1 Overview

Scientists worldwide are putting together massive efforts to understand how the biodiversity that we see on Earth evolved from a single-cell organism at the origin of life. This diversification process is represented by the Tree of Life which in mathematical terms is a fully bifurcating tree in which internal nodes represent ancestral species that over time differentiate into two separate species giving rise to its two children nodes. Recently, scientists have challenged the notion that evolution can be represented with a fully bifurcating process, as this process cannot capture important biological realities like hybridization, introgression or horizontal gene transfer. Thus, recent years have seen an explosion of methods to reconstruct phylogenetic networks, which naturally account for reticulate evolution. This proposal will produce a novel suite of statistical theory and computational methods for the reconstruction of phylogenetic networks that will extend existing methods in three ways (Figure 1): 1) by allowing metagenomes as input data and appropriately propagating statistical error in the assembly and classification into the phylogenetic estimates, 2) by estimating more complex networks suitable for eukaryotes and prokaryotes alike with solid statistical guarantees of identifiability and convergence, 3) by improving the scalability via pseudolikelihood and divide-and-conquer techniques to produce large networks with hundreds or thousands of taxa.

Objective 1. Estimation of phylogenetic networks from metagenomic data accounting for statistical uncertainty in the pre-processing steps. Existing work on phylogenetic networks use genomic data as input and assumes that the genomes have been assembled and aligned without error. We will design the first inference method to reconstruct phylogenetic networks from metagenomic data by explicitly modeling statistical uncertainty in every stage of the pipeline from raw data to phylogeny which is expected to produce more robust phylogenetic estimates. Our new method will be highly applicable given the frequent use of metagenomic data in microbial studies.

Objective 2. Extension of pseudolikelihood estimation to broader classes of phylogenetic networks. Pseudolikelihood estimation is among the more scalable alternatives for the reconstruction of phylogenetic networks. However, existing pseudolikelihood theory is restricted to level-1 networks [55, 56, 57] which are not biologically reasonable by not allowing hybridization events to intersect. We will extend our pseudolikelihood model to level-2 and tree-child networks by studying the statistical identifiability of these classes of networks and identifying efficient traversal strategies in network space that can guarantee a solid inference framework when estimating complex networks suitable for eukaryotes, prokaryotes and viruses.

Objective 3. Scalable divide-and-conquer algorithms for the inference of large phylogenetic networks. Despite the scalability of the pseudolikelihood model, inference is still restricted to dozens of taxa. We will produce a novel divide-and-conquer algorithm to reconstruct large phylogenetic networks from smaller subnetworks by decomposing networks on sets of parental trees and utilize existing merging strategies on trees. We will produce the first algorithm to build a phylogenetic network from its set of parental trees.

Objective 4. Implementation of novel easy-to-use, open-source software. New theory will be implemented in new user-friendly software, with extensive documentation and step-by-step tutorials. The new software will serve the broad evolutionary biology scientific community (as existing software by the PI already does [55, 56]), and will require no technical programming expertise.

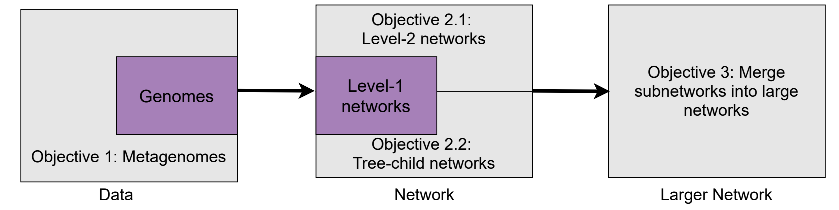
Objective 1: Metagenomes
Genomes
Objective 2.1: Level-2 networks
Level-1 networks
Objective 2.2: Tree-child networks
Objective 3: Merge subnetworks into large networks
Data
Network
Larger Network

Figure 1: **Research Objectives in the context of existing work.** This proposal will extend existing work by the PI (purple rectangles) focused on the estimation of level-1 phylogenetic networks from genomic data [58, 55, 56, 6, 57]. While there are indeed other network methods that infer complex networks [69, 71, 51, 66], i) none use metagenomes as input data (Obj. 1); ii) no method has studied the identifiability or statistical guarantees of network inference on these complex networks (Obj. 2); iii) there are no existing network methods that can handle large datasets of hundreds or thousands of taxa (Obj. 3).

Intellectual Merit. The development of statistical theory that can help reconstruct the evolutionary history of life, especially when flexible to incorporate reticulate evolution and scalable to big data, is paramount in evolutionary biology, systematics, and biodiversity conservation efforts. This proposal will contribute to the fundamental research of the evolutionary history of life by producing three entirely novel scientific outcomes with broad scientific reach: 1) the first phylogenomic inference method tailored to metagenomic data to estimate the evolutionary history of complex fungal, prokaryotic or viral communities, 2) the first statistical theory on identifiability of complex phylogenetic networks and divide-and-conquer algorithms to produce the most scalable to date inference procedures to meet the ever growing needs of biological big data, and 3) open-source publicly available software with broad applicability and outreach that will allow evolutionary biologists to apply our new methods on their own data.

Education Plan: Improving statistical education in biological sciences and enhacing diversity in STEM. Computational and statistical skills are no longer optional learning topics for biology-oriented students in the XXI century. We will contribute to the development of novel statistical educational tools for middle, high school and college level biology students via the following innovative goals:

Goal 1. Creation of Data Science educational modules for the WI Fast Plants K-16 educational framework. WI Fast Plants [49] were developed as a research tool at the University of Wisconsin-Madison and have been used by K-16 teachers around the world for nearly 30 years with estimates of over 20,000 classrooms using fast plants as an educational model-organism [46]. Yet data collected from plant experiments tend not to be analyzed due to the lack of user-friendly statistical tools for K-16 teachers and students. Our early process created the foundation of WI Fast Stats [40, 60], the first and only statistical web apps tailored to the WI Fast Plants educational objectives. Now, we will produce statistical educational modules to accompany the WI Fast Stats web apps and WI Fast Plants evolutionary lessons so that K-16 teachers and students can learn the basics of Data Science with educational YouTube videos and materials to study statistical concepts with or without access to physical WI Fast Plants.

