# Inferring Gene Presence in Incomplete Data via Phylogenetic Occupancy Modeling

**John S.A. Mattick**[† 1], **Wesley C. DeMontigny**[† 1], **and Charles F. Delwiche**[1]

[†] Equally contributing authors
[1] Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, 20742, USA

**Increasing access to genomic data has revolutionized our understanding of biology. Organisms that were previously unculturable or otherwise difficult to study have been investigated using metagenomic sequencing and bioinformatic assemblies, illuminating biological diversity that was previously invisible. However, as the availability of genomic data has grown, so has the challenge posed by incomplete genomes. Many genomes obtained from metagenomic assemblies or mixed cultures are of poor quality and establishing fully complete genomes requires substantial effort. Incomplete genomes pose difficulties for several common analyses, particularly gene-inventory and core-genome analyses. When genomes are incomplete, distinguishing true gene absence from non-detection becomes difficult. For relatively complete genomes, gene absences are inferred to be true absences, but for highly incomplete genomes, researchers often exclude such data entirely. Probabilistic models have attempted to address this issue in core genome analyses. In particular, mOTUpan utilizes an iterative algorithm to infer genome completeness and categorize genes as core or accessory. In this work, we substantially improve upon this approach by integrating a well-established class of ecological models, called occupancy models, with evolutionary modeling. Our "phylogenetic occupancy model" defines a probability distribution over gene presence that accounts for shared information across related genomes. This framework simultaneously estimates genome completeness and the probability that a gene is present but unobserved. This model substantially outperforms competing methods for core genome inference and enables inference of single-gene presence/absence and ancestral-state reconstruction. Alongside this paper, we provide our model as a Python package.**

**Incomplete Data | Occupancy Modeling | Core Genome | Gene Inventory**
**Correspondence:** ***delwiche@umd.edu***

## Introduction

Our understanding of the microbial world has been revolutionized several times, most recently by the metagenomics-driven expansion of the catalog of Earth's diversity (1). The advent of the microscope enabled the discovery and morphological characterization of microorganisms, which were eventually classified into two major categories: eukaryotes and prokaryotes (2). Later, the discovery of DNA and the advent of sequencing technologies enabled the characterization of the genomic content of these organisms and the use of those genomes to define relationships based on shared genetic history, in addition to shared morphology (3). Small subunit ribosomal RNA (16S rRNA), which has been used for over 50 years to define the relationships between different species of bacteria (4, 5), was able to identify a third domain of life, the Archaea, whose morphology superficially resembles bacteria, but that are genetically closer to eukaryotes (6, 7). As sequencing costs have decreased, genomics has shifted from the characterization of isolated single organisms to broader sequencing projects that study entire microbial communities. Metagenomics is the study of genetic material directly recovered from environmental samples and has redefined our understanding of microbial diversity on Earth (8–10). The number of cultured (or culturable) microorganisms remains unclear, but some studies estimate that as few as 0.5% of bacteria and archaea can be cultured using standard techniques (11). Metagenomic analysis has led to the discovery of thousands of novel taxa and clades, and revolutionized our understanding of the tree of life.

One of the most powerful tools in metagenomics is the ability to assess the genomic content of hundreds of genomes from a specific taxon without isolating and growing them individually (12). However, the genomes acquired by these methods are often incomplete, either due to low sequencing depth or difficulty in taxonomic binning (13); even deep sequencing can yield partial genomes of rare taxa. In fact, this is representative of a broader trend in genomics, in which the number of partial genomic assemblies vastly outnumbers that of complete or near-complete assemblies (14).

To understand the functional capabilities of a given clade, gene inventories and core genome analyses were developed. Gene inventories examine the presence or absence of specific gene families to characterize the functional capacities of individual taxa, whereas core genome analyses identify the set of genes shared by all members of a clade to infer their common biological capabilities (15). However, incomplete genomic data make these analyses challenging, since the absence of a gene from a genome assembly does not necessarily imply its absence from the true genome (16). Explicit treatment of incomplete information in gene inventories remains uncommon, and core genome analyses have largely relied on inclusion thresholds or empirical gene presence thresholding to accommodate variability in genome completeness. Other researchers have attempted to leverage the shared information between genomes to make more informed inferences about the core genomes in the presence of incomplete data. PPanGGOLiN (17), for example, uses synteny networks to identify clusters of co-occurring genes and classifies genes into "persistent," "shell," or "cloud" populations. Although this approach is highly scalable, its reliance on synteny is problematic for genomes that are highly fragmented or so distantly

DRAFT

related that synteny is no longer conserved. mOTUpan (18) adopts a probabilistic approach to core genome inference, using an iterative algorithm that classifies genes as either core or accessory. Core gene presence is modeled as a function of genome completeness, whereas accessory genes are assumed to be independently distributed across genomes, following a gene-frequency-dependent distribution informed by inferred genome sizes. While this performs well at the low evolutionary distances for which it was designed, the assumption that core genes are those whose presence-absence patterns are explained entirely by genome completeness becomes problematic at deeper evolutionary depths. Moreover, if the dataset is heavily biased toward particular sub-clades, this approach is likely to produce a substantial number of false-positive core genes.

The model used by mOTUpan can be viewed as an occupancy model. These models are popular in ecology for assessing the probability that a species occupies a particular site given incomplete (or non-existent) survey data at that site, given repeated observations (19–22). In general, occupancy models assign a probability that species $i$ occupies site $j$ given an observation $x_{ij}$ as:

$$p(x_{ij} = 1 \mid z_{ij}) = \begin{cases} 0, & z_{ij} = 0 \\ p_j, & z_{ij} = 1 \end{cases}$$

$$p(x_{ij} = 0 \mid z_{ij}) = \begin{cases} 1, & z_{ij} = 0 \\ 1 - p_j, & z_{ij} = 1 \end{cases}$$

where $p_j$ indicates the completeness of the survey at site $j$ and $z_{ij}$ is the true occupancy state, with $z_{ij} = 1$ if the species truly occupies that site and $z_{ij} = 0$ otherwise. In the case of mOTUpan, there is a mixture of models: one component models genes as being distributed according to their frequency of occurrence (the accessory component), and another component acts as an occupancy model, where the true state $z_{ij}$ is 1 for all genes in that mixture component. There are several areas where this model could be improved. First, even genes that exhibit seemingly random distributions may contain valuable information about the completeness of a given genome. Second, not all genes whose presence is approximately predictable under an occupancy model belong in the "core" set. Indeed, the core set in the context of mOTUpan is simply the set of genes whose presence is predictable by genome completeness, rather than the set of genes that are universal across all genomes.

Ideally, we would like to pool information across all genomes $j$ according to their similarity. In ecological occupancy modeling, this can be accomplished using Gaussian processes to leverage spatial covariance among survey sites (23) or ecological covariates of the sites. In the context of genomes, however, these approaches are not particularly useful. There are no especially meaningful notions of distance between genomes, and designing covariates for each metagenomic dataset would require complex dataset-specific choices. An alternative approach is to use a belief network. Belief networks define complex joint distributions over random variables through a set of dependence relationships that form a directed acyclic graph (24). Such approaches have been used to model between-species dependence in ecological occupancy models (25), but are less commonly applied to survey sites. In the context of genomes, a natural choice for the belief network structure is a phylogenetic tree. Although statistical phylogenetic models are typically formulated as stochastic processes evolving along the branches of a tree, they may equivalently be viewed as belief networks. In this representation, two genomes are conditionally independent given the state of their most recent common ancestor, and the joint distribution over ancestral and descendant states is governed by a global set of shared parameters (i.e., the substitution rate parameters) together with relationship-specific parameters (i.e., the branch lengths) that control the divergence of the joint distribution from some limiting stationary distribution.

Designing our occupancy model as a belief network allows us to draw on a broad body of well-established literature and algorithms from probabilistic graphical modeling. In particular, it provides several options for the model's output. In our case, both the sum-product and max-product algorithms seem particularly useful. These algorithms compute the marginal posterior probability that a gene is in a genome and the joint maximum *a posteriori* state across all genomes, respectively (26, 27). Because these algorithms propagate across the entire phylogenetic tree, they also identify the marginal probability and the maximum *a posteriori* state for each ancestral node. This allows us to better understand genomic content in extinct organisms when the data from their descendants is incomplete.

Therefore, to improve probabilistic assessments of gene presence-absence in incomplete genomic data, we developed a probabilistic model in which a phylogenetic belief network defines the joint distribution of the true presence-absence states across genomes, while a full occupancy model specifies the observation process. We benchmark this model for internal coherence using simulations, and evaluate its performance on empirical datasets from $\alpha$-proteobacteria and $\gamma$-proteobacteria. We also showcase the ability of the model to reconstruct ancestral genome content on the phylogeny in the Asgardarchaea. Our approach outperforms comparable methods, such as mOTUpan and simple completeness thresholding, while enabling richer downstream analyses. In particular, the model assigns posterior probabilities to the presence of every gene in every genome (rather than merely assigning genes to categories) and naturally supports ancestral state reconstruction within the same probabilistic framework.

## Methods

**Phylogenetic Occupancy Modeling.** In a phylogenetic occupancy model, our goal is to infer the true occupancy state $z_{ij}$ for each gene $i$ in each genome $j$. Each genome is associated with a completeness parameter $p_j$, which governs the conditional distribution of the observed presence–absence data $x_{ij}$ given the latent occupancy state. As described above, these conditional probabilities follow the stan-

DRAFT

dard occupancy model,

$$p(x_{ij}=1 \mid z_{ij}) = \begin{cases} 0, & z_{ij}=0 \\ p_j, & z_{ij}=1 \end{cases}$$
$$p(x_{ij}=0 \mid z_{ij}) = \begin{cases} 1, & z_{ij}=0 \\ 1-p_j, & z_{ij}=1\,. \end{cases}$$

This construction already defines a simple belief network, but with dependencies restricted to pairs of variables $(z_{ij}, x_{ij})$. Our aim is to extend this structure so that, for a fixed gene $i$, the latent occupancy states $(z_{i1}, \ldots, z_{iN})$ across genomes are coupled through a joint distribution that reflects their evolutionary relatedness, with more closely related genomes being more likely to share the same state.

