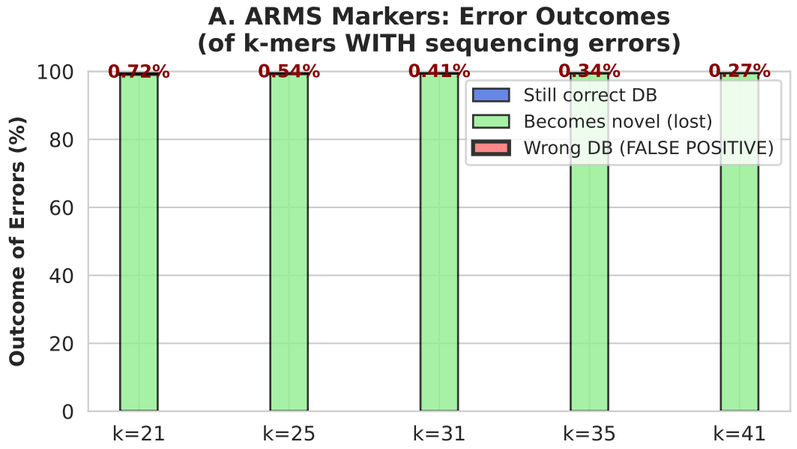
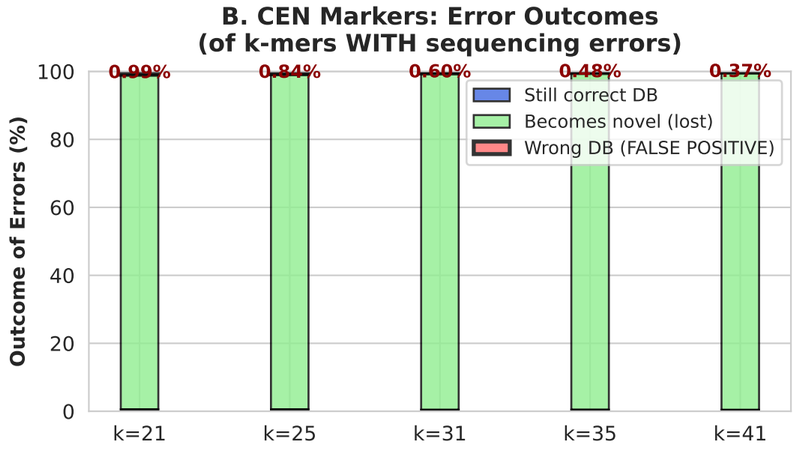
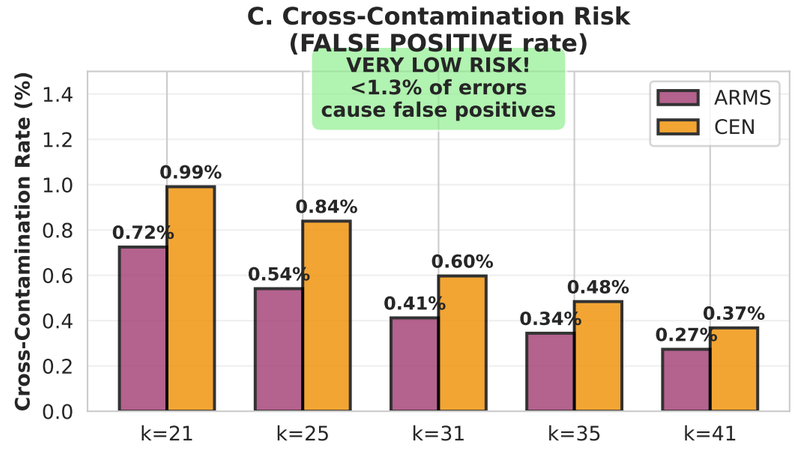
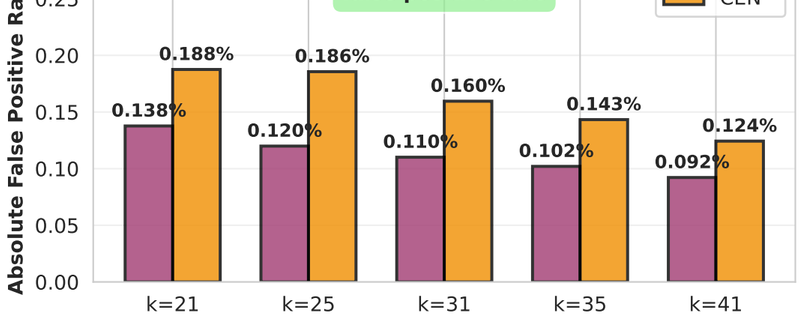
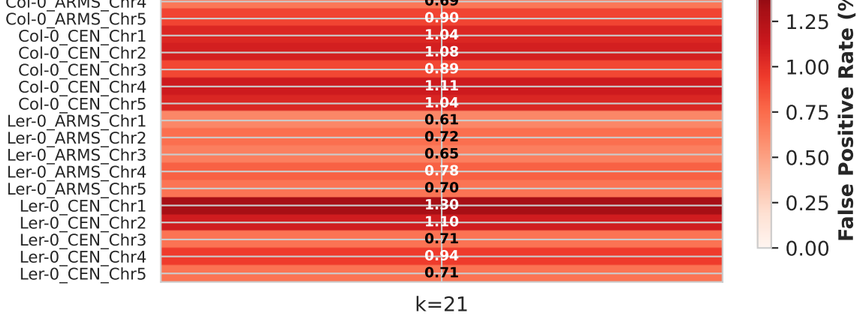
Cross-Contamination Analysis: False Positive Risk from Sequencing Errors
1% per-base error rate (ONT-like sequencing)












```
CROSS-CONTAMINATION SUMMARY

When sequencing errors occur:

✓ 98-99% become "novel" (just lost)
✗ <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:

1. FALSE POSITIVE RISK IS VERY LOW
   (<0.3% of all k-mers)

2. Most errors just LOSE reads
   (become novel, not misclassified)

3. CEN markers have slightly higher
   cross-contamination (but still <1.3%)

4. All k-mer sizes perform similarly
   for specificity

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