Rapid evolutionary innovation during an Archaean genetic expansion

SUPPLEMENTARY INFORMATION

Contents:

1.1. *AnGST* algorithm

1.1.1. Overview

We developed a phylogenomic method that we named *AnGST* (Analyzer of Gene and Species Trees), which "reconciles" any observed differences between a gene tree and a reference tree (species tree) by inferring a minimal set of evolutionary events, including horizontal gene transfer (HGT), gene duplication (DUP), gene loss (LOS), speciation (SPC) and exactly one gene birth or genesis event (GEN). Each event type is assigned a unique cost, and the overall sum of costs associated with a reconciliation is minimized (*i.e.*, we use a generalized parsimony criterion). We address previously described shortcomings of similar parsimony-based models of host-parasite evolution[30] by accounting for phylogenetic uncertainty (using a new approach described below) and directly estimating event costs from our large dataset. We divide the gene-tree/species-tree reconciliation process into two components:

- The **basic reconciliation** step assumes a known gene tree and species tree and identifies the set of evolutionary events (HGT, DUP, LOS, SPC, GEN) needed to explain any discordance between the trees
- The **tree amalgamation** step accounts for gene tree uncertainty by incorporating tree construction into the reconciliation process: multiple gene tree bootstraps are provided to *AnGST* and the algorithm retains and combines bootstrap subtrees which yield the most conservative reconciliation consistent with the sequence data.

The estimation of **event costs** from the input data is based on reducing large fluctuations in ancient genome sizes. This method is presented in Methods Section 1.2 together with a **sensitivity analysis** for the resulting parameters.

The *AnGST* software package is implemented in the Python programming language and will be freely available for download from the following website: http://almlab.mit.edu/ALM/Software/Entries/2009/3/29_AnGST.html.

1.1.2. Basic Reconciliation Algorithm

First assumptions

The **basic reconciliation** step requires a rooted, strictly bifurcating gene tree G and species tree S. Each tree is composed of a set of nodes linked to one another by a set of connecting edges. We assume that each node g in G can be mapped to a node s in S, a mapping we abbreviate as g:s. This mapping describes which (extant or ancestral) genome hosted a given (extant or ancestral) gene copy. Maps are known with certainty for extant genes, but must be inferred for ancestral gene copies.

Algorithm explanation

Our goal in gene/species tree reconciliation is to recover the optimal set of evolutionary events that explain any topological discordance between the gene and species trees. A brute-force search through all possible evolutionary histories is intractable, as the number of possible histories grows exponentially with increasing tree size[31]. However, for a given gene and species tree pair, there are only $|S|$ possible mappings for the root node of the gene tree, g_r. If the optimal reconciliation is already known for each possible mapping g_r:s_r, where s_r is a node in S, a new outgroup for the gene tree can be added (making g_r a child of the new root node g_n), and optimal reconciliations for the larger gene tree can be quickly computed using the following method:

I. For each possible pair of mappings (g_r:s_r, g_n:s_n) where s_r and s_n are nodes in S

A. Choose the most parsimonious explanation for how a gene copy in s_n descended into s_r.

B. Concatenate this history to the known optimal reconciliation for g_r:s_r, to produce the optimal reconciliation for the (g_r:s_r, g_n:s_n) pair

II. Identify optimal reconciliations for each mapping g_n:s_n by selecting the minimal overall reconciliation cost associated with (g_r:s_r, g_n:s_n) as s_r is varied over the nodes of S.

Using the above method, the reconciliation problem can be formulated in a dynamic programming framework, yielding computational complexity that is a polynomial-time function of gene tree size. The *AnGST* program implements this algorithm as a post-fix traversal of the gene tree. At each node, reconciliations from child subtrees are combined in "mini-reconciliations," which explain how the gene copy at g coalesced from two child copies c_1 and c_2 (*i.e.*, whether HGT, speciation, or duplication occurred), assuming the mappings g:s, c_1:s_1, and c_2:s_2. This is repeated for each s, s_1, and $s_2 \in S$. Mini-reconciliations return optimal duplication-loss or HGT scenarios if s is the last common ancestor of s_1 and s_2, or if s is identical to either s_1 or s_2. All other combinations of s, s_1, and s_2, yield mini-reconciliations that we refer to as "complex scenarios." We include these scenarios in the pseudocode below to aid understanding of basic reconciliation design, but we do not provide a method for their solution since complex scenarios can be safely ignored without loss of reconciliation optimality (see *Running Time* discussion below). If g is a leaf node, mini-reconciliations are unnecessary since the true mapping from g to the species tree is known. Once all combinations have been evaluated, we retain the optimal reconciliation associated with <u>each</u> possible mapping of g to the species tree. Pseudocode for the reconciliation algorithm is provided on the following page in Python style.

Pseudocode:

% Main %

- Reconcile(gene_tree.root)

% Methods %

- **define Reconcile(node):**
  - child_1, child_2 = ChildNodes(node) **%strictly bifurcating tree**
  - if child_1 AND child_2 are null**: %is a leaf node**
    - for node_map in AllNodes(species_tree)**:**
      - if node_map is KnownHostGenome(node)**:**
        - node.reconciliation_cost(node_map) = 0 **%correct answer is known for leaves**
      - else**:**
        - node.reconciliation_cost(node_map) = maxint
    - return
  - Reconcile(child_1) **%post-fix traversal**
  - Reconcile(child_2)
  - for node_map in AllNodes(species_tree)**: %try all possible hosts for ancestor**
    - for child_1_map in AllNodes(species_tree)**: %try all possible hosts for children**
      - for child_2_map in AllNodes(species_tree)**:**
        - events = MiniReconcile(node_map, child_1_map, child_2_map)
        - prior_events_1 = child_1.reconciliation_cost(child_1_map)
        - prior_events_2 = child_2.reconciliation_cost(child_2_map)
        - overall_cost = Cost(events + prior_events_1 + prior_events_2)
        - cost_matrix(node_map, child_1_map, child_2_map) = overall_cost
  - for node_map in AllNodes(species_tree)**:**
    - node.reconciliation_cost(node_map) = Min(cost_matrix(node_map, **:, :**))
  - return

- **define MiniReconcile(node_map, child_1_map, child_2_map):**
  - **% compute DupLoss scenarios**
  - if node_map is ancestral to child_1_map AND child_2_map**:**
    - if node_map is last_common_ancestor of child_1_map AND child_2_map**:**
      - **%%% See Page**[32] **for DupLoss pseudocode**
      - duploss_events = DupLoss(node_map, child_1_map, child_2_map)
    - else**:**
      - duploss_events = ComplexScenario()
      - **%ComplexScenario() not implemented -- see Methods Section 1.1.2 *Running Time* discussion for explanation**
  - else**:**
    - duploss_events = maxint **% impossible to reconcile with only dup-loss**
  - **% compute HGT scenarios**
  - if node_map is child_1_map**:**
    - hgt_events = {HGT from node_map to child_2_map}
  - elif node_map is child_2_map**:**
    - hgt_events = {HGT from node_map to child_1_map}
  - else**:**
    - hgt_events = ComplexScenario()
  - return MinCost(hgt_events,duploss_events)

***An example reconciliation*:**

An *AnGST* reconciliation of two simple, but discordant, gene and species trees is provided in Supplementary Fig. 1. Here, we assume that we know the true mappings from the leaves in G to S: $\{g_1{:}s_A, g_2{:}s_C, g_3{:}s_B\}$. Because *AnGST* uses a post-fix traversal of G and the mapping of G's leaves to S is trivial, we first investigate how g_4 is mapped to nodes in S. We initialize the algorithm by assigning infinite reconciliation cost to leaf mappings which deviate from the known leaf mappings (e.g. $g_1{:}s_B$); thus, there is only one valid mapping for g_1 and g_2.

In Scenario α, g_4 is mapped to s_A ($g_4{:}s_A$) and we infer one HGT event using the mini-reconciliation algorithm (since g_4 is mapped to the same lineage as one of its child nodes). Similarly, if we consider $g_4{:}s_C$, we infer one HGT from s_C to s_A (Scenario β). In the case of $g_4{:}s_E$ (Scenario γ), g_4 is mapped to the LCA of s_A and s_E and a duplication-loss scenario is invoked by the mini-reconciler. Other more complex scenarios exist (*e.g.,* $s_4{:}s_D$), but these can be ignored without affecting overall reconciliation optimality (see *Running Time* section below). Once optimal reconciliations have been found for each possible $g_4{:}s$ mapping, *AnGST* recurses to g_5 and repeats the process. In the next mini-reconciliation, there are multiple valid $g_4{:}s$ mappings. Thus *AnGST* must iterate through prospective mappings for both g_5 and g_4 (although for the sake of illustrative simplicity, we only enumerate a fraction of these scenarios).

In the first mapping shown for g_5 ($g_5{:}s_D$,$g_4{:}s_A$.$g_3{:}s_B$), there is 1 SPC (since s_A and s_B are direct vertical descendants of s_D) and this cost is added to the 1 HGT already inferred in Scenario α, which resulted in $g_4{:}s_A$. For the combination ($g_5{:}s_C$,$g_4{:}s_C$.$g_3{:}s_B$), a cost of 1 HGT (because g_5 and g_4 share the same mapping) is added to the cost for Scenario β. The last mapping shown is ($g_5{:}s_E$,$g_4{:}s_E$.$g_3{:}s_B$). A mini-reconciliation that posits HGT will imply forward-

in-time gene transfers -- an evolutionary event we do not allow (see *Temporal constraints on HGT* below). Instead, a DUP in s_E and subsequent losses among s_A and s_C are needed to correctly explain the mapping of s_5:s_E, s_4:s_E, and g_3:s_B. The g_5 mapping that leads to the optimal reconciliation is a function of the chosen evolutionary event costs. With a cost structure: C_{SPC}=0, C_{HGT}=1, C_{LOS}=2, C_{DUP}=3, the optimal mappings would be g_5:s_D, g_4:s_A, and the associated reconciliation would be a GEN event at s_D, followed by SPC at s_D, and an HGT from s_A to s_C. However, if C_{SPC}=0, C_{HGT}=10, C_{LOS}=2, C_{DUP}=3, the optimal mapping would be g_5:s_E, g_4:s_E, and the associated reconciliation would be an initial GEN event at s_E, followed by a DUP in s_E, 2 SPCs each at s_E and s_D, and LOS in lineages s_A, s_B, and s_C.

