Stephanie J. Spielman
Research Statement

Biology in the 21st century faces a unique, historically unprecedented challenge: We are generating a large amount at a rate that far outpaces our ability to analyze it in concert. We find ourselves at a critical junction in an ongoing scientific transition, requiring researchers who can extract meaningful information from massive amounts of biological data. It is therefore crucial to have robust and reliable analysis methods, and similarly to have a firm understanding how these analysis methods should (and should not) be used.

These fundamental goals motivate my research program. I primarily work in the fields of *evolutionary genomics and phylogenetics*. In my research, I develop, test, and apply computational techniques to reveal evolutionary processes from comparative sequence data. I address questions that are both empirical (analyzing the evolutionary constraints operating on clinically-relevant natural sequences) and theoretical (interrogating the statistical underpinnings of evolutionary models), and I additionally develop software platforms to enable and facilitate such analyses.

While the bulk of my research is in evolutionary biology, my expertise more broadly falls within the scope of "data science," i.e. computational analysis of large-scale datasets. I have a keen interest in merging complex data and inferences from interrelated biological systems, as well as identifying overarching biological trends. My skills translate into tremendous research flexibility, making me an asset to biology departments in need of collaborators who can process high volumes of biological data. For example, I previously collaborated with the Schmidt lab at The University of Texas-Southwestern to study the distribution and role of sequence motifs that induce the formation of clathrin-coated pits during cellular endocytosis (Kadlecova et al, J. Cell Biol., 2017). During my graduate work, I was involved in several collaborative initiatives in my graduate lab to study how protein structure influences a given protein's evolutionary trajectory (Shahmoradi et al, J. Mol. Evol. 2014; Jackson et al, Prot. Sci, 2016). I have two well-developed research programs, as described below.

Modeling sequence evolution
For my Ph.D. dissertation, I studied the properties and limitations of *models of coding-sequence evolution*, that is statistical models that reveal how natural selection shapes genome sequences and diversity. These models have a wide array of applications, such as i) identifying positively-selected (i.e. evolving faster than a neutral expectation) genetic regions in pathogens, such as Influenza, Ebola, and HIV to inform vaccine development, ii) investigating how genes evolve when organisms adapt to novel environments, iii) providing the methodological basis for modern-day molecular phylogenetics and systematics, and iv) discovering novel genes or gene-family members from large genomic datasets.

Past Research. My Ph.D. dissertation, which the UT Austin Office of Graduate Studies awarded the "Outstanding Dissertation Award" among all dissertations in math, science, and engineering, focused on "best practices" in the development and usage of models of coding-sequence evolution. I established a unifying mathematical framework to compare two distinct models (known respectively as *dN/dS* and *mutation–selection* models), and I used this framework to demonstrate how certain biological processes can induce systematic biases model inference (Spielman & Wilke, Mol. Biol. Evol. 2015; Spielman & Wilke, Mol. Biol. Evol. 2016; Spielman & Wilke, Genetics 2016). For example, I showed how the genome phenomenon of codon usage bias may cause false positive results (Spielman & Wilke, Mol. Biol. Evol. 2015). My work offered work offered specific recommendations for how users can avoid pitfalls and

reliably interpret inferences to make robust conclusions about the activity of natural selection over long evolutionary timescales. To complement this work, I developed a popular software platform, *pyvolve*, for simulating evolutionary genetic data (Spielman & Wilke, PLOS ONE 2015).

Current Research. I am now working to develop new approaches for modeling evolution at the *protein-sequence* level. Protein models of evolution have revolutionized the field of phylogenetics, for example by providing researchers with tools to study "deep evolutionary divergence." For example, protein models are used almost exclusively when studying early evolution of mammals and, even more ancient, the evolution and origins of all animals. In my current position at Temple University, I am developing a new algorithmic approach to build these protein models so that researchers can have modeling frameworks tailored to their data. For example, a researcher studying HIV would not want to apply a mammalian-specific model to their data. My framework will provide a customizable platform such that the evolution of specific model systems can be interrogated with precision and accuracy. I am applying this framework to study *transmembrane proteins,* as described below.

