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 is often 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* 2008). In contrast, a 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 and in which genes convergent molecular evolution occurs and 2) whether molecular convergence is a primary driver of phenotypic convergence.** Most 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); to date, no genome-wide analysis involving a wide variety of species has been completed. 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). Thus, a genome-wide search for convergent mutations across multiple species is now possible, and might reveal new evidence of adaptive molecular convergence.

I plan to develop and apply a novel computational framework 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 framework will extend the boundaries of biological knowledge by quantifying the frequency of convergent evolution, and will augment existing phylogenetic methods in a broadly applicable framework that can be extended to other genomes in the future.

Aim 1: Develop a novel computational framework for identifying convergent evolution between genomes. My framework will be generalizable to any data set with the following inputs: 1) a phylogenetic tree for a set of species, and 2) the pairwise sequence alignments of all proteins in those species. I will limit my analysis to genes unambiguously alignable to a one-to-one human ortholog in at least 80% of mammalian genomes. I have verified that even after filtering on these criteria, >50% of all human genes are retained. My analysis will integrate these data as follows:

- Step 1: Infer ancestral sequences. For every amino acid in every sequence, I will compute amino acid sequences at each ancestral node in the phylogeny using the linear-time maximum parsimony method implemented in PAML 4 (Yang, *Mol Biol Evol* 2007).
- Step 2: Infer patterns of adaptive convergent evolution between species pairs. For every pair of species within our analysis, for each gene, I will identify the set of amino acids that converged in that pair of species with probability >0.9, as well as those that diverged with probability >0.9, based on the ancestral sequences inferred in Step 1. I will report a

convergence score for the gene in this pair of species: the ratio of convergent mutations to divergent mutations. Thus, genes with high rates of both divergent and convergent mutations (such as rapidly evolving immune genes) will score lower than those with relatively higher rates of convergent mutations. After computing the full distribution of convergence scores for every gene in every species pair, I will denote the upper outliers as *convergence events*.

- Step 3: Evaluate model performance on simulated data. I will simulate evolution of protein sequences in which a small fraction of the sequences evolve under a non-neutral convergence pattern and the rest evolve neutrally. I will measure my framework's precision and recall in inferring which sequences evolved under the convergent model. If my framework is able to detect convergence events in the simulated data at a low false discovery rate, but observes no convergence events in the biological data, then these results will cast doubt on the hypothesis that adaptive 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 genes evolving in convergence.

- Step 1: Identify specific genes that have evolved in convergence across species. Using public alignment tools (Kent et al., *PNAS* 2003), my collaborators in the Bejerano Lab at Stanford have provided cross-species sequence alignments and a phylogenetic tree for 61 mammalian species. I will apply my framework from Aim 1 to all pairs of sufficiently diverged mammalian species and identify genes within each species-species pairing that exhibit non-neutral convergence. **I hypothesize that I will find evidence of adaptive convergent evolution within many genes and within almost all species-species pairings.**
- Step 2: Identify convergent genotypes potentially responsible for known convergent phenotypes. If evolutionary parallelism 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 and manatees, we might expect high convergence of skin-expressed genes involved in thermoregulation. Indeed, my preliminary analysis has identified 4 convergent mutations in the gene TGM1 in dolphins and manatees, a significantly greater number than expected by chance. Human mutations in this gene cause a skin disease called ichthyosis, characterized by fish-like, scaly skin (Laiho et al., *AJHG* 1997). In order to find similar cases, I have identified a set of 20 convergent phenotypes that arose independently in mammals, such as adaptation to high altitudes or dry environments. For each of these phenotypes, I will identify genetic convergence events that may be responsible for the phenotype in question. 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. The specific amino acid substitutions that I identify will then be prime targets for further examination in functional assays, as described by Liu et al. (*Mol Biol Evol* 2014).

Broader Impacts: This project will help us to understand whether phenotypic convergence is a direct result of genotypic convergence. I will create a publicly available, user-friendly visualization for the UCSC Genome Browser that highlights hotspots of convergent evolution. This tool will allow biologists to visually explore instances of adaptive molecular convergence and to ask even deeper questions about the specific functional roles and phenotypic effects of these convergent mutations.