Introduction & Background Phylogenomic estimation, in which genomes of different organisms are compared to *understand evolutionary relationships*, provides insight into functional genomics, underlines priorities for conservation efforts, and illuminates the history of life on Earth [1]. However, accurate methods are not yet able to scale to *large datasets* with many species. Biological processes including incomplete lineage sorting (ILS) [2], horizontal gene transfer, hybridization, introgression, and recombination [3], along with limitations in sequencing accuracy and statistical errors due to limited sequence length, introduce *noise* into the *phylogenetic signal* and add to the challenge of species tree estimation. Highly accurate Bayesian estimators [4] can generate a probability distribution over the output tree space with several parameters, including population size, branch length, and evolutionary histories of individual genes, but cannot scale beyond small datasets. On the other hand, *fast estimators* can scale to medium or large datasets while providing good accuracy and theoretical guarantees of correctness [2], but may not scale to extremely large datasets and do not provide as much information about the output as Bayesian methods do.

ILS will be the focus of my work because, unlike other sources of gene tree discord, it appears in almost all genome-scale datasets. Therefore, my main research goals involve the development of scalable methods that function in the presence of high levels of ILS, as well as other sources of gene tree discord. In particular, I will develop

1. *Fast and accurate methods* that provide greater accuracy than existing methods on datasets with hundreds of species and can scale to genome-wide datasets with tens of thousands of species
2. *Meta-methods* that can increase the scalability and accuracy of existing methods using *divide-and-conquer techniques* on *high-performance computing* infrastructure. These will allow fast methods to scale to extremely large datasets, and accurate Bayesian methods to scale to moderately sized datasets.

To evaluate these methods I will:

1. Use *simulation protocols* from the phylogenetic method development literature [2] that generate datasets with realistic sources of noise to benchmark accuracy and running time of methods
2. Work with domain experts to analyze *biological data* and provide insight into important scientific questions. Working with domain experts will also improve the *usability* of my methods, as they will be able to provide feedback on difficulties using the software.

Development of new methods Since most phylogenetic estimation problems are NP-hard, it is impractical to solve them exactly. Therefore, scalable methods typically either use *heuristic search* techniques like hill-climbing or find exact solutions over a *constrained solution space*. The two existing most accurate scalable algorithms are ASTRID [5], which uses a heuristic search, and ASTRAL [2], which uses an exact algorithm in a constrained search space. I am the chief developer for ASTRID, and ASTRAL is developed by my advisor's group. I am conducting preliminary research into using heuristic search and exact constrained algorithms in tandem to improve the performance of both types of methods. That is, better heuristic search algorithms can be used to choose constrained search spaces, and subroutines of heuristic search algorithms can be rewritten as exact constrained searches. Furthermore, ASTRID and ASTRAL are both easy to use and available on GitHub under free, open source

licenses, and I will distribute new methods this way as well.

Meta-methods to boost scalability Methods that increase the scalability and accuracy of existing tree estimation algorithms can enable the analysis of extremely large datasets with fast methods as well as moderately sized datasets with accurate but slow Bayesian estimators.

Iterative meta-methods like DACTAL [6] have been developed that *divide a large dataset into overlapping subsets* based on an initial estimate of the phylogeny, estimate trees on each subset, then combine them with a *supertree method*. This supertree is then used to find better overlapping subsets to form the basis of the next iteration. This approach has the advantage of being *highly parallelizable*, making it particularly amenable to running on high performance NSF supercomputers, notably UIUC's Blue Waters.

There is a lot of room for improvement for both supertree methods and iterative meta-methods. My initial focus on this project will be developing improved supertree methods, as these are useful for more than just iterative meta-methods. Existing supertree methods use heuristics, but we are developing polynomial-time exact constrained algorithms that can solve a variety of optimization criteria, including the Robinson-Foulds supertree problem [7].

Biological analyses I will participate in several ongoing collaborations aimed at estimating species trees for specific taxonomic groups. My advisor is a member of the Avian Phylogenomics Project (http://avian.genomics.cn) and the 1000 Plants Initiative (http://onekp.com), which have produced results and are expanding to larger analyses. She is also on the executive committee for the Genome 10K Project (http://genome10k.soe.ucsc.edu), which will produce datasets with whole genomes for over 10,000 vertebrate species. I will help develop phylogenetic methods suitable for these projects and work with domain experts to analyze their data.

Broader Impact The true test of a method is if it is used for biological analyses. I will run tutorials and workshops at major meetings of systematic biologists, including the annual Evolution and SMBE conferences, each of which has over a thousand attendees. I led one such session this June at the Hemipteroid Insect Phylogeny workshop at UIUC to train 30 researchers in the use of ASTRAL.

With the support of the NSF, this work will enable a wide range of biological analyses and support domain experts in research critical to our understanding of genomics, biodiversity, and natural history.

[1] N. J. Wickett, *et al.*, *Proceedings of the National Academy of Sciences* **111**, E4859 (2014).
[2] S. Mirarab, T. Warnow, *Bioinformatics* **31**, i44 (2015).
[3] W. P. Maddison, *Systematic Biology* **46**, 523 (1997).
[4] J. Heled, A. J. Drummond, *Molecular Biology and Evolution* **27**, 570 (2010).
[5] P. Vachaspati, T. Warnow, *BMC Genomics* **16**, S3 (2015).
[6] S. Nelesen, K. Liu, L.-S. Wang, C. R. Linder, T. Warnow, *Bioinformatics* **28**, i274 (2012).
[7] M. S. Bansal, J. G. Burleigh, O. Eulenstein, D. Fernández-Baca, *Algorithms for Molecular Biology* **5**, 18 (2010).