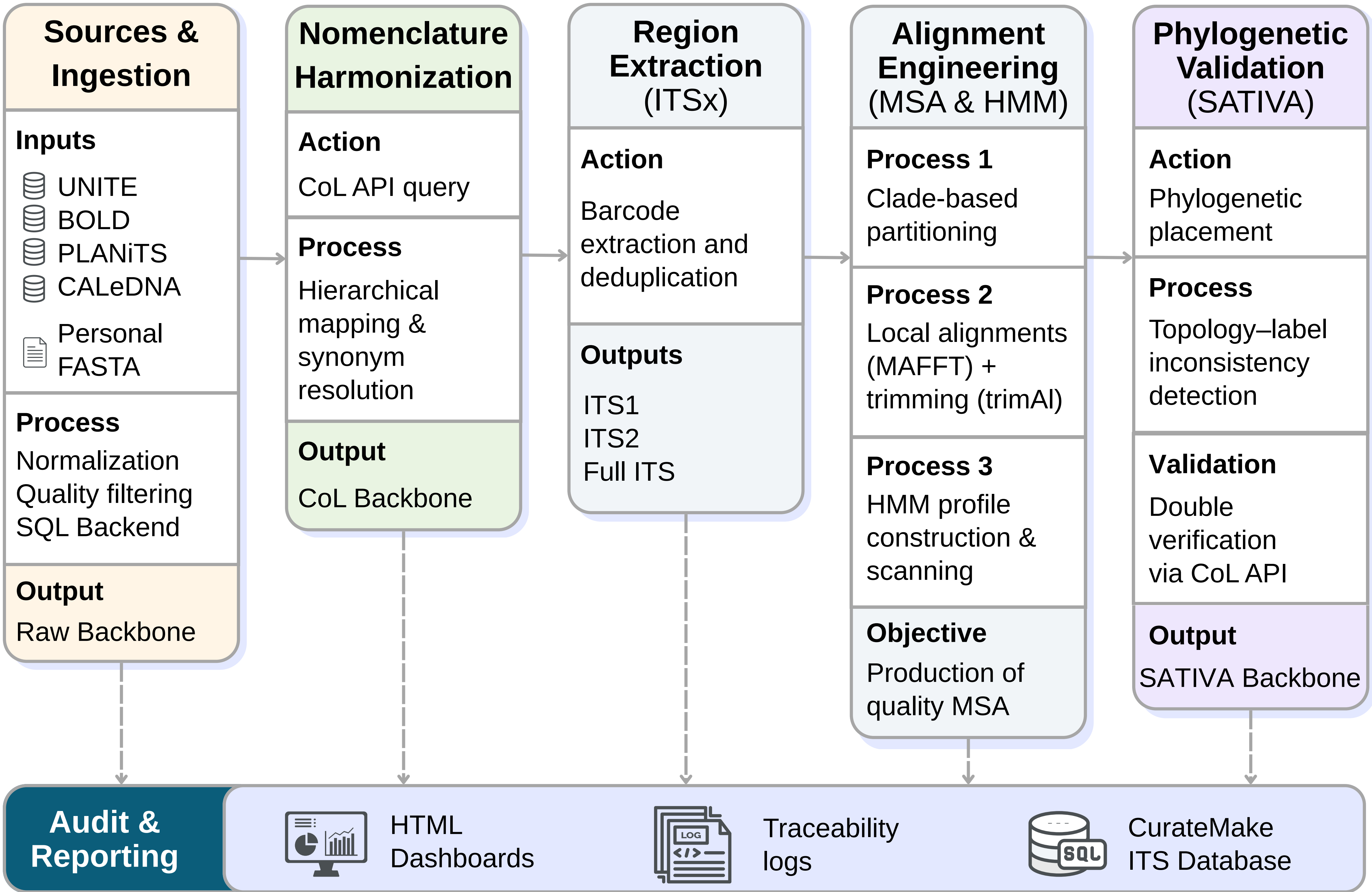
Inconsistent annotations
Nomenclatural conflicts, mislabels and contradictions propagate into downstream assignments [2].

[2] Cheng et al. The devil is in the details: problems in DNA barcoding practices indicated by systematic evaluation of insect barcode; Frontiers in Ecology and Evolution; DOI: 10.3389/fevo.2023.1149839

Incomplete coverage
Missing ranks and unbalanced representation limit reliable assignment [3].

