Postdoc position available: Bayesian phylogenetics in the densely sampled regime in the Matsen group, Fred Hutchinson Cancer Center

The project  Statistical phylogenetic (evolutionary tree) methods have been essential for understanding the SARS-CoV-2 epidemic, whether for understanding origins, global spread, or lineage dynamics of the virus. These methods are extremely mature, with optimized code and software packages implementing complex models. However, these methods were developed with the “classical” sampling regime in mind: a relatively small number of sequences with relatively large divergences between them.

Methods for the classical sampling regime work to integrate out the uncertainty we have in ancestral sequences. Although the Felsenstein algorithm does allow for efficient calculation and updating of phylogenetic likelihoods, even this is not enough to handle the massive trees we would like to use for SARS-CoV-2. Furthermore, the Felsenstein algorithm only works for IID models between sites.

With SARS-CoV-2 we are in a completely different sampling regime, with over 2 million genomes for a virus without very much evolutionary divergence. That means that we frequently sample identical viruses, and we often sequence the direct ancestor of a given virus. This greatly limits the uncertainty that we have in the ancestral states of the genome. However, the transmission history is quite uncertain, motivating a Bayesian approach.

There are some interesting opportunities in this new regime. For example, du Plessis, McCrone, Zarebski, Hill, Ruis, et al, (2021) replace the classical phylogenetic likelihood with a proxy that counts the number of substitutions that could have happened along a branch. This reduces computation time by orders of magnitude, and allows the model to focus on the important aspects of uncertainty: how the virus spread between individuals.

I think that there are many more opportunities in this new regime, including for substitution model complexity (think whole-sequence modeling), online (i.e. incremental updating) inference, integration with epidemiological models, and for inference (it’s not going to be MCMC).
There are other settings that we care about for which we have dense sampling, and for which complex sequence substitution models are quite important. Specifically, I’m thinking about the evolution of antibodies that happens inside of our lymph nodes when we are vaccinated or infected. Our collaborator Gabriel Victora and his lab sample these evolutionary histories in great depth. We are also very interested in the interplay between sequence and fitness.

It’s still early stages, but thus far it looks like this will become a collaborative project with:

- **Trevor Bedford**, an evolutionary biologist and genomic epidemiologist known for his co-development of the [nextstrain](https://nextstrain.org) platform
- **JT McCrone**, a postdoc in the Rambaut group working on scaling Bayesian phylogenetics for SARS-CoV-2
- **Vladimir Minin**, a leading statistician especially known for his work in “phyldynamics:” the intersection of phylogenetics, immunology, and epidemiology
- **Marc Suchard**, another leading statistician working on phylodynamics, who has developed much of the statistical framework for complex data integration in BEAST, as well as high performance algorithms

and hopefully many others in the phylogenetics community.

**Environment**   The position will come with a competitive postdoc-level salary with great benefits for two years, with the ability to extend if things are going well. The environment is lively yet casual, with a strong emphasis on collaborative work. The collaborative aspect should make this position especially interesting for cross-disciplinary work: postdocs can work with programmers in the group to implement inferential models, and with statisticians in the group to formulate them as needed.

The Center is housed in a lovely campus on Lake Union a short walk from downtown, and a slightly longer walk from the University of Washington. The Matsen group is in the newly-remodeled Steam Plant building overlooking the lake. Powerful computing resources and helpful IT staff await. Ideally you’d want to be on campus but long-term remote work is possible from these states: Alabama, Alaska, Arizona, California, Colorado, Hawaii, Idaho, Maryland, Minnesota, Montana, New York, Ohio, Oregon, South Carolina, and Texas.

We believe that science is for everyone. We have had researchers with a variety of backgrounds, including Latinx, Black, Asian, and Middle Eastern. We have had women, men, gay, and straight, and we welcome people of all sexual orientations and gender identities. We have had successful high schoolers, postdocs, people who were the first in their family to attend college, and one who had decided that college wasn’t for them. We have had researchers with backgrounds in biology, physics, statistics, math, and computer science.

We acknowledge the historical and present barriers for underrepresented groups,
and work to increase diversity, equity and inclusion in computational biology. Members of underrepresented groups are especially encouraged to apply.

Please read our expectations of group members. By applying for this position, I expect that you will fulfill these expectations. I enthusiastically solicit feedback on these expectations or requests for clarification.

You can find out more about our group by visiting our group website.

**Qualifications** This position is most suited to someone a PhD in statistics, computer science, biology, or another relevant field. However, we will consider exceptionally skilled candidates with BS or MS degrees.

**Essential skills** We are looking for someone who has:

- experience doing methods development for Bayesian inference, or experience doing high-performance software development using graphs
- clear ability to perform independent research
- the ability to work and collaborate with a team.

**Additional helpful skills** Ideally the candidate would have:

- knowledge of Bayesian phylogenetics and genomic epidemiology
- experience with C++, and with modern C++ idioms
- desire to improve software development skills towards clean and tested code
- experience with a modern git-based workflow
- experience with Docker and continuous integration
- experience developing in a Linux environment

**Applying** If you are interested in this position, please submit the following materials:

- Two representative publications.
- A CV summarizing your education and work experience so far.
- The names and email addresses of three references.
- A code sample showing work that you are proud of. This has to be nontrivial, but doesn’t have to be long. Ideally it would be publicly accessible, e.g. on GitHub, but if that’s not possible an emailed attachment is fine too.

Please send these materials to ematsen@gmail.com.