We encode this dependence through a set of conditional independence relationships defined on a phylogenetic tree. For any two genomes $a$ and $b$, the corresponding occupancy states for gene $i$ are conditionally independent given the state at their most recent common ancestor. Conditional probabilities of occupancy state along each branch of the tree are modeled using a symmetric two-state process, with the conditional probabilities

$$P(\text{ChildState} = \text{AncestralState}) = \tfrac{1}{2}\left(1+e^{-2tr_i}\right),$$
$$P(\text{ChildState} \neq \text{AncestralState}) = \tfrac{1}{2}\left(1-e^{-2tr_i}\right),$$

where $t$ is a branch-specific parameter measuring the evolutionary distance between a node and its ancestor, traditionally interpreted as the elapsed time of a continuous-time Markov chain; this parameterization is chosen to ensure increasing agnosticism about the occupancy state as evolutionary distance grows, with the descendant state approaching a uniform distribution for large $t$. To account for the fact that gene families exhibit different levels of transience in a genome, we use a mixture of rate multipliers for $t$. In each mixture component, we assign a new parameter $r_i$ that acts as a multiplicative factor for the branch lengths $t$. In this case, we set the values of $r_i$ to come from a discretized standard log-normal distribution. This approach can be viewed as a modification of the phylogenetic method for Gamma site-rate heterogeneity (28). We found that simultaneously estimating the parameters of the discretized mixture distribution led to unstable optimization in NumPyro, whereas the standard log-normal mixture performed well. In all of our analyses, we utilize a 10-component mixture.

Likelihoods are computed by propagating these conditional probabilities down the phylogenetic tree using Felsenstein's pruning algorithm (29). This computation is equivalent to exact marginalization via variable elimination on a tree-structured belief network or the sum-product algorithm, where the query node is only the root (26, 27, 30). At the root of the tree, instead of weighting by the stationary distribution of a Markov chain, we introduce a prior vector $\boldsymbol{\pi}$ that represents the proportion of genes present at the ancestral genome of the entire dataset. Thus, depending on the rooting of the tree, we are likely to infer a slightly different $\pi$, since it represents a prior probability assigned to a different node.

The model was implemented in Python using NumPyro (31, 32), enabling efficient maximum likelihood estimation via automatic differentiation and the ADAM optimizer. Preliminary experiments with Hamiltonian Monte Carlo in NumPyro indicated that the posterior distribution was tightly concentrated around the maximum likelihood estimate, but was much more computationally intensive. Given the large number of benchmarking runs performed here, we therefore report maximum likelihood estimates. We simultaneously estimate genome completeness $\boldsymbol{p}$, branch lengths $\boldsymbol{t}$, and ancestral proportion $\boldsymbol{\pi} = [\pi_0, 1-\pi_0]$ given a dataset and a fixed phylogeny. Initial experimentation indicated that 1,000 epochs of optimization yielded a good fit on the datasets we investigated. Following parameter estimation, we apply the sum-product algorithm to compute marginal posterior probabilities of occupancy for each gene at each node and genome, and the max-product algorithm to obtain a joint maximum *a posteriori* reconstruction of occupancy states across all genomes. On a single NVIDIA L4 GPU, inference for a single dataset takes minutes. This makes data curation the most time-consuming part of these analyses.

The phylogenetic occupancy model can be interpreted as a statistical phylogenetic model with taxon-specific state ambiguity (e.g., utilizing IUPAC ambiguity codes in an amino-acid phylogeny). In this view, an observed $x_{ij}=1$ corresponds to an unambiguous presence state, whereas $x_{ij}=0$ is ambiguous, representing either true absence or unobserved presence, with the degree of ambiguity determined by the genome-specific occupancy model. In terms of a generative probabilistic model, this encodes a process whereby some proportion of genes $1-\pi_0$ are present at the root of the phylogeny and are gained or lost across the tree according to a continuous-time Markov chain to produce a final true gene occupancy state $z_{ij}$ for each gene $i$ and each genome $j$. Finally, if the true occupancy state is $z_{ij}=0$, we will observe $x_{ij}=0$, however if $z_{ij}=1$, we will observe $x_{ij}=1$ with probability $p_j$ and $x_{ij}=0$ with probability $1-p_j$.

**Simulation Analysis.** Datasets with varying numbers of genomes and degrees of completeness were simulated under the model to characterize its behavior when the model was well specified. Specifically, datasets with 10, 50, 100, 500, and 750 genomes were simulated, each containing 5,000 potentially acquirable genes, with genome completeness values drawn from a Beta distribution. We considered 8 different parameterizations of the Beta distribution, spanning a wide range of completeness profiles, using the following treatments: $(\alpha,\beta) \in \{(30,10),(5,1),(3,0.5),(1,1),(0.25,0.25),(1,5),(0.5,3.0)\}$. For each combination of completeness profile and number of genomes, we simulated 10 datasets, each with a randomly generated phylogenetic tree, yielding a total of 400 analyses. The tree diameter of each phylogeny was set to be 0.5 for all simulations, meaning each ortholog has been gained/lost approximately 0.5 times from the root to the longest tip of the tree. For each simulation, model outputs were recorded, and precision and recall were computed under the different conditions. Here, we do not consider observed presence

($x_{ij} = 1$) in our precision/recall calculations, as these inferences are guaranteed to be correct. The results were visualized using ggplot2.

**Empirical Evaluations.** $\alpha$-proteobacterial and $\gamma$-proteobacterial genomes (442 and 694 genomes, respectively) were pulled from the NCBI genome database by selecting all genomes that were categorized under these taxonomic designations in NCBI and had the following qualities: reference genomes, annotated by NCBI RefSeq, from type materials, excluding atypical genomes, excluding metagenomic assemblies, and excluding genomes from large multi-isolate projects. Proteins from these genomes were assigned to orthogroups via the emapper tool in the EggNog2 suite (33, 34). Concatenated amino acid phylogenies were generated by identifying orthogroups that were present in 95% of genomes with a copy number of 1 or 2. When 2 copies were present, the longest ortholog was selected. Each orthogroup was aligned with MAFFT v.7.5.20 (35) using the –auto parameter and trimmed with TrimAl using the -automated1 parameter (36). These trimmed alignments were concatenated, and phylogenies were inferred with IQTree2 v. 2.2.7 (37), using models selected by IQTree2's model finder. These phylogenies and high-quality gene presence/absence data were treated as ground truth to evaluate our model's performance on biological data.

We artificially depleted orthogroups from our genomes using the same completeness profiles (Beta distributions) described above and analyzed the resulting datasets under our model. For each genome $j$, we draw a completeness parameter $p_j$ and mask $(1-p_j)\%$ of the present orthogroups from the data. After each dataset was depleted, we discarded orthogroups present in fewer than 5 depleted genomes to simulate the removal of extremely rare orthogroups. We performed this procedure 10 times per completeness profile, yielding 80 analyses per dataset.

Using both the marginal probabilities of occupancy and the maximum *a posteriori* estimates, we inferred core genomes for major $\alpha$-proteobacterial and $\gamma$-proteobacterial clades and assessed their performance against the observed ground truth. We compared the precision and recall of our core genome estimates with those of empirical presence-thresholding methods (i.e., genes present in 90% or 95% of observed incomplete genomes) and with mOTUpan predictions. We investigated both definitions of a true core genome, in which genes are present in all members of a clade, and a relaxed core genome, in which genes are present in at least 90% of members of a clade. The results were visualized using ggplot2.

**Asgardarchaea Analysis.** The Asgardarchaeal dataset was generated from all Asgard genomes present in Genbank as of October 2025. Unlike the proteobacterial genomes, there was no discrimination in genome quality: all genomes were downloaded, including 304 with predicted protein sequences and 437 with only nucleic acid sequences. Genomes without proteins had their protein content inferred using Prokka with the –kingdom Archaea parameter (38). In addition, 206 reference-quality and complete genomes from the TACK group of Archaea were downloaded. A phylogeny based on a concatenated amino-acid alignment of the combined TACK and Asgard dataset was inferred, as was done for the $\alpha$-proteobacterial and $\gamma$-proteobacterial datasets. Two Asgard genomes were excluded from the analysis because 90% of the orthogroups used to generate the phylogeny were missing from these genomes. The TACK outgroup was excluded from downstream analyses and included solely to infer a root for ancestral reconstruction within the Asgard clade. We analyzed this dataset in our phylogenetic occupancy model with particular focus on the ancestral states of the major lineages of the Asgards. We specifically focused on orthogroups identified as belonging to the set of eukaryotic-specific proteins (ESPs) identified by Liu et al (39). Proteins from the Asgard dataset were assigned to AsCoG designations and filtered to only include AsCogs identified as ESPs. These proteins then filtered the orthogroup list, retaining only orthogroups containing ESP-designated AsCog proteins for analysis in the ancestral state reconstruction. Joint reconstruction was used to infer ancestral genome content at each node on the phylogeny. Gains and losses along the tree were plotted using ggtrees for visualization (40). Specific nodes were selected as roots of major Asgard clades using NCBI species designations. The operational clade definitions for Heimdal, Loki, Hel, Thor, Odin, Hod, Njord, Wukong, Jord, Freya, Baldr, and Sif were chosen to maximize the number of tips that match the species designation while minimizing those that do not. This algorithm was employed to account for the placement of some species outside their monophyletic groups, either due to incorrect species designations or other anomalies. Joint reconstruction was used to determine the number of ESP-containing orthogroups at each node, and these counts were then visualized against the total number of identified ESP-containing orthologs using the pheatmap function in R (41).