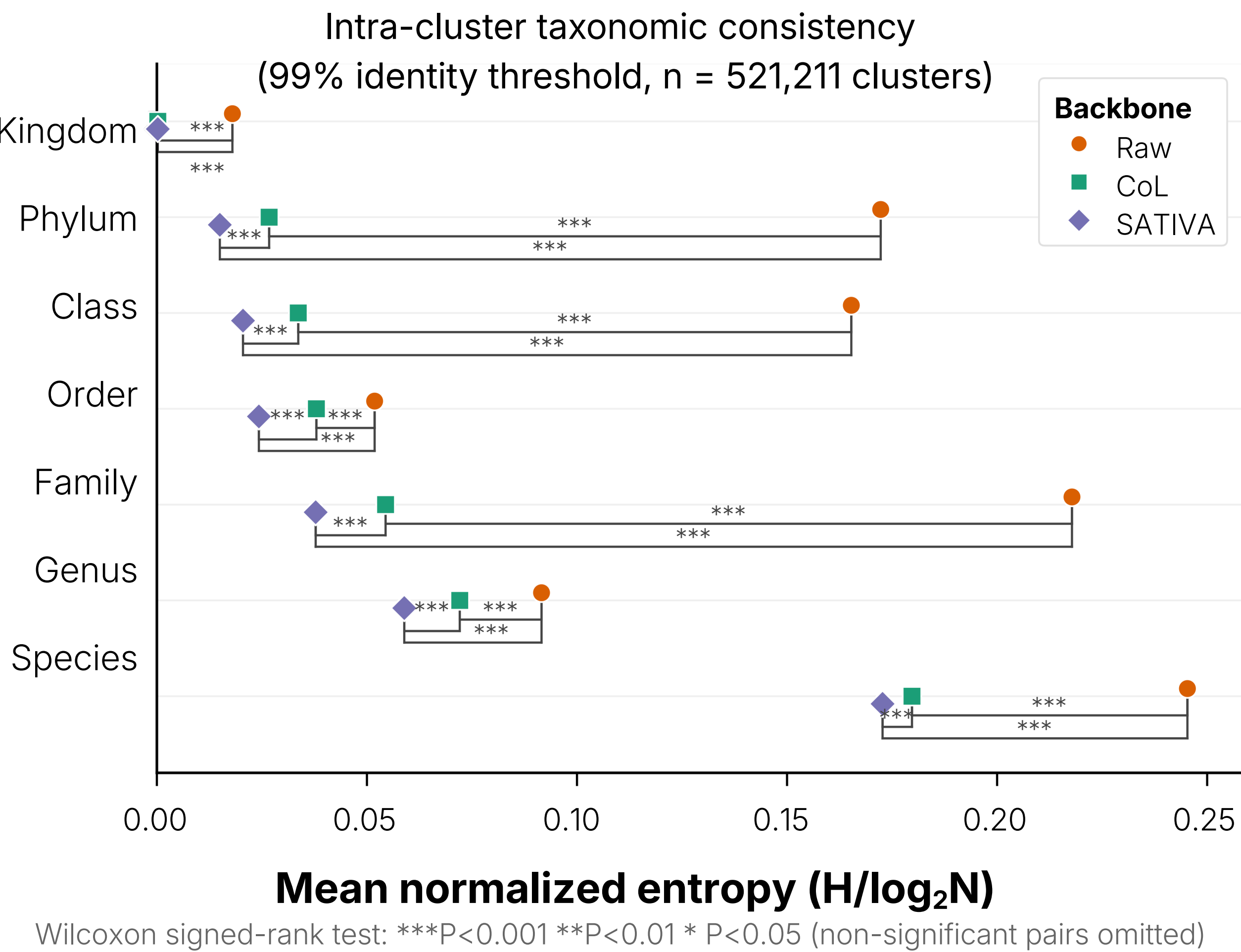
[3] Keck et al. Navigating the seven challenges of taxonomic reference databases in metabarcoding analyses; Molecular Ecology Resources; DOI: 10.1111/1755-0998.13746

CurateMake Pipeline Overview



Benchmark 1 - Entropy drops after curation

Similar sequences become taxonomically more consistent.