Running time

O($|G|*|S|^3$) is an upper bound on run-time complexity of *AnGST*, where $|S|$ and $|G|$ are the number of nodes in those trees, respectively. Running times can be significantly reduced without loss of reconciliation optimality, however, with a simple speedup. When performing mini-reconciliations on all combinations of g:s, c_1:s_1, and c_2:s_2 for s, s_1, and $s_2 \in S$, any complex scenario (s is not s_1 or s_2, and s is not the last common ancestor of both s_1 or s_2) will require at least two HGT (one to s_1 and another to s_2), or one HGT to the last common ancestor of s_1 and s_2 followed by a duplication-loss scenario originating at that ancestor. These more complex scenarios will therefore always be suboptimal with respect to non-complex scenarios and their evaluation can be skipped during the reconciliation process. The resulting reduction in mapping search space lowers *AnGST* run-time complexity to O($|G|*|S|^2$). When temporal constraints on HGT are enforced (see below), this speedup cannot be fully exploited, as nodes ancestral to s_1 and s_2 are potentially optimal values for s in HGT scenarios.

In practice, on 3.0Ghz single-cores with access to 8GB of memory, an *AnGST* run reconciling 100 bootstrap trees from one gene family against a reference tree of 100 species would take roughly: 0.1 minutes for gene trees with ~10 leaves, 4 minutes for gene trees with ~50 leaves, 13 minutes for gene trees with ~100 leaves, 27 minutes for gene trees with ~150 leaves, 37 minutes for gene trees with ~200 leaves.

Temporal constraints on HGT

If provided a chronogram as a reference tree, *AnGST* will restrict the set of possible inferred gene transfers to only those between contemporaneous lineages. This feature eliminates the possibility of inferring multiple HGT events which are chronologically impossible[33]. Any non-zero chronological overlap is sufficient to allow transfers. But, if a gene transfer is inferred from node s_1 to node s_2, subsequent transfers of the gene copy in s_2 may only occur with lineages which exist during the range $T_1 \cap T_2$, where T_1 and T_2 are the times spanned by the parent edges of s_1 and s_2, respectively. A feature enabling transfers forward in time (which may represent "phantom transfers" from unsampled taxa[34]) has been built into *AnGST*, but remains off by default and was not used in our analyses.

Gene tree rooting

Bootstrap trees are assumed to be unrooted. All possible rootings of these bootstrap trees are evaluated during the reconciliation process. The resulting gene tree is rooted on the branch that results in the overall lowest reconciliation score.

1.1.3. Bootstrap tree amalgamation

Errors or uncertainty in gene phylogenies can lead to the inference of spurious macroevolutionary events[35] and is a particular concern for deeply branching phylogenies[36].

AnGST resolves uncertainty by incorporating reconciliation into the tree-building process: the tree with the lowest reconciliation cost is chosen from a large ensemble of trees consistent with the sequence data. To generate an ensemble of suitable trees, *AnGST* considers the set of all trees that contains only bipartitions observed in a set of input trees, which we generate with non-parametric bootstrapping. Thus, *AnGST* typically outputs "chimeric" trees that do not match any of the input bootstrap trees exactly, although every bipartition in the *AnGST* tree occurs in at least one of the bootstraps. In simulations, we observe these trees to be significantly more accurate than trees based on sequence likelihood alone, although they generally have lower likelihood (see *Chimeric tree fidelity* below and Supplementary Fig. 3). Any number of bootstrap trees can be used, but we found limited increase in accuracy in simulated data as a result of using more than 10 (data not shown).

We implement this approach in the following manner (see Supplementary Fig. 2 for an example). Given n gene tree bootstraps $\{G_1, G_2, \ldots G_n\}$ and a reference tree S, *AnGST* will begin the basic reconciliation algorithm starting on tree G_1. Each time *AnGST* evaluates an internal node g of G_1, it also evaluates the set of internal nodes $I = \{g_1, g_2, \ldots g_k\}$ in other bootstrap gene trees that define the same bipartition as g. The optimal reconciliation at this node is the lowest scoring scenario/topology observed in any of the bootstrap trees. That is, a distinct solution is computed for each possible mapping (g_i:s for $g_i \in I$, $s \in S$), and only the best solution is retained for each value of s. These $|S|$ optimal mappings and their reconciliations are subsequently shared across all the nodes in I. This last step creates "chimeric" gene trees, as the reconciliation at g in G_1 may now refer to a topology found in bootstrap G_i.

1.1.4. Benchmarking accuracy

We used simulations to benchmark the performance of *AnGST*. Ten independent gene trees births were simulated on each of the 199 extant and ancestral lineages of the reference tree. A simple Poisson statistics-based model of HGT, DUP, and LOS was used to generate random gene histories and associated gene trees; the average simulated gene family underwent 0.21 HGT, 0.05 DUP, 0.76 SPC, and 0.26 LOS per extant gene copy. (For comparison, our analysis of the COG dataset inferred 0.29 HGT, 0.10 DUP, 0.83 SPC, and 0.27 LOS per extant gene copy.) Synthetic amino acid sequences were generated using these simulated trees and the SeqGen software (v.1.3.2)[37]. Trees were reconstructed from the synthetic sequences using either the BIONJ algorithm (implemented in PhyML), PhyML (v.2.4.5)[38], or *AnGST* via 100 PhyML-generated bootstrap topologies (see Methods Section 1.4 for PhyML parameters). A subset of 75 gene families were used to learn costs for HGT and DUP (see Methods Section 1.2). A cost combination of C_{HGT}=4, C_{DUP}=3 minimized genome size flux using this gene family subset (compared to C_{HGT}=3, C_{DUP}=2 learned for the COG dataset).

Chimeric tree fidelity

Following reconciliation, nodes deep in the interior of the resultant gene tree can contain topologies not found in any of the inputted bootstraps (although all possible bipartitions of these subtrees will exist in at least one of the bootstraps). Thus, the potential search space of topologies is vast. We tested the fidelity of the chimeric gene trees learned during the reconciliation process using the Robinson-Foulds (RF) statistic[39], which measures the number of bipartitions not shared by a pair of trees. A 0 RF score indicates perfect concordance (all bipartitions of the candidate and reference tree are identical) and increasing RF scores denote higher phylogenetic discordance. Analysis of the 225 gene trees with a minimal level of

complexity (more than 10 leaves) demonstrates that *AnGST* trees are significantly more accurate than trees generated by BIONJ ($p=9.1\times10^{-8}$ Wilcoxon rank sum test) or PhyML alone ($p=1.8\times10^{-2}$ Wilcoxon rank sum test). Interestingly, this increase in topological accuracy comes with a likelihood tradeoff in comparison to the PhyML algorithm ($p=2.7\times10^{-39}$ Wilcoxon rank sum test). As an aside, we note that the PhyML likelihoods in these analyses are in agreement with previous simulations which showed PhyML capable of constructing trees with higher likelihood than the true topologies[38].

Inferred birth date accuracy

We benchmarked the accuracy of gene family birth dates predicted by *AnGST* using the 747 synthetic gene families that included more than one extant gene copy. A comparison of inferred birth events and the simulated age of birth events is shown in Supplementary Fig. 4A. There is a strong correlation between inferred and simulated ages (0.88) and 76% of births are predicted to within 250 My of their simulated age. These results are especially promising given the noisy processes (sequence simulation and phylogenetic inference) separating simulation of a gene family and its reconciliation. Moreover, we see no obvious evidence of inference bias which may lead to the false inference of a birth spike. Direct comparison of birth counts during the Archean Expansion (2.9-3.3 Ga) to simulated births during the same period (Supplementary Fig. 4B) did not show a bias towards over-counting births. We did, however, observe a bias toward gene birth prior to the Archean Expansion, suggesting that our set of very ancient genes (born prior to 3.3 Ga) may be inflated.

1.2. Fitting *AnGST* parameters

1.2.1. Minimizing genome size flux

We address the problem of assigning the costs to each event type in a manner similar to some previous studies[24,25,40]: we use predictions of ancestral genome sizes to constrain the costs C_{DUP} and C_{HGT} (these are the only free parameters as we can assume C_{LOS}=1 and C_{SPC}=0 without loss of generality). However, we chose to minimize differences in genome size between parent and child nodes (a metric we refer to as genome size flux) rather than constraining overall genome size over time for two reasons: first, gene acquisition rates may not have been constant over time and ancient genomes may have been smaller (or larger) than modern day genomes[40]; second, the extinction of ancient gene families would lead to a trend of smaller inferred ancestral genome sizes at earlier times even if actual genome sizes were constant. A grid search of cost space showed genome size flux to be minimized at: C_{HGT}=3 and C_{DUP}=2 (Supplementary Figs. 5A, 6).

1.2.2. Sensitivity analysis

We investigated the extent to which the high fraction of overall gene birth detected during the Archean Gene Expansion was dependent on model parameters (Supplementary Fig. 5B). Gene birth patterns were invariant over a broad range of C_{DUP}. Gene birth from 2.8-3.4 Ga dissipated only at low C_{HGT} values. However, this regime of C_{HGT} resulted in unrealistic genome size distributions: ancestral genomes were much smaller than present day ones, and most genes were predicted to have been born on terminal branches and spread *via* HGT.

1.3. Reference tree construction

1.3.1. Building a Chronogram of Life:

We used a previously reported Tree of Life as the template for a reference chronogram[26]. This template was constructed using a concatenation of 31 translation-related orthologs. All of the species represented in our gene family dataset were present in this template tree. Divergence times were estimated using PhyloBayes (v.2.3c)[21]. Since autocorrelated molecular clock models have been shown to outperform uncorrelated ones in some cases similar to this study[41], we ran PhyloBayes with a CIR process model of rate correlation. Eight sets of temporal constraints that could be directly linked to fossil or geochemical evidence were used and are displayed in Supplementary Fig. 7. Benchmarking PhyloBayes runs in parallel (n=95) established that predicted divergence times and model likelihood converged after a burn-in of roughly 1500 model cycles. Final divergence time estimates were estimated following a burn-in of 2500 cycles, after which trees were sampled every 20 cycles until the 3500th cycle.