## Results

**Simulation Analysis.** Across all conditions in our simulations, performance improves with genome count; larger datasets exhibit consistently higher precision at a given recall (**Figure 1**). When the number of genomes is small, precision declines rapidly as recall increases, indicating reduced robustness to incomplete observations in low-data scenarios. In contrast, datasets with several hundred genomes maintain high precision across most of the recall range, with degradation occurring primarily near perfect recall. Within each genome-count, variation across completeness profiles yields only modest differences in the precision–recall curves. Overall, the model performs strongly on idealized datasets.

**Empirical Evaluations.** At the per-gene inference level, our model maintains favorable precision and recall across a wide range of completeness profiles (**Figure 3**). Although the model generally achieves higher precision and recall on incomplete datasets, these datasets contain more true positives available for detection. Under most completeness profiles, around 90% precision is obtainable at 40% recall. Overall,

performance is generally lower than that of simulation analyses using a well-specified model. Still, it shows that reasonable per-gene inferences can be obtained when both the phylogeny and orthologous group assignments are inferred from real data.

Across both $\alpha$- and $\gamma$-proteobacterial subclades (**Figure 2**), our model's joint reconstruction and marginal probability thresholding consistently achieved higher recall for core genome inference than empirical thresholding approaches, while maintaining competitive precision. For the strict (present in 100%) definition of the core genome, the joint reconstruction exhibited near-perfect recall across most subclades, while the other methods frequently failed to recall anything (**Figure 4a, 4b**). Similar patterns were observed with the relaxed core definition (90% present) (**Figure 5a, 5b**); all methods showed modest improvements in recall, with our model achieving significantly higher precision.

We used pairwise superiority matrices to simplify performance comparisons among methods (**Figure 4c, 5c**). Rather than assessing the benefits of greater/lesser recall at the cost of greater/lesser precision, we count the number of times each model is strictly superior to another (i.e., better in terms of both precision and recall). Across both strict and relaxed core definitions, the joint reconstruction output outperforms empirical thresholding and mOTUpan in the majority of comparisons, with marginal probability thresholding outperforming empirical thresholding, albeit with lower recall than mOTUpan and joint reconstruction. Precision distributions show that while mOTUpan achieves high recall, it does so at the cost of substantially reduced precision, whereas our methods allow for higher precision.

**Asgardarchaea Analysis.** To apply our model's ancestral reconstruction capabilities, we used 741 Asgardarchaeal metagenomes of varying degrees of completeness. Two genomes were excluded from this analysis due to extremely poor gene content (more than 95% gaps in the alignment used to infer phylogeny). An outgroup of 206 TACK archaea was used to root the Asgard tree. The inferred phylogeny was monophyletic within TACK, allowing us to identify a potential root for ancestral state reconstruction (**Figure 6**). Eukaryotic signature proteins (ESPs) (39) were matched to eggnog annotations via BLAST matches, and the presence or absence of these specific orthologs across the Asgardarchaeal tree was observed (**Figure 6**). The ancestral reconstruction suggests significant gains and losses of ESPs after the divergence of the major Asgard groups, and often near the tips of the tree (**Figure 7c**). This is notably true in Heimdalarchaea and Hodarchaea, which have approximately the same number of ESPs in their common ancestor (158 ESPs in the reconstructed ancestor of Heimdalarchaea (**Figure 7a**), 166 in the reconstructed ancestor of Lokiarchaea (**Figure 7b**), and 169 in the reconstructed ancestor of Hodarchaea) as other major Asgard groups, but gain significantly more towards the tips (the maximum number of ESPs in any Asgard tip species after modeling is Heimdallarchaeota GCA011364945.1 with 241 ESPs). Based on this reconstruction, no Asgard group is able to claim more than 40% of identified Asgard ESPs in the common ancestor.

## Discussion

**Model Performance.** Our initial simulation analyses indicate that the model performs very well when well-specified. Furthermore, its performance tends to improve as the number of taxa in the dataset increases. This behavior can be understood as arising from the model's increasing number of imperfect "views" of each genome; each genome serves as an evolutionary replicate of the others. However, these data were simulated under the model itself, which is not the generative process that operates in real biological systems. Therefore, we sought to evaluate our model on high-quality biological data, for which we could be as confident as possible that gene presence/absence in each genome reflected true biological presence or absence.

We pulled from NCBI's database of $\alpha$- and $\gamma$-proteobacterial genomes, requiring that genomes be complete (circular, single-chromosome), RefSeq-standard, and fully annotated. Phylogenies were constructed from concatenated orthologs that had fewer than two copies across all genomes and were present in at least 90% of the included taxa. This procedure produced phylogenies for both $\alpha$- and $\gamma$-proteobacteria (**Figure 2**), which were then used to evaluate model performance. We simulated incomplete data by artificially depleting the presence/absence matrices. A variety of completeness profiles were considered; although incompleteness arising from sequencing and assembly is likely more complex, this procedure preserves the generative process of evolutionary gene presence/absence followed by bioinformatic clustering.

At the per-gene level, our model achieves favorable precision and recall across a range of completeness profiles. The model generally performed better on $\gamma$-proteobacterial than on $\alpha$-proteobacterial datasets (**Figure 1**). It is unclear whether this difference is driven by the extreme genomic reduction observed in some $\alpha$-proteobacterial clades (42) or by the fact that the $\gamma$-proteobacterial dataset contains 252 more genomes. In general, precision increases as datasets become more incomplete; however, this effect is largely attributable to class imbalance rather than to the model itself. Nearly complete datasets contain fewer true positives that can be recovered. Nonetheless, across many completeness profiles, we can achieve approximately 40% recall at 90% precision, indicating that a substantial fraction of missing data can be recovered with reasonable confidence. In large-scale genomic inventory analyses, reporting these marginal probabilities may allow researchers to assign confidence to gene presence in individual genomes.

For core genome inference, our approach substantially outperforms alternative methods (**Figure 4, 5**). Joint reconstruction more reliably recovers core genomes than competing approaches, albeit with lower precision, whereas marginal-probability-based thresholding achieves superior precision and recall relative to empirical thresholding based on observed presence/absence. Although our methods outperform the others we considered here, reconstructing the core

genome from incomplete data remains a very difficult problem. For example, four taxa in our $\gamma$-proteobacterial dataset are missing COG0858: *Ahniella affigens*, *Billgrantia tianxiuensis*, *Luteimonas granuli*, and *Pseudomonas purpurea*. These taxa are not closely related on the phylogenetic tree, and in all four genomes, the gene has been pseudogenized through one or more point mutations. This orthogroup is typically present as a single copy and encodes the protein rbfA, which assists in maturation of the 30S ribosomal subunit in *Escherichia coli* (43). A reasonable prediction by our algorithm (or even by a biologist) would be that this gene is present in all sampled $\gamma$-proteobacterial genomes, but it is apparently absent in any functional form. Further modifications to our model may improve performance, but recent evolutionary events of this kind will likely remain difficult to predict. In its strictest formulation, core genome inference and gene presence/absence estimation are inherently difficult; some irreducible error is likely unavoidable in the strict core genome inference problem.

We would like to emphasize that although mOTUpan performs poorly in our evaluations, the deep-time phylogenetic relationships considered here are substantially different from the use case for which it was designed. Indeed, its original benchmarks focused on within-genus datasets of *Prochlorococcus* genomes (18). In such settings, genome-content heterogeneity may be limited enough to justify their modeling decisions. We anticipate that our model would also perform well in this setting.

**Asgardarchaea Analysis.** Eukaryotes are thought to root within the Asgardarchaea, more specifically as a sister to either the Heimdalarchaea or Hodarchaea, and the divergence of these species occurred approximately 2 billion years ago (44, 45). Previous analyses (46, 47) that examined the ancestral state of Asgardarchaea focused primarily on the metabolic nature of Heimdal- and Hodarchaea, rather than the proteins in their genomes that were previously thought to be eukaryote-specific (ESPs). These analyses suggest that the ancestors of Heimdal and Hod had metabolic capabilities similar to those of the crown species in those clades, consistent with our ancestral state reconstructions of those metabolic pathways. However, the metabolic conservation between crown and ancestral Asgard metabolic capabilities was not carried through to the eukaryotic-specific proteins found in crown Asgards (**Figure 7c**). These proteins were previously considered present exclusively in eukaryotes, until their identification in at least one Asgardarchaeal species, and are thought to have been inherited by eukaryotes from an Asgard ancestor. Our reconstruction indicates that all Asgard common ancestors had a similar number of ESPs ( 40% of the total ESPs identified in Asgard genomes), and although Hodarchaea have the most ESPs, there remains a significant gap between their ancestral and modern gene content. As shown in (**Figures 7a,7b**), many of these ESPs are present in modern crown Asgardarchaea but are absent in ancestral-state reconstructions, suggesting patchy gain and loss of these proteins across all Asgardarchaea families. Notably, functional roles of gene families in the ancestral-state reconstructions included actin regulation, ESCRT transport, and cell cycle checkpoint proteins, suggesting that ancestral Asgardarchaea were capable of membrane remodeling similar to that observed in *Lokiarchaeum ossiferum*. An analysis of gene families specific to the ancestor of Heimdalarchaea that were absent from the ancestor of all Asgard identified 33 ESPs with diverse functions, including vesicular transport, GTPase, and gyrase protein families. Although some discrepancies may be attributable to errors in the phylogeny, this pattern is observed throughout the tree. Further refinement of the model, as well as more complete genomes of diverse Asgard, will be able to improve this analysis to create a more accurate picture of the shared ancestor between Asgardarchaea and eukaryotes, though it is worth noting that Asgardarchaea are not fixed fossils in time, and it makes sense that they would continue to evolve with respect to gene content just as eukaryotes did.

