

Molecular Systematics &amp; Complexity | Signal Rating: 5/5 Stars | Oct 30, 2026 | 12 min read

## Separating Orthologs from Paralogs Across Divergent Genomes (RBH vs. DIAMOND)

Benchmarking exact Smith-Waterman BLOSUM62 local alignment, Emms & Kelly gene-length normalization, and Reciprocal Best-Hit (RBH) orthogroup clustering against DIAMOND v2.2.8 across 80 proteins, 21 species, and 1,632 pairwise comparisons

Building accurate multi-gene phylogenetic trees requires comparing true orthologs -- genes that diverged when ancestral species split -- while excluding paralogs created by ancient gene duplications. Across four bacterial and vertebrate datasets spanning 80 proteins, 21 species, and 1,632 pairwise comparisons, our native Smith-Waterman BLOSUM62 and length-normalized Reciprocal Best-Hit (RBH) orthology engine matches exact DIAMOND v2.2.8 ( --swipe ) with  $r = 1.0000$  bit-score correlation and 100% orthogroup agreement. Furthermore, while default seed-indexed DIAMOND ( fast mode ) misses 100% (8/8) of deeply diverged cross-phylum bacterial replication-origin orthologs in the 24%-28% amino-acid identity twilight zone, full dynamic-programming alignment recovers every single-copy core orthogroup.

### 1. Biological Context: Orthologs, In-Paralogs, Out-Paralogs, and the Twilight Zone

In comparative genomics and molecular systematics, homologous proteins -- proteins that share a common evolutionary ancestor -- fall into distinct functional and evolutionary categories depending on how they diverged (Fitch, 1970; Koonin, 2005). Orthologs are homologous genes in different species that diverged from a single ancestral gene during a speciation event; because they trace the splitting of species lineages directly, single-copy orthologs are the foundational markers used to reconstruct species trees. By contrast, paralogs arise through gene duplication events within a genome. When a gene duplication occurs after two species have already split, the duplicated copies are called in-paralogs (or co-orthologs relative to an outgroup genome). When a duplication occurred before the speciation event -- such as the ancient vertebrate globin duplications that separated Myoglobin ( MB ), Alpha-Globin ( HBA ), and Beta-Globin ( HBB ) hundreds of millions of years ago -- the resulting genes in modern species are out-paralogs .

Mistaking an ancient out-paralog for a 1:1 ortholog is one of the most damaging errors in phylogenomics: instead of measuring when two species diverged, the tree inadvertently measures when the ancestral gene family duplicated. Separating true orthologs from out-paralogs across whole proteomes faces three recurring quantitative challenges:

\* The Protein Twilight Zone (<35% Amino-Acid Identity): Over hundreds of millions of years of evolution -- such as between Gram-positive Firmicutes ( Bacillus , Staphylococcus ) and Gram-negative Gammaproteobacteria ( Pseudomonas ) -- essential single-copy proteins can diverge until only 20%-30% of their amino-acid residues remain identical (Rost, 1999). Fast heuristic aligners that rely on exact or spaced

k-mer seed matches often fail to trigger an alignment in this twilight zone, causing core genes to drop out of supermatrix analyses.

\* Gene-Length Bias in Alignment Bit-Scores: Because local alignment scores accumulate approximately linearly with the number of aligned residues, a 938-amino-acid enzyme such as Isoleucyl- tRNA synthetase ( IleS ) generates a raw bit-score more than three times higher than a 313-amino-acid enzyme such as Riboflavin kinase ( RibF ) at the exact same evolutionary divergence. As shown by Emms & Kelly (2015) in OrthoFinder, unnormalized bit-scores systematically bias cross-proteome clustering against short proteins and favor long multi-domain proteins.

\* Genomic Island Insertions and Synteny Breakpoints: Horizontal gene transfer (HGT) frequently inserts tens of kilobases of accessory DNA into conserved chromosomal backbones. An orthology engine must verify that conserved genes flanking an insertion remain true 1:1 positional orthologs despite large shifts in genomic coordinates.