1.3.2. Alternative reference trees/chronograms:

We tested the extent to which the Archean Expansion was sensitive to the topology of the reference tree and to the molecular clock model used in chronogram construction. We built 10 separate reference phylogenies using non-parametric bootstrapping of the Ciccarelli *et al.* gene alignment[26] (see Methods Section 1.4 for PhyML parameters) and rooted each with either the Bacteria, the Archaea, or the Eukarya as the outgroup. Unequivocal errors in phylogeny that may be due to sequence alignment construction errors were observed for the *Bdellovibrio*, *Shigella*, *Treponema*, and *Helicobacter pylori* taxa; these were resolved by manual pruning and re-grafting. Each phylogeny was then converted to an ultrametric tree using

r8s (v.1.71)[22] under a penalized likelihood model (with an additive penalty function, truncated Newton nonlinear optimization, cross-validation enabled, a cross validation start value of 10, a cross validation smoothing increment of 3, and the number of smoothing values tried set to 4). The same set of temporal constraints was used as for PhyloBayes. For the purposes of computational economy, a subset of 250 COGs was randomly selected from our dataset and reconciled against each of the 10 alternative chronograms.

Predicted birth ages were robust to the usage of alternative chronograms. The median gene family birth date difference between any two alternative chronograms is 0.09 Ga (Supplementary Fig. 8) Elevated rates of gene birth during the Late Archean were observed in all 30 of the alternative chronograms and on average, 19% of the 250 chosen COGs were predicted to be born during a 200-My window (Supplementary Fig. 9). However, the timing of Archean Expansion-like window diverged from that reported by PhyloBayes and spanned 2.7-2.5 Ga (compared to 3.3-2.9 Ga for PhyloBayes). Inspection of the r8s chronograms suggests this temporal discrepancy may be related to differences in dating the cyanobacteria under the two models. For both the r8s and PhyloBayes chronograms, Archean Expansion-like events coincided with the relatively brief period during which the major bacterial phyla, such as the Firmicutes and Proteobacteria, diverged from the last bacterial common ancestor. This period of compressed cladogenesis predates the appearance of the cyanobacteria by roughly 100-200 My in both models. Our r8s analysis places the initial occurrence of the cyanobacteria at 2.5 Ga, which is precisely the minimum age constraint for the appearance of this clade (see Section 1.3.1). Using the same constraints, PhyloBayes predicts the cyanobacteria to have emerged 3.0 Ga. We confirmed the importance of the cyanobacteria in dating the Archean Expansion by

reducing the minimum constrained age of this clade during r8s chronogram construction; the resultant model yielded a younger Archean Expansion (data not shown).

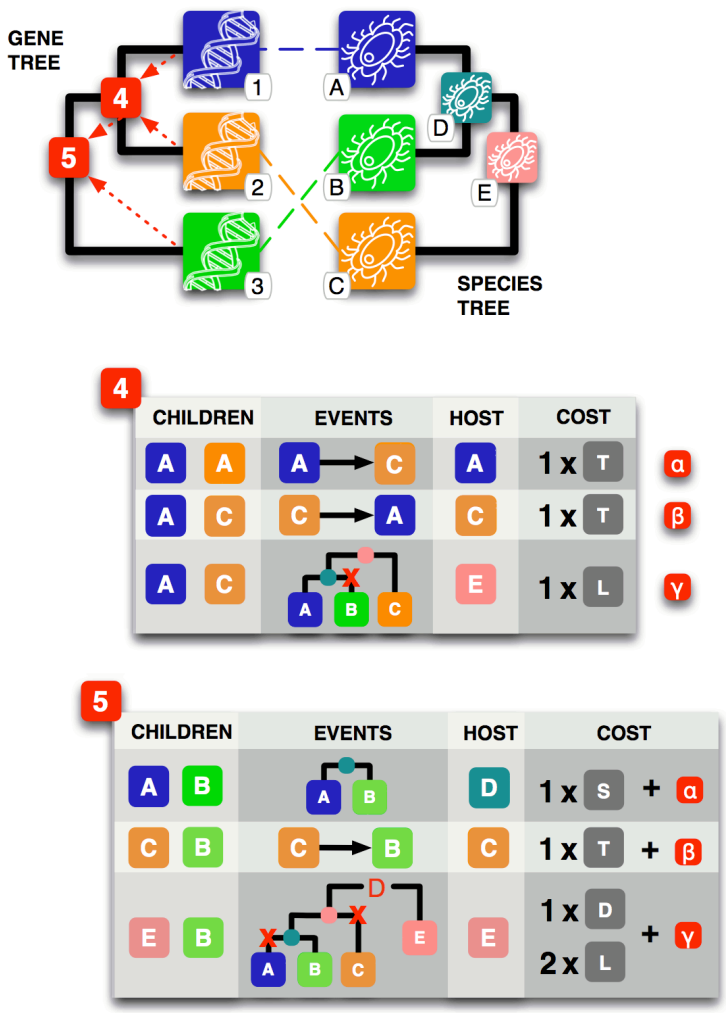
1.4. Gene tree construction

Families of orthologous genes used in this study are based upon functionally annotated orthologous groups from the COG database[42], as extended to a wider set of genomes in the eggNOG database[43]. Due to computational limitations, we restricted this study to a subset of 100 of these genomes (11 eukaryotic, 12 archaeal, and 67 bacterial) broadly distributed across the Tree of Life. Sequences were downloaded from the eggNOG database in September of 2008.

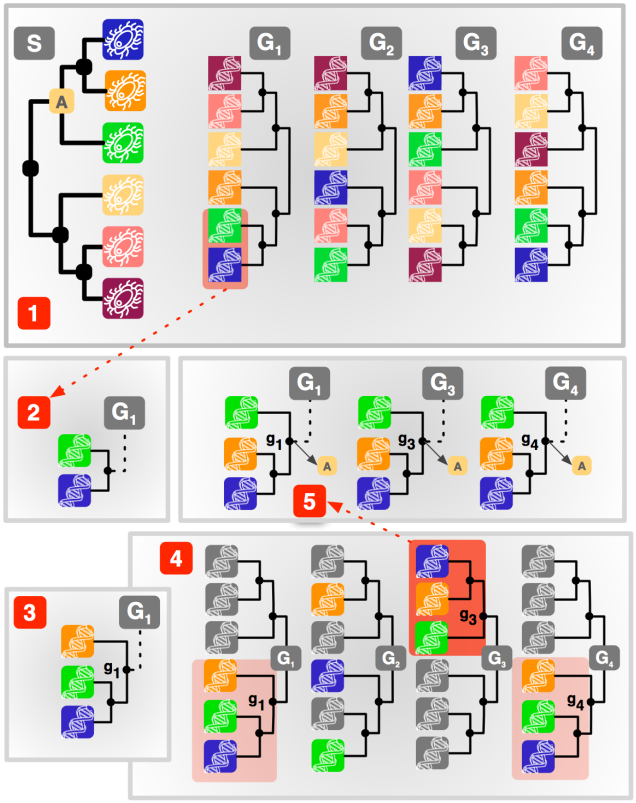
eggNOG-derived families were filtered to ensure usable levels of sequence conservation with the aim of excluding the most error-prone phylogenies. We performed this filtering in an iterative fashion: First, we excised poorly aligned regions of sequence[26,44], using Gblocks (0.91b)[45] with the minimum number of sequences for a flank position set to half the number of sequences in the alignment, the maximum number of contiguous non-conserved positions set to 8, the minimum length of a block set to 2, and the allowed gap positions set to all. Second, we excluded genes with more than 20% of their sequence in these excised regions from each gene family. Third, Muscle (v3.7)[46] was used with default settings to realign the remaining sequences. This process then returned to the first step, unless no sequences or regions were removed in the first or second steps in which case the process terminated. Of the original 4872 COGs, 788 lost more than 25% of their original gene copies during this process; these COGs were considered likely to be error-prone and thus excluded from further analysis. Another 101 COGs were not analyzed due to their high gene copy numbers and the extreme computational demands of running *AnGST* on those large families. A distribution of gene copy numbers within each gene family is shown in Supplementary Figure 11.

Phylogenetic trees were constructed for the remaining gene families using version 2.4.5 of PhyML[38] and the following parameters: 100 bootstrap trees, a JTT substitution model, 0.0 percentage of the sites were invariable, 4 substitution rate categories, a gamma distribution parameter of 1.0, a BIONJ-based starting tree, and both tree topology and branch length optimization were enabled.