**Future Directions.** In this work, our model design was guided primarily by interpreting the phylogeny as a belief network rather than by strict adherence to standard phylogenetic formulations. Our departures from classical phylogenetic models are relatively modest, most notably our decision not to use the stationary distribution of the transition matrix $\boldsymbol{Q}$ as the root distribution and our choice of transition probabilities that reflects agnosticism in the limit of time, rather than encoding dynamics. We anticipate that further deviations from classical phylogenetic assumptions may yield additional improvements in model performance. More broadly, we believe phylogenetic occupancy models represent a promising approach to improving the inference of gene presence-absence in incomplete genomic datasets. In addition to performing better on core genome inference for deep-time evolutionary datasets, these models also enable inference of the presence of individual genes in particular genomes and ancestral state reconstruction under data uncertainty. Although our phylogenetic occupancy model improves on existing approaches, several plausible directions could yield further improvements. These include:

- Explicitly accounting for uncertainty in phylogenetic structure. Reasonable approaches include "cut" Bayesian approaches and fully Bayesian approaches. Cut Bayesian models use the posterior distribution from an initial analysis as the true posterior distribution for a subsequent analysis (48). These methods are particularly useful when the secondary analysis contains information you do not want to flow into the posterior distribution of the first analysis. This is reasonable here because we may not believe that gene presence-absence should strongly sway our beliefs about the evolutionary relationships between organisms. Alternatively, fully Bayesian approaches may be reasonable and have been applied to several mixed-data scenarios (49).

- Incorporating gene co-occurrence information into the model. This poses a serious challenge, as it requires defining conditional distributions over binary strings

rather than simple binary states. In the simulations we explored, this would involve defining conditional distributions over $2^{5000}$ potential states. Approaches for phylogenetic modeling in this context do exist (50, 51), but remain computationally difficult. A potentially interesting approach would be to infer small cliques of genes that influence one another's presence/absence, which could help prevent the state space from becoming prohibitively large.

## Code Availability

All the code required to run the phylogenetic occupancy model can be found at https://github.com/Wesley-DeMontigny/Phylogenetic-Occupancy-Model. Additional files can be found at https://doi.org/10.5281/zenodo.18786143.

**ACKNOWLEDGEMENTS**

This research is funded in part by the Gordon and Betty Moore Foundation through Grant GBMF11481 (https://doi.org/10.37807/GBMF11481) to the University of Maryland (UMD). WCD was partially funded through a TAship from UMD. We would like to thank the attendees of the *Archaea: Ecology, Metabolism and Molecular Biology* Gordon conference for the helpful discussion on the model.

## Bibliography

1. Andrey V. Mardanov, Vadim V. Kadnikov, and Nikolai V. Ravin. Metagenomics: A paradigm shift in microbiology. In Meena Nagarajan, editor, *Metagenomics*, pages 1–15. Academic Press, 2025. doi: 10.1016/B978-0-323-91631-8.00006-8.
2. Robert Hooke. *Micrographia: Or Some Physiological Descriptions of Minute Bodies Made by Magnifying Glasses*. Jo. Martyn and Ja. Allestry, London, 1665.
3. Robert D. Fleischmann et al. Whole-genome random sequencing and assembly of Haemophilus influenzae rd. *Science*, 269:496–512, 1995. doi: 10.1126/science.7542800.
4. K. J. Pechman and Carl R. Woese. Characterization of the primary structural homology between the 16s ribosomal rnas of Escherichia coli and Bacillus megaterium. *Journal of Molecular Evolution*, 1:230–240, 1972. doi: 10.1007/BF01653965.
5. Carl R. Woese et al. Conservation of primary structure in 16s ribosomal rna. *Nature*, 254: 83–86, 1975. doi: 10.1038/254083a0.
6. George E. Fox, Linda J. Magrum, William E. Balch, Ralph S. Wolfe, and Carl R. Woese. Classification of methanogenic bacteria by 16s ribosomal rna characterization. *Proceedings of the National Academy of Sciences USA*, 74:4537–4541, 1977. doi: 10.1073/pnas.74.10.4537.
7. Carl R. Woese and George E. Fox. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences USA*, 74:5088–5090, 1977. doi: 10.1073/pnas.74.11.5088.
8. Gene W. Tyson et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428:37–43, 2004. doi: 10.1038/nature02340.
9. Jo Handelsman, Michelle R. Rondon, Sean F. Brady, Jon Clardy, and Robert M. Goodman. Molecular biological access to the chemistry of unknown soil microbes. *Chemistry & Biology*, 5:R245–R249, 1998. doi: 10.1016/S1074-5521(98)90108-9.
10. S M Barns, C F Delwiche, J D Palmer, and N R Pace. Perspectives on archaeal diversity, thermophily and monophyly from environmental rrna sequences. *Proceedings of the National Academy of Sciences*, 93(17):9188–9193, 1996. doi: 10.1073/pnas.93.17.9188.
11. Andrew D. Steen et al. High proportions of bacteria and archaea across most biomes remain uncultured. *ISME Journal*, 13:3126–3130, 2019. doi: 10.1038/s41396-019-0484-y.
12. Miguel Pignatelli et al. Metagenomics reveals our incomplete knowledge of global diversity. *Bioinformatics*, 24:2124–2125, 2008. doi: 10.1093/bioinformatics/btn343.
13. Itai Sharon and Jillian F. Banfield. Genomes from metagenomics. *Science*, 342:1057–1058, 2013. doi: 10.1126/science.1247023.
14. Paul A. Kitts, Deanna M. Church, Françoise Thibaud-Nissen, Jinna Choi, Vichet Hem, Victor Sapojnikov, Robert G. Smith, Tatiana Tatusova, Charlie Xiang, Andrey Zherikov, Michael DiCuccio, Terence D. Murphy, Kim D. Pruitt, and Avi Kimchi. Assembly: a resource for assembled genomes at ncbi. *Nucleic Acids Research*, 44(D1):D73–D80, 11 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv1226.
15. Duccio Medini, Claudio Donati, Hervé Tettelin, Vega Masignani, and Rino Rappuoli. The microbial pan-genome. *Current Opinion in Genetics & Development*, 15:589–594, 2005. doi: 10.1016/j.gde.2005.09.006.
16. Andrew J. Page et al. Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31:3691–3693, 2015. doi: 10.1093/bioinformatics/btv421.
17. Guillaume Gautreau, Adelme Bazin, Mathieu Gachet, Rémi Planel, Laura Burlot, Mathieu Dubois, Amandine Perrin, Claudine Médigue, Alexandra Calteau, Stéphane Cruveiller, Catherine Matias, Christophe Ambroise, Eduardo P. C. Rocha, and David Vallenet. PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLOS Computational Biology*, 16(3):e1007732, March 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007732.
18. Matthias Buck, Maliheh Mehrshad, and Stefan Bertilsson. motupan: A robust bayesian approach to leverage metagenome-assembled genomes for core-genome estimation. *NAR Genomics and Bioinformatics*, 4:lqac060, 2022. doi: 10.1093/nargab/lqac060.
19. Darryl I. MacKenzie, James D. Nichols, Gideon B. Lachman, Sam Droege, J. Andrew Royle, and Catherine A. Langtimm. Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8):2248–2255, 2002. doi: https://doi.org/10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2.
20. Darryl I. MacKenzie et al. *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence*. Academic Press, Burlington, MA, 2006.
21. Larissa L. Bailey, Darryl I. MacKenzie, and James D. Nichols. Advances and applications of occupancy models. *Methods in Ecology and Evolution*, 5(12):1269–1279, 2014. doi: https://doi.org/10.1111/2041-210X.12100.
22. Kadambari Devarajan, Toni Lyn Morelli, and Simone Tenan. Multi-species occupancy models: review, roadmap, and recommendations. *Ecography*, 43(11):1612–1624, 2020. doi: https://doi.org/10.1111/ecog.04957.
23. Wilson J. Wright, Kathryn M. Irvine, Thomas J. Rodhouse, and Andrea R. Litt. Spatial Gaussian processes improve multi-species occupancy models when range boundaries are uncertain and nonoverlapping. *Ecology and Evolution*, 11(13):8516–8527, July 2021. ISSN 2045-7758, 2045-7758. doi: 10.1002/ece3.7629.
24. Judea Pearl. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29(3):241–288, 1986. ISSN 0004-3702. doi: https://doi.org/10.1016/0004-3702(86)90072-X.
25. Peter R. Thompson, William F. Fagan, and Phillip P. A. Staniczenko. Predictor species: Improving assessments of rare species occurrence by modeling environmental co-responses. *Ecology and Evolution*, 10(7):3293–3304, April 2020. ISSN 2045-7758, 2045-7758. doi: 10.1002/ece3.6096.
26. Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008. ISSN 1935-8237. doi: 10.1561/2200000001.
27. Kevin P. Murphy. *Probabilistic machine learning: advanced topics*. Adaptive computation and machine learning series. The MIT Press, Cambridge, Massachusetts, 2023. ISBN 978-0-262-37599-3 978-0-262-37600-6.
28. Ziheng Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, 39(3):306–314, September 1994. ISSN 1432-1432. doi: 10.1007/BF00160154.
29. Joseph Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, November 1981. ISSN 1432-1432. doi: 10.1007/BF01734359.
30. N. L. Zhang and D. Poole. Exploiting causal independence in bayesian network inference, 1996.
31. Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20:28:1–28:6, 2019.
32. Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro, December 2019. arXiv:1912.11554 [stat].
33. Hernandez-Plaza Cantalapiedra et al. eggnog-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution*, 38:5825–5829, 2021. doi: 10.1093/molbev/msab293.
34. Szklarczyk Huerta-Cepas et al. eggnog 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47:309–314, 2019. doi: 10.1093/nar/gky1085.
35. Katoh and Standley. Mafft multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30:772–780, 2013. doi: 10.1093/molbev/mst010.
36. Silla-Martinez Capella-Gutierrez et al. trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25:1972–1973, 2009. doi: 10.1093/bioinformatics/btp348.
37. Schmidt Minh et al. Iq-tree 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, 37:1530–1534, 2020. doi: 10.1093/molbev/msaa131.
38. Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30:2068–2069, 2014. doi: 10.1093/bioinformatics/btu153.
39. Yang Liu et al. Expanded diversity of asgard archaea and their relationships with eukaryotes. *Nature*, 593:553–557, 2021. doi: 10.1038/s41586-021-03517-8.
40. Smith Yu et al. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8:28–36, 2016. doi: 10.1111/2041-210X.12628.
41. Raivo Kolde. *pheatmap: Pretty Heatmaps*, 2025. R package version 1.0.13.
42. Björn Sällström and Siv GE Andersson. Genome reduction in the α-proteobacteria. *Current Opinion in Microbiology*, 8(5):579–585, 2005. ISSN 1369-5274. doi: https://doi.org/10.1016/j.mib.2005.08.002. Antimicrobials / Edited by Malcolm Page and Christopher T Walsh · Genomics / Edited by Stephan C Schuster and Gerhard Gottschalk.
43. Bing Xia, Haiping Ke, Ujwal Shinde, and Masayori Inouye. The role of rbfa in 16s rrna processing and cell growth at low temperature in escherichia coli. *Journal of Molecular Biology*, 332(3):575–584, 2003. ISSN 0022-2836. doi: https://doi.org/10.1016/S0022-2836(03)00953-7.
44. Tom Williams et al. Phylogenomics provides robust support for a two-domains tree of life. *Nature Ecology & Evolution*, 4:138–147, 2020. doi: 10.1038/s41559-019-1040-x.
45. Holly Betts et al. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nature Ecology & Evolution*, 2:1556–1562, 2018. doi: 10.1038/s41559-018-0644-x.
46. Laura Eme et al. Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature*, 618:992–999, 2023. doi: 10.1038/s41586-023-06133-1.
47. Feng Zhang et al. Deep origin of eukaryotes outside heimdallarchaeia within asgardarchaeota. *Nature*, 642:990–998, 2025. doi: 10.1038/s41586-025-08955-7.
48. Yang Liu and Robert J B Goudie. A general framework for cutting feedback within mod-