## 2. Four-Stage Architecture and Head-to-Head Calibration Against DIAMOND v2.2.8

To evaluate how exact dynamic programming compares with heuristic seed-and-extend protein search, we benchmarked our four-stage protein orthology engine against DIAMOND v2.2.8 (Buchfink et al., 2015, 2021) executed in both its exact vectorized Smith-Waterman mode ( `diamond blastp --swipe --comp-based-stats 0` ) and its default double-indexed seed heuristic mode ( `diamond blastp fast mode` ). Our native architecture processes candidate proteomes through four sequential stages:

\* Stage 1 -- Exact Gotoh 3-State Affine-Gap Smith-Waterman Alignment ( BLOSUM62 ): All-vs-all pairwise local protein alignments are computed using the standard 24x24 NCBI BLOSUM62 substitution matrix (Henikoff & Henikoff, 1992) with affine gap penalties (gap open = 11, gap extension = 1, where a gap of length  $L \geq 1$  costs  $11 + 1 * L$  ) and bit-packed traceback to recover exact aligned intervals, percent amino-acid identity, mismatches, and gap openings.

\* Stage 2 -- Karlin-Altschul Edge-Corrected Bit-Score and E-Value Statistics : Raw Smith-Waterman scores (  $S$  ) are converted into database-independent bit-scores (  $S' = (\lambda S - \ln K) / \ln 2$  ) and expectation values (  $E = m_{eff} * n_{eff} * 2^{-S'}$  ) using the standard ungapped/gapped BLOSUM62 Karlin-Altschul parameters ( $\lambda = 0.267$ ,  $K = 0.0410$ , relative entropy  $H = 0.140$ ; Karlin & Altschul, 1990; Altschul & Gish, 1996).

\* Stage 3 -- Emms & Kelly (2015) Log-Linear Gene-Length Normalization: Pairwise hits are binned by the product of query and target protein lengths (  $L_i * L_j$  ), and the top hits in each length bin are fit by log-linear regression (  $\log_{10} S'_{ij} = a + b \log_{10} (L_i L_j)$  ). Dividing raw bit-scores by the length trend (  $B_{ij} = S'_{ij} / (10^{a + b \log_{10} (L_i L_j)})$  ) removes gene-length bias so short and long proteins compete on an equal footing.

\* Stage 4 -- Reciprocal Best-Hit (RBH) & Paralog Graph Partitioning into Supermatrix Phylogenetics : Cross-species pairs that represent mutual best length-normalized hits (above minimum coverage  $\geq 50\%$  and significance  $E \leq 10^{-5}$  ) are linked as

1:1 RBH orthologs. Within-species duplicates whose self-species score exceeds their strongest cross-species RBH threshold (  $S_{\text{within}} \geq 0.85 * S_{\text{RBH}}$  ) are grouped as lineage-specific co-ortholog in-paralogs , whereas ancient duplicates falling below the speciation threshold are rejected as out-paralogs . Connected components yielding strictly one gene per species (1:1 core orthogroups) are then aligned at the nucleotide level and passed into maximum-likelihood supermatrix tree reconstruction.

| Cohort ID || Biological System & Taxonomic Scope || Proteins || Taxa ||  
 All-vs-All Pairs || Bit-Score  $r$  (vs. --swipe) || Mean |DeltaBit| (bits) || RBH  
 Jaccard (--swipe) || RBH Jaccard (Fast) || Orthogroups (1:1 Core) || Twilight  
 (<35% ID) Recall (Native / --swipe) || Twilight (<35% ID) Recall (DIAMOND Fast) ||  
 Native Wall Time (ms) || DIAMOND --swipe Time (ms) |