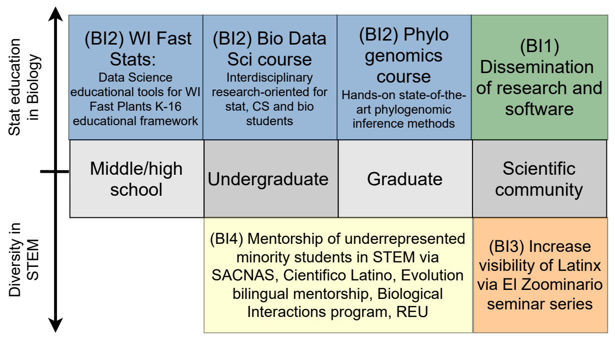
Stat education in Biology
Diversity in STEM
(BI2) WI Fast Stats:
Data Science educational tools for WI Fast Plants K-16 educational framework
(BI2) Bio Data Sci course
Interdisciplinary research-oriented for stat, CS and bio students
(BI2) Phylo genomics course
Hands-on state-of-the-art phylogenomic inference methods
(BI1) Dissemination of research and software
Middle/high school
Undergraduate
Graduate
Scientific community
(BI4) Mentorship of underrepresented minority students in STEM via SACNAS, Cientifico Latino, Evolution bilingual mentorship, Biological Interactions program, REU
(BI3) Increase visibility of Latinx via El Zoominario seminar series

Figure 2: **Proposed Broader Impact activities.** Under two principles (improvement of statistical education in biology and enhancement of diversity in STEM), our Broader Impacts goals involve 1) the creation of Data Science educational modules for the WI Fast Plants K-16 educational framework and 2) design of a novel Biological Data Science interdisciplinary class for undergraduate students in statistics, CS and biology. In addition, I will continue my efforts on graduate student training, dissemination of research and mentorship of underrepresented minority students in STEM. My proposed activities target four NSF outcomes for Broader Impacts (BI1 in blue, BI2 in green, BI3 in orange, BI4 in yellow; see main text).

Goal 2. Creation of a new interdisciplinary research-based undergraduate-level class for Biological Data Science. Scientific research is becoming increasingly interdisciplinary yet undergraduate-level education has not adapted fast enough. In particular, computer science, statistics and biology students rarely interact in the same courses. We will create a class that will combine students from these three fields to tackle research projects together in powerful interdisciplinary teams. The class will focus on good scientific and computational practices, reproducibility, science communication in addition to data science and programming skills.

In addition to these two innovative goals, we will continue the following Broader Impacts activities:

1. Training of graduate students in and outside the lab in statistical phylogenomics, reproducibility and best computing practices for data science on big data
2. Creation of open-source, easy-to-use, publicly available software for the biological scientific community
3. Organization of workshops and tutorials on our software tools for maximum outreach
4. Research dissemination via publications and conference presentations targeting organismal conferences in addition to standard broad conferences like Evolution
5. Mentorship of females and underrepresented minorities in STEM
6. Public engagement via El Zoominario, a seminar series that I co-created to increase visibility of Latinx people in STEM and to inspire the next generation of Latinx STEM scientists

Broader Impacts. My proposed activities target four NSF outcomes for Broader Impacts (BI) (Figure 2):

BI1. Enhanced infrastructure for research and education.
BI2. Improved STEM education and educator development at any level.
BI3. Increased public scientific literacy and public engagement with science and technology.
BI4. Full participation of women, persons with disabilities, and underrepresented minorities in STEM.

Unifying principles in the research and education plan. Both the research and the education plan stem from the notion of incorporating solid and rigorous statistical theory into evolutionary biology via 1) the study and development of novel statistical theory and computational methods that will serve the broader evolutionary biology community in their efforts to reconstruct the Tree of Life and 2) the production of Data Science educational tools for K-16 education and beyond that provide necessary tools and training to meet the statistical needs of biological students in the XXI century.

2 Background and significance

The abundance of gene flow in the Tree of Life [4, 5, 45] challenges the notion that evolution can be represented with a fully bifurcating process, as this process cannot capture important biological realities like hybridization, introgression or horizontal gene transfer. Thus, recent years have seen an explosion of methods to reconstruct phylogenetic networks, which naturally account for reticulate evolution [30, 12, 14, 9]. Phylogenetic networks are thus an extension to the tree structure by the explicit modeling of gene flow (see Figure 3 and Notation below). Among the many methods to estimate phylogenetic networks, probabilistic alternatives tend to be more accurate given that they aim to explicitly model all the biological forces that shape the genomes that we observe, e.g. mutations, incomplete lineage sorting (ILS), gene flow, duplications, losses, and recombination. Currently, there is no method that can account for all biological processes simultaneously, but network methods that simultaneously account for ILS and gene flow under the coalescent model are among the most widely used [38, 41, 69, 55, 66, 65, 71, 51, 15, 35].

Despite their popularity, coalescent-based network methods have two main weaknesses. First, these methods are still not scalable enough to meet the big data needs of evolutionary biology in the XXI century. While tree estimation methods are more scalable and able to consistently estimate trees with hundreds of taxa [36], it is not advised to first reconstruct a tree and then add gene flow events between lineages given the lack of robustness of tree methods under gene flow [58, 73] and thus, we need network methods to simultaneously estimate the underlying tree topology and gene flow events. Unfortunately, maximum likelihood [69] and Bayesian network methods [66, 71] can only tackle less than a dozen taxa at a time, and even pseudolikelihood approaches [55, 70] have not been able to reconstruct networks with more than 50 taxa. Second, the identifiability of phylogenetic networks under these models is not fully understood. Lack of identifiability implies that the data is unable to detect the network model that generated it. There has been a lot of work on the identifiability of phylogenetic networks from displayed (sub)trees [64, 27, 16, 48], yet displayed trees are not relevant for the network coalescent model due to gene tree discordance [74], and thus, studies on the identifiability of phylogenetic networks under the coalescent model are still needed.

The PI has pioneered the development of statistical theory on the identifiability of phylogenetic networks under the pseudolikelihood model which is more scalable than the standard likelihood model. The PI created the software SNaQ [55] within the Julia package PhyloNetworks [56] (third most widely used Julia package in genomics) which, based on citations and personal communications, is constantly used by evolutionary biologists worldwide. The pseudolikelihood model in SNaQ has five main advantages compared to standard likelihood or Bayesian network methods: 1) it is built on a strict statistical theory of network identifiability, 2) it is more scalable by bypassing the computation of the whole likelihood through the agglomeration of quartet likelihoods, 3) it is more robust to gene tree estimation error by using concordance factors (see [7] and Notation below) as input rather than assuming the gene trees are perfecly known, 4) it is more robust to rooting error by not requiring the input gene trees to be rooted, and 5) it is more robust to molecular clock deviations by not requiring branch lengths on input gene trees.