ularized bayesian inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4):1171–1199, March 2025. ISSN 1467-9868. doi: 10.1093/jrsssb/qkaf012.

49. Max R Tolkoff, Michael E Alfaro, Guy Baele, Philippe Lemey, and Marc A Suchard. Phylogenetic Factor Analysis. *Systematic Biology*, 67(3):384–399, May 2018. ISSN 1063-5157, 1076-836X. doi: 10.1093/sysbio/syx066.
50. D. M. Robinson. Protein Evolution with Dependence Among Codons Due to Tertiary Structure. *Molecular Biology and Evolution*, 20(10):1692–1704, June 2003. ISSN 0737-4038, 1537-1719. doi: 10.1093/molbev/msg184.
51. Monir Hajiaghayi, Bonnie Kirkpatrick, Liangliang Wang, and Alexandre Bouchard-Côté. Efficient continuous-time markov chain estimation, 2014.

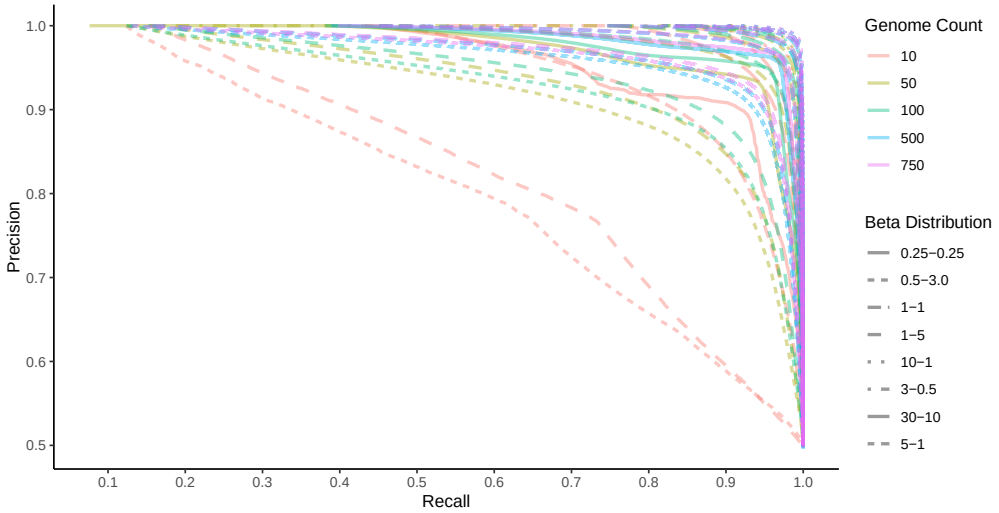
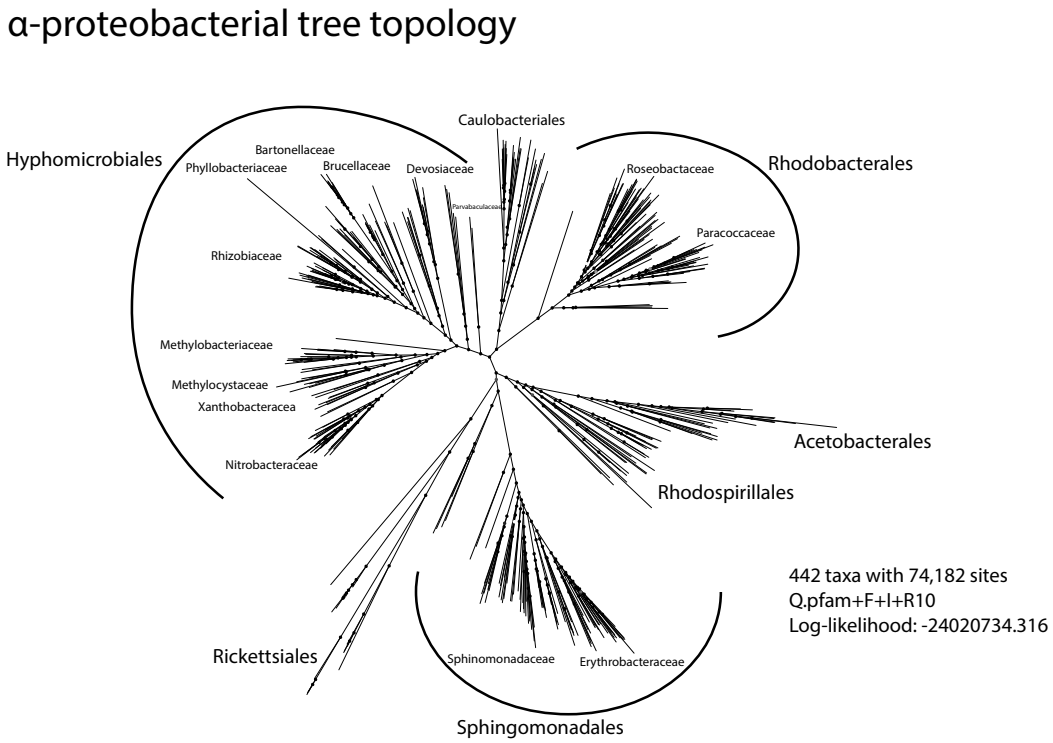
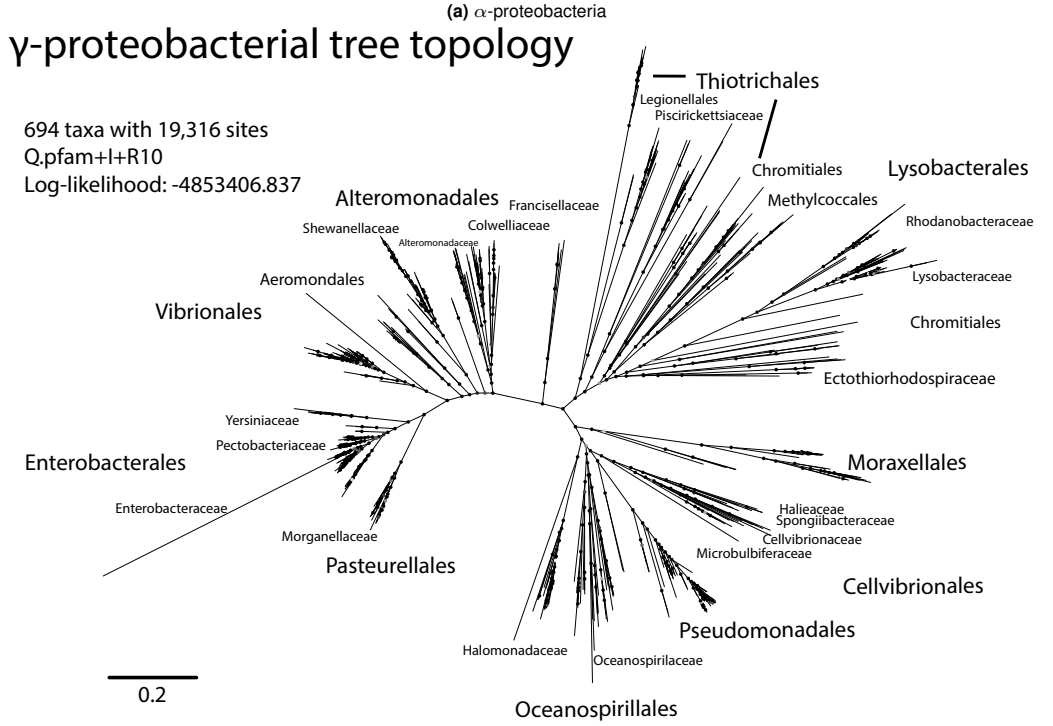



**Fig. 1.** Precision and recall of presence/absence data simulated and run under our model at varying completeness profiles (Beta distributions) and number of genomes in the dataset. These metrics were only calculated on unobserved orthogroups; known presences do not contribute toward the precision and recall. This reflects the model's performance in single-gene presence/absence imputation under idealized conditions.