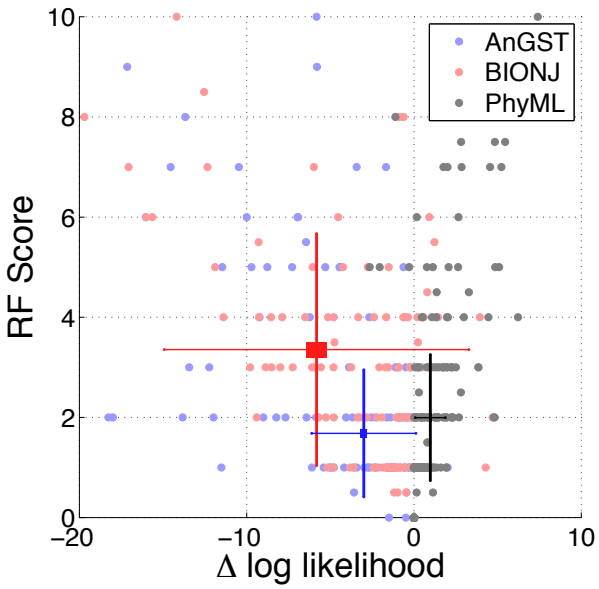
2. Supplementary Figures



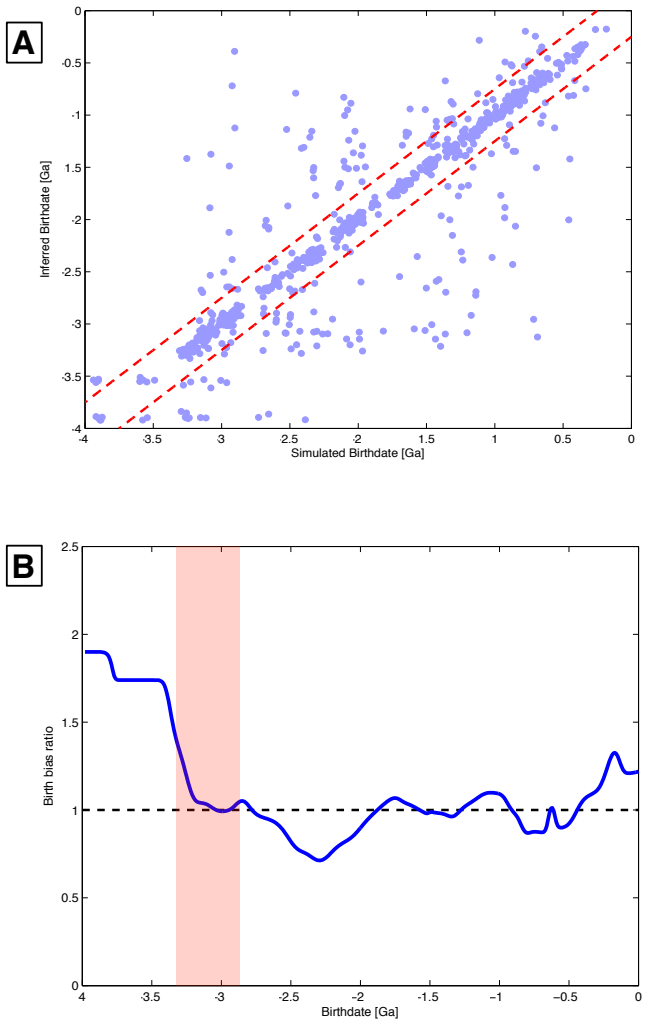
Supplementary Figure 1: Example of a basic reconciliation. An *AnGST* reconciliation of two simple, but discordant, gene (*G*) and species (*S*) trees is shown. The mapping of leaves of *G* to *S*: $\{g_1{:}s_A, g_2{:}s_C, g_3{:}s_B\}$ is indicated with color (*e.g.*, g_1 and s_A are both shown in blue). Reconciliation proceeds in a post-fix manner through the gene tree, first evaluating possible mappings from g_4 to nodes in the *S*. Once the reconciliation process is completed at g_4, the algorithm continues at g_5. A detailed explanation of this reconciliation is provided in Methods Section 1.1.2.



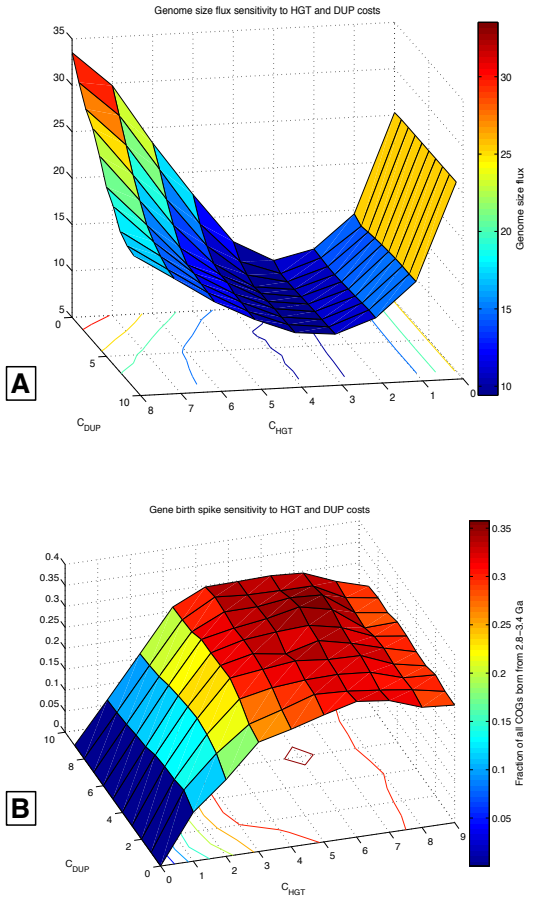
Supplementary Figure 2: Amalgamation algorithm for phylogenetic uncertainty. An *AnGST* reconciliation of four gene tree bootstrap topologies {G_1, G_2, G_3, and G_4} and a species tree S is shown. Leaf nodes on each bootstrap map to leaves on *S* according to color. The reconciliation begins on one of the bootstrap trees, G_1 (Step 1) and proceeds to an interior node (Step 2). The reconciliation does not consider other topologies for this subtree, as it only contains two leaves. When the reconciliation reaches the parent node node g_1 (Step 3), *AnGST* considers subtrees from other bootstraps with alternative topologies (but identical leaves). Corresponding subtrees are found on G_3 and G_4 and rooted at nodes g_3 and g_4 respectively (Step 4). Reconciliations are performed in parallel at g_1, g_3, and g_4. For the mapping of these internal nodes to lineage *A* on the species tree, the reconciliation at g_3 is optimal (since its topology matches the reference one) and the corresponding subtree in G_3 is substituted for the mappings g_1:s_A and g_4:s_A.



Supplementary Figure 3: *AnGST* trees are more accurate than likelihood trees in simulation studies. We simulated the evolution of sequence data using 225 randomly generated gene trees with more than 10 leaves. Gene trees were reconstructed from synthetic sequence data using either BIONJ (red), PhyML (black), or *AnGST* (blue). Phylogenetic accuracy was evaluated by Robinson-Foulds (RF) score. A 0 RF score indicates perfect concordance (all bipartitions of the candidate and reference tree are identical) and increasing RF scores denote higher phylogenetic discordance. The logarithm of sequence likelihood given each tree model, relative to the likelihood calculated with the true gene topology, is plotted on the X axis. Mean RF scores and relative log likelihoods are drawn with rectangles whose height and width reflect standard errors of the mean; protruding lines are standard deviations. PhyML-based trees enjoy significantly higher likelihood scores than the *AnGST* chimeric trees ($p=2.7\times10^{-39}$ Wilcoxon rank sum test), but the *AnGST*-based trees are significantly more similar to the correct gene tree topologies ($p=1.8\times10^{-2}$ Wilcoxon rank sum test). Outlying points beyond axes were not drawn to facilitate viewing mean values, but were included in mean and significance estimations.

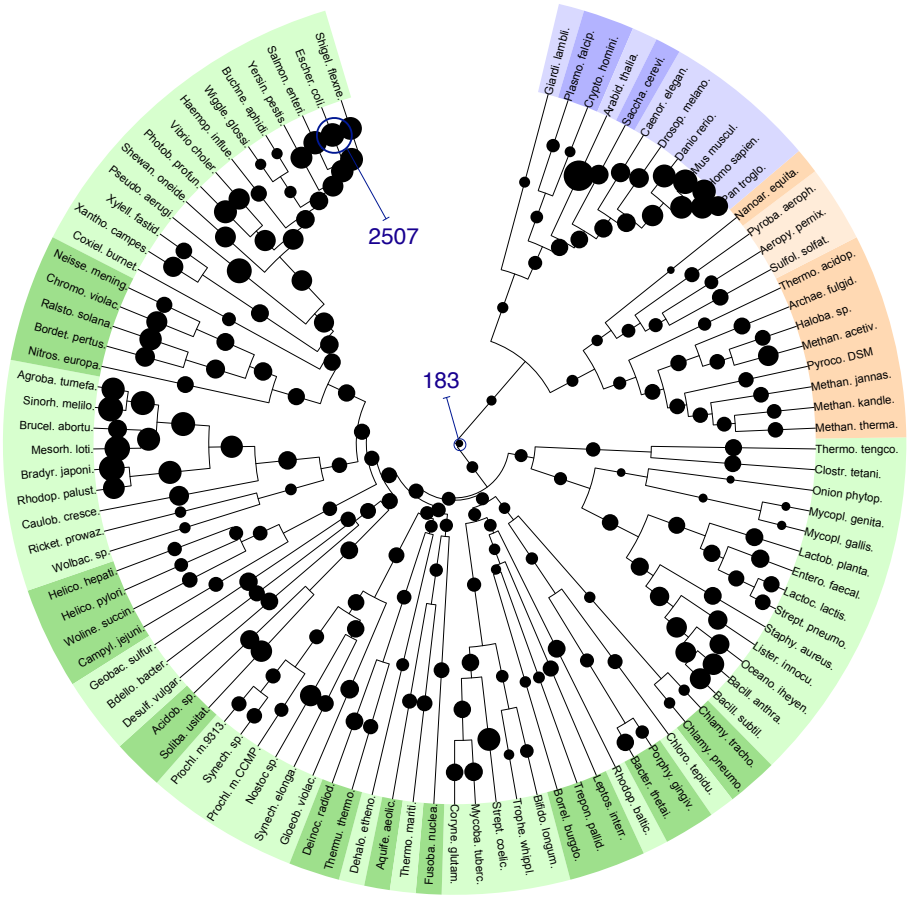


Supplementary Figure 4: Benchmarking *AnGST* inference accuracy. A) A scatter plot of simulated gene family birth dates and inferred birth dates. Points drawn signify midpoints of branches associated with birth events. A slight amount of Gaussian noise with distribution $N(\mu=0,\sigma=0.025)$ has been added to each point so that overlapping points can be distinguished. The correlation coefficient is 0.88 and 76% of predicted births are within 250 My (bounded by red dashed lines) of their true ages. B) Birth prediction bias is plotted as a function of time. Predicted births have been normalized by the number of simulated births associated with a given age. The Archean Expansion (2.9-3.3 Ga) is highlighted in pink.