Despite its advantages and broad use, the SNaQ pseudolikelihood model has three main areas of improvement that will be addressed in this proposal. First, the model currently needs perfect alignments as input which limits its applicability to microbial data generally metagenomic in nature (Objective 1). Second, the model has a limiting assumption on the type of network that it can reconstruct (more in Section 2.2) which limits its applicability to prokaryotes or viruses who have a complex evolutionary history with rampant gene flow at all depths of the network (Objective 2). Third, even the improved scalability of the pseudolikelihood model is not enough to tackle large networks of hundreds of taxa (Objective 3).

Significance. The successful completion of the research objectives will generate statistical tools with theoretical guarantees to estimate phylogenetic networks that are complex enough to represent the evolutionary history of all organisms across the Tree of Life. The novel tools will be scalable enough to estimate truly large networks in the order of hundreds or thousands of taxa which are currently prohibited by all existing network methods and will be implemented in open source, easy to use, publicly available software that will profit the broad scientific community in evolutionary biology. Accurate, robust and scalable inference of phylogenetic networks will also serve for downstream analysis of computation of phylogenetic diversity or UniFrac scores in microbiome studies and trait evolution in comparative methods [34, 6].

2.1 Notation

A rooted phylogenetic network $\mathcal{N}$ on taxon set X ($|X| = n$) is a connected directed acyclic graph with vertices $V = \{r\} \cup V_L \cup V_H \cup V_T$, edges $E = E_H \cup E_T$ and a bijective leaf-labeling function $f : V_L \to X$ with the following characteristics. The root r has indegree 0 and outdegree 2. Any leaf $v \in V_L$ has indegree 1 and outdegree 0. Any tree node $v \in V_T$ has indegree 1 and outdegree 2. Any hybrid node $v \in V_H$ has indegree 2 and outdegree 1. A tree edge $e \in E_T$ is an edge whose child is a tree node. A hybrid edge $e \in E_H$ is an edge whose child is a hybrid node. Unrooted phylogenetic networks are typically obtained by suppressing the root node and the direction of all edges. In our work, we focus on *semi-directed networks*, where the root node is suppressed and we ignore the direction of all tree edges, but we maintain the direction of hybrid edges, thus keeping information on which nodes are hybrids (Figure 3 bottom). The placement of the root is then constrained, because the direction of the two hybrid edges to a given hybrid node informs the direction of time at this node: the third edge must be a tree edge directed away from the hybrid node and leading to all the hybrid's descendants. Therefore, the root cannot be placed on any descendant of any hybrid node, although it might be placed on some hybrid edges. Like phylogenetic trees, semi-directed networks can be rooted by a known outgroup.

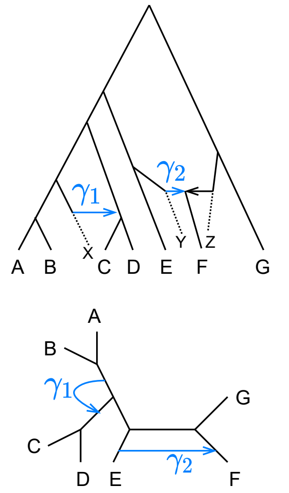

Figure 3: Example of rooted and semi-directed phylogenetic networks with $h = 2$ hybridization events and $n = 7$ sampled taxa. Inheritance probabilities γ represent the proportion of genes contributed by each parental population to a given hybrid node. Top: rooted network modeling several biological processes. Taxon F is a hybrid between two non-sampled taxa Y and Z with $\gamma_2 \approx 0.50$, and the lineage ancestral to taxa C and D has received genes introgressed from a non-sampled taxon X, for which $\gamma_1 \approx 0.10$. Bottom: semi-directed network for the biological scenario just described.

Each hybridization event in the network creates a hybridization cycle (or blob [29]). Figure 4 illustrates the difference between a level-1 network in which hybridization cycles cannot intersect, and a level-2 network in which the cycles can (up to a certain degree). Intuitively, a biconnected component can be seen as an intersection of cycles (in red in Figure 4), and a level-2 network is a network whose biconnected components have at most 2 hybridization nodes [29]. Tree child networks are networks in which every internal node has at least one child node that is a tree node [29]. For example, a network in which a given internal node has two hybrid nodes as children is not a tree child network.

Throughout this work, we denote by n the number of taxa, and h the number of hybridization events in the network. For example, in Figure 3 $n = 7, h = 2$. The main parameter of interest is the topology $\mathcal{N}$ of the semi-directed network, the vector of branch lengths ($\mathbf{t}$), and the vector of inheritance probabilities (γ), describing the proportion of genes inherited by a hybrid node from one of its hybrid parent.

Note that phylogenetic networks can describe various biological processes causing gene flow from one population to another such as hybridization, introgression, or horizontal gene transfer. Hybridization occurs when individuals from 2 genetically distinct populations interbreed, resulting in a new separate population. Introgression, or introgressive hybridization, is the integration of alleles from one population into another existing population, through hybridization and backcrossing. Genes are horizontally transferred when acquired by a population through a process other than reproduction, from a possibly distantly related population. Although these three processes are biologically different, we do not make the distinction when modeling them with a network. In other words, our model takes into account all three biological scenarios, but those scenarios are not distinguishable in the estimated phylogenetic network unless more biological information is provided. Note that this is a standard assumption of all coalescent-based network methods.

2.2 Prior work

In this proposal, we focus on the PI's pseudolikelihood inference method [55]. The standard phylogenetic analysis pipeline involves 1) the estimation of gene trees $\hat{\mathbf{G}} = \{\hat{G}_1, \hat{G}_2, \dots, \hat{G}_L\}$ from sequence alignments of L loci ($\mathbf{D} = \{D_1, D_2, \dots, D_L\}$) using a likelihood-based approach like RAxML [59, 37] or MrBayes [28] that rely on different models of evolution, and 2) the estimation of a phylogenetic network $\hat{\mathcal{N}}$ with branch lengths ($\mathbf{t}$) and inheritance probabilities (γ) from the estimated gene trees using the pseudolikelihood model under the multispecies coalescent model on networks.