| COHORT-01-PROK-ORIC-PARALOGS || Bacterial oriC Core + GyrB / ParC Paralogs ( *B. subtilis* , *S. aureus* , *P. aeruginosa* ) || 16 || 3 || 256 || 1.0000 || 0.5179 |  
 | 1.0000 || 0.6923 || 6 (3) || 100.0% (8/8) || 0.0% (0/8) || 240.3 || 138.9 |

| COHORT-02-ENTERIC-SYNTENY-10LOCI || Enterobacterial 60 kb Syntenic Backbone vs.  
 +31.4 kb Island ( *E. coli* K-12 vs. *S. enterica* LT2) || 20 || 2 || 400 || 1.0000  
 || 0.3000 || 1.0000 || 1.0000 || 10 (10) || 100.0% (0/0) || 100.0% (0/0) ||  
 664.2 || 159.1 |

| COHORT-03-VERTEBRATE-MITO-COX1-COX2 || Jawed Vertebrate Mitochondrial Complex IV  
 COX1 + COX2 (12 Mammals & Birds, icode = 1 ) || 24 || 12 || 576 || 1.0000 ||  
 0.5920 || 1.0000 || 1.0000 || 2 (2) || 100.0% (0/0) || 100.0% (0/0) || 324.4 |  
 | 135.4 |

| COHORT-04-MAMMAL-GLOBIN-APOE-PARALOGS || Mammalian Globin ( HBA1/HBA2/HBB/MB ) &  
 Apolipoprotein ( APOE/APOA1 ) Superfamilies || 20 || 4 || 400 || 1.0000 ||  
 0.4397 || 1.0000 || 1.0000 || 5 (1) || 100.0% (62/62) || 90.3% (56/62) || 66.4  
 || 115.4 |

| Global Summary || All 4 Multi-Clade Empirical Benchmarks Combined || 80 || 21 |  
 | 1632 || 1.0000 || 0.4624 || 1.0000 || 0.9231 || 23 (16) || 100.0% (70/70) |  
 | 80.0% (56/70) || 1295.3 || 548.8 |

Generated by: Native Rust 4-Stage Smith-Waterman BLOSUM62, Emms-Kelly  
 Length-Normalized RBH & Paralog Graph Engine ( orthology.rs ) \* DIAMOND v2.2.8 (   
 blastp --swipe / fast : Buchfink et al. 2021)

Across all 1,632 all-vs-all pairwise protein comparisons in Table 1, our native Smith-Waterman engine achieves an exact raw alignment score match (  $S_{\text{native}} = S_{\text{diamond}}$  ) on every pair and a Pearson bit-score correlation of  $r = 1.0000$  against exact DIAMOND v2.2.8 --swipe (mean absolute bit-score difference |Delta  $S_{\text{prime}}$ | = 0.4624 bits, reflecting only the minor effective-length edge-correction term between standalone pairwise scoring and database-level scoring). Both engines achieve a 1.0000 Jaccard index across all 185 reciprocal best hits and a 1.0000 Rand concordance index across all 23 orthogroups (including all 16 single-copy core orthogroups). On the 20-protein mammalian globin and apolipoprotein cohort ( COHORT-04 ), native in-memory execution completes in 66.4 ms compared to 115.4 ms for DIAMOND --swipe (a 1.74x speedup by avoiding external binary database formatting

and disk I/O overhead on targeted gene panels).

### 3. Twilight-Zone Sensitivity and Paralog Discrimination in Bacterial *oriC* Loci

The biological importance of exact dynamic programming is illustrated in COHORT-01-PROK-ORIC-PARALOGS, which compares 16 proteins encoded near the chromosomal replication origin ( *oriC* : *dnaA* , *dnaN* , *recF* , *gyrB* , and *gyrA* ) plus topoisomerase paralogs across two Gram-positive Firmicutes ( *Bacillus subtilis* 168 and *Staphylococcus aureus* NCTC 8325) and the Gram-negative Gammaproteobacterium *Pseudomonas aeruginosa* PAO1. While the replication initiator *DnaA* remains moderately conserved across phyla (43.5%-45.1% amino-acid identity, 360-399 bits), both the beta sliding-clamp subunit *DnaN* (25.8%-26.5% identity, 154-159 bits,  $E \leq 1.5 \times 10^{-46}$ ) and the recombination repair ATPase *RecF* (24.3%-27.8% identity, 115-117 bits,  $E \leq 3.5 \times 10^{-32}$ ) reside deep inside the protein twilight zone when comparing Firmicutes against *P. aeruginosa*.

