RESEARCH INTERESTS

POPULATION GENOMICS OF TRANSCRIPTIONAL REGULATION

Selective, demographic, and random processes all determine the frequency of alleles in a population and differences between species. One of the major goals of population genetics has been to uncover which of these processes is acting in natural populations through a combination of directed empirical studies and theoretical models that provide expectations under a variety of conditions. While most of the work in the field has involved single loci or limited multiple locus studies and models, the availability of genomic-scale data will begin to require new genomic-scale approaches. Within the next year a population of whole genomes will be sequenced; the approach population genetics takes now may determine how soon this data becomes informative and what information it gives us.

My research program is aimed at developing the empirical, computational, and statistical tools necessary for studying variation in whole genomes in an evolutionary context. More specifically, I am studying the evolution of *cis*-regulatory sequences—the DNA necessary for directing the time, level, and place of transcription of protein-coding genes. The sequencing of whole genomes and a growing number of studies into *cis*-regulatory variation have shown that the effects of natural selection reach far beyond the start and stop codons. Through a combination of directed empirical studies, new computational techniques, and improved statistical tools, my goal is to contribute to an understanding of the role *cis*-regulatory variation plays in evolution.

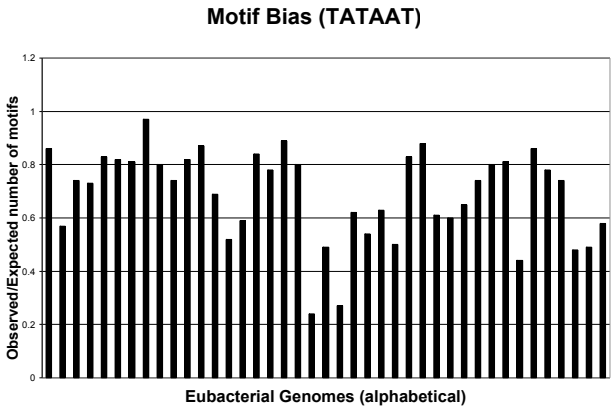
ONGOING RESEARCH

My current research can be divided into three main areas: computational approaches to studying populations of genomes, the development of models for statistical inference of natural selection, and empirical studies of *cis*-regulatory variation in humans. Below I briefly describe my prior research in these areas and the future work I aim to conduct.

Computational Biology

The availability of whole genome sequences means that we can examine not just natural selection on thousands of individual genes, but also at the level of the genome itself. Because cells must regulate the transcription of suites of genes expressed together but located throughout the genome, and because transcription factors may control hundreds of target genes expressed at different times and places, selection for improved transcriptional efficiency may have its basis in selection throughout the whole genome. Transcription factor binding sites are very short (generally 6-10 base pairs), and thus binding site sequence motifs are expected to appear frequently throughout a genome. Frequent creation of new binding sites may be an important way in which novel

transcriptional patterns evolve, but it may also introduce a large degree of noise into the efficient functioning of a cell. I hypothesized that the binding of transcription factors to spurious binding sites—the correct sequence of nucleotides in inappropriate genomic locations—could drive natural selection to eliminate binding site motifs from a genome. Using available computational tools, as well as creating new tools, I studied the effects of selection on 52 whole genomes in Eubacteria and Archaea (Figure 1). Examining the sequence motifs necessary for polymerase binding in each group of organisms, I found that spurious binding sites appear less frequently than expected under a random model in every genome but one (Hahn et al. 2003). It appears that both functional and non-functional sequences are constrained to avoid mutating to binding site motifs. In addition, I developed a model of binding site evolution that allows an estimate of the strength of selection against binding sites. Selection intensity appears to be weak, similar to that of codon bias.



While the finding that spurious binding sites are significantly under-represented in Eubacteria and Archaea represents a novel discovery of selection at the level of the whole genome, we would also like to know whether it is a general phenomenon found in more complicated, eukaryotic genomes. In my postdoctoral research I am extending these studies of motif bias to eukaryotic genomes, where heterochromatin, gene-rich and gene-poor regions, recombination, and multiply-represented binding sites greatly complicate both the effects and detection of natural selection. These studies will greatly expand our understanding of natural selection, patterns of variation, and the regulation of transcription.