Instead of calculating the likelihood of an n-taxon network, the pseudolikelihood is based on the expected concordance factors (CFs) of every 4-taxon subnetwork (sometimes denoted *quarnet* [25]) and the observed CFs of every 4-taxon subtree (*quartet*) from the sample of gene trees [55]. The concordance factor (CF) of a given quartet (or split) is the proportion of genes whose true tree displays that quartet (or split) [7]. For example, for taxon set $s = \{a, b, c, d\}$, there are only three possible quartets, represented by the splits $q_1 = ab|cd$, $q_2 = ac|bd$ and $q_3 = ad|bc$. The *observed CFs* depend on the proportion of estimated gene trees that match each of the three quartets: $(\widehat{CF}_{q_1}, \widehat{CF}_{q_2}, \widehat{CF}_{q_3})$. The *expected CFs* $(CF_{q_1}, CF_{q_2}, CF_{q_3})$ under the multispecies coalescent network model [42, 68] depend on the probability of observing the quartet under a given 4-taxon network with branch lengths ($\mathbf{t}$) and inheritance probabilities (γ).

Assuming unlinked loci, the vector $L * (\widehat{CF}_{q_1}, \widehat{CF}_{q_2}, \widehat{CF}_{q_3})$ follows a multinomial distribution with probabilities given by the expected CFs $(CF_{q_1}, CF_{q_2}, CF_{q_3})$, where L is the total number of gene trees. For a network on $n \geq 4$ taxa, we consider all 4-taxon subsets s and combine the likelihood of each 4-taxon subnetworks to form the full network pseudolikelihood:

$$\tilde{L}(\mathcal{N}|\mathbf{G}) \propto \prod_{s \in \mathcal{S}} (CF_{q_1})^{L*\widehat{CF}_{q_1}} (CF_{q_2})^{L*\widehat{CF}_{q_2}} (CF_{q_3})^{L*\widehat{CF}_{q_3}} \tag{1}$$

where $\mathcal{S}$ is the collection of all 4-taxon sets and $q_i = q_i(s)$ $(i = 1, 2, 3)$ are the 3 possible quartets on s.

We use the term 'CF' as opposed to 'probability' to emphasize that CFs measure genomic support. Probabilities (such as posterior probabilities or bootstrap values) are most often thought to measure statistical uncertainty [2]. Intuitively, splits between natural evolutionary groups of organisms are recovered by most or all genes, and thus have high CFs. On the other hand, the presence of a hybrid would be captured by intermediate CFs. For example, if a is a hybrid intermediate between b and c, the CFs of $ab|cd$ and $ac|bd$ would be around 0.5 while the CF of $ad|bc$ would be near 0.

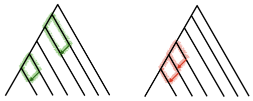

Figure 4: Left: Level-1 network in which the hybridization cycles do not intersect. Right: Level-2 network in which the hybridization cycles intersect.

Under the pseudolikelihood model, the PI pioneered the study of network identifiability by proving that certain hybridization events on level-1 networks are identifiable from observed CFs. The identifiability proofs involve the comparison between the theoretical formulas of the expected CFs of a species network with h hybridization events to the theoretical formulas of the expected CFs of a species network without the hybridization event of interest (that is, with $h - 1$ hybridization event). By proving that these equations do not share any feasible solutions, the same set of observed CFs cannot have been generated by both the species network with h hybridizations, and the species network with $h - 1$ hybridizations (see PI's work in [55, 57]). In addition, it was proven that the direction of the hybrid edges is detectable, but the root is not, which is why we estimate semi-directed networks. Given that we only have theoretical guarantees of identifiability for level-1 networks, the pseudolikelihood estimation in SNaQ is currently restricted to this class of networks [55]. While this type of networks can suit well to represent the evolutionary relationships of certain eukaryotic organisms, more complex evolutionary histories with widespread gene flow events cannot be properly captured.

In prior work, the PI has successfully provided a theoretical framework for the estimation of level-1 phylogenetic networks from multilocus sequence alignments [55], and a scalable software implementation that is widely used by evolutionary biologists worldwide [56]. Here, we will produce a framework to estimate the observed CFs from short reads instead of estimated gene trees (Objective 1) and we will extend the identifiability proofs beyond level-1 networks into level-2 and tree child networks (Objective 2). Lastly, we will develop a divide-and-conquer algorithm to merge smaller estimated networks into a large network

(Objective 3). All PI's work is always accompanied by open source, easy to use software to serve the broader evolutionary biology community (Objective 4).

Note that recently there have appeared proofs of identifiability of level-1 [19, 20] and level-2 phylogenetic networks [3]. This work, however, focuses on models of evolution with DNA sequences as input data (not gene trees), so they are not studying the identifiability of networks under the multispecies coalescent model which is the focus in this proposal.

3 Research plan

Objective 1: Extension of the pseudolikelihood network model to short unassembled reads as input. This objective includes several stages (Figure 5): 1) unsupervised clustering of reads to taxa, 2) reference-free assembly and alignment, and 3) estimation of CFs from assembled and aligned sequences.

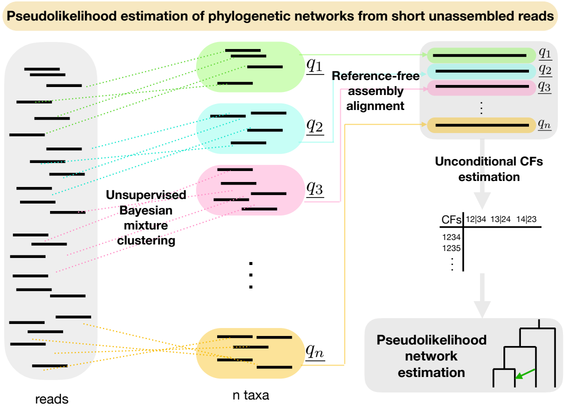
Pseudolikelihood estimation of phylogenetic networks from short unassembled reads
q_1
q_2
q_3
q_n
Reference-free assembly alignment
q_1
q_2
q_3
q_n
Unsupervised Bayesian mixture clustering
Unconditional CFs estimation
CFs
12|34
13|24
14|23
1234
1235
Pseudolikelihood network estimation
reads
n taxa

Figure 5: **Objective 1 pipeline:** 1) Unsupervised clustering of reads to taxa via Bayesian mixture, 2) reference-free assembly and alignment, 3) estimation of CFs from assembled and aligned sequences taking into account the uncertainty in the clustering stage via the posterior probabilities ($\underline{q_i}$ in the figure), and 4) estimation of phylogenetic network from CFs.