**(a)** $\alpha$-proteobacteria




**(b)** $\gamma$-proteobacteria

**Fig. 2.** Phylogenetic relationships of $\alpha$- and $\gamma$-proteobacteria. Maximum-likelihood phylogenies were reconstructed using concatenated single-copy genes present in 98% of taxa analyzed, and models were inferred using IQTree ModelFinder. Top: $\alpha$-proteobacterial phylogeny inferred from 442 taxa and 74,182 aligned amino acid positions (log-likelihood = 2,402,0734.316). Major orders are indicated, including Hyphomicrobiales, Rhodobacterales, Caulobacterales, Rhodospirillales, Sphingomonadales, Rickettsiales, and Acetobacterales. Bottom: $\gamma$-proteobacterial phylogeny inferred from 694 taxa and 19,316 aligned amino acid positions (log-likelihood = 4,853,406.837). Major orders are labeled, including Enterobacterales, Vibrionales, Alteromonadales, Pseudomonadales, Oceanospirillales, Moraxellales, Cellvibrionales, Chromatiales, and Thiotrichales. Branch lengths represent the expected number of substitutions per site (scale bar shown). Trees are displayed as unrooted radial topologies for visualization of overall clade structure.

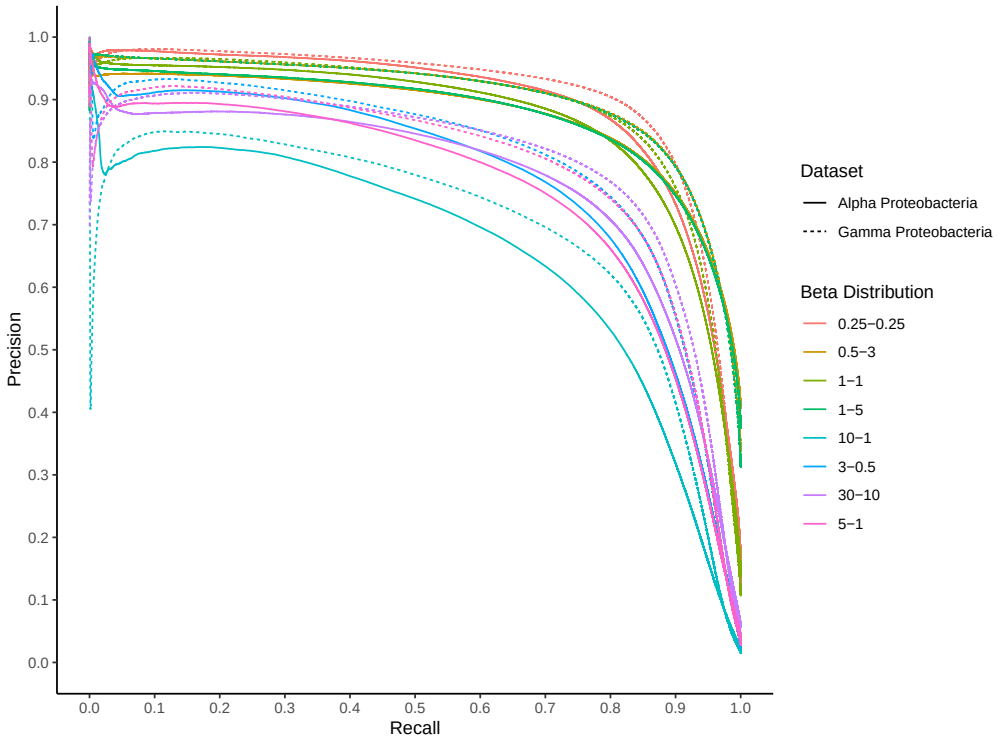
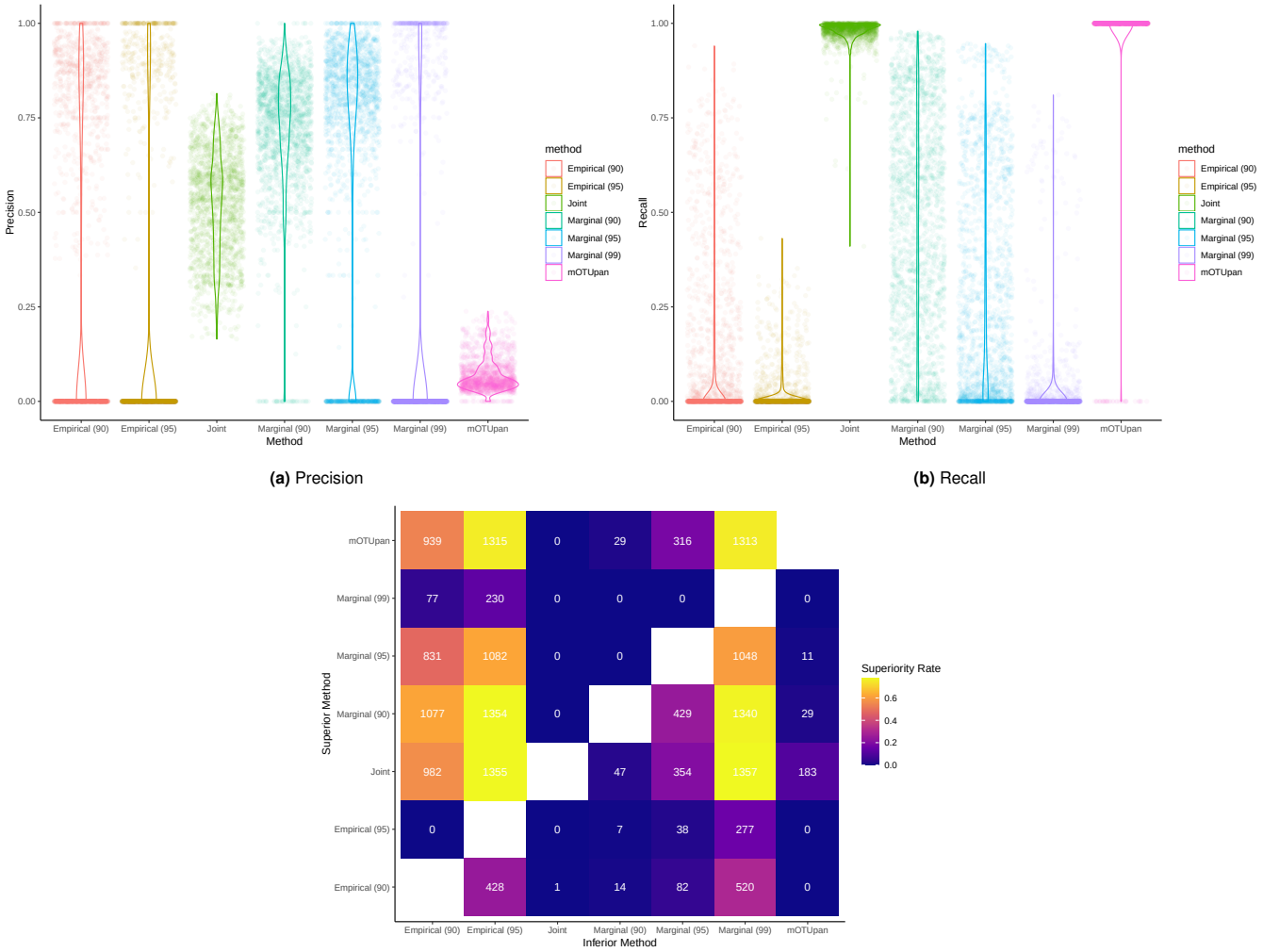



**Fig. 3.** Precision and recall of our model for artificially depleted $\alpha$-proteobacteria and $\gamma$-proteobacteria genomes under varying completeness profiles (Beta distributions). These metrics were only calculated on unobserved orthogroups; known presences do not contribute toward the precision and recall. This reflects the model's performance in single-gene presence/absence imputation on two real datasets with some gene content masked.