When run in its default fast double-indexed seed mode, DIAMOND v2.2.8 misses 100% (0/8) of these cross-phylum twilight-zone alignments, breaking the 3-species reciprocal best-hit graph for both OG\_*DnaN* and OG\_*RecF* and reducing the number of recovered single-copy core orthogroups from 3 down to 1 ( OG\_*DnaA* alone, RBH Jaccard = 0.6923). By contrast, both our native Gotoh Smith-Waterman engine and DIAMOND --swipe recover all 8 cross-phylum twilight-zone hits with highly significant expectation values (  $E \leq 10^{-31}$  ), preserving all three 1:1 core *oriC* orthogroups.

Generated by: Native Rust Gotoh 3-State Affine-Gap & 9-mer Anchor Multiple Sequence Aligner ( phylo.rs ) \* Native Rust 4-Stage Smith-Waterman BLOSUM62, Emms-Kelly Length-Normalized RBH & Paralog Graph Engine ( orthology.rs )

Figure 1 also shows how Stage 4 separates recent antibiotic-resistance in-paralogs from ancient topoisomerase out-paralogs. *P. aeruginosa* encodes both a housekeeping DNA gyrase subunit B ( PSEUD\_GyrB1 ) and a secondary copy ( PSEUD\_GyrB2 ) that carries the Arg138Lys substitution in the GHKL ATP-binding lid associated with self-resistance to aminocoumarin antibiotics. Because PSEUD\_GyrB1 and PSEUD\_GyrB2 share 89.8% amino-acid identity ( 1,498.0 bits ), far exceeding the cross-phylum GyrB speciation score against *B. subtilis* ( 692.0 bits , 48.1% ID), the orthology graph correctly merges PSEUD\_GyrB2 into OG\_GyrB as a lineage-specific co-ortholog in-paralog. Conversely, the Topoisomerase IV subunit BACIL\_ParC matches BACIL\_GyrA only across a short 76-residue local alignment in the winged-helix catalytic domain ( 102.0 bits , 9.2% query coverage) and is cleanly rejected as an ancient out-paralog.

### 4. Correcting Gene-Length Bias and Navigating a +31.4 kb Genomic Island Insertion

In COHORT-02-ENTERIC-SYNTENY-10LOCI, we extracted the first 60 kb downstream of the replication origin in *Escherichia coli* K-12 MG1655 ( NC\_000913.3:1-60000 ) and *Salmonella enterica* serovar Typhimurium LT2 ( NC\_003197.2:1-60000 ). Across this region, both genomes share 10 core metabolic and chaperone enzymes ( *ThrA* , *ThrB* , *ThrC* , *YaaJ* , *TalB* , *DnaK* , *DnaJ* , *NhaA* , *RibF* , and *IleS* ) with 76.3% to 96.9%

amino-acid identity. However, between dnaJ and nhaA , S. enterica LT2 carries a +31.4 kb lineage-specific genomic island insertion that shifts the coordinates of nhaA-ribF-ileS from 17.4-25.2 kb in E. coli out to 48.6-56.7 kb in S. enterica . Similarly, in COHORT-03-VERTEBRATE-MITO-COX1-COX2 , we translated the adjacent mitochondrial Cytochrome c Oxidase subunits COX1 (513-523 aa) and COX2 (220-239 aa) across 12 jawed vertebrates using the vertebrate mitochondrial genetic code ( icode = 1 , where UGA encodes Tryptophan and AUA encodes Methionine).

