Our current sequencing capabilities outstrip our computational resources, a gap solved through algorithm improvement, increased computational capabilities, and multivariate models. An area this is particularly relevant to is infectious diseases. Statistical advancements are necessary as biotechnological advances generate infectious disease datasets that are massive in terms of sheer size and variables. The NIAID influenza sequencing database has over 6000 complete genomes available and the LANL HIV database over 300,000, far greater than the number of sequences used in most studies. My research interests focus on bridging this divide between the massive datasets available and what is currently analyzable. I explore Bayesian hierarchical modeling, constrained covariance matrices, nonparametric analyses, and importance sampling for this purpose.

Infectious diseases emerge and re-emerge in a population. In the case of influenza A, there are yearly epidemics and occasional pandemics despite a native human immune response and a yearly re-engineered vaccine. Evolutionary mechanisms RNA viruses use include a high mutation rate (due to lack of proofreading mechanism and exacerbated by a fast replication time), recombination, and reassortment. Each HIV virus contains two duplicate copies of RNA and recombination occurs when two different viruses infect a single host and produce a virus with one of each viral RNA strand. The RNA polymerase responsible for viral replication may switch between the two RNA templates. Switching in place is a homologous recombination and switching to a new location on the other RNA template is nonhomologous. While recombination in influenza A is rare, reassortment, in which two entire segments of RNA exchange, has been responsible for the Spanish, Asian, Hong Kong, and current swine flu pandemics. By couching these evolutionary mechanisms in the language of phylogenetics, we can detect mutation rates, reassortment, and recombination events.

Rich infectious disease sequencing databases include viruses sampled from different geographical locations, times, viral subtypes, and subjects inspiring novel biological hypotheses involving correlation, covariates, and their interactions through time. Formally testing these models, potentially based on thousands of taxa, demands statistical ingenuity due to the computational intractability of working with a massive modeling space. Ideally, a researcher interested in reconstructing the evolutionary history of a virus identifies the isolates of interest, sequences them, and uses readily available Bayesian software to infer a phylogenetic tree from these sequences. Unfortunately, due to the computational complexity of inferring phylogenies compounded by the
large number of sequences, the researcher is often forced to partition the
taxa by a covariate (e.g., sampling location) and run independent or strati-
fied analysis. This stratification, while facilitating fast estimation, results in
overparameterization and ignores the correlation between parameters across
strata. Additionally, stratification fails to profit from the massive amounts of
data available because parameters are estimated from siloed strata, removed
from the implicit context that motivated the initial data collection. Using
the intermediate realizations from these stratified analysis, I infer hierarchical
models based on importance sampling. This strategy yields improved esti-
mators due to shrinkage towards the mean and the ability to assign Bayes
factors. I have successfully applied this methodology to unresolved biologi-
cal hypothesis concerning influenza A using both a (1) mixture model with
patterned covariance matrices and (2) a nonparametric wavelet based model.

Future Directions: The fast evolutionary dynamics and extensive sam-
pling of the HIV virus enables a detailed reconstruction of its evolutionary
history. However, the considerable complexity makes understanding these
dynamics and the creation of a successful vaccine elusive. Similar to the
HA influenza glycoprotein, the surface glycoprotein env is often a target of
drug development. Amino acid substitutions within the env open reading
frame indicate chemokine co-receptor conversion from CCR5 to CXCR4, a
harbinger of a more virulent stage of the disease motivating the need to
identify covariates indicative of conversion.

The Bayesian software BAli-Phy jointly infers sequence alignment and
phylogeny and promises to yield new insight into evolutionary dynamics.
However, it is hindered by a tremendous modeling space that severely limits
the number of taxa that can be analyzed. I am particularly interested in
the covariates of alignment length and testing insertion and deletion (indel)
rates in particular regions of the HIV genome. High indel rates are indicative
of regions undergoing positive selection pressure and may pinpoint regions
important in enabling early immune escape and determining HIV subtype.

The computational issues described for infectious diseases are small when
compared against human sequences in the 1000 genomes and cancer genome
project. While current sequences maintain the imprint of past demographic
events, informing these histories with molecular epidemiology data is also
a key direction. The importance of developing bioinformatics tools for the
genomic community means addressing practical issues so it is a priority of
mine to implement these methods in freely available software.