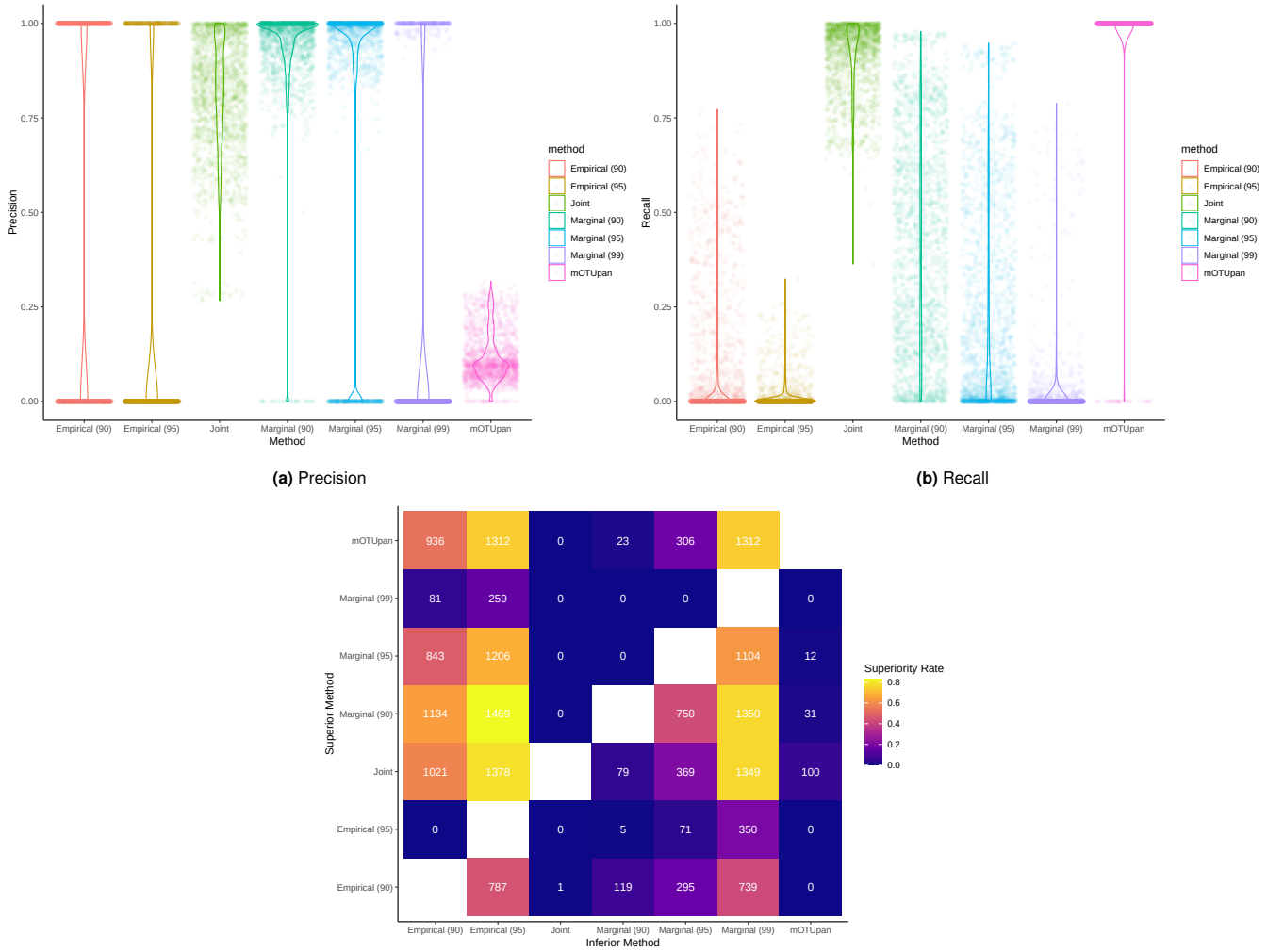



**(a)** Precision

**(b)** Recall

**(c)** Superiority Matrix

**Fig. 4.** Performance metrics for strict core genome inference in the $\alpha$- and $\gamma$-proteobacteria analyses. Core genomes were inferred for clades within the proteobacteria trees and predicted using: empirical thresholding (denoted "Empirical $(\cdot)$") of observed orthogroup presence/absence (i.e., the orthogroup is observed in 90% of incomplete genomes); joint maximum *a posteriori* reconstruction (denoted "Joint") from our model (i.e., the maximum *a posteriori* reconstruction assigns the gene to all genomes in the clade); thresholding of marginal presence/absence probabilities (denoted "Marginal $(\cdot)$") from our model (i.e., the gene has greater than 90% marginal probability of presence in every genome in the clade); and mOTUpan core genome inference. Precision and recall for each method are shown in (a) and (b). The superiority matrix (c) summarizes pairwise comparisons, indicating the frequency with which each method is strictly better than the others (better in precision and recall).




**(a)** Precision

**(b)** Recall

**(c)** Superiority Matrix

**Fig. 5.** Performance metrics for relaxed core genome inference in the $\alpha$- and $\gamma$-proteobacteria analyses. In these analyses, the definition of a "core" gene is any gene that is truly present in over 90% of the clade. Core genomes were inferred for clades within the proteobacteria trees and predicted using: empirical thresholding (denoted "Empirical $(\cdot)$") of observed orthogroup presence/absence (i.e., the orthogroup is observed in 90% of incomplete genomes); joint maximum *a posteriori* reconstruction (denoted "Joint") from our model (i.e., the maximum *a posteriori* reconstruction assigns the gene to all genomes in the clade); thresholding of marginal presence/absence probabilities (denoted "Marginal $(\cdot)$") from our model (i.e., the gene has greater than 90% marginal probability of presence in every genome in the clade); and mOTUpan core genome inference. Precision and recall for each method are shown in (a) and (b). The superiority matrix (c) summarizes pairwise comparisons, indicating the frequency with which each method is strictly better than the others (better in precision and recall).

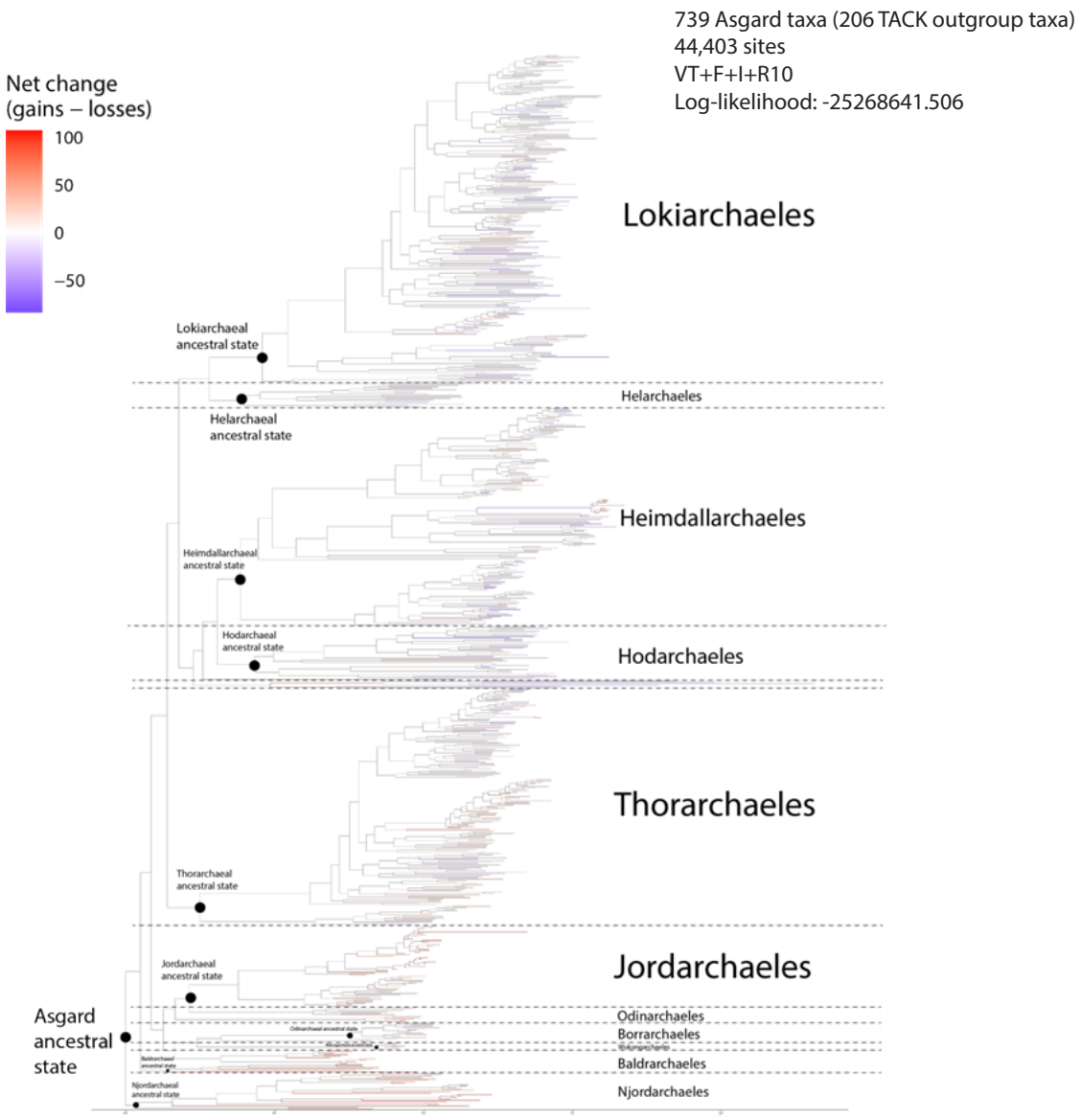
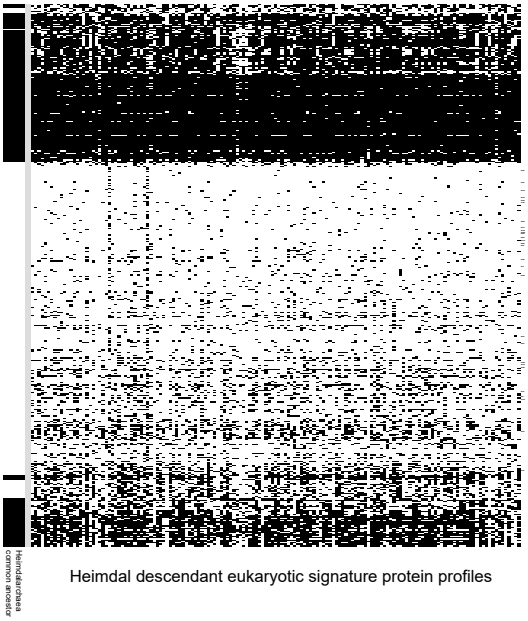
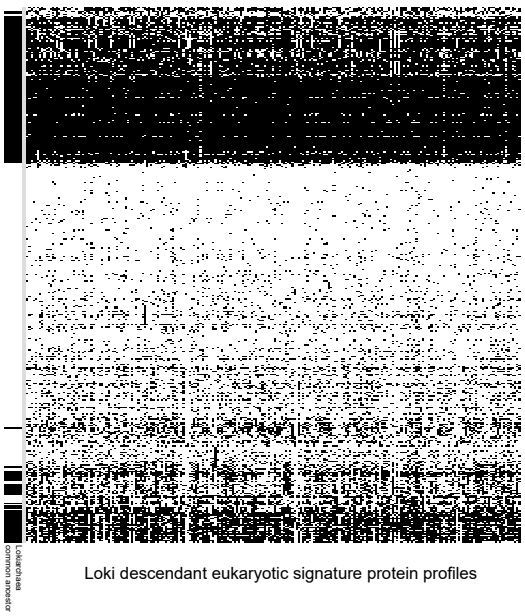