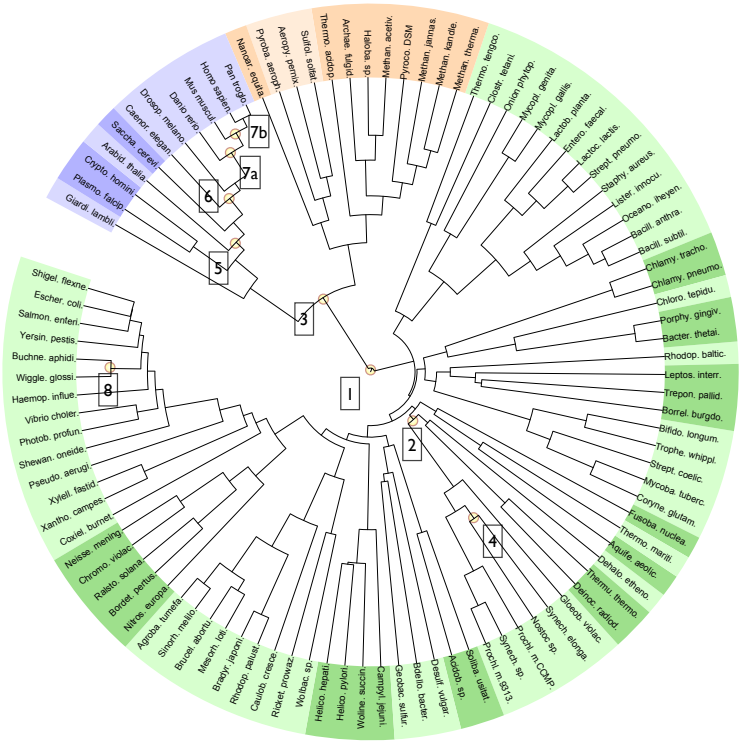
Supplementary Figure 5: *AnGST* parameter learning and sensitivity analysis. **A)** We performed a grid search over the costs C_{HGT} and C_{DUP} with the intention of minimizing average genome size flux between inferred ancestral genomes. The costs C_{LOS} and C_{SPC} were fixed at 1.0 and 0.0, respectively. Flux can clearly be minimized along the C_{HGT} axis, but is less sensitive to changes in C_{DUP}. A minimum point does exist, however, at C_{HGT}=3.0, C_{DUP}=2.0. **B)** A sensitivity analysis for our detection of a high fraction of births from 2.8-3.4 Ga was performed over the same parameter space evaluated in A). Comparable fractions of overall gene birth to the Archean Expansion were detected in parameter space near the genome size flux minimum. Note

that **axes are reversed** in panels A and B in order to facilitate viewing parameter sensitivity landscapes.



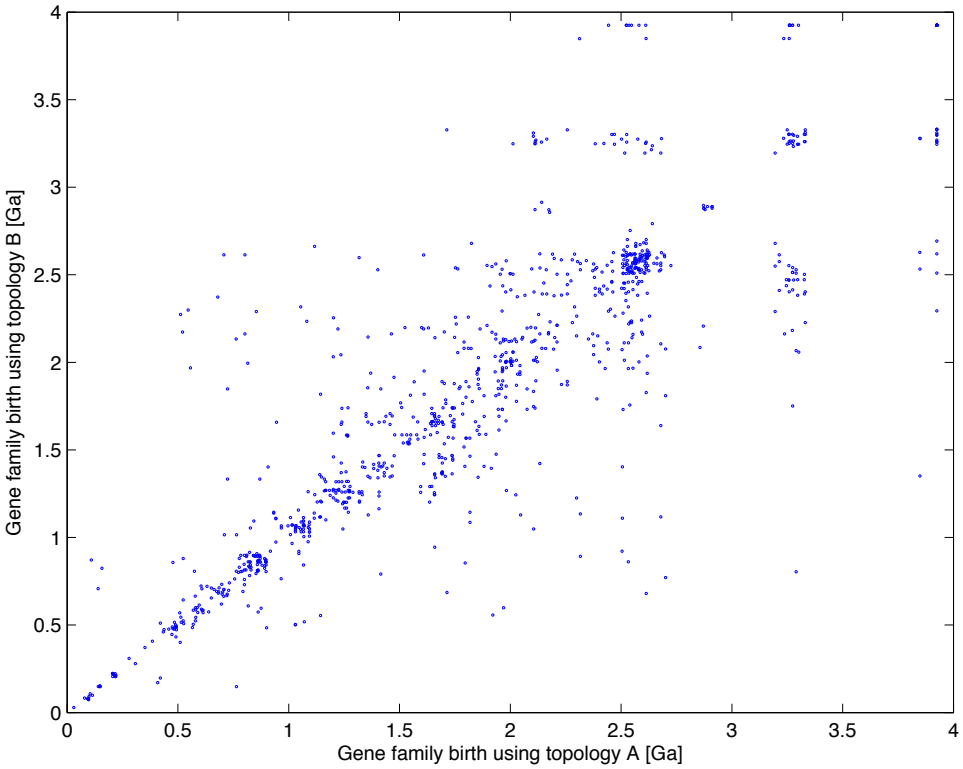
Supplementary Figure 6: Inferred ancient genome sizes for C_{HGT}=3.0, C_{DUP}=2.0. Circle areas scale absolutely with genome sizes. Our optimization metric, genome size flux, aims to minimize the average difference between parent and child genomes. Ancestral genomes are predicted to be smaller than modern day ones; this may reflect the evolution of increasingly complex genomes, and/or the extinction of ancestral gene families. Genome sizes for the LUCA (183 genes) and the modern-day genome of *E. coli* (2507 genes) are labeled in dark blue. Metazoan genomes appear only slightly larger than prokaryotic ones because the COGs used in

this study were originally defined using only unicellular organisms[47], which thus biased our analyses of eukaryotic genomes towards only microbially-related genes.

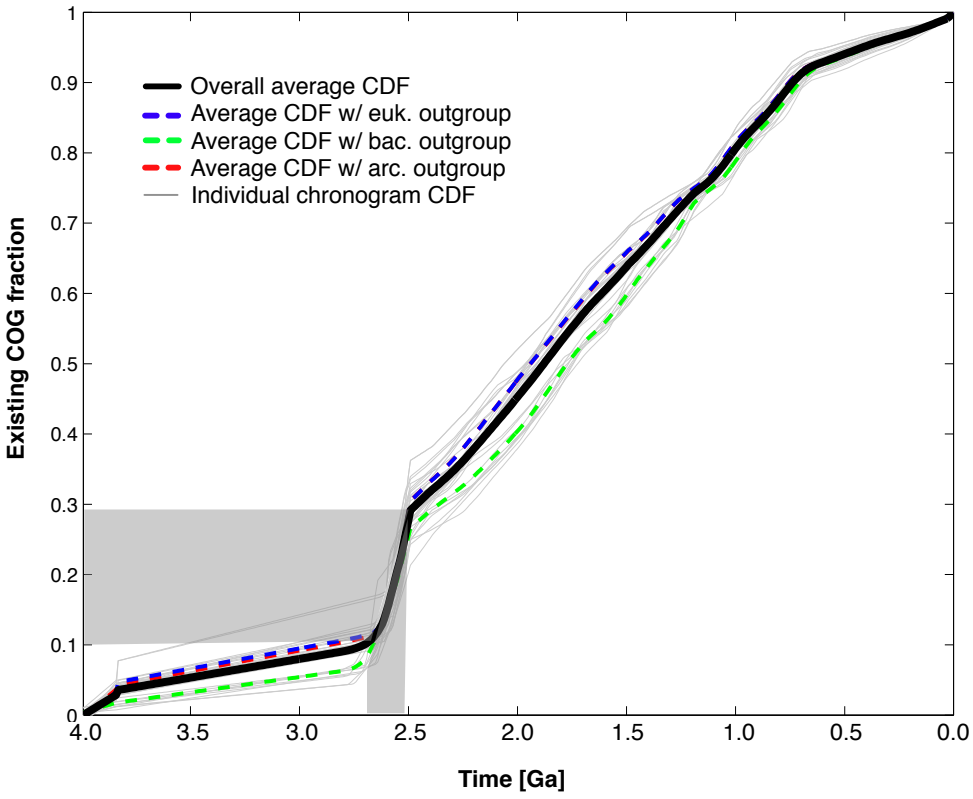


	Event	Constraint
1	Last Universal Common Ancestor arises	< 3850 Ma
2	Cyanobacteria emerge	> 2500 Ma
3	Eukaryotes diverge from Archaea	> 2670 Ma
4	Akinetes diverge from cyanobacteria lacking cell differentiation	> 1500 Ma
5	Archaeplastida emerge	> 1198 Ma
6	Animals emerge	> 635 Ma
7	Tetrapods emerge	(a) < 385 Ma (b) > 359 Ma
8	*Buchnera* diverge from *Wigglesworthia*	> 160 Ma

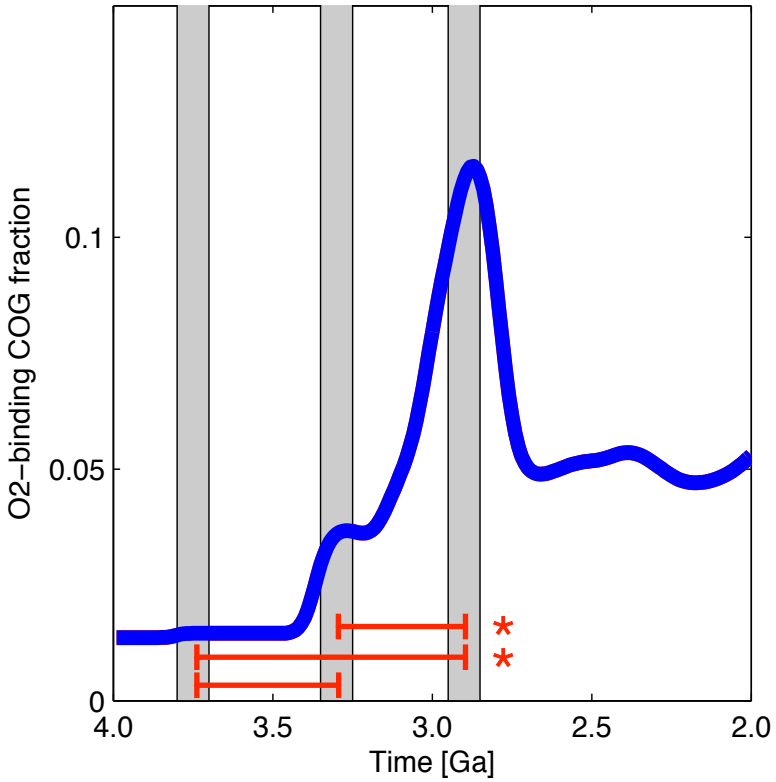
Supplementary Figure 7: Temporal constraints. Eight fossil and biogeochemical constraints were used to constrain the chronogram (evidence cited can be found in Supplementary Table 1). Those constraints are overlaid onto the reference phylogeny.



