Dyr, a Program for Bayesian Inference of Language Phylogenies

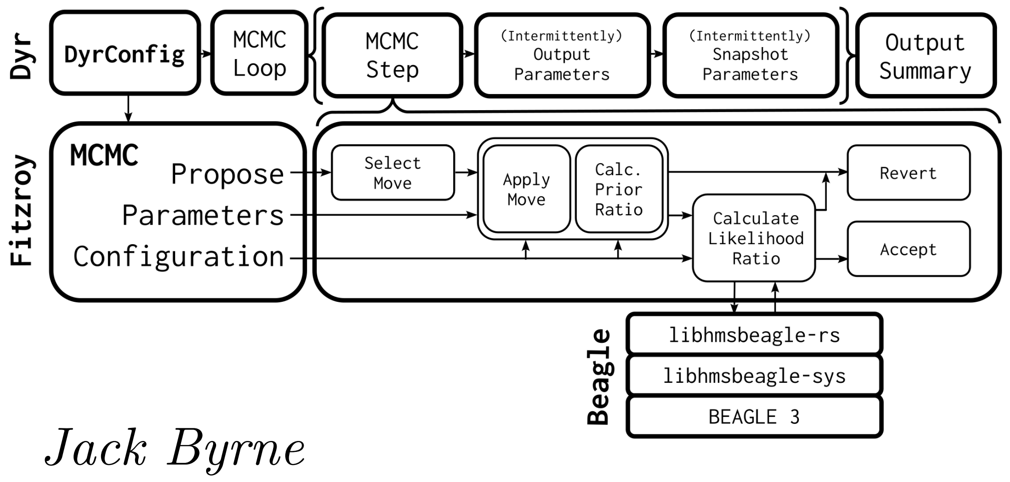
Languages evolve over time. It is common to model language evolution with a descent tree called a **phylogeny**. When two related languages both refer to a particular concept with words from the same root, we say they share a **lexical trait**. From a database of these shared traits, we seek to quantitatively infer the most plausible dated phylogeny of a language family.

Our approach is a technique called Markov Chain Monte Carlo (**MCMC**), a stochastic process that steps through the space of possible phylogenies and tends towards sampling from their probability distribution according to a Bayesian statistical model. This distribution is called the '**posterior**', and it is defined as the product of a specified prior distribution and the likelihood of the trait data given the phylogeny.