| Cohort    | Orthogroup ID | Protein Symbol                                    | Length (aa) | E. coli / Query Coord (bp) | S. enterica / Target Coord (bp) | Coord Shift (bp) | Mean AA Identity (%) | Raw Bit-Score (bits) | Length-Normalized Score (bits) | Pre-Norm Slope b | Post-Norm Slope b |        |
|-----------|---------------|---------------------------------------------------|-------------|----------------------------|---------------------------------|------------------|----------------------|----------------------|--------------------------------|------------------|-------------------|--------|
| COHORT-02 | OG_ThrA       | ThrA (Aspartokinase I)                            | 820         | 337..2799                  | 337..2799                       | 0                | 94.5%                | 1525.0               | 936.4                          | 0.4923           | 0.0000            |        |
| COHORT-02 | OG_ThrB       | ThrB (Homoserine kinase)                          | 310         | 2801..3733                 | 2801..3730                      | -3               | 93.5%                | 574.0                | 922.1                          | 0.4923           | 0.0000            |        |
| COHORT-02 | OG_DnaK       | DnaK (Chaperone Hsp70)                            | 638         | 12163..14079               | 12183..14099                    | +20              | 96.9%                | 1186.0               | 928.5                          | 0.4923           | 0.0000            |        |
| COHORT-02 | OG_DnaJ       | DnaJ (Co-chaperone Hsp40)                         | 376         | 14168..15298               | 14193..15326                    | +28              | 95.3%                | 721.0                | 959.2                          | 0.4923           | 0.0000            |        |
| COHORT-02 | OG_NhaA       | NhaA (Na <sup>+</sup> /H <sup>+</sup> antiporter) | 388         | 17489..18655               | 48604..49770                    | +31115           | 93.2%                | 709.0                | 914.3                          | 0.4923           | 0.0000            |        |
| COHORT-02 | OG_RibF       | RibF (Riboflavin kinase)                          | 313         | 21407..22348               | 52835..53776                    | +31428           | 90.0%                | 553.0                | 880.0                          | 0.4923           | 0.0000            |        |
| COHORT-02 | OG_IleS       | IleS (Isoleucyl-tRNA synthetase)                  | 938         | 1693.0                     | 22391..25207                    | 53819..56689     | +31482               | 87.1%                | 1693.0                         | 913.8            | 0.4923            | 0.0000 |
| COHORT-03 | OG_COX1       | MT-CO1 (Cytochrome c oxidase I)                   | 515         | 1..1545 (12 Taxa)          | 1..1545 (12 Taxa)               | 0                | 90.1%                | 979.2                | 586.5                          | 0.6332           | 0.0000            |        |
| COHORT-03 | OG_COX2       | MT-CO2 (Cytochrome c oxidase II)                  | 227         | 1685..2368 (12 Taxa)       | 1685..2368 (12 Taxa)            | 0                | 75.7%                | 353.9                | 590.7                          | 0.6332           | 0.0000            |        |

Generated by: Native Rust 4-Stage Smith-Waterman BLOSUM62, Emms-Kelly Length-Normalized RBH & Paralog Graph Engine ( orthology.rs )

Table 2 illustrates how Stage 3 corrects gene-length bias across both datasets. Before normalization, the 938-residue IleS enzyme yields a raw bit-score of 1,693.0 bits whereas the 313-residue RibF enzyme immediately upstream yields 553.0 bits -- a 3.06x raw score disparity driven purely by polypeptide length (log-length regression slope b = 0.4923). Similarly, across the 12 vertebrate mitogenomes in COHORT-03 , COX1 (515 aa) averages 979.2 raw bits compared to 353.9 raw bits for COX2 (227 aa), a 2.77x disparity ( b = 0.6332). Applying Emms & Kelly log-linear normalization brings IleS ( 913.8 normalized bits) and RibF ( 880.0 normalized bits) into direct parity with their percent identities, and equalizes COX1 ( 586.5 normalized bits)

and COX2 ( 590.7 normalized bits), reducing the residual length-bias slope to b = 0.0000 in both cohorts.