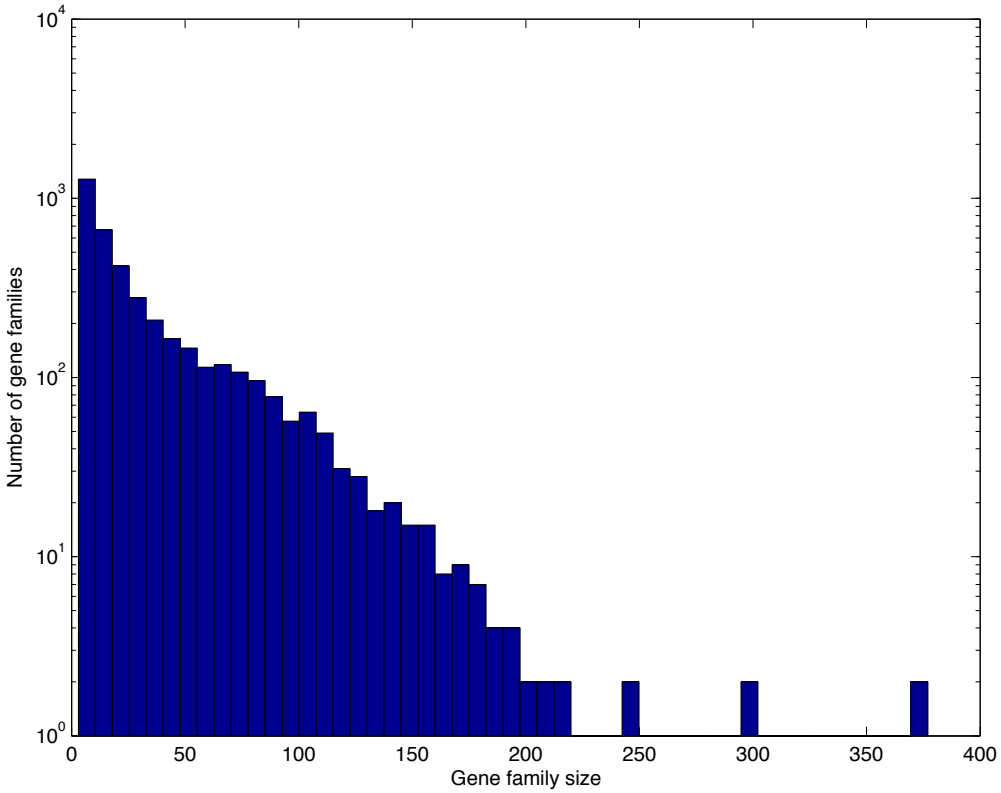
Supplementary Figure 8: Sensitivity of predicted birth ages to variation in reference tree topology. Ten bootstraps of the reference tree were rooted using either the Bacteria, Archaea, or Eukarya as an outgroup and subsequently processed with *r8s*, producing 30 alternative reference chronograms. Birth dates were inferred for 250 gene families using each chronogram. We graph 1000 random combinations of alternative chronogram pairs and gene families in the scatter plot above (correlation coefficient = 0.86). The median gene family birth date difference between any two alternative chronograms is 0.09 Ga.



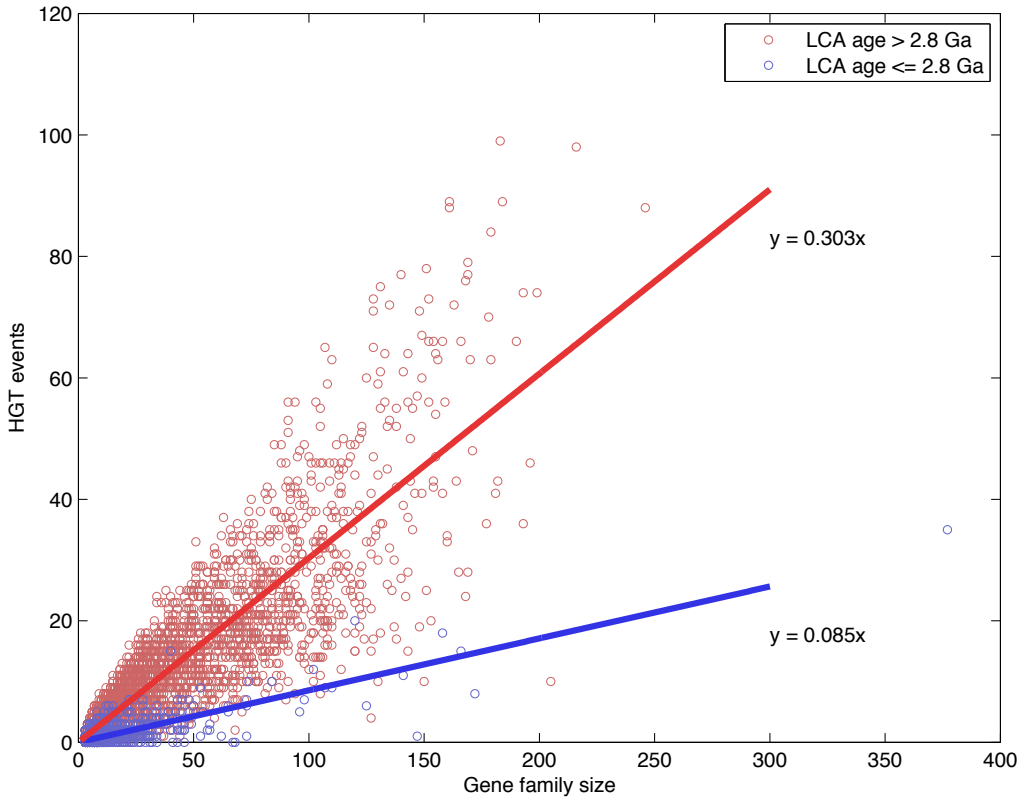
Supplementary Figure 9: Gene family birth using 30 alternative reference tree topologies. Shown above are cumulative distribution functions (CDFs) of total COG birth over time for the 30 alternative reference chronograms (light gray lines). Mean CDFs for the Bacteria, Archaea, and Eukarya as outgroups are shown using green, red, and blue dashed lines, respectively. Overall (solid black line), the period 2.7-2.5 Ga witnesses a gene family birth spike of on average 0.23 families born per 1 Ma and accounts for the birth of 19% of the COG families studied. By contrast, birth rates average 0.07 families born per 1 Ma from 2.5 Ga-present day.



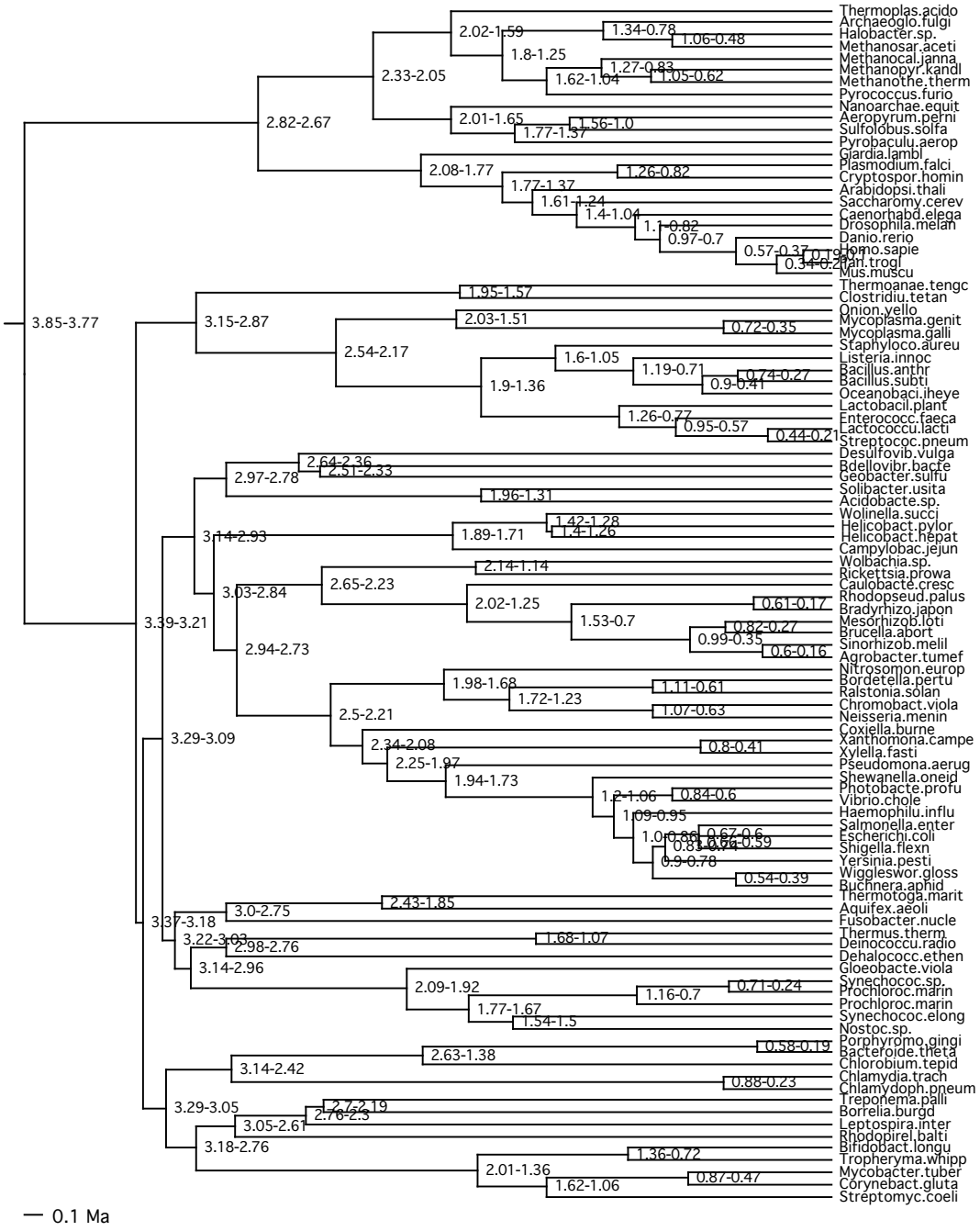
Supplementary Figure 10: O_2 utilizing gene birth over time. The fraction of compound-binding COG births which bind O_2 is shown over time. A chi-square test was used to compare the overall number of COGs born and the number of O_2-binding COGs born in 100 My windows: prior to the Archean Expansion (3.7 Ga), at the height of the Archean Expansion (3.25 Ga), and at the tail of the Archean Expansion (2.85 Ga). Comparisons with $p < 0.05$ are denoted with asterisks on the graph. These data suggest that changes in O_2 usage came toward the end of the Archean Expansion.