1. Unsupervised clustering of reads to taxa. Let $\mathbf{R} = \{R_1, \dots, R_s\}$ denote a collection of short unassembled reads from metagenomic data. Following [50], we can estimate the posterior probabilities of reads belonging to taxa through an unsupervised naive Bayesian mixture model. In this model, we need to know the number of taxa (n), but see Limitations for relaxations of this assumption. Let $\mathbf{Y} = \{Y_1, \dots, Y_s\}$ be the (unknown) cluster labels so that $Y_i = m$ implies that read i belongs to the m^{th} taxon for $m = 1, \dots, n$. Let a_m be the proportion of reads belonging to the m^{th} taxon so that $P(Y_i = m) = a_m$. Each read will be converted into a $p-$dimensional vector where each component corresponds to the count of a specific k-mer. That is, for a given k, there are $p = 4^k$ possible k-mers, denoted $\mathbf{W} = \{W_1, \dots, W_p\}$. The read R_i is then represented as p-dimensional vector $R_i = (X_{i1}, \dots, X_{ip})$ where X_{ij} represents the count of the j^{th} k-mer (W_j) in the i^{th} read. The distribution of k-mers within the m^{th} taxon is modeled by a Poisson distribution with parameter $\lambda_m = (\lambda_{m1}, \dots, \lambda_{mp})$ so that the probability of observing the k-mer counts in read R_i given that it belongs to the m^{th} taxon is

$$P(R_i = (x_{i1}, \dots, x_{ip}) | Y_i = m) = \prod_{j=1}^{p} \frac{e^{-\lambda_{mj}} \lambda_{mj}^{x_{ij}}}{x_{ij}!} \tag{2}$$

Then, we cluster the reads to taxa based on the the posterior probability (q_{im}) of read R_i belonging to the m^{th} taxon which is given by $q_{im} = P(Y_i = m | R_i) \propto a_m P(R_i = (x_{i1}, \dots, x_{ip}) | Y_i = m)$. All parameters are estimated by an EM algorithm [50]. After this step, we have a clustering of reads to taxa with associated posterior probabilities for each read ($\{q_{im}\}$ for $i = 1, \dots, s$ and $m = 1 \dots, n$).

2. Reference-free assembly and alignment. After the first step, we have the estimated cluster labels for each read: $\hat{\mathbf{Y}} = \{\hat{Y}_1, \dots, \hat{Y}_s\}$ based on the posterior probabilities of belonging to a given taxon: $\mathbf{Q} = \{q_{ij}\}$ for reads $i = 1 \dots, s$ and taxa $j = 1, \dots, n$. That is, $\hat{Y}_i = m$ if q_{im} is the maximum probability among $\{q_{i1}, \dots, q_{in}\}$. We keep the posterior probabilities to measure uncertainty in the clustering stage, not just the cluster labels. Following standard *de novo* sequence assemblers [39], we utilize de Brujin graphs to assemble the single genome reads, and heuristic algorithms for multiple sequence alignment. After this step, we will have aligned sequences per taxa. The limitation of this stage is that propagation of statistical uncertainty on the assembly and alignment remains unknown to this date (at least for large datasets) and beyond the scope of the present proposal (but see Limitations).

3. Estimation of concordance factors from newly clustered, assembled and aligned sequences. We extend the work in [47] which estimates concordance factors from SNPs by computing the proportion of sites supporting each of the three possible quartet resolutions from a sample of biallelic SNPs. For example, in Table 1, we

observe that 2 out of 5 loci (locus 1 and 2) agree with the split 12|34, while locus 3, 4 and 5 agree with the split 13|24. Thus, the estimated CFs from this toy dataset would be ($\widehat{CF}_{12|34} = 2/5, \widehat{CF}_{13|24} = 3/5, \widehat{CF}_{14|23} = 0$).

The difficulty from metagenomes is that there is some uncertainty in the cluster of reads to taxa, which is represented by the posterior probabilities described before (q_{im}). We need to account for this uncertainty in the estimation of the CFs. We define the *conditional* CFs as the concordance factors conditioned on the first sequence labeled "taxon 1", the second sequence labeled "taxon 2", the third sequence labeled "taxon 3" and the fourth sequence labeled "taxon 4". In this manner, we can describe the unconditional concordance factors by the law of total probability:

$$CF(12;34) = \sum_{1 \leq i,j,k,l \leq n} CF(12;34|Z_i = 1, Z_j = 2, Z_k = 3, Z_l = 4) P(Z_i = 1, Z_j = 2, Z_k = 3, Z_l = 4) \tag{3}$$

where Z_i represents the taxon label for the i^{th} assembled and aligned sequence and let $b_{im} = P(Z_i = m)$ for $i = 1, \dots, n$ and taxon m. Note that we change the notation slightly to make the distinction between conditional and unconditional CFs clearer (we now write $CF_{12|34} = CF(12;34)$).

Ideally, we want to propagate the uncertainty at the read-level through $\mathbf{Y}$ (cluster labels for reads) and $\mathbf{Q}$ (posterior probabilities of reads belonging to taxa) into the probabilities $\{b_{im}\}$. Through assembly and alignment, the i^{th} aligned sequence is a combination of multiple reads assigned to the i^{th} cluster. That is, let $\{R_j | Y_j = i\}$ represent the set of reads assigned to the i^{th} cluster. For simplicity, we will define the probability of the i^{th} sequence belonging to taxon m ($b_{im} = P(Z_i = m)$) as the average of the posterior probabilities of the reads belonging to taxon m: $P(Z_i = m) = \frac{1}{s_m} \sum_{\{j : \hat{Y}_j = m\}} q_{jm}$, where s_m represents the number of reads assigned to taxon m. We also assume that these probabilities are independent: $P(Z_i = 1, Z_j = 2, Z_k = 3, Z_l = 4) = P(Z_i = 1) P(Z_j = 2) P(Z_k = 3) P(Z_l = 4)$ in Equation 3.

Locus	1	2	3	4	5
1	1	0	1	1	0
2	1	0	0	0	1
3	0	1	1	1	0
4	0	1	0	0	1

Table 1: Example of computation of concordance factors from biallelic SNPs from [47]. Rows are taxa and columns are loci.

