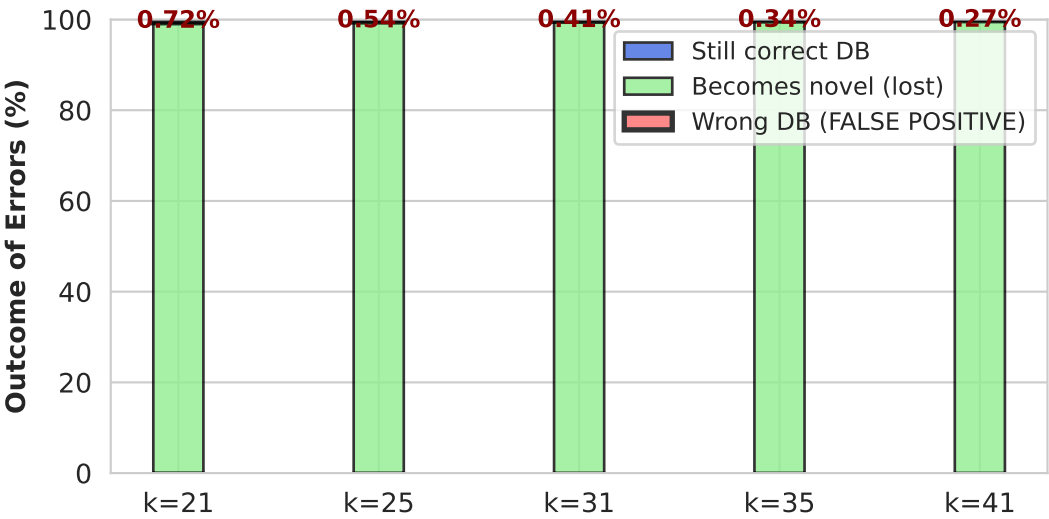
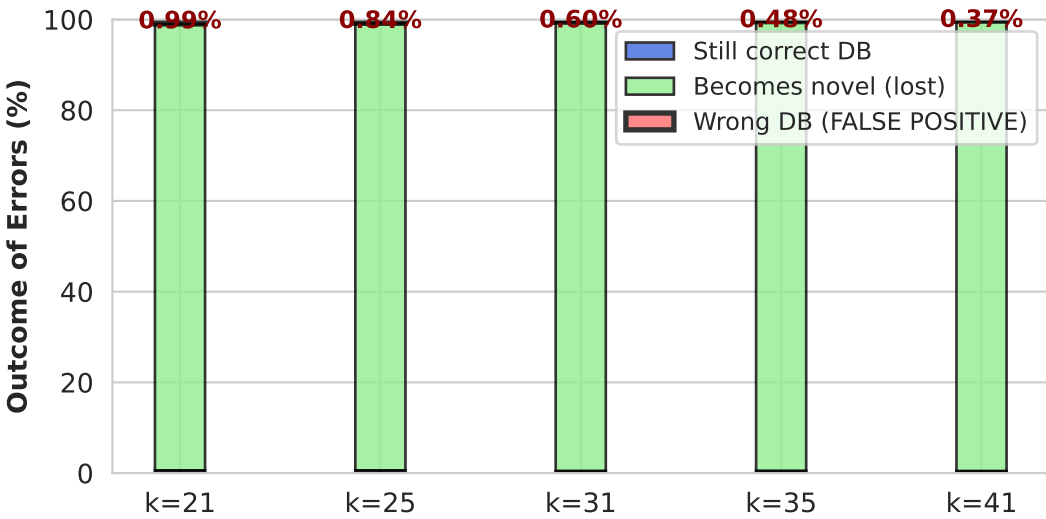


Cross-Contamination Analysis: False Positive Risk from Sequencing Errors
1% per-base error rate (ONT-like sequencing)

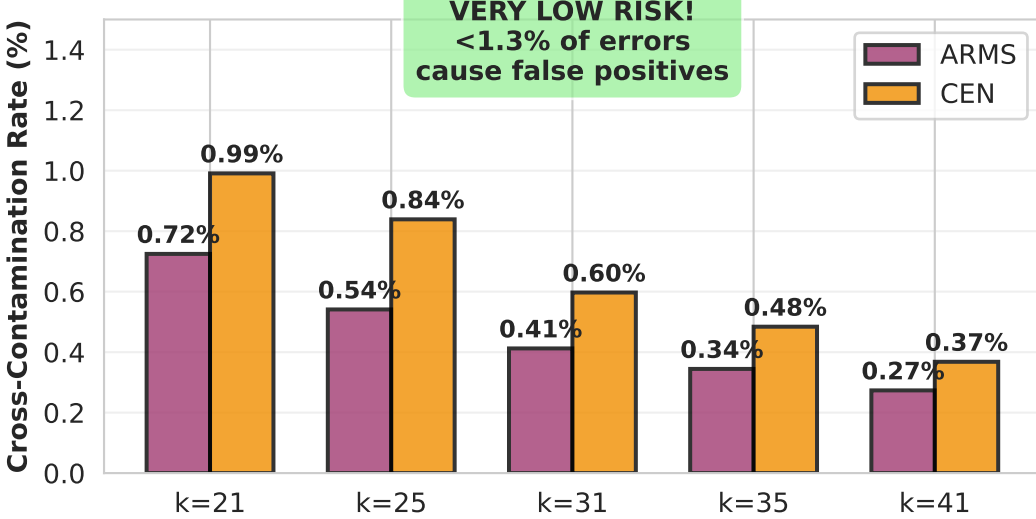
A. ARMS Markers: Error Outcomes
(of k-mers WITH sequencing errors)