## 5. Mammalian Globin and Apolipoprotein In-Paralog vs. Out-Paralog Separation

To test paralog resolution in vertebrate multi-gene families, COHORT-04-MAMMAL-GLOBIN-APOE-PARALOGS analyzes 20 proteins across four mammals ( Homo sapiens , Pan troglodytes , Macaca mulatta , and Mus musculus ) spanning the classic globin superfamily (Alpha-Globin HBA1 / HBA2 , Beta-Globin HBB , and monomeric Myoglobin MB ) and the lipid-transport apolipoprotein family ( APOE and APOA1 ).

| Query Protein || Target Protein || Evolutionary Relationship || Aligned Length (aa) || AA Identity (%) || Raw Bit-Score (bits) || Normalized Bit-Score (bits) || Expectation E-Value || Assigned Orthogroup || Graph Decision |

| HUMAN\_HBA1 || PANTR\_HBA1 || 1:1 Speciation Ortholog (Human-Chimp) || 142 || 100.0% || 287.0 || 664.8 || 1.63e-89 || OG\_AlphaGlobin || Accepted (RBH Core Edge) |

| HUMAN\_HBA1 || HUMAN\_HBA2 || Concerted-Evolution In-Paralog (Chr16p13.3) || 142 || 100.0% || 287.0 || 664.8 || 1.63e-89 || OG\_AlphaGlobin || Merged as Co-Ortholog In-Paralog |

| HUMAN\_HBA1 || HUMAN\_HBB || Ancient Tetrameric Globin Out-Paralog (~450 MYA) || 139 || 43.2% || 115.1 || 263.8 || 1.04e-37 || Separate ( OG\_BetaGlobin ) || Rejected Out-Paralog (< RBH Gate) |

| HUMAN\_HBA1 || HUMAN\_MB || Ancient Monomeric Globin Out-Paralog (~600 MYA) || 140 || 26.4% || 55.8 || 125.9 || 7.45e-20 || Separate ( OG\_Myoglobin ) || Rejected Out-Paralog (Twilight Zone) |

| HUMAN\_APOE3 || PANTR\_APOE || 1:1 Speciation Ortholog (Human-Chimp) || 317 || 97.5% || 602.4 || 554.7 || 1.18e-183 || OG\_APOE || Accepted (RBH Core Edge) |

| HUMAN\_APOE3 || HUMAN\_APOE4 || Allelic / Isoform Variant (Cys112Arg) || 317 || 99.7% || 610.6 || 562.3 || 4.12e-186 || OG\_APOE || Merged as Within-Taxon Variant |

| HUMAN\_APOE3 || HUMAN\_APOA1 || Tandem Amphipathic Helix Out-Paralog || 192 || 22.4% || 68.2 || 68.9 || 6.20e-23 || Separate ( OG\_APOA1 ) || Rejected Out-Paralog (Twilight Zone) |

Generated by: Native Rust 4-Stage Smith-Waterman BLOSUM62, Emms-Kelly Length-Normalized RBH & Paralog Graph Engine ( orthology.rs )

As shown in Table 3, human HBA1 and HBA2 on chromosome 16p13.3 encode identical 142-amino-acid alpha-globin chains (100.0% identity, 287.0 raw bits, 664.8 normalized bits) due to ongoing concerted evolution via non-allelic homologous gene conversion. Both our native engine and DIAMOND --swipe cluster HBA1 and HBA2 into OG\_AlphaGlobin as co-ortholog in-paralogs while cleanly separating them from Beta-Globin ( OG\_BetaGlobin , 43.2% identity, 115.1 raw bits) and Myoglobin ( OG\_Myoglobin , 26.4% identity, 55.8 raw bits). Likewise, APOE and APOA1 -- which

share homologous 22-residue amphipathic alpha-helical lipid-binding repeats in the 20.3%-23.1% identity twilight zone ( 58.5-71.6 bits) -- are partitioned without false-positive cross-bridging into OG\_APOE and OG\_APOA1 .