Similar to Equation 3, we can compute $CF(13;24)$ and $CF(14;23)$ for every subset of 4 taxa. The formulas for the unconditional CFs naturally account for the uncertainty in the assignment of reads to taxa. However, the formula can become intractable for a big n given that its complexity is $O(n^4)$ (see Limitations for scalable alternatives). Note that an alternative approach to the direct estimation of the CFs from the aligned assembled sequences is to use these sequences to estimate gene trees, and then estimate CFs from gene trees as the standard SNaQ approach. The downside of the estimation of gene trees is the lack of propagation of the statistical error in the clustering of reads to taxa which is accounted for with the unconditional CFs.

4. Estimation of phylogenetic network from unconditional CFs. The estimated unconditional CFs will be then used as input in the pseudolikelihood inference pipeline in SNaQ [55, 56]. We will test our methodology with extensive simulations varying branch lengths to control the amount of ILS, the length and noise in the input reads to control the estimation error in the CFs, the number of taxa, of reads and the complexity of the network to estimate. We will begin with level-1 networks which have been widely studied first, and then, we will estimate level-2 and tree child networks with the novel theory developed in Objective 2.

Implementation: We will implement this pipeline as a novel Julia package (see Objective 4) and we will assess the accuracy in the estimation of CFs and phylogenetic networks via extensive *in silico* experiments under a variety of biological scenarios on the simulation of reads (amount of missingness, sampling bias, read length, number of taxa, species heterogeneity and amount of gene flow among species). We will further test our methods on public and collaborations data (Lankau, Koch, Rioux; see LoC and Facilities).

Limitations: The proposed work in Objective 1 has three main limitations: 1) it assumes a known number of taxa (n), which is not true in practice. We will explore the scalability of cross-validation or other model selection strategies to select the best n, as well as alternative clustering strategies that do not require a pre-specified number of clusters. We will utilize these alternative clustering strategies simply to estimate n, not to replace the Bayesian mixture model. 2) Our method does not allow for propagation of uncertainty in the assembly and alignment stage. This is beyond the scope of the current proposal, but this limitation could serve as motivation of a future proposal. The future work will need to draw inspiration from simultaneous estimation of alignment and phylogeny in a joint framework (see [23]). 3) The formula for the unconditional CFs is intractable for large number of taxa as it grows at rate $O(n^4)$. We will explore the ranking of the subsets $\{1 \leq i, j, k, l \leq n\}$ to only consider those such that the probability of that subset is a above a certain threshold (τ): $P(Z_i = 1)P(Z_j = 2)P(Z_k = 3)P(Z_l = 4) > \tau$.

Objective 2: Extension of pseudolikelihood network model to broader classes of networks. We will study the identifiability of level-2 and tree child networks [29] from CFs under the pseudolikelihood in SNaQ [55] which is built on the multispecies network coalescent model [68]. A level-2 network is a network whose biconnected components (intersections of cycles) have at most 2 hybridization nodes and tree child networks are networks in which every internal node has at least one child node that is a tree node.

In this research plan, we describe the process to study the identifiability of level-2 networks, but the steps are the same for tree child networks. The study of identifiability of level-2 networks follows the same steps as the original theoretical proofs of identifiability of level-1 networks in [55, 57]. We first investigate the identifiability of what we denote the *base case* (5-taxon networks with 2 hybridizations), and then, we extend our findings (*extended case*) to the case of n taxa and h hybridization events.

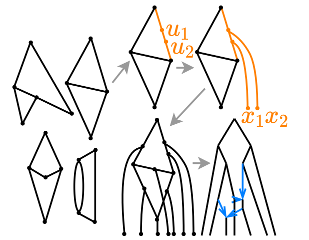

Figure 6: Left: Four generators for level-2 networks [63]. Right: Take one generator to produce a level-2 network. Each edge in the generator is replaced by a path graph (in orange with two internal nodes u_1, u_2). A leaf is connected to each of the new internal nodes (x_1 and x_2 respectively). The process is repeated until all seven leaves are added. Finally, we show a more standard visual representation of the 7-taxon level-2 network with hybrid edges in blue.

Base case: 5-taxon level-2 network with 2 hybridization events. We begin by listing all 5-taxon level-2 networks with 2 hybridization events via an enumeration algorithm based on the level-2 generators [63] (Figure 6). A level-k generator is a biconnected graph with parallel edges that has exactly k reticulation vertices. Any level-2 network can be obtained from the level-2 generator by replacing an edge in the generator with a path graph (linear chain of edges and nodes), and then a new leaf x and a leaf edge (u, x) is added for every internal node u in the path. After all the leaves connected to internal nodes in path graphs are added, new leaves need to be added connected to any internal node in the generator with indegree 2 and outdegree 0. We illustrate this process graphically in Figure 6.

Let $\mathbf{N}_{5,2}$ be the list of all level-2 networks with 5 taxa and 2 hybridization events. Let $\mathcal{N} \in \mathbf{N}_{5,2}$ with branch lengths t and inheritance probabilities γ.

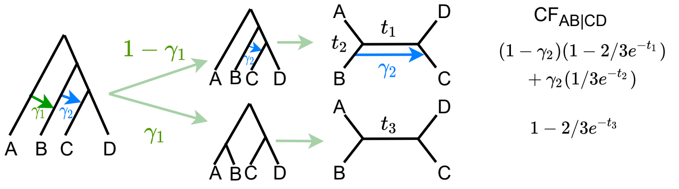

Figure 7: Computation of expected $CF_{AB|CD}$ for a 4-taxon network as weighted average of expected CFs from tree quartets as in [55].

To investigate if $\mathcal{N}$ is identifiable: 1) we calculate the expected CFs for all possible 4-taxon subnetworks of $\mathcal{N}$ and we denote these equations by $CF(\mathcal{N}, t, \gamma)$ which represent a set of polynomial equations with variables t and γ. Figure 7 illustrates the computation of the expected CFs of a given 4-taxon subnetwork. 2) We compute the corresponding polynomial equations of the expected CFs from the networks $\mathcal{N}_1$ and $\mathcal{N}_2$ which represent the subnetworks obtained from $\mathcal{N}$ by removing one hybridization event at a time. Note that both $\mathcal{N}_1$ and $\mathcal{N}_2$ only have one hybridization event. 3) We search for shared solutions to the polynomial equations: $CF(\mathcal{N}, t, \gamma) = CF(\mathcal{N}_1, t_1, \gamma_1)$ and $CF(\mathcal{N}, t, \gamma) = CF(\mathcal{N}_2, t_2, \gamma_2)$ using Mathematica [31] or Macaulay2 [18]. For example, if we can find branch lengths and inheritance probabilities (t_1, γ_1) on $\mathcal{N}_1$ such that the set of CFs are the same $CF(\mathcal{N}, t, \gamma) = CF(\mathcal{N}_1, t_1, \gamma_1)$ for any (t, γ), then $\mathcal{N}$ and $\mathcal{N}_1$ are not identifiable from the set of CFs. The process is illustrated in Figure 8. Note that we do not need to compare the CFs of the level-2 network to those of its major tree (tree obtained by removing all minor hybrid edges in the network) because this work has been done in [55, 57], so we only focus on the comparison between 2-hybridization networks to 1-hybridization networks. From this work, we conclude which is the set of identifiable 5-taxon level-2 networks with 2 hybridization events under the pseudolikelihood model.