Evolution of transmembrane proteins

Transmembrane proteins comprise up to 35% of a given organism's proteome, but they remain comparatively understudied relative to cytosolic proteins. I merged techniques from structural biology and comparative sequence analysis to discern the evolutionary constraints operating on transmembrane protein domains, compared to extramembrane (intra- and extra-cellular) domains.

Past Research. I use G protein-coupled receptors (GPCRs) as a model system to understand how evolution operates across levels of cellular organization. GPCRs are cell-surface receptor proteins that represent the largest receptor family in *Metazoa* and are targeted by nearly 40% of pharmaceuticals. Through a large-scale survey of hundreds of mammalian GPCRs, I showed that transmembrane protein domains tend to evolve more slowly compared to extramembrane domains. This work revealed that the membrane environment uniquely influences protein evolution above and beyond cellular localization (Spielman & Wilke, J. Mol. Evol. 2013). After establishing these governing principles, I analyzed in-depth one of the largest and most pharmaceutically important vertebrate GPCR sub-families: the biogenic amine (e.g. serotonin, dopamine, histamine, etc.) receptors. I developed a novel algorithm for obtaining structurally-aware GPCR sequence alignments to construct a comprehensive phylogeny of biogenic amine receptors (Spielman, Kumar, & Wilke, PeerJ 2015). This phylogeny provided unique insights into the complicated evolutionary patterns in GPCRs. For example, I discovered novel receptor families, identified patterns in gene family contraction/expansion events across different species, and uncovered motif shifts accompanying gene duplication events.

Current Research. All available statistical models of protein-sequence evolution have been exclusively designed to study cytosolic (i.e. non-membrane) proteins. As such, current analyses of transmembrane proteins all suffer from poorly-specified and likely incorrect models. I am applying my algorithm for building protein evolutionary models to establish a robust model tailored to transmembrane sequence data, using a massive training dataset of sequences from online databases such as NCBI and Ensembl. The resulting model will capture the unique sequence properties inherent to transmembrane proteins and will have a wide

variety of applications, ranging from phylogenetic inference to the identification of novel transmembrane proteins from genomic datasets.

Undergraduate Research Involvement

During my graduate work, I directly mentored three undergraduate students in the lab and worked in tandem with two other undergraduate students on lab-wide projects, resulting in five publications with undergraduate co-authors. Three of these students have since begun Ph.D. programs in biochemistry and/or computational biology. Because my research expertise is highly interdisciplinary (featuring components from biology, computer science/software design, biostatistics, and biophysics), I am well-equipped to mentor students with a variety of academic backgrounds and interests. For example, a student interested in developing computational tools for analyzing biological data (broadly interpreted), a student interested in studying the evolutionary patterns of their favorite gene, and a student interested in next generation sequencing analysis would be equally suited to work in my lab group. This flexibility encourages scientific diversity by allowing students to explore and collaborate on range of research avenues, ultimately fostering their development into well-rounded scientists.

Commitment to Open Science

A key component of my research philosophy is to maintain scientific openness via data and code sharing, thereby encouraging scientific reproducibility. To this end, I make all of my computer code freely available on the Github (https://github.com/sjspielman/) repository, and similarly all data from my published research is freely available either on Github or a data repository such as Data Dryad (https://datadryad.org/). Before publication in a peer-reviewed journal, I deposit my manuscripts in an online pre-print server, such as *bioRxiv* or *PeerJ Preprints*. Students in my lab will be expected to maintain scientific openness through these practices. Publishing pre-prints will also serve as crucial training in scientific writing and manuscript preparation. Indeed, even if a given student's project is not intended for publication, the student will still be able to write a proper scientific manuscript and share it with the scientific community at large.