Tier 1: Exploratory | Deep-Time Phylogenomics & Molecular Systematics | Signal Rating: 2/5 Beakers | Nov 2, 2026 | 12 min read

# Syntenic Ribosomal Protein RP16 Supermatrix and Dayhoff-6/SR4 Compositional Recoding Benchmark Against IQ-TREE 2

Evaluating amino-acid recoding, Felsenstein supermatrix pruning, compositional long-branch attraction rescue, and CPR/DPANN metabolic auxotrophy profiling across 12 deep-time phylogenomic cohorts (6 Development + 6 Untouched Reserve)

Deep-time evolutionary trees built from standard 20-letter amino-acid alignments often group unrelated organisms together when they share extreme AT-rich genomes or high-temperature amino-acid preferences -- a systematic error known as compositional long-branch attraction (LBA). Across 12 deep-time phylogenomic cohorts partitioned into 6 development and 6 untouched reserve sets (VAL-SET-008), our calibrated Rust ribosomal supermatrix and Dayhoff-6 / Susko-Roger-4 (SR4) recoding engine matches IQ-TREE v2.4.0 log-likelihoods across 20-state, 6-state, and 4-state models to within 0.000038 nats (Pearson r = 1.000000), reduces inter-clade compositional bias by 93.13%, and rescues 10 of 10 deep-time clades from compositional LBA while running 630.9x faster (9.23 ms vs. 5,820.29 ms).

## 1. Biological & Mathematical Context: Compositional Long-Branch Attraction in Deep-Time Phylogenomics

Reconstructing the deepest branches of the Tree of Life -- such as the emergence of Eukaryota from within Asgard Archaea , the placement of the Candidate Phyla Radiation ( CPR / Patescibacteria) and DPANN symbiotic archaea, or the endosymbiotic origins of mitochondria and plastids -- relies on concatenating universally conserved, colinear ribosomal proteins encoded in the S10-spc-alpha syntenic block (the 16-gene RP16 supermatrix: uL2 , uL3 , uL4 , uL5 , uL6 , uL14 , uL16 , uL18 , uL22 , uL24 , uS3 , uS8 , uS10 , uS17 , uS19 , plus uL15 / uS11 ).

However, across billion-year timescales, standard 20-state amino-acid maximum-likelihood models ( AA-20 ) are vulnerable to a pervasive systematic error: compositional long-branch attraction (LBA) . Two distinct evolutionary pressures drive unrelated lineages toward convergent amino-acid frequencies:

* AT-Rich Mutational Drift in Obligate Endosymbionts and Organelles: Host-restricted bacteria (such as Rickettsiales, Buchnera , and CPR ), DPANN nanoarchaea, and intracellular parasites (such as Microsporidia) experience reduced effective population sizes ( N e ) and loss of DNA repair enzymes , shifting their genomes toward extreme AT richness and enriching their proteomes in AT-encoded residues ( F , Y , M , I , N , K ).

* Thermophilic Purine/Hydrophobic Stabilization ( IVYWREL ): Hyperthermophilic bacteria (such as Aquifex and Thermotoga ) and thermophilic Crenarchaeota independently accumulate hydrophobic and charged residues ( I , V , Y , W , R , E , L ) that stabilize ribosomal cores against thermal denaturation.

Because most nucleotide-driven and thermal substitutions occur within biochemically

conservative exchange groups (for example, swapping AT-rich Isoleucine I for GC-rich Valine V , or Lysine K for Arginine R ), physicochemical state recoding collapses the 20 amino acids into reduced alphabets that mask within-class compositional drift while preserving deep between-class replacements:

* Dayhoff-6 Recoding ( k = 6; Hrdy et al., 2004; Embley & Martin, 2006): Groups the 20 amino acids into six Dayhoff PAM exchange classes: 0 = {A, G, P, S, T} (small/neutral), 1 = {D, E, N, Q} (acidic/amide), 2 = {H, K, R} (basic), 3 = {F, Y, W} (aromatic), 4 = {I, L, M, V} (aliphatic hydrophobic), and 5 = {C} ( cysteine ).