Extended case: n-taxon level-2 network with h hybridization events. We generalize our findings on the 5-taxon 2-hybridization level-2 networks (base case) to detect which n-taxon level-2 networks with h hybridization events are identifiable. The key insight is that level-2 networks can be viewed as level-1 networks if we think of a biconnected component as one hybridization cycle. That is, in level-1 networks, the hybridization cycles do not intersect and thus, we are able to study the detectability of a given cycle ignoring all other cycles in the network. Similarly, in a level-2 network, each biconnected component (with exactly two reticulation nodes) does not intersect with other biconnected components because if they did, the network will no longer be of level-2. Therefore, we could focus on each biconnected component at a time regardless of other biconnected components in the network, and thus, we can write the expected CF polynomial equations for any level-2 network with n taxa and h hybridizations as a level-2 network with n taxa and 2 hybridizations (the ones on the biconnected component of interest). The CF polynomial equations of the network $\mathcal{N}$ ($CF(\mathcal{N}, t, \gamma)$) will similarly be contrasted to those coming from the two subnetworks $\mathcal{N}_1$ and $\mathcal{N}_2$ obtained from $\mathcal{N}$ by removing one minor hybrid edge from the biconnected component of interest at a time. By focusing on a given biconnected component at a time, this extension to n taxa follows the same structure as the identifiability proofs of level-1 networks with n taxa already studied in [55, 57].

Implementation: After the level-2 (and tree child) networks of n taxa and h hybridizations are proven to be identifiable, we extend the search in SNaQ to include the region in network space that includes these classes of networks by incorporating novel network moves [26, 24, 17, 10, 32, 33]. Most of the existing network moves are for rooted network or unrooted network, so we plan to combine ideas in order to traverse the space of semi-directed networks. In addition to the theoretical proofs and software implementation, we will perform *in silico* experiments under a variety of different scenarios: short vs long branch lengths to control the amount of ILS, short vs long sequences to control the amount of estimation error in the gene trees, number of taxa, of loci and of hybrids to assess the accuracy of the pseudolikelihood estimation of level-2 and tree child networks from multilocus sequence data. We will test our methods on public and collaborations data (Baum, Rakotondrafara; LoC and Facilities).

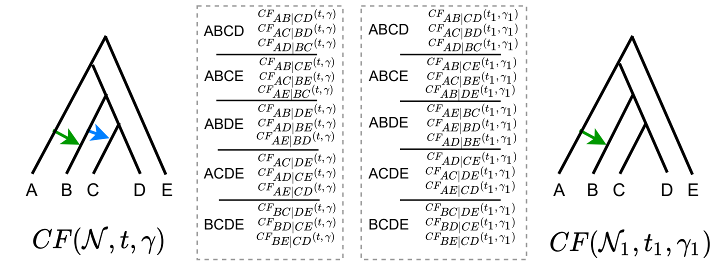
A B C D E
$CF(\mathcal{N}, t, \gamma)$
ABCD
$CF_{AB|CD}(t,\gamma)$
$CF_{AC|BD}(t,\gamma)$
$CF_{AD|BC}(t,\gamma)$
ABCE
$CF_{AB|CE}(t,\gamma)$
$CF_{AC|BE}(t,\gamma)$
$CF_{AE|BC}(t,\gamma)$
ABDE
$CF_{AB|DE}(t,\gamma)$
$CF_{AD|BE}(t,\gamma)$
$CF_{AE|BD}(t,\gamma)$
ACDE
$CF_{AC|DE}(t,\gamma)$
$CF_{AD|CE}(t,\gamma)$
$CF_{AE|CD}(t,\gamma)$
BCDE
$CF_{BC|DE}(t,\gamma)$
$CF_{BD|CE}(t,\gamma)$
$CF_{BE|CD}(t,\gamma)$
ABCD
$CF_{AB|CD}(t_1,\gamma_1)$
$CF_{AC|BD}(t_1,\gamma_1)$
$CF_{AD|BC}(t_1,\gamma_1)$
ABCE
$CF_{AB|CE}(t_1,\gamma_1)$
$CF_{AC|BE}(t_1,\gamma_1)$
$CF_{AB|DE}(t_1,\gamma_1)$
ABDE
$CF_{AE|BC}(t_1,\gamma_1)$
$CF_{AE|BD}(t_1,\gamma_1)$
$CF_{AD|BE}(t_1,\gamma_1)$
ACDE
$CF_{AD|CE}(t_1,\gamma_1)$
$CF_{AC|DE}(t_1,\gamma_1)$
$CF_{AE|CD}(t_1,\gamma_1)$
BCDE
$CF_{BC|DE}(t_1,\gamma_1)$
$CF_{BD|CE}(t_1,\gamma_1)$
$CF_{BE|CD}(t_1,\gamma_1)$
A B C D E
$CF(\mathcal{N}_1, t_1, \gamma_1)$

Figure 8: **Objective 2:** To prove identifiability of level-2 networks with 5 taxa and 2 hybridizations, we extract the expected CF polynomial equations from $\mathcal{N}$ and from $\mathcal{N}_1$ and search if there exist solutions in terms of branch lengths and inheritance probabilities that would simultaneously solve both systems of equations: $CF(\mathcal{N}, t, \gamma) = CF(\mathcal{N}_1, t_1, \gamma_1)$. If those solutions exist, then the networks $\mathcal{N}$ and $\mathcal{N}_1$ are not identifiable from the set of CFs.

