Abstract:
For decades, it has been hypothesized that gene regulation has played a central role in human evolution, yet much remains unknown about the genome-wide impact of regulatory mutations. In this talk, I will describe work my group has been pursuing over the last few years to better characterize the evolution of gene regulatory elements in humans and nonhuman primates, based on patterns of polymorphism and divergence in complete genome sequences. First, I will review a new probabilistic method, called INSIGHT, that we developed to measure the influence of selection on collections of short, interspersed noncoding elements across the genome. Using INSIGHT, we showed that natural selection has profoundly influenced human transcription factor binding sites since the divergence of humans from chimpanzees 4-6 million years ago, and that regulatory elements contribute substantially to both adaptive substitutions and deleterious polymorphisms, with key implications for human evolution and disease. Next, I will describe how we have adapted the INSIGHT framework for use in estimating the probability that a point mutation at each position in a genome will influence fitness. These fitness consequence (fitCons) scores serve as evolution-based measures of potential genomic function. We have generated fitCons scores for three human cell types based on public data from ENCODE. Compared with conventional conservation scores, fitCons scores show considerably improved prediction power for cis-regulatory elements. In addition, they indicate that 4.2–7.5% of nucleotides in the human genome have influenced fitness since the human-chimpanzee divergence, and they suggest only modest impact from recent evolutionary turnover on the functional content of the genome. Finally, I will describe our recent work on extending fitCons to accommodate much larger collections of genomic covariates. These methods are based on an information theoretic measure of the additional information obtained from each candidate functional genomic data type, which serves as an objective function in a hierarchical divisive clustering algorithm. Preliminary results indicate that these methods substantially improve the genomic resolution and predictive power of the fitCons scores.