* Susko-Roger-4 Recoding ( SR4 , k = 4; Susko & Roger, 2007): Collapses the 20 residues into four broad physicochemical classes optimized to minimize compositional heterogeneity: 0 = {A, G, N, P, S, T} , 1 = {C, H, W, Y} , 2 = {D, E, K, Q, R} , and 3 = {F, I, L, M, V} .

## 2. Validation Dataset & Cohort Provenance Tracking ( VAL-SET-008-RP16-DEEP-PHYLOGENOMICS-12COHORT )

As part of our structured Validation Dataset & Provenance Registry ( validation-datasets-registry.json ), every Lab Discovery explicitly documents the origin, partition structure, and replication scope of its underlying datasets across our three research tiers:

* Tier 1 (Exploratory): Seeks out and explores a mix of new, classic, or previously unseen datasets, or applies an existing validation cohort in a new analytical way as part of a calibration sweep.

* Tier 2 (Validation): Uses curated validation sets ( VAL-SET-001 through VAL-SET-008 ) that have been repeatedly replicated and validated across multiple benchmarks with locked Development vs. Untouched Reserve metadata.

* Tier 3 (Scientific): Combines new and validated datasets to test new hypotheses and establish a new scientific interpretation, conclusion, or cross-scale synthesis.

This Tier 1: Exploratory benchmark establishes and documents VAL-SET-008-RP16-DEEP-PHYLOGENOMICS-12COHORT , a 12-cohort deep-time phylogenomic and metabolic completeness suite (98 taxa across 12 concatenated 64-residue focal RP16 alignments, plus 10-pathway biosynthetic presence/absence matrices) anchored to published phylogenomic studies (Hug et al., 2016; Spang et al., 2015; Zaremba-Niedzwiedzka et al., 2017; Ponce-Toledo et al., 2017):

| Cohort ID | | Partition Split | | Evolutionary Clade & Phylogenomic Test Regime | | Taxa | | Focal RP16 Sites | | Compositional LBA Trap Tested | | Primary Literature & NCBI / GTDB Reference Anchor |

| DEV-01 | | Development | | Asgard Archaea (Loki/Thor/Heimdall) + Eukaryota (2D vs. 3D) | | 8 | | 64 | | 3-Domain Thermophilic Archaea vs. 2-Domain Eocyte | | Spang et al. (2015) Nature ; Zaremba-Niedzwiedzka et al. (2017) |

| DEV-02 | | Development | | Candidate Phyla Radiation ( CPR / Patescibacteria) Monophyly | | 8 | | 64 | | AT-Rich Symbiont Polyphyly vs. Monophyletic CPR | | Brown et al. (2015) Nature ; Hug et al. (2016) Nat. Microbiol. |

| DEV-03 | | Development | | DPANN Symbiotic Archaea ( Nanoarchaeum / Parvarchaeota)

| | 8 | | 64 | | Fast-Evolving AT-Rich Outgroup Attraction | | Rinke et al. (2013) Nature ; Castelle et al. (2015) |

| DEV-04 | | Development | | Hyperthermophilic Bacteria ( Aquifex / Thermotoga ) | | 8 | | 64 | | Thermophilic IVYWREL Convergence with Crenarchaeota | | Boussau et al. (2008) Nature ; Williams et al. (2013) |

| DEV-05 | | Development | | Alphaproteobacteria & Mitochondrial Endosymbiotic Origin | | 8 | | 64 | | AT-Rich Rickettsiales + Mitochondrial LBA | | Wang & Wu (2015); Martijn et al. (2018) Nature |

| DEV-06 | | Development | | Symmetrically Permuted Within-Class Drift Control ( 0 Gate) | | 8 | | 64 | | Pure Within-Dayhoff-Class Shuffle (Deltaln L D6 = 0) | | Susko & Roger (2007) Mol. Biol. Evol. Null Control |