Interpretation

Mean entropy **decreases monotonically** from Raw → CoL → SATIVA across all taxonomic ranks. CoL removes nomenclatural noise; SATIVA reduces residual inconsistencies where phylogenetic signal is informative

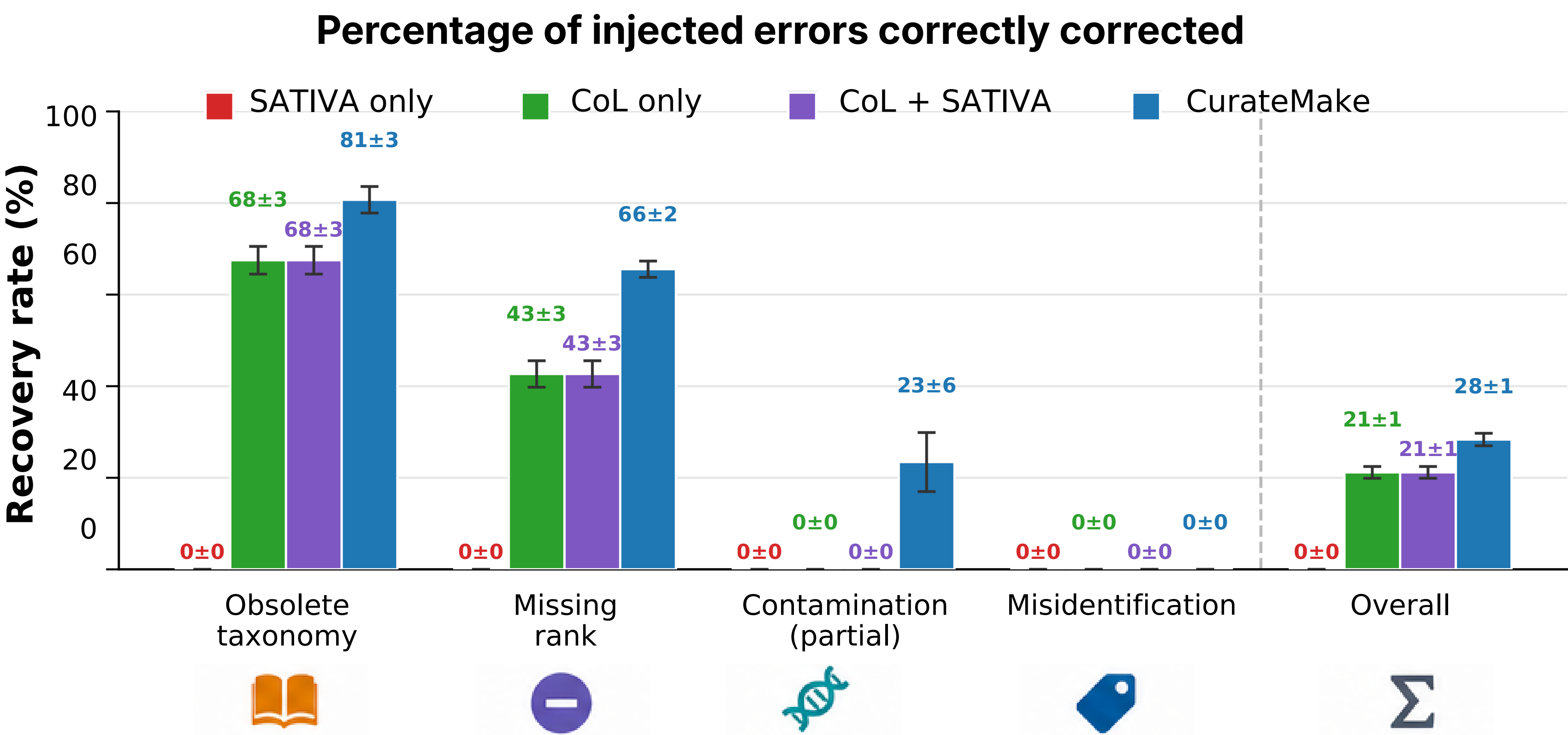
Limit

Residual species-level entropy (0.17) reflects the intrinsic limits of ITS-based species resolution.

Benchmark 2 - Component ablation under controlled errors

The full workflow corrects more injected errors than any component alone.

Step	Description	Error type	Example
1	Clean References 20 ITS of 20 different species		
2	Simulation of sequences Badreads : 100 sequences per species		
3	Inject taxonomic errors Corrupt 15% of sequence taxonomies		
4	Measure recovery Test each tool separately		
		Misidentification	Species A → Species B
		Missing rank	Genus information removed
		Obsolete taxonomy	Use of outdated synonym
		Contamination	Complete or partial switch in annotation (plantae → Fungi)



Interpretation



At 15% corruption (the reference level), CurateMake restores 28.1% of corrupted annotations, compared with 21.1% for CoL alone and 0% for SATIVA alone. SATIVA only adds value when CurateMake provides taxon-coherent alignments.