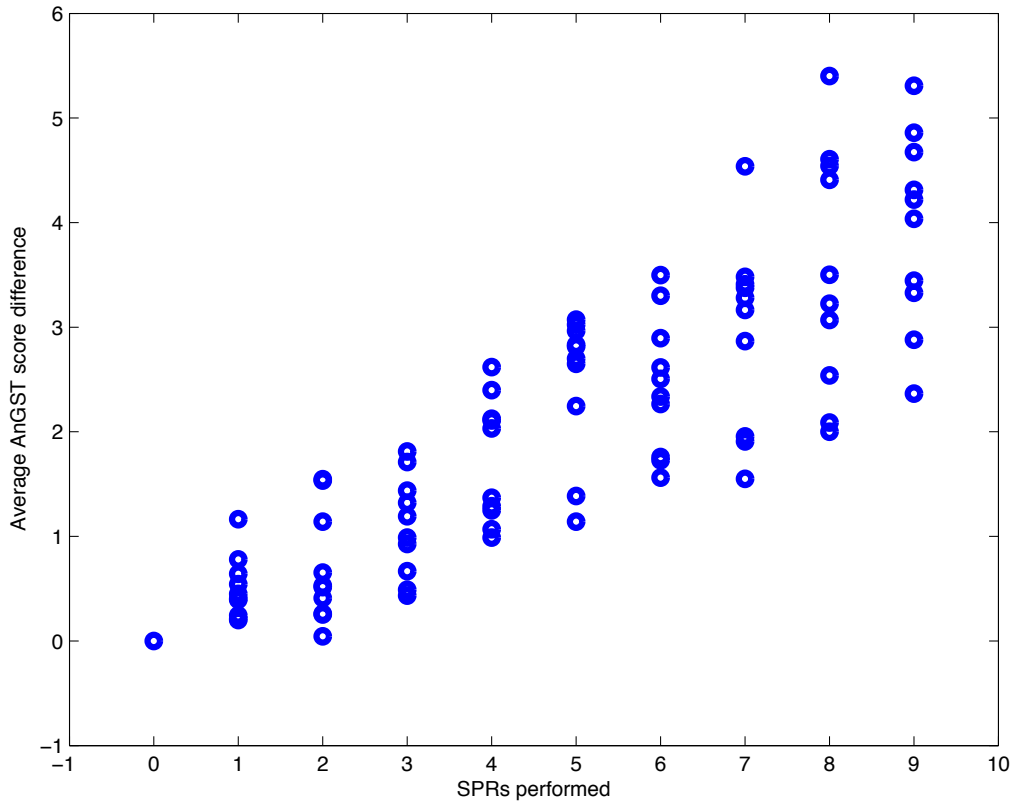
Supplementary Figure 11: Histogram of COG family sizes. The median COG family in our dataset possesses 18 gene copies, and 93% of COG families have 100 or fewer gene copies.



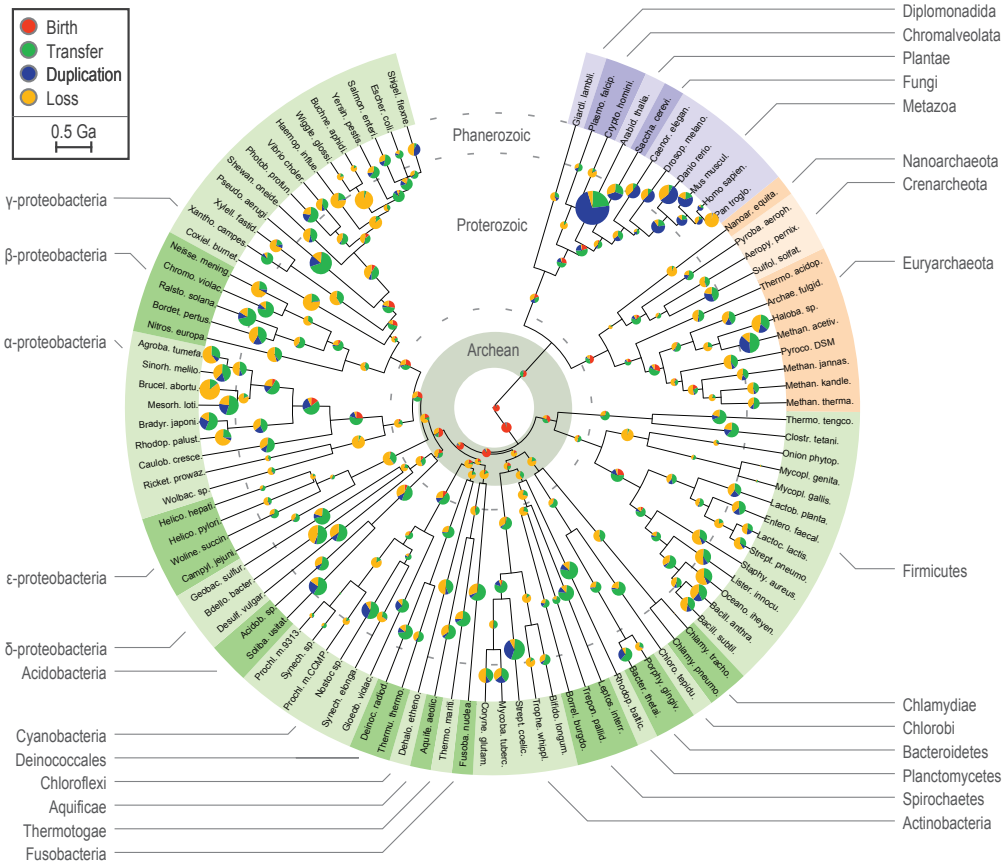
Supplementary Figure 12: HGT counts vs. gene family size. The average gene family reconciliation yields 9.7 inferred HGT events. The number of HGT events inferred grows with the number of gene copies in a COG family. Gene family HGT counts also grow with the age of the last common ancestor of all genomes represented in the family, suggesting that HGT is more frequent among gene families spanning wider phyletic range. We note that y-intercepts for the above line fittings have been forced to equal 0.



Supplementary Figure 13: Confidence intervals for divergence times on reference chronogram. Confidence intervals (95%) were estimated by PhyloBayes and are shown next to each divergence point on the tree. Values are in units of Ga.



Supplementary Figure 14: Average *AnGST* gene tree reconciliation scores as species tree randomization increases. Between 1-9 random Subtree Prune and Regraft (SPR) moves were made to 90 copies of the Tree of Life[26], producing a continuum of species tree accuracy. Each of the these species trees was reconciled against a set of 250 randomly chosen gene trees. The differences between the average *AnGST* scores for each randomized reference tree and the original reference tree are plotted on the y-axis ($r^2 = 0.73$). The monotonic increase in *AnGST* score as a function of species tree permutation suggests the Tree of Life is at least a locally optimal tree representation of the evolution of the sampled gene families.



Supplementary Figure 15: Evolutionary events by lineage. The number of macroevolutionary events is mapped to each lineage on an ultrametric Tree of Life and visualized using the iToL website[48]. Pie chart area denotes the number of events, and color indicates event type: gene birth (red), duplication (blue), HGT (green), and loss (yellow). The Archean Expansion period (3.33-2.85 Ga) is highlighted in green.

3. Supplementary Tables

	Event	Constraint	Evidence
1	Last Universal Common Ancestor arises	< 3850 Ma	Carbon isotope fractionation[49,50]
2	Cyanobacteria emerge	> 2500 Ma	Traces of an aerobic nitrogen cycle[18], changes in redox-metal enrichments[51], and sulfur isotope fractionation data[52,53] indicate oxygenic photosynthesis; traces of 2α-methylhopane biomarkers[8] indicate cyanobacterial presence
3	Eukaryotes diverge from Archaea	> 2670 Ma	Preserved sterane biomarkers[8,20]
4	Akinetes diverge from cyanobacteria lacking cell differentiation	> 1500 Ma	Akinete microfossils[54]
5	Archaeplastida emerge	> 1198 Ma	Red algae microfossils[55]
6	Animals emerge	> 635 Ma	Preserved demosponge steranes[56]
7	Tetrapods emerge	(a) < 385 Ma (b) > 359 Ma	(a) Tetrapod precursor dating[57] (b) Tetrapod fossil dating[58]
8	*Buchnera* diverge from *Wigglesworthia*	> 160 Ma	Fossil history of *Buchnera*'s aphid hosts[59]

Supplementary Table 1: Temporal constraints used to construct chronogram. Eight temporal constraints that could be directly linked to fossil or geochemical evidence were used to estimate divergence times on the Tree of Life (Supplementary Fig. 7).