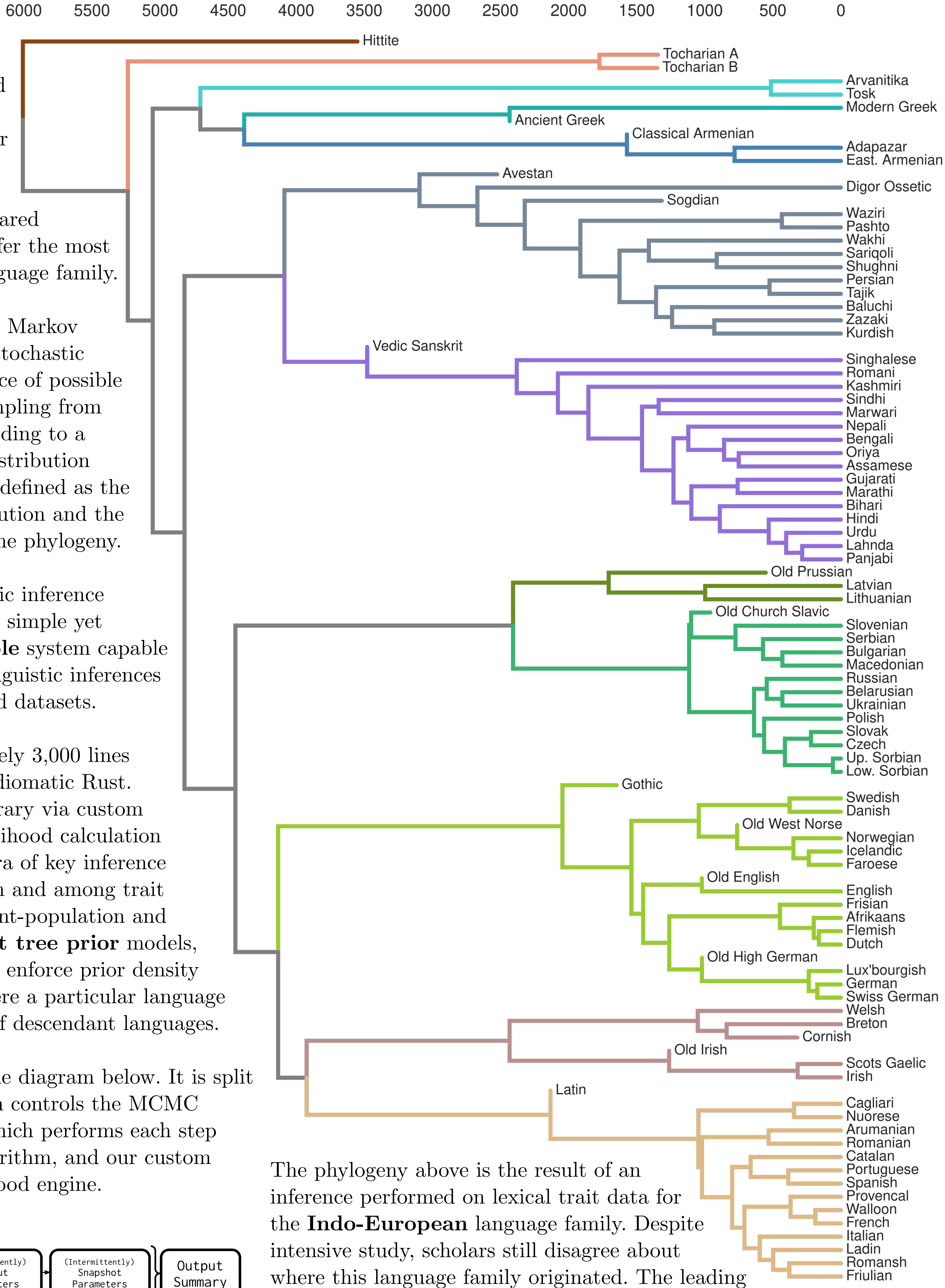
We have created a new phylogenetic inference program called **Dyr**, designed as a simple yet **high-performance** and **extensible** system capable of running state-of-the-art phylolinguistic inferences on real-world problem domains and datasets.

Dyr is implemented in approximately 3,000 lines of well-commented, modern, and idiomatic Rust. It incorporates the BEAGLE 3 library via custom bindings, enabling accelerated likelihood calculation using CUDA. It supports a plethora of key inference techniques including among branch and among trait evolutionary rate variation, constant-population and generalised skyline plot **coalescent tree prior** models, and **ancestry constraints**, which enforce prior density penalties on those phylogenies where a particular language is not ancestral to a specified set of descendant languages.

The design of Dyr is outlined in the diagram below. It is split into three layers, a front-end which controls the MCMC mainloop, a core library *Fitzroy* which performs each step of the **Metropolis-Hastings** algorithm, and our custom interface to the BEAGLE 3 likelihood engine.



Jack Byrne
Durham University
Department of Computer Science



The phylogeny above is the result of an inference performed on lexical trait data for the **Indo-European** language family. Despite intensive study, scholars still disagree about where this language family originated. The leading hypothesis is that the speakers of Proto-Indo-European were inhabitants of the Pontic-Caspian **Steppe** around 6000 years before the present. However, a competing theory suggests an earlier origin in **Anatolia** around 8500 years before the present.

Using a constant-time coalescent prior model for tree evolution, we performed Bayesian inferences on three datasets, previously used by Chang et al. in a similar study. Our analyses yielded a median root age of 6008 years for the Broad dataset (see above), 7692 years for the Medium dataset, and 6307 years for the Narrow dataset. **We therefore re-affirm the conclusion of Chang et al. that the Steppe hypothesis is the most plausible.**