Title: Fast, scalable inference of fitness consequences for noncoding mutations

Yi-Fei Huang – yihuang@cshl.edu

Authors: Yi-Fei Huang [1], Brad Gulko [1,2], Adam Siepel [1]

Affiliations:
1. Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA
2. Graduate Field of Computer Science, Cornell University, Ithaca, New York, USA

Genome-wide association studies and evolutionary studies suggest that a large proportion of variants important to human phenotypes, diseases, and evolutionary adaptation are located in noncoding regions. However, methods for identifying causal noncoding variants important to human evolution and diseases have only achieved limited success. Here we report a scalable framework, LINSIGHT, for inferring fitness consequences of noncoding mutations, i.e., the probabilities that the mutations are under natural selection, by integrating a variety of comparative and functional genomic data. This novel framework is based on a previously developed model, INSIGHT, which was used in the fitCons framework to infer fitness consequences of noncoding mutations in human genome. As a generalized linear regression model motivated by the evolutionary model in INSIGHT, LINSIGHT significantly improves the scalability of the fitCons method and allows the integration of vastly larger numbers of genomic features. By integrating a large number of conservation scores and functional genomic data, LINSIGHT provides a high resolution map of the fitness consequences of mutations in the human noncoding genome. Unlike most existing methods, our framework is explicitly defined in evolutionary terms and integrates different functional annotations based on the principles of evolution. Using noncoding Mendelian disease variants in the ClinVar and HGMD databases, we show that our new method performs favorably compared to the state-of-the-art in the prioritization of disease variants. Furthermore, we show that pleiotropy and target genes associated with enhancers are important determinants of the evolutionary constraints on enhancers, which implies that the evolution of enhancer sequences is highly context dependent. Therefore, LINSIGHT is a powerful tool not only for prioritizing noncoding disease variants but also for estimating the strength of natural selection acting on noncoding sequences.