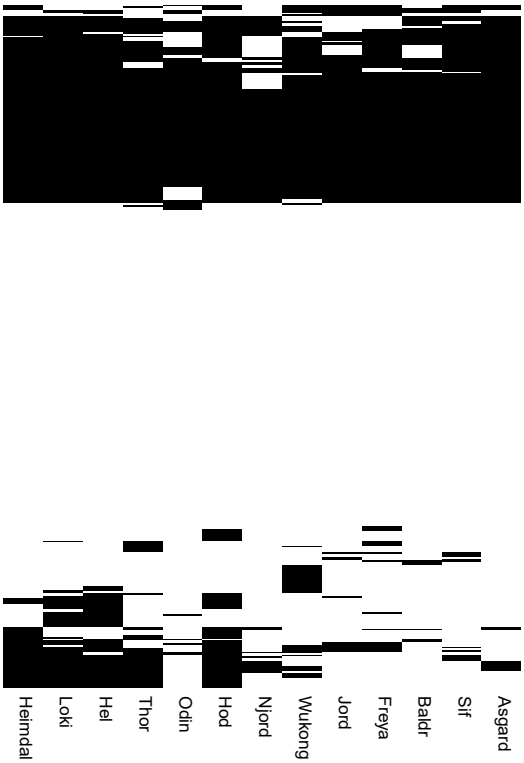

**Fig. 6.** Asgard ancestral reconstruction of eukaryotic-specific protein gain-loss. Maximum-likelihood phylogeny of 739 Asgard taxa (including 206 TACK outgroup taxa not displayed in this tree) inferred from 44,403 aligned amino acid positions under the VT+F+I+R10 substitution model (log-likelihood = 25,268,641.506). Major Asgard clades are indicated, including Lokiarchaeales, Heimdallarchaeales, Hodarchaeales, Thorarchaeales, and Jordarchaeales, with additional subclades labeled where relevant. Net changes in eukaryotic-specific protein families were constructed using ancestral state inference and plotted using branch colors to indicate the direction and magnitude of the change in ESP presence. Black circles denote inferred ancestral nodes for major clades analyzed in Figure 7, including the overall Asgard ancestral state and order-level ancestors. Branch lengths are proportional to the expected number of substitutions per site.




**(a)** Heimdalarchaeota Eukaryotic Signature Proteins Predicted in the Ancestor and Modern Crown Species




**(b)** Lokiarchaeota Eukaryotic Signature Proteins Predicted in the Ancestor and Modern Crown Species

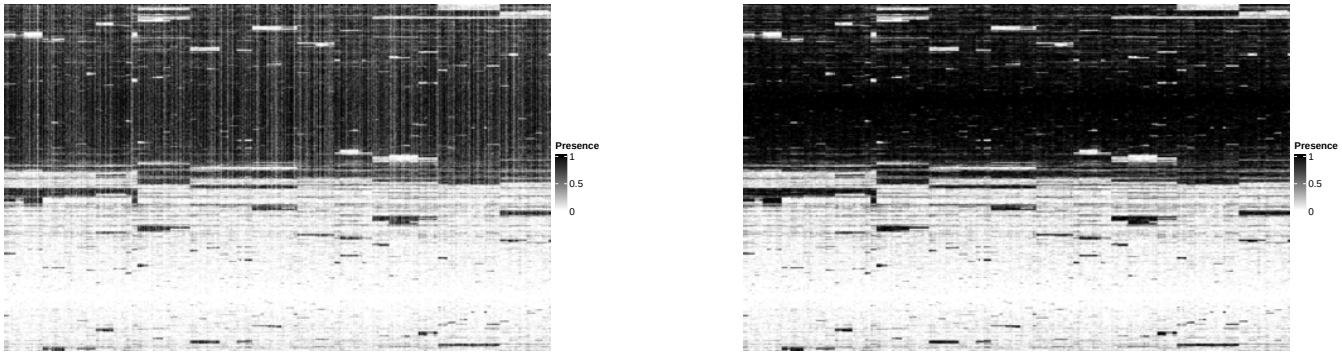



**(c)** Predicted Eukaryotic Signature Protein Profile of All Major Asgard Clades

**Fig. 7.** Ancestor and Crown Eukaryotic Signature Protein Profiles in Asgard Archaea after Occupancy Model Predictions of Crown gene content. Heatmaps summarize the presence–absence patterns of predicted eukaryotic signature proteins (ESPs) inferred using occupancy-based ancestral reconstruction of crown gene content. Black cells indicate predicted presence, white cells indicate absence. Rows correspond to individual ESP families; columns correspond to taxa or reconstructed ancestral nodes, as indicated below each panel. (a) ESP profiles for Heimdallarchaeota, showing predicted content in the Heimdallarchaeal ancestor alongside extant crown species. (b) ESP profiles for Lokiarchaeota, showing predicted content in the Lokiarchaeal ancestor compared to modern crown representatives. (c) Condensed ESP profile matrix summarizing reconstructed ancestral and crown gene complements across all major Asgard clades. Notably, in all Asgard ancestral cases a large proportion of ESPs are absent from the inferred ancestor despite their presence in one or more crown species members.

# Supplemental Materials: Core Genome Inference with Phylogenetic Occupancy Modeling

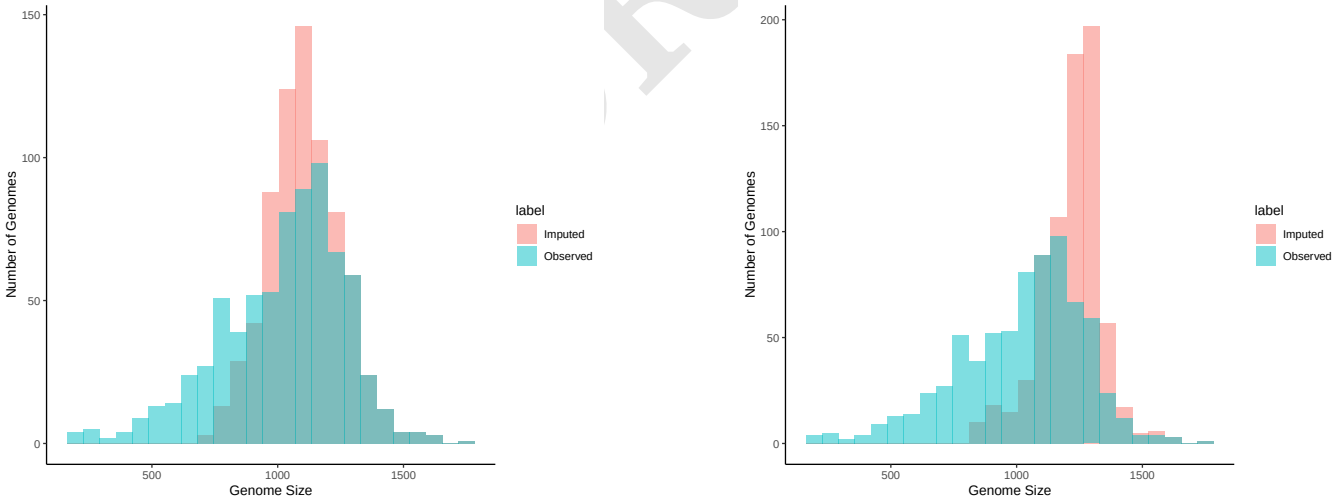
DRAFT




**(a)** Observed Presence/Absence **(b)** Ground Truth

**Fig. S1.** Example of observed data (a) and ground truth (b) from our simulation analyses. Specifically, this simulation had 750 genomes and 5,000 genes and was depleted according to a Beta$(5, 1)$ distribution.

DRAFT




**(a)** Greater 90% Marginal Probability Thresholding **(b)** Maximum *a Posteriori*

**Fig. S2.** Histogram representing frequencies of genome size (in orthogroup counts) of Asgardarchaea genomes before and after gene imputation using (a) 90% marginal probability thresholding and (b) maximum *a posteriori* reconstruction.