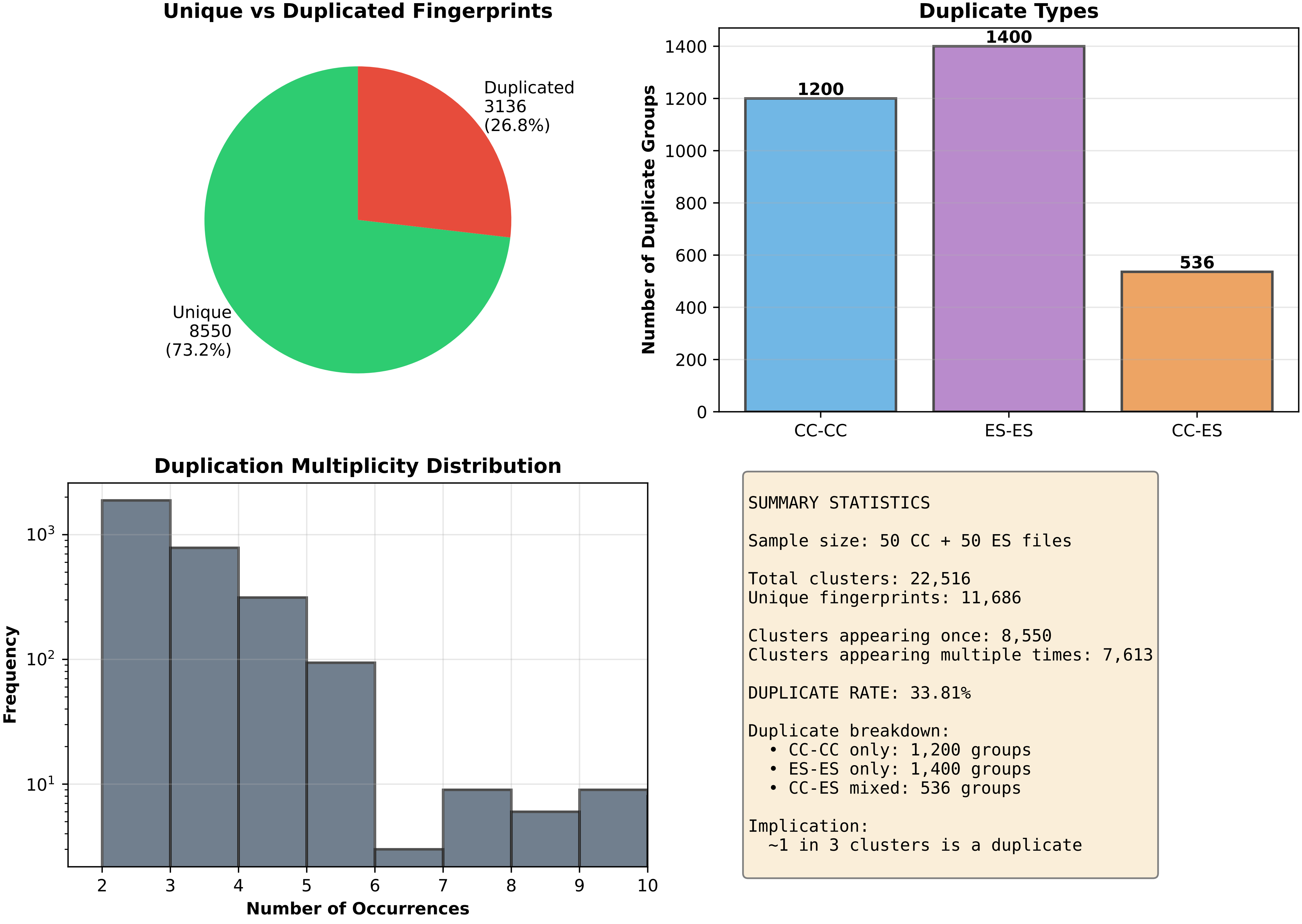
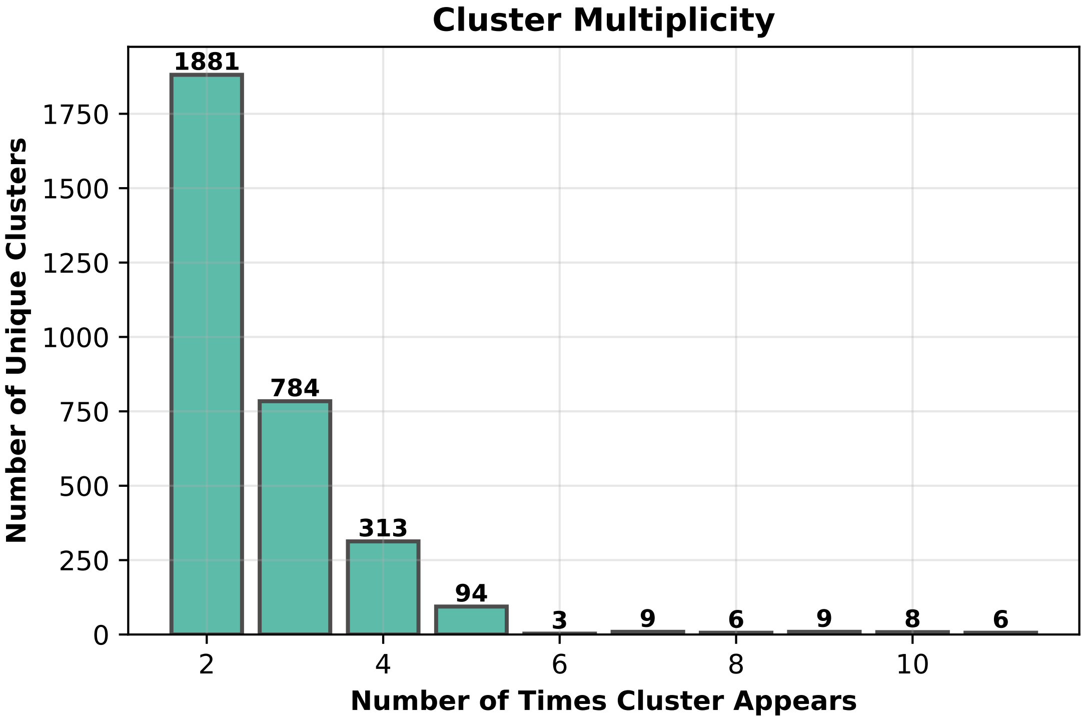
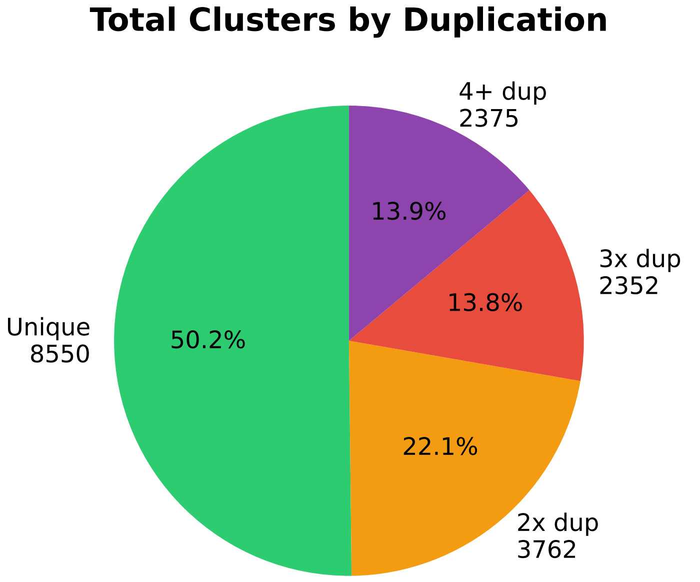
Duplicate Cluster Analysis
(50 CC + 50 ES files sample)

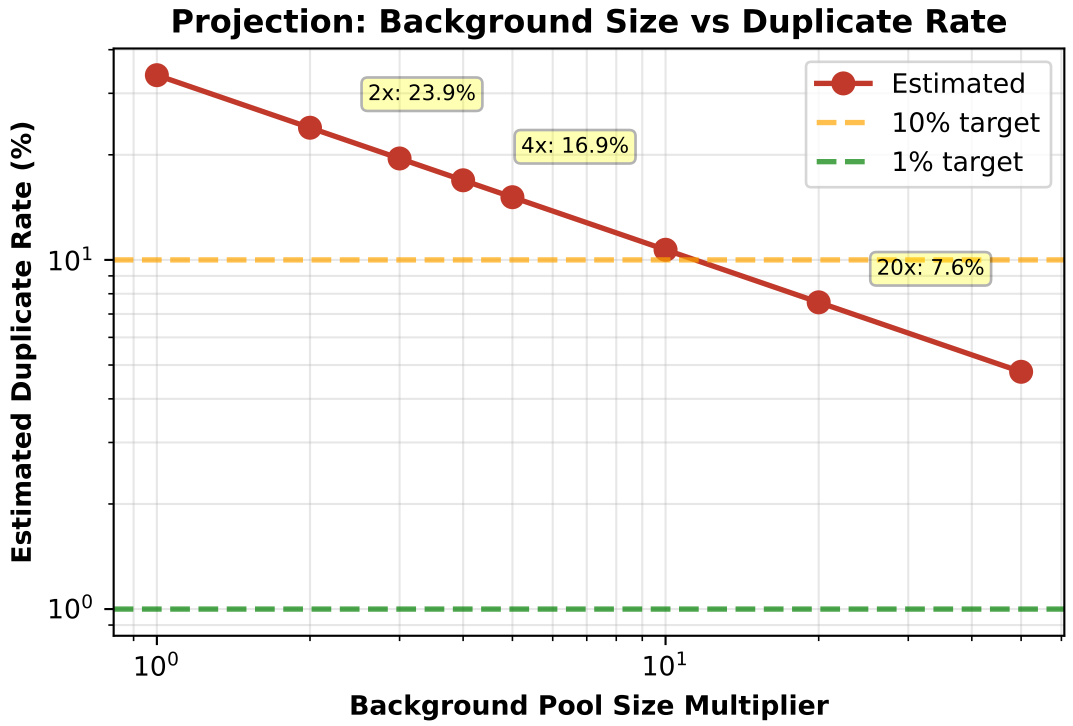


```
SUMMARY STATISTICS

Sample size: 50 CC + 50 ES files

Total clusters: 22,516
Unique fingerprints: 11,686

Clusters appearing once: 8,550
Clusters appearing multiple times: 7,613

DUPLICATE RATE: 33.81%

Duplicate breakdown:
  • CC-CC only: 1,200 groups
  • ES-ES only: 1,400 groups
  • CC-ES mixed: 536 groups

Implication:
  ~1 in 3 clusters is a duplicate
```

Duplication Patterns and Projections







```
RECOMMENDATIONS

Current status:
  • Duplicate rate: 33.8%
  • ~1 in 3 clusters is duplicated

To reduce to <10%:
  • Need ~11x larger background pool
  • From ~100 files to ~1,100 files

To reduce to <1%:
  • Need ~1,100x larger background pool
  • Practically challenging

Impact assessment:
  ✓ Less critical than file-level leakage
  ✓ Signal portions are different
  ✓ Model sees clusters in varied contexts

  ⚠ May cause overfitting to specific
    background patterns

Best practices:
  1. Monitor validation performance
  2. Use strong regularization
  3. File-level train/val/test split (✓ done)
  4. Consider data augmentation
```