| RES-01 | | Untouched Reserve | | Primary Plastid Endosymbiosis ( Gloeomargarita + Chloroplast) | | 8 | | 64 | | AT-Rich Plastid + Non-Photosynthetic Melainabacteria LBA | | Ponce-Toledo et al. (2017) Curr. Biol. |

| RES-02 | | Untouched Reserve | | Microsporidia ( Encephalitozoon / Nosema ) Fungal Placement | | 9 | | 64 | | Basal "Archezoa" AT-Rich Long-Branch Artifact | | Hirt et al. (1999) PNAS ; Capella-Guti rrez et al. (2012) |

| RES-03 | | Untouched Reserve | | Insect Endosymbiont Gammaproteobacteria ( Buchnera / Wigglesworthia ) | | 9 | | 64 | | AT-Rich Tenericutes ( Mycoplasma ) Convergence | | Husnik et al. (2011); Williams et al. (2010) |

| RES-04 | | Untouched Reserve | | Non-Bilaterian Metazoan Base (Porifera / Ctenophora / Cnidaria) | | 8 | | 64 | | Distant Choanoflagellate / Ichthyosporean Outgroup LBA | | Feuda et al. (2017) Curr. Biol. ; Simion et al. (2017) |

| RES-05 | | Untouched Reserve | | TACK Archaeal Superphylum (Thaum-/Aig-/Cren-/Korarchaeota) | | 8 | | 64 | | Mesophilic Thaumarchaeota vs. Hyperthermophilic Crenarchaeota | | Guy & Ettema (2011) Trends Microbiol. |

| RES-06 | | Untouched Reserve | | Reserve Symmetrically Permuted Within-Class Control ( 0 Gate) | | 8 | | 64 | | Pure Within-Dayhoff-Class Reserve Decoy (Deltaln L D6 = 0) | | Reserve Specificity Gate Control |

Generated by: NCBI Entrez E-Utilities * CRV Science Labs PoliteCacheManager * Native 4-Stage RP16 Syntenic Supermatrix, Dayhoff-6 / SR4 Compositional Recoder & Candle Felsenstein Pruning Engine

## 3. Baseline Audit: Replacing Hardcoded Placeholders with Exact Multi-Alphabet Felsenstein Pruning

Before wiring our universal-rp16-supermatrix module into multi-stage workflows, we audited our initial prototype against live external IQ-TREE v2.4.0 (running -st AA -m POISSON+FQ on 20-state alignments and -st MORPH -m MK -keep-ident on 6-state Dayhoff-6 and 4-state SR4 recoded alignments). That audit exposed two limitations in the initial prototype:

* Hardcoded Scalar Log-Likelihoods: The initial prototype in our tool registry returned a static summary constant (−148,924.62 nats) rather than executing Felsenstein's tree-pruning algorithm over actual amino-acid or recoded character

matrices, producing a 148,141.6-nat discrepancy on DEV-01 .

* Unrecoded 20-State Vulnerability to Within-Class Compositional Drift: When evaluated under standard 20-state Equal-Input ( POISSON+FQ ) amino-acid likelihood without Dayhoff-6 or SR4 recoding, all 10 empirical deep-time cohorts favored the compositionally converged LBA trap tree over the true biological clade by Deltaln L AA20 = −51.44 to −244.99 nats.

We upgraded the native Rust engine to perform (1) exact Dayhoff-6 ( k = 6) and SR4 ( k = 4) character recoding, (2) pairwise Jensen-Shannon compositional divergence ( JSD ) quantification before and after recoding, (3) closed-form k -state Jukes-Cantor / Equal-Input transition probability tensor pruning ( P ij ( t ) = pi j + (delta ij − pi j ) e −u t ) via candle-core , and (4) 10-pathway amino-acid/cofactor metabolic completeness scanning to separate obligate CPR and DPANN symbionts from free-living bacteria and archaea .

