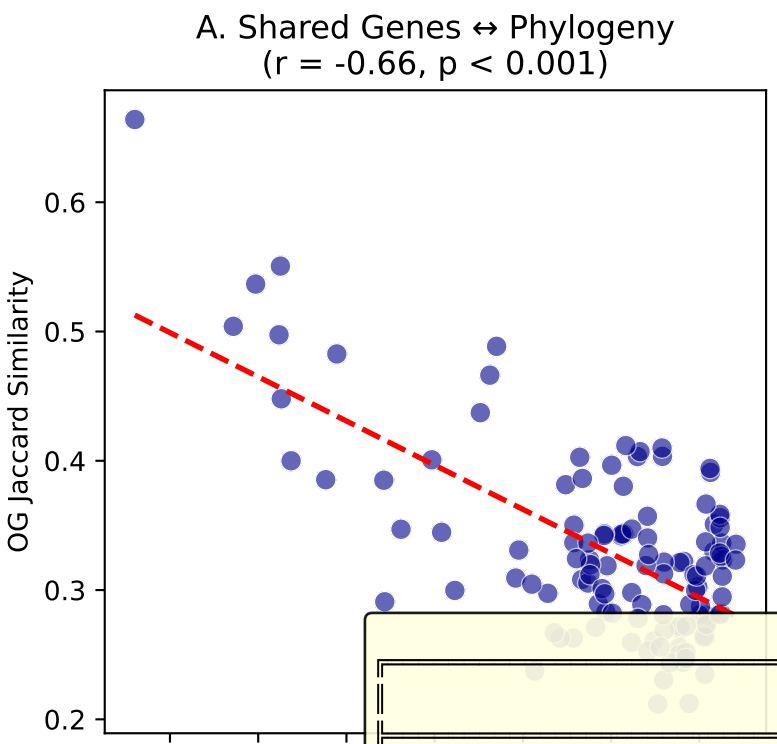
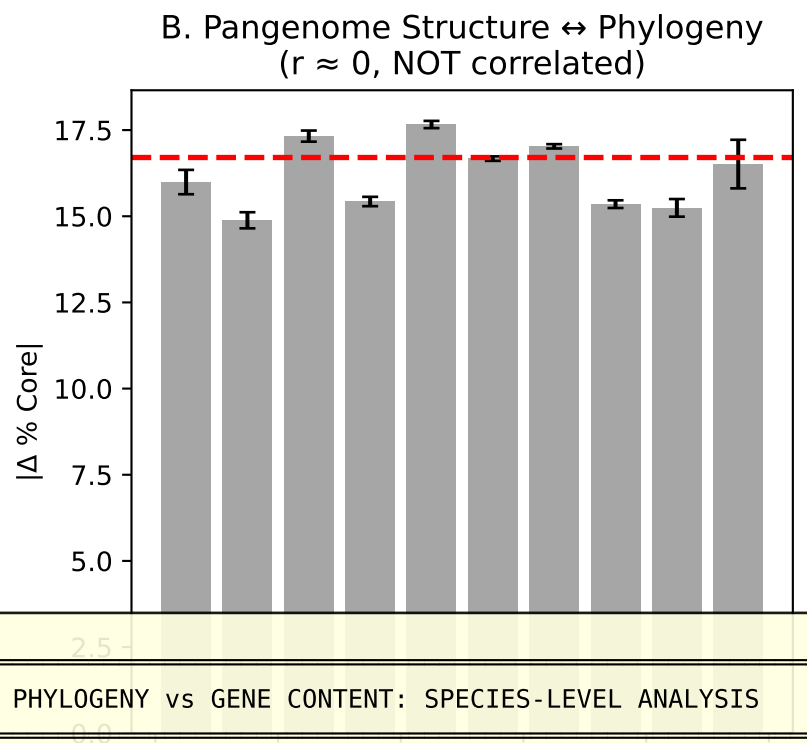
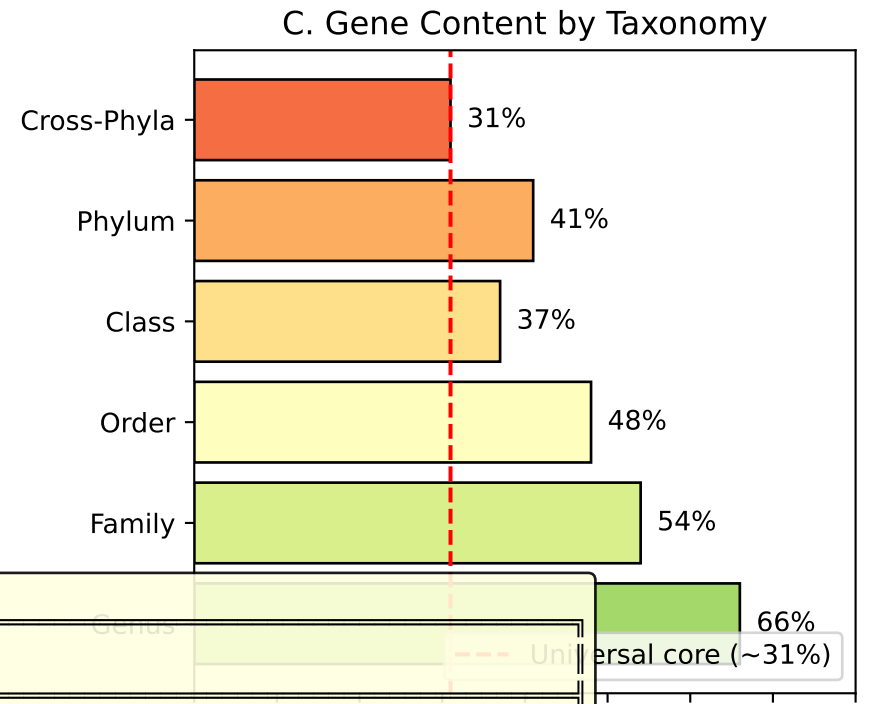
Species-Level Phylogeny ↔ Gene Content Analysis








```
PHYLOGENY vs GENE CONTENT: SPECIES-LEVEL ANALYSIS

MAIN FINDING: Two aspects of gene content show OPPOSITE relationships with phylogeny

┌──────────────────────────────────┬──────────────────────────┬──────────────────────────────────────────────┐
│ METRIC                           │ CORRELATION WITH PHYLOGENY │ INTERPRETATION                             │
├──────────────────────────────────┼──────────────────────────┼──────────────────────────────────────────────┤
│ Shared gene content (OG Jaccard) │ r = -0.66 (STRONG)       │ WHAT genes → determined by ancestry          │
│ Pangenome structure (% core)     │ r ≈ 0 (NONE)             │ HOW genes organized → determined by ecology  │
└──────────────────────────────────┴──────────────────────────┴──────────────────────────────────────────────┘

BIOLOGICAL INTERPRETATION:

• Gene content (which genes a species has) is largely inherited from ancestors
  → Closely related species share ~66% of core genes (same genus)
  → Distantly related species still share ~31% (universal bacterial machinery)

• Pangenome structure (how genes are partitioned into core/accessory) reflects ecological adaptation
  → Two sibling species can have very different % core if they adapted to different niches
  → Distantly related species can converge on similar pangenome structures if in similar ecological roles

IMPLICATIONS FOR METAGENOMICS:

• Taxonomic classification DOES predict gene content: knowing a species' genus tells you ~66% of its core genes
• But taxonomy does NOT predict pangenome "openness" or flexibility
• Functional inference from taxonomy is reliable for conserved housekeeping genes
• Accessory gene content requires strain-level resolution

Data: 16 species with OG profiles, 500 species for structure analysis | Source: BERDL GTDB Pangenome Database
```