Limitations: The identifiability proofs for level-2 networks with n taxa and h hybridization is straight-forward as it builts on the level-1 identifiability proofs in [55]. The identifiability proofs of tree child networks will be more challenging. Mainly, there is not a definition for semi-directed tree child networks as they are currently defined only for rooted networks. We will explore the use of artificial roots in the definition of tree child networks to assess whether the root is identifiable under the pseudolikelihood model. If the root is not identifiable, then the estimated networks will be indeed semi-directed. If the root is identifiable, however, the estimation of tree child networks will differ that of level-1 and level-2 by producing rooted (instead of semi-directed) networks. Another complexity is the computational expense of solving systems of polynomial equations. Even in the case of level-1, some scenarios were prohibited to run on standard computers. We will rely on the computational power of the Center for High Throughput Computing (CHTC, see LoC).

Objective 3: Network-merging algorithm to reconstruct large networks. In this proposal, we explore the potential of merging techniques to reconstruct large networks of hundreds or thousands of taxa. In particular, we build on existing merging algorithms for trees [52, 8, 44, 67]. Existing merging methods take a collection of trees as input $\{T_1, \dots, T_N\}$ and returns a compatibility supertree $\mathcal{T}$. We will merge collections of parental trees obtained from the input networks, and then use the merged parental trees as building blocks for the merged network. We note that existing divide-and-conquer network algorithms require a guide phylogeny [22], are not scalable for more than dozens of taxa [22] or can only build networks from 4-taxon subnetworks [1] so they are not suitable to merge subnetworks of any size into a larger network as we propose here. We propose the following pipeline (Figure 9):

1. Estimate a collection of networks using the methodology in Objectives 1 and 2 (or any network method-

ology) for K subsets of taxa: $\mathbf{N} = \{N_1, \dots, N_K\}$ so that each network N_i has fewer taxa (n_i) that the total number of taxa: $|n_i| < n$.

2. Decompose every network into its set of parental trees as described in [74]: $N_i \to \{T_1^{(N_i)}, \dots, T_{k_i}^{(N_i)}\}$ where k_i represents the number of parental trees from network N_i. Note that parental trees take the coalescent model into account and are a broader set than the displayed trees of a network (which are obtained by simply turning off certain minor hybrid edges). When there is only one allele sampled below a hybrid node, then the parental tree agrees with the displayed tree. If there are multiple alleles sampled below a hybrid node, however, then different alleles can follow different paths towards the root producing more trees than those displayed by the network (see [41, 74]).
3. Create a superset of parental trees: $\mathbf{T}(\mathbf{N}) = \bigcup_i \{T_1^{(N_i)}, \dots, T_{k_i}^{(N_i)}\}$.
4. Cluster the parental trees in $\mathbf{T}(\mathbf{N})$ into C groups: $\{\mathbf{T}(\mathbf{N})_1, \mathbf{T}(\mathbf{N})_2, \dots, \mathbf{T}(\mathbf{N})_C\}$ using a Dirichlet process as in BUCKy [2]. The rationale is that each cluster of parental trees will represent a single parental tree in the merged network.
5. Merge all the parental trees within a given cluster a supertree algorithm: $\mathbf{T}(\mathbf{N})_i \to \bar{T}_i$.
6. Combine the set of merged parental trees $\{\bar{T}_1, \dots, \bar{T}_C\}$ into a merged network: $\bar{N}$ with all n taxa.

While most of the steps in the pipeline are building on existing work – yet, these methods have never been used together to merge phylogenetic networks –, the last step of combining parental trees into a phylogenetic network has never been done. Existing work provides merging algorithms of *displayed trees* into split networks [29]. However, these methods are not suited for phylogenetic networks which are explicit networks (internal nodes represent actual biological process like speciation or hybridization), and they tend to overestimate the number of hybridization events by not accounting for other sources of tree discordance like ILS.

Aside for the consolidation of the full pipeline, the innovation of Objective 3 will be the development of an algorithm to combine a set of parental trees into a phylogenetic (explicit) network. This work will require the mathematical proof that networks are *identifiable* from the set of parental trees, which is currently unknown (see Limitations for alternative approaches if we prove that some networks are *not* identifiable from the set of parental trees).

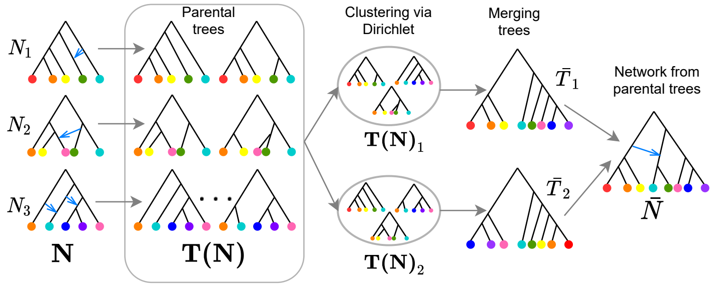

Figure 9: **Objective 3:** We merge three 5-taxon networks ($\mathbf{N} = \{N_1, N_2, N_3\}$) into a larger 8-taxon network ($\bar{N}$) via the decomposition of the small networks into parental trees, clustering and merging of those trees into the parental trees ($\bar{T}_1, \bar{T}_2$) of the large network which is then obtained from this set of parental trees. The procedure is illustrated for small networks, but it will be scalable to produce networks with hundreds or thousands of taxa.

Implementation: We will implement the network-merging algorithm as a novel Julia package (see Objective 4) and we will test our method extensively with *in silico* experiments under different biological conditions in terms of number of taxa and complexity of the networks (number of hybridization events). We will further test our methods on public and collaborations data (Baum, Rakotondrafara; see LoC and Facilities).

Limitations: The work proposed in Objective 3 has the limitation that a phylogenetic (explicit) network might not be identifiable from the set of parental trees. Under this situation, our algorithm will return a set of (unidentifiable) networks that can later be compared via likelihood or model selection tools like [11].

Objective 4: Implement the new methods in an easy-to-use, open-source software. This proposal will produce two novel Julia packages: 1) to estimate CFs from metagenomes (Objective 1) and 2) to merge small networks into a large network (Objective 3) as well as an extension to the widely used SNaQ software [55] in the PhyloNetworks Julia package [56] to level-2 and tree child networks (Objective 2). The PI has vast experience in creating open source publicly available software that is easy to use by the evolutionary biology community. New software will be accompanied by documentation and tutorials guiding every step from raw input files into desired output as well as a google user group to provide end-user support. Furthermore, the novel tools will be broadly advertised in conferences like Evolution and organismal conferences for maximum exposure and outreach via hands-on tutorials and in standard molecular evolution workshops like MBL