## 4. Head-to-Head Benchmark Across Development and Untouched Reserve Cohorts

Across all 12 cohorts (6 Development + 6 Untouched Reserve), the calibrated native Rust engine reproduces IQ-TREE v2.4.0 log-likelihoods across all three alphabets ( AA-20 , Dayhoff-6 , and SR4 ) to four decimal places, while reversing compositional LBA on 10/10 empirical clades and rejecting both 0 within-class permuted negative controls:

| Cohort | | Split | | Native AA20 ln L | | IQ-TREE AA20 ln L | | Native Dayhoff-6 ln L | | IQ-TREE Dayhoff-6 ln L | | Native SR4 ln L | | IQ-TREE SR4 ln L | | Deltaln L AA20 (True−LBA) | | Deltaln L Day6 (True−LBA) | | JSD Bias Reduction (%) | | Symbiont Metabolic (%) | | Concordant |

| DEV-01 | | Development | | −783.0438 | | −783.0438 | | −301.8976 | | −301.8976 | | −261.8651 | | −261.8651 | | −76.9101 | | +104.9958 | | 90.59 | | 0.00 | | YES |

| DEV-02 | | Development | | −728.0698 | | −728.0698 | | −280.2544 | | −280.2544 | | −244.0087 | | −244.0087 | | −70.0936 | | +90.3294 | | 91.59 | | 26.67 | | YES |

| DEV-03 | | Development | | −732.6117 | | −732.6118 | | −286.1882 | | −286.1882 | | −249.8424 | | −249.8424 | | −65.8834 | | +88.0002 | | 91.59 | | 23.75 | | YES |

| DEV-04 | | Development | | −750.1609 | | −750.1609 | | −277.5942 | | −277.5942 | | −241.8530 | | −241.8530 | | −51.4419 | | +122.4783 | | 91.00 | | 0.00 | | YES |

| DEV-05 | | Development | | −747.8681 | | −747.8681 | | −280.3068 | | −280.3068 | | −244.4696 | | −244.4697 | | −73.8472 | | +100.3193 | | 91.99 | | 29.44 | | YES |

| DEV-06 | | Development | | −1352.6160 | | −1352.6161 | | −249.9731 | | −249.9731 | | −221.7693 | | −221.7693 | | 0.0000 | | 0.0000 | | 99.97 | | 0.00 | | YES |

| RES-01 | | Reserve | | −748.4302 | | −748.4302 | | −281.1993 | | −281.1993 | | −245.2155 | | −245.2155 | | −73.2437 | | +100.2724 | | 91.99 | | 30.83 | | YES |

| RES-02 | | Reserve | | −759.5584 | | −759.5585 | | −314.2564 | | −314.2564 | | −277.4971 | | −277.4972 | | −68.0171 | | +88.4644 | | 92.31 | | 33.67 | | YES |

| RES-03 | | Reserve | | −925.5726 | | −925.5726 | | −302.0240 | | −302.0240 | | −265.6413 | | −265.6412 | | −244.9884 | | +90.3556 | | 92.31 | | 29.00 | | YES |

| RES-04 | | Reserve | | −730.6192 | | −730.6193 | | −285.4413 | | −285.4413 | | −249.0513 | | −249.0513 | | −65.7936 | | +86.9607 | | 91.59 | | 0.00 | | YES |

| RES-05 | | Reserve | | −731.8415 | | −731.8415 | | −273.1783 | | −273.1783 | | −237.6145 | | −237.6145 | | −104.7319 | | +72.7814 | | 92.62 | | 48.33 | | YES |

| RES-06 | | Reserve | | −1389.4308 | | −1389.4308 | | −234.0195 | | −234.0195 | | −206.1329 | | −206.1329 | | 0.0000 | | 0.0000 | | 99.97 | | 0.00 | | YES |

Generated by: Native 4-Stage ML Tree Builder (Multi-Distance BIONJ + Exact Yang Gamma4 + Inside-Outside Pulley NNI) * IQ-TREE v2.4.0 (Minh et al.) * Native 4-Stage RP16 Syntenic Supermatrix, Dayhoff-6 / SR4 Compositional Recoder & Candle Felsenstein Pruning Engine * IQ-TREE v2.4.0 (-st AA POISSON+FQ & -st MORPH MK -keep-ident: Minh et al. 2020)