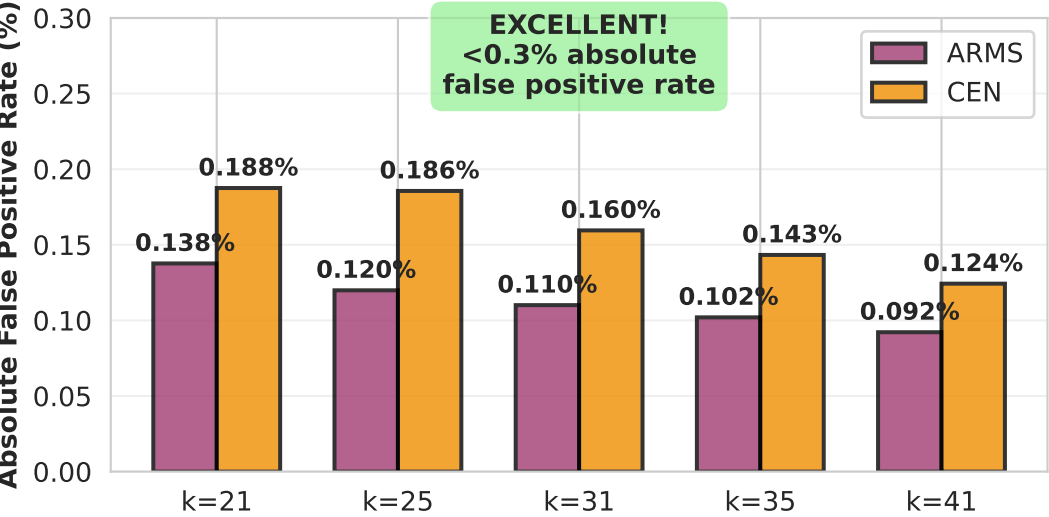
B. CEN Markers: Error Outcomes
(of k-mers WITH sequencing errors)



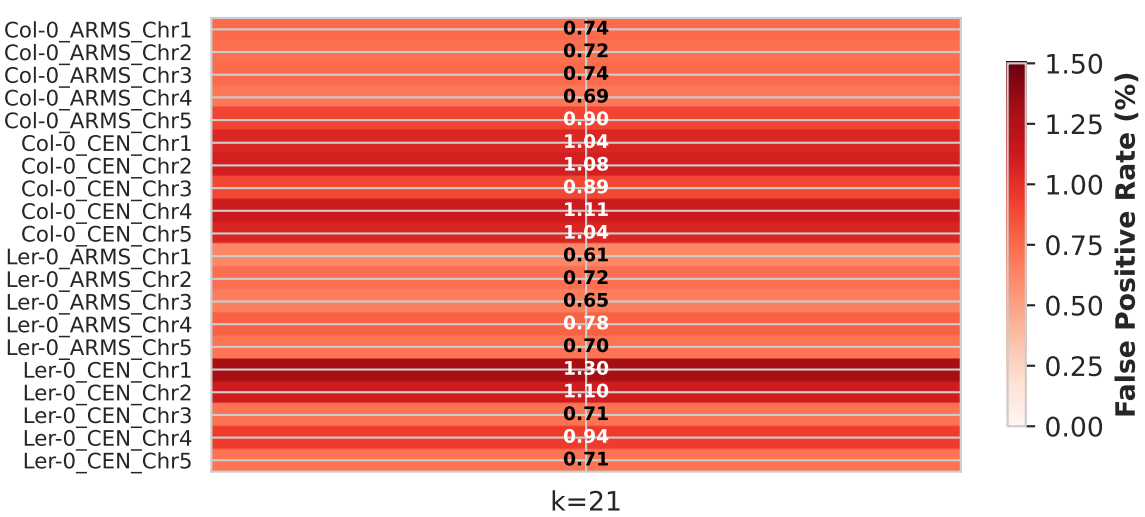
C. Cross-Contamination Risk
(FALSE POSITIVE rate)



D. Overall False Positive Risk
(% of ALL k-mers that become false positives)



E. Per-Database Cross-Contamination
(k=21 only)



CROSS-CONTAMINATION SUMMARY

When sequencing errors occur:

- ✓ 98-99% become "novel" (just lost)
- x <1.3% match wrong database (false +)

Absolute False Positive Rates:

	ARMS	CEN
k=21:	0.138%	0.188%
k=25:	0.120%	0.186%
k=31:	0.110%	0.160%
k=35:	0.102%	0.143%
k=41:	0.092%	0.124%

KEY FINDINGS:

- FALSE POSITIVE RISK IS VERY LOW (<0.3% of all k-mers)
- Most errors just LOSE reads (become novel, not misclassified)
- CEN markers have slightly higher cross-contamination (but still <1.3%)
- All k-mer sizes perform similarly for specificity

RECOMMENDATION: k=21 for best balance of read retention + specificity