Meta-function	Function	COG Code	Number of COGs	Fraction of genes studied	Fraction of AGE births	AGE birth enrichment	Fraction of pre-AGE births	pre-AGE birth enrichment	pre-AGE vs. AGE p-value
Information storage & processing	Translation	J	197	0.049	0.061	1.234	0.150	3.042	0.000
	RNA proc.	A	17	0.004	0.001	0.219	0.002	0.377	1.000
	Transcription	K	173	0.043	0.030	0.681	0.042	0.962	0.246
	Replication, recombination	L	155	0.039	0.034	0.868	0.068	1.746	0.002
	Chromatin struct.	B	11	0.003	0.000	0.000	0.002	0.655	0.338
Cellular processes & signaling	Cell cycle control	D	56	0.014	0.013	0.903	0.012	0.854	1.000
	Defense mech.	V	29	0.007	0.008	1.083	0.001	0.097	0.033
	Signal transduction	T	106	0.027	0.028	1.044	0.020	0.762	0.405
	Cell wall/membrane	M	141	0.035	0.050	1.424	0.060	1.702	0.487
	Cell motility	N	82	0.021	0.031	1.493	0.015	0.718	0.064
	Cytoskeleton	Z	5	0.001	0.000	0.000	0.000	0.000	1.000
	Intracell. trafficking	U	122	0.031	0.032	1.047	0.022	0.710	0.274
	Post-trans. modification	O	157	0.039	0.043	1.101	0.042	1.070	1.000
Metabolism	Energy prod. & conv.	C	211	0.053	0.079	1.499	0.068	1.289	0.490
	Carb. trans. & met.	G	186	0.047	0.056	1.205	0.051	1.087	0.730
	Amino acid trans. & met.	E	226	0.057	0.079	1.394	0.109	1.923	0.054
	Nucleotide trans. & met.	F	83	0.021	0.026	1.228	0.059	2.835	0.001
	Coenzyme trans. & met.	H	155	0.039	0.056	1.439	0.069	1.772	0.326
	Lipid trans. & met.	I	72	0.018	0.024	1.306	0.032	1.795	0.334
	Inorganic ion trans. & met.	P	182	0.046	0.058	1.276	0.046	0.998	0.298
	Secondary metabolites	Q	70	0.018	0.015	0.847	0.004	0.216	0.045
Poorly characterized	Func. unknown	S	1186	0.298	0.183	0.615	0.071	0.238	0.000
	General func. pred.	R	560	0.141	0.142	1.010	0.113	0.804	0.105

Supplementary Table 2: Function of gene births prior to and during the Archean Expansion. Functional enrichment of gene birth from 2.8-3.3 Ga is shown for the 20 COG functional categories. A two-tailed Fisher exact test was used to compute the p-value of a difference in total COG births prior to the Archean Expansion vs. during the Archean Expansion, for each functional category (last column).

4. Supplementary Notes

Additional References:

1. Nisbet, E. G. & Sleep, N. H. The habitat and nature of early life. *Nature* **409**, 1083-1091 (2001).
2. Rasmussen, B., Fletcher, I. R., Brocks, J. J. & Kilburn, M. R. Reassessing the first appearance of eukaryotes and cyanobacteria. *Nature* **455**, 1101-1104 (2008).
3. Dupont, C. L., Yang, S., Palenik, B. & Bourne, P. E. Modern proteomes contain putative imprints of ancient shifts in trace metal geochemistry. *Proc Natl Acad Sci USA* **103**, 17822-17827 (2006).
4. Dupont, C. L., Butcher, A., Valas, R. E., Bourne, P. E. & Caetano-Anollés, G. History of biological metal utilization inferred through phylogenomic analysis of protein structures. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 10567-10572 (2010).
5. Saito, M. A., Sigman, D. M. & Morel, F. M. M. The bioinorganic chemistry of the ancient ocean: the co-evolution of cyanobacterial metal requirements and biogeochemical cycles at the Archean-Proterozoic boundary? *Inorganica Chimica Acta* **356**, 308-318 (2003).
6. Zerkle, A. L., House, C. H. & Brantley, S. L. Biogeochemical signatures through time as inferred from whole microbial genomes. *American Journal of Science* **305**, 467-502 (2005).
7. De Marais, D. J. Evolution. When did photosynthesis emerge on Earth? *Science* **289**, 1703-1705 (2000).
8. Brocks, J. J., Logan, G. A., Buick, R. & Summons, R. E. Archean molecular fossils and the early rise of eukaryotes. *Science* **285**, 1033-1036 (1999).
9. Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. Prokaryotic Evolution in Light of Gene Transfer. *Molecular Biology and Evolution* **19**, 2226 (2002).
10. Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci USA* **96**, 3801-3806 (1999).
11. Ragan, M. A. & Beiko, R. G. Lateral genetic transfer: open issues. *Philos Trans R Soc Lond, B, Biol Sci* **364**, 2241-2251 (2009).
12. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299-304 (2000).
13. Fischer, D. & Eisenberg, D. Finding families for genomic ORFans. *Bioinformatics* **15**, 759-762 (1999).
14. Yang, D., Oyaizu, Y., Oyaizu, H., Olsen, G. J. & Woese, C. R. Mitochondrial origins. *Proc Natl Acad Sci USA* **82**, 4443-4447 (1985).
15. Giovannoni, S. J. et al. Evolutionary relationships among cyanobacteria and green chloroplasts. *J Bacteriol* **170**, 3584-3592 (1988).
16. Scott, C. et al. Tracing the stepwise oxygenation of the Proterozoic ocean. *Nature* **452**, 456-459 (2008).
17. Konhauser, K. O. et al. Oceanic nickel depletion and a methanogen famine before the Great Oxidation Event. *Nature* **458**, 750-753 (2009).
18. Garvin, J., Buick, R., Anbar, A. D., Arnold, G. L. & Kaufman, A. J. Isotopic evidence for an aerobic nitrogen cycle in the latest Archean. *Science* **323**, 1045-1048 (2009).
19. Canfield, D. E. A new model for Proterozoic ocean chemistry. *Nature* **396**, 450-453 (1998).

20. Waldbauer, J. R., Sherman, L. S., Sumner, D. Y. & Summons, R. E. Late Archean molecular fossils from the Transvaal Supergroup record the antiquity of microbial diversity and aerobiosis. *Precambrian Research* **169**, 28-47 (2009).
21. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution* **21**, 1095-1109 (2004).
22. Sanderson, M. J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**, 301-302 (2003).
23. Alm, E., Huang, K. & Arkin, A. The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. *PLoS Comput Biol* **2**, e143 (2006).
24. Kunin, V. & Ouzounis, C. A. The balance of driving forces during genome evolution in prokaryotes. *Genome Res* **13**, 1589-1594 (2003).
25. Snel, B., Bork, P. & Huynen, M. A. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res* **12**, 17-25 (2002).
26. Ciccarelli, F. D. et al. Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283-1287 (2006).
27. Dagan, T. & Martin, W. The tree of one percent. *Genome Biol* **7**, 118 (2006).
28. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).
29. Alm, E. J. et al. The MicrobesOnline Web site for comparative genomics. *Genome Res* **15**, 1015-1022 (2005).
30. Huelsenbeck, J. P. & Rannala, B. A Statistical Perspective for Reconstructing the History of Host-Parasite Associations. *… (RDM Page* (2003).
31. Arvestad, L., Berglund, A. C., Lagergren, J. & Sennblad, B. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. *RECOMB'04* 326-335 (2004).
32. Page, R. D. M. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst. Biol* **43**, 58-77 (1994).
33. Charleston, M. A. Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Mathematical biosciences* **149**, 191-223 (1998).
34. MacLeod, D., Charlebois, R. L., Doolittle, F. & Bapteste, E. Deduction of probable events of lateral gene transfer through comparison of phylogenetic trees by recursive consolidation and rearrangement. *BMC Evol Biol* **5**, 27 (2005).
35. Hahn, M. W. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol* **8**, R141 (2007).
36. Bergsten, J. A review of long-branch attraction. *Cladistics* **21**, 163-193 (2005).
37. Rambaut, A. & Grass, N. C. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics* **13**, 235-238 (1997).
38. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**, 696-704 (2003).
39. Robinson, D. & Foulds, L. Comparison of phylogenetic trees. *Mathematical Biosciences* **53**, 131-147 (1981).
40. Dagan, T. & Martin, W. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci USA* **104**, 870-875 (2007).
41. Lepage, T., Bryant, D., Philippe, H. & Lartillot, N. A general comparison of relaxed molecular clock models. *Molecular Biology and Evolution* **24**, 2669-2680 (2007).

42. Tatusov, R. L. et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
43. Jensen, L. J. et al. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res* **36**, D250-4 (2008).
44. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**, 564-577 (2007).
45. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* **17**, 540-552 (2000).
46. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797 (2004).
47. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**, 33-36 (2000).
48. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127-128 (2007).
49. Mojzsis, S. J. et al. Evidence for life on Earth before 3,800 million years ago. *Nature* **384**, 55-59 (1996).
50. Rosing, M. T. 13C-Depleted carbon microparticles in >3700-Ma sea-floor sedimentary rocks from west greenland. *Science* **283**, 674-676 (1999).
51. Anbar, A. D. et al. A whiff of oxygen before the great oxidation event? *Science* **317**, 1903-1906 (2007).
52. Kaufman, A. J. et al. Late Archean biospheric oxygenation and atmospheric evolution. *Science* **317**, 1900-1903 (2007).
53. Reinhard, C. T., Raiswell, R., Scott, C., Anbar, A. D. & Lyons, T. W. A late Archean sulfidic sea stimulated by early oxidative weathering of the continents. *Science* **326**, 713-716 (2009).
54. Golubic, S., Sergeev, V. N. & Knoll, A. H. Mesoproterozoic Archaeoellipsoides: akinetes of heterocystous cyanobacteria. *Lethaia* **28**, 285-298 (1995).
55. Butterfield, N. J. Bangiomorpha pubescens n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/ Neoproterozoic radiation of eukaryotes. *Paleobiology* **26**, 386-404 (2000).
56. Love, G. D. et al. Fossil steroids record the appearance of Demospongiae during the Cryogenian period. *Nature* **457**, 718-721 (2009).
57. Daeschler, E. B., Shubin, N. H. & Jenkins, F. A. A Devonian tetrapod-like fish and the evolution of the tetrapod body plan. *Nature* **440**, 757-763 (2006).
58. Daeschler, E. B., Shubin, N. H., Thomson, K. S. & Amaral, W. W. A Devonian Tetrapod from North America. *Science* **265**, 639-642 (1994).
59. Moran, N. A., Munson, M. A., Baumann, P. & Ishikawa, H. A Molecular Clock in Endosymbiotic Bacteria is Calibrated Using the Insect Hosts. *Proceedings of the Royal Society B: Biological Sciences* **253**, 167-171 (1993).