## 5. Focal Sequence Alignment: How Dayhoff-6 Recoding Masks AT/GC Drift While Preserving Deep Clade Synapomorphies

To illustrate why Dayhoff-6 recoding reverses compositional LBA in DEV-01 (Asgard Archaea + Eukaryota) and RES-01 (Primary Plastid Endosymbiosis), Figure 1 shows a 24-column focal window from the concatenated uL2-uL14-uS8 ribosomal core. In the raw 20-letter alignment, AT-rich lineages substitute I/M/K/N/F where GC-rich and thermophilic lineages carry V/L/R/D/Y -- creating artificial pairwise identity between unrelated AT-rich or thermophilic taxa. Once mapped to Dayhoff-6 classes ( 0..5 ), those within-class swaps collapse to identical states ( 4 for {I,L,M,V} , 2 for {H,K,R} , 1 for {D,E,N,Q} , 3 for {F,Y,W} ), exposing the genuine deep synapomorphies uniting Asgard Archaea with Eukaryota:

Generated by: Native Gotoh 3-State Affine-Gap & 9-mer Anchor Multiple Sequence Aligner * Native 4-Stage RP16 Syntenic Supermatrix, Dayhoff-6 / SR4 Compositional Recoder & Candle Felsenstein Pruning Engine

## 6. Universal RP16 Phylogenomic Cladogram & Newick Topology

Combining the Dayhoff-6 recoded RP16 supermatrix across our deep-time reference cohorts recovers the 2-Domain Eocyte tree (Eukaryota nesting sister to Heimdallarchaeota within Asgard Archaea, Deltaln L Day6 = +105.00 nats over the 3-Domain LBA topology), monophyletic CPR (Patescibacteria) and DPANN symbiotic

clades characterized by severe de novo biosynthetic auxotrophy (23.8%-26.7% pathway completeness vs. 87.5%-95.0% in free-living relatives), and the endosymbiotic placements of mitochondria within Alphaproteobacteria (+100.32 nats) and plastids sister to Gloeomargarita (+100.27 nats):

Generated by: Native K2P / TN93 Distance, Neighbor-Joining & ABGD Systematics Engine * Native 4-Stage RP16 Syntenic Supermatrix, Dayhoff-6 / SR4 Compositional Recoder & Candle Felsenstein Pruning Engine

Generated by: Native K2P / TN93 Distance, Neighbor-Joining & ABGD Systematics Engine * Native 4-Stage RP16 Syntenic Supermatrix, Dayhoff-6 / SR4 Compositional Recoder & Candle Felsenstein Pruning Engine

## 7. Aggregate Summary: Development vs. Untouched Reserve Concordance

| Benchmark Metric | | Gen-1 Uncalibrated Baseline | | Development Split (6 Cohorts) | | Untouched Reserve Split (6 Cohorts) | | Combined 12-Cohort Concordance |

| Mean |Deltaln L | on 20-State AA-20 ( POISSON+FQ ) | | 148,075.56 nats | | 0.000034 nats | | 0.000027 nats | | 0.000030 nats ( r = 1.000000) |

| Mean |Deltaln L | on 6-State Dayhoff-6 ( MK ) | | N/A (Not Implemented) | | 0.000023 nats | | 0.000030 nats | | 0.000027 nats ( r = 1.000000) |

| Mean |Deltaln L | on 4-State SR4 ( MK ) | | N/A (Not Implemented) | | 0.000025 nats | | 0.000038 nats | | 0.000031 nats ( r = 1.000000) |

| Compositional JSD Bias Reduction (Dayhoff-6 vs. AA-20) | | 0.00% | | 92.79% | | 93.46% | | 93.13% |

| Empirical Clade LBA Rescue Rate (True vs. LBA Trap) | | 0.0% (0/10 under AA-20) | | 100.0% (5/5) | | 100.0% (5/5) | | 100.0% (10/10 Rescued) |

| 0 Within-Class Permuted Negative Controls Rejected | | 0.0% | | 100.0% ( DEV-06 ) | | 100.0% ( RES-06 ) | | 100.0% (2/2 Rejected) |

| Total Wall-Clock Execution Time (ms) | | 1.10 ms (Static) | | 5.06 ms | | 4.16 ms | | 9.23 ms vs. 5,820.29 ms (630.9x) |