## 6. Downstream Concatenated Supermatrix Phylogeny Across 12 Jawed Vertebrates

Once Stage 4 identifies the single-copy core orthogroups ( OG\_COX1 and OG\_COX2 ) across all 12 vertebrates in COHORT-03 , the engine automatically concatenates their nucleotide coding sequences into a 2,229 bp supermatrix (1,545 bp COX1 + 684 bp COX2 ), selects the best-fit substitution model ( TIM2+I+G / GTR+F+I+G4 ), and reconstructs the maximum-likelihood species tree ( ln L = -14,295.07 , total tree length = 2.6606 substitutions/site, mean SH-aLRT branch support = 98.4%).

Generated by: Native Rust 4-Stage ML Tree Builder ( tree\_builder.rs : Multi-Distance BIONJ + Exact Yang Gamma4 + Inside-Outside Pulley NNI) \* Native Rust 4-Stage Smith-Waterman BLOSUM62, Emms-Kelly Length-Normalized RBH & Paralog Graph Engine ( orthology.rs )

Generated by: Native Rust K2P / TN93 Distance, Neighbor-Joining & ABGD Systematics Engine ( phylo.rs ) \* Native Rust 4-Stage Smith-Waterman BLOSUM62, Emms-Kelly Length-Normalized RBH & Paralog Graph Engine ( orthology.rs )

## 7. Verified Specimen & Sequence Provenance Ledger

| Cohort ID || Organism (Scientific Name) || Clade / Phylum || NCBI RefSeq Accession || Genomic Window / Loci || Proteins Extracted || Genetic Code || Database Verification |

| COHORT-01 || *Bacillus subtilis* subsp. *subtilis* str. 168 || Bacteria (Firmicutes) || NC\_000964.3 || oriC Window (1..25000: dnaA, dnaN, recF, gyrB, gyrA, parC ) || 6 || Bacterial (Table 11) || NCBI RefSeq Verified |

| COHORT-01 || *Staphylococcus aureus* subsp. *aureus* NCTC 8325 || Bacteria (Firmicutes) || NC\_007795.1 || oriC Window (1..25000: dnaA, dnaN, recF, gyrB, gyrA ) || 5 || Bacterial (Table 11) || NCBI RefSeq Verified |

| COHORT-01 || *Pseudomonas aeruginosa* PAO1 || Bacteria (Gammaproteobacteria) || NC\_002516.2 || oriC Window (1..25000: dnaA, dnaN, recF, gyrB1, gyrB2 ) || 5 || Bacterial (Table 11) || NCBI RefSeq Verified |

| COHORT-02 || *Escherichia coli* str. K-12 substr. MG1655 || Bacteria (Enterobacterales) || NC\_000913.3 || Syntenic Backbone (1..60000: thrA-thrC, yaaJ, talB, dnaK, dnaJ, nhaA, ribF, ileS ) || 10 || Bacterial (Table 11) || NCBI RefSeq Verified |

| COHORT-02 || *Salmonella enterica* subsp. *enterica* serovar Typhimurium LT2 || Bacteria (Enterobacterales) || NC\_003197.2 || Syntenic Backbone + 31.4 kb Island (1..60000: thrA-ileS ) || 10 || Bacterial (Table 11) || NCBI RefSeq Verified |

| COHORT-03 || 12 Jawed Vertebrates ( Pan, Gorilla, Pongo, Mus, Rattus, Equus, Bos, Balaenoptera, Gallus, Struthio, Columba, Taeniopygia ) || Vertebrata (Mammalia & Aves) || NC\_001643.1 - NC\_013978.1 || Mitochondrial Complex IV Window ( MT-CO1 + MT-CO2 ) || 24 || Vertebrate Mitochondrial (Table 2) || NCBI RefSeq Verified |

| COHORT-04 || Homo sapiens, Pan troglodytes, Macaca mulatta, Mus musculus ||  
Mammalia (Euarchontoglires) || NM\_000558.5 / NM\_000041.4 || Globin ( HBA1, HBA2,  
HBB, MB ) & Apolipoprotein ( APOE, APOA1 ) Panel || 20 || Standard Nuclear (Table  
1) || NCBI RefSeq Verified |

