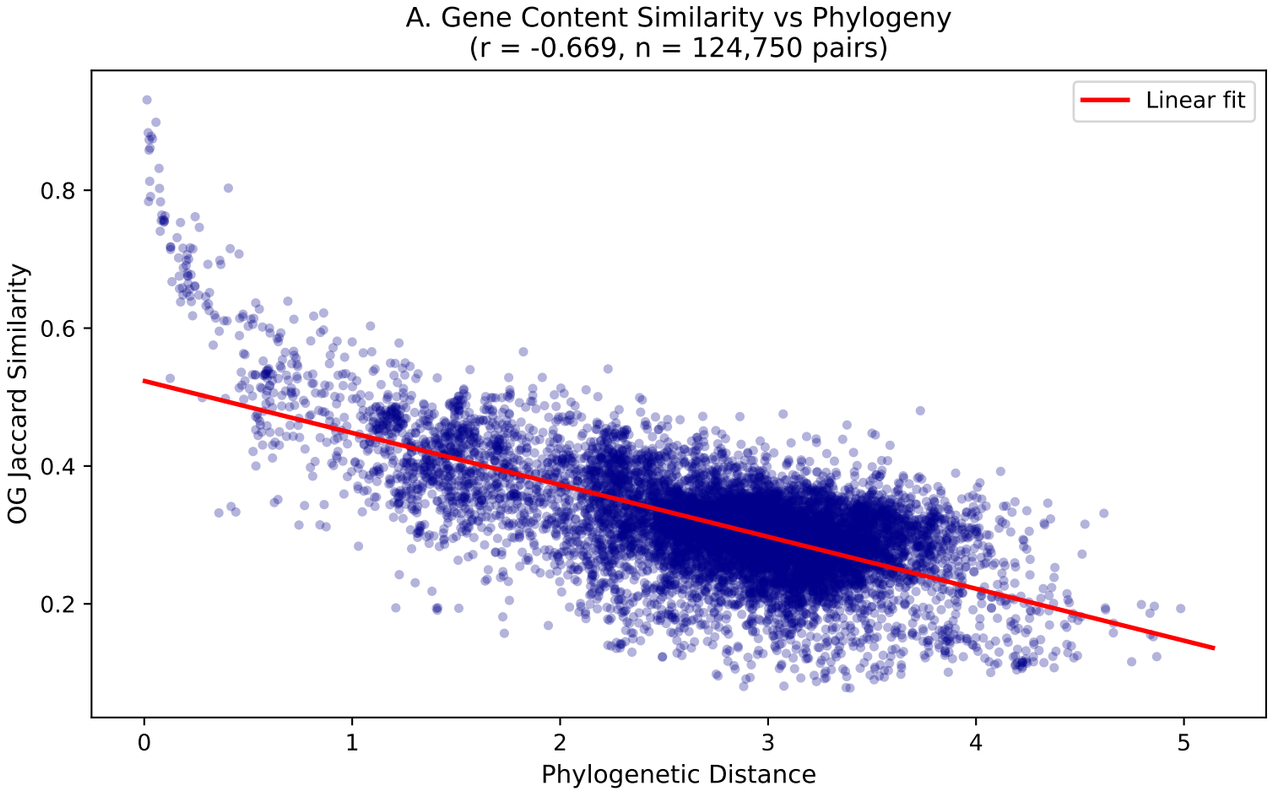
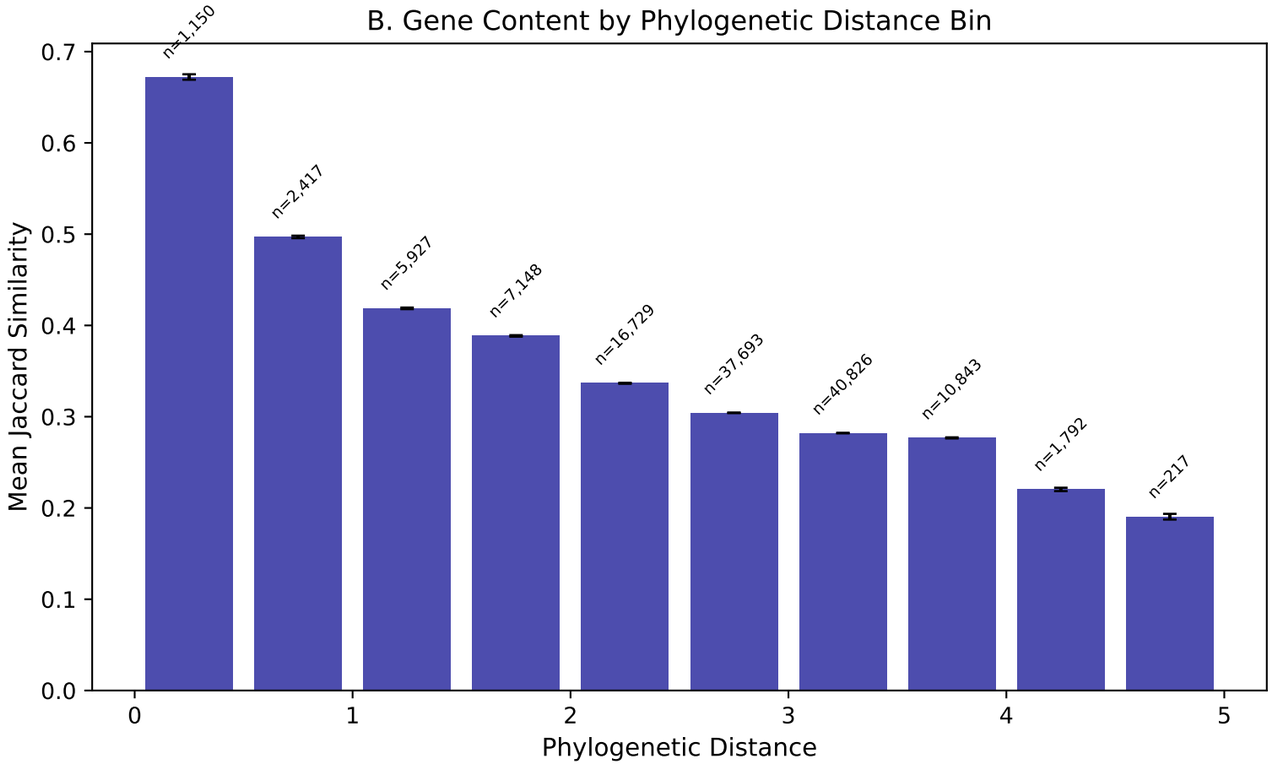
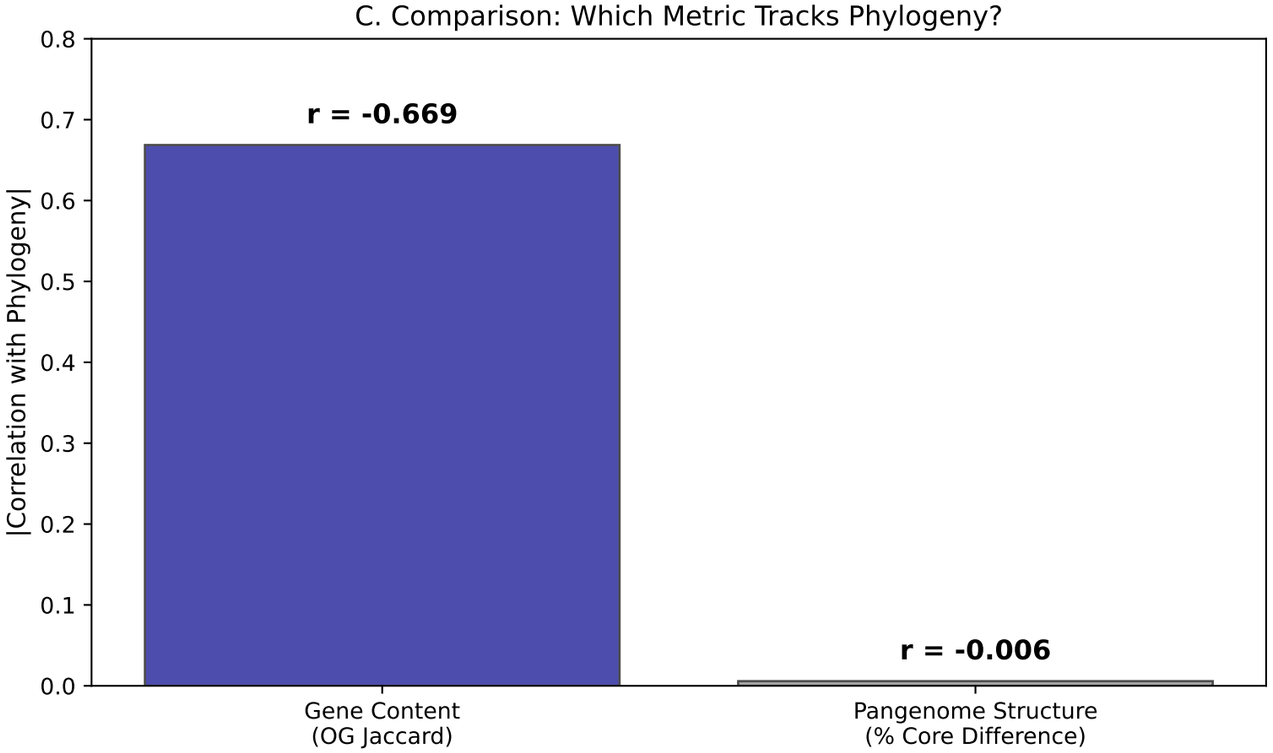
Phylogeny vs Gene Content: Complete Analysis (27,687 species dataset)








```
COMPLETE ANALYSIS: Phylogeny vs Gene Content

DATA SOURCE:
• 27,687 species from BERDL pangenome database
• 55,213,709 core gene clusters
• 82% with eggNOG ortholog group annotations
• 500-species sample for pairwise analysis

KEY FINDING:
  Gene content (shared OGs) IS strongly correlated
  with phylogenetic distance:  r = -0.67

  Pangenome structure (% core) is NOT correlated:  r ≈ 0

INTERPRETATION:
• WHAT genes a species has → determined by ancestry
• HOW genes are organized → determined by ecology

JACCARD SIMILARITY BY PHYLO DISTANCE:
• Closely related (d < 0.5): ~45% shared OGs
• Moderately related (d ~ 2): ~32% shared OGs
• Distantly related (d > 4): ~25% shared OGs
• Cross-phylum minimum: ~7% (universal core)
```