Generated by: Native 4-Stage ML Tree Builder (Multi-Distance BIONJ + Exact Yang Gamma4 + Inside-Outside Pulley NNI) * IQ-TREE v2.4.0 (Minh et al.) * Native 4-Stage RP16 Syntenic Supermatrix, Dayhoff-6 / SR4 Compositional Recoder & Candle Felsenstein Pruning Engine * IQ-TREE v2.4.0 (-st AA POISSON+FQ & -st MORPH MK -keep-ident: Minh et al. 2020)

## 8. Signal Rating Assessment (2/5 Beakers -- Technical Validation)

This study is rated 2/5 Beakers ( Technical Validation) under Tier 1: Exploratory scope. The calibrated native Rust engine achieves four-decimal-place log-likelihood parity with IQ-TREE v2.4.0 across 20-state, 6-state Dayhoff, and 4-state SR4 alphabets on both the 6 development cohorts and 6 untouched reserve cohorts ( VAL-SET-008 ). Because each cohort uses a compact 64-residue focal supermatrix window evaluated under homogeneous Equal-Input/Mk rate matrices rather than full-length 2,500-residue profile-mixture models ( LG+C60+F+G4 ), we classify this report as a 2 technical tool validation rather than a 4 - 5 genome -scale

phylogenomic synthesis.

## Discussion & Key Takeaways

Benchmarking our native RP16 supermatrix module against IQ-TREE v2.4.0 illustrates why deep-time phylogenomic workflows benefit from pairing standard 20-state amino-acid models with reduced physicochemical alphabets ( Dayhoff-6 and SR4 ). When distantly related clades share strong AT-rich mutational drift or hyperthermophilic IVYWREL amino-acid enrichment, unrecoded 20-state likelihoods strongly favor convergent long-branch attraction topologies (Deltaln L AA20 = −51.4 to −245.0 nats). Collapsing within-class exchanges removes 93.13% of pairwise compositional Jensen-Shannon divergence and restores the established biological topology across all 10 empirical deep-time cohorts (Deltaln L Day6 = +72.8 to +122.5 nats).

By registering VAL-SET-008-RP16-DEEP-PHYLOGENOMICS-12COHORT in our public Validation Dataset Registry alongside locked Development and Untouched Reserve splits, this module is now ready to chain directly with upstream Plan-7 profile HMM marker extraction and downstream coalescent and morphological tip-dating pipelines.

## References & Primary Sources

* [1] Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., ... & Banfield, J. F. (2016). A new view of the tree of life. Nature Microbiology, 1(5), 16048. https://doi.org/10.1038/nmicrobiol.2016.48

* [2] Zaremba-Niedzwiedzka, K., Caceres, E. F., Saw, J. H., B ckstr m, D., Juzokaite, L., Vancaester, E., ... & Ettema, T. J. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature, 541(7637), 353-358. https://doi.org/10.1038/nature21031

* [3] Susko, E., & Roger, A. J. (2007). On reduced amino acid alphabets for phylogenetic inference. Molecular Biology and Evolution, 24(9), 2139-2150. https://doi.org/10.1093/molbev/msm144

* [4] Ponce-Toledo, R. I., Deschamps, P., L pez-Garc a, P., Zivanovic, Y., Benzerara, K., & Moreira, D. (2017). An early-branching freshwater cyanobacterium at the origin of plastids. Current Biology, 27(3), 386-391. https://doi.org/10.1016/j.cub.2016.11.056

* [5] Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Molecular Biology and Evolution, 37(5), 1530-1534. https://doi.org/10.1093/molbev/msaa015