Generated by: NCBI Entrez E-Utilities ( `efetch.fcgi` ) \* CRV Science Labs  
PoliteCacheManager ( `fetch.rs` )

## Discussion & Key Takeaways

This benchmark highlights three practical principles for computational orthology inference and multi-locus phylogenetics:

\* Exact Dynamic Programming Prevents Twilight-Zone Gene Dropout: While double-indexed k-mer seed heuristics ( DIAMOND fast ) provide rapid screening for closely related proteomes, they missed 100% (8/8) of cross-phylum bacterial oriC orthologs ( DnaN and RecF at 24.3%-27.8% amino-acid identity). For deep comparative genomics across phyla or kingdoms, exact Gotoh Smith-Waterman alignment (or DIAMOND --swipe ) is essential to prevent false gene-loss calls and preserve core orthogroups.

\* Length Normalization Equalizes Short and Long Proteins: Raw local alignment bit-scores scale linearly with protein length, creating a 2.77x to 3.06x bias between adjacent genes on the same chromosome or mitogenome (such as IleS vs. RibF or COX1 vs. COX2 ). Applying the Emms & Kelly (2015) log-linear length correction removes this length dependency (residual slope  $b = 0.0000$ ) so orthogroup clustering thresholds reflect sequence conservation rather than polypeptide size.

\* Reciprocal Best-Hit Thresholds Separate Recent In-Paralogs from Ancient Out-Paralogs: Comparing within-species duplication scores against cross-species RBH speciation thresholds provides a transparent, parameter-light test that merges adaptive or concerted in-paralogs (such as *P. aeruginosa* GyrB2 and human HBA1 / HBA2 ) while excluding ancient out-paralogs ( ParC , HBB , MB , and APOA1 ) before supermatrix tree inference.

## 8. References & Primary Literature

- \* Altschul, S. F., & Gish, W. (1996). Local alignment statistics. *Methods in Enzymology* , 266, 460-480. [https://doi.org/10.1016/S0076-6879\(96\)66029-7](https://doi.org/10.1016/S0076-6879(96)66029-7)
- \* Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods* , 12(1), 59-60. <https://doi.org/10.1038/nmeth.3176>
- \* Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* , 18(4), 366-368. <https://doi.org/10.1038/s41592-021-01101-x>
- \* Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* , 16, 157. <https://doi.org/10.1186/s13059-015-0721-2>
- \* Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic Zoology* , 19(2), 99-113. <https://doi.org/10.2307/2412448>
- \* Gotoh, O. (1982). An improved algorithm for matching biological sequences. *Journal*

of Molecular Biology , 162(3), 705-708. [https://doi.org/10.1016/0022-2836\(82\)90398-9](https://doi.org/10.1016/0022-2836(82)90398-9)

\* Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* , 89(22), 10915-10919. <https://doi.org/10.1073/pnas.89.22.10915>

\* Karlin, S., & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences* , 87(6), 2264-2268. <https://doi.org/10.1073/pnas.87.6.2264>

\* Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics* , 39, 309-338. <https://doi.org/10.1146/annurev.genet.39.073003.114725>

\* Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering* , 12(2), 85-94. <https://doi.org/10.1093/protein/12.2.85>

\* Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology* , 147(1), 195-197. [https://doi.org/10.1016/0022-2836\(81\)90087-5](https://doi.org/10.1016/0022-2836(81)90087-5)