

Identification of functional SNPs in noncoding regions of the human genome

Daniel C Koboldt, Raymond D Miller

Department of Genetics, Washington University School of Medicine, 4566 Scott Avenue, Box 8232, St. Louis, MO 63110, United States of America

Single nucleotide polymorphisms (SNPs) are the most abundant form of DNA sequence variation in humans. As the number of reported SNPs in public databases continues to grow, identifying putative functional variants has become an important goal of human genetics. While the majority of known disease-causing mutations are nonsynonymous SNPs, it is critical to realize that a substantial portion of functional sequence variation lies outside the exons of protein-coding genes. We developed an approach to systematically identify and evaluate SNPs with possible functional relevance in the human genome. First, we constructed a refined set of single-base, bi-allelic, uniquely mapped SNPs from public databases. Our set contained over 9 million SNPs, of which 57,356 were classified as nonsynonymous. To expand the pool of putative functional variants, we mapped SNPs to gene structures as well as annotated regions of functional significance, including transcription factor binding sites, exon splicing enhancers or silencers, non-protein-coding RNA genes, regulatory-potential sequences, microRNA binding sites, and human-rodent conserved or human-vertebrate conserved elements. For each class of putative functional SNP, we examined the allele frequency distribution and population-specificity in using data from the International HapMap Project (results in SNPseek database available at <http://snp.wustl.edu>). As expected, nonsynonymous SNPs exhibited significantly decreased allele frequencies and increased population-specificity consistent with purifying selection. Intriguingly, we also observed this pattern of excess rare variation among noncoding SNPs that map to transcription factor binding sites, microRNA binding motifs, regulatory-potential sequences, and splice sites. In fact, we found that substitutions in 3' UTR microRNA binding sites or intronic splice sites were more likely to be deleterious than nonsynonymous substitutions. Our results demonstrate the effectiveness of allele frequency and population-based tests to infer selection against a set of SNPs, and also highlight the fact that noncoding SNPs are important contributors to the functional variation of the human genome.