Statistical Inference

Understanding the roles of selection, mutation, and drift in shaping within species variation requires that we have explicit population genetic models and predictions. It is through these models that we can then make statistical inferences about the forces acting on genes or genomes. A major part of my research, therefore, aims to develop population models appropriate for the analysis of both individual genes and whole genomes.

In addition to creating new computational tools for studying selection against spurious transcription factor binding site motifs, I have been working to develop statistical models of motif bias among a population of genomes. Using analytical and simulation techniques, I have shown (Hahn and Rausher, *submitted*) that individuals will vary in the number of any particular nucleotide motif found in their genomes simply because of

single nucleotide polymorphisms. Natural selection can be added to the same analyses to show both that the mean number of motifs in the population is lowered and that variation is lost among individuals. I have subsequently used the results of this model to create a likelihood ratio test for the action of natural selection on any particular binding site motif. I am currently working on new models of motif bias to relax some of the necessary assumptions of the first model (such as no recombination), and to incorporate additional features of *cis*-regulatory evolution.

For the analysis of individual genes, use of coalescent genealogies has made for more efficient and more precise statistical inference. However, both demographic and various selective mechanisms will cause significant deviations from the neutral model. In order to distinguish demographic from selective effects, and among different selective mechanisms, I developed an improved statistical test based on the coalescent (Hahn et al. 2002). The method uses coalescent simulations to test for differences in frequency distributions between neutral and selected mutations (such as synonymous and nonsynonymous coding sequence mutations). Application of this test to published data has revealed hidden instances of natural selection on coding and *cis*-regulatory sequences not detected before (Hahn, *in prep*), and has enabled researchers to distinguish between multiple evolutionary processes (Schaeffer 2002; McDonald and Shaw, *in press*; Mes, *in press*; Balhoff and Wray, *in prep*.).

Regulatory Variation

In order to understand the role of natural selection, it is also important to study the individual mutations with phenotypic effects that are visible to selection. In an ongoing study of functionally characterized *cis*-regulatory polymorphisms in humans and the other primates, I have been collecting sequence data among human populations and from chimpanzee, bonobo, gorilla, orangutan, and baboon. This work focuses on binding sites within a regulatory region that have been shown in biochemical studies to contain polymorphisms that significantly affect transcription levels in humans. Working with Dr. David Goldstein at University College London and Dr. Gregory Wray at Duke University, we continue to find instances of population-specific selection (Rockman et al. *in press*) and selection in human history (Hahn et al. *in prep*) on *cis*-regulatory mutations. The enormous amount of functional *cis*-regulatory polymorphism in humans implies that there is not one, static regulatory network seen by natural selection. This work will have important implications for both the evolution of modern humans and for the evolution of transcription factor:DNA interactions.

To further characterize the evolution of transcription factor binding sites in humans, Dr. Wray and I are currently developing a silicon chip-based technology that will enable us to look at the DNA binding properties of human transcription factors. By modifying existing technologies, we believe that this method will allow us to not only find the binding sites bound by human transcription factors, but will also allow us to detect changes in binding affinity among the primates and along the human lineage.

FUTURE WORK

In addition to pursuing the research topics outlined above, I also have a number of additional studies planned that will complement this research. One major study aims to look at variation in genome sequences among individuals of the mosquito, *Anopheles gambiae*, an important vector of human disease. With the availability of a population of genomes, we will be able to examine the patterns of variation and selection at every gene and across the genome. I am collaborating as the primary population geneticist on a project to sequence multiple genomes of *A. gambiae* with a group of researchers from the European Molecular Biology Laboratory, the European Bioinformatics Institute, and the University of Notre Dame. We have already collected and begun to analyze data from 1x coverage of a second *A. gambiae* genome. On a finer scale, I have begun a project with mosquito researchers here at UC-Davis to look at polymorphism and divergence in a number of *A. gambiae* loci thought to be important for disease resistance.

Understanding the enormous amount of data generated by genome sequencing and functional genomics projects will require an evolutionary framework. In the next few decades the field of biology will no longer be data-limited; the limiting factor will be in creating tools necessary for analyzing data in an evolutionary context. My research aims to both create and facilitate the use of evolutionary biology and population genetics in the analysis of whole genome data, and in so doing to stay at the forefront of biological discovery. I believe that a vibrant research laboratory comes through having a mix of people with different backgrounds and expertise. In order to maintain an active lab that produces work with broad implications for the field, I will continue to support work that uses empirical, statistical, and computational approaches to understand the evolution of diversity.