Title: Quantifying convergent molecular evolution across mammalian genomes

The evolution of similar traits in distantly related species is one of nature's great surprises. Convergent evolution of traits has been widely observed throughout the animal kingdom; however, it remains unclear whether this phenotypic convergence results from convergent evolution of genes (Stern, *Nat Rev Genet* 2013). On a molecular level, convergent evolution occurs when the amino acids of a specific protein preferentially undergo mutations that produce similar or even identical amino acid sequences within distinct evolutionary lineages. A high level of adaptive convergent evolution – that is, convergence due to positive selection of beneficial mutations – would suggest that some genes have "optimal configurations," which evolution uses and reuses across species. If such genes exist, then evolution is somewhat predictable, proceeding by one of a small number of possible paths (Stern & Orgogozo, *Science* 2009). In contrast, the complete absence of adaptive convergence would indicate that protein configurations beneficial to one species are seldom optimal within other species, and that evolution may proceed by a much wider set of paths. **Thus, two fundamental questions in evolutionary molecular biology are 1) how often convergent molecular evolution occurs and 2) whether molecular convergence is a primary driver of phenotypic convergence.** To adequately answer these questions, genome-wide analyses are needed; however, most previous analyses of convergent evolution have been limited to single genes (Li et al., *Curr Biol* 2010) or small taxa (Bazykin et al., *Biol Direct* 2007). Recently, the advent of whole-genome sequencing has opened new opportunities for exploring molecular convergence. Published genomes now exist for over 100 animal species, and 177 more are currently underway (Koepfli et al., *Annu Rev Anim Biosci* 2015). Additionally, phenotypes for 4,541 mammalian traits have been documented in a public database (O'Leary et al., *Cladistics* 2011). Together, these data enable a **genome-wide search for convergent mutations throughout mammalian species** followed by a **test for association between convergent genes and specific convergent phenotypes**, which might achieve strong statistical power for detecting adaptive convergence.

I plan to develop and apply a statistical model to test for convergent evolution among 61 sequenced mammalian species. *I hypothesize that convergent molecular evolution occurs at a higher rate than has been previously observed, and that this genetic convergence drives convergence of observable traits.* This model will extend the boundaries of biological knowledge by measuring the frequency of adaptive convergence, and will augment existing statistical methods in a broadly applicable framework that can be extended to other genomes in the future.

Aim 1: *Develop a statistical framework for inferring convergent evolution between genomes.* I will make my framework **broadly generalizable** to any data set with the following inputs: 1) a phylogeny containing the accepted evolutionary relationships within a set of species, and 2) the pairwise sequence alignments of all proteins in those species for which unique human orthologs exist. My analysis will integrate these data as follows:

- Step 1: *Infer ancestral sequences*. For every amino acid position in every sequence, I will employ a linear-time approach to compute probability distributions of DNA and amino acid sequences at each ancestral node in the phylogeny (Yang, *Mol Biol Evol* 2007).
- Step 2: *Infer patterns of adaptive convergent evolution between species pairs.* At the protein sequence level, I will apply a statistical analysis adapted from the CONVERGE algorithm described by Zhang & Kumar (*Mol Biol Evol* 1997). For every pair of species within our analysis, I will examine each amino acid and compute the probability that this amino acid converged between the two species, using the ancestral sequences from the previous step as

reference. I will then determine the likelihood of convergence within each exon and within the gene. Because rapidly evolving genes often converge by random chance, called neutral convergence, I will control for gene-specific mutation frequencies in order to distinguish adaptive convergence from neutral convergence (Thomas & Hahn, *Mol Biol Evol* 2015).

- Step 3: *Evaluate performance on simulated data*. I will simulate evolution of protein sequences in which a small pre-selected subset of sequences evolve under a non-neutral convergence pattern and the rest evolve neutrally. I will then test my model's precision and recall in inferring which of the sequences evolved under the convergent model, correcting for false discovery. If my model is able to detect evidence of convergent evolution in the simulated data at a low false discovery rate, but I observe no signs of convergent evolution in my analysis of the biological data, then these results will cast doubt on the hypothesis that non-neutral convergence is a significant driving force in molecular evolution.

Aim 2: ***Quantify the levels of convergent evolution within mammalian genomes, and identify functions enriched within convergent genes.***

- Step 1: *Identify specific genes that have evolved in convergence across species.* Using public alignment tools (Kent et al., *PNAS* 2003), the Bejerano Lab has obtained cross-species sequence alignments and a phylogenetic tree for 61 mammalian species. I will apply my statistical framework to all pairs of sufficiently diverged mammalian species and identify particular genes within each species-species pairing that exhibit non-neutral convergence. **I hypothesize that I will find widespread evidence of adaptive convergent evolution.**
- Step 2: *Identify potential phenotypic effects of convergent genotypes*. If evolutionary convergence is truly adaptive, this would suggest that genes converge when they undergo similar selective pressures within different species (Castoe et al., *PNAS* 2009). Therefore, **I hypothesize that genes facing similar pressures within independent species will be more likely to evolve in convergence within those species**. For example, in aquatic mammals such as dolphins, manatees, and walruses, we might expect high convergence of skin-expressed genes involved in thermoregulation. I have identified a set of 20 convergent phenotypes that arose independently in mammals. For each of these phenotypes, I will conduct Gene Ontology enrichment analysis on genes that show elevated levels of convergence among the animals possessing the phenotype in question. As a null model, I will apply the same enrichment tests within groups of species that do not exhibit phenotypic convergence. If my hypothesis is correct, this analysis will suggest functions for which similar selective pressures across species have necessitated similar courses of protein evolution across these species. Importantly, in order to confirm these putative links between convergent genotypes and convergent phenotypes, future experiments will be necessary. The specific amino acid substitutions that I identify will be prime targets for further examination in functional assays, as described by Liu et al. (*Mol Biol Evol* 2014).

My experience developing statistical models to quantify bias in genome-wide DNase-seq experiments (Gloudemans, 2015 Honors Thesis) makes me well-qualified to develop a genome-wide statistical model of evolution. This project will help us to understand whether phenotypic convergence is a direct result of genotypic convergence. To further facilitate **broader impacts**, I will create a user-friendly visualization that highlights hotspots of convergent evolution, and I will make this tool publicly available in the UCSC genome browser. This tool will allow biologists to visually explore instances of molecular convergence and to ask even deeper questions about the specific functional roles of these convergent mutations, ultimately increasing our understanding of how these proteins